

1 **Towards a simulation framework for optimizing infectious disease**
2 **surveillance: An information theoretic approach for surveillance**
3 **system design**

4
5 Qu Cheng^{1†}, Philip A. Collender^{1†}, Alexandra K. Heaney^{1†}, Xintong Li², Rohini Dasan¹, Charles
6 Li¹, Joseph A. Lewnard³, Jon Zelner⁴, Song Liang⁵, Howard H. Chang², Lance A. Waller²,
7 Benjamin A. Lopman², Changhong Yang⁶, Justin V. Remais^{1*}

8
9 ¹ Division of Environmental Health Sciences, School of Public Health, University of California,
10 Berkeley, Berkeley 94720, USA

11 ² Rollins School of Public Health, Emory University, Atlanta 30322, USA

12 ³ Division of Epidemiology, School of Public Health, University of California, Berkeley,
13 Berkeley 94720, USA

14 ⁴ Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor
15 48109, USA

16 ⁵ Department of Environmental and Global Health College of Public Health and Health
17 Professions, University of Florida, Gainesville, 32611, USA

18 ⁶ Institute of Health Informatics, Sichuan Center for Disease Control and Prevention, Chengdu
19 610041, China

20

21

22 [†] Denotes shared first authorship

23 ^{*} Correspondence to: Prof. Justin V. Remais, jvr@berkeley.edu; 510-643-8900 (office); 510-
24 643-5056 (fax)

25

26

27

28

29 **Abstract**

30 Infectious disease surveillance systems provide vital data for guiding disease prevention and
31 control policies, yet the formalization of methods to optimize surveillance networks has largely
32 been overlooked. Decisions surrounding surveillance design parameters—such as the number
33 and placement of surveillance sites, target populations, and case definitions—are often
34 determined by expert opinion or deference to operational considerations, without formal analysis
35 of the influence of design parameters on surveillance objectives. Here we propose a simulation
36 framework to guide evidence-based surveillance network design to better achieve specific
37 surveillance goals with limited resources. We define evidence-based surveillance design as a
38 constrained, multi-dimensional, multi-objective, dynamic optimization problem, acknowledging
39 the many operational constraints under which surveillance systems operate, the many
40 dimensions of surveillance system design, the multiple and competing goals of surveillance, and
41 the complex and dynamic nature of disease systems. We describe an analytical framework for
42 the identification of optimal designs through mathematical representations of disease and
43 surveillance processes, definition of objective functions, and the approach to numerical
44 optimization. We then apply the framework to the problem of selecting candidate sites to expand
45 an existing surveillance network under alternative objectives of: (1) improving spatial prediction
46 of disease prevalence at unmonitored sites; or (2) estimating the observed effect of a risk factor
47 on disease. Results of this demonstration illustrate how optimal designs are sensitive to both
48 surveillance goals and the underlying spatial pattern of the target disease. The findings affirm
49 the value of designing surveillance systems through quantitative and adaptive analysis of
50 network characteristics and performance. The framework can be applied to the design of
51 surveillance systems tailored to setting-specific disease transmission dynamics and surveillance
52 needs, and can yield improved understanding of tradeoffs between network architectures.

54 **Author summary**

55 Disease surveillance systems are essential for understanding the epidemiology of
56 infectious diseases and improving population health. A well-designed surveillance system can
57 achieve a high level of fidelity in estimates of interest (e.g., disease trends, risk factors) within its
58 operational constraints. Currently, design parameters that define surveillance systems (e.g.,
59 number and placement of the surveillance sites, target populations, case definitions) are
60 selected largely by expert opinion and practical considerations. Such an informal approach is
61 less tenable when multiple aspects of surveillance design—or multiple surveillance objectives—
62 need to be considered simultaneously, and are subject to resource or logistical constraints.
63 Here we propose a framework to optimize surveillance system design given a set of defined
64 surveillance objectives and a dynamical model of the disease system under study. The
65 framework provides a platform to conduct *in silico* surveillance system design, and allows the
66 formulation of surveillance guidelines based on quantitative evidence, tailored to local realities
67 and priorities. The approach facilitates greater collaboration between health planners and
68 computational and data scientists to advance surveillance science and strengthen the
69 architecture of surveillance networks.

70 1 Introduction

71 Infectious disease surveillance systems provide vital information on patterns of disease
72 occurrence across space, time, and populations of interest, and ultimately provide the basis for
73 evidence-based disease control policy decisions [1]. Considerable progress has been made
74 supporting infectious disease control decision-making with computational approaches to
75 evaluate the outcomes of alternative decisions [2]. Examples include optimizing when, where,
76 and among which populations to allocate public health resources [3, 4], determining the optimal
77 balance between multiple intervention approaches (e.g., case detection, treatment, vaccination,
78 and sanitation improvement) [5-8], and optimizing the start time, duration, and dose of drug
79 treatment programs [9, 10]. In contrast, little attention has been paid to the development of tools
80 for improving infectious disease surveillance system designs, and formalization of methods to
81 optimize surveillance networks has largely been overlooked.

82 The ‘design parameters’, which are the high-level characteristics that define infectious
83 disease surveillance networks—such as the number and locations of surveillance sites,
84 sampling frequency for laboratory testing or community-based surveys, and selection of
85 diagnostic techniques—can greatly influence the degree to which the resulting surveillance data
86 serves public health objectives, including early detection of outbreaks (e.g., the coronavirus
87 disease outbreak in 2020) [11], improved understanding of disease emergence and spread [12],
88 and accurate measurement of the impact of interventions [13]. Thus, key design parameters can
89 be modified in a manner informed by optimization analysis such that the system better achieves
90 specific surveillance goals. Examples include relocating and adding reporting sites to predict the
91 temporal trend of diseases more accurately [14]; changing diagnostic approaches/case
92 definitions to increase the chance of detecting cases [15]; and targeting of sampling towards
93 specific subpopulations to improve the timeliness of outbreak detection [16, 17].

94 In practice, surveillance system design parameters are often set in an *ad hoc* fashion
95 based on operational considerations (e.g., budget, convenience, political agendas), rather than
96 through quantitative evaluation of how alternative designs might impact surveillance system
97 objectives. For instance, World Health Organization (WHO) recommends selection of influenza
98 surveillance sites based on the facilities’ willingness to participate, availability of necessary
99 laboratory and information infrastructure, ability to cover the surveillance cost, and
100 representativeness of the general population. Notably absent from these criteria is the degree to
101 which the network’s performance on specific surveillance objectives will be enhanced [13]. The
102 absence of objective criteria and methods to evaluate and iteratively reconfigure surveillance
103 system design can lead to inefficient use of limited resources. For example, in China, current
104 requirements specify that 5-15 influenza-like illness (ILI) cases are required to be sampled per
105 week at each of the 556 influenza sentinel hospitals for laboratory confirmation [18]. If the total
106 sample size is fixed, it may be that reducing the number of sentinel sites (e.g., prioritizing sites
107 in populous regions and with high levels of population movement), while increasing the sample
108 sizes at the remaining sites, could yield more timely detection of outbreaks with the same level
109 of resources. What is more, because disease surveillance systems generally operate in pursuit
110 of multiple objectives, decision-making surrounding optimal design can be highly
111 counterintuitive.

112

113 Recent research has provided some early examples of quantitative infectious disease
114 surveillance design optimization [19, 20]. In one study, researchers estimated that an optimal
115 relocation of Iowa’s existing 22 ILINet sentinel sites could increase population coverage of the
116 network from 56% to 75% [21]. As another example, targeted surveillance of pregnant women
117 over blood donors for compulsory diagnostic testing was estimated to increase the weekly
118 probability of detecting at least one Zika case from 11% to 40% [15]. While these and other
119 studies serve as foundational examples, the methods utilized in these analyses are targeted
120 towards narrow, study-specific objectives and specific networks, and are challenging to
121 generalize to other—even closely related—surveillance design optimization problems. What is
122 more, prior studies have not attempted to articulate a general theory of surveillance design
123 optimization and decision-making.

124 Here, we present for the first time a unified analytical framework for quantitative
125 infectious disease surveillance system optimization, accommodating multiple surveillance
126 design parameters, objectives, operational constraints, and underlying disease processes. A
127 common framework and standard terminology can enable closer collaboration between and
128 among computational researchers, public health officials, and other stakeholders regarding the
129 design and implementation of infectious disease surveillance systems. This in turn can
130 accelerate the pace of methodological innovations and facilitate the development of surveillance
131 design theories that anticipate and respond to current and future epidemiological challenges.
132 Furthermore, a generalized framework can inspire the application of quantitative surveillance
133 optimization across broader settings, resulting in system designs better aligned with local
134 realities and public health priorities.

