PNAS
www.pnas.org

Supplementary Information for

Identifying asymptomatic spreaders of antimicrobial-resistant pathogens in hospital settings

Sen Pei[a,1], Fredrik Liljeros[b,c], and Jeffrey Shaman[a,1]

[a]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, United States

[b]Department of Sociology, Stockholm University, Stockholm, Sweden

[c]Department of Public Health Sciences, Karolinska Institutet, Solna, Sweden

[1]To whom correspondence should be addressed. E-mail: sp3449@cumc.columbia.edu or jls106@cumc.columbia.edu.

This PDF file includes:

Supplementary text
Figures S1 to S9
Table S1
SI References

Supplementary Information Text

The agent-based model

The ODD description (1) of the agent-based model is as follows.

Purpose: The purpose of the model is to simulate the spread of AMROs in healthcare systems.

State variables and scales: The model spans three hierarchical levels in hospital: individual, ward, and total population. Model dynamics are defined at the individual level, but outcomes can be aggregated to any of the three scales. Individuals are characterized by state variables: colonization status (susceptible or colonized), hospitalization time, and the location (ward) where the patient resides. Hospitalization time and location data provide information on patient transfer within healthcare systems. For each ward, an additional state variable is defined: the force of infection attributed to environmental contamination ε. Six parameters are introduced: 1) the baseline transmission probability upon contact, β; 2) the importation probability of colonization, γ; 3) the baseline environmental contamination coefficient, θ; 4) the mean environmental decolonization period, D; 5) the patient decolonization probability, α; and 6) the probability that a carrier is observed, ρ.

Process overview and scheduling: The model proceeds in daily time steps. Within each time step, five modules are processed: transmission between patients staying in the same room, environmental contamination within each room, transmission from the environment to patients, importation of colonized patients from the community, and observation of carriers. Transmission in the community is not explicitly simulated but is reflected by the importation rate of colonization.

Design concepts: *Emergence*: HAI outbreaks emerge from the movement of individuals within healthcare systems. *Interaction*: HCW-mediated contacts between patients facilitate transmission of AMROs. We model HCW-mediated transmission indirectly by assuming that AMROs can spread between all pairs of patients staying in the same room at the same time. Colonization of patients may spillover to contaminate the environment resulting in indirect transmission to patients admitted to the same room at a later time. Nosocomial transmission interacts with the community through the admission and discharge of colonized patients. *Stochasticity*: Transmission, importation, decolonization and laboratory testing are all run stochastically according to predefined probabilities. Distributions of outcomes can be generated through repeated model simulations. *Observation*: Carriers of AMR pathogens are observed with a given probability.

Initialization: Each patient in hospital at the beginning of a simulation is randomly assigned as a carrier with a probability drawn uniformly from [0, 5%]. The environmental force of infection in each room is set to zero initially. If colonization/infection information is available for each ward, we can randomly assign carriers in wards with reported colonization/infection.

Input: We use the daily admission-discharge-transfer record from the Swedish hospital dataset to inform patient movement within the model.

Submodels: The contact network within a collection of hospitals is represented by a time-varying graph *G* constructed using the actual hospitalization records (see an example in Fig. 1C). In this contact network, nodes represent uniquely labeled patients, connected by undirected links among individuals sharing a room at a given time. Individuals are classified into two categories: Susceptible (*S*) and Colonized (*C*). Within hospital, transitions between these states are governed by model transmission dynamics. ***Contact transmission*:** A susceptible individual *i* can be colonized, with probability $\beta/(n_{r_i}-1)$, upon contact with a colonized person *j* who is directly linked to *i* in the contact network *G*. Here n_{r_i} is the capacity of the room in which patient i resides. We use a frequency dependent transmission model here as the chance of person-to-person contact decreases in larger rooms (with denominator $n_{r_i}-1$, to exclude the focal patient) (2). ***Environmental contamination*:** Each colonized patient in a given room contributes a daily θ/n_{r_i} increment to the environmental force of infection ε_{r_i}. Meanwhile, ε_{r_i} decays to $1/D$ of its

prior value per day. A susceptible individual in room r_i becomes colonized with probability ε_{r_i} due to environmental contamination. ***Importation*****:** For new admissions, patients are colonized with a probability γ. ***Observation*****:** Each carrier in hospital is tested and observed with probability ρ.

Master equation derivation

Denote S_i^t and C_i^t as the probability of individual i being susceptible and colonized on day t, and $E_{r_i}^t$ as the colonization probability of ward r_i at time t. If the states of neighboring individuals in the contact network are independent, the evolution of S_i^t and C_i^t can be described by a set of ordinary differential equations:

$$\frac{dS_i^t}{dt} = \underbrace{\alpha C_i^t}_{Decoloniztion} - \underbrace{\frac{\beta}{n_{r_i}-1} S_i^t \sum_{j\in\partial i} C_j^t}_{Contact\ transmission} - \underbrace{\beta_e E_{r_i}^t S_i^t}_{Environmental\ contamination}, \qquad [S1]$$

$$\frac{dC_i^t}{dt} = \underbrace{\frac{\beta}{n_{r_i}-1} S_i^t \sum_{j\in\partial i} C_j^t}_{Contact\ transmission} + \underbrace{\beta_e E_{r_i}^t S_i^t}_{Environmental\ contamination} - \underbrace{\alpha C_i^t}_{Decoloniztion}, \qquad [S2]$$

$$\frac{dE_{r_i}^t}{dt} = \underbrace{-\frac{1}{D} E_{r_i}^t}_{Environmental\ decolonization} + \underbrace{\frac{\delta}{n_{r_i}} \sum_{j\ in\ r_i} C_j^t}_{Environmental\ colonization}. \qquad [S3]$$

Here α is the patient decolonization rate, β is the baseline contact transmission rate, r_i is the ward in which patient i resides, n_{r_i} is the occupancy of ward r_i, β_e is the environmental transmission rate, δ is the baseline environmental colonization rate, D is the mean environmental decolonization period, and ∂i is the set of patients in contact with patient i on day t. Note in Eqs. [S1-S2] we omitted the high-order term – the probability that patient i is simultaneously colonized by contact and environmental contamination. The environmental force of infection in ward r_i can be defined as $\varepsilon_{r_i}^t = \beta_e E_{r_i}^t$. Using $\varepsilon_{r_i}^t$, Eqs. [S1-S3] can be re-written as

$$\frac{dS_i^t}{dt} = \alpha C_i^t - \frac{\beta}{n_{r_i}-1} S_i^t \sum_{j\in\partial i} C_j^t - \varepsilon_{r_i}^t S_i^t, \qquad [S4]$$

$$\frac{dC_i^t}{dt} = \frac{\beta}{n_{r_i}-1} S_i^t \sum_{j\in\partial i} C_j^t + \varepsilon_{r_i}^t S_i^t - \alpha C_i^t, \qquad [S5]$$

$$\frac{d\varepsilon_{r_i}^t}{dt} = -\frac{1}{D}\varepsilon_{r_i}^t + \frac{\beta_e \delta}{n_{r_i}} \sum_{j\ in\ r_i} C_j^t. \qquad [S6]$$

Define $\theta = \beta_e \delta$ as the baseline environmental contamination coefficient and discretize Eqs. [S4-S6] using an interval of $\Delta t = 1$ day, we obtain Eqs. [1-3] in the main text. Note here θ is the product of the environmental transmission rate β_e (the probability that a susceptible individual is colonized by the environment) and the baseline environmental colonization rate δ (the probability that the environment is colonized by a colonized patient). Assuming the units of β_e and δ are 1/day, the unit of θ is therefore (1/day)^2.

