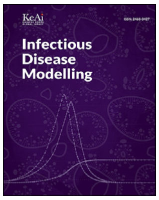
Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm



Information-guided adaptive learning approach for active surveillance of infectious diseases

Qi Tan [a, b], Chenyang Zhang [a, b], Jiwen Xia [a, b], Ruiqi Wang [e], Lian Zhou [f], Zhanwei Du [c, d], Benyun Shi [a, b, *]

[a] *College of Computer and Information Engineering, Nanjing Tech University, Nanjing, Jiangsu Province, China*
[b] *College of Artificial Intelligence, Nanjing Tech University, Nanjing, Jiangsu Province, China*
[c] *WHO Collaborating Center for Infectious Disease Epidemiology and Control, School of Public Health, The University of Hong Kong, Hong Kong SAR, China*
[d] *Laboratory of Data Discovery for Health Limited, Hong Kong SAR, China*
[e] *Faculty of Arts and Social Sciences, Hong Kong Baptist University, Hong Kong SAR, China*
[f] *Jiangsu Provincial Center for Disease Control and Prevention, Nanjing, China*

ARTICLE INFO

Article history:
Received 12 July 2024
Received in revised form 1 October 2024
Accepted 24 October 2024
Available online 26 October 2024
Handling Editor: Dr Yiming Shao

Keywords:
Active surveillance
Adaptive learning
Incomplete data
Information guide

ABSTRACT

The infectious disease surveillance system is a key support tool for public health decision making. Current research concentrates on optimizing static sentinel deployment to address the problem of incomplete data due to the lack of sufficient surveillance resources. In this study, we introduce an information-guided adaptive learning strategy for the dynamic surveillance of infectious diseases. The goal is to improve monitoring effectiveness in situations where it is possible to adjust the focus of surveillance, such as serial surveys and allocation of testing tools. Specifically, we develop a probabilistic neural network model to learn spatio-temporal correlations among the numbers of infections. Based on a probabilistic model, we evaluate the information gain of monitoring a spatio-temporal target and design a greedy selection algorithm for monitoring targets selection. Moreover, we integrate two major surveillance objectives, i.e., informativeness and coverage, in the monitoring target selection. The experimental results on the synthetic dataset and two real-world datasets demonstrate the effectiveness of our approach, showcasing the promise of further exploration and application of dynamic adaptive active surveillance.

1. Introduction

Infectious disease surveillance is critical to evidence-based public health decision making (Organization et al., 2022; Zeng et al., 2021). Through the continuous collection of health-related data and the dissemination of epidemic status information, surveillance systems benefit numerous evidence-based decision-making processes, including early detection and response to outbreaks, identifying the priorities of prevention strategies and allocation of health resources (Groseclose & Buckeridge, 2017; Jamison et al., 2006). Surveillance systems employ two main surveillance strategies: passive surveillance and active

[*] Corresponding author. College of Computer and Information Engineering, Nanjing Tech University, Nanjing, Jiangsu Province, China.
E-mail address: benyunshi@outlook.com (B. Shi).
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2024.10.005

surveillance. Passive surveillance mechanisms accumulate data from those who come to them, such as hospital visits of patients and disease reports forwarded by physicians to a central authority (Longbottom et al., 2021). However, passive surveillance systems often suffer from the limitations of under-reporting and delays, which are intricately intertwined with healthcare-seeking patterns and accessibility to medical services (Sarti et al., 2016; Vitale et al., 2020). This issue is particularly pronounced in distant, mountainous, and developing areas. To seek complete and timely information, active surveillance systems adopt a proactive approach, collecting data from specific subsets of the target population. Typical examples of active surveillance practice include sentinel surveillance, health survey, and testing program. An example lies in the malaria disease, which has posed a persistent threat to various countries over the decades. Through sustained long-term efforts, many countries are advancing toward the phase of malaria elimination (Cao et al., 2021). At this critical stage, border screening becomes significantly important for detecting and containing imported infection. Given that these screenings require substantial resources and labor, it is suggested to prioritize high-risk immigrants (Khamsiriwatchara et al., 2011; Sturrock et al., 2015). Another example is the COVID-19 pandemic. During the recent COVID-19 pandemic, targeted active testing of specific subpopulations emerged as an effective strategy to identify infections and mitigate the spread of the virus (Li et al., 2020; Litwin et al., 2022). Although active surveillance offers the benefit of collecting high-quality data, it often comes with significant costs. It is challenging to secure enough resources to maintain extensive surveillance efforts in different regions or large populations (Pei et al., 2021; Polgreen et al., 2009).

There is increasing interest in optimizing surveillance planning to collect adequate data with limited resources. Current research focuses primarily on static sentinel selection, which involves continuously collecting data from predetermined sentinel locations. These studies are approached from two perspectives: coverage and informativeness. In a notable work focusing on the coverage perspective, Polgreen (Polgreen et al., 2009) proposed a maximal coverage model to optimize the selection of sentinel networks for influenza-like illnesses. This model aims to maximize the number of people located within a certain geographic region around the chosen sites, thereby enhancing the effectiveness of the surveillance network. Research focusing on the second aspect (that is, informativeness) highlights the ability to forecast the state of unmonitored epidemics through data collected from specific sentinel sites (Pei et al., 2021; Scarpino et al., 2012). Pei et al. developed a strategy for selecting sentinel sites in infectious disease monitoring employing a group sparsity technique (Pei et al., 2020). This method aims to identify key nodes within a transmission network to facilitate inference of epidemic statuses in unmonitored locations. Gaussian process model is extensively used in non-parametric statistical spatial modeling (Cressie, 2015; Laurent & Cowlagi, 2023). Leveraging Gaussian process models, classic tools from information theory are employed to assess the importance of monitoring specific spatial locations, such as utilizing entropy (Cressie, 2015), cross-entropy (Wang et al., 2004), and mutual information (Krause et al., 2008).