135 **2 Surveillance design as a multi-objective, multi-dimensional,** 136 **constrained and dynamic optimization problem**

137 The search for optimal disease surveillance designs is a highly complex problem. This
138 results from the multiple, often competing goals of surveillance data collection, idiosyncratic
139 surveillance network design, the need to represent operational constraints that govern
140 surveillance systems, and the complexity and dynamic nature of diseases under surveillance.
141 Simple optimization problems involving a single design parameter and objective for a given
142 target disease—such as the optimal placement of a new surveillance site to maximize the
143 proportion of influenza cases detected—may be solved in relatively straightforward fashion by
144 testing all possible designs and choosing the design that generates optimal network
145 performance (e.g., the new site location that results in the highest proportion of cases detected
146 overall). However, surveillance network optimization quickly becomes non-trivial when the
147 design space increases (e.g., selecting 10 sites out of 200 alternative sites), when multiple
148 objectives (such as increasing case detection, improving spatial and temporal trend coverage,
149 and risk factor identification) are subject to simultaneous analysis and optimization, or when
150 optimization is subject to constraints regarding resource limitations and operational plausibility.
151 Uncertainty regarding the functioning of the epidemiologic system and shifts in patterns of
152 diseases further complicate matters. Hence, our optimization goals are multidimensional,
153 dynamic, and stochastic. In this section, we describe the relevance of surveillance objectives,

154 network design parameters, operational constraints and dynamic disease systems to the pursuit
155 of surveillance optimization.

156 **Multiple objectives.** Disease surveillance systems are established and designed for
157 diverse purposes, including to collect data to understand variations in disease frequency across
158 populations, space, and time, to monitor pathogen composition over time, to detect outbreaks
159 and forecast epidemics, to assess the impact of interventions, and to determine risk factors
160 associated with diseases. Most surveillance systems operate with multiple public health
161 objectives and multiple logistical constraints. Hence, surveillance system designs should
162 generally be subject to multi-objective optimization, and tradeoffs between different objectives
163 must be considered. For instance, if the goals of a system are to both estimate prevalence and
164 assess the impact of risk factors, the network design should be subjected to optimization that
165 considers both objectives using a framework that is capable of capturing tradeoffs between
166 designs with respect to achieving the two objectives.

167 **Multiple design parameters.** Surveillance system structure and design can be
168 decomposed into a multitude of characteristics, operational details, and features that influence
169 the performance of surveillance networks. Design parameters can lend themselves to
170 representation and simulation within models. Multiple design parameters may be amenable to
171 optimization analysis, either through single or multivariate optimization. For example, to improve
172 estimation of disease incidence, either the accuracy of diagnostics at existing reporting facilities
173 or the number of facilities in the reporting network, or both, can be modified. Other design
174 parameters, such as when, where, and among which populations to implement targeted
175 sampling efforts may also be entered into the analysis, greatly expanding the dimensionality of
176 the problem. Moreover, the set of design parameters to optimize depends on the surveillance
177 goals. For example, when the surveillance goal is accurate estimation of the temporal trend of a
178 disease, it may be that the placement of sites is less important than sampling frequency. Table
179 1 shows examples of design parameters from select real world infectious diseases surveillance
180 systems, and their potential impacts on surveillance system performance.

181

182 **Table 1.** Example surveillance system design parameters and their potential impacts on surveillance performance.

Design parameter	Definition	Potential impacts on surveillance performance	Example designs	Example surveillance system	Ref.
Target population	Population to be monitored for outcomes of interest.	Target populations representative of a general population provide a means of tracking overall disease incidence and trends in the population as a whole. Target populations informed by demographic differences in disease risk, transmission potential, or detection probability may provide advantages for monitoring outcomes in vulnerable populations, anticipating outbreaks, or tracking rare diseases.	All persons >2 years of age residing in homes	Republic of South Africa HIV prevalence survey	[22]
			Pregnant women and infants	US Zika Pregnancy and Infant Registry	[23]
Site enrollment	The inclusion of hospitals and other facilities in passive reporting networks, or selection of locations for active or environmental surveillance	Site selection influences factors such as population coverage and representativeness, diagnostic quality, the speed at which spreading outbreaks may be detected, and informational redundancy due to spatial proximity or other sources of similarity between locations	Hospitals in Maluku, North Sulawesi, East Kalimantan, North Sumatra, Yogyakarta and West Nusa Tenggara	Indonesia influenza sentinel surveillance system	[24]
			Health centers in Dembi, Asendabo, Tulubolo, Guangua, Bulbula, Dhera, Welenchity, Metahara, Asebot, and Kersa	Ethiopia malaria sentinel surveillance	[25]
Sampling strategy	Type of sampling used to identify cases among the target population.	Sampling strategies influence the representativeness of surveillance data, as well as the ability of surveillance systems to detect rare or underreported conditions. Strategies that adequately characterize a general population may be biased with respect to critical subpopulations, especially those facing stigma.	Hospital-based convenience sampling (e.g. every fourth patient meeting case definition)	Bangladesh rotavirus surveillance system	[26]
			Respondent-driven sampling, which uses existing samples in high-risk groups (e.g., intravenous drug user, men who have sex with men) to recruit new samples, then uses a model to correct for potential bias in the nonprobability sampling	Central America sexual behaviors and HIV prevalence survey	[27]
			Multistage cluster sampling, which selects districts first according to their population size, then selects communes within the selected districts by simple random sampling,	Viet Nam national survey of tuberculosis prevalence	[28]

			and selects all residents aged ≥ 15 years in the selected communes		
Sampling intensity	Number of samples per sampling interval	Under operational constraints, the choice between sampling more frequently but with low intensity or less frequently with higher intensity represents a tradeoff between the ability to resolve high frequency changes in outcomes of interest, or timeliness of detection, and reducing statistical uncertainty	3 adults and 2 children per week	Malaysia laboratory-based influenza surveillance system	[29]
			5 mild cases serotyped per month per site	China hand foot mouth disease sentinel surveillance system	[30]
Sampling seasonality	Pre-determined changes in sampling intensity over time	Year-round sampling increases the chances of detecting unexpected changes in disease incidence. However, if disease seasonality is static and well-understood, resources may be better used for intensive seasonal sampling	Year-round	New Zealand virological surveillance system	[31]
			Transmission season (June-October)	China dengue virological surveillance system	[32]
Laboratory diagnostics	Methods used to determine the presence of a pathogen or syndrome of interest.	Diagnostic tests and other related factors such as specimen types, the quality of the specimen, and the time from onset to specimen collection can influence the sensitivity and specificity of the surveillance system.	Isolation of <i>Bordetella pertussis</i> from clinical specimen and/or a four-fold or greater increase in titer of antibody against <i>B. pertussis</i> between acute and convalescent sera	China pertussis surveillance system	[33]
			Isolation of <i>B. pertussis</i> from clinical specimen and positive polymerase chain reaction (PCR) for <i>B. pertussis</i>	US CDC pertussis surveillance system	[34]
Case definition	Diagnostic criteria to classify outcomes of interest.	Case definitions can influence factors such as the severity and characteristics of cases identified, and the sensitivity and specificity of the system.	Influenza-like illness, defined as an acute respiratory infection with measured fever of ≥ 38 °C and cough with onset within the last 10 days	WHO global influenza surveillance	[35]
			Severe acute respiratory infection, defined as an acute respiratory infection with history of fever or measured fever of ≥ 38 °C and cough with onset within the last 10 days and requires hospitalization		

184
185 **Operational constraints.** Operational restrictions on surveillance system designs—due
186 to budgetary, logistical, political and cultural considerations—add critical constraints to the
187 optimization problem. Absent constraints, the optimal design may be self-evident, e.g., sampling
188 at maximal frequency and intensity. Yet when there is a fixed budget for samples, the optimal
189 balance between design parameters—say, number of samples and sampling frequency—
190 depends on the relative value of precise cross-sectional estimates of disease prevalence versus
191 characterizing disease incidence over time, which in turn depends on the specific objectives of
192 surveillance and the dynamics of the underlying disease system.