The sequential individual-level inference algorithm

Framework. In the SILI algorithm, an ensemble of system states, which represent the distribution of probabilities S_i^t and C_i^t for all patients, are sequentially adjusted using individual-level diagnostic information. At each time t, the SILI algorithm proceeds with the following three steps (see an illustration in Fig. S2). In this example, two patients are observed positive after time t (Fig. S2A).

1) Backward temporal propagation: We use the Bayes' rule to propagate information backward and estimate the colonization probability of observed carriers at time t (Fig. S2B).

2) Covariability adjustment: We use cross-ensemble covariability to adjust the colonization probability of patients who have contact with observed carriers (Fig. S2C). Covariablity arises from the dynamical coupling between neighbors connected in the contact network and can be computed directly from the ensemble.

3) Forward propagation: We integrate the model to time $t+1$ and propagate information forward to the neighbors of patients whose colonization probabilities have been adjusted (Fig. S2D).

We repeat the three steps sequentially at each time until the most recent observation. Details of the above three procedures are reported in the following subsections.

Backward temporal propagation. We use $X_i^t \in \{\mathcal{S}, \mathcal{C}\}$ to represent the state of patient i at time t, where $\mathcal{S}$ and $\mathcal{C}$ are the events of being susceptible and colonized. $X_i^t = \mathcal{S}$ or $\mathcal{C}$ means patient i is susceptible or colonized at time t. Denote the diagnosis records as $\mathcal{D} = \left\{\left(i_k, t_d^{i_k}, R\right) | \, k = 1, \cdots n\right\}$, where n is the number of positive patients, $R \in \{\mathcal{S}, \mathcal{C}\}$ represents the binary test result, and $t_d^{i_1} \leq t_d^{i_2} \leq \cdots \leq t_d^{i_n}$. Here $\left(i_k, t_d^{i_k}, R\right)$ means patient i_k has test result R at time $t_d^{i_k}$. Denote the diagnosis records after time t as $\mathcal{D}_{>t} = \left\{\left(i_k, t_d^{i_k}, R\right) | \, t_d^{i_k} > t\right\}$. For each patient i_k, tested after time t (i.e., $t_d^{i_k} > t$), we aim to compute his/her colonization probability at time t: $P(X_{i_k}^t = \mathcal{C} | \mathcal{D}_{>t})$.

Using Bayes' rule, we have

$$P\left(X_{i_k}^t = \mathcal{C} \middle| \mathcal{D}_{>t}\right) \propto P\left(X_{i_k}^t = \mathcal{C}\right) P\left(\mathcal{D}_{>t} \middle| X_{i_k}^t = \mathcal{C}\right), \quad [S7]$$

$$P\left(X_{i_k}^t = \mathcal{S} \middle| \mathcal{D}_{>t}\right) \propto P\left(X_{i_k}^t = \mathcal{S}\right) P\left(\mathcal{D}_{>t} \middle| X_{i_k}^t = \mathcal{S}\right). \quad [S8]$$

Here $P\left(X_{i_k}^t = \mathcal{C}\right)$ and $P\left(X_{i_k}^t = \mathcal{S}\right)$ are the prior probabilities of patient i_k to be colonized and susceptible at time t, obtained from inference prior to time t for each ensemble member. In order to compute the posterior, we need to calculate the likelihoods $P(\mathcal{D}_{>t} | X_{i_k}^t = \mathcal{C})$ and $P(\mathcal{D}_{>t} | X_{i_k}^t = \mathcal{S})$.

We provide details of the likelihood computation here. Suppose $i_{k'}$ is the first patient diagnosed positive after time t. We re-write $\mathcal{D}_{>t}$ in the following form: $\mathcal{D}_{>t} = \left\{\left(i_{k'}, t_d^{i_{k'}}, R\right), \left(i_{k'+1}, t_d^{i_{k'+1}}, R\right), \cdots, \left(i_n, t_d^{i_n}, R\right)\right\} = \left\{X_{i_{k'}}^{t_d} = R, X_{i_{k'+1}}^{t_d} = R, \cdots, X_{i_n}^{t_d} = R\right\}$. Note here we drop the superscript of t_d for notational simplicity. The likelihood $P\left(\mathcal{D}_{>t} \middle| X_{i_k}^t = \mathcal{C}\right)$ can be presented as a product of multiple conditional probabilities:

$$\begin{aligned} P\left(\mathcal{D}_{>t} \middle| X_{i_k}^t = \mathcal{C}\right) &= P\left(\left(X_{i_{k'}}^{t_d} = R, X_{i_{k'+1}}^{t_d} = R, \cdots, X_{i_n}^{t_d} = R\right) \middle| X_{i_k}^t = \mathcal{C}\right) \\ &= P\left(X_{i_{k'}}^{t_d} = R \middle| X_{i_k}^t = \mathcal{C}\right) P\left(X_{i_{k'+1}}^{t_d} = R \middle| X_{i_{k'}}^{t_d} = R, X_{i_k}^t = \mathcal{C}\right) \times \cdots \\ &\times P\left(X_{i_n}^{t_d} = R \middle| X_{i_{n-1}}^{t_d} = R, \cdots, X_{i_{k'}}^{t_d} = R, X_{i_k}^t = \mathcal{C}\right). \quad [S9] \end{aligned}$$

We compute the likelihood using Eqs. [1-3] in the main text. Specifically, at time t, we set $P\left(X_{i_k}^t = \mathcal{C}\right) = 1$ in the estimated model state for each ensemble member (i.e., set $C_{i_k}^t = 1$ and $S_{i_k}^t = 0$). We then integrate Eqs. [1-3] until the next diagnosis at time t_d for $i_{k'}$, and calculate $P\left(X_{i_{k'}}^{t_d} = R \middle| X_{i_k}^t = \mathcal{C}\right)$, the first term of r.h.s. of Eq. [S9], using the master equations. Note the characteristics of laboratory tests can be incorporated into the likelihood to reflect imperfect observations. Specifically, for tests with 100% sensitivity and 100% specificity, $P\left(X_{i_{k'}}^{t_d} = R \middle| X_{i_k}^t = \mathcal{C}\right)$ can be directedly computed using Eqs. [1-3]. For imperfect tests with a% sensitivity and b% specificity, we multiply $P\left(X_{i_{k'}}^{t_d} = R \middle| X_{i_k}^t = \mathcal{C}\right)$ obtained from Eqs. [1-3] by a% if $R = \mathcal{C}$ and b% if $R = \mathcal{S}$. For this study, we assumed 100% sensitivity and specificity, as information on test accuracy was not available.

