

Design Strategies for Efficient Arbovirus Surveillance

Samuel V. Scarpino, Lauren Ancel Meyers,
Michael A. Johansson

As public health agencies struggle to track and contain emerging arbovirus threats, timely and efficient surveillance is more critical than ever. Using historical dengue data from Puerto Rico, we developed methods for streamlining and designing novel arbovirus surveillance systems with or without historical disease data.

Mosquitoborne viruses in the families *Flaviviridae* and *Togaviridae* cause substantial illness and death worldwide (1,2). Dengue is the most widespread arboviral disease, with an estimated 70–140 million cases occurring annually (3). Despite the large public health and economic costs of arboviruses, effective medical countermeasures are limited (1). Globally, primary arbovirus prevention and control efforts include personal protection, mosquito control, and clinical treatment. The success of these efforts depends on timely and accurate situational awareness: knowing spatiotemporal patterns of exposure, infection, and severity.

Puerto Rico has an islandwide passive dengue surveillance system similar to those found in other regions with endemic dengue (4). Healthcare providers (clinics or hospitals) report suspected dengue cases and submit blood samples for laboratory diagnosis. This comprehensive system captures spatiotemporal variation in incidence and enables characterization of circulating viruses, but it requires substantial resources and may lack efficiency.

Here, we extend a previous approach (5) to designing dengue surveillance systems with 4 sets of specific public health objectives: real-time estimation of island-wide dengue cases, regional dengue cases, island-wide cases of each dengue virus serotype, and all three preceding quantities combined. Using dengue case data from 1991 through 2005, we identified a surveillance system including a subset of Puerto Rican providers that was expected to achieve these objectives efficiently and demonstrated the robustness of that system with data for 2006–2012.

Author affiliations: University of Vermont, Burlington, Vermont, USA (S.V. Scarpino); Santa Fe Institute, Santa Fe, New Mexico, USA (S.V. Scarpino, L.A. Meyers); University of Texas at Austin, Austin, Texas, USA (L.A. Meyers); Centers for Disease Control and Prevention, San Juan, Puerto Rico, USA (M.A. Johansson); Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA (M.A. Johansson)

DOI: <http://dx.doi.org/10.3201/eid2304.160944>

The Study

Across Puerto Rico, we analyzed the weekly number of suspect cases, laboratory-positive cases, and cases of each serotype reported during 1991–2012. For each case, we considered the patient's municipality of residence and the identity of the reporting provider.

In designing a multipurpose dengue surveillance system, we sought to identify a small subset of providers that can provide accurate situational awareness. However, it is computationally unfeasible to evaluate all possible combinations of providers. Our procedure for solving this computational issue is described in the following sections, with a detailed description in the online Technical Appendix (<https://wwwnc.cdc.gov/EID/article/23/4/16-0944-Techapp1.pdf>).

Building from previous research (6), we design surveillance systems by sequentially adding providers that most improve system performance. To evaluate the performance of a system with respect to an objective, we repeatedly perform the following: fit multilinear models to historical reported dengue cases, use the fitted models to estimate dengue cases in another historical time period (one not used in model fitting), and quantify accuracy by using the coefficient of determination (R^2) resulting from a linear regression of the estimated on the actual time series. In each repetition, we used a different combination of training data and testing data, and average all the scores across repetitions (denoted as \hat{R}^2). That is, we chose the set of providers that achieved the highest average out-of-sample performance (see, e.g., online Technical Appendix Figure 1).

We compared our results to 3 systems in which providers were selected without historical disease data. Specifically, we selected providers on the basis of the population within 20 miles of a provider (proposed by Polgreen et al. [7]), the total number of patients seen (proposed by Mandl et al. [8]), and the diversity of the municipality of residence for patients, which does not require that each provider see an even distribution of patients; rather, providers are incorporated sequentially to achieve geographic complementarity.