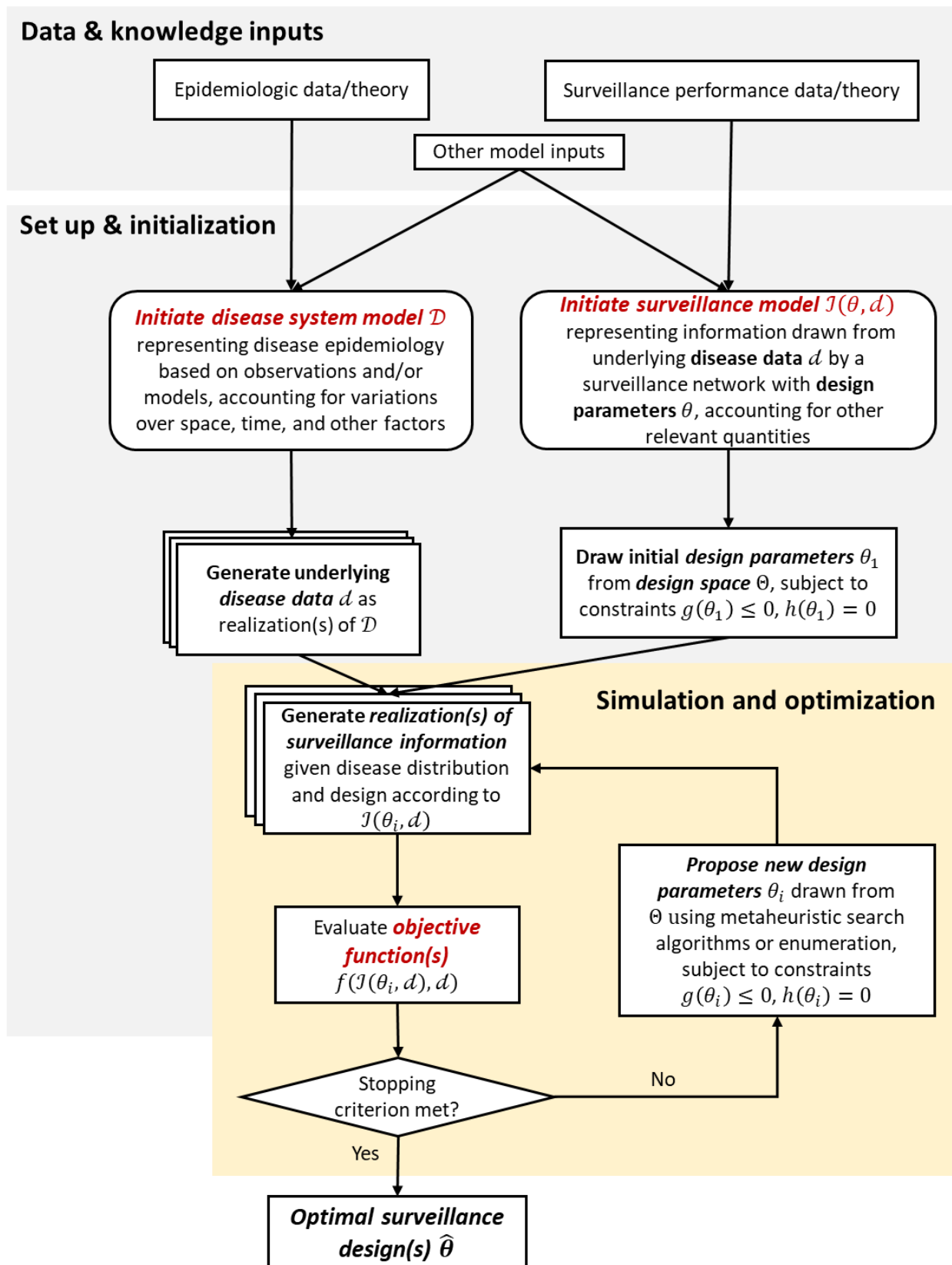
193 **Dynamic and imperfectly understood disease systems.** Surveillance systems must
194 respond to shifts in the epidemiology of target infections. Optimal designs will likely shift in
195 response to the evolution of underlying epidemiology and available knowledge. For instance, as
196 infections emerge, become endemic, or approach elimination within populations or
197 subpopulations, the goals of surveillance, and the resulting optimal designs, can (and must)
198 evolve alongside them. The dynamic nature of optimal surveillance design may be especially
199 important in emerging economies that are undergoing epidemiologic transitions. For instance,
200 as a region or nation approaches elimination of a particular infectious disease, surveillance
201 goals generally shift from enumeration of endemic cases occurring in the general population to
202 detection of nexuses of sporadic transmission. This may require new designs (e.g., shifting to a
203 more sensitive diagnostic test within a limited area, or increasing the coverage of
204 subpopulations involved in ongoing transmission), and adjustment of system objectives (e.g.,
205 maximize detection of the few remaining cases instead of minimizing false positive detections).
206 Additionally, as cases caused by novel pandemics (e.g., the 2020 coronavirus disease
207 pandemic, or 2009 H1N1 pandemic) start to increase exponentially, surveillance systems may
208 need to switch from tracking individual cases to population-based surveillance (e.g., pathogen
209 testing for a proportion of patients with a non-specific syndrome) in order to monitor the
210 progression of the outbreak and develop mitigation strategies without depleting public health
211 resources.

212 **3 A framework for surveillance simulation and optimization**

213 The aforementioned challenges of surveillance optimization—multiple objectives,
214 combinatorial complexity of relevant design parameters, operational constraints, and dynamic
215 and uncertain epidemiology of target diseases—suggest the need for a generalized framework
216 for surveillance network optimization. Advances in computation for simulation-based studies
217 have benefitted many related fields, including optimal disease control [36-39], yet applications of
218 simulation optimization to the design of disease surveillance networks have scarcely been
219 pursued. In the following sections, we detail a simulation and optimization framework for
220 designing infectious disease surveillance networks, and demonstrate its application in a site
221 selection context. Our framework facilitates a quantitative approach to designing surveillance
222 systems tailored to local disease transmission dynamics and surveillance needs, as well as a
223 more general study of optimal network design principles under varying objectives and
224 epidemiological circumstances.

225 Broadly, our framework (Figure 1) allows for evaluation of surveillance system
226 performance across a predefined design space under different epidemiologic scenarios
227 (disease system model) and network characteristics (surveillance model). Numerical
228 optimization algorithms are applied to efficiently identify the region(s) of design space that yield
229 superior network performance based on one or more specific surveillance goals (simulation
230 optimization search). The optimization procedure (Figure 1 and Box 1) yields a set of network
231 designs (i.e., optimal design parameter values) that maximize performance with respect to the
232 specified public health goal(s), according to the specified data and models.
233

234



235

236
237 **Figure 1. Schematic of surveillance system optimization.** The surveillance system optimization
238 procedure uses data and knowledge about disease transmission and case ascertainment to identify
239 optimal surveillance designs with regard to predefined surveillance goals. First, a disease system model
240 \mathcal{D} is defined, using observed epidemiologic data and/or theory, and taking into account relevant factors
241 influencing disease dynamics or distribution. Multiple realizations of disease data (d) may be generated to
242 explore optimal designs under uncertainty or variability of the underlying system (Section 3.1).
243 Furthermore, an ensemble of disease models can be combined to reduce the chance of model
244 misspecification. Next, a surveillance model is defined to represent how information on the state of the
245 disease system is captured as a function of design parameters θ and any other relevant variables (e.g.,
246 factors known to affect the sensitivity and specificity of a diagnostic test, or estimated underreporting
247 rates for an area; Section 3.2). To initiate the optimization process, an initial design parameter set, θ_1 , is
248 drawn from the design space subject to operational constraints $g(\theta_i) \leq 0, h(\theta_i) = 0$ and, along with
249 underlying disease data d , input to the surveillance model to generate a realization of surveillance
250 information, $\mathcal{J}_1 = \mathcal{J}(\theta_1, d)$. The objective function, f , is evaluated based on the disease data d , and
251 surveillance information \mathcal{J}_1 (Section 3.3). If a stopping criterion (e.g., reaching a large number of iterations;
252 *de minimis* improvement in objective function) is not met, a new design parameter set, θ_i , is proposed
253 from the design space using metaheuristic search algorithms (e.g., simulated annealing, genetic
254 algorithm, particle swarm algorithm) when the design space is large, or enumeration when the design
255 space is small. This new design parameter set is then used to generate a new realization of surveillance
256 information and evaluation of the objective functions (Section 3.4). After a stopping criterion is met,
257 design parameter sets with the best objective function values are output as optimal surveillance designs.
258

Box 1. Surveillance System Optimization Procedure

Input: Epidemiologic data and/or theory, surveillance performance data and/or theory, and other auxiliary data (e.g., disease risk factors)

Output: the design parameter set with the highest/lowest (i.e., optimal) objective function value

Initialization:

Define a disease system model to represent the underlying dynamics of the target disease system in the spatial, temporal, and demographic context of interest

Generate disease distributions d as realization(s) of the system

Sample initial design parameter set, θ_1 , within the design space subject to constraints $g(\theta_i) \leq 0, h(\theta_i) = 0$

Generate realization(s) of surveillance information, \mathcal{J}_1 , given d and θ_1

Evaluate objective function(s) f given \mathcal{J}_1 and d

while stop criterion is not met **do**

Propose a new design parameter set, θ_i , within the design space using metaheuristic search algorithms or enumeration

Generate realization(s) of surveillance information, \mathcal{J}_i , given d and θ_i

Evaluate objective function(s), f , given \mathcal{J}_i and d

end while

return the best design parameter set, $\hat{\theta}$ (i.e., with the optimal objective function value)

259

260 **3.1 Specify and parameterize disease system model**

261 An accurate representation of epidemiologic characteristics of the target disease(s) is
262 essential for a successful optimization. This representation can be generated using
263 observational data, outputs of mechanistic transmission models, or other approaches, and
264 represents the best estimate of the disease's epidemiology that is used to evaluate surveillance
265 network performance using objective functions (defined in section 3.3 *Define objective*
266 *function(s)*, below). To avoid potential model-misspecification, an ensemble of disease models
267 and multiple realizations of disease models (i.e., with varying epidemiologic parameter values)
268 can be utilized in the framework. The structure of the disease system model output—such as
269 spatial and temporal resolution—should be tailored to the surveillance objectives and design
270 parameters. For instance, if a surveillance objective is to better estimate the spatial distribution
271 of a disease, the target disease data must include geographical information about cases.