In addition to static sentinel surveillance, dynamically adapting target sites is a viable approach in numerous health surveillance activities, such as serial health surveys (Losos, 1996), testing kits resource allocation (Buhat et al., 2021) and vector monitoring (Case et al., 2024). Temporal correlations are inherent in epidemic dynamics, allowing the prediction of infection cases in one subpopulation based on its recent infection data. Consequently, gathering data from other subpopulations can provide significant benefits in these contexts, enhancing the overall understanding and prediction of the spread of the epidemic. Du et al. proposed a dynamic sensing strategy that employs matrix completion techniques to reconstruct the entire monitoring data matrix using data from only a few spatio-temporal monitoring points (Du et al., 2013). This strategy has been applied to develop a city traffic monitoring system using a limited number of mobile scout vehicles. However, this strategy is primarily based on passive monitoring and does not actively determine target sites. By dynamically adapting monitored sites, it is possible to significantly enhance monitoring effectiveness, as this allows the collection of important data from critical areas at key times. Despite these potential benefits, current active surveillance strategies often do not explore adaptive monitoring scenarios, which is hindered by a series of challenges:

- How to capture the spatio-temporal correlations within the epidemic dynamics based on incomplete data? When inferring the full scope of an epidemic based on partial observation, it is crucial to utilize the correlations among infection data in both spatial and temporal dimensions. However, in situations where resource are constrained, surveillance data often have missing elements, posing great challenges in learning spatio-temporal correlations.
- How to evaluate the gain of gathering spatio-temporal data? It requires quantitative measures to assess how the acquisition of infection data improves predictability and fulfills other surveillance objectives. These metrics are crucial in determining the effectiveness of data collection strategies, yet they are difficult to design and calculate.
- How to optimally select a set of observations and accurately infer transmission states? It is imperative to develop a selection algorithm capable of identifying mutually beneficial surveillance targets from a wide range of options at each decision juncture, thus maximizing the efficiency of surveillance efforts.

In addition, to evaluate the entire epidemic situation, a reliable inference procedure is required to utilize data from varying surveillance targets to infer the unobserved infection.

In this paper, we introduce an information-guided adaptive learning approach that offers a potential solution to address the aforementioned challenges. Specifically, considering the prevalent presence of missing elements in spatio-temporal data in surveillance planning contexts, we first design a deep spatio-temporal model with latent probabilistic variable to handle the inference uncertainty stemming from noisy incomplete data. Second, we assess the information gain of spatio-temporal data points, combining this analysis with the extent of infection coverage to ascertain the value of collecting these data samples. This method allows for a more strategic approach to data collection, focusing on the most informative and impactful data points. Finally, contrasting with current approaches in sentinel station selection, our proposed approach dynamically evaluates the utility of spatio-temporal data points and adjusts its inferences about infection risks based on the observation from varying target sites. By adaptively modifying monitor targets, our approach aims to enhance the efficiency and accuracy of infectious disease surveillance systems.

The remainder of this article is organized as follows. Section 2 formulates the surveillance selection problem and associated computational challenges mathematically. Section 3 elaborates the details of our proposed computational approach, including learning with incomplete data, target selection, and epidemic inference, tailored to address the aforementioned computational challenges correspondingly. Section 4 presents extensive experimental results on a synthetic dataset and two real-world datasets to validate the effectiveness of our proposed approach. Section 5 concludes and discusses this article.

2. Epidemic active surveillance

We first provide a formal description of the epidemic active surveillance problem, followed by an overview of the computational challenges associated with it. Consider an epidemic spread among N groups of subpopulations. Let matrix $\mathbf{O}_{t-L:t} \in \mathbb{R}^{L\times N} = [\mathbf{O}_{t-L};\ldots;\mathbf{O}_t]$ be the observation of epidemic dynamics during a time window of time length L. Specifically, for a time slice, the row vector is constructed as $\mathbf{O}_{t_1} = [\mathbf{O}_{t_1,1};\ldots;\mathbf{O}_{t_1,N}]^T$, where the element $\mathbf{O}_{t_1,i}$ denotes the observation on subpopulation i at time t_1. Due to limited surveillance capacity, some of the entries in $\mathbf{O}_{t-L:t}$ are empty. Let $\mathbf{M}_{t-L:t}$ denote the observing indicator matrix, where $\mathbf{M}_{t_1,i} = 1$ if $\mathbf{O}_{t_1,i}$ is not empty, and 0 otherwise. The active disease surveillance planning can be formulated as follows: given the data $\mathbf{O}_{t-L:t}$, select K target subpopulations to surveillance and infer the number of infection cases in the remain subpopulations:

$$\mathbf{O}_{t+1,\sim\mathbb{S}_{t+1}} = f(\mathbf{O}_{t-L:t}, \mathbf{O}_{t+1,\mathbb{S}_{t+1}}). \tag{1}$$

where $\mathbb{S}_{t+1}$ denotes the set of select target subpopulations and $\sim\mathbb{S}_{t+1} = \{\forall i \in (1:N) \text{ and } i \notin \mathbb{S}_{t+1}\}$. For simplicity of notion, we omit the subscript of $\mathbb{S}_{T+1}$ in the remainder of the paper. Note that the entire epidemic period is partitioned into multiple time windows, and we make surveillance decisions for every time window. In order to realize adaptive surveillance, as outlined in the three challenges in Section 1, there exist three consecutive computational tasks:

(T1) Learning with incomplete data: How to extract the spatial-temporal interaction based on incomplete data on epidemic dynamics $\mathbf{O}_{t-L:t}$?
(**T2**) Target selection for active surveillance: How to select the surveillance targets according to a given budget K to maximize the effectiveness of the monitoring?
(**T3**) Epidemic inference: How to infer the unobserved dynamics $\mathbf{O}_{t+1,\sim\mathbb{S}_{t+1}}$ from the available surveillance data $\mathbf{O}_{t-L:t}$, $\mathbf{O}_{t+1,\mathbb{S}_{t+1}}$?

3. Proposed approach

In this section, we introduce our proposed approach, which consists of three key modules, designed to tackle the computation tasks **T1**, **T2**, and **T3** in the context of active surveillance. We start with a general overview of the approach, followed by a detailed description of each module throughout the section.

