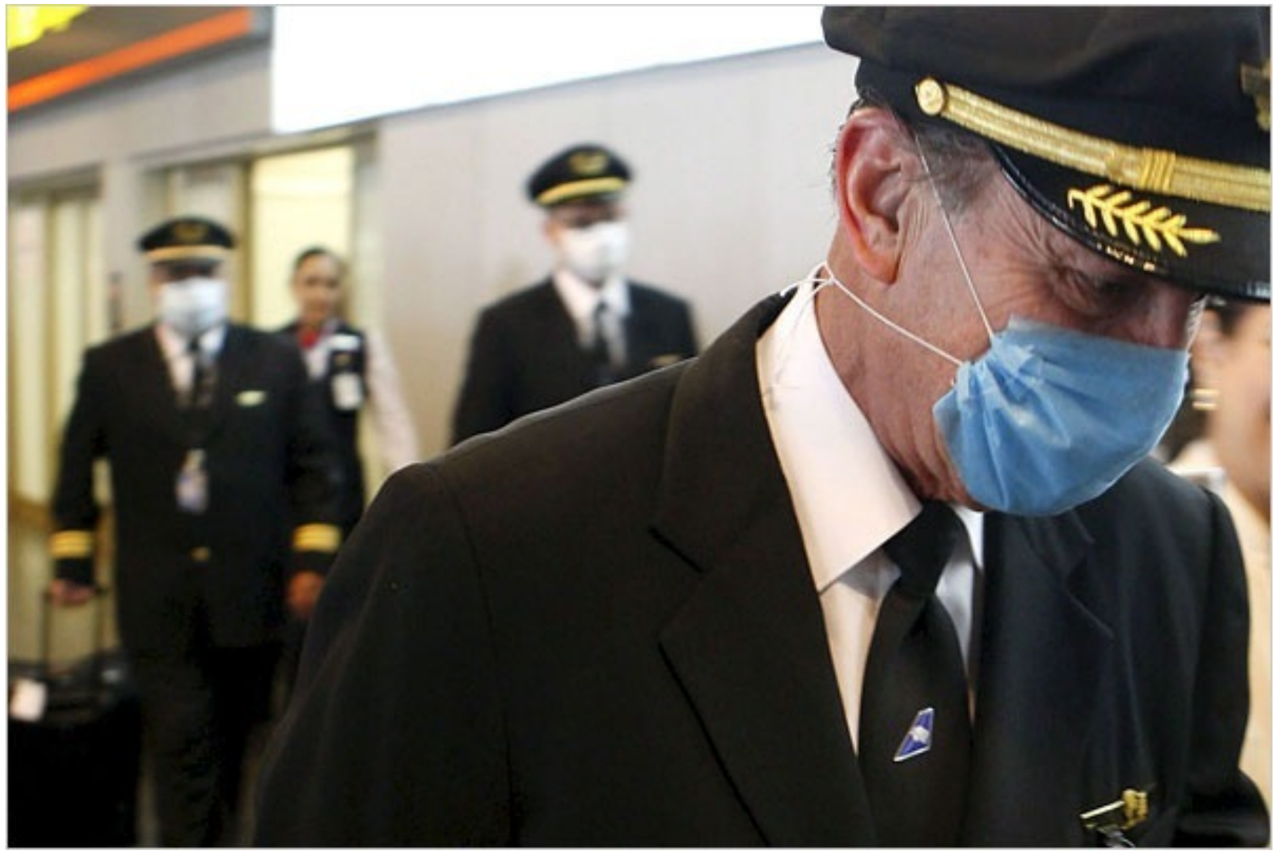
Designing multi-objective surveillance systems

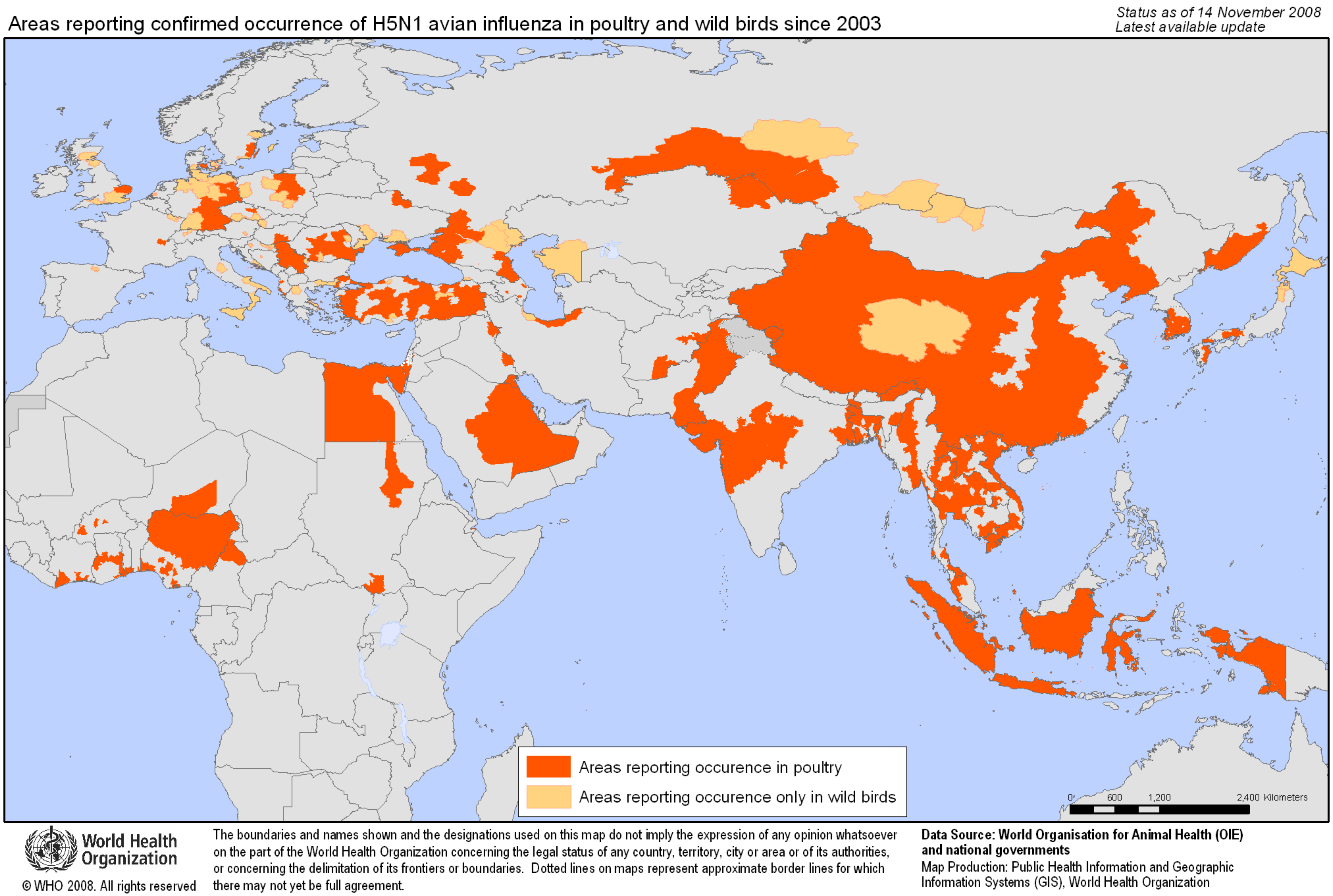
Samuel V. Scarpino
Omidyar Fellow
Santa Fe Institute

scarpino@santafe.edu
@svscarpino
scarpino.github.io

Where will the next pandemic emerge?



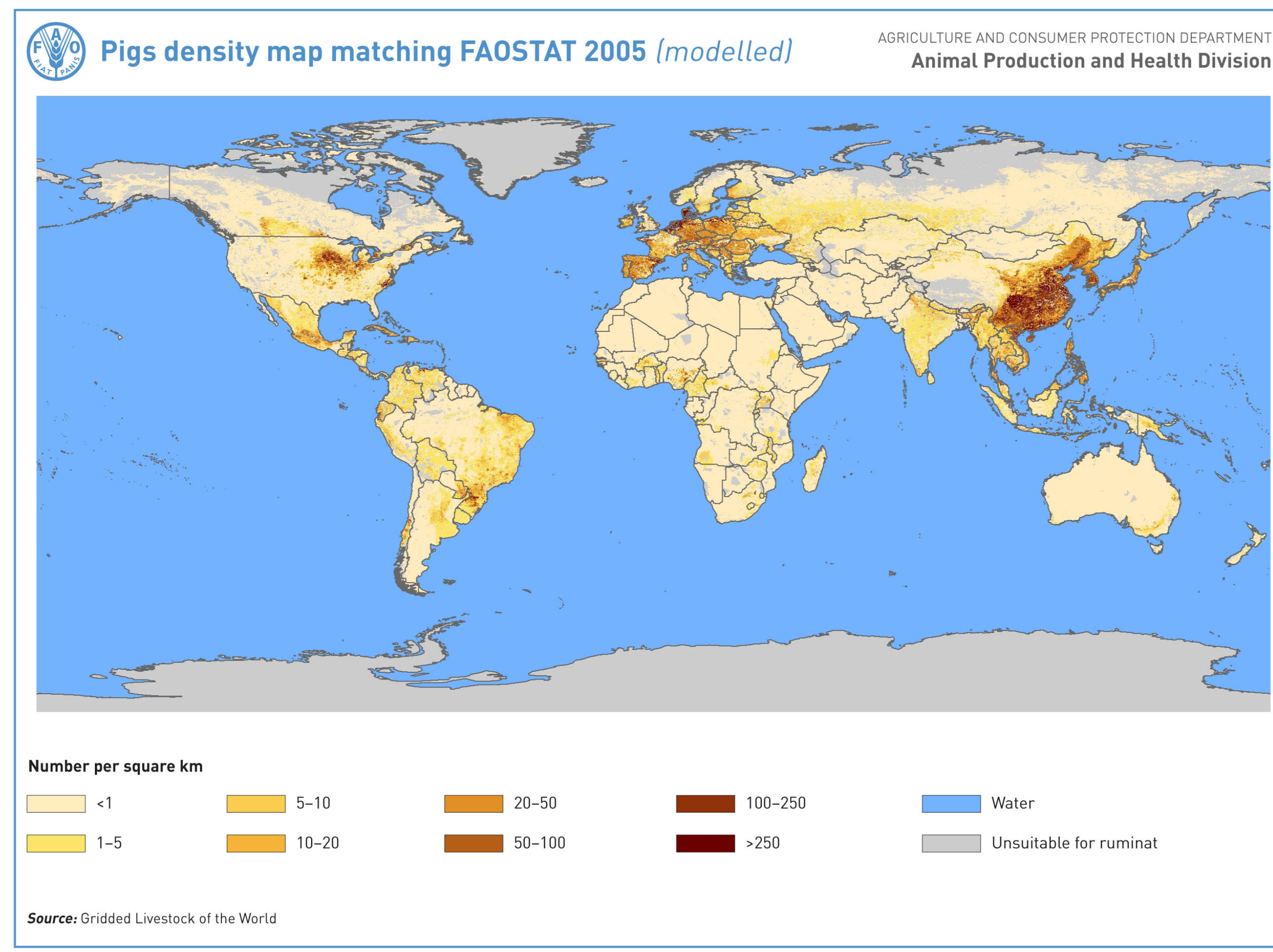
Where will the next **influenza** pandemic emerge?



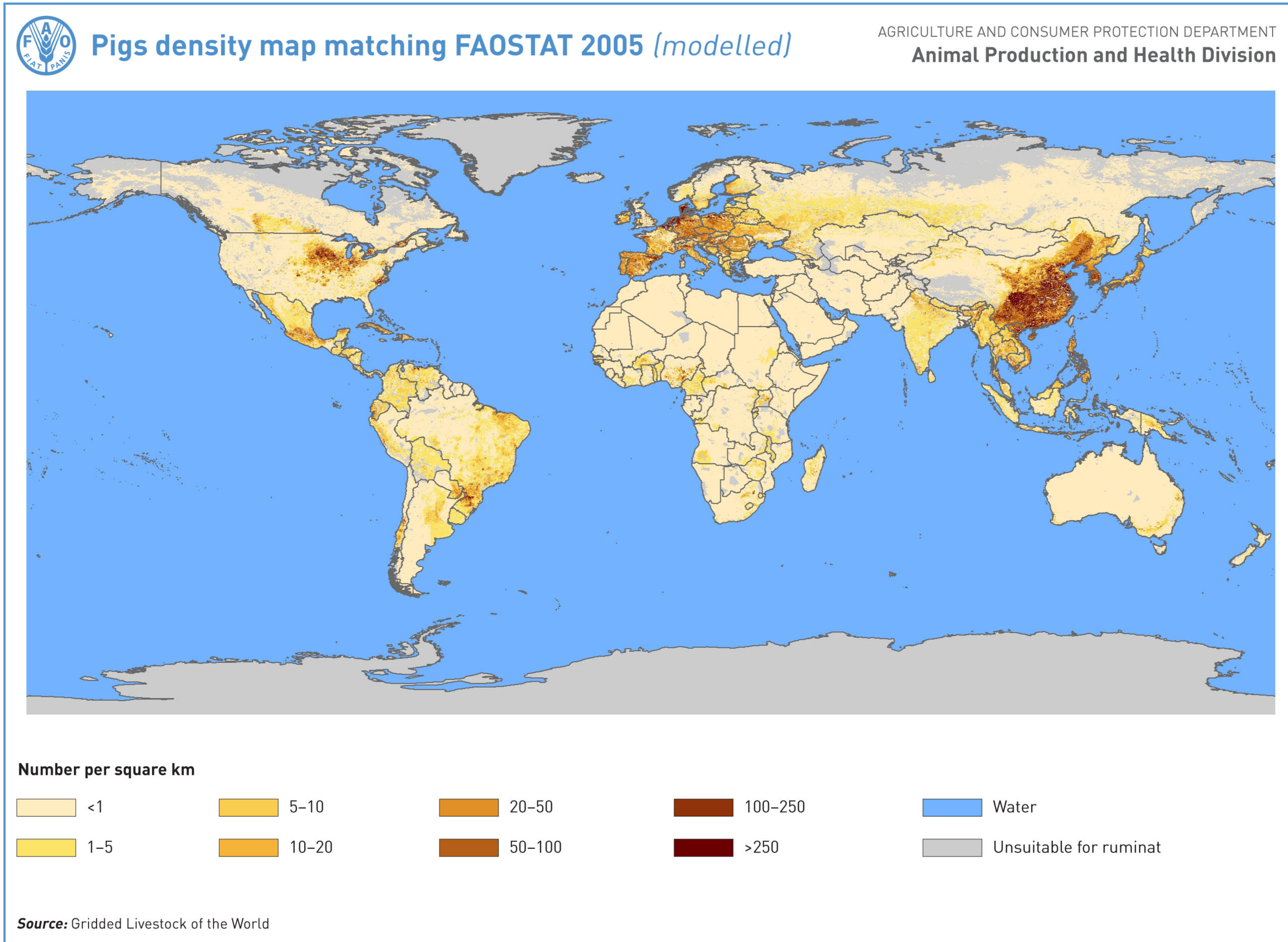
What factors are important for influenza?