272 **3.2 Specify and parameterize surveillance model**

273 In order to identify optimal network designs, a model representing key aspects of the
274 sampling of and extraction of information from underlying disease processes by the surveillance
275 system is needed. The surveillance model must represent the mechanisms through which
276 variation in network design parameters is expected to impact the epidemiologic information
277 obtained and thus governs optimization with respect to system objectives. Surveillance models
278 generally comprise a set of probability distributions relating target estimands to the underlying
279 disease distribution, conditional on network design and other relevant considerations. For
280 example, to optimize the diagnostic protocol for minimal bias in reporting, a surveillance model
281 may be constructed for the distribution of reported cases conditioned on diagnostic method,
282 background prevalence of the target disease and conditions with similar clinical presentation,
283 and the distribution across subpopulations of factors that impact diagnostic sensitivity and
284 specificity. When random errors contributed by surveillance processes are not explicitly taken
285 into account, as may be the case when seeking to maximize the size of the population covered
286 by a surveillance network, the surveillance model becomes a set of conditional Dirac delta
287 distributions, and is deterministic. During the process of surveillance model specification,
288 aspects of surveillance design that will be allowed to vary during optimization (i.e., the
289 parameters to be optimized), and those that will be fixed (i.e., design aspects that are relevant
290 to performance, but which it is not feasible or desirable to change) must be decided upon.
291 Surveillance models may be as granular (e.g., modeling the full sequence of events necessary
292 for each individual case to be reported) or abstract (e.g., modeling the overall proportion of
293 cases detected in a population) as is deemed necessary for the optimization procedure,
294 recognizing, however, that computational complexity may limit the feasibility of certain
295 representations.

296 **3.3 Define objective function(s)**

297 Changes to design parameters can be analyzed in relation to their influence on network
298 performance in the context of specific surveillance system objectives. That is, performance is
299 evaluated with respect to achieving a specific goal or goals. This evaluation is formalized by

300 defining objective functions, which define the specific minimization or maximization problem to
 301 be solved, based on the design parameters and surveillance goals of interest. Thus, network
 302 performance is estimated through the iterative evaluation of objective functions, which are
 303 minimized (or maximized) as the design parameter space is searched. Table 2 presents
 304 canonical objective functions available for use in surveillance network optimization. Our
 305 examples do not explicitly include operational considerations within objective functions, but
 306 these can easily be taken into account. For example, the objective function could be established
 307 so as to yield the marginal information gain per added site or sample, or per dollar spent on
 308 surveillance.

309

310 **Table 2.** Examples of objective functions for optimization analysis of surveillance networks

Objective function type	Description	Example objective functions
Minimize mean error magnitudes	On average, how different a quantity, Q_i , measured or estimated from the ascertained data $\mathcal{J}(\theta, d)$, is from the same quantity, Q_D , estimated or measured from the underlying disease data D . Includes mean squared error, mean squared percentage error, root mean squared error, mean absolute error, mean absolute percentage error, or other expressions.	To better characterize geographic, temporal, or demographic distribution of disease , the objective function may be expressed as: $f = \sum_{i=1}^n (C_{I,i} - C_{D,i})^2 / n$ n – number of subpopulations $C_{I,i}$ – number of ascertained cases in subpopulation i $C_{D,i}$ – number of true cases from D in subpopulation i
		To assess the impact of interventions more accurately, the objective function may be expressed as: $f = \left \frac{D_I}{F_I} - \frac{D_D}{F_D} \right $ D_I – observed number of cases in ascertained cases after intervention F_I – predicted number of ascertained cases in absence of the intervention D_D – observed number of cases in all cases after intervention F_D – predicted number of all cases in absence of the intervention
Minimize uncertainty of surveillance estimands	If bias in surveillance sampling and estimation is not a concern (e.g. for asymptotically unbiased estimators), then minimizing uncertainty may be the primary goal. Uncertainty can be represented by standard error, standard deviation, inter-quantile range, or other expressions.	To determine the effect of a risk factor on infection more precisely when assuming a linear relationship between the risk factor and disease rate, the objective function may be expressed as: $f = \text{var}(\hat{\beta}_I)$ $\hat{\beta}_I$ – estimated regression coefficient of the effect of the risk factor on the disease rate from the ascertained data
		To forecast the peak case count more precisely, the objective function may be expressed as: $f = \text{var}(P_I)$ P_I – forecasted peak case count based on ascertained data overall or for a specific area
Maximize log-likelihood	If a probability distribution $Q_I \sim Q(\theta, \dots)$ can be expressed by the surveillance model, then maximizing the likelihood of true data Q_D under the estimated distribution can be used to simultaneously address bias and variance.	To better estimate the effect of a risk factor on infection rates when assuming a linear relationship between the risk factor and disease rate, the objective function may be expressed as: $f = \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\hat{\beta}_I - \beta_D}{\sigma} \right)^2} \right),$ if a normal distribution with a variance of σ^2 is assumed for the true effect of a risk factor β_D

		<p>To improve estimation of outbreak probabilities, the objective function may be expressed as: $f = \sum_{t=1}^T [Y_t \log(\hat{p}_t) + (1 - Y_t) \log(1 - \hat{p}_t)]$, if outbreak probabilities in subsequent weeks are assumed to be conditionally independent. \hat{p}_t – estimated outbreak probability in time period t Y_t – indicator (0 or 1) for actual occurrence of an outbreak in time period t</p>
<p>Maximize classification performance</p>	<p>When Q_I and Q_D are categorical, the performance of the surveillance system can be measured by classification evaluation metrics, such as sensitivity, specificity, positive predictive value, F1 scores, area under the receiver operating characteristic curve, etc.</p>	<p>To improve our ability to discriminate outbreaks from false alarms, the objective function may be expressed as the area under the ROC curve: $f = \int_0^1 \pi_{tp}(\pi_{fp}) d\pi_{fp}$ π_{tp} – proportion of true outbreaks correctly identified π_{fp} – proportion of non-outbreak time periods falsely identified as outbreaks</p>
		<p>To improve our ability to detect a rare disease, the objective function may be expressed as the maximum of the average F_1 score: $f = 2 \int_0^1 \frac{\pi_{tp p}(\tau) \times \pi_{tp}(\tau)}{\pi_{tp p}(\tau) + \pi_{tp}(\tau)} d\tau$ π_{tp} – proportion of true cases reported $\pi_{tp p}$ – proportion of reported cases that are true τ – threshold condition for reporting a case, assumed in this example to represent a probability</p>

311

312 3.4 Simulation optimization search

313 The goal of the optimization process (*while* block in Box 1; the loop in *Simulation*
 314 *optimization search* component of Figure 1) is to thoroughly explore the response surface of the
 315 objective function(s) over the design space so as to identify designs likely to yield optimal or
 316 near-optimal surveillance performance. Candidate surveillance designs are drawn from the
 317 design space, and the expectations of resulting objective function values across realizations are
 318 evaluated with respect to the simulated true and ascertained disease data; this process is
 319 repeated iteratively until a stopping criterion is reached, e.g., the convergence on estimated
 320 optimum(a); exhaustive sampling of the design space; or the exceedance of a computational
 321 budget. When the design parameter space is small, exhaustive evaluation of objective function
 322 values across the entire design parameter space is possible. Sufficient and efficient searching
 323 of large design parameter spaces, by contrast, requires heuristic or metaheuristic optimization
 324 algorithms (e.g., simulated annealing, genetic algorithms, particle swarm optimization, or
 325 Bayesian model based optimization).

326 Multiple surveillance objectives can be optimized simultaneously through multi-objective
 327 optimization approaches, such as through weighted sums of objective functions or Pareto
 328 optimization [40]. Generating weighted sums of objective function values allows for the
 329 specification of relative importance of different objectives. If one objective is less important, it
 330 would be assigned a smaller weight when compared with other objectives. In this way, optimal
 331 designs are not overly influenced by less important objectives. Pareto optimization outputs a set
 332 of optimal solutions (Pareto optimal set) for which no other solutions can perform better under
 333 all objectives. That is, improving the performance on one objective leads to worsening at least
 334 one of the other objectives. Decision makers are then tasked with choosing the “best” design

335 from the Pareto optimal set based on other considerations. Multi-objective optimization in the
336 presence of a large design space can be handled by modified metaheuristic algorithms [41]. For
337 example, to accommodate multiple objectives, Pareto simulated annealing approaches seek to
338 express the acceptance probability of a new design as a function of its improvements in all
339 objectives when compared with the current best design [42].
340

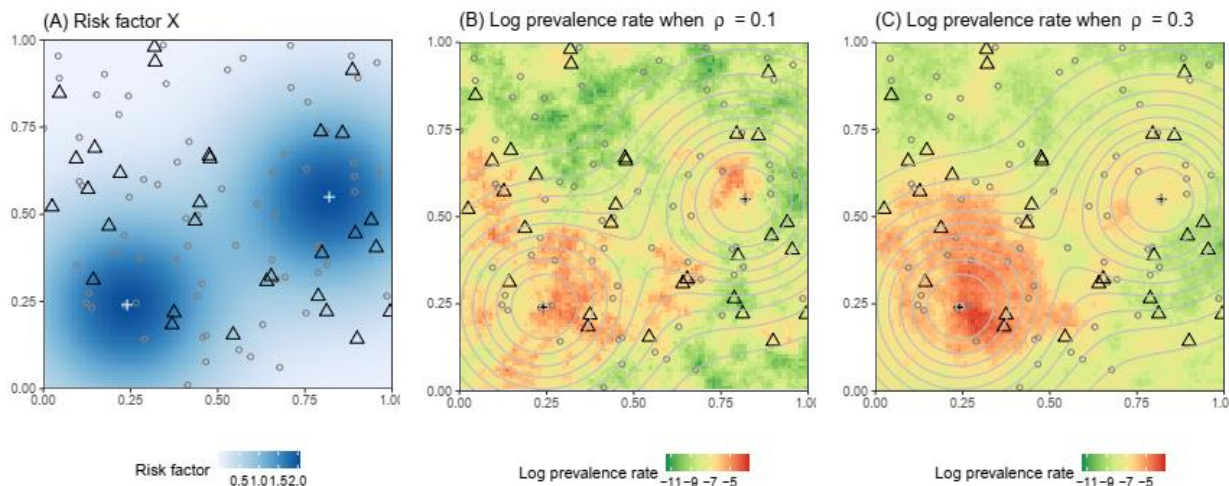
341 **4 Demonstration of the surveillance simulation and optimization** 342 **framework: optimal selection of new surveillance sites**

343 Here, we demonstrate an application of our surveillance system optimization framework
344 in the context of selecting candidate sites to add to an existing cross-sectional survey network.
345 We consider two surveillance design objectives—optimal prediction of the geographical
346 distribution of the disease (hereafter referred to as *spatial prediction*) and optimal estimation of
347 the effect of a risk factor (hereafter referred to as *effect estimation*). We demonstrate how
348 optimal designs can vary in relation to epidemiological characteristics of the target disease; in
349 this case, the rate of decrease in correlation of disease prevalence rates over distance, which
350 determines whether prevalence changes abruptly or smoothly over the spatial domain.