At time t_d for $i_{k'}$, we set $P\left(X_{i_{k'}}^{t_d} = R\right) = 1$, integrate Eqs. [1-3] until the next diagnosis time t_d for $i_{k'+1}$, and calculate $P\left(X_{i_{k'+1}}^{t_d} = R \middle| X_{i_{k'}}^{t_d} = R, X_{i_k}^{t} = \mathcal{C}\right)$, the second term of the r.h.s. of Eq. [S9]. We repeat this process until the last diagnosis time t_d for i_n, and calculate $P\left(X_{i_n}^{t_d} = R \middle| X_{i_{n-1}}^{t_d} = R, \cdots, X_{i_{k'}}^{t_d} = R, X_{i_k}^{t} = \mathcal{C}\right)$, the last term of the r.h.s. of Eq. [S9]. The likelihood $P\left(\mathcal{D}_{>t} \middle| X_{i_k}^{t} = \mathcal{C}\right)$ is computed as the product of those conditional probabilities. The likelihood $P(\mathcal{D}_{>t} | X_{i_k}^{t} = \mathcal{S})$ can be computed similarly.

For a small system with a short observation time window, the likelihoods can be computed separately for each ensemble member. However, for a large system with a long observation time window, computing likelihoods for all ensemble members could be computationally expensive. In our implementation, we used the ensemble mean state to compute likelihoods, which are the same across ensemble, but allowed the priors $P\left(X_{i_k}^{t} = \mathcal{C}\right)$ and $P\left(X_{i_k}^{t} = \mathcal{S}\right)$ to vary for each ensemble member. This approximation yields satisfactory performance in both synthetic tests and real-world application to the outbreak in Swedish healthcare facilities.

Finally, we can compute $P(X_{i_k}^{t} = \mathcal{C} | \mathcal{D}_{>t})$ using normalization:

$$P\left(X_{i_k}^{t} = \mathcal{C} \middle| \mathcal{D}_{>t}\right) + P\left(X_{i_k}^{t} = \mathcal{S} \middle| \mathcal{D}_{>t}\right) = 1. \quad [S10]$$

In the above process, we propagate information from observations made after time t (i.e., $\mathcal{D}_{>t}$) backward to estimate the colonization probability of observed carriers at time t.

Covariability adjustment. At time t, we use cross-ensemble variability to update the colonization probability of patients who have contacts with carriers observed after time t. Through the dynamical coupling between two neighbors in the contact network, information on the colonization probability of one patient can inform the colonization probability of his/her close contacts. For instance, if one patient is a confirmed carrier, his/her neighbor in the contact network should have a higher colonization probability. Such dynamical coupling can be quantified using cross-ensemble covariability.

Denote the ensembles of the prior and posterior colonization probability for an observed carrier i as $\{C_i^t\}_{prior}$ and $\{C_i^t\}_{post}$, where $\{C_i^t\}_{post}$ is estimated using backward propagation. For a patient j who is connected to patient i at time t, denote the ensemble of the prior colonization probability as $\left\{C_j^t\right\}_{prior}$. The cross-ensemble covariability is computed as $cov\left(\{C_i^t\}_{prior}, \left\{C_j^t\right\}_{prior}\right)$, the covariance between $\{C_i^t\}_{prior}$ and $\left\{C_j^t\right\}_{prior}$. The posterior colonization probability for patient j is updated through:

$$\left\{C_j^t\right\}_{post}^{\ell} = \left\{C_j^t\right\}_{prior}^{\ell} + \frac{cov\left(\{C_i^t\}_{prior}, \left\{C_j^t\right\}_{prior}\right)\left(\{C_i^t\}_{post}^{\ell} - \{C_i^t\}_{prior}^{\ell}\right)}{var\left(\left\{C_j^t\right\}_{prior}\right)}. \quad [S11]$$

Here $\left\{C_j^t\right\}_{post}^{\ell}$ is the ℓth member of the posterior ensemble $\left\{C_j^t\right\}_{post}$, $\left\{C_j^t\right\}_{prior}^{\ell}$ is the ℓth member of the prior ensemble $\left\{C_j^t\right\}_{prior}$, and $var\left(\left\{C_j^t\right\}_{prior}\right)$ is the variance of the prior ensemble $\left\{C_j^t\right\}_{prior}$. This update scheme is routinely used in the ensemble adjustment Kalman filter (3, 4). The probability to be susceptible is updated by $S_j^t = 1 - C_j^t$.

Forward propagation. We integrate the updated model state from time t to time $t + 1$ using Eqs. [1-3]. Model integration propagates the updated information (i.e., posteriors) forward in time and through the system dynamics transmission to the neighbors of patients whose colonization probabilities have been updated may occur.

The inference algorithm is computationally efficient. The backward temporal propagation—the most computationally intensive component of the algorithm—is performed only for the sparsely observed cases. For the present synthetic tests in a 52-week period, integration of the full model-

inference system can be completed on a regular laptop within two hours. Further, computation of the likelihood, the most computationally intensive step of the inference algorithm, could be paralleled to shorten run times. As a result, the algorithm can be potentially scaled up for application to large-scale contact networks.

We tested a second version of the inference algorithm in which the backward temporal propagation is performed only once upon the first appearance of a patient in hospital (as opposed to adjusting the states of observed cases every time step sequentially). This implementation substantially reduces the computation cost of the algorithm. However, while the inference yielded similar accuracy for synthetic outbreaks, the performance for the real-world data in Fig. 3 degraded compared to the inference with sequential adjustment. There are several advantages using sequential adjustment in real-world applications. First, the likelihoods in the backward temporal propagation can be better estimated when information is propagated to other unobserved nodes; update of the colonization probability of observed cases based on the most recent estimates is therefore more accurate. Second, the agent-based model cannot perfectly represent real-world transmission processes; the sequential adjustment can correct the errors introduced by model misspecification and reduce the accumulation of error over time.

Pseudo-code.

Input: A model M to update colonization probability, ensembles of initial states $\{C_i^1\}$ and $\{S_i^1\}$, observations of MRSA carriers $\mathcal{D}$ during time 1 to T

For $t = 1$ to T

  For each carrier i_k observed after time t

    Update $\{C_{i_k}^t\}$ and $\{S_{i_k}^t\}$ using $\mathcal{D}_{>t}$ and backward temporal propagation

    For each neighbor j of i_k at time t

      Update $\{C_j^t\}$ and $\{S_j^t\}$ using covariability adjustment

    End For

  End For

  Compute $\{C_i^{t+1}\}$ and $\{S_i^{t+1}\}$ by integrating M to time $t+1$

End For

Output: Colonization probability $\{C_i^T\}$ for all patients at time T

Competing methods

Free simulation. We ran free simulations with binary states using the agent-based model to rank the colonization risk of patients in hospital. Specifically, we initiated the system by randomly assigning a patient as a carrier with a probability drawn uniformly from [0, 5%], and ran model simulations using parameters estimated from population-level incidence numbers. Model simulation was initiated 52 weeks prior to the most recent observation. We repeated simulations for 100 times and quantified the colonization risk for each patient as the fraction of simulations in which the patient is colonized. Note that model simulation does not use individual-level diagnostic information.