We constructed surveillance systems ranging from 1 through 75 providers by using the selection algorithm for 4 objectives: island-wide cases (*Island*), island-wide cases for each of the 4 dengue virus serotypes (*Serotype*), health region-specific cases for all 8 health service regions (*Regional*), and all objectives combined (*Multi-objective*). We assessed 3 alternative systems: population coverage (*Population*), patient volume (*Volume*), and patient geographic diversity (*Diversity*). The Multi-objective system reached 99% of maximum accuracy with just 22 providers (online

Technical Appendix Figure 2) and performed almost as well as the systems designed specifically to achieve each objective individually (Figure 1). The Diversity system achieved 99%, 92%, and 90% of the performance of the systems specifically engineered for estimating island-wide, serotype, and regional cases, respectively, and showed similar geographic patterns to the Multi-objective system (online Technical Appendix Figure 3). For individual serotypes and regions, performance was best for objectives with less sparse data (online Technical Appendix Figure 4).

Finally, we assessed the robustness of the Multi-objective system, which offered the strongest combination of efficiency

and performance. We tested it against 7 additional years' worth of data that were withheld from the analysis. The system performed well for each of the objectives (Figure 2), achieving average values of 0.86 and 0.78 for surveillance of individual serotypes and regions, respectively, and 0.97 for surveillance of island-wide cases. Among individual serotypes and regions, all had values greater than 0.75, except for the Fajardo region, where cases were particularly sparse.

Conclusions

Surveillance systems are widely used to support public health efforts, but they are rarely designed systematically to achieve clear, quantifiable objectives or surveillance goals, and to do so efficiently. Articulating such public health objectives is a critical first step toward evaluating, improving, and streamlining surveillance. Here, we applied a rigorous, quantitative approach to design a dengue surveillance system that efficiently achieves several distinct public health objectives. The method flexibly and robustly maximizes information collected while minimizing the effort required. In this application, we built a multi-objective system that efficiently tracks the spatiotemporal patterns of dengue in Puerto Rico. This system is almost as informative as the systems we optimized to achieve individual objectives, and it maintained its expected performance on recent data that were withheld during the design stage.

Although surveillance goals and resources may be highly specific to the disease threat and region of concern, the proposed optimization method can be applied broadly to enhance the detection of infectious disease threats, as we have shown now for both dengue and influenza (5). We hypothesize that the systems we designed for dengue in Puerto Rico may also serve well for other arboviruses transmitted by *Aedes* spp. mosquitoes, given their similar transmission mechanisms and the strong out-of-sample performance of the system. In some cases, additional data (e.g., mosquito or nonhuman host surveillance) and public health goals (e.g., vector density) could be integrated into the systems. Such data were not available for this study. For newly emerging arboviruses, when historical data are not available, systems optimized for similar pathogens may provide reasonable coverage. Nonetheless, emergence dynamics may have more sporadic and explosive characteristics that may not be captured by a system designed to track spatiotemporal patterns of an endemic disease.

Public health authorities seek situational awareness at multiple geopolitical scales as well as early warning of anomalous events across a wide spectrum of biologic threats beyond arboviruses. The method we present can also be used to redesign existing surveillance systems by manually including or excluding providers during optimization. Additionally, the method is well suited to integrating diverse data streams, such as climatic, mosquito vector, pharmacy, or digital data (9).