Air travel

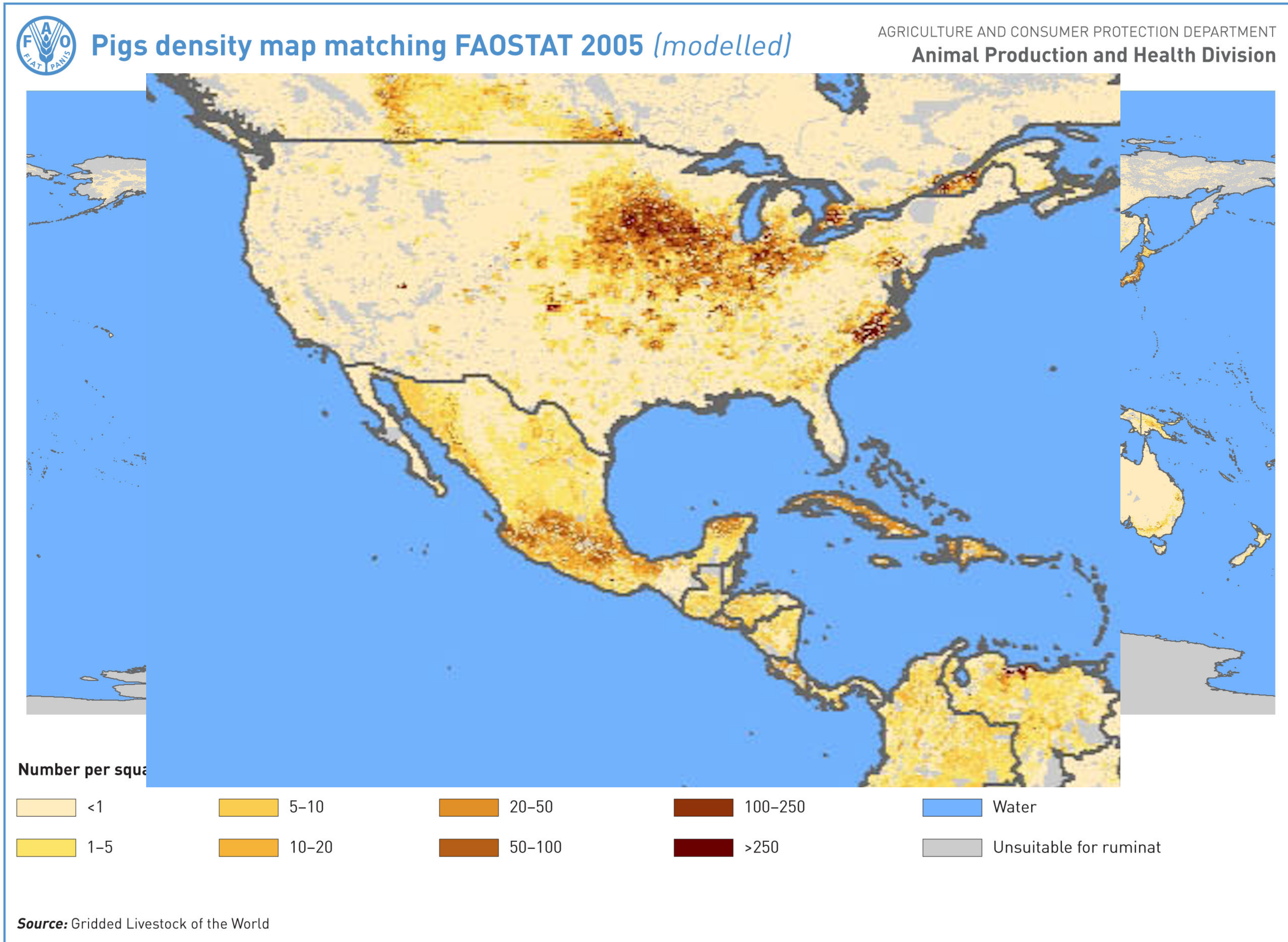
Density of pig farms



Worldwide pig density



Worldwide pig density



and it happened again with Ebola

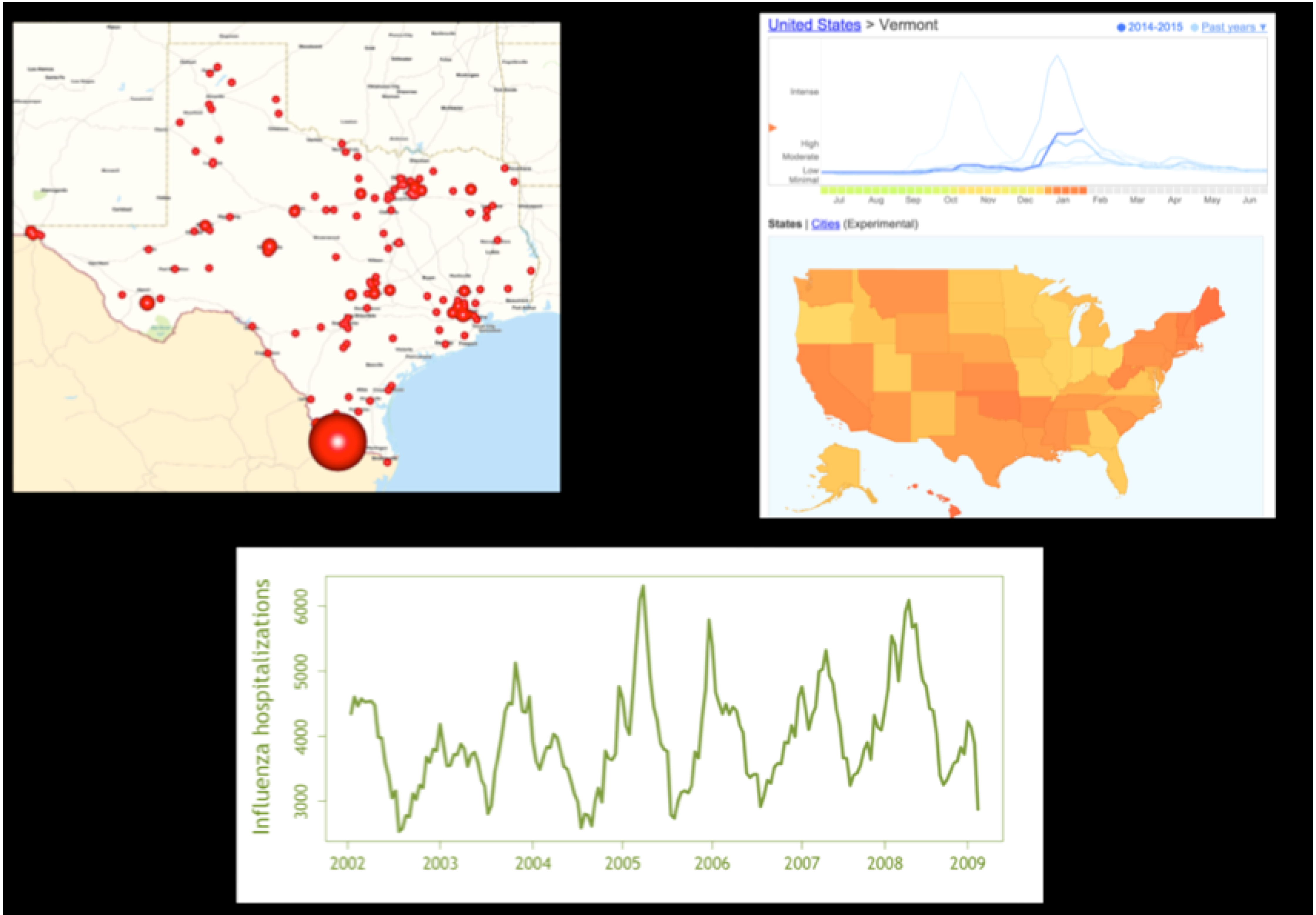
and it happened again with Ebola

TABLE I. — **Age distribution of persons positive for either Lassa (LAS), Ebola (EBO) or Marburg (MAR) virus antibodies.**

Age (years)	Nb tested	LAS-positive (prevalence %)	EBO-positive (prevalence %)	MAR-positive (prevalence %)
0-9	49	5 (10 %)	2 (4 %)	0
10-19	68	11 (16 %)	5 (7 %)	0
20-29	108	21 (19 %)	6 (6 %)	1
30-39	94	16 (17 %)	5 (5 %)	1
40-59	88	9 (10 %)	6 (7 %)	1
60 plus	26	5 (16 %)	2 (8 %)	2
Total	433	67 (16 %)	26 (6 %)	5 (1 %)

Knobloch et al. 1982

Disease surveillance stands at a crossroads



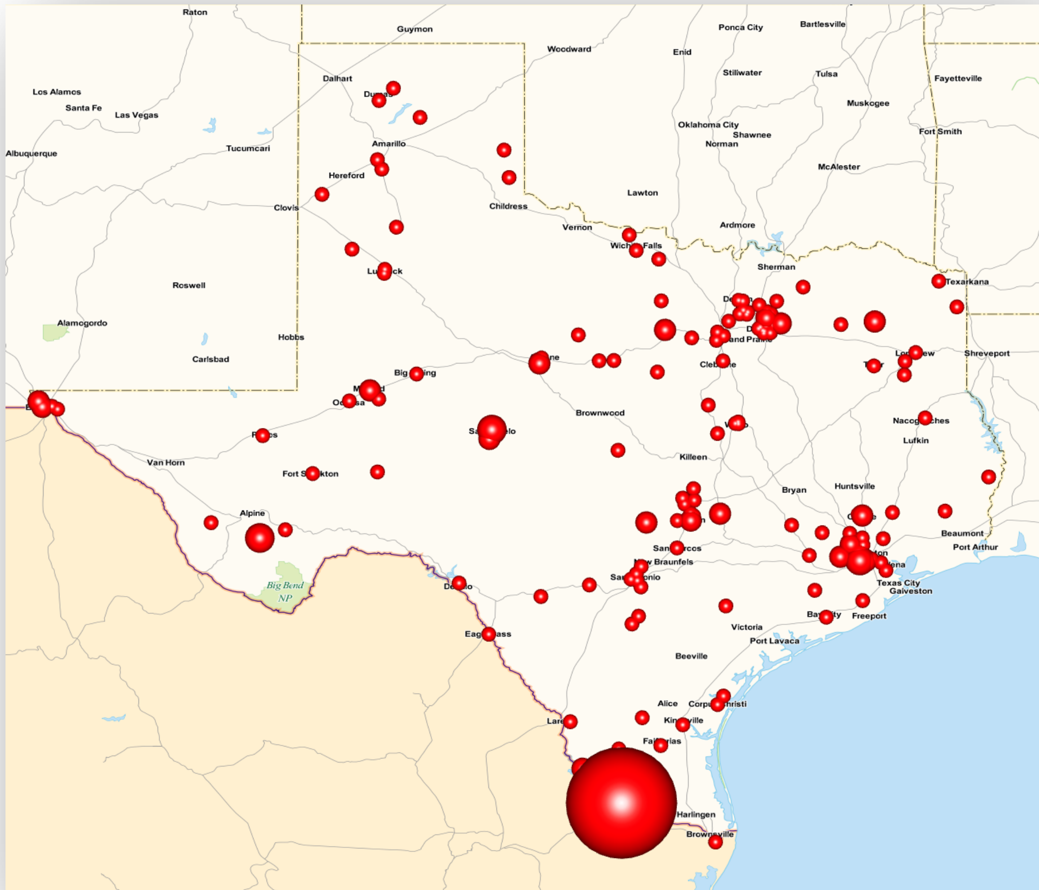
Four Step Program

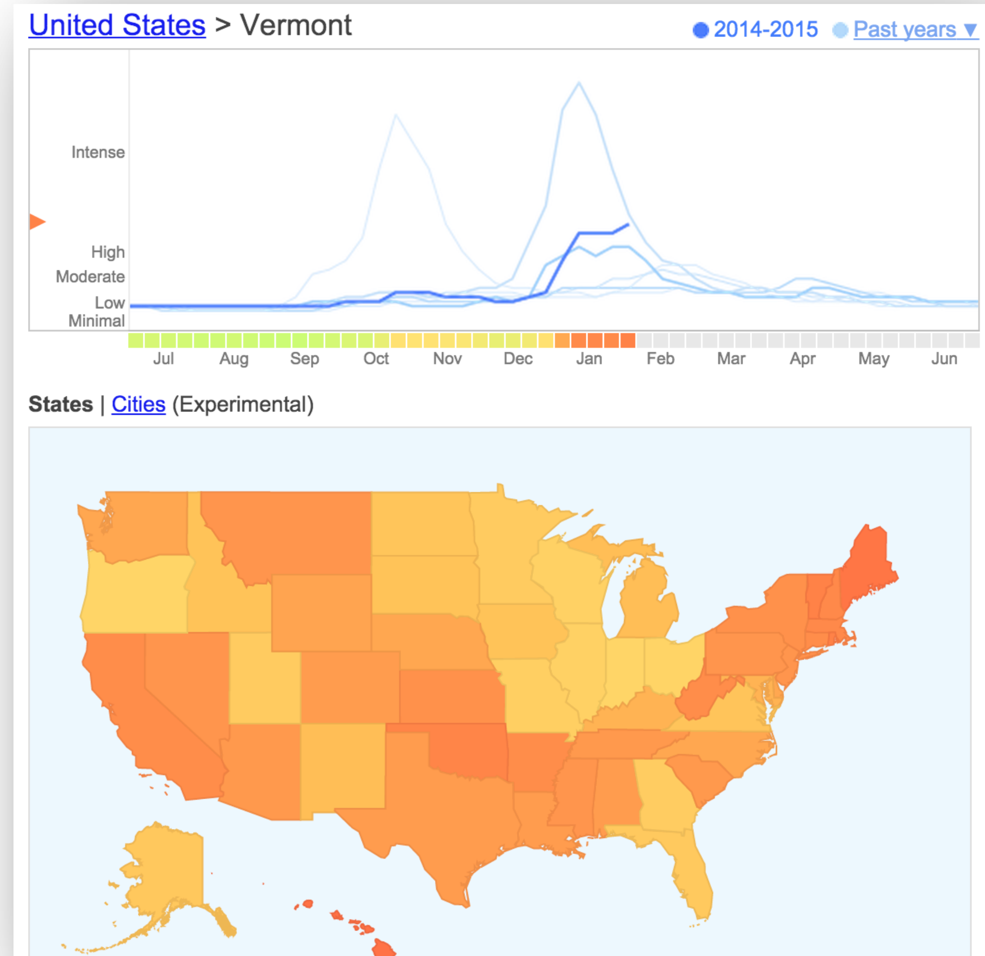
1. Formalize the public health objectives
2. Specify candidate data sources & acquire historical data
3. Simulate data where missing
4. Select the most informative data sources and validate!

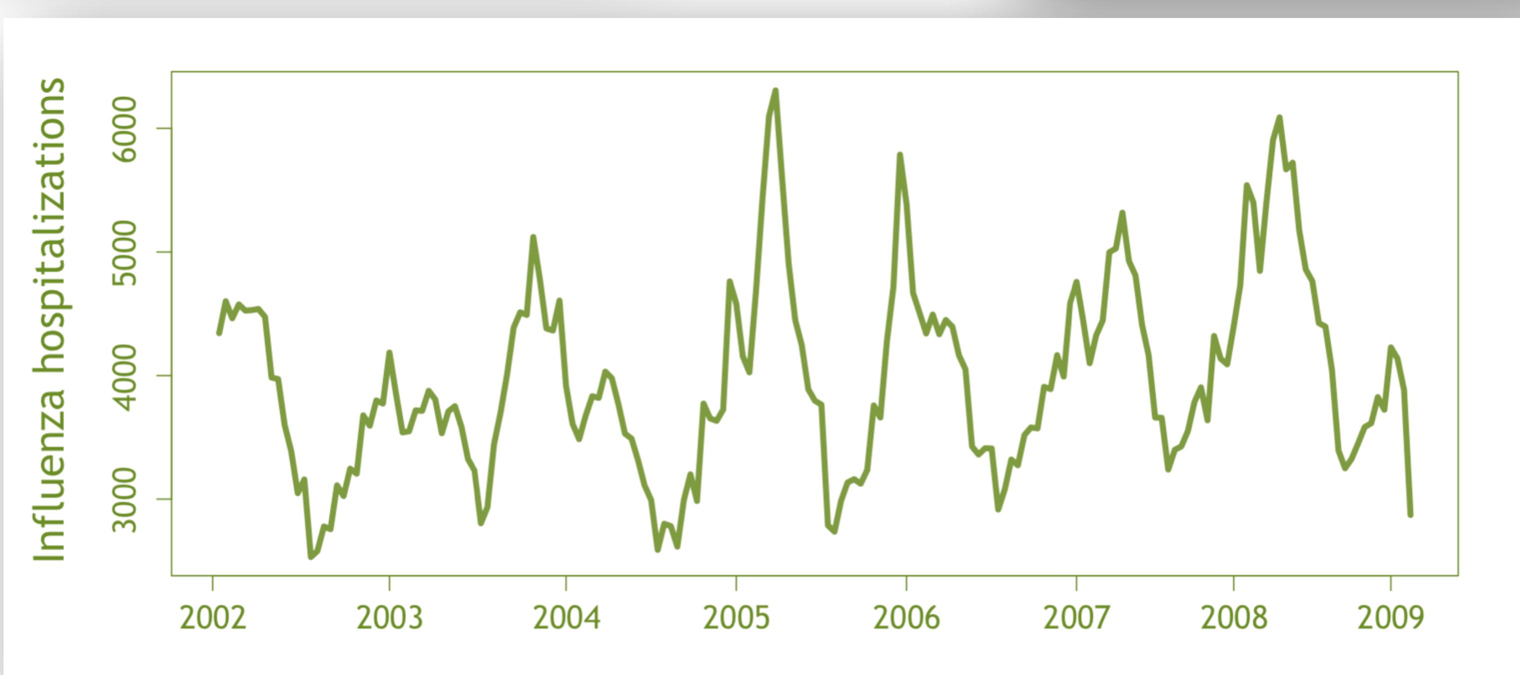
Scarpino, Dimitrov, Meyers 2012

1. Surveillance goal - influenza in Texas

2. Data - ILINet, Hospitalizations, & G.F.T.







3. Simulate data

1. Obtain a hospitalization time series.

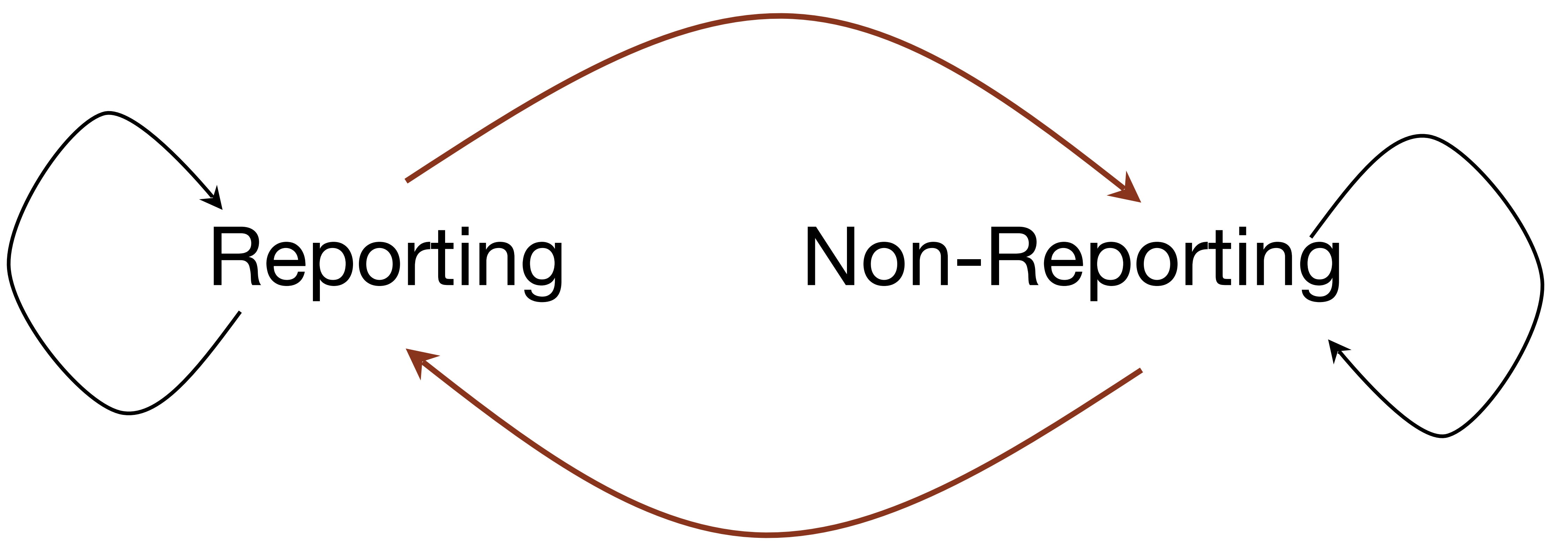
2. Integrate over published estimates for the hospitalization rate of influenza.