Length of stay. Previous studies suggest that hospitalization is a risk factor for MRSA colonization (5–8). We used hospitalization records up to 52 weeks prior to the most recent observation to compute the length of stay (days) in hospital for each patient. Individuals with a longer length of stay are suspected of having a higher risk of colonization.

Number of contacts. We computed the total number of contacts (person-day) in hospital using hospitalization records up to 52 weeks prior to the most recent observation. Patients who have more contacts with other individuals are suspected of having a higher risk of colonization.

Contact tracing. We tracked the patients who have direct contact with the observed carriers prior to their diagnosis up to 52 weeks before the most recent observation. The colonization risk is ranked by the number of contacts (person-day) with identified colonized patients. Due to the sparsity of observations of colonization, only a limited number of patients can be tracked and ranked using contact tracing.

A multivariate logistic regression model. We also used a multivariate logistic regression model that incorporates multiple predictors to rank patient colonization risk:

$$\log\frac{c_i}{1-c_i} = a_0 + a_1 LOS_i + a_2 N_i^{contact} + a_3 N_i^{tracing}. \quad [S12]$$

Here c_i is the colonization probability of patient i, LOS_i is the length of stay, $N_i^{contact}$ is the total number of contacts, and $N_i^{tracing}$ is the number of contacts with confirmed colonization. We fit the model using the real-world data from the Swedish healthcare facilities. However, this dataset contained only a small number of positive patient test results and no records of negative test results. To solve this problem, we randomly sampled $k \times n$ patients in hospital, excluding the observed positive patients (n is the number of observed positive patients and k is a multiplier for sample size), and labeled these individuals negative patients. As the majority of patients should be negative, we presume the sampled patients are unlikely to be positive. We tested a range of sample size for the negative patients, $k = 1,2,\ldots,10$, and used a two-fold out-of-sample validation to avoid overfitting. The model with $k = 9$ yielded the best performance in out-of-sample validation. In synthetic testing and experiments using real-world data, we compared this regression model with other approaches.

For the synthetic tests, we used information from the simulated 52 weeks to rank colonization risk for all patients. In application to the real-world outbreak, hospitalization history during the 52 weeks prior to the diagnosis of each carrier was used to inform free simulation, length of stay, number of contacts, contact tracing and regression ranking.

Synthetic tests

We generated three synthetic outbreaks to validate the SILI algorithm. In those three outbreaks, the majority of colonization is attributed to contact transmission (Fig. S3A), community importation (Fig. S5A) and environmental contamination (Fig. S6A), respectively. Model parameters are reported in Table S1. We ran the agent-based model for 52 weeks to initiate the system and continued the simulation for another 52 weeks. Data in the first 52 weeks were discarded to remove any transient dynamics. Synthetic incidence numbers used for inference were obtained from the simulations over the latter 52 weeks.

We also tested inference using both positive and negative testing results. Specifically, we generated a synthetic outbreak. For each patient testing positive, we randomly selected a susceptible individual in hospital and labeled him/her as a patient testing negative. Both positive and negative results were used in the individual-level inference presented in Fig. S7.

Parameter inference

We used iterated filtering to estimate model parameters. To improve system identifiability, several techniques were employed. First, we kept the transmission model parsimonious and introduced

only a small number of unknown parameters. We assumed that a single transmission rate, importation rate, environmental contamination coefficient and environmental decolonization rate were common to all patients and rooms in the network (rather than varying these parameters by patient or room) and used published findings to assign the patient decolonization rate and case detection rate (Table S1). Second, we assigned a wide initial prior range for each parameter in order to better explore parameter space. Third, we ran multiple realizations of parameter inference and found the results remain robust. In synthetic tests, we found the transmission rate $\boldsymbol{\beta}$, importation rate $\boldsymbol{\gamma}$, and environmental contamination coefficient $\boldsymbol{\theta}$ are well identified. The estimated mean environmental decolonization rate $\boldsymbol{D}$ has a small bias; however, as the model dynamics are less sensitive to $\boldsymbol{D}$, this estimation bias does not severely impact the ranking of colonization risk, as demonstrated by the superior performance of the SILI algorithm.

Control experiment

For the *in silica* control experiment, targeted screening can be implemented at any frequency: daily, weekly, monthly, etc., to reflect differences in the prevalence of MRSA and resources presumed available for intervention. In the main text, testing and isolation was enforced every 4 weeks as MRSA has a low prevalence in Swedish healthcare facilities. In an additional experiment, we implemented a weekly targeted screening and isolation. Specifically, every week, after each sliding 52-week window, we selected 1% or 5% of patients present in hospital with the highest colonization risk, ranked by the different approaches. The selected, high-risk individuals were screened and put into isolation in the following week if they remained hospitalized. During isolation, the targeted individuals will neither be colonized nor transmit MRSA to other patients. We track the numbers of observed incidence and all colonized patients during the 320-week period under each control strategy and compare these results to outcomes without any control. The numbers of observed cases and colonized patients were further reduced with this more frequent testing and isolation (Fig. S9); however, the improvement is limited. This is possibly due to a large number of imported cases from the community that cannot be precisely identified by any of the methods considered here. These results indicate that the inference framework can be applied at a variety of time scales; however, the most cost-effective testing frequency and coverage will need to be determined.

SI References

1. V. Grimm, *et al.*, The ODD protocol: A review and first update. *Ecol. Model.* **221**, 2760–2768 (2010).
2. M. Begon, *et al.*, A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiol. Infect.* **129**, 147–153 (2002).
3. J. L. Anderson, An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Wea. Rev.* **129**, 2884–2903 (2001).
4. S. Pei, F. Morone, F. Liljeros, H. Makse, J. L. Shaman, Inference and control of the nosocomial transmission of methicillin-resistant Staphylococcus aureus. *eLife* **7**, e40977 (2018).
5. H. Grundmann, S. Hori, B. Winter, A. Tami, D. J. Austin, Risk Factors for the Transmission of Methicillin-Resistant Staphylococcus aureus in an Adult Intensive Care Unit: Fitting a Model to the Data. *J. Infect. Dis.* **185**, 481–488 (2002).
6. A. I. Hidron, *et al.*, Risk Factors for Colonization with Methicillin-Resistant Staphylococcus aureus (MRSA) in Patients Admitted to an Urban Hospital: Emergence of Community-Associated MRSA Nasal Carriage. *Clin. Infect. Dis.* **41**, 159–166 (2005).
7. S. Harbarth, *et al.*, Evaluating the Probability of Previously Unknown Carriage of MRSA at Hospital Admission. *Am. J. Med.* **119**, 275.e15-275.e23 (2006).
8. E. Girou, G. Pujade, P. Legrand, F. Cizeau, C. Brun-Buisson, Selective Screening of Carriers for Control of Methicillin-Resistant Staphylococcus aureus (MRSA) in High-Risk Hospital Areas with a High Level of Endemic MRSA. *Clin. Infect. Dis.* **27**, 543–550 (1998).