351 We first describe the demonstration setting, the data available for design optimization,
352 the specification and parameterization of the disease and surveillance system models, and the
353 resulting formalized objective functions for optimizing spatial predictions and effect estimation.
354 We demonstrate the use of an exhaustive search strategy to find the single most optimal site to
355 add to the existing network for both goals, as well as the Pareto-optimal set of single sites to
356 add when considering both objectives simultaneously. We simulate the addition of an arbitrary
357 number of sites, acknowledging that in real-world applications of the framework, the number of
358 sites might be determined by budgetary constraints and/or the marginal informational gains per
359 added site. We conclude our demonstration by considering the best set of three sites to add,
360 which introduces substantial combinatorial complexity, motivating the use of a metaheuristic
361 algorithm to efficiently search for optimal regions of design space.

362 363 **4.1 Demonstration setting**

364 We generated a set of 100 potential surveillance sites scattered uniformly at random
365 across a unit grid, and randomly selected 30 sites to represent a virtual existing surveillance
366 network. We seeded two point sources for a risk factor influencing expected disease prevalence
367 rates (Figure 2A), then simulated disease prevalence under two scenarios of spatial auto-
368 correlation by adjusting the scale parameter (ρ) of a log-Gaussian spatial process centered on a
369 linear function of the risk factor. We refer to these as spatially patchy ($\rho = 0.1$; Figure 2B) and
370 spatially smooth ($\rho = 0.3$; Figure 2C) disease scenarios. Additional details of data generation
371 are provided in Text S1.
372



373
374 **Figure 2. Simulated data used for surveillance system optimization.** Spatial variation of (A)
375 the risk factor X and (B) log prevalence when $\rho = 0.1$ and (C) $\rho = 0.3$. Triangles represent the
376 existing 30 surveyed sites; dots represent the 70 candidate sites; crosses represent two point
377 sources of the risk factor of interest (e.g. locations of mass gatherings); background color in
378 Panel A and contour lines in panels B and C represent the levels of risk factor X . Three
379 realizations of the log prevalence surface when $\rho = 0.1$ or 0.3 are shown in Figure S1.

380

381 4.2 Data and knowledge inputs

382 Available epidemiologic data to characterize the relevant aspects of the disease system
383 include simulated prevalence rates observed at the 30 sites enrolled in the surveillance network.
384 Data characterizing the surveillance system and design space include the coordinates of the 30
385 enrolled and 70 candidate sites. Additional data to support optimization include levels of risk
386 factor X at each sampling location. Theoretical inputs include the assumption of a log-linear
387 relationship between X and disease prevalence, and that spatial disease prevalence residuals
388 follow a Gaussian process with exponential covariance function.

389

390 4.3 Set up and initialization

391 **Disease system model specification and simulation.** In this demonstration, relevant
392 aspects of the disease system include the correlation of disease outcomes over space, as well
393 as the association of disease outcomes with risk factor X . Based on the observed disease
394 prevalence at participating sites, we assume the log of the prevalence value Y is generated from
395 an underlying random spatial process with an *i.i.d* mean-zero normally distributed noise ε with a
396 variance of σ_d^2 , and can be modelled by:

$$397 Y = \exp(\beta_0 + \beta_1 X + \eta + \varepsilon),$$

398 where β_0 represents log of the overall mean prevalence rate, β_1 represents the effect of a unit
399 increase in risk factor X , and η represents a mean-zero Gaussian process accounting for spatial
400 correlation induced by additional dependence not captured by X . The spatially correlated error
401 term η follows a multivariate normal distribution with a variance-covariance matrix \mathbf{C} , in which
402 each entry c_{ij} represents the covariance between the residuals at the i th and the j th location
403 when $i \neq j$, and the variance of the residuals at the i th location when $i = j$. Covariance between

404 sites i and j is specified $c_{ij} = \sigma_s^2 e^{-d_{ij}/\rho}$, where d_{ij} is the distance between sites i and j , and ρ is the
405 scale parameter as before; and the variance at site i is $\sigma_d^2 + \sigma_s^2$. The correlation of the residuals
406 between two sites as a function of the distance between them is shown in Figure S1.
407 Parameters β_0 , β_1 , σ_s , σ_d , and ρ were estimated based on the prevalence rates and risk factor
408 levels at the 30 in-network sites, after which 1000 realizations of log-prevalence rates at the 70
409 candidate sites were drawn according to the fitted parameters, observed prevalence at in-
410 network sites, risk factor levels at candidate sites, and distance matrix between in-network and
411 candidate sites.

412 **Surveillance model specification.** Relevant aspects of information captured by the
413 surveillance system in this demonstration pertain to the extrapolation of prevalence from
414 enrolled to unenrolled sites, as well as the variance in X at enrolled sites. Assuming perfect
415 enumeration of disease prevalence at each enrolled site, as well as known values of the risk
416 factor X for all sites, information drawn by each candidate design is represented by
417 improvements in estimates of β_1 and predictions at 70- n out-of-network sites obtained by fitting
418 a universal kriging predictor to data from enrolled sites [43].

419 **Design space.** In our hypothetical example, we have an existing network of 30
420 surveillance sites $\{s_1 \dots s_{30}\}$, and 70 additional locations $\{s_{31} \dots s_{100}\}$ from which we may select n
421 new sites to be added to the network. Therefore, our design parameter s_θ is the set of n sites to
422 be added to the network, and the discrete design space is all possible sets of n sites chosen
423 from 70.

424

425 4.4 Optimization

426 **Objective functions: Spatial interpolation.** The first surveillance function we wish to
427 optimize is prediction of the geographical distribution of the disease. Therefore, we define the
428 objective function as the mean squared error (MSE) of log predicted prevalence at the 70- n out-
429 of-network locations after adding s_θ to the network:

$$430 f_1(s_\theta) = \sum_{k=1}^{1000} \sum_{j=1}^{70-n} \left(Y_{d_{k,j}} - \hat{Y}_{d_{k,j}}(s_\theta) \right)^2 / ((70 - n) * 1000),$$

431 where $Y_{d_{k,j}}$ represents the log prevalence rate at the j th unenrolled site in the k th disease
432 system model realization, while $\hat{Y}_{d_{k,j}}(s_\theta)$ represents the predicted log prevalence rate at the j th
433 site after augmenting the existing network with s_θ in the k th realization. We denote the objective
434 function value for this objective as OFV1. Other objective functions, such as the MSE of log
435 predicted prevalence rate at the existing 30 sites or across all 100 sites, can also be used.
436 Existing literature on optimal spatial design provides more options for relevant objective
437 functions [44-46].

438 **Objective functions: Effect estimation.** Our second surveillance goal is precise
439 estimation of the effect of the risk factor X on the disease outcome, so the objective function is
440 formalized as:

$$441 f_2(s_\theta) = \sum_{k=1}^{1000} \text{var}(\hat{\beta}_{1d_k}(s_\theta)) / 1000,$$

442 where $\hat{\beta}_{1d_k}(s_\theta)$ represents the estimate of β_1 after augmenting the existing network with s_θ in the
443 k th disease system model realization.

444

445

446 **Search algorithms.** When a single site is to be added to the network, the design space
447 is small enough to allow for enumeration of objective function values at all possible designs.
448 Therefore, the algorithm for proposing new designs simply steps sequentially through sites
449 $\{s_{31} \dots s_{100}\}$. However, when the optimization question is shifted to the best three sites to add,
450 the design space expands to 54,740 combinations, making sequential enumeration a
451 prohibitively expensive search strategy. Under these conditions, heuristic (greedy) or
452 metaheuristic algorithms play an important role in finding the optimal or near-optimal solution
453 within a reasonable amount of time [47]. Moreover, the evaluation of objective function values
454 across realizations can be paralleled to further reduce computational time.