3. Account for noise.

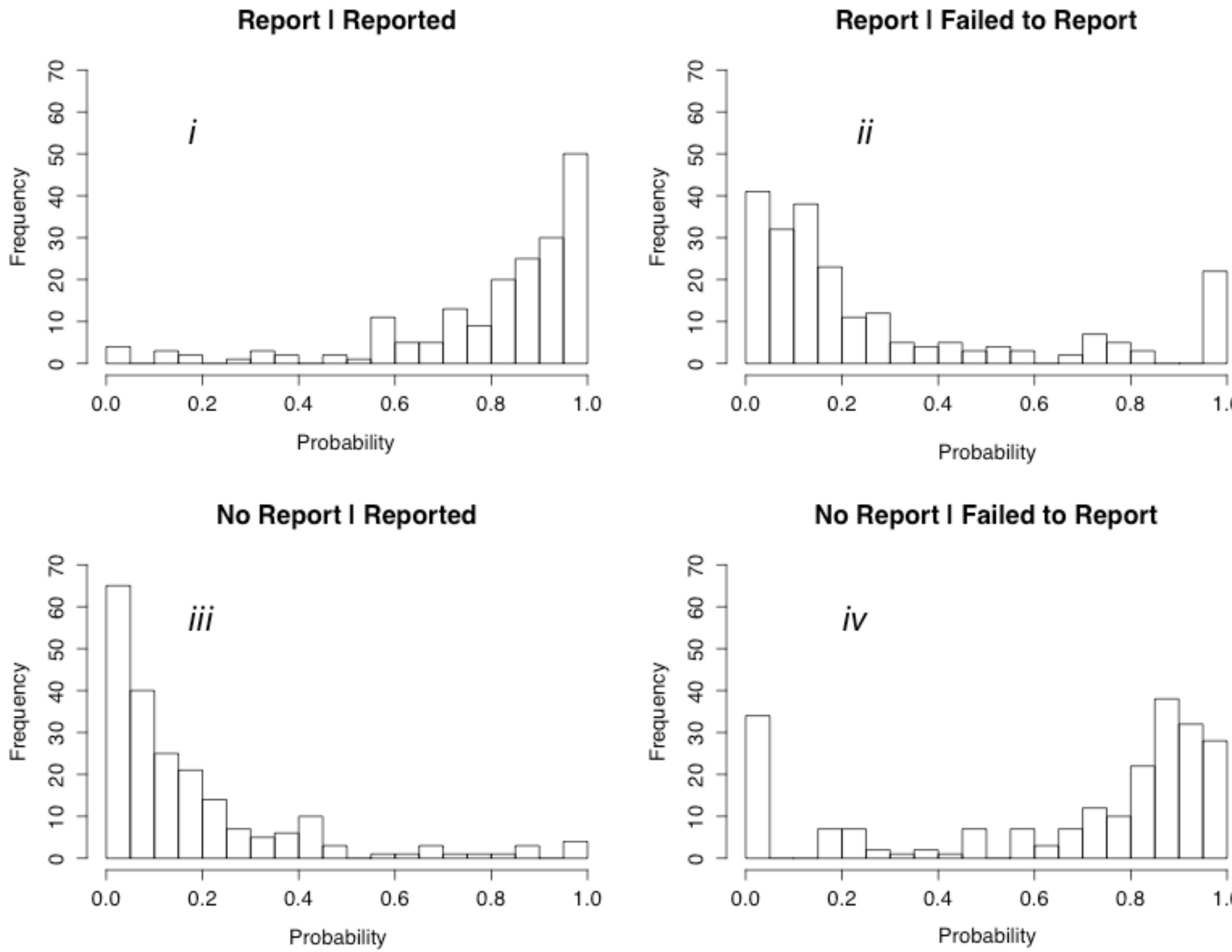
$$ILI = C1 \times PerfectInformation + C0 + N(0, \sigma^2)$$

4. Account for imperfect reporting.

3. Simulate data - provider behavior



3. Simulate data - empirical reporting rates



4. Select providers

$$R^2(Hosp, S, \hat{\xi}) =$$

$Hosp$ = goal time series (state-wide hospitalizations)

S = subset of selected providers

$\hat{\xi}$ = observational noise and reporting

4. Select providers

$$R^2(Hosp, S, \hat{\xi}) = \frac{\mathrm{Var}(Hosp) - \mathrm{Var}\left(Hosp - \sum_{i \in S} \alpha_i P_i(\hat{\xi}) \right)}{\mathrm{Var}(Hosp)}$$

$Hosp$ = goal time series (state-wide hospitalizations)

S = subset of selected providers

$\hat{\xi}$ = observational noise and reporting

α_i = regression coefficients

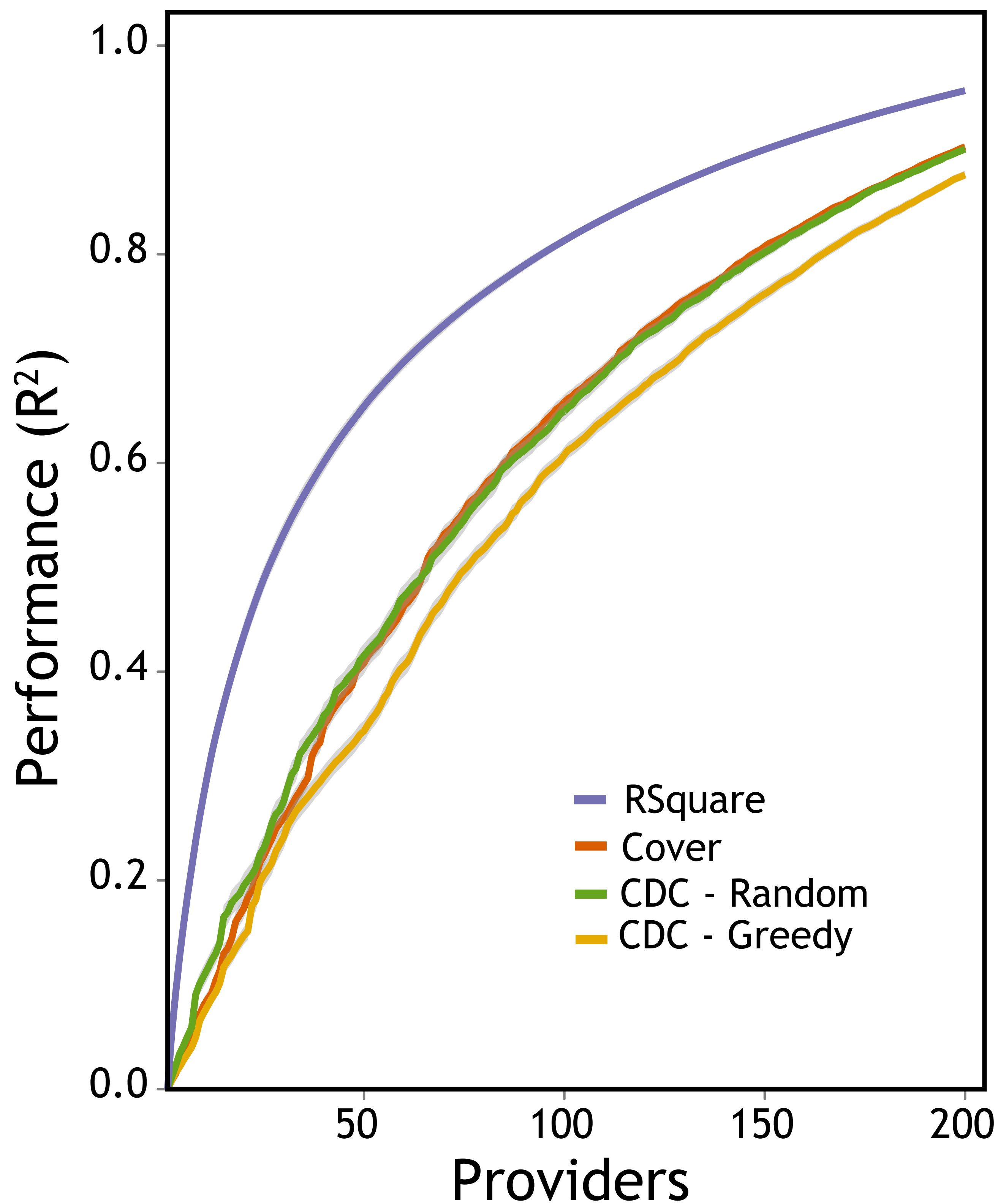
P = mock provider pool

4. Select providers - account for provider behavior

$$R^2(Hosp, S, \hat{\xi}) = \frac{\mathrm{Var}(Hosp) - \mathrm{Var}\left(Hosp - \sum_{i \in S} \alpha_i P_i(\hat{\xi}) \right)}{\mathrm{Var}(Hosp)}$$

$$\max_{S \subseteq P} E_{\bar{\xi}} \left[R^2(Hosp, S, \hat{\xi}) \right]$$

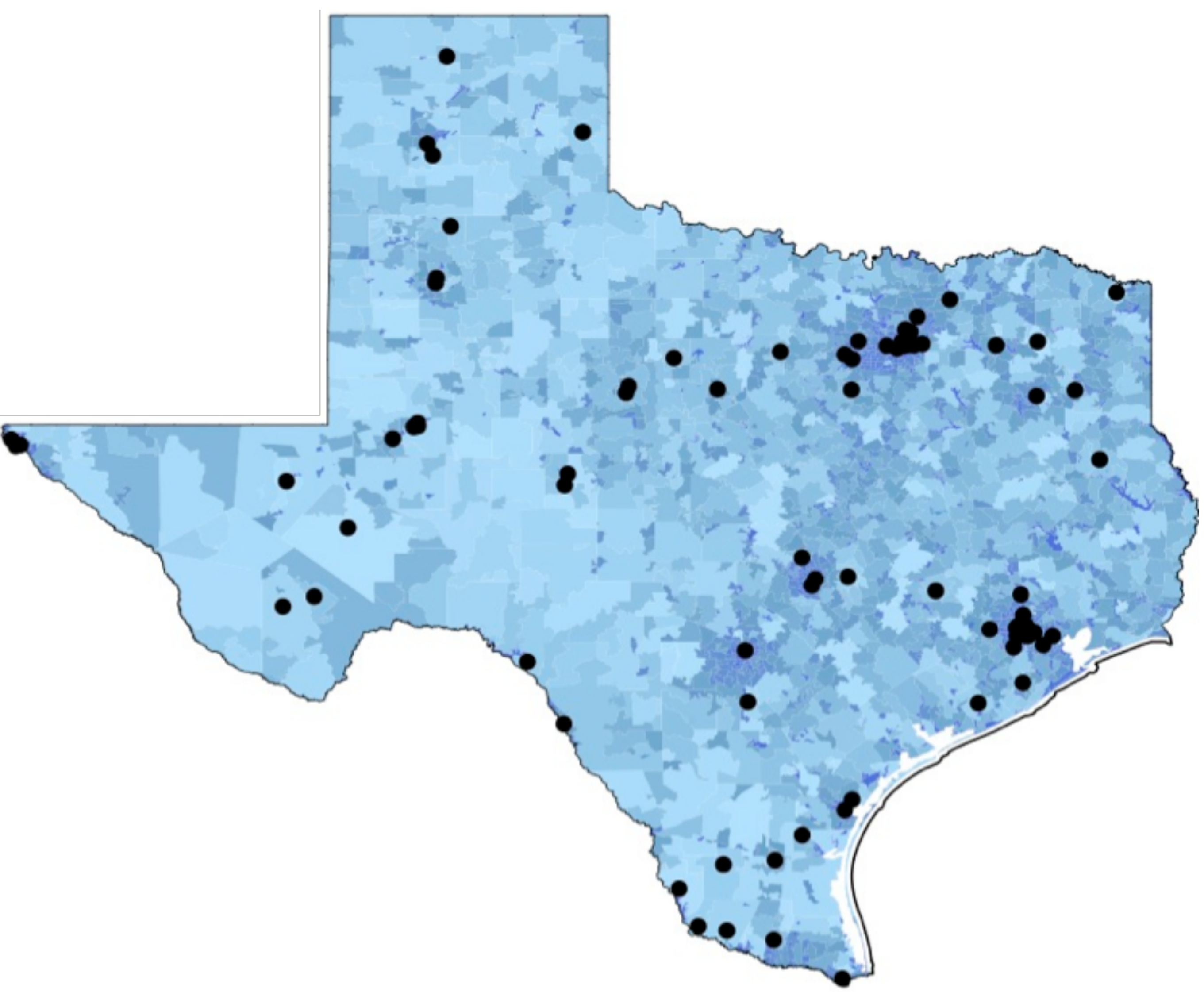
An optimized network is more informative