9. B. S. Cooper, *et al.*, Methicillin-resistant Staphylococcus aureus in hospitals and the community: Stealth dynamics and control catastrophes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10223–10228 (2004).
10. M. C. J. Bootsma, O. Diekmann, M. J. M. Bonten, Controlling methicillin-resistant Staphylococcus aureus: Quantifying the effects of interventions and rapid diagnostic testing. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5620–5625 (2006).
11. M. Eveillard, *et al.*, Consideration of age at admission for selective screening to identify methicillin-resistant Staphylococcus aureus carriers to control dissemination in a medical ward. *Am. J. Infect. Control* **34**, 108–113 (2006).
12. X. Wang, S. Panchanathan, G. Chowell, A Data-Driven Mathematical Model of CA-MRSA Transmission among Age Groups: Evaluating the Effect of Control Interventions. *PLOS Comput. Biol.* **9**, e1003328 (2013).
13. C. M. Macal, *et al.*, Modeling the transmission of community-associated methicillin-resistant Staphylococcus aureus: a dynamic agent-based simulation. *J. Transl. Med.* **12**, 124 (2014).
14. A. Jarynowski, F. Liljeros, Contact Networks and the Spread of MRSA in Stockholm Hospitals in *2015 Second European Network Intelligence Conference*, (2015), pp. 150–154.
15. E. Kajita, J. T. Okano, E. N. Bodine, S. P. Layne, S. Blower, Modelling an outbreak of an emerging pathogen. *Nat. Rev. Microbiol.* **5**, 700–709 (2007).

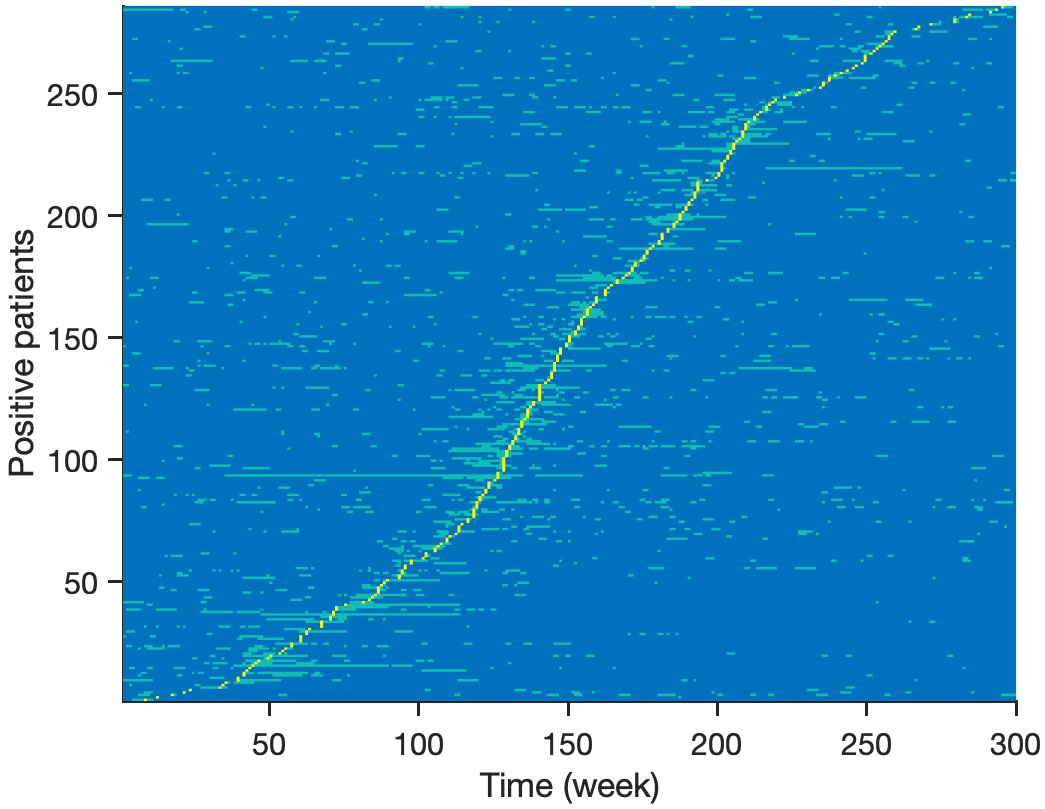

Fig. S1. Hospitalization history of the 289 observed colonized patients. Blue, green and yellow colors represent out-of-hospital, hospitalized and tested positive. Patients are ranked according to their week of confirmation from bottom to top.