455 We illustrate the use of a simulated annealing (SA) meta-heuristic algorithm popular in
456 spatial sampling network design [48, 49] to more efficiently explore the design space when three
457 sites are to be added. In SA, a random initial design is proposed, after which, at each iteration, a
458 new design is sampled from the neighborhood of the current design and the objective function
459 value (OFV) for the new design is evaluated. Here, the neighborhood of a set of n sites to enroll
460 is defined as designs sharing $n-1$ sites with the current design. If the new OFV is superior to the
461 current OFV, the new design is accepted as the next design; otherwise, it is accepted with a
462 probability of $e^{-\frac{\Delta OFV}{T}}$, where ΔOFV is the change in the OFV and T is a tuning parameter
463 analogous to temperature [50]. T decreases at a rate α after each iteration, causing SA to
464 accept deterioration in the OFV more frequently at the beginning of the run and rarely towards
465 the end. Probabilistically accepting worse solutions early in the search enables the algorithm to
466 escape local optima. For our demonstration, we set the initial temperature T_0 and cooling rate α
467 separately for each objective and epidemiologic scenario, following suggestions by Sait and
468 Youssef [50], and set the stopping criteria is to be $T \leq 10^{-6}$. We repeat the SA process 3 times to
469 examine the convergence of the result.

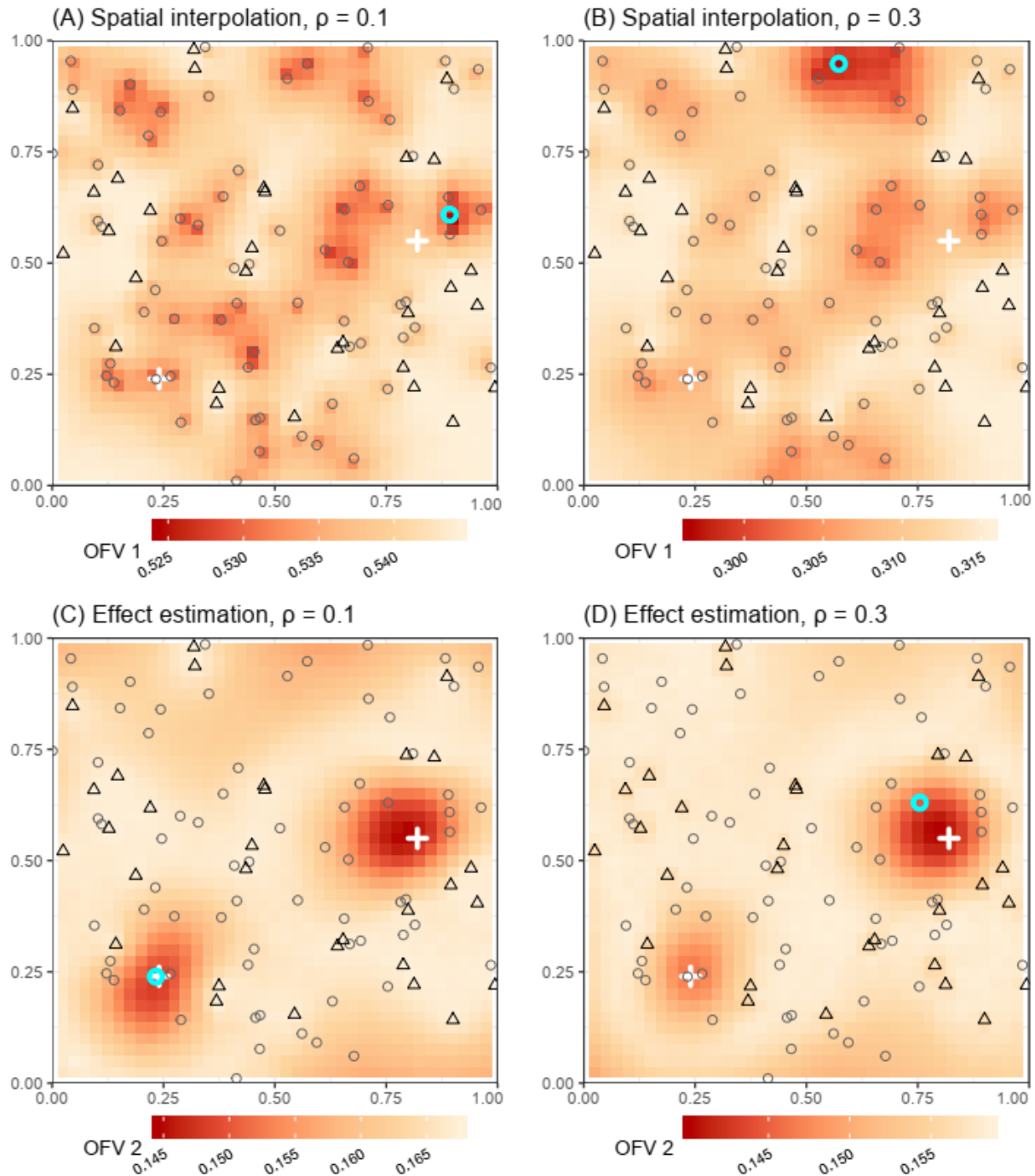
470

471 **4.5 Optimal surveillance designs**

472 **Selecting one additional site to optimize spatial prediction.** The mean squared error
473 of spatial predictions across unenrolled sites (OFV1) is minimized by enrolling sites that are in
474 close proximity to multiple out-of-network sites — especially clusters of unmeasured sites at
475 long distances from existing network locations (Figure 3, panels A and B). These optimal
476 placements address informational gaps by enrolling sites that increase the average covariance
477 between measured and unmeasured locations, and tend to fall in areas close to several
478 unenrolled sites but away from the initially enrolled locations. Furthermore, the amount of
479 information that can be inferred from the same set of neighboring sites increases with the scale
480 parameter ρ . Thus, in the spatially patchy disease scenario, where the scale of spatial
481 autocorrelation is small, optimal placement occurs in the center of a tight cluster of unenrolled
482 sites (Figure 3, panel A). Under the spatially smooth scenario, the same cluster is correlated
483 with initially enrolled sites, and optimal site placement falls in the center of a loose cluster of
484 unmeasured sites located quite far from the initial network (Figure 3, panel B). Under the
485 parameter values used to generate demonstration data, there is no clear influence of risk factor
486 level X on site selection to optimize spatial prediction.

487 **Selecting one additional site to optimize effect estimation.** The variance of the effect
488 of risk factor X on log disease prevalence (OFV2) is lower when the value of X at the cell to be

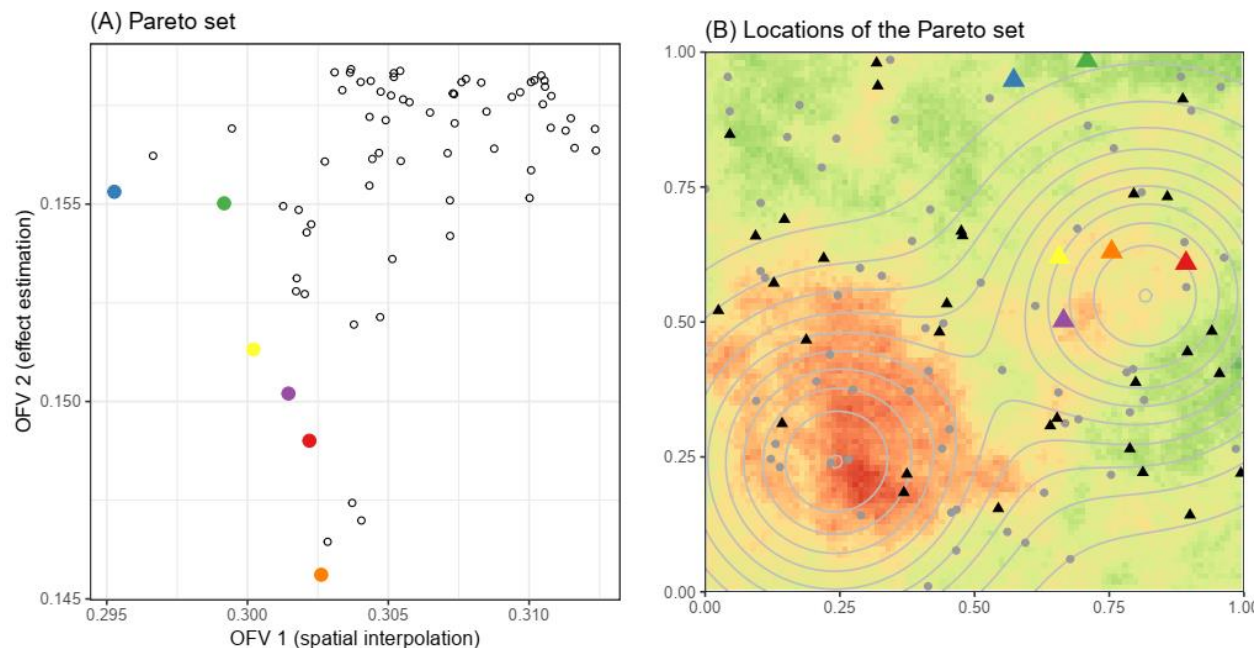
489 added lies towards an extreme of X 's observed range and when the site to be added is relatively
490 uncorrelated with (i.e., distant from) initially enrolled sites (Figure 3, panels C and D). In the
491 spatially patchy disease scenario, where the scale of spatial autocorrelation is limited, optimal
492 site placement is dominated by the level of risk factor X , and the available site with highest X is
493 chosen (Figure 3C). In the spatially smooth scenario, with an extended scale of spatial
494 autocorrelation, the correlation of outcomes between the site with the highest X level and
495 nearby initially enrolled sites results in selection of an alternative location where the value of X is
496 less extreme, but prevalence is expected to be more independent of previously observed
497 outcomes (Figure 3D).
498



499
500
501
502
503
504
505
506
507

Figure 3. Optimal site placement to augment a surveillance network for spatial prediction or effect estimation under scenarios of spatially patchy or smooth disease distributions. Black triangles represent initially enrolled sites, gray circles represent unselected candidate sites, and the cyan circle indicates the optimal site to add to the network. White crosses represent point sources for risk factor X. Raster colors represent objective function values for hypothetical sites added across a regular 41*41 grid in order to visualize the response surface in relation to initial network locations and the underlying risk factor.