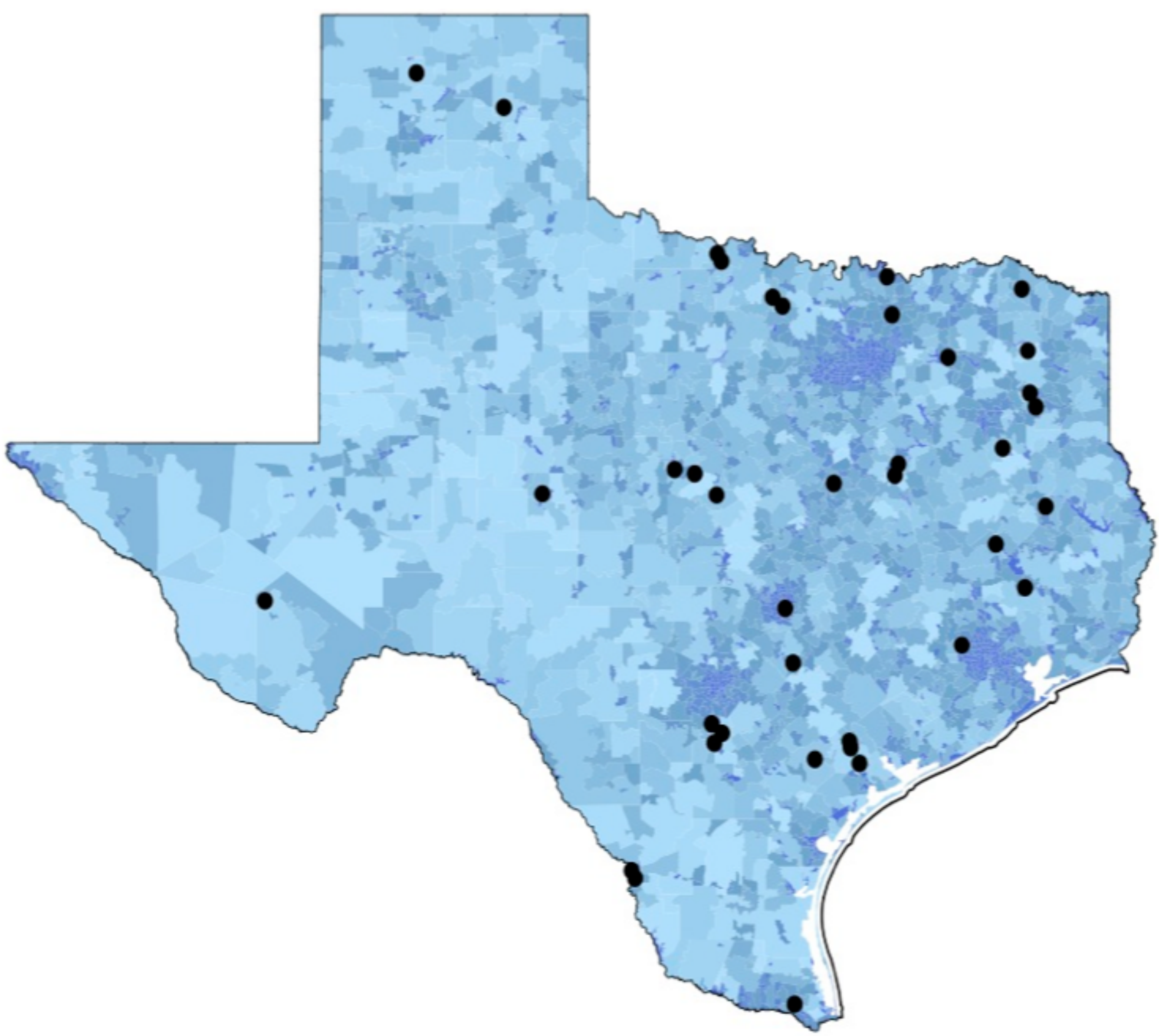
Scarpino et al. 2012

An optimized network is more efficient

Obs. ILINet – 82 Providers – R^2 ~ 0.6



Rsquare – 38 Providers – R^2 ~ 0.6

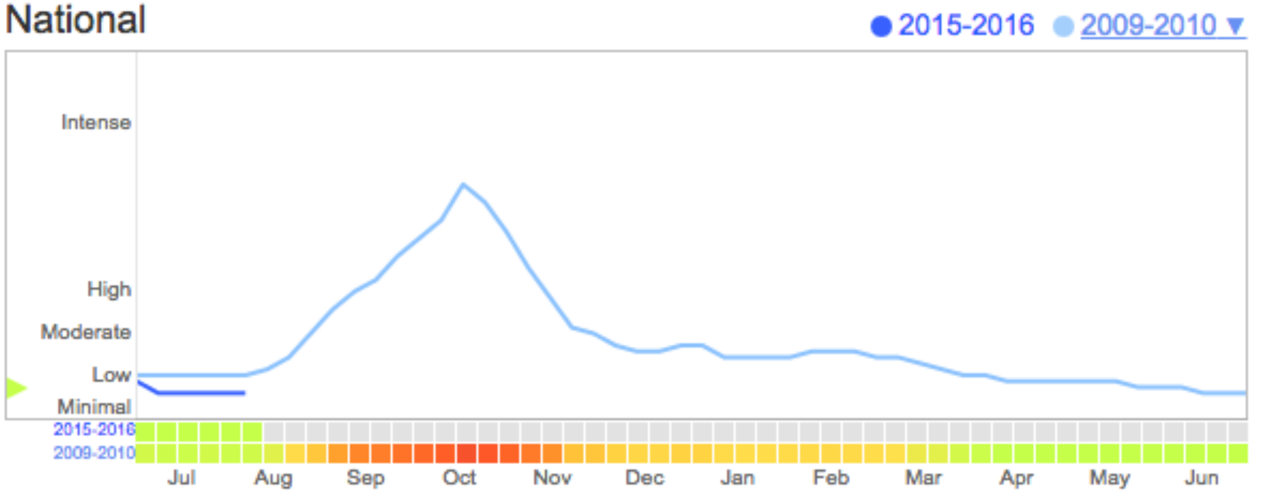


Scarpino et al. 2012

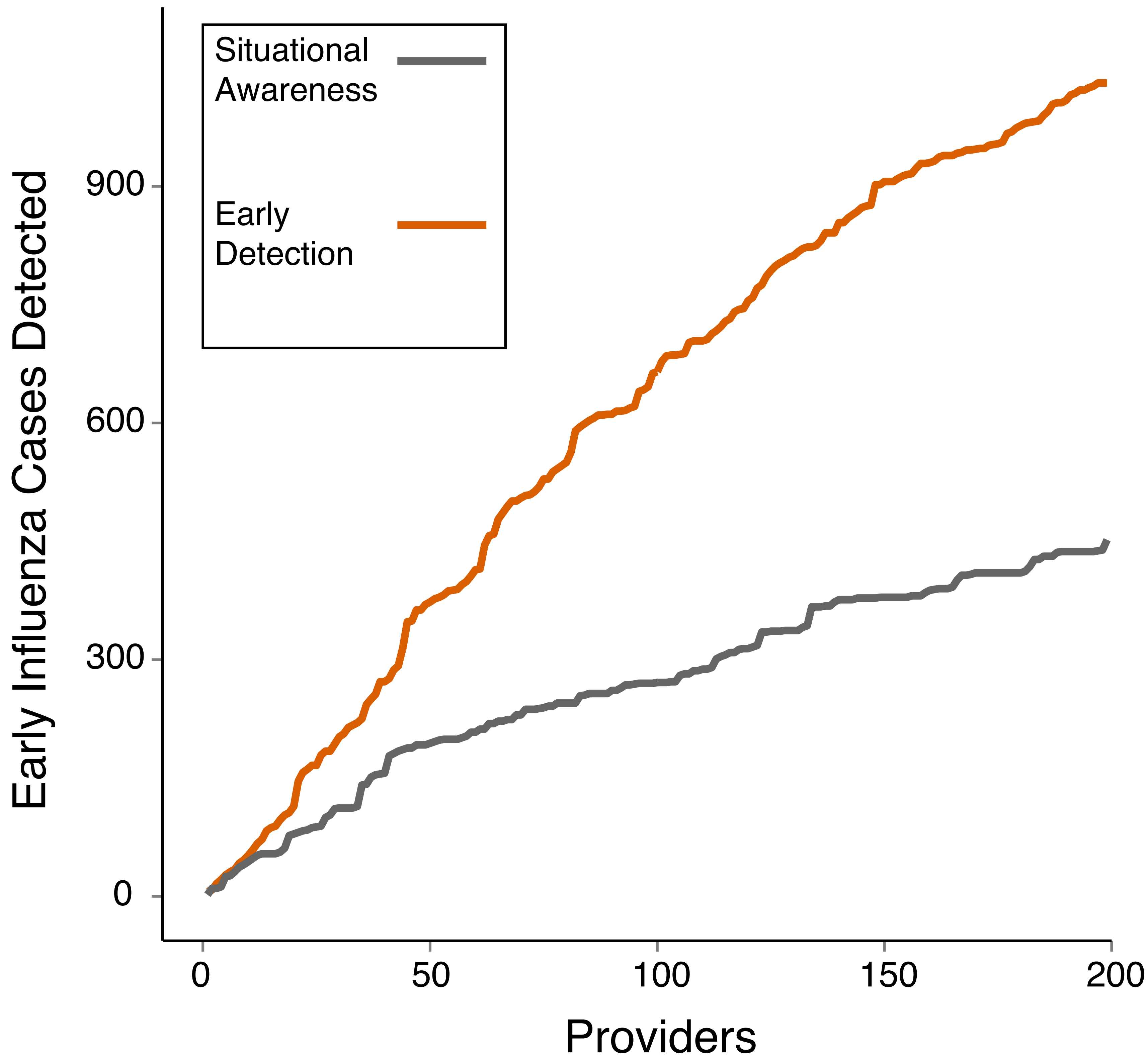
But what about pandemic influenza?

Early Detection

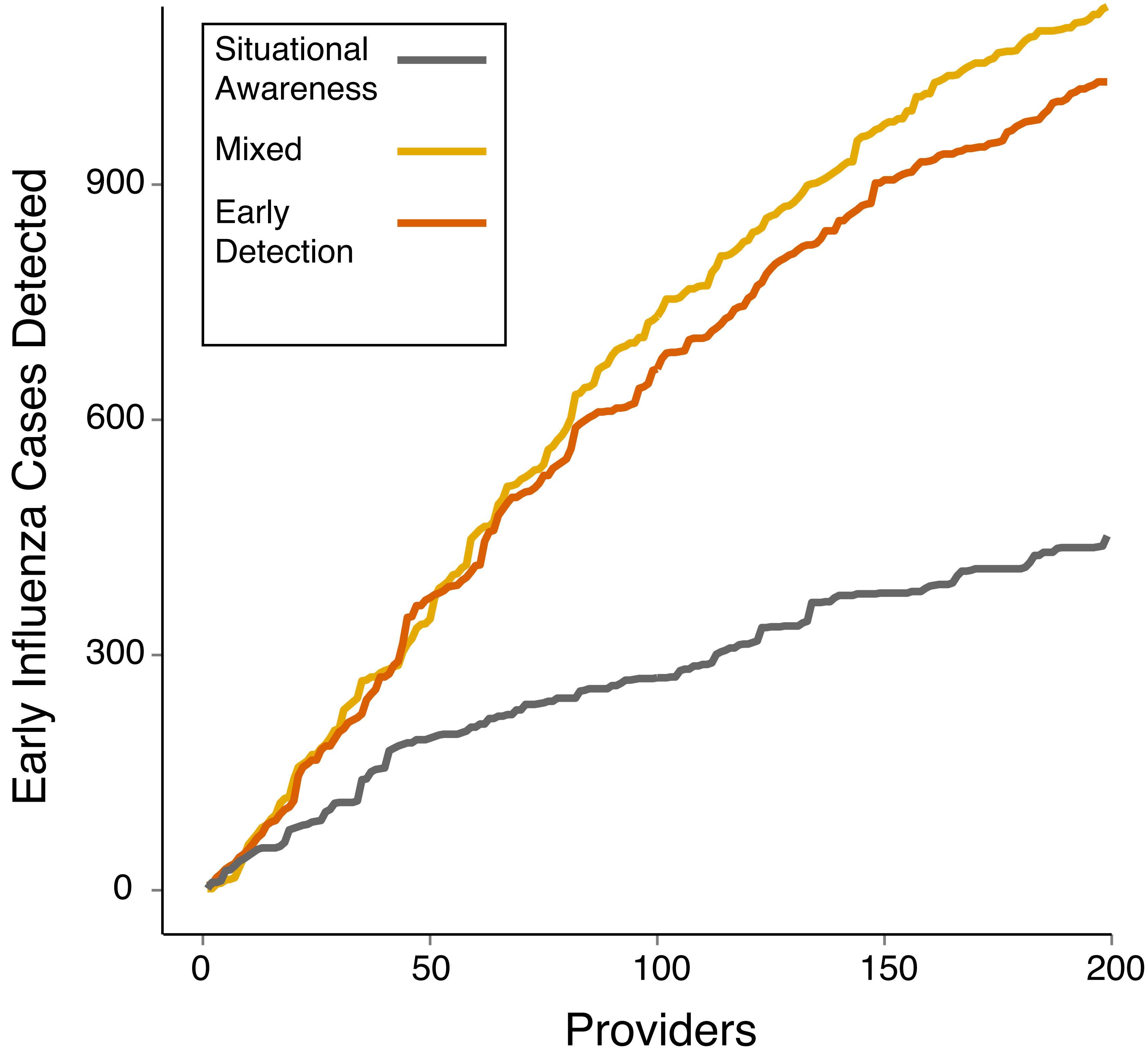
Situational Awareness



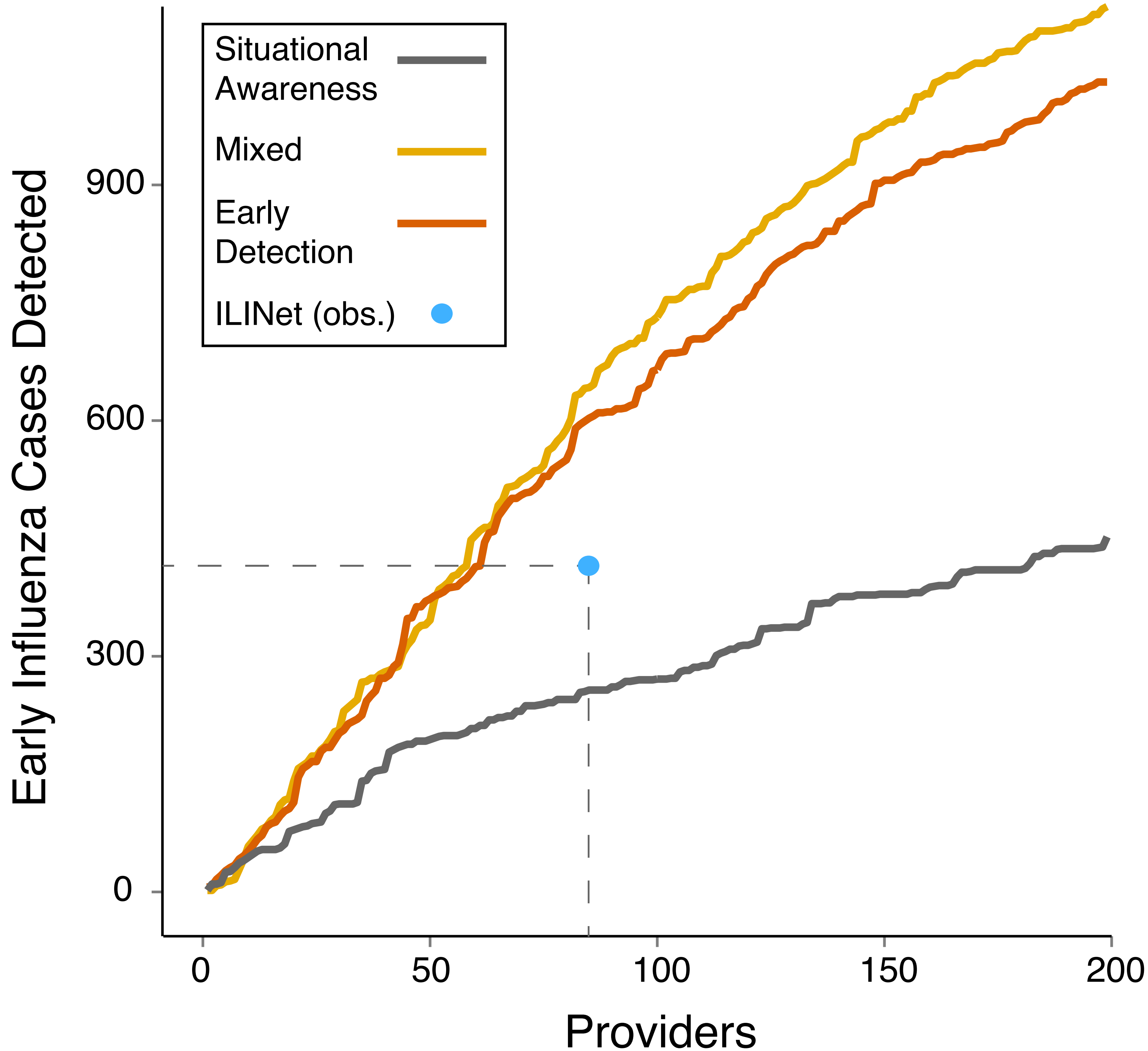
The RSquare network doesn't detect early cases



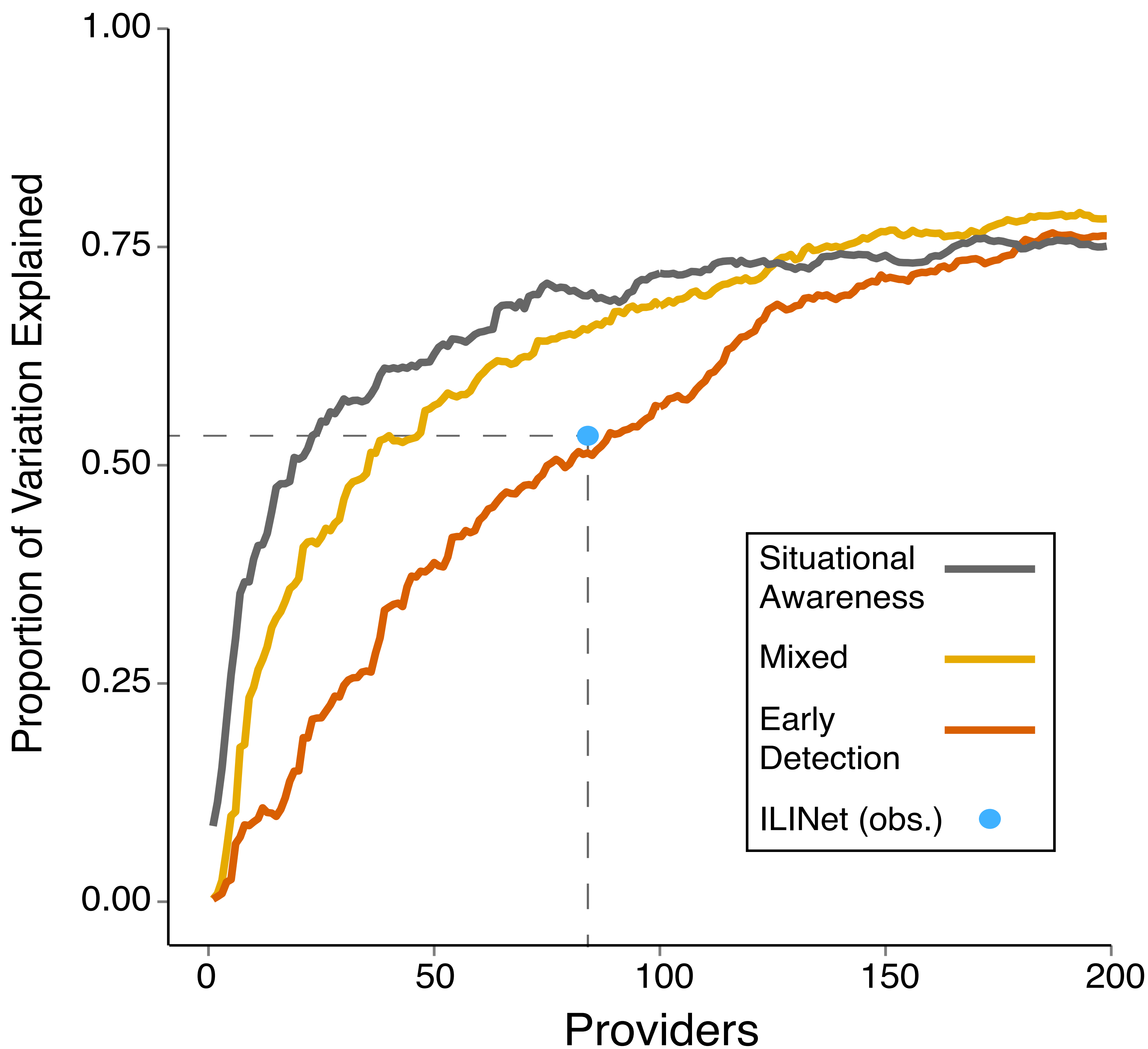
But a mixed network can!