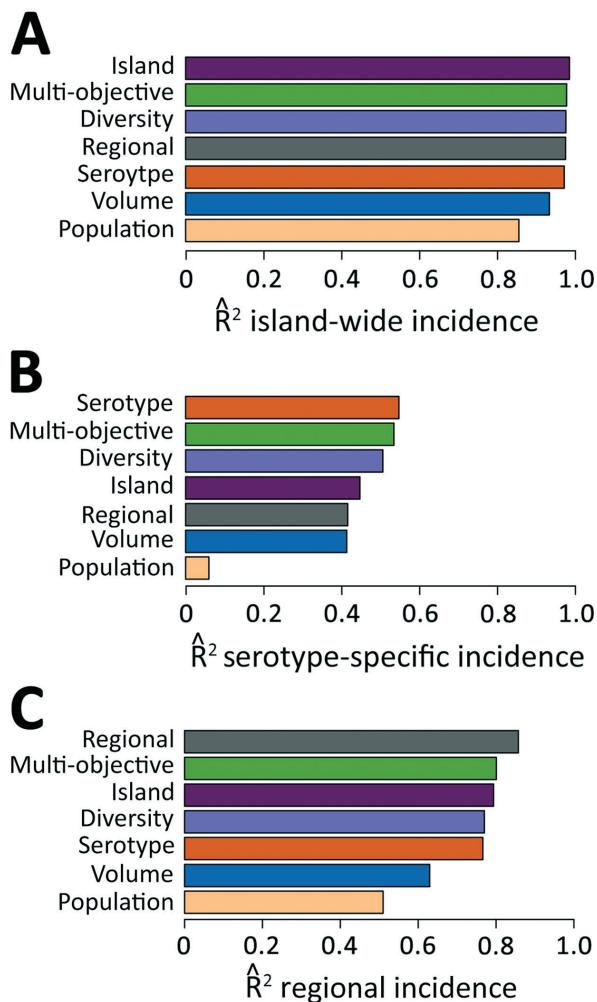


Figure 1. Relative surveillance system performance. The performance of the 4 optimized surveillance systems (Island, Regional, Serotype, and Multi-objective) compared with 3 alternative designs (Population, Volume, and Diversity), with respect to estimating A) island-wide cases, B) serotype-specific cases, and C) regional cases. Each system contains 22 providers. Systems are ordered from highest to lowest performance in each graph. Performance is measured by average out-of-sample across 100 different 3-year periods, resulting from linear regression of target time series (e.g., island-wide cases) on time series of cases occurring within the specified surveillance system.

In an era of “right-sizing,” quantitative development and evaluation are critical to the design, redesign, justification, and benchmarking of surveillance efforts. Given

limited public health budgets on all scales, methods such as the one we present are critical to the future reliability and sustainability of infectious disease surveillance.

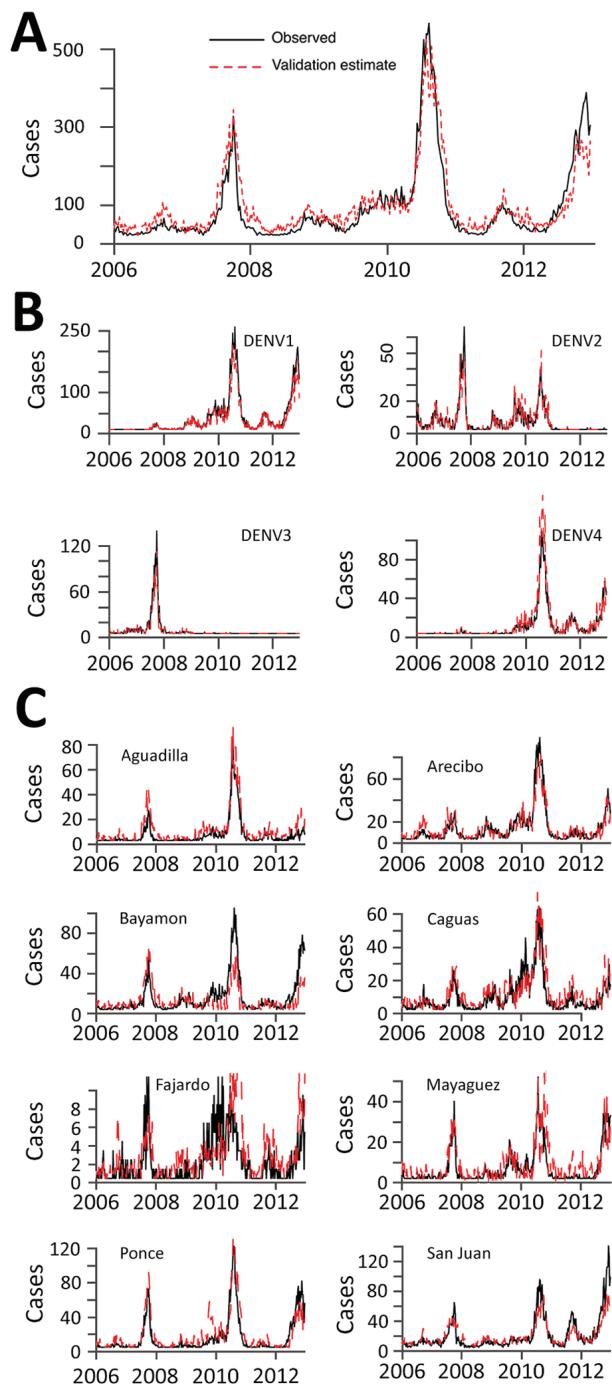


Figure 2. Independent evaluation of performance. The 22-provider Multi-objective surveillance system was designed using data before 2006 and then evaluated on data for 2006–2012 with respect to surveillance of A) island-wide, B) serotype-specific, and C) regional cases. Surveillance estimates from the 22-provider system (red) are compared with raw data from the complete passive surveillance system of 105 providers (black).

