**Main Manuscript for**

Quantifying likelihoods of detection and nosocomial transmission of pathogenic bacteria in hospital settings using patient records and culture data

Jaime Cascante Vega1, Tal Robin1, Rami Yaari1, Lingsheng Wen2, Jason Zucker2, Anne-Catrin Uhlemann2, Sen Pei1,\*, Jeffrey Shaman1,3,\*

1. Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA.

2. Division of Infectious Diseases, Department of Medicine, Columbia University, College of Physicians and Surgeons, New York, New York, USA.

3. Columbia Climate School, Columbia University, New York, NY, USA.

\* Corresponding authors: Sen Pei, Jeffrey Shaman

**Email:**  sp3449@cumc.columbia.edu, jls106@cumc.columbia.edu

**Author Contributions:** J.C.V., A.C.U., S.P. and J.S. conceived the study and managed the project. J.C.V. performed the analysis. L.W. and J.Z. curated the data, J.C.V. and R.Y. processed and compiled the data, J.C.V., T.R., R.Y., A.C.U., J.Z., S.P. and J.S. analyzed the results. J.C.V drafted the manuscript. All authors revised and reviewed the manuscript.

**Competing Interest Statement:** J.S. and Columbia University disclose partial ownership of SK Analytics. J.S. discloses consulting for BNI. ACU discloses research support from Merck. All other authors declare no competing interests.

**Classification:** Major: Biological Sciences. Minor: Biophysics and Computational Biology.

**Keywords:** epidemiology, microorganisms, antimicrobial resistance, Bayesian inference, individual-based model, agent-based model, inverse problems

**This PDF file includes:**

Main Text

Figures 1 to 6

Tables 1 to 2

**Abstract (244/250 words)**

One of the most important problems in epidemiology is to quantify and understand transmission.

Antimicrobial-resistant organisms (AMROs) are a major threat to public health. These organisms increase mortality, hospital length of stay, and in consequence healthcare-associated costs. AMROs are present in both hospitals and the broader community; however, information on AMRO carriage rates and transmission in both settings is limited. In this work, we leverage electronic health records from a major New York City hospital system collected during 2020-2021 to support simulation-based inference of detection and nosocomial transmission quantities for eight pathogens. We develop an agent-based model to simulate the admission, transfer, and discharge of patients at the ward facility level in the hospital system, AMRO importation from the community, and patient-to-patient transmission of AMROs. The model is coupled with a Bayesian inference data assimilation algorithm to estimate effective sensitivity and nosocomial transmission rates. We evaluate parameter identifiability for this model-inference system and investigate if bias in the inference is caused by the structural identifiability or by the inherent stochasticity of the model. We applied the framework to estimate both quantities for seven prevalent microorganisms species: *Escherichia coli, Klebsiella pneumoniae, Pseudomonas aeruginosa, Staphylococcus aureus* (both sensitive, MSSA, and resistant, MRSA, phenotypes), *Staphylococcus epidermidis*, *Enterococcus faecium* and *Enterococcus faecalis*.

The parameter estimates reveal substantially higher community prevalence (i.e., importation rate) for *E. coli* and *K. pneumoniae* than the other six organisms. Nosocomial transmission rates are found to be highest for *P. aeruginosa* and MSSA. This work highlights how fine-scale patient data can support simulation based inference of epidemiological properties of microorganisms. How the design of a detailed individual-level observational model does improve parameter identifiability, compated to a population level observational model.

Evaluation of the community prevalence and transmission potential for different pathogens could ultimately support the development of in-hospital control measures that limit the spread of these pathogens.