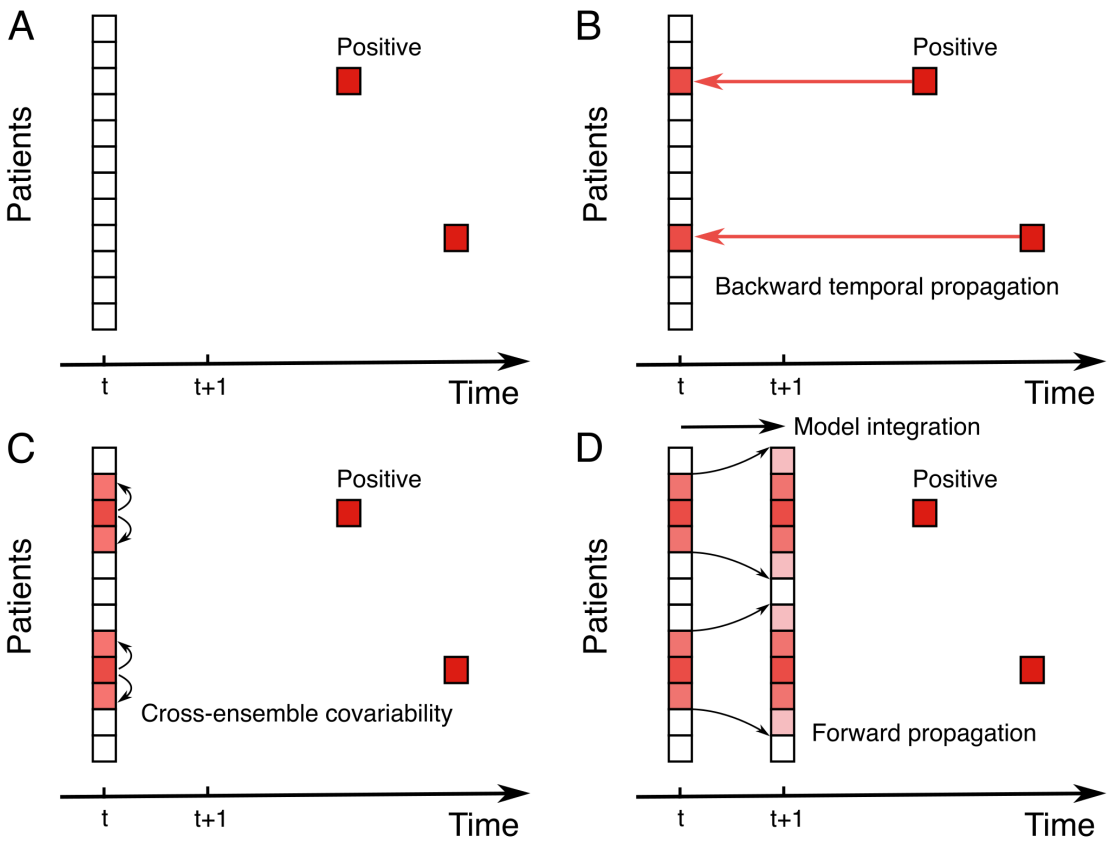

Fig. S2. An illustration of the SILI algorithm. (A) The x-axis represents time, and the y-axis shows patients connected in a one-dimensional chain. Two patients are observed positive at later times. The red color shows the amount of information obtained from observations (here quantified by colonization probability). (B) We use Bayes' rule and model simulation to propagate information backward in time and estimate the colonization probability of observed carriers at time t. (C) We use cross-ensemble covariability to adjust the colonization probability of patients who have contact with observed carriers. (D). We integrate the model to time $t + 1$ and propagate information forward to neighbors.

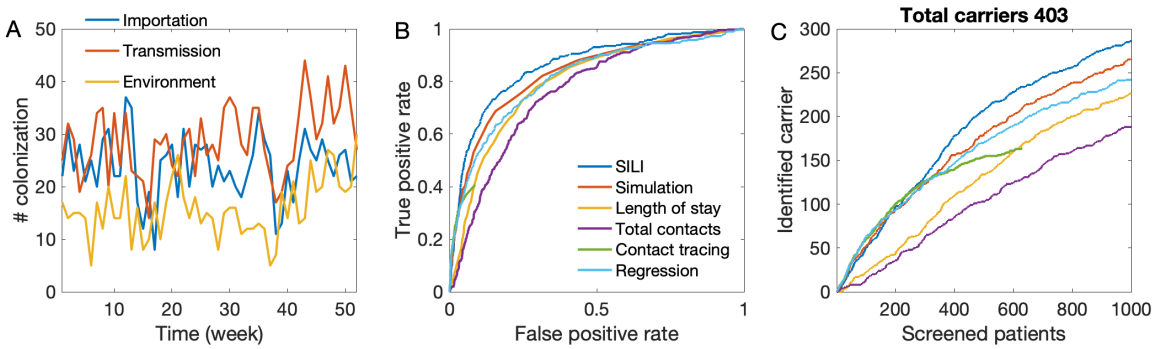

Fig. S3. Synthetic test for an outbreak with the majority of colonization caused by contact transmission (parameter setting: $\beta = 0.028$, $\gamma = 0.005$, $\theta = 0.005$ and $D = 1.5$). (A) The numbers of colonized patients attributed to importation, contact transmission and environmental contamination. (B) The ROC curves for identification of MRSA carriers using different methods at week 52. (C) The number of MRSA carriers identified by screening a given number of patients selected using different approaches at week 52.

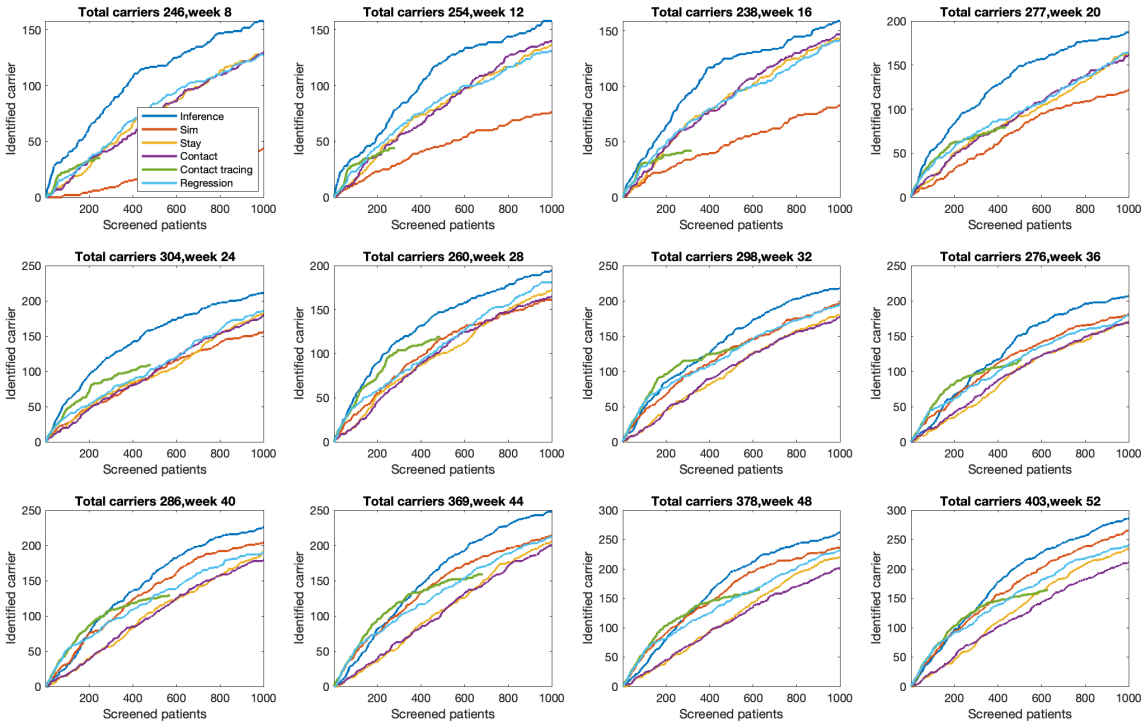

Fig. S4. The number of MRSA carriers identified by screening a given number of patients selected using various approaches at different weeks during the sequential inference. The SILI algorithm consistently outperforms other competing methods.