3.1. Overview

Fig. 1 illustrates the overall design of our proposed approach. We designate our proposed approach as Information-Guided Dynamic Active Surveillance (IGDAS). Firstly, to capture spatio-temporal dynamics in task **T1**, IGDAS learns a deep neural network with latent probabilistic variable using incomplete data to capture the correlations between the infection case numbers in different subpopulations over time. Subsequently, with regard to task **T2**, IGDAS evaluates the inference effect of monitoring specific epidemic infection data on other spatio-temporal epidemic cases of the target, referred to spatio-temporal information gain, based on the spatial-temporal correlations learned in the first step. Then, IGDAS selects a set of monitoring points at every decision time step, considering both the information gain and coverage in a greedy manner.

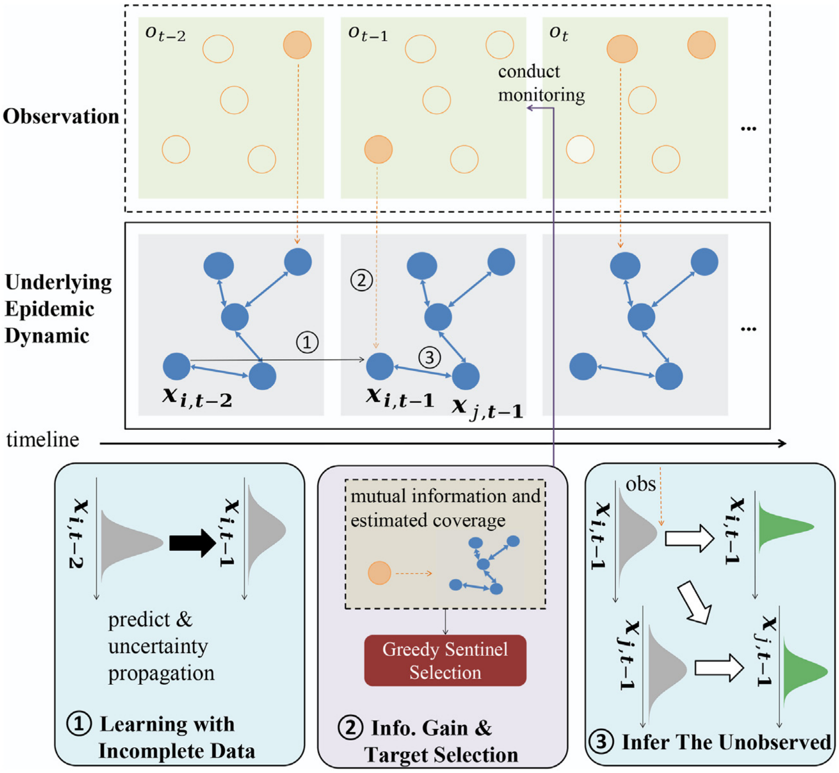


Fig. 1. An overview of our proposed approach, IGDAS, involves a structured process consisting of three key steps: (1) learning a probabilistic deep neural network to capture relationships between spatial-temporal infection numbers with incomplete data, (2) evaluating the information gain of collecting specific data and optimizing surveillance targets over time to enhance effectiveness with considering both informativeness and coverage, as well as (3) inferring the uncovered infection numbers based on monitoring data.

Finally, in task **T3**, IGDAS uses a conditional Gaussian model to infer the number of infections in unobserved sub-populations based on monitoring data.

3.2. Learning with incomplete data

We develop a variational neural network model with latent probabilistic variable to learn the spatio-temporal dependency (Kingma & Welling, 2013):

$$\begin{aligned}[\boldsymbol{\mu}_{t+1},\boldsymbol{\sigma}_{t+1}] &= g(\widehat{\mathbf{O}}_{t-L:t},\theta),\\ \mathbf{z}_{t+1} &\sim \mathbb{N}(\boldsymbol{\mu}_{t+1},\boldsymbol{\sigma}_{t+1}),\\ \widehat{\mathbf{y}}_{t+1} &= W\mathbf{z}_{t+1},\end{aligned} \tag{2}$$

where θ is the parameter of the neural networks encoder, $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\sigma}_{t+1}$ is the mean and variance of the Gaussian latent variables, and $\widehat{\mathbf{y}}_{t+1}$ is the estimation output. Through the latent probabilistic variable $\mathbf{z}_{t+1}$, we represent the belief in the underlying states of epidemic diffusion inferred from the incomplete and noisy data.

Real-world monitoring data often contain measurement missingness and errors, which pose significant challenges to model training and prediction. To address the issues caused by data noise and missingness, we utilize data masking, which is a common practice in deep learning techniques. It is common and simple to fill the missing data with zero in deep learning practice (Wang et al., 2019). As there are some observations 0 in the epidemic data, we fill the missing entries with -1 in the model input feature $\widehat{\mathbf{O}}_{t-L:t} = \mathbf{M}_{t-L:t} \circ \mathbf{O}_{t-L:t} + (1 - \mathbf{M}_{t-L:t})* - 1$, where $\circ$ is the Hadamard product. Moreover, the learning objective is calculated with respect to the observed data only:

$$\mathrm{loss}_p = \sum_t \|\mathbf{M}_{t+1} \circ \widehat{\mathbf{y}}_{t+1} - \mathbf{M}_{t+1} \circ \mathbf{O}_{t+1}\|. \tag{3}$$

We learn the model parameters via minimizing loss_p with stochastic gradient variational Bayes technique (Kingma & Welling, 2013).

3.3. Information gain and surveillance target selection

With respective to surveillance target selection, our considerations are twofold: firstly, we aim for the chosen data to be insightful in deducing the transmission status within the unobserved subpopulations; secondly, we aim to cover more epidemic infections in order to monitor and control the diffusion of infections. To address the first objective, we use the information criterion. When monitoring a data point, the uncertainty about the hidden true variables diminishes. Based on the correlations between the variables, the uncertainty of other variables in the epidemic diffusion system can also effectively decrease. IGDAS evaluates the reduction in overall system uncertainty following the observation of specific data points, using it as an assessment of the importance of that spatio-temporal data. Specifically, we utilize mutual information between specific monitoring variables and the remaining variables to measure their information gain. The estimated prior covariance matrix is employed for mutual information estimation.