**Main Text**

**Introduction**

Antibiotic resistance microorganisms (ABRO) are a major threat to human health worldwide and has emerged as one of the leading public threats of the 21st century 1. An estimated 4.95 million deaths were associated with bacterial ABR in 2019 globally, and mortality caused by ABR is projected to reach 10 million by 2050 2, although more rigorous projections XXX. Hospital-acquired (nosocomial) infections by bacterial pathogens including those resistant to antibiotics are a major contributor to mortality, length of stay in hospital and health-care associated costs 2. Understanding the burden and spread of ABR pathogens, and microorganisms in general, within hospital settings, is critical for effective control planning and design of testing/culture protocols; however, quantification of these characteristics remains challenging due to limited observation of microorganisms carriage, difficulty assessing interventions in real-world hospital settings, and incomplete understanding of the mechanisms shaping the coexistence (both within and between hosts) of different microbial species as well as same species antibiotic-resistant and antibiotic-sensitive microorganism in hospitals 3,4. It is well-recognized and assumed in modeling studies that between-host transmission is an important driver of the maintenance of resistance in hospitals and of microorganism circulation 5, among other routes of transmission, such as environmental or fomite transmission (cite), or in the case of resistant strains within the host processes, such as horizontal gene transfer also referred as endogenous transmission 6. These mechanisms are thought to be differential per microorganisms species, across body sites as suggested by Niche theory 1–3 and heterogeneous across geographical space, although some consistency in human prevalence is expected. For example, methicillin-resistant *Staphylococcus aureus* (MRSA) is known to have a unique geographical distribution classified principally by the complex clone(s) circulating 7,8, with human prevalences from 0.4% to 5%, but the general prevalence of *S. aureus* is consistently between 20% and 50%, depending on the body sites and populations considered 9–15.

(discussion: how valid is this assumption - across all microorganisms species? Perhaps I could cite those hypothetical ecological mechanisms that determine population density?)

To circumvent these difficulties, mathematical models have been applied to theoretically study pathogen transmission in hospital settings, especially to quantify and understand the relative roles of different routes of transmission 4 or to characterize the condition of the hospital settings to sustain transmission of both resistant and sensitive strains 16. In the context of ABR, theory has been used to understand the emergence of resistance and its interplay with community-acquired infections 17, to evaluate antibiotic treatment protocols 6, to assess control measures to reduce nosocomial transmission 16,18–21, and has been extended recently in pair with empirical observations to assess the role of competition of different strains at between-host level and the role of within-host microbiome pathogen interactions 22.

However, most existing modeling studies focus on general theoretical frameworks of ABR organisms 6,16,17,22–24 where empirical observations presumably drive the modeling but interrogation with empirical data lacks. Despite the recognized fact that process-based models must be questioned with empirical data before being used as reliable tools for public health 5 few ABRO modeling studies constrained models with data 4,18,21. Furthermore, investigations using simulation-based inference 25 tools usually rely on a single ABROs of interest such as (MRSA) 7,18,21,26, *P. aeuruginosa* 4*, C. difficil* 27*, Vancomycin-resistant enterococci* (VRE)28and used carriage detection 18,21,26,28, intensive care unit (ICU) clinical culture data 4,26 or incorporated sequencing data 27 to constrain the modeling exercises. Some of these studies are absent of interrogation about the structural identifiability of the system 29, some probably a consequence of designing individual-based models where the state space dimension is high and model-inference methods hard to interrogate within a mathematical framework. For example, 4 infers whether nosocomial transmissions were via host-to-host or an environmental reservoir for *P. aeuruginosa* using ICU admissions.We hypothesized that as patient-to-patient and environmental transmission apport linearly to nosocomial transmission and act in the same time-scales there might be multiple solutions that cannot be stressed out without observing additional variables of the process. However, the study used nosocomial infection data from two ICUs of 9 and 6 years respectively and possibly the length of the time series and interventions in the hospital enriched dynamics and improved the parameter estimation. We suggest that investigations relying on simulation-based inference tools need to start including in their cookbook systematical assessment of the identifiability of the system, especially when relying on stochastic transmission models.