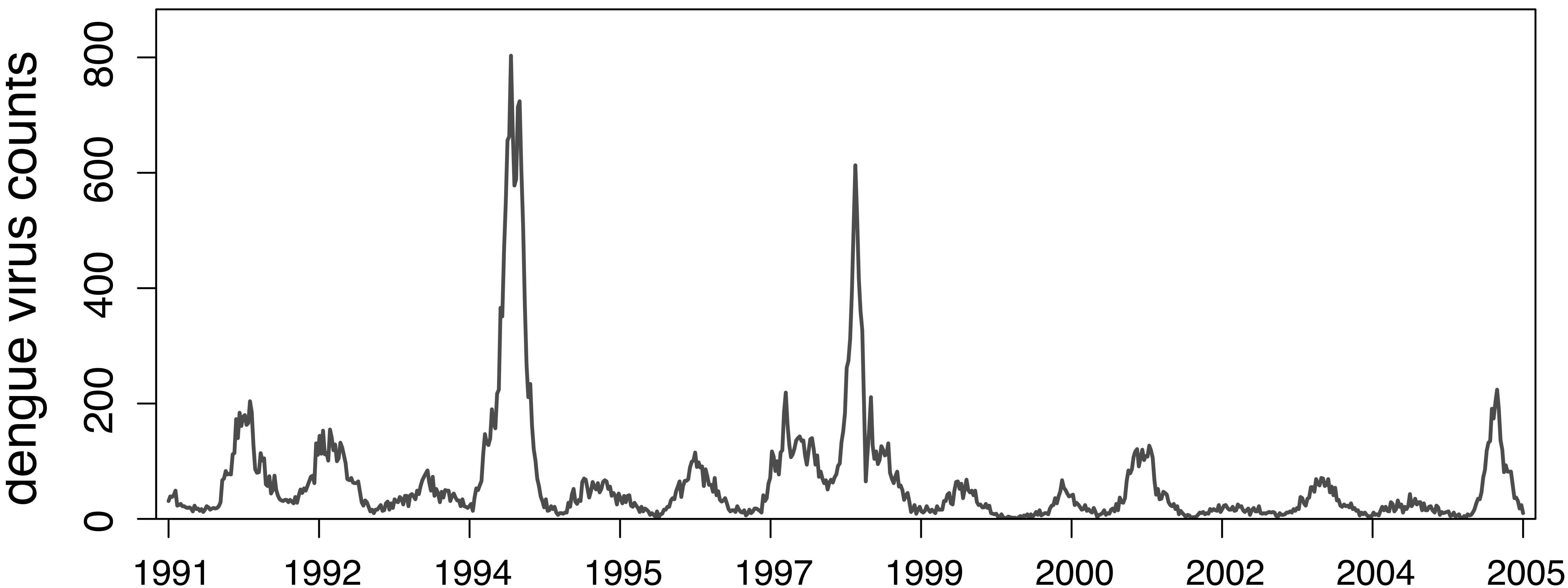
And we can benchmark ILINet



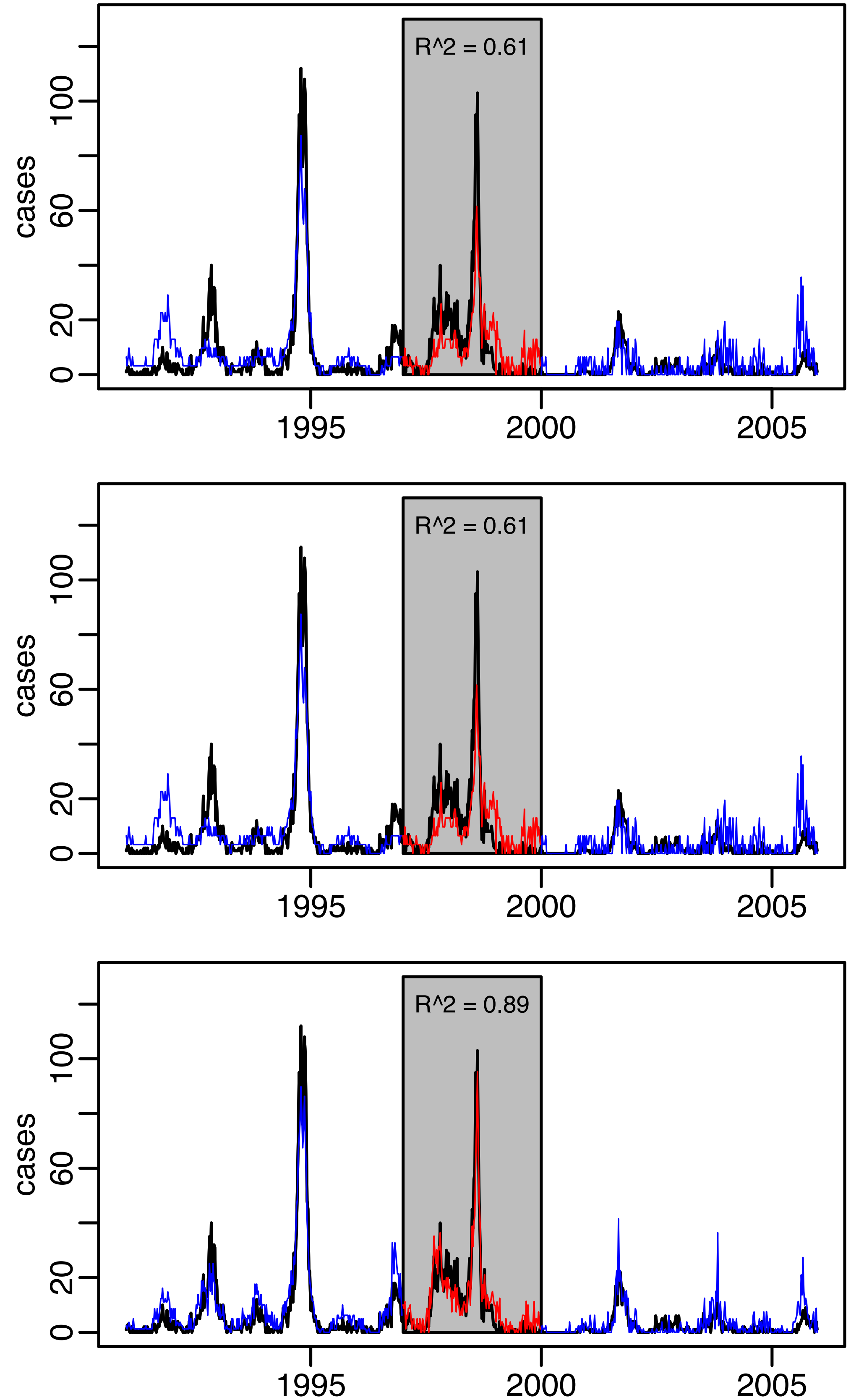
And the results hold for situational awareness