Specifically, the joint distribution of $\mathbf{O}_{t+1}$ is a multivariate Gaussian distribution $\mathbf{O}_{t+1} \sim \mathbb{N}(\mathbf{W}\boldsymbol{\mu}_{t+1},\ \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{W}\mathrm{diag}(\sigma_{t+1})\mathbf{W}^T$ and $\mathrm{diag}(\sigma_{t+1})$ denotes the diagonal matrix with the main diagonal as σ_{t+1}. The information gain at a spatio-temporal point is expressed as:

$$I(\mathbf{O}_{t+1,j}; \mathbf{O}_{t+1,\sim j}) = \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_{j,j}||\boldsymbol{\Sigma}_{\sim j,\sim j}|}{|\boldsymbol{\Sigma}|}\right), \tag{4}$$

where $\boldsymbol{\Sigma}_{r,c}$ denote the submatrix of $\boldsymbol{\Sigma}$ index by row index r and column index c. Similarly, we further derive the information gain when a selected spatio-temporal point set is already being considered from conditional mutual information $I(\mathbf{O}_{t+1,j}; \mathbf{O}_{t+1,\sim[j,\mathbb{S}^k]}|\mathbf{O}_{t+1,\mathbb{S}^k})$, where $\mathbb{S}^k$ is the set of selected subpopulations up to the k_{th} step, $k = 1, \ldots, K$. Let $\mathbf{c}$ denote the combined vector of the conditional mutual information value of the candidates.

For the second objective, we utilize the estimated mean to assess the infection risk $\mathbf{r} = \mu_{t+1,\sim\mathbb{S}^k}$. To combine the above two metrics, we normalize them into range $[0, 1]$ (denoted as $\bar{\mathbf{c}}$ and $\bar{\mathbf{r}}$ respectively) and calculate the weighted metrics as follow:

$$\mathbf{s} = a_1 * \bar{\mathbf{c}} + a_2 * \bar{\mathbf{r}}. \tag{5}$$

We use the greedy algorithm to select K surveillance targets.

3.4. Epidemic inference

After monitoring by the sentinel section, we process to infer the unobserved subpopulations. As the joint distribution of number of infections case $\mathbf{O}_{t+1}$ is a multivariate Gaussian distribution, with obtaining the observation of the selected sentinel $\mathbf{O}_{t+1,\mathbb{S}}$, the posterior distribution of the epidemic dynamic on the unobserved subpopulations is written as a conditional Gaussian distribution:

$$\begin{aligned} \bar{\mathbf{y}}_{t+1,\sim\mathbb{S}} &\sim \mathbb{N}(\widehat{\mu}, \widehat{\boldsymbol{\Sigma}}), \\ \widehat{\boldsymbol{\mu}} &= \widehat{\mathbf{y}}_{t+1,\sim\mathbb{S}} + \boldsymbol{\Sigma}_{\sim\mathbb{S},\mathbb{S}}\boldsymbol{\Sigma}_{\mathbb{S},\mathbb{S}}^{-1}(\mathbf{O}_{t+1,\mathbb{S}} - \widehat{\mathbf{y}}_{t+1,\mathbb{S}}), \\ \widehat{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{\sim\mathbb{S},\sim\mathbb{S}} - \boldsymbol{\Sigma}_{\sim\mathbb{S},\mathbb{S}}\boldsymbol{\Sigma}_{\mathbb{S},\mathbb{S}}^{-1}\boldsymbol{\Sigma}_{\sim\mathbb{S},\mathbb{S}}^T. \end{aligned} \tag{6}$$

Through the probabilistic modeling, we are enable to access the uncertainty of the estimation. As observed from the last line of Equation (6), compared to prior covariance, the magnitude of posterior covariance could be reduced with the information derived from observation $\mathbf{O}_{t+1,\mathbb{S}}$. Moreover, the mean estimation with respect to the observed subpopulations $\widehat{\boldsymbol{\mu}}$ is adjusted by the discrepancy between the observed values in the selected subpopulations $\mathbf{O}_{t+1,\mathbb{S}}$ and its prior expectation $\widehat{\mathbf{y}}_{t+1,\mathbb{S}}$. In the following validation section, we use the mean estimation for accuracy evaluation for comparison with other methods. The algorithm 1 formally describes the entire procedure of IGDAS.

Algorithm 1. Information Guided Dynamic Active Surveillance

```
input  : Training Data $\mathcal{D}$, score weights $\mathbf{a}$
output: Set of selected sentinel $\mathbb{S}$, estimation of unobserved sites
        $\bar{\mathbf{y}}_{t_f+1,\sim\mathbb{S}}$ for each testing time step

/* Dynamic Learning */
for $i \leftarrow 1$ to $|\mathcal{D}|$ do
    $\mathbf{O}_{t-L,t}, \mathbf{M}_{t-L,t} \leftarrow \mathcal{D}_i$
    Generate $\hat{\mathbf{y}}_{t+1}$ using Equation 2
    Calculate the prediction error loss using Equation 3
    Update model parameter by stochastic gradient variational Bayes
      technique

/* Selection and Inference */
while not stop do
    /* Dynamic Surveillance Target Selection */
    $\mathbb{S} \leftarrow [\,]$ ; /* Init */
    Generate prior prediction $\hat{\mathbf{y}}_{t_f+1}$, $\mu_{t_f+1}$, $\sigma_{t_f+1}$ using Equation 2
    for $i \leftarrow 1$ to $k$ do
        $\mathbf{c} \leftarrow [\,]; \mathbf{r} \leftarrow [\,]$
        for $j \leftarrow 1$ to $N$ do
            if $j \notin \mathbb{S}$ then
                Calculate conditional mutual information:
                  $c_j = I(\mathbf{O}_{t_f+1,j}; \mathbf{O}_{t_f+1,\sim[j,\mathbb{S}^k]} | \mathbf{O}_{t_f+1,\mathbb{S}^k})$
                $\mathbf{c} \leftarrow [\mathbf{c}, c_j]$
                $\mathbf{r} \leftarrow [\mathbf{r}, \hat{\mathbf{y}}_{t_f+1,j}]$
        Normalize $\mathbf{c}$ and $\mathbf{r}$ to $\bar{\mathbf{c}}$ and $\bar{\mathbf{r}}$
        Calculate the weighted score $\mathbf{s}$ using Equation 5 with $\bar{\mathbf{c}}$ and $\bar{\mathbf{r}}$
          and $\mathbf{a}$
        $\mathbb{S} \leftarrow [\mathbb{S}, \arg\max_j \mathbf{s}]$

    /* Inference */
    Monitor the $\mathbb{S}$ subpopulations and obtain observation $\mathbf{O}_{t_f+1,\mathbb{S}}$
    Infer the unobserved dynamic $\bar{\mathbf{y}}_{t_f+1,\sim\mathbb{S}}$ using Equation 6
```

4. Validation

We evaluate the efficiency of our proposed method in tackling the active surveillance issue using one synthetic and two real-world datasets.