Studies on the epidemiological characteristics of multiple co-circulating organisms in a single hospital network supported by real-world data are absent. The epidemiology and transmission dynamics of healthcare-associated infections (HAI) differ from community-acquired infections in a number of important ways. First, hospital networks are open systems with significant variations in the daily and weekly admission rate ranging from 23% to 52% (See Figure 1A), explained mostly by variations at the building scale in the hospital network with some ranging from 0-50% (See building traffic for Allen, Harkness Pavillion and Milstein hospitals in Supplementary Information Figure S1), and others with faster patient replacement between 80% to 150%, consequence of admitting mostly outpatients (See Presbyterian hospital and Rest in SI Figure S1). This heterogeneity at the building scale is in turn explained by variation across the ward traffic composing each building (See SI Figure S2 ward traffic). Discharges or outflux of patients at ward facility scale follows similar patterns with some few wards per building admitting the majority of patients per week and with stable patient traffic (See Figure 1A and SI Figure S2). This heterogeneity in admissions and hospitalizations is principally dictated by ward size that was also variable at the building scale (See SI Figure S3 for ward size distribution and SI Figure S4 for the linear relationship). In contrast, communities are closed systems with individuals moving in or out at much lower rates. Second, as many bacterial species exist as commensals on the human hosts' detection of colonization is difficult and differential by body site. Indeed, patients with infections of the bloodstream or lower respiratory tract are more likely to be detected in hospitals than those colonized at other sites 17,30. Third, facilities within a single hospital system (e.g., infusion, pediatric, emergency wards, surgical, among others) may differ substantially in their control and detection of microorganisms as well as in the hospital traffic features, a complication compounded by patient transfers between wards and hospital buildings (see SI Figure S5) 31,32. Fourth, a lack of observational data for communities of microorganisms spreading among patients and the environment in the same hospital system often imposes further challenges in comparing the epidemiological features of co-circulating organisms.

In this study, we use a process agent-based model (ABM) and detailed observational data for eight prevalent pathogenic microbial species collected in a major New York City (NYC) hospital system to address these challenges. The ABM is informed by real-world patient movement in the hospital system and incorporates importation of microorganisms from the community, nosocomial transmission, patient transfer across hospital wards and decolonization via host clearance. To account for the heterogeneity of testing frequency and microorganism prevalence among facilities, we used clinical testing records to design a patient-level observational model that mimics the detection of microorganisms in hospitals.

We searched the literature to set the importation rate of each microorganism as a proxy of the community prevalence. Our search terms included 'prevalence', 'colonization', 'carriage rate', and used reviews for some of the organisms that reported pooled estimates across different geographical locations. In the Supplementary Information section **Prevalence estimates** we included the different values, sources, and a small description of the study, with the geographical location and population of interest; in Table 1 we consigned a resume with the values. We couple the ABM with a Bayesian inference algorithm, validate the system against simulated outbreaks and estimate the likelihood of detection given carriage and nosocomial transmission rates for the eight microorganisms, which cause substantial mortality associated with AMR worldwide 2. We explored whether bias in the inferences was inherent to structural identifiability 29 or caused by Monte Carlo error 33. We used 3 different values for the importation rates to both account for the uncertainty in prevalence and to study the sensitivity of the model-inference construct to this parameter.