Acknowledgment

We thank Ned Dimitrov and Ben Althouse for productive discussions on surveillance and dengue.

S.V.S. acknowledges funding from the Omidyar Group and the Santa Fe Institute. M.A.J. acknowledges partial support from the Models of Infectious Disease Agent Study program (Cooperative Agreement no. 1U54GM088558).

Dr. Scarpino is an assistant professor in the department of mathematics and statistics and is a core faculty member in the Complex Systems Center at the University of Vermont. He investigates questions at the intersection of biology, behavior, and disease and works collaboratively with clinical and public health decision makers to improve disease surveillance.

References

1. World Health Organization and the Special Programme for Research and Training in Tropical Diseases. *Dengue: guidelines for diagnosis, treatment, prevention and control*. Geneva: The Organization; 2009.
2. LaBeaud AD, Bashir F, King CH. Measuring the burden of arboviral diseases: the spectrum of morbidity and mortality from four prevalent infections. *Popul Health Metr*. 2011;9:1. <http://dx.doi.org/10.1186/1478-7954-9-1>
3. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. *Nature*. 2013;496:504–7. <http://dx.doi.org/10.1038/nature12060>
4. Beatty ME, Stone A, Fitzsimons DW, Hanna JN, Lam SK, Vong S, et al.; Asia-Pacific and Americas Dengue Prevention Boards Surveillance Working Group. Best practices in dengue surveillance: a report from the Asia-Pacific and Americas Dengue Prevention Boards. *PLoS Negl Trop Dis*. 2010;4:e890. <http://dx.doi.org/10.1371/journal.pntd.0000890>
5. Scarpino SV, Dimitrov NB, Meyers LA. Optimizing provider recruitment for influenza surveillance networks. *PLoS Comput Biol*. 2012;8:e1002472. <http://dx.doi.org/10.1371/journal.pcbi.1002472>
6. Waterman SH, Novak RJ, Sather GE, Bailey RE, Rios I, Gubler DJ. Dengue transmission in two Puerto Rican communities in 1982. *Am J Trop Med Hyg*. 1985;34:625–32.
7. Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, Rushton G. Optimizing influenza sentinel surveillance at the state level. *Am J Epidemiol*. 2009;170:1300–6. <http://dx.doi.org/10.1093/aje/kwp270>
8. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. *J Am Med Inform Assoc*. 2003;11:141–50. <http://dx.doi.org/10.1197/jamia.M1356>
9. Althouse BM, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, Baumbach J, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science*. 2015;4:17. <http://dx.doi.org/10.1140/epjds/s13688-015-0054-0>

Address for correspondence: Michael A. Johansson, Centers for Disease Control and Prevention, 1324 Calle Cañada, Mailstop P01, San Juan, PR 00920, USA; email: eyq9@cdc.gov

Design Strategies for Efficient Arbovirus Surveillance

Methods

Simulating Provider Data, 1991–1998

The identities of the submitting providers were not included in dengue reports before 1999. However, these identities are critical to our optimization methods. Thus, we used a simulation method to assign each pre-1999 case to a specific provider, based on post-1999 data linking providers to patient municipalities. Each pre-1999 report was either assigned to a known provider or designated as unknown, as follows:

1) Estimating the fraction of cases with known and unknown providers for each municipality. During 1999–2005, a weekly average of 10% of the reported suspected cases did not identify a provider (range 0%–37%), with the proportion varying by municipality and increasing at times of high case volumes. For each municipality, we stratified the 1999–2005 reports into 5 equal-width bins based on the observed island-wide cases when each case was reported. Then, for each municipality (m)-cases bin (b) combination, we calculated the proportion of cases with a known provider ($k_{m,b}$).

2) Estimating distribution of cases across providers, for each municipality. Using 1999–2005 case reports with known providers, we calculated the fractions of cases in each of the 78 municipalities (m) that were reported by each of the 105 providers (p) ($a_{m,p}$).