4.1. Validation on synthetic dataset

Datasets The synthetic dataset is constructed as follows. We run the SEIR models on N locations. The equations of SEIR model can be written as follow (Wan et al., 2014):

$$
\begin{aligned}
\frac{dS_i}{dt} &= -\frac{\beta \sum_j^N C_{j,i} I_j S_i}{P}, \\
\frac{dE_i}{dt} &= \frac{\beta \sum_j^N C_{j,i} I_j S_i}{P} - \sigma E_i, \\
\frac{dI_i}{dt} &= \sigma E_i - \gamma I_i, \\
\frac{dR_i}{dt} &= \gamma I_i,
\end{aligned}
$$

where β is the transmission rate, γ is the recovery rate and $C_{j,i}$ is the contact rate between location i and j. We set the population sizes P to $1,000,000$, $\beta = 3$, $\gamma = 0.1428$, $\sigma = 0.25$. We generate the contact pattern using the Erdos-Renyi random network model (Newman, 2018) with edge probability $p = 0.3$.

Comparative methods In our comparative analysis, we use three baseline approaches: the group sparse learning approach, the Gaussian process approach, and the linear inverse problem (LIP) based approach. SNMA emerges as a state-of-the-art method for sentinel selection using group sparsity selection to shape a sentinel network marked by sparsity of rows (Pei et al., 2020). The LIP-based methodology, grounded in traditional experimental methodology, strategically deploys sensors to tackle linear inverse challenges. FrameSense (FSense for short) (Ranieri et al., 2014), MNEP (Jiang et al., 2016), and MPME (Jiang et al., 2016) are the representative methods of the LIP-based approach. Using principal component analysis according to (Pei et al., 2020), we extract an estimated sensing matrix from past data. Additionally, in GP-based methodology, sensor placements are optimized based on information theory principles (e.g., entropy or mutual information) to facilitate predictions for unobserved areas through GP interpolation. For this group, GP-based mutual information (GPMI) (Krause et al., 2008) is the primary representative.

Evaluation setting In the scenarios where the surveillance resource is limited, the observation used for model training is often incomplete. Accordingly, we generate the missing pattern with random monitoring probability K/N for model learning. We evaluate the performance of surveillance methods in terms of accuracy of estimation for unknown epidemic data by two commonly used metrics root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) (Pei et al., 2021; Tan et al., 2021):

$$
\begin{aligned}
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \\
MAE &= \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|
\end{aligned}
\tag{7}
$$

where y_i is real value and $\widehat{y}_i$ is the estimation. These two metrics evaluate the discrepancy between the estimate and the ground-truth values.

Performance of inference Table 1 shows the evaluation results on the synthetic dataset and real-world malaria data-set.Our IGDAS method generally outperformed the other methods in the majority of cases. For the synthetic dataset scenarios, with 20% monitoring data, IGDAS had a much lower RMSE of 1910, performing better than the other methods. At the same data level, IGDAS also had a lower MAE of 850, indicating better performance. With 40% monitoring data, IGDAS continued to lead with an RMSE of 1260 and an MAE of 577. With 60% observing data, IGDAS achieves the results of an RMSE of 1118 and an MAE of 526.

Table 1
Performance evaluation of our method and existing inference methods with different surveillance budgets K/N. The best result for each scenario is highlighted in bold and the second best is underlined.

Datasets	Scenarios		Methods					
			GPMI	FSense	MNEP	MPME	SNMA	IGDAS
Synthetic	20%	RMSE	41713	24251	13518	13194	11076	**1910**
		MAE	24172	10249	6268	6055	5161	**850**
	40%	RMSE	39294	53393	11350	10599	6043	**1260**
		MAE	25475	23349	4910	4932	2933	**577**
	60%	RMSE	37810	15847	7834	7299	3640	**1118**
		MAE	27465	7561	3486	3257	1682	**526**

Table 2
Performance evaluation of our method and existing inference methods with different surveillance budgets K/N. The best result for each scenario is highlighted in bold and the second best is underlined.

Datasets	Scenarios		Methods					
			GPMI	FSense	MNEP	MPME	SNMA	IGDAS
Malaria	20%	RMSE	1.25	20.11	0.91	0.88	1.18	**0.29**
		MAE	0.88	9.74	0.48	0.47	0.58	**0.02**
	40%	RMSE	1.03	8.16	0.98	0.87	1.30	**0.69**
		MAE	0.64	4.57	0.59	0.53	0.64	**0.28**
	60%	RMSE	1.39	129.67	1.02	1.15	1.14	**0.90**
		MAE	0.81	73.66	0.66	0.77	0.60	**0.48**

4.2. Validation on malaria infection survey

We conduct empirical validation in the scenario of a malaria infection survey. Malaria is still one of the most threatening infectious diseases, causing 249 million infections worldwide in 2022 (Venkatesan, 2024). China has been certified malaria-free in 2021, while Yunnan Province, located in the China-Myanmar border region, is at high risk for imported malaria infections, requiring timely imported case surveillance. Local CDC experts should conduct house-to-house visits in villages to inquire about any instances of fever. In addition, they should conduct biweekly surveys to identify additional secondary cases before the start of the next cycle based on the incubation period (Yang et al., 2014). Yunnan province has 18 border counties, but the limited number of health experts poses a great challenge in comprehensive coverage of all regions.

To evaluate the capacity of IGDAS to learn correlations between malaria infection and the selection of surveillance targets, we use malaria case data from 2005 to 2009 collected from 62 townships in Yunnan (Tan et al., 2021). Similarly to the synthetic scenario, we generated the missing pattern with random monitoring probability in consistent with the level of surveillance resource. We varied the number of surveillance regions and evaluated the performance of estimating infection epidemic in remain regions based on the target surveillance data.