As suggested in (cite) host-to-host transmission is thougth to be low

**Results**

### *Empirical patterns and heterogeneity of microorganism burden*

In Figure 1A, we plot weekly incidence color coded by building for *E. coli* (total positives cultures, n=), *K. pneumoniae* (n=), *P. aeruginosa* (n=), MSSA (n=), MRSA (n=), *S. epidermidis* (n=), *E. faecalis* (n=) and *E. faecium (n=)* . We deduplicated multiple positives results during a patient visit, and consistently deduplicated multiple tests in the observational model. During the study period, the daily number of hospitalized patients fluctuated between 1,000 and 2,500 with most of the variation explained by differences in the day of the week (See solid line for daily and dashed line for weekly in Figure 1B). Daily numbers of new admissions ranged between 100 and 1,000, including outpatients. During the first COVID-19 wave in New York City, the numbers of in-patients and admissions were generally lower; after June 2020, patient traffic was higher and relatively stationary. Ward size, defined as the average occupancy per day during the study period, was heterogeneous (SI Figure. S4) with the majority of wards experiencing an occupancy below 10 patients. However, a few wards (e.g., emergency rooms) could admit over 100 patients each day. To visualize patient traffic within each ward, we investigated temporal occupancy. Fig 1C shows the weekly number of patients relative to average occupancy for each ward and highlights the 5 most and least populated wards (see red and blue lines respectively); we only included wards with sizes greater than one for better visualization. The least populated wards were empty most of the time with irregular occupancy during the study period, whereas hospitalizations in the most populated wards were relatively stable. The relationship between newly admitted and in-hospital patients (Figures 1D and 1E) was linear and depended on ward size (see SI Figure S3). Clinical culture numbers were also heterogeneous across wards. Most cultures were sampled from a small subset of wards, as shown by the weekly number of cultures collected in each (see Fig. 1D).

The process model, as is an ABM, tracks the state of the patients at daily time scales. Patients were either susceptible to colonization or carriers with a microorganism (See *The transmission model* in Methods). The observational model designed at the individual level allows us to map from the carriers to detected individuals via a simulated culture. We parametrize the patient observational model with a likelihood of detection given carriage upon testing - (See *The individual observational model* inMethods). Data assimilation is conducted at the building level therefore we consistently aggregate simulated patient detections at this scale. Weekly and monthly incidences of positive cultures for each building are shown in Fig 1A, faded and strong solid lines respectively. We subsequently used these six building time series to perform parameter estimation.

### *Identifiability: synthetic simulations and inference*

To investigate the identifiability of the model-inference system, we explored inference on a simulated trajectory with known parameters. We first investigated in the model inference is capable of recovering parameters fixing the importation rate to 25% and 50%, and varying the likelihood of and the nosocomial transmission rate uniformly covering the prior range, but without forcing the simulated carriage detection to match the observed one (See SI section *Model inference framework)*. The posterior parameter estimates of the inference consistently capture the true parameter values for both detection rates and nosocomial transmission rates (See Figure 2A and 2B for equal to 25 and 50% respectively). Figure 2A highlights two bits of information, first how the model inference is able to explore different regions of the prior range (indicated in the limits of each axis), and second how the posterior captures the true parameter values. Interestingly inferences with are more biased and less sharp than inferences with . An important question to ask is whether the inverse inference problem is well posed; i.e. if perturbations to the observations change the answer of the inverse problem (See Chapter 7 of 34). Inference is conducted on a single stochastic trajectory, we investigated if the stochasticity has a substantial effect on the posterior inference (also referred to as Monte Carlo error). Guided by numerical results SI Figure S5A shows how the goodness-of-fit of the posterior can be linearly predicted by how probable the observation used to conduct inference was (see Methods). This was also a function of with showing a poorer goodness-of-fit than for a more common inferred time series but a less pronounced relationship with the probability of observing the time-series used for inference (slopes of purple line and red line respectively in SI Figure S5A). When controlled by the Monte Carlo error (see Methods) we found that the goodness-of-fit decreased exponentially with the proba, and it's not impacted by (See SI Figure S5B).

We ran ensemble simulations with the posterior parameter estimate and compared them to the ensemble simulations with the true parameters, highlighting the observation used to conduct inference (See SI Figure S6A and S6B, for respectively) and found that qualitatively even biased inferences reproduced well both the true ensemble simulations and the observation used to conduct inference. We quantified the calibration displaying the reliability plot of both the posterior and truth ensemble simulation with the inferred time-series (See SI Figure S7A and S7B, for respectively) and finish this

Last approach allowed us to understand the ability of the model inference to solve the inverse problem in the prior parameter range. We also study the ability of the method to solve the inverse problem in combination of parameters that reproduce the microorganism's nosocomial data at the hospital level (aggregation across buildings of time series shown in Figure 1A). We used the middle prevalence of each bacterial microorganism reported in Table 1, see Methods for further details on the 9 scenarios selected for each pathogenic bacteria. SI Figure S8 shows the posterior estimates for each microorganism, note that the value of the 9 scenarios of each pathogen are very similar. We found the method is able to discriminate between the scenarios. This analysis covers a broad range of prevalences from 'low' to 'high' importation rates; 4% and 16% for MRSA and *K. pneumoniae* to 50% and 70% for *E. faecium* and *E. coli* respectively.