3) Assigning pre-1999 cases to providers. For the 1991–1998 data, which contain the patient's municipality of residence but not the reporting provider, we simulated provider identities by multiplying the quantities described in the first two steps. That is, the number of cases from a given municipality (m) assigned to a given provider (p) for a given time period depended on the island-wide cases at the time (b), and was set equal to the product of the total number of dengue cases reported in m during that period, the estimated fraction of cases from that municipality with a known provider ($k_{m,b}$), and the estimated fraction of dengue cases from m seeking care from p ($a_{m,p}$). Last, fractional cases were rounded to integers.

Surveillance Objectives

For each of the surveillance objectives, we formulated specific quantities to be estimated from the surveillance data: for island-wide cases, the total number of laboratory-positive cases by week; for serotype cases, the total number of cases of each of the 4 serotypes reported by week; and for regional cases, the number of confirmed cases in each of the 8 health service regions by week.

In designing a multipurpose dengue surveillance system, we sought to identify a relatively small subset of providers that could provide accurate real-time estimates of these quantities. However, it is computationally unfeasible to evaluate all possible combinations of providers. For example, an exhaustive analysis of all subsets of 75 providers from the full set of 105 providers would require 1.6×10^{26} evaluations. Rather than performing an exhaustive search, we used a more efficient procedure for identifying providers for inclusion in the surveillance system, as described in the following sections.

Surveillance System Optimization

We designed surveillance systems using a greedy algorithm that sequentially adds providers that most improve the performance of the system. Starting with the set of all possible providers P (in this case, all clinics in Puerto Rico historically reporting dengue cases), we selected a set of providers, S , which initially has no members. At each step, we added the provider that is expected to yield the highest value of our objective function, f .

The Objective Function

To evaluate the performance of a given system (set of providers S) with respect to the surveillance objectives listed previously, we repeatedly performed the following three-step procedure: 1) fit multiple linear models relating historical data from the surveillance system in question to actual dengue cases, 2) use the fitted models to estimate dengue cases in another historical time period (that was not included in the model fitting procedure), and 3) quantify the accuracy of those estimates. In each repetition, we used a different combination of training data and testing data, and ultimately combined all the accuracy estimates (across all objectives and repetitions) into a single objective function.

Our overarching surveillance objective was a set, G , of up to 13 different subobjectives, g (estimating dengue cases across the whole island, in each of the 8 geographic regions, and for each of the 4 different serotypes). When evaluating a subset of providers, S , we fit multiple linear models (one per subobjective) given by

$$S\theta(g, S): Y_{g,t} = a_g + \sum b_{s,g} X_{s,g,t} + \varepsilon_{t,s}$$

to the testing data, where $Y_{g,t}$ are the actual cases with respect to objective g at time t (for example, island-wide dengue virus serotype 1 (DENV-1) cases in a particular week), $X_{s,g,t}$ are the cases with respect to objective g reported by provider s at time t , a_g and $b_{s,g}$ are the model coefficients, and ε_t is a zero-mean normally distributed error term.

After estimating the coefficients of each subobjective model by using the training data, we used the models to predict the case quantities during the testing period and quantify the accuracy of the predictions by calculating R^2 values. Ultimately, we aggregated the accuracy measurements into the single objective function given by $f(S, G, D)$.

$$\sum R^2(\theta(g, S), d) w_g w_d, g \in G, d \in D$$

where $R^2(\theta(g, S), d)$ represents the out-of-sample performance of system S on objective g with testing-training data combination d ; D represents the set of testing-training data combinations used in the evaluation; and w_g and w_d are weights indicating the contribution of each objective and dataset to the objective function, where each sum to 1, $\sum_{g \in G} w_g = 1$ and $\sum_{d \in D} w_d = 1$. For simplicity we refer to $R(\theta(g, S), d)$ as R' .

We built surveillance networks for 4 different objective functions (each consisting of a distinct combination of subobjectives): 1) overall island-wide cases (*Island*), 2) island-wide cases for each of the 4 DENV serotypes, with each serotype given a 1/4 weight (*Serotype*), 3) regional dengue cases for each of the 8 health service regions, with each region given a 1/8 weight (*Regional*), and 4) the 3 prior objectives weighted equally (*Multi-objective*), resulting in weights of 1/12 for each serotype case, 1/32 for each regional case, and 1/3 for each island-wide case.