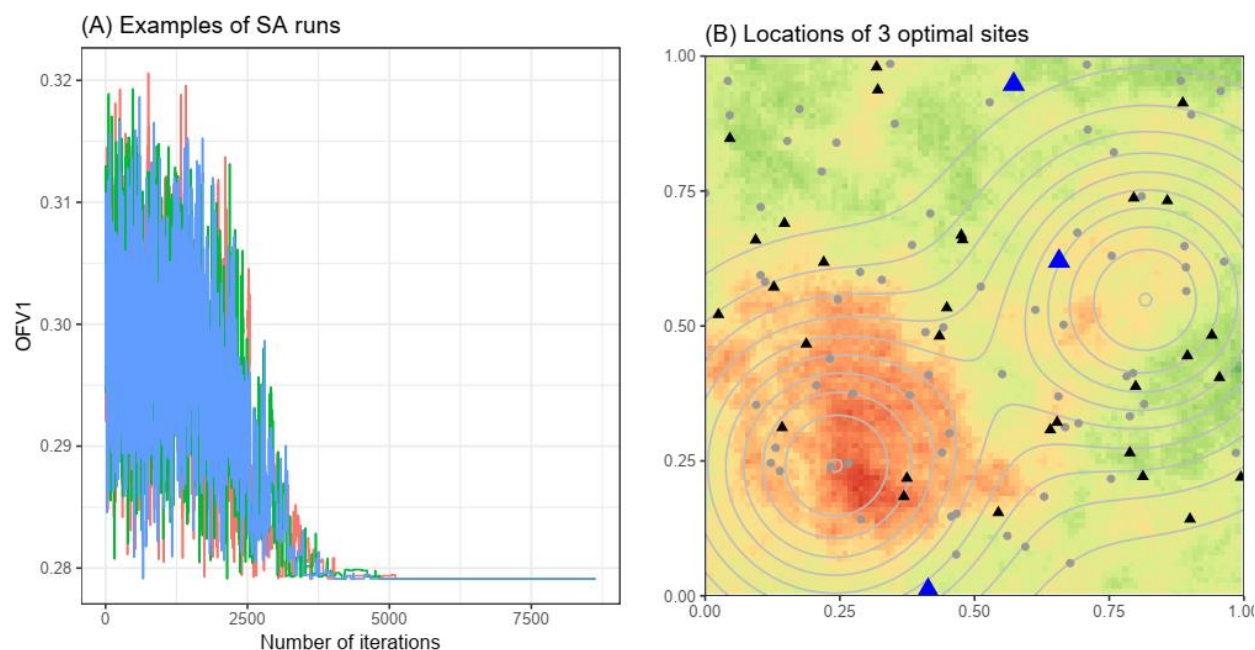
508 **Single site selection based on multiple objectives.** When simultaneously optimizing site
509 enrollment for spatial prediction and effect estimation, the output is a Pareto optimal set
510 containing designs that are considered equally optimal because no objective function value can
511 be improved without impairing the other objective function values. A set of six candidate sites
512 emerges for the spatially smooth disease scenario, including four alternative selections to the
513 optimal locations for each single objective (Figure 4). The Pareto optimal set for the spatially
514 patchy scenario includes only one non-dominated site in addition to the optimal locations for
515 either objective individually (Figure S2). Since the solution given by Pareto optimization is not
516 unique, some way of reconciling the objective criteria, such as a weighted sum or expression of
517 total cost may be required to choose the optimal design. Notably, we did not incorporate cost
518 associated with adding sites in our analysis, but this could be accomplished by including a third
519 objective function representing the marginal information gain per added site. In this case, the
520 spatial prediction OFV, effect estimation OFV, and the cost-effectiveness OFV would be jointly
521 optimized.
522
523



524
525 **Figure 4. Results from Pareto optimization under the spatially smooth disease scenario**
526 **($\rho=0.3$).** (A) Mean squared error of log predicted disease prevalence (OFV1) and variance of
527 causal effect estimate (OFV2) of the Pareto set (colored dots) and all other candidate sites
528 (hollow dots). (B) Locations of the Pareto set (colored triangles) colored coded as in Panel A.
529 Black triangles represent initially enrolled sites, and gray dots represent unchosen candidate
530 sites. Background color in Panel B represents log prevalence when $\rho = 0.3$ using the same
531 color scheme as in Figure 2C, while contour lines represent levels of risk factor X.
532

533 **Selecting three additional sites to optimize spatial prediction.** As a final example, we
534 demonstrate the use of metaheuristic algorithms to search larger design spaces, applying
535 simulated annealing to select three additional sites out of seventy candidate sites

536 simultaneously. Simulated annealing optimizations seeded with different initial designs
537 converged to the same best set of three additional sites to enroll for enhanced spatial prediction
538 under the spatially smooth disease scenario (Figure 5). All three SA runs (Figure 5A, colored
539 lines) converged to the same optimal design within 6,000 iterations. Given the parameters and
540 the stopping criteria we used, each run terminated after 8,630 iterations. Even with three runs,
541 the total number of objective function evaluations was 25,890, less than half of what would be
542 required if using enumeration. Figure 5B shows the location of the optimal three-site set. The
543 results effect estimation, as well as for the spatially patchy outcome scenario are shown in
544 Figures S3-5.
545



546
547 **Figure 5. Metaheuristic optimization with simulated annealing (spatial prediction,**
548 **spatially smooth disease scenario).** (A) Mean squared error of predicted log prevalence
549 (OFV1) across iterations of three SA runs. (B) The locations of the optimal 3 sites. Black
550 triangles represent existing sites, blue triangles represent the optimal additional sites, and gray
551 dots represent unchosen alternative sites. Background color in Panel B represents log
552 prevalence when $\rho = 0.3$ using the same color scheme as in Figure 2C, while contour lines
553 represent levels of risk factor X.
554

555 5. Conclusion

556 Surveillance system designs that provide reliable, timely estimates of the spatial-
557 temporal distributions of endemic and epidemic diseases, are critical to the efficient allocation of
558 resources for public health responses. However, opportunities to apply numerical optimization to
559 surveillance system design have heretofore been overlooked in the literature. In this paper, we
560 have presented a framework for surveillance optimization via simulation to enhance design
561 decision making and facilitate research into optimal design principles under uncertain or

562 changing epidemiological conditions. While we focus on surveillance of human disease, the
563 framework could also be applied to the optimization of vector or environmental surveillance.

564 The framework presented can arrive at improved surveillance system designs through
565 the incorporation of data and models of local disease transmission status, diverse surveillance
566 goals, resource and operational constraints, and by stimulating collaboration between health
567 planners, researchers, and software developers. However, it should also be recognized that the
568 rationality of the output optimal design will be highly dependent on the accuracy and relevance
569 of data or models used to represent disease and surveillance processes during optimization, as
570 well as the performance of the optimization search algorithm. There is much future work to be
571 done to develop and validate simulation models that can represent relevant case generating
572 and measurement processes accurately; to analyze the sensitivity of optimal design to the
573 specification of disease system models and changes in disease epidemiology; and to adopt
574 optimization approaches from related fields—such as environmental monitoring network design
575 and signal processing [51-53]—to disease surveillance design applications.

576

577 **References**

- 578 1. Teutsch SM, Churchill RE. Principles and practice of public health surveillance: Oxford
579 University Press, USA; 2000.
- 580 2. Fournet F, Jourdain F, Bonnet E, Degroote S, Ridde V. Effective surveillance systems
581 for vector-borne diseases in urban settings and translation of the data into action: a scoping
582 review. *Infectious diseases of poverty*. 2018;7(1):99.
- 583 3. Venkatramanan S, Chen J, Fadikar A, Gupta S, Higdon D, Lewis B, et al. Optimizing
584 spatial allocation of seasonal influenza vaccine under temporal constraints. *PLoS computational
585 biology*. 2019;15(9):e1007111.
- 586 4. Zhao Z, Zhao J, Ma J. Conception of an integrated information system for notifiable
587 disease communicable surveillance in China. *Disease Surveillance*. 2018;33(5):423-7.
- 588 5. Neilan RLM, Schaefer E, Gaff H, Fister KR, Lenhart S. Modeling optimal intervention
589 strategies for cholera. *Bulletin of mathematical biology*. 2010;72(8):2004-18.
- 590 6. Bowong S, Alaoui AA. Optimal intervention strategies for tuberculosis. *Communications
591 in Nonlinear Science and Numerical Simulation*. 2013;18(6):1441-53.
- 592 7. Eisenberg JN, Scott JC, Porco T. Integrating disease control strategies: balancing water
593 sanitation and hygiene interventions to reduce diarrheal disease burden. *American Journal of
594 Public Health*. 2007;97(5):846-52.
- 595 8. Cooper BS, Stone S, Kibbler C, Cookson B, Roberts J, Medley G, et al. Systematic
596 review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus
597 aureus*: a review of the literature with epidemiological and economic modelling. *Health
598 Technology Assessment (Winchester, England)*. 2003;7(39):1-194.
- 599 9. Kirschner D, Lenhart S, Serbin S. Optimal control of the chemotherapy of HIV. *Journal of
600 mathematical biology*. 1997;35(7):775-92.
- 601 10. Murphy SA. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society:
602 Series B (Statistical Methodology)*. 2003;65(2):331-55.
- 603 11. Center for Disease Control. Swine influenza A (H1N1) infection in two children—
604 Southern California, March–April 2009. *Morbidity and Mortality Weekly Report*. 2009;58:400-2.
- 605 12. Lau MS, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, et al. Spatial and temporal
606 dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *Proceedings
607 of the National Academy of Sciences*. 2017;114(9):2337-42.
- 608 13. World Health Organization. Global epidemiological surveillance standards for influenza.
609 2013.