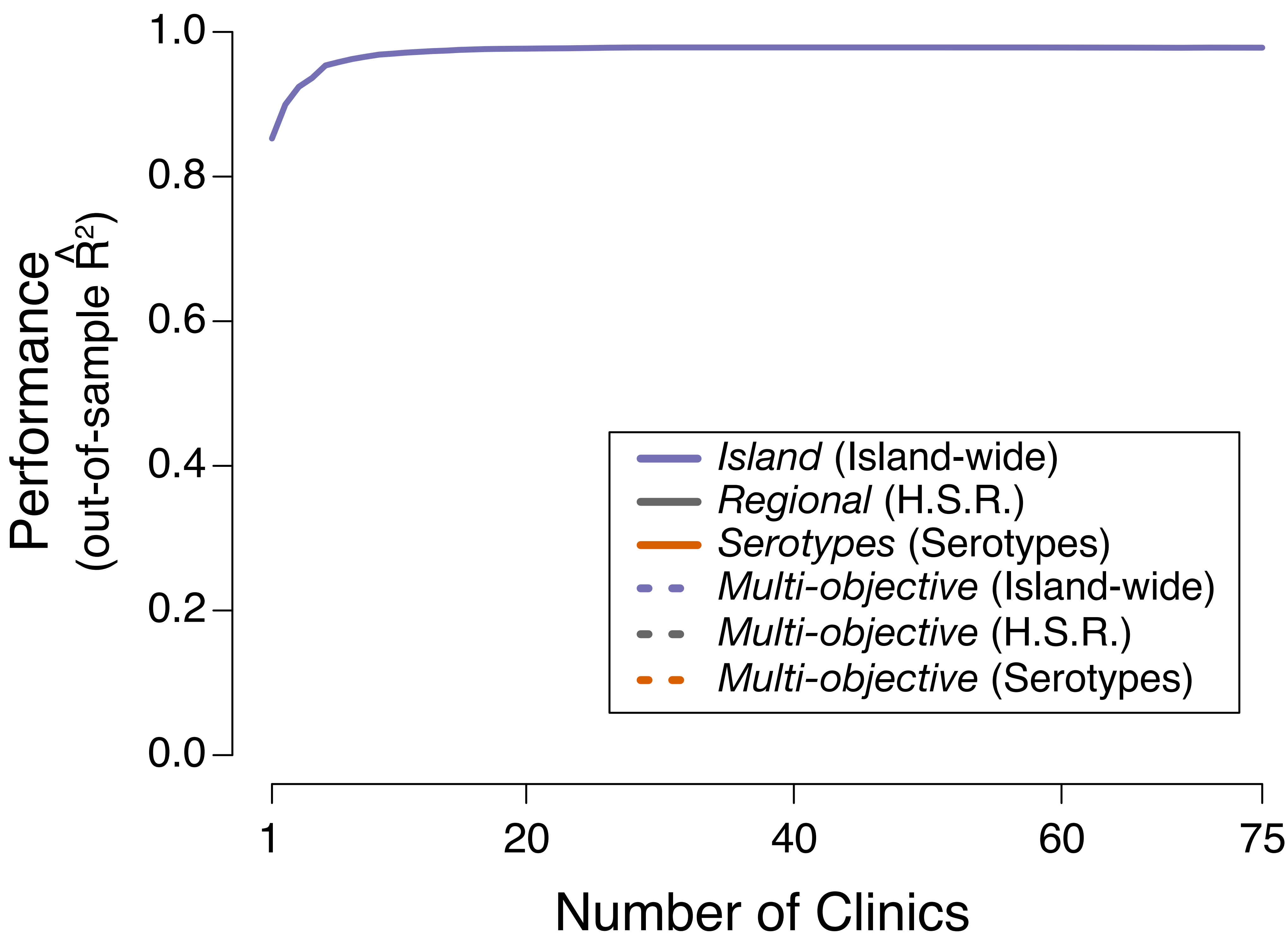
Dengue in Puerto Rico



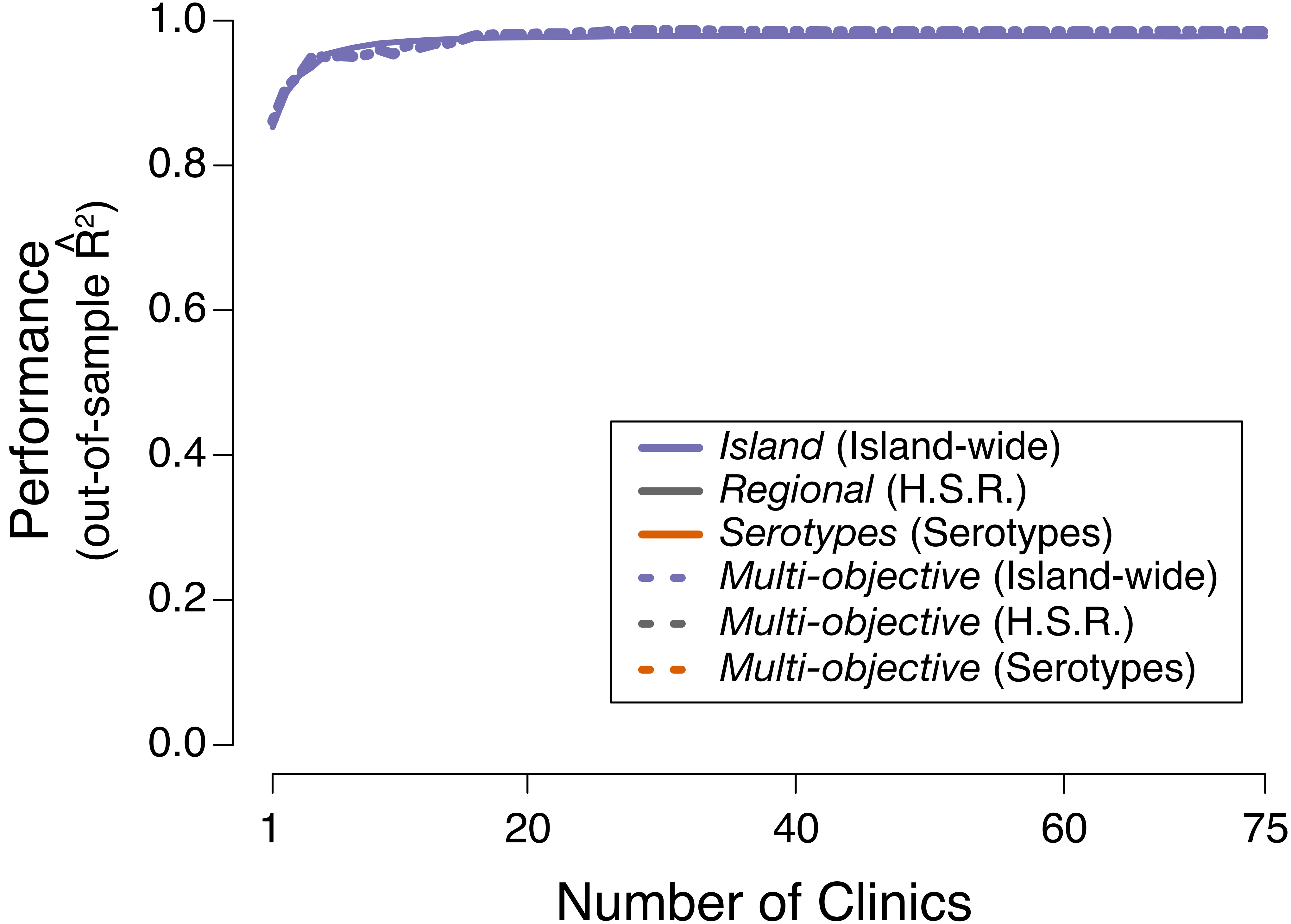
Greedy selection providers



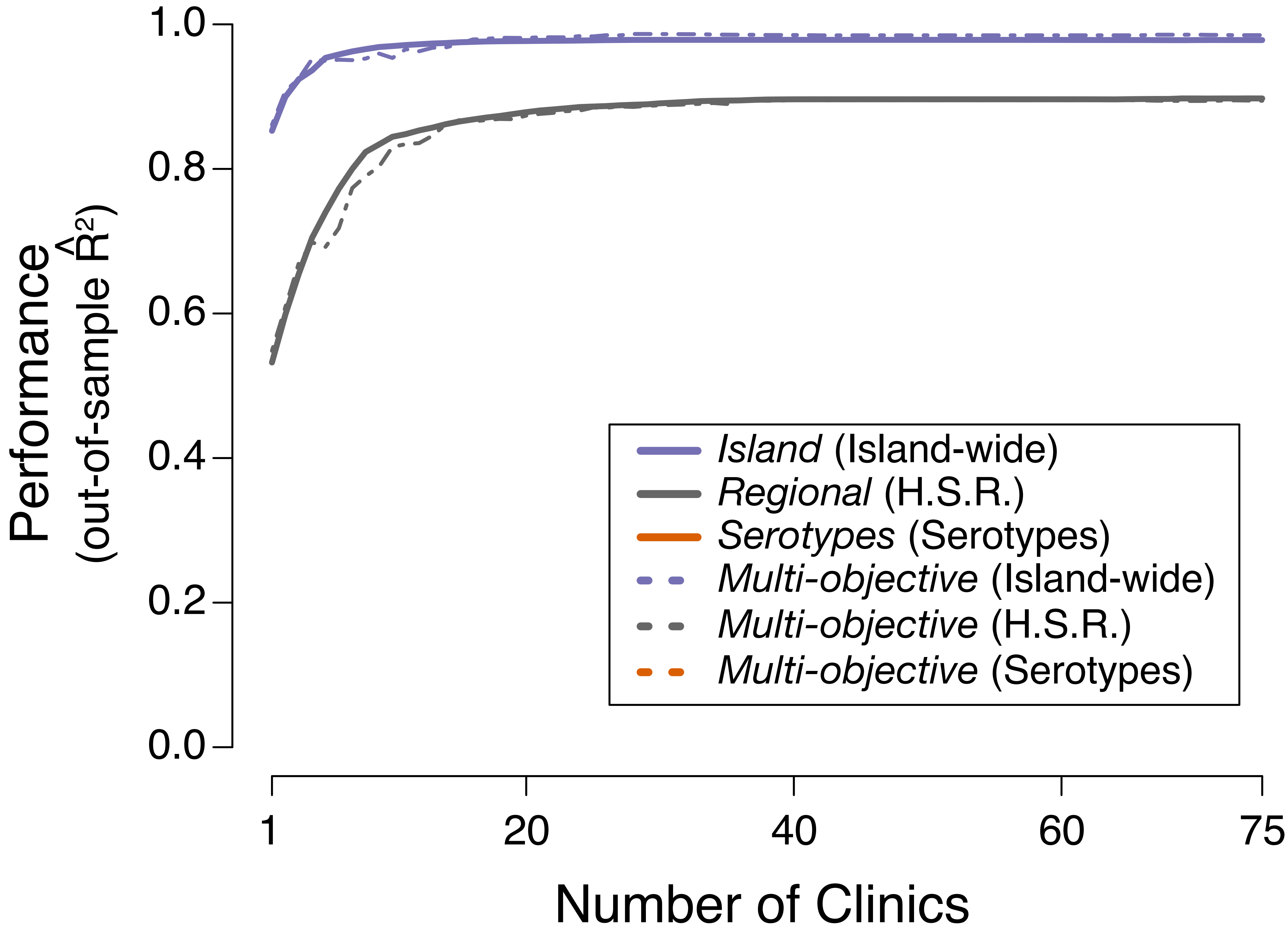
Island-wide incidence



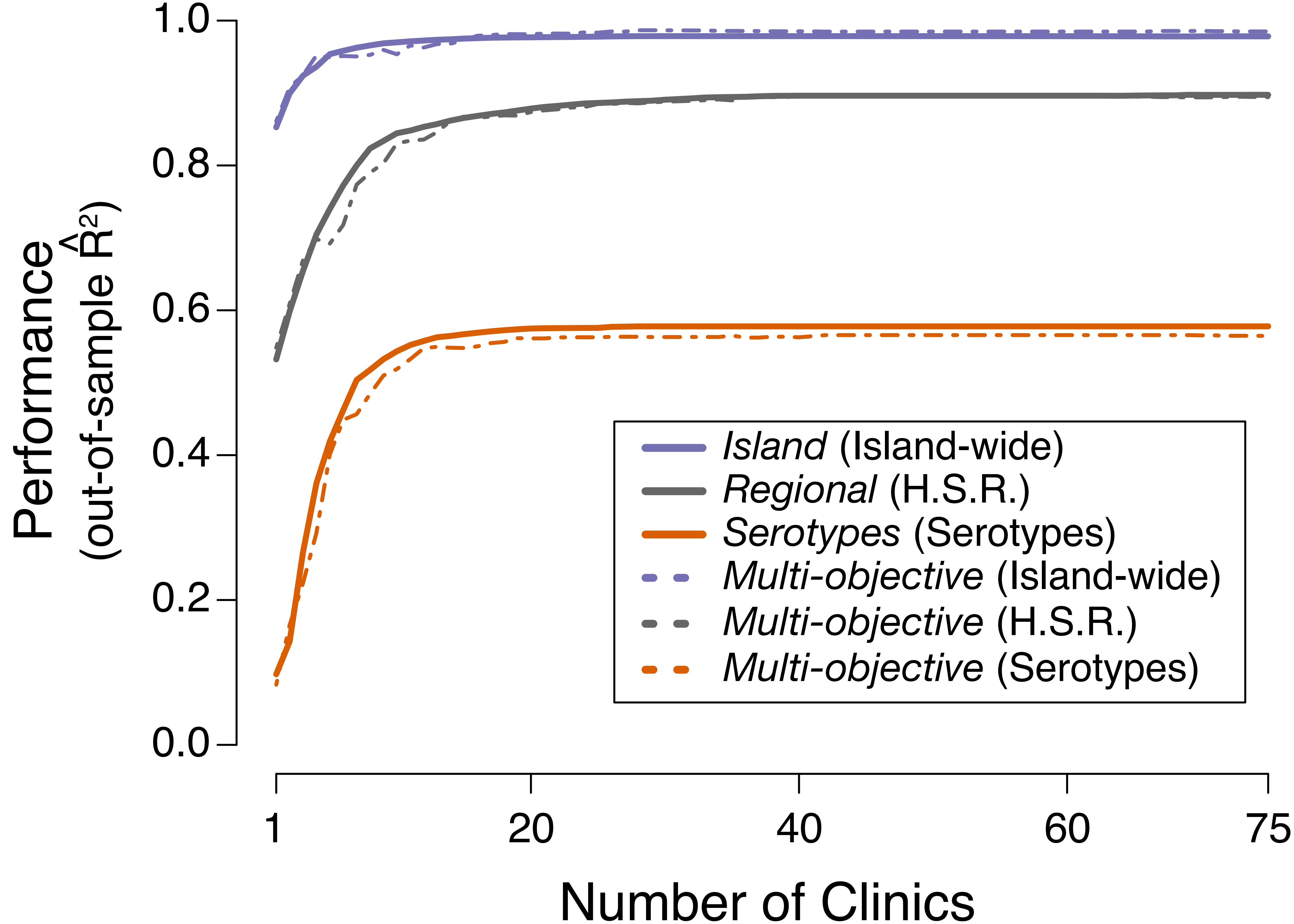
Multi-Objective Surveillance



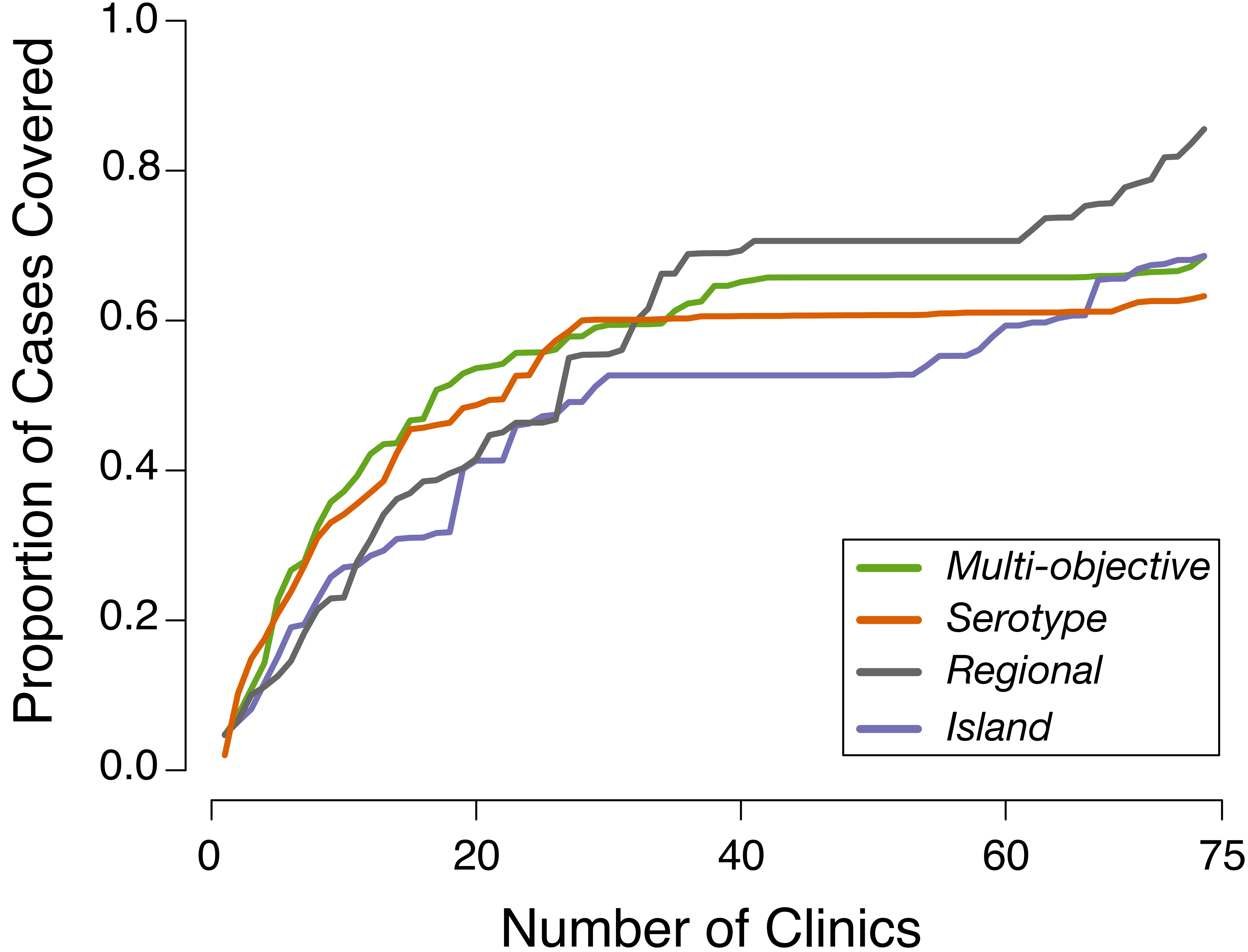
Regional surveillance



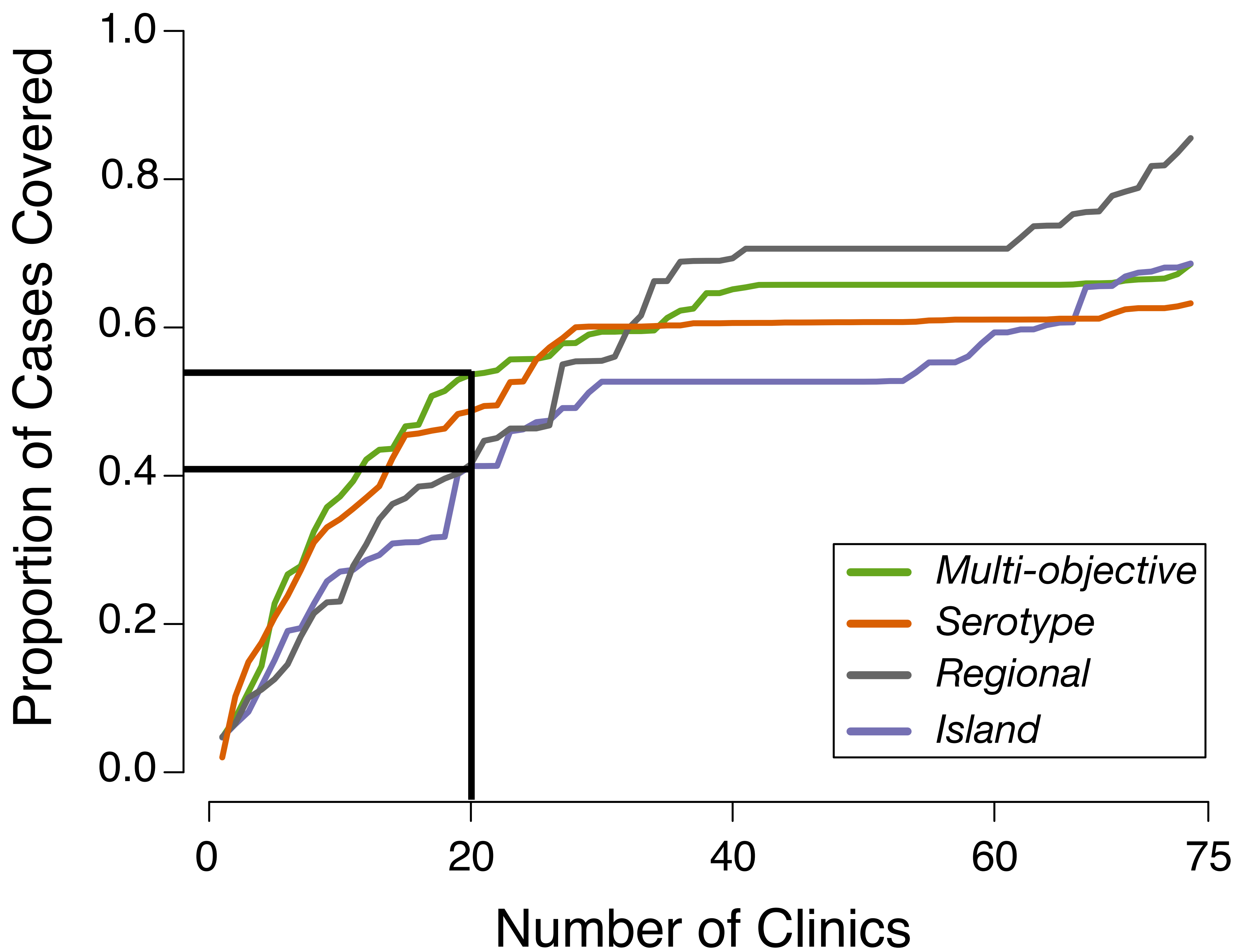
Serotype surveillance



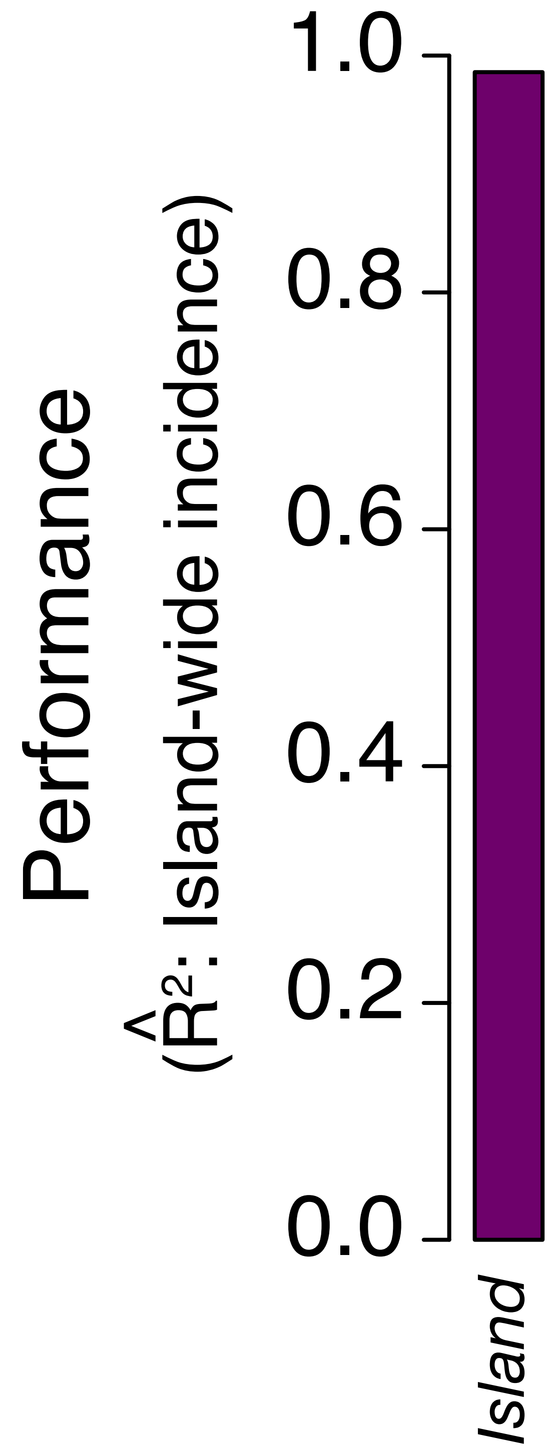
Performance without redundancy



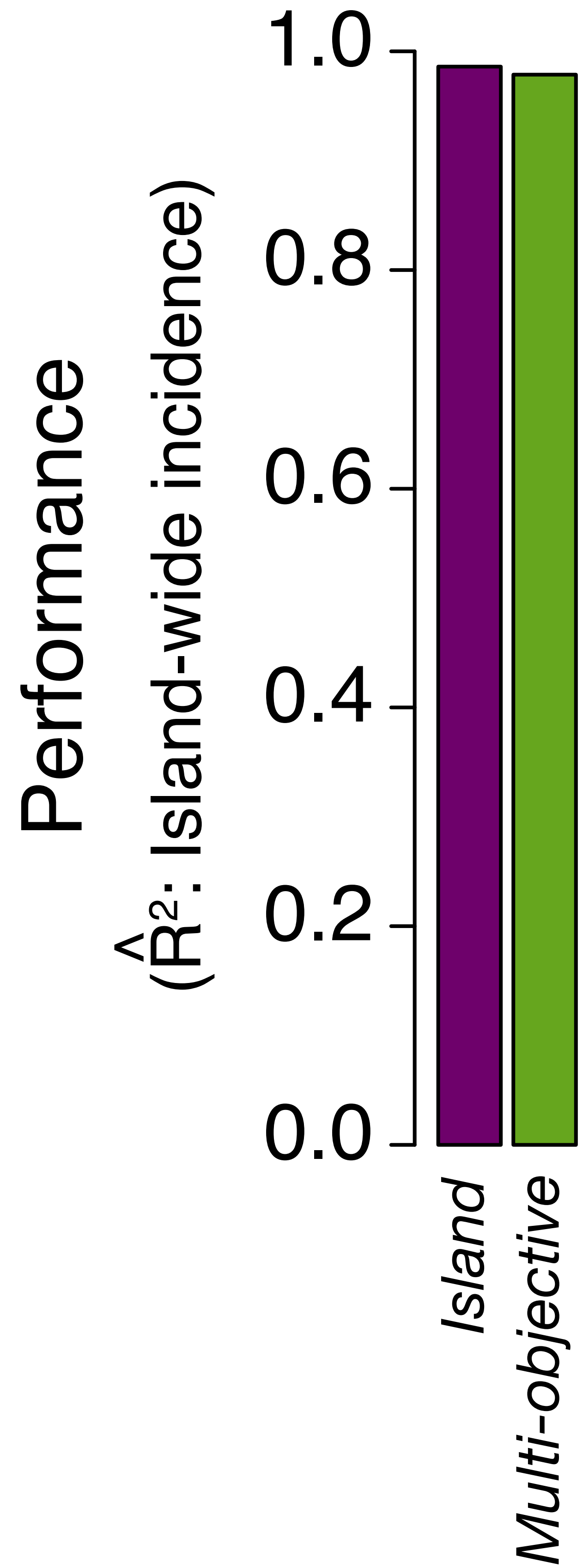
Performance without redundancy



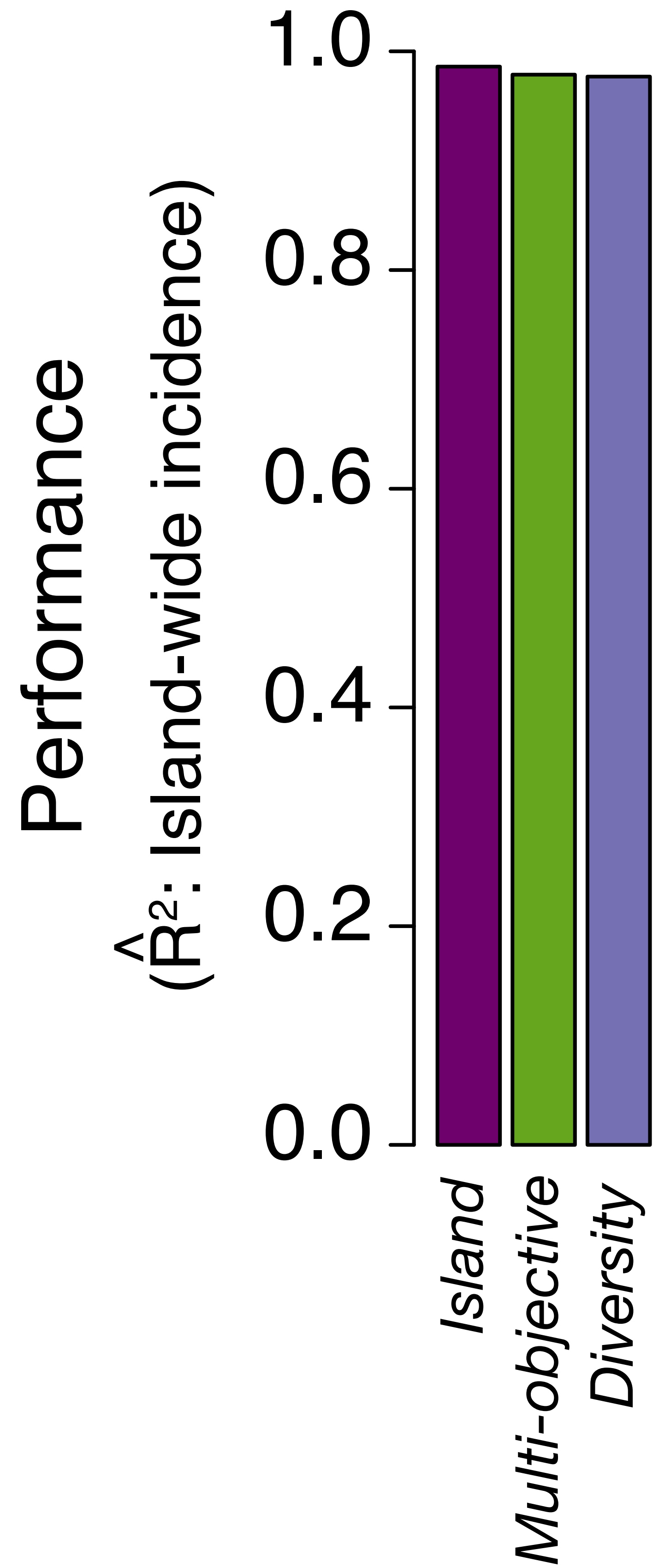