Provider Selection Algorithm

At each step in the optimization, we considered all providers that had not yet been included in the system, and selected the one that produced the maximum value of f . Let S_n denote the surveillance system at step n in the optimization. The iterative selection proceeds as follows:

1) For each provider $x \in P/x \notin S_n$, create a candidate system $S_{n,x} = \{S_n, x\}$ and calculate $f(S_{n,x}, G, D)$.

2) Identify the provider x that maximizes the expected improvement in performance $f(S_{n,x}, G, D) - f(S_n, G, D)$.

3) Add x to S .

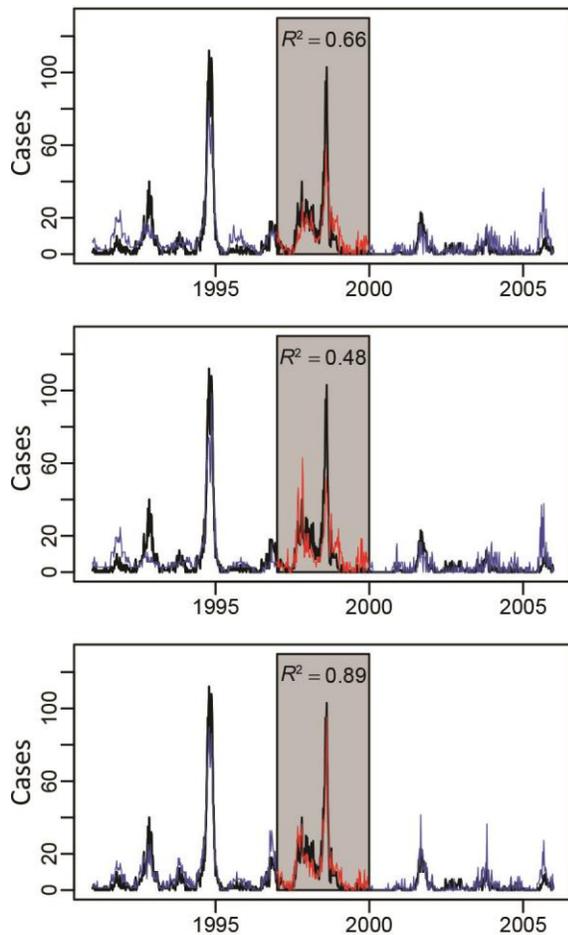
4) Repeat.

Volume-based design: We selected providers sequentially based on the total number of patients seen during 1990–2005.

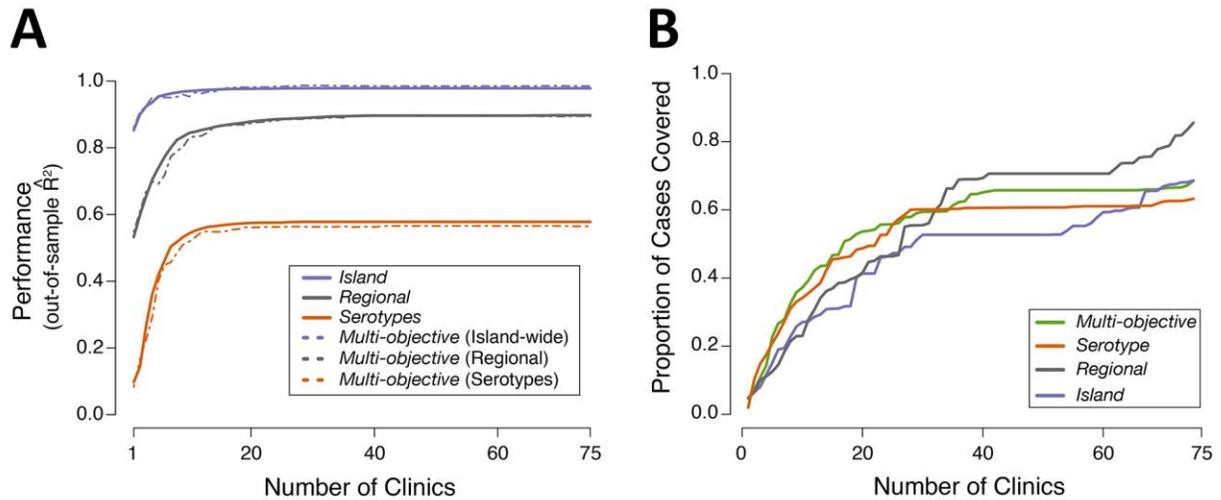
Diversity-based design: We used a greedy algorithm to maximize the Shannon diversity index, H , of the municipality of residence for patients captured by the surveillance system S . If there are M municipalities and the proportion of patients in S from municipality i is p_i , then the Shannon diversity for S is:

$$H_S = -\sum p_i \ln p_i$$

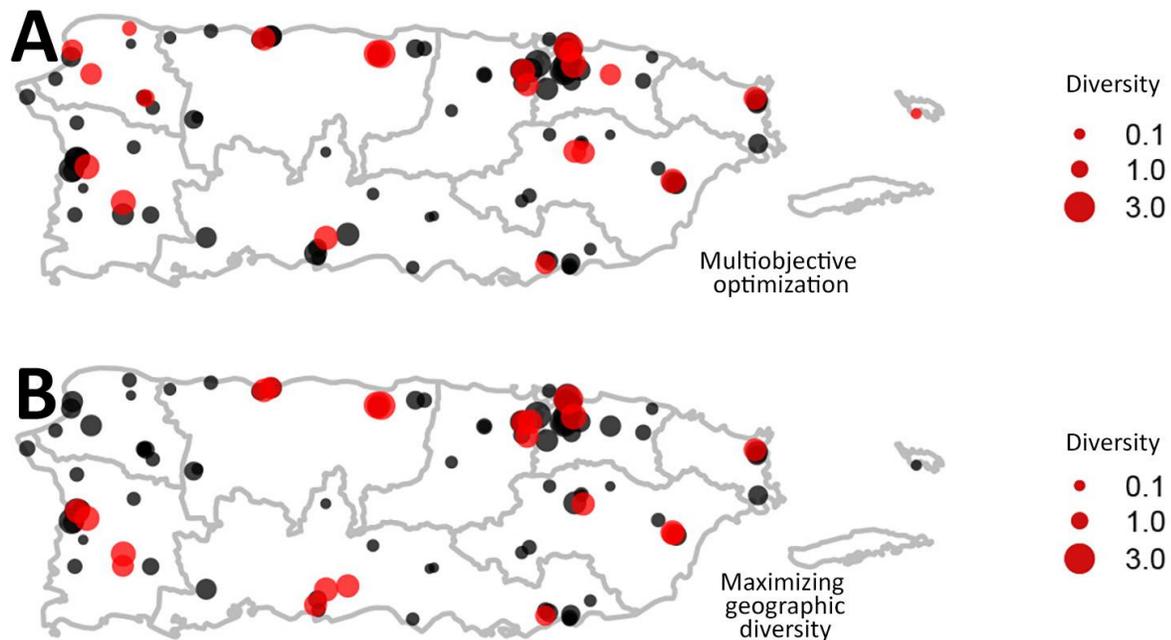
This quantity is maximized when $p_i = 1/M$, that is, when the municipalities contribute equal numbers of patients to S . This does not require that each provider see a uniform distribution of patients, and providers are incorporated sequentially to achieve geographic complementarity. This procedure resembles the *Population*-based model, but maximizes the diversity of patient residences rather than the number of patients.



Technical Appendix Figure 1. Provider selection for dengue surveillance in the Mayaguez health service region. To design a system for regional-level dengue surveillance, we evaluated the performance of different combinations of providers across each of the 8 health service regions. This figure illustrates a single step in the selection process for one of the health service regions after 1 provider has already been included. Dengue incidence in the Mayaguez region is shown in black for the period of 1991–2005, and serves as the response variable in the regression-based provider selection method. The panels show 3 candidate providers under evaluation for subsequent inclusion as the second provider selected for the system. We combined data from each candidate with data from the first provider already incorporated in the system. We then performed linear regression of total Mayaguez incidence on the combined data during the 1991–1996 and 2000–2005 time periods (blue), and made out-of-sample predictions (red). Performance is quantified by the out-of-sample R^2 . This process is repeated 100 times with random 3-year intervals withheld for out-of-sample evaluation. The candidate provider delivering the highest average R^2 across the 100 trials is selected for inclusion. In this example, the provider associated with the bottom panel is more informative than the alternatives.