Table 2 shows the results of the inference with different surveillance budgets. Regarding the malaria dataset scenarios, IGDAS consistently outperformed, even at lower monitoring rates like 20%, with a very low RMSE of 0.29 and a MAE of 0.02. This trend continued at the 40% surveillance rate, where IGDAS showed good results with an RMSE of 0.69 and an MAE of 0.28. At the 60% surveillance rate, IGDAS proved to be the best performer, demonstrating better results on both RMSE and MAE metrics. FSense and SNMA become unstable on the real-world malaria infection, which may be caused by the complex spatio-temporal pattern and the missing data.

4.3. Validation on COVID testing resource allocation

We further perform empirical validation in the scenario of the allocation of the test toolkit. The spread of SARS-CoV-2 in the United States has been noticeable since 2020. Timely response to SARS-CoV-2 relies heavily on rapid testing, yet during the initial phases, testing resources are notably scarce. We investigate the resource allocation strategy in the early phase of the epidemic for an accurate assessment of the status of the epidemic using a limited amount of monitoring data. We use the cases reported for COVID infections among 51 states in 2021 for experiments .[1] The experimental setting is consistent with the synthetic scenario and the malaria survey scenario.

Fig. 2 shows the inference performance of different methods in different scenarios (surveying 5 states, 10 states, and 20 states) in the diffusion of COVID between 51 states. For the RMSE metric, the GPMI method achieved scores of 2165, 2246, and 4177 for scenarios of 5, 10 and 20 sentinel states, respectively. MNEP method scored 1773, 1271, and 1171 for the corresponding scenarios. MPME method obtained scores of 1773, 1271, and 1171 for the respective scenarios. SNMA method scored 2624, 2253, and 7225 for the different scenarios. IGDAS method achieved scores of 1360, 1136, and 1047 for the respective scenarios. For the MAE metric, the GPMI method achieved scores of 1078, 1127, and 1536 for scenarios of 5, 10, and 20 sentinel states, respectively. MNEP method scored 1032, 683, and 590. MPME method obtained scores of 1032, 683, and 1171 for the respective scenarios. SNMA method scored 1257, 956, and 2024 for the different scenarios. IGDAS method achieved scores of 720, 573, and 526 for the respective scenarios. We omit the result of FSense in the figure, as FSense is not robust to complex noisy real-world data, reaching rmse 976813, 96084 and 170685 respectively. MNEP and MPME perform close in this scenario.

4.4. Informativeness and coverage

To examine the balance between informativeness and coverage, we varied the weights of informativeness a_1 in Equation (5) and, without loss of generality, set $a_2 = 1 - a_1$. Fig. 3 shows the accuracy and coverage of inference with variation a_1 in malaria surveillance. From the first two rows, we observe that the best inference performances appear in the middle range: in

[1] Available at https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

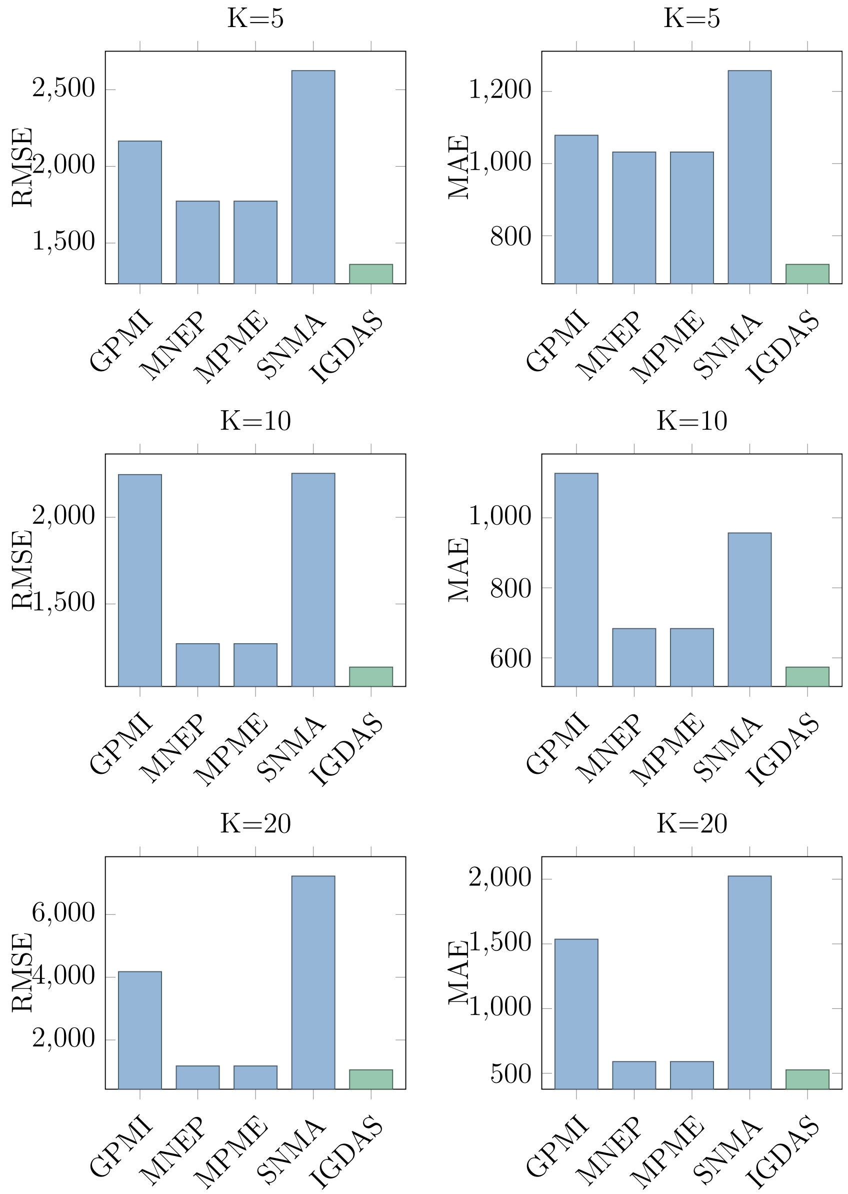


Fig. 2. Comparing results in COVID-19 testing resource allocation.

terms of RMSE, $a_1 = 0.6$, 0.2, 0.4 produce the best results, for $K/N = 20\%$, 40%, 60%, respectively; and in terms of MAE, $a_1 = 0.4$, 0.2, 0.2 produce the best results, respectively. These findings indicate that taking coverage into account benefit reducing the inference error. This might be because by monitoring high-incidence locations, there are fewer cases left unnoticed. This means that the overall mistakes we make in our conclusions are limited, as shown in the results in the third row. From the third row, we can see that generally with increasing a_1, the number of infections decreases. Moreover, as the number of surveillance locations increases, the difference in coverage with varying a_1 decreases. In $K/N = 20\%$, with higher a_2 (that is, less a_1) the coverage increases from 7.5 to 10.8, while in $K/N = 60\%$, the coverage just increases from 19.8 to 20.4.