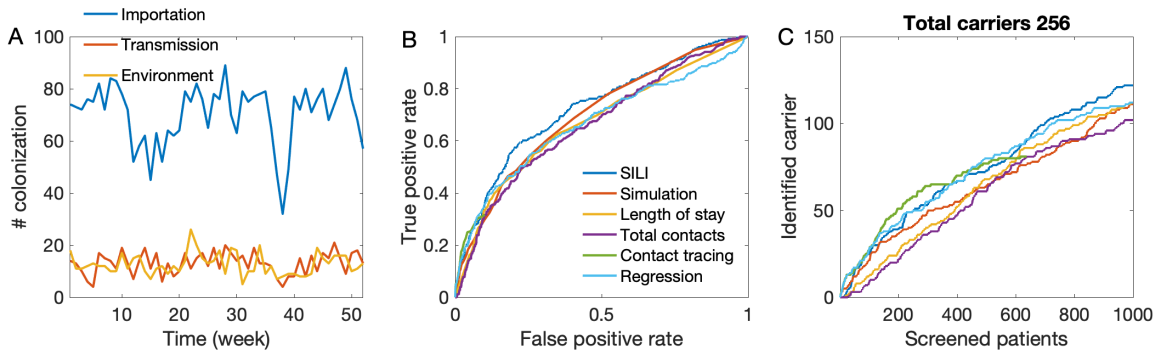

Fig. S5. Synthetic test for an outbreak with the majority of colonization caused by importation from the community (parameter setting $\beta = 0.015$, $\gamma = 0.02$, $\theta = 0.005$ and $D = 1.5$). (A) The numbers of colonized patients attributed to importation, contact transmission and environmental contamination. (B) The ROC curves for identification of MRSA carriers using different methods at week 52. (C) The number of MRSA carriers identified by screening a given number of patients selected using different approaches at week 52.

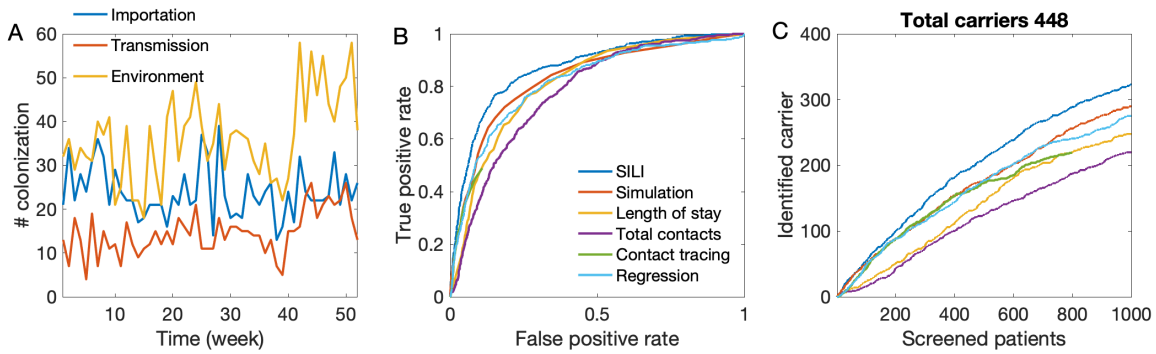

Fig. S6. Synthetic test for an outbreak with the majority of colonization caused by environmental contamination (parameter setting $\beta = 0.015$, $\gamma = 0.005$, $\theta = 0.016$ and $D = 1.5$). (A) The numbers of colonized patients attributed to importation, contact transmission and environmental contamination. (B) The ROC curves for identification of MRSA carriers using different methods at week 52. (C) The number of MRSA carriers identified by screening a given number of patients selected using different approaches at week 52.

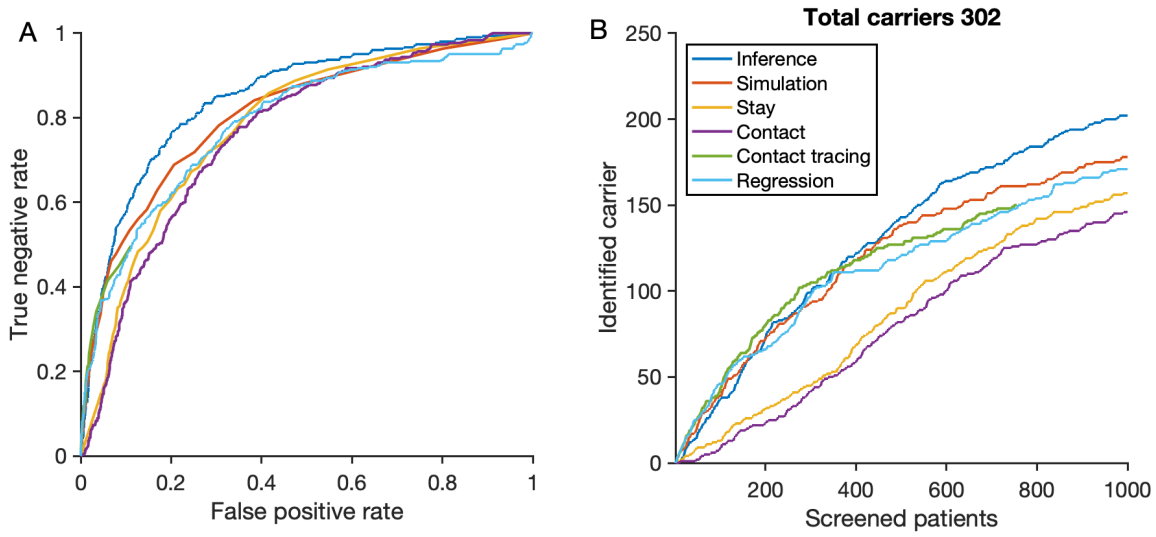

Fig. S7. Synthetic test for an outbreak using both positive and negative test results (parameter setting: $\beta = 0.028$, $\gamma = 0.005$, $\theta = 0.005$ and $D = 1.5$). (A) The ROC curves for identification of MRSA carriers using different methods at week 52. (B) The number of MRSA carriers identified by screening a given number of patients selected using different approaches at week 52.