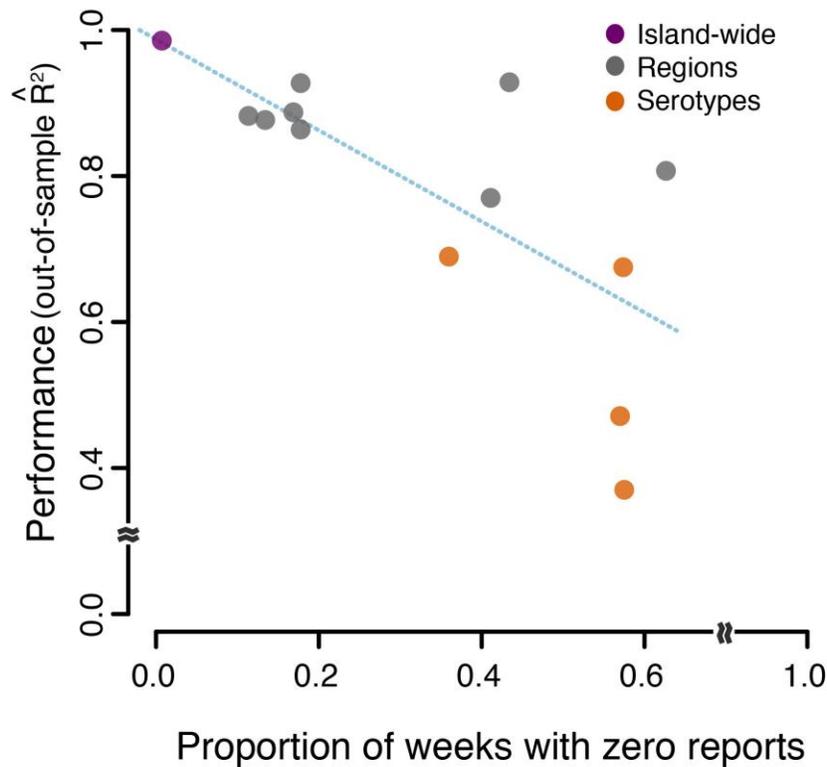


Technical Appendix Figure 2. Performance curves for optimized dengue surveillance systems. As providers are added to the systems, A) estimation of island-wide, regional, and serotype cases improves and then levels, as quantified by average out-of-sample R^2 , whereas B) the proportion of DENV cases occurring at providers within the system increases more gradually. Some providers were estimated to have reported either zero or very nearly zero cases during the study period, which is why the proportion of cases covered does not increase with each additional provider. The maximum performance remains less than 1.



Technical Appendix Figure 3. Location and patient geographic diversity of selected providers. The 22 providers selected A) under multi-objective optimization (*Multi-Objective*) and B) when maximizing the geographic diversity of patients (*Diversity*) are indicated in red and the remaining 92 providers in black. Circle size reflects the Shannon diversity of patient municipalities of a given provider. The lines indicate

the boundaries of the Puerto Rico health regions. The islands of Culebra and Vieques (right) are part of the Fajardo health region (northeastern corner).



Technical Appendix Figure 4. Performance decreases as sparsity of training data increases. For each of the 13 different surveillance subobjectives (island-wide incidence, incidence in each of the 8 health service regions, and incidence for each of the 4 serotypes), we plot the average out-of-sample R^2 for the best combination of 22 providers against the proportion of weeks with zero reported cases in the training period time series data (1991–2005). For example, for dengue virus serotype 1 (DENV-1), we find the combination of 22 providers that maximizes performance, and plot the resulting performance against the proportion of weeks during 1991–2005 without a reported case of laboratory-confirmed DENV-1. Performance is measured by average out-of-sample R^2 across 100 different 3-year periods, resulting from linear regression of a target average series (e.g., all DENV-1 cases) on the time series of cases occurring within the candidate set of providers. The least-squares regression line relating performance to data quality is plotted (blue dashes).