5. Conclusion and discussion

In this work, we have developed an adaptive surveillance approach based on learning with incomplete epidemic data. By adopting probabilistic modeling, we are able to evaluate the informativeness of collecting a spatio-temporal point in a rigorous manner and inform the uncertainty of the estimation. We have also validated our proposed approach using both a synthetic dataset and two real-world malaria datasets, demonstrating that our approach produces more accurate surveillance

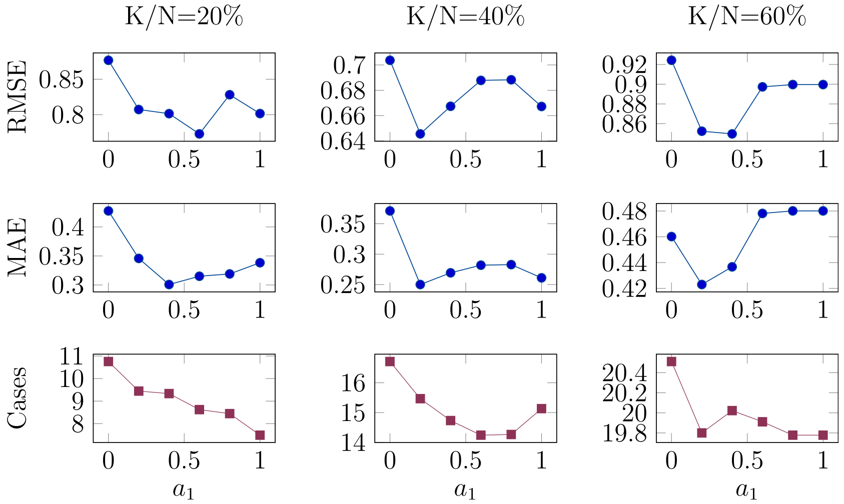


Fig. 3. RMSE, MAE and covered infection with varying weighting of informativeness a_1 on malaria dataset. Covered infection indicates the average number of infections monitored in the selected sites.

performance than existing static sentinel-based methods. The results provide a good foundation for investigating the adaptive planning of many health surveillance activities.

Compared with several baseline methods, our method is capable of capturing nonlinear spatio-temporal dependency for adaptive planning. Previous methods mainly focus on spatial sentinel deployment and infer the epidemic of uncovered locations using only spatial dependency. While our method is capable of adaptive planning the sentinel and utilizes the temporal dependency when previous observations are available. The linear inverse methods, namely MNEP and MPME, presuppose linearity in spatial interconnection, whereas GPMI, SNMA and IGDAS can grasp the nonlinear spatial interdependency. Moreover, SNMA and our method have the capacity to integrate prior knowledge of the epidemic, thereby enhancing its efficacy. Given that our method involves nonlinear dependency modeling that encompasses both temporal and spatial dependency, the computational overhead is elevated. However, our method completes a single-step planning and inference in the COVID testing context with $K = 10$ in a mere 0.78 s, which remains admissible in applications of infectious disease surveillance planning, especially when the planning phase typically spans a day or more.

As a notable point, our approach integrates informativeness and coverage in the selection of surveillance sites. We have also shown that there are delicate relationships between these two objectives. With increasing coverage, the number of remaining epidemic infections decreases, which improves inference performance. However, if only coverage is considered, the inference performance declines. The balance between informativeness and coverage depends on the specific surveillance goals aligned with decision-making processes.

There are also some limitations to the current work. First, incomplete data are a big challenge for spatio-temporal correlations learning in surveillance selection. In this study, we employ a straightforward and effective approach to managing missing data in model input. Future research can investigate more sophisticated techniques to handle missing data. Secondly, during the process of selecting surveillance sites, targets are chosen without taking into account the limitations imposed by employing sentinels. For example, selecting certain locations for monitoring could make it difficult to access other distant areas. These constraints could be important in certain situations.

CRediT authorship contribution statement

Qi Tan: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Chenyang Zhang:** Validation, Investigation, Data curation. **Jiwen Xia:** Software, Formal analysis, Data curation. **Ruiqi Wang:** Writing – review & editing, Visualization, Project administration. **Lian Zhou:** Validation, Project administration, Formal analysis. **Zhanwei Du:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition. **Benyun Shi:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the funding support from the National Natural Science Foundation of China and the Research Grants Council (RGC) of Hong Kong Joint Research Scheme (No. 62261160387, N_HKBU222/22) and Shenzhen-Hong Kong-Macau Science and Technology Project (Category C) (Project no: SGDX20230821091559022). The authors would like to express their gratitude to Dr. Dong QIAN for his valuable advice on experimental configuration management.

References

Buhat, C. A. H., Duero, J. C. C., Felix, E. F. O., Rabajante, J. F., & Mamplata, J. B. (2021). Optimal allocation of covid-19 test kits among accredited testing centers in the Philippines. *Journal of Healthcare Informatics Research, 5*, 54–69.

Cao, J., Newby, G., Cotter, C., Hsiang, M. S., Larson, E., Tatarsky, A., Gosling, R. D., Xia, Z., & Gao, Q. (2021). Achieving malaria elimination in China. *The Lancet Public Health, 6*(12), e871–e872.