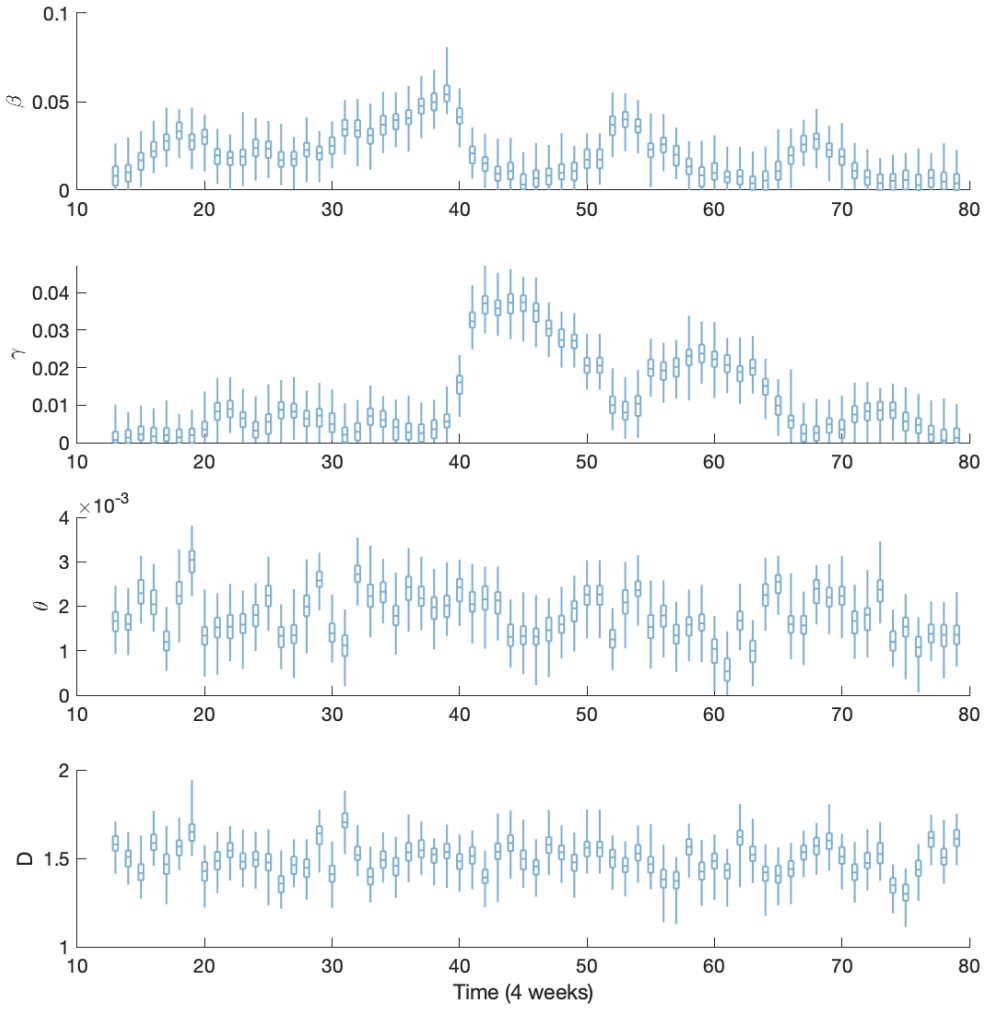

Fig. S8. The estimated posterior parameters for each 4-week time window for the real-world outbreak in Swedish hospitals after assimilating data from the prior 52 weeks. ($\boldsymbol{\beta}$: the baseline transmission rate; $\boldsymbol{\gamma}$: the importation rate; $\boldsymbol{\theta}$: the baseline environmental contamination coefficient; $\boldsymbol{D}$: the mean environmental decolonization rate). Boxes and whiskers show the interquartile and 95% CIs.

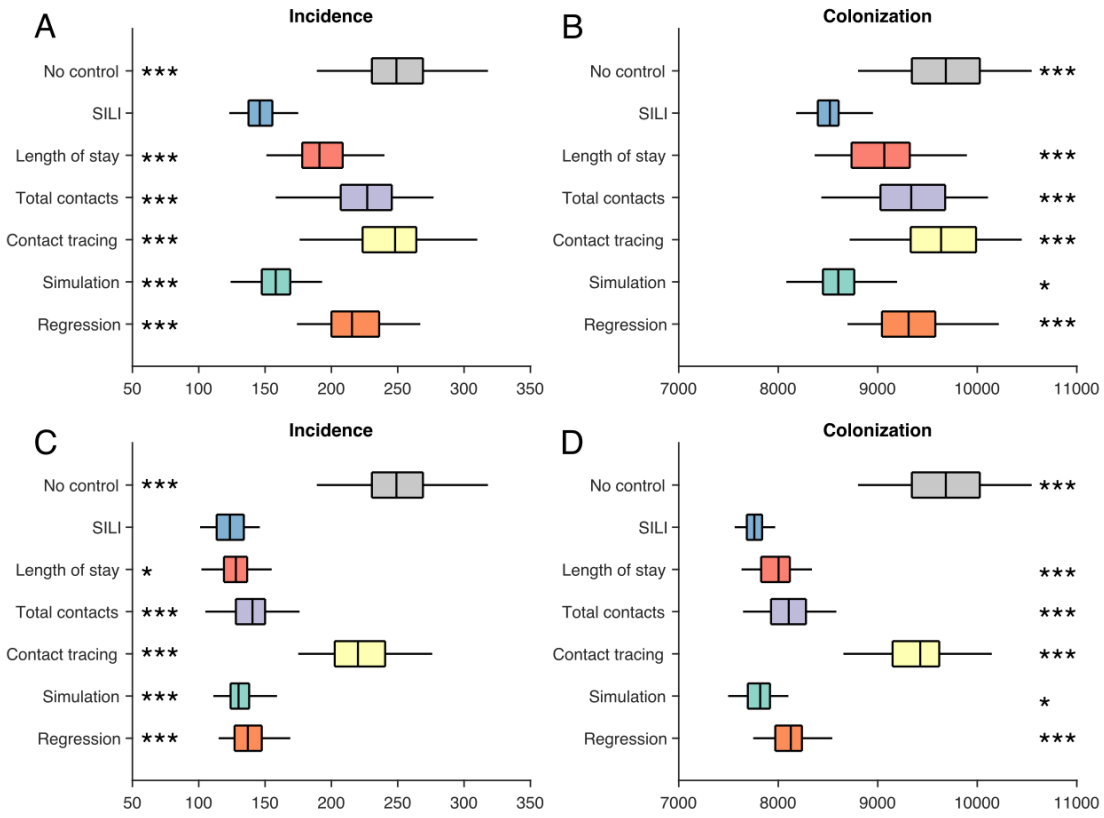

Fig. S9. Retrospective control experiment in 66 Swedish healthcare facilities under weekly control intervention. (A and B) Distributions of the observed incidence and total colonization by isolating 1% patients in hospital selected by different methods every week. Results are obtained from 100 independent control experiments. Asterisks indicate statistical significance that the SILI algorithm outperforms other approaches, obtained from the Mann-Whitney test (*** $p < 10^{-5}$, ** $p < 0.005$, * $p < 0.05$). (C and D) Results for isolating 5% patients in hospital selected by different methods every week.

Parameter	Description	Range/Value	Unit	Reference
α	Patient decolonization rate	[1/365, 1/175]	Per day	(4, 9–14)
β	Baseline transmission rate	0.028; 0.015; 0.015	Per day	Assigned
γ	Importation rate	0.005; 0.02; 0.005	Per admission	Assigned
θ	Baseline environmental contamination coefficient	0.005; 0.005; 0.016	(Per day)^2	Assigned
D	Mean environmental decolonization period	1.5; 1.5; 1.5	Day	Assigned
ρ	Observation rate	[0.15α,0.25α]	Per day	(4, 14, 15)

Table S1. Parameter settings in synthetic tests. Settings for β, γ, θ and D used to generate three synthetic outbreaks are separated by semicolons. For each individual, the patient decolonization rate α is randomly drawn from the pre-defined range; the observation rate ρ is drawn after α is specified.