Alternative design algorithms



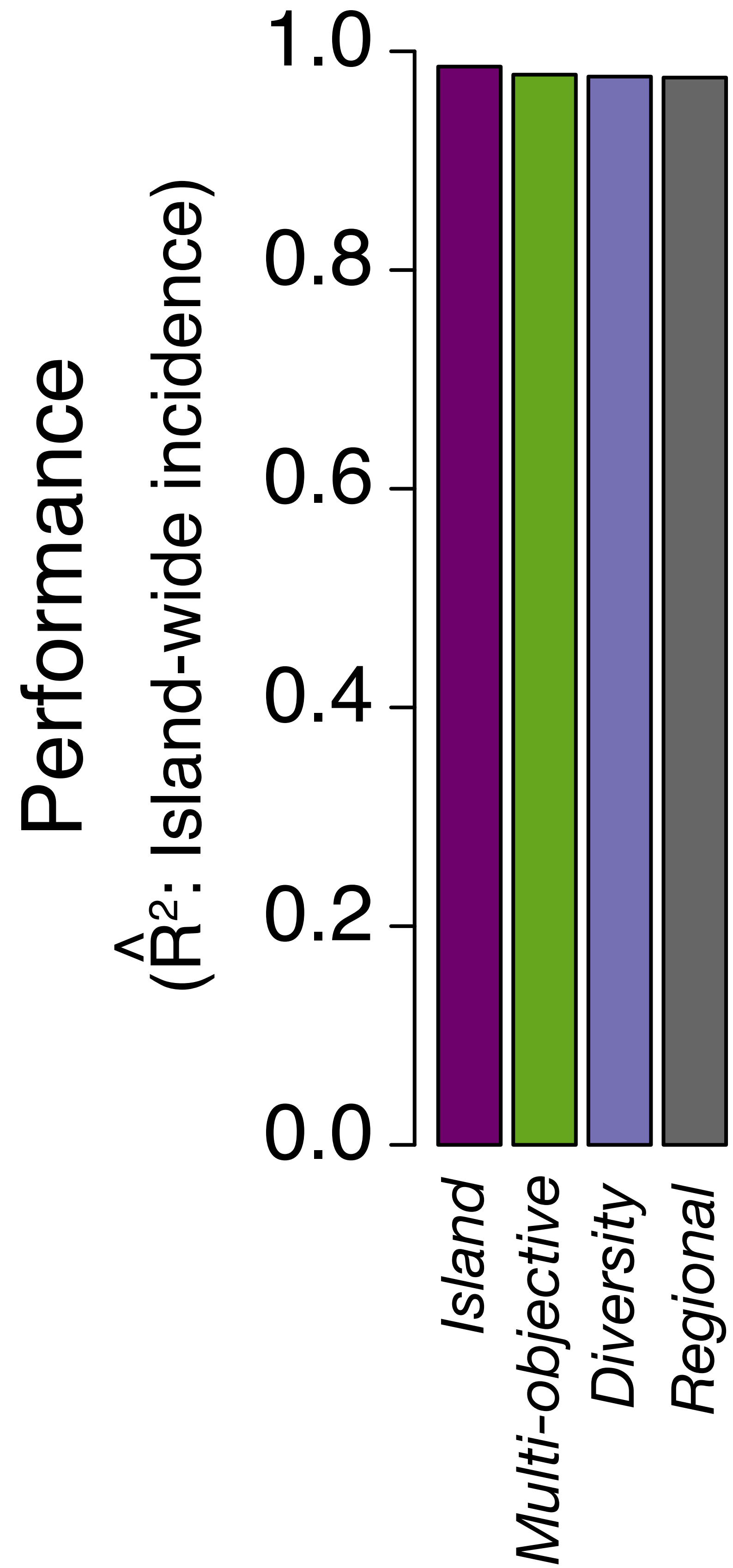
Alternative design algorithms



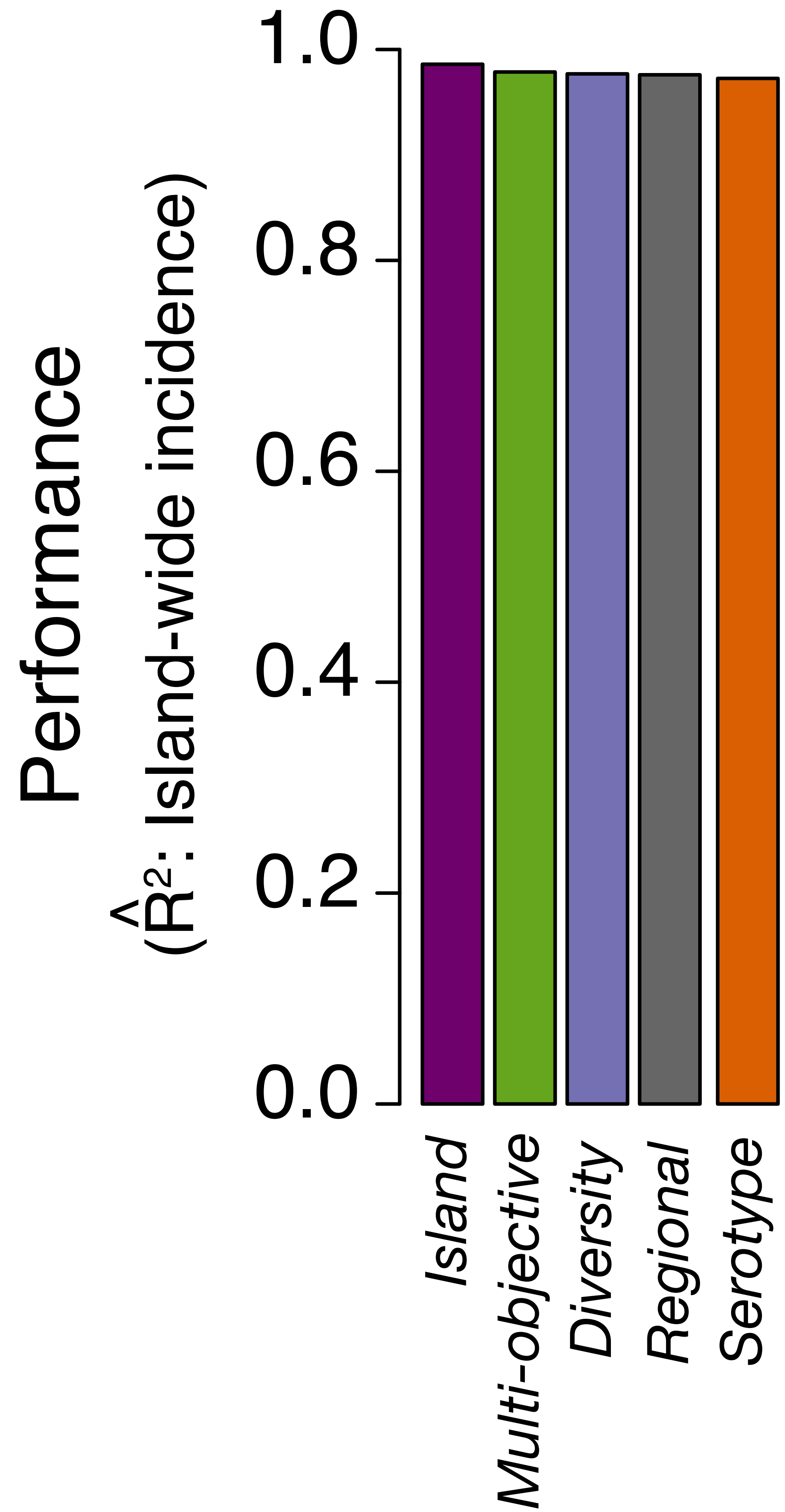
Alternative design algorithms



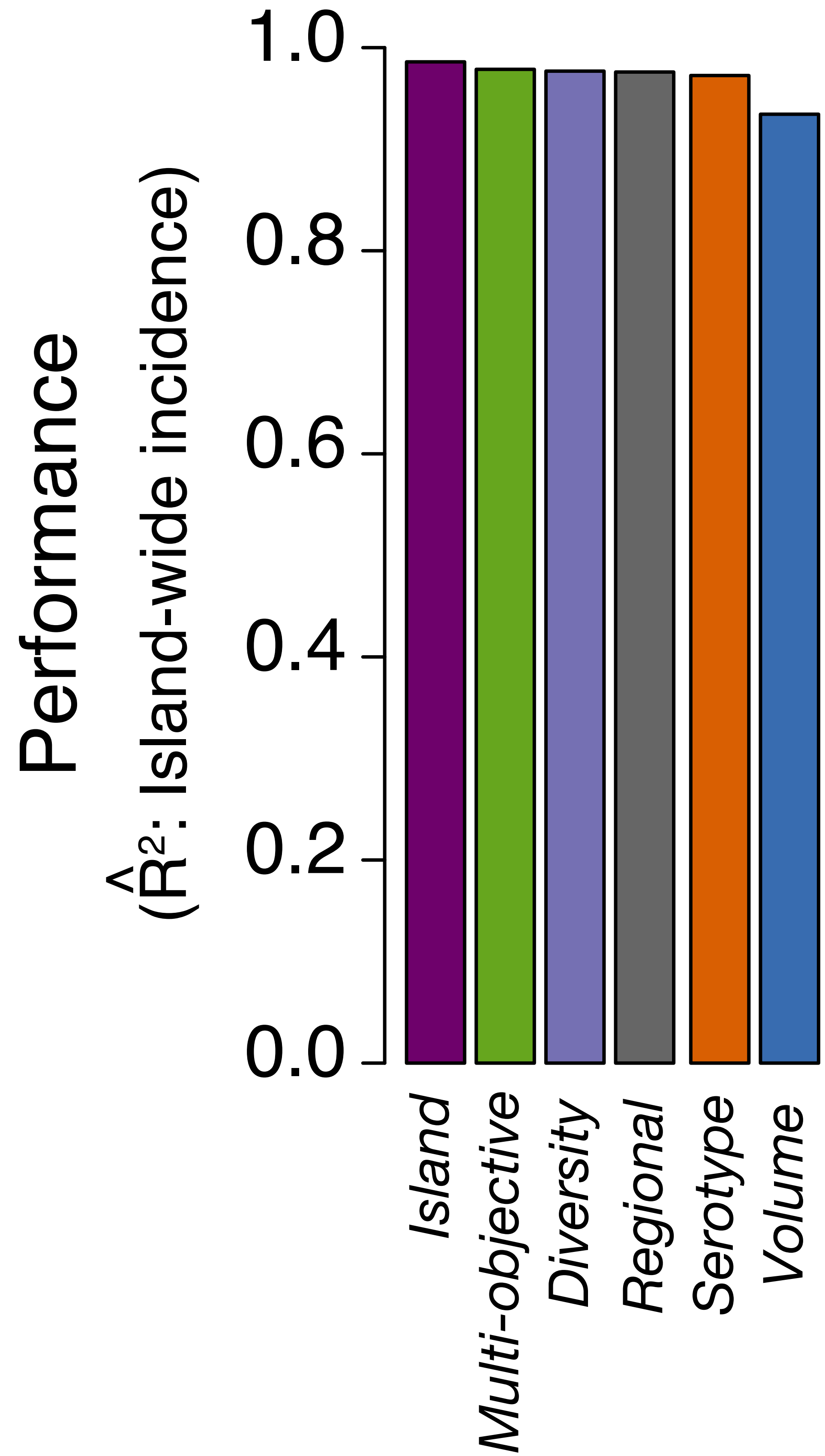
Alternative design algorithms



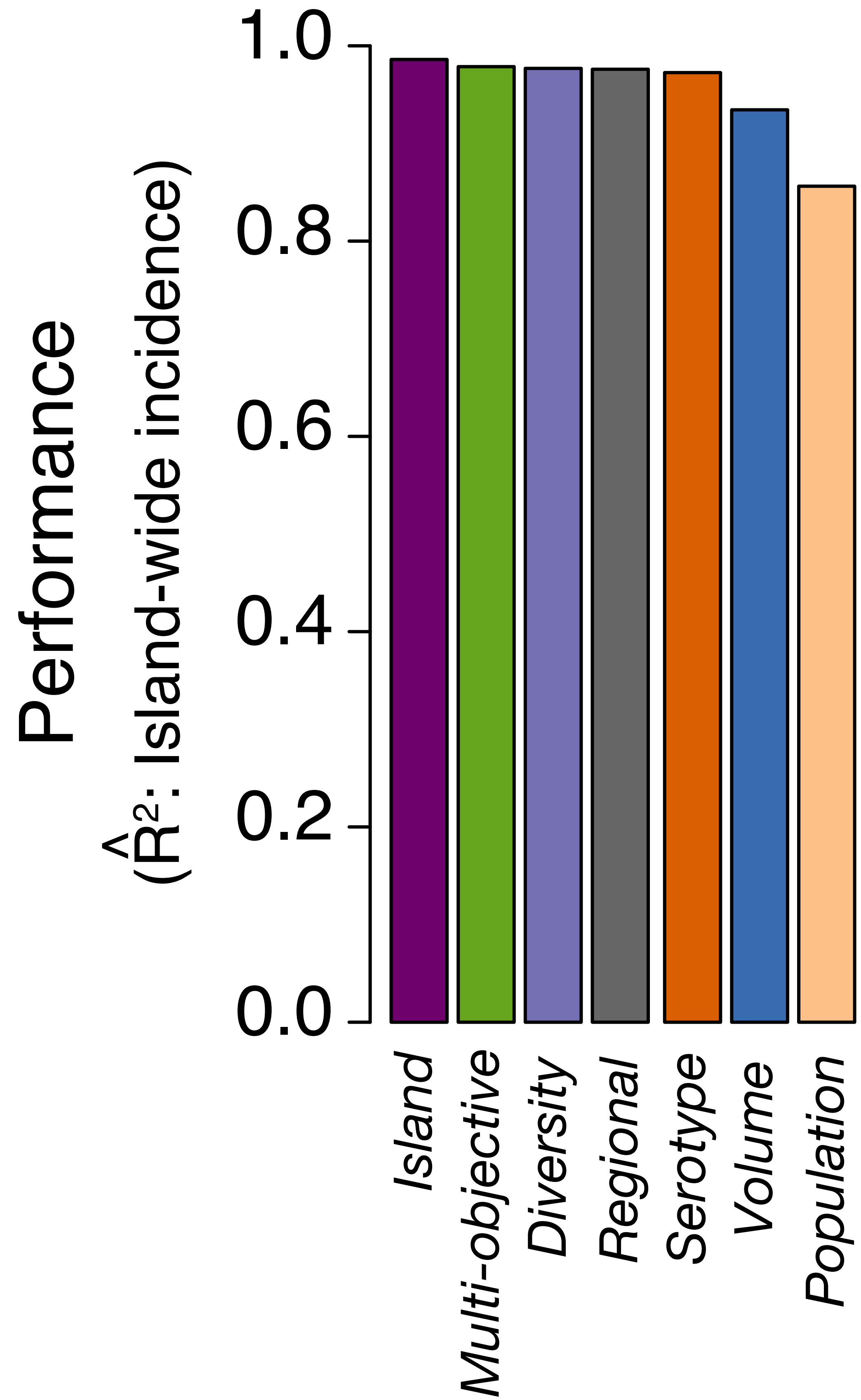
Alternative design algorithms



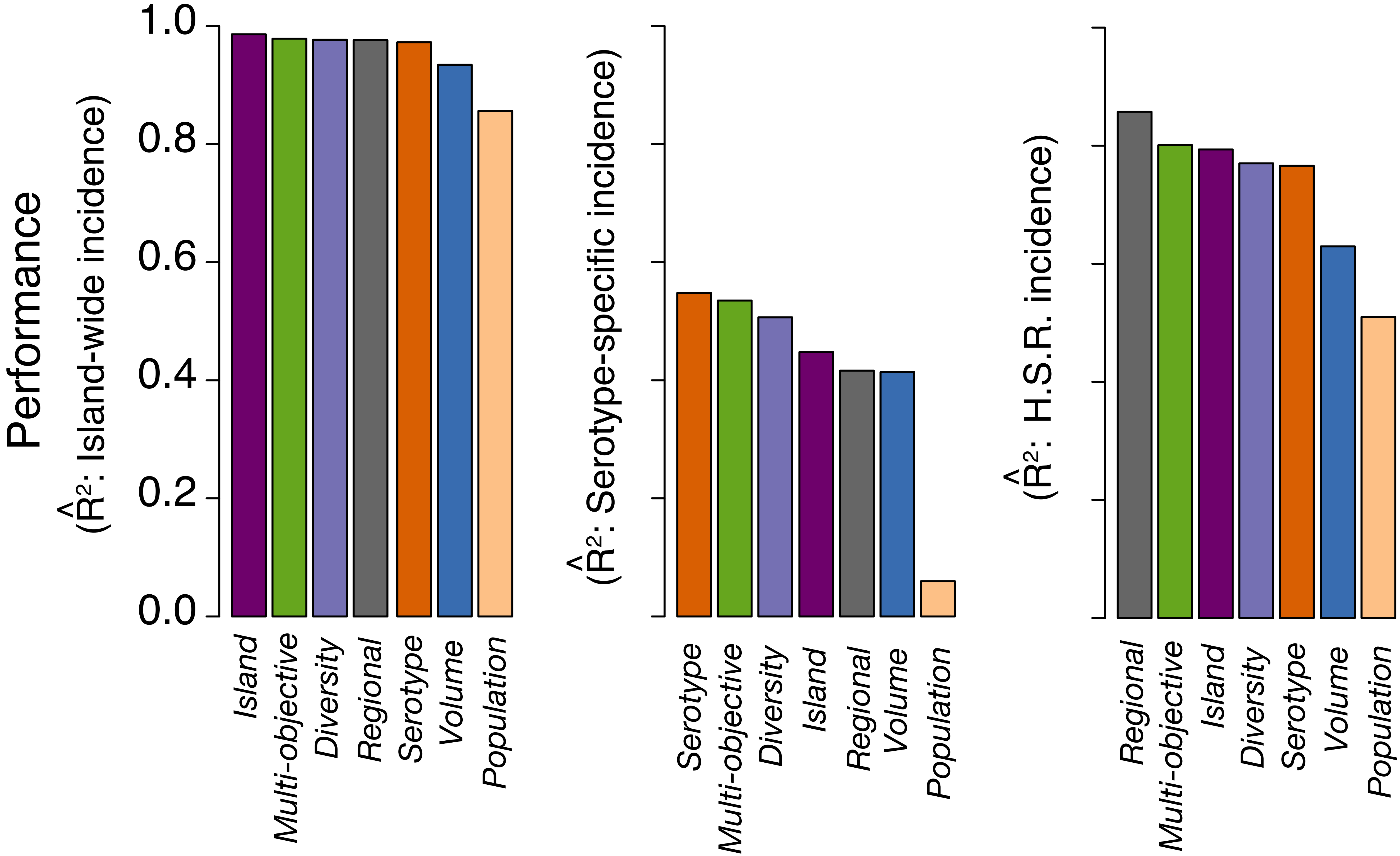
Alternative design algorithms



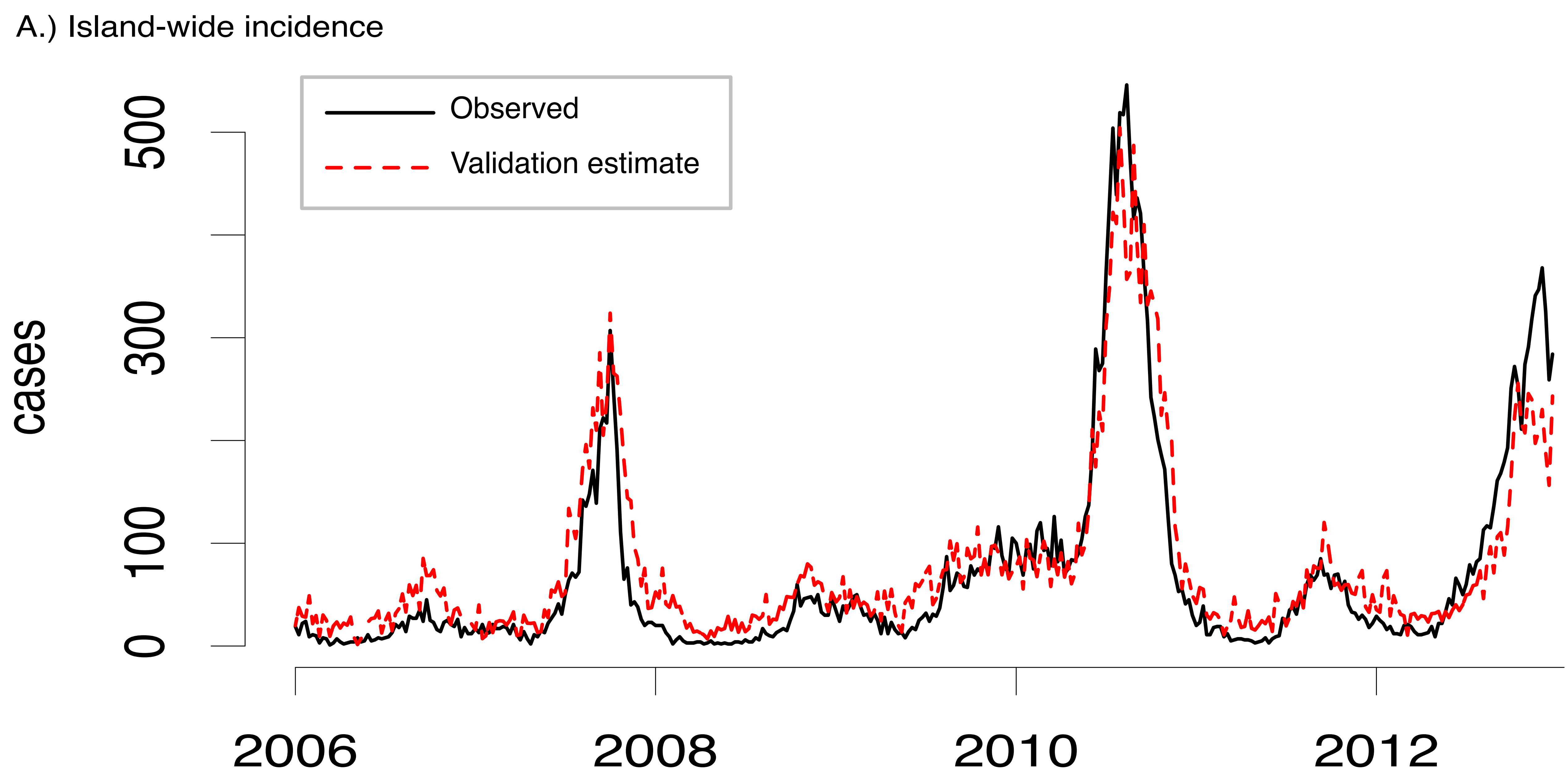
Alternative design algorithms



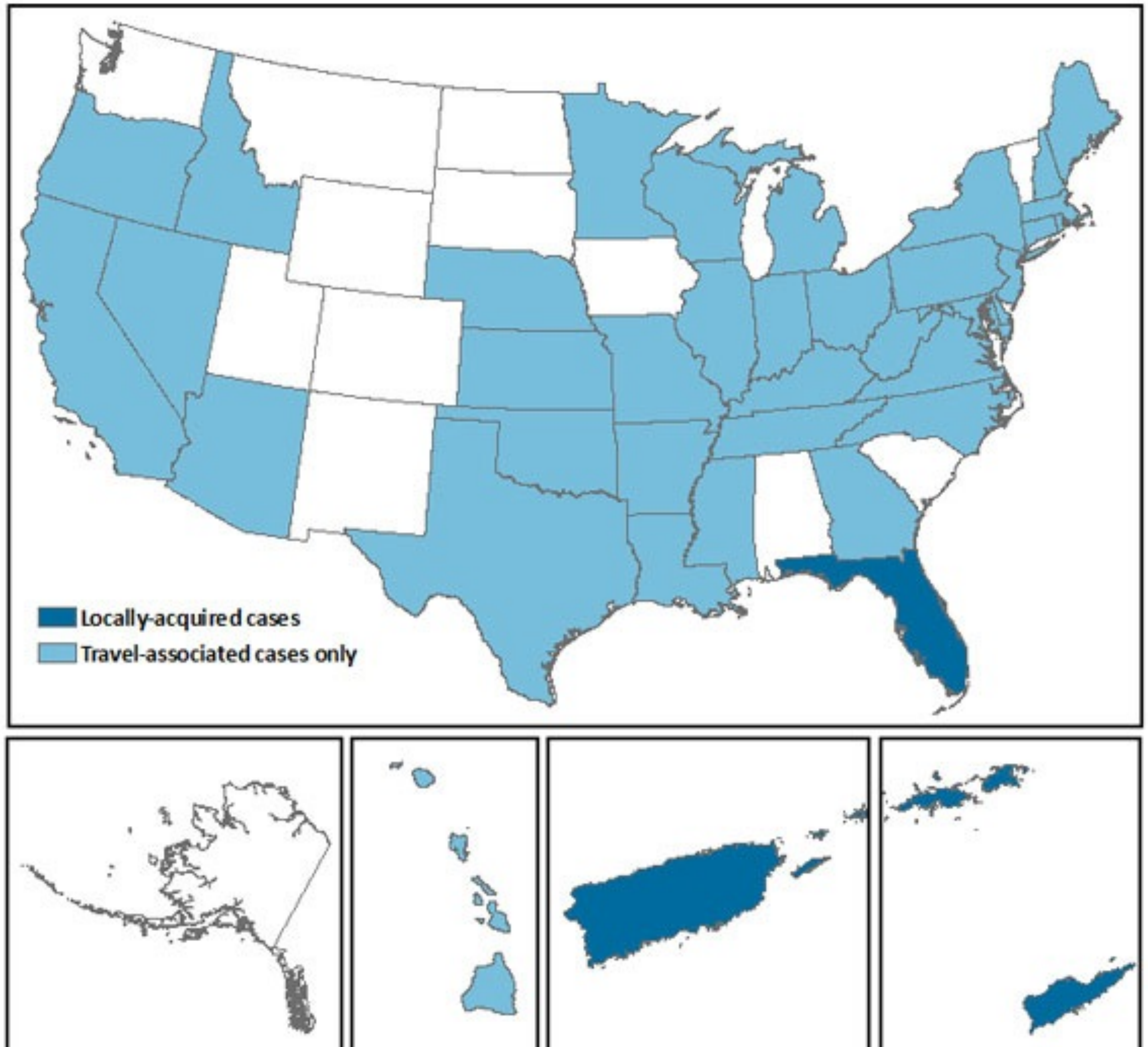
Similar pattern for the other objectives



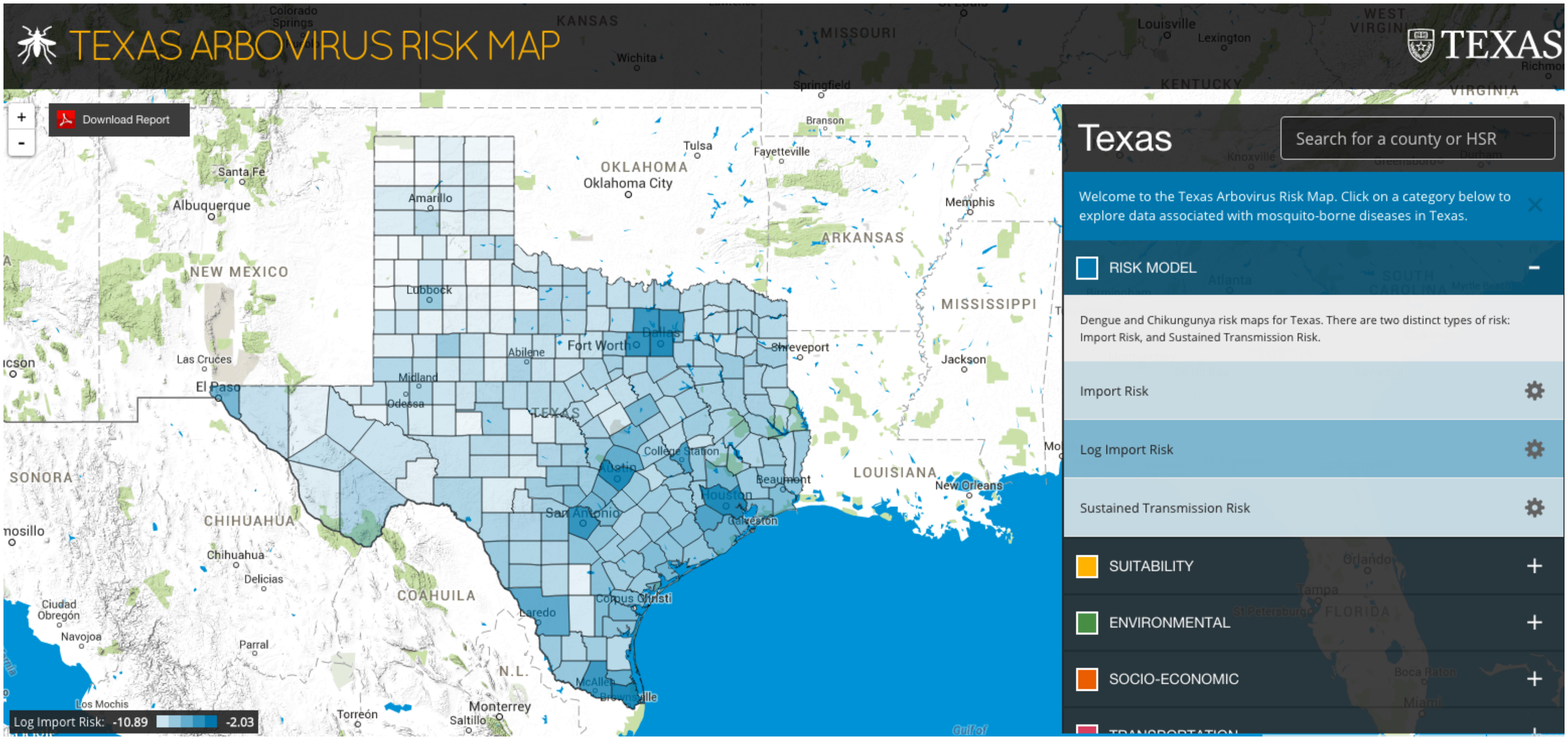
Hidden validation



Chikungunya Surveillance



Texas Arbovirus Risk Map



Four Step Program

1. Formalize the public health objectives

2. Specify candidate data sources & acquire historical data

3. Simulate data where missing

4. Select the most informative data sources and validate!

Scarpino, Dimitrov, Meyers 2012

Acknowledgments

Ned Dimitrov

Lauren Ancel Meyers

Michael Johansson

Bruce Clements

Lyn Finelli

Alison Galvani

NIH MIDAS

DTRA

The Santa Fe Institute

The Omidyar Group

Questions?

Samuel V. Scarpino
Omidyar Fellow
Santa Fe Institute

scarpino@santafe.edu
@svscarpino
scarpino.github.io