- 610 14. Scarpino SV, Dimitrov NB, Meyers LA. Optimizing provider recruitment for influenza
611 surveillance networks. *PLoS computational biology*. 2012;8(4):e1002472.
- 612 15. Russell S, Ryff K, Gould C, Martin S, Johansson M. Detecting local Zika virus
613 transmission in the continental United States: a comparison of surveillance strategies. *PLoS*
614 *currents*. 2017;9.
- 615 16. Herrera JL, Srinivasan R, Brownstein JS, Galvani AP, Meyers LA. Disease surveillance
616 on complex social networks. *PLoS computational biology*. 2016;12(7):e1004928.
- 617 17. Adhikari B, Lewis B, Vullikanti A, Jiménez JM, Prakash BA. Fast and near-optimal
618 monitoring for healthcare acquired infection outbreaks. *PLoS computational biology*.
619 2019;15(9):e1007284.
- 620 18. Chinese National Influenza Center. National Influenza Surveillance Guidelines.
621 Beijing2010.
- 622 19. Gonsalves GS, Copple JT, Johnson T, Paltiel AD, Warren JL. Bayesian adaptive
623 algorithms for locating HIV mobile testing services. *BMC medicine*. 2018;16(1):155.
- 624 20. Yaesoubi R, Cohen T. Identifying dynamic tuberculosis case-finding policies for HIV/TB
625 coepidemics. *Proceedings of the National Academy of Sciences*. 2013;110(23):9457-62.
- 626 21. Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, Rushton G. Optimizing
627 influenza sentinel surveillance at the state level. *American journal of epidemiology*.
628 2009;170(10):1300-6.
- 629 22. Shisana O, Rehle T, Simbayi L. South African national HIV prevalence, HIV incidence,
630 behaviour and communication survey, 2005: HSRC press; 2005.
- 631 23. Centers for Disease Control and Prevention. US Zika Pregnancy and Infant Registry
632 2019 [cited 2019 December 30th]. Available from:
633 <https://www.cdc.gov/pregnancy/zika/research/registry.html>.
- 634 24. Susilarini NK, Sitorus M, Praptaningsih CY, Sampurno OD, Bratasena A, Mulyadi E, et
635 al. Application of WHO's guideline for the selection of sentinel sites for hospital-based influenza
636 surveillance in Indonesia. *BMC health services research*. 2014;14(1):424.
- 637 25. Yukich JO, Butts J, Miles M, Berhane Y, Nahusenay H, Malone JL, et al. A description of
638 malaria sentinel surveillance: a case study in Oromia Regional State, Ethiopia. *Malaria journal*.
639 2014;13(1):88.
- 640 26. Satter SM, Gastanaduy PA, Islam K, Rahman M, Rahman M, Luby SP, et al. Hospital-
641 based surveillance for rotavirus gastroenteritis among young children in Bangladesh: defining
642 the potential impact of a rotavirus vaccine program. *The Pediatric infectious disease journal*.
643 2017;36(2):168.
- 644 27. Zoni AC, Gonzalez MA, Sjoegren HW. Syphilis in the most at-risk populations in Latin
645 America and the Caribbean: a systematic review. *International Journal of Infectious Diseases*.
646 2013;17(2):e84-e92.
- 647 28. Hoa NB, Sy DN, Nhung NV, Tiemersma EW, Borgdorff MW, Cobelens FG. National
648 survey of tuberculosis prevalence in Viet Nam. *Bulletin of the World Health Organization*.
649 2010;88:273-80.
- 650 29. Surveillance Sector. Malaysia Influenza Surveillance Protocol. In: Disease Control
651 Division MoH, Malaysia, editor. 2015.
- 652 30. Wang J, Teng Z, Cui X, Li C, Pan H, Zheng Y, et al. Epidemiological and serological
653 surveillance of hand-foot-and-mouth disease in Shanghai, China, 2012–2016. *Emerging*
654 *microbes & infections*. 2018;7(1):1-12.
- 655 31. Public Health Surveillance. Virological Surveillance 2018 [cited 2019 December 30th].
656 Available from: <https://surv.esr.cri.nz/virology/virology.php>.
- 657 32. National Institute for Viral Disease Control and Prevention. National Dengue
658 Surveillance Guideline. Beijing2011.
- 659 33. National Health and Family Planning Commission of the People's Republic of China.
660 Diagnostic criteria for pertussis (WS 274-2007). Beijing2007.

- 661 34. Centers for Disease Control and Prevention. Pertussis (Whooping Cough) Surveillance
662 & Reporting 2019 [cited 2019 December 17th]. Available from:
663 <https://www.cdc.gov/pertussis/surv-reporting.html>.
- 664 35. World Health Organization. WHO surveillance case definitions for ILI and SARI 2014
665 [cited 2019 December 17th]. Available from:
666 https://www.who.int/influenza/surveillance_monitoring/ili_sari_surveillance_case_definition/en/.
- 667 36. Rowthorn RE, Laxminarayan R, Gilligan CA. Optimal control of epidemics in
668 metapopulations. *Journal of the Royal Society Interface*. 2009;6(41):1135-44.
- 669 37. Blayneh K, Cao Y, Kwon H-D. Optimal control of vector-borne diseases: treatment and
670 prevention. *Discrete and Continuous Dynamical Systems B*. 2009;11(3):587-611.
- 671 38. Medlock J, Galvani AP. Optimizing influenza vaccine distribution. *Science*.
672 2009;325(5948):1705-8.
- 673 39. Bussell EH, Dangerfield CE, Gilligan CA, Cunniffe NJ. Applying optimal control theory to
674 complex epidemiological models to inform real-world disease management. *Philosophical
675 Transactions of the Royal Society B*. 2019;374(1776):20180284.
- 676 40. Ehrgott M. *Multicriteria optimization*: Springer Science & Business Media; 2005.
- 677 41. Gandibleux X, Sevaux M, Sörensen K, T'kindt V. *Metaheuristics for multiobjective
678 optimisation*: Springer Science & Business Media; 2004.
- 679 42. Czyżżak P, Jaskiewicz A. Pareto simulated annealing—a metaheuristic technique for
680 multiple - objective combinatorial optimization. *Journal of Multi - Criteria Decision Analysis*.
681 1998;7(1):34-47.
- 682 43. Gelfand AE, Diggle P, Guttorp P, Fuentes M. *Handbook of spatial statistics*: CRC press;
683 2010.
- 684 44. Zimmerman DL. Optimal network design for spatial prediction, covariance parameter
685 estimation, and empirical prediction. *Environmetrics: The official journal of the International
686 Environmetrics Society*. 2006;17(6):635-52.
- 687 45. Diggle P, Lophaven S. Bayesian geostatistical design. *Scandinavian Journal of
688 Statistics*. 2006;33(1):53-64.
- 689 46. Mateu J, Müller WG. *Spatio-temporal design: Advances in efficient data acquisition*:
690 John Wiley & Sons; 2012.
- 691 47. Blum C, Roli A. *Metaheuristics in combinatorial optimization: Overview and conceptual
692 comparison*. *ACM computing surveys (CSUR)*. 2003;35(3):268-308.
- 693 48. Van Groenigen J, Stein A. Constrained optimization of spatial sampling using continuous
694 simulated annealing. *Journal of Environmental Quality*. 1998;27(5):1078-86.
- 695 49. Brus D. *Sampling for digital soil mapping: A tutorial supported by R scripts*. *Geoderma*.
696 2019;338:464-80.
- 697 50. Sait SM, Youssef H. *Iterative computer algorithms with applications in engineering:
698 Solving Combinatorial Optimization Problems* Wiley-IEEE Computer Society Press; 2000.
- 699 51. Le ND, Zidek JV. *Statistical analysis of environmental space-time processes*: Springer
700 Science & Business Media; 2006.
- 701 52. Thomopoulos SC, Viswanathan R, Bougoulas DC. Optimal decision fusion in multiple
702 sensor systems. *IEEE Transactions on Aerospace and Electronic Systems*. 1987;(5):644-53.
- 703 53. Lin CY, Waller LA, Lyles RH. The likelihood approach for the comparison of medical
704 diagnostic system with multiple binary tests. *Journal of Applied Statistics*. 2012;39(7):1437-54.
705