Case, B., Dye-Braumuller, K. C., Evans, C., Li, H., Rustin, L., & Nolan, M. S. (2024). Adapting vector surveillance using bayesian experimental design: An application to an ongoing tick monitoring program in the southeastern United States. *Ticks and Tick-borne Diseases, 15*(3), Article 102329.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Du, R., Chen, C., Yang, B., & Guan, X. (2013). Vanet based traffic estimation: A matrix completion approach. In *2013 IEEE global communications conference* (pp. 30–35).

Groseclose, S. L., & Buckeridge, D. L. (2017). Public health surveillance systems: Recent advances in their use and evaluation. *Annual Review of Public Health, 38*, 57–79.

Jamison, D. T., Breman, J. G., Measham, A. R., Alleyne, G., Claeson, M., Evans, D. B., Jha, P., Mills, A., & Musgrove, P. (2006). *Disease control priorities in developing countries*.

Jiang, C., Soh, Y. C., & Li, H. (2016). Sensor placement by maximal projection on minimum eigenspace for linear inverse problems. *IEEE Transactions on Signal Processing, 64*(21), 5595–5610.

Khamsiriwatchara, A., Wangroongsarb, P., Thwing, J., Eliades, J., Satimai, W., Delacollette, C., & Kaewkungwal, J. (2011). Respondent-driven sampling on the Thailand-cambodia border. i. can malaria cases be contained in mobile migrant workers? *Malaria Journal, 10*, 1–11.

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational bayes*. arXiv preprint arXiv:1312.6114.

Krause, A., Singh, A., & Guestrin, C. (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research, 9*(2).

Laurent, C. S., & Cowlagi, R. V. (2023). Near-optimal task-driven sensor network configuration. *Automatica, 152*, Article 110966.

Li, Z., Chen, Q., Feng, L., Rodewald, L., Xia, Y., Yu, H., Zhang, R., An, Z., Yin, W., Chen, W., et al. (2020). Active case finding with case management: The key to tackling the covid-19 pandemic. *The lancet, 396*(10243), 63–70.

Litwin, T., Timmer, J., Berger, M., Wahl-Kordon, A., Müller, M. J., & Kreutz, C. (2022). Preventing covid-19 outbreaks through surveillance testing in healthcare facilities: A modelling study. *BMC Infectious Diseases, 22*(1), 105.

Longbottom, J., Wamboga, C., Bessell, P. R., Torr, S. J., & Stanton, M. C. (2021). Optimising passive surveillance of a neglected tropical disease in the era of elimination: A modelling study. *PLoS Neglected Tropical Diseases, 15*(3), Article e0008599.

Losos, J. Z. (1996). Routine and sentinel surveillance methods. *EMHJ-Eastern Mediterranean Health Journal, 2*(1), 46–50, 1996.

Newman, M. (2018). *Networks*. Oxford University Press.

Organization, W. H., et al. (2022). *10 proposals to build a safer world together–strengthening the global architecture for health emergency preparedness, response and resilience*.

Pei, S., Teng, X., Lewis, P., & Shaman, J. (2021). Optimizing respiratory virus surveillance networks using uncertainty propagation. *Nature Communications, 12*(1), 222.

Pei, H., Yang, B., Liu, J., & Chang, K. (2020). Active surveillance via group sparse bayesian learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(3), 1133–1148.

Polgreen, P. M., Chen, Z., Segre, A. M., Harris, M. L., Pentella, M. A., & Rushton, G. (2009). Optimizing influenza sentinel surveillance at the state level. *American Journal of Epidemiology, 170*(10), 1300–1306.

Ranieri, J., Chebira, A., & Vetterli, M. (2014). Near-optimal sensor placement for linear inverse problems. *IEEE Transactions on Signal Processing, 62*(5), 1135–1146.

Sarti, E., L'Azou, M., Mercado, M., Kuri, P., Siqueira, J. B., Jr., Solis, E., Noriega, F., & Ochiai, R. L. (2016). A comparative study on active and passive epidemiological surveillance for dengue in five countries of Latin america. *International Journal of Infectious Diseases, 44*, 44–49.

Scarpino, S. V., Dimitrov, N. B., & Meyers, L. A. (2012). Optimizing provider recruitment for influenza surveillance networks. *PLoS Computational Biology, 8*(4), Article e1002472.

Sturrock, H. J., Roberts, K. W., Wegbreit, J., Ohrt, C., & Gosling, R. D. (2015). Tackling imported malaria: An elimination endgame. *The American Journal of Tropical Medicine and Hygiene, 93*(1), 139.

Tan, Q., Liu, Y., Liu, J., Shi, B., Xia, S., & Zhou, X.-N. (2021). Heterogeneous neural metric learning for spatio-temporal modeling of infectious diseases with incomplete data. *Neurocomputing, 458*, 701–713.

Venkatesan, P. (2024). The 2023 who world malaria report. *The Lancet Microbe, 5*(3), e214.

Vitale, M., Lupone, C. D., Kenneson-Adams, A., Ochoa, R. J., Ordoñez, T., Beltran-Ayala, E., Endy, T. P., Rosenbaum, P. F., & Stewart-Ibarra, A. M. (2020). A comparison of passive surveillance and active cluster-based surveillance for dengue fever in southern coastal Ecuador. *BMC Public Health, 20*, 1–10.

Wan, X., Liu, J., Cheung, W. K., & Tong, T. (2014). Inferring epidemic network topology from surveillance data. *PLoS One, 9*(6), Article e100661.

Wang, S., Li, B., Yang, M., & Yan, Z. (2019). Missing data imputation for machine learning. In *IoT as a service* (pp. 67–72). Cham.

Wang, H., Yao, K., Pottie, G., & Estrin, D. (2004). Entropy-based sensor selection heuristic for target localization. In *Proceedings of the 3rd international symposium on information processing in sensor networks* (pp. 36–45).

Yang, B., Guo, H., Yang, Y., Shi, B., Zhou, X., & Liu, J. (2014). Modeling and mining spatiotemporal patterns of infection risk from heterogeneous data for active surveillance planning. *Proceedings of the AAAI Conference on Artificial Intelligence, 28*.

Zeng, D., Cao, Z., & Neill, D. B. (2021). Chapter 22 - artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control. In L. Xing, M. L. Giger, & J. K. Min (Eds.), *Artificial intelligence in medicine* (pp. 437–453). Academic Press.