### *Inference using real data*

We applied the model-inference system to estimate the epidemiological quantities of interest for eight different microbial species. Posterior estimates of the importation rate and nosocomial transmission rate are compared in Figure 4A. Species are sorted from the most abundant (*E. coli)* to the least (*E. faecalis).* We found *E. coli* importation was the highest with an estimated 64.18% of patients entering the hospital system colonized. This high importation likely reflects high community prevalence, which is in line with empirical findings 35. *E. coli* was followed by *K. pneumoniae* with an importation rate of 25.61%, and MSSA at 14.48% (Figure 4A). We found similar importation rates for MRSA, *C. albicans* and *S. epidermidis* of around 10.5%. *P. aeruginosa* and *E. faecalis* had slightly lower estimated importation rates of 9.9%. While the estimates for importation rates across organisms match the frequency of detection across the hospital network, nosocomial transmission rates differ. We estimated *P. aeruginosa* had the highest nosocomial transmission rate, followed by *C. albicans* and MSSA. We found nosocomial transmission rates for *E. coli* and *E. faecalis* were the lowest with rates less than 0.005 per capita per day. Table 2 shows all the posterior estimates for importation and nosocomial contact rates.

To assess the goodness-of-fit of the model to the data, we simulated the dynamics using the posterior estimates of parameters and found that simulated detected positive cases span the observed numbers at the hospital and ward cluster level (Figure 4B-C right column and Figure S6). We also produced reliability plots to examine whether the uncertainty in the observed data can be reproduced by model simulations. Specifically, we computed the percentage of observations falling within a given credible interval (CI) of the quantity obtained from model simulations. For a perfectly calibrated simulation, X% of observed values should fall within the X% CI generated by model simulations, producing a diagonal line (y=x) in the reliability plot. We found that the coverage of simulated CIs is slightly below expected values (under the diagonal line), indicating an uncertainty level biased low (Fig. 5D), possibly due to stochastic processes not represented in the model. The model simulations allow estimation of the relative contributions of importation and nosocomial transmission to the overall burden of different species. We found that importation plays a more important role than nosocomial transmission for all species examined (Fig. 5).

To assess the robustness of the parameters estimates to specimens collected from specific body sites, we repeated the previous analyses using subsets of the data that excluded observations from one of the 5 body sites (See Methods sections). We found, excepting one instance, that the estimates remain similar. The exception corresponds to the instance in which the body site with the most observed positive cultures of each species was excluded (See Figures S7-S8). For example, the estimates of both importation and nosocomial transmission rates changed when urine observations were dropped for *E. coli, C. albicans, K. pneumoniae*, and *E. faecalis*, leaving only a small fraction of observations for use in the inference. Finally, we found the parameter estimates were unchanged when excluding outpatient wards (See Figures S9-S10).

**Discussion**

The epidemiological properties of microorganisms present in hospitals are difficult to quantify. Estimating these epidemiological properties can distinguish the roles of different mechanisms of transmission, such as community importation and nosocomial transmission, and help support improved control in healthcare systems. In this paper, we used an agent-based model informed by patient hospitalization records and culture testing results from a large hospital system in New York City to study eight co-circulating microbial species. We coupled this model with an inference algorithm to estimate importation and nosocomial transmission rates for these organisms and found substantial variation in community prevalence and within-hospital transmission.

Our findings echo several previous observations about the transmission of different microorganisms. The difference in nosocomial transmission rates among resistant and sensitive strains of *S. aureus* is potentially explained by the relative fitness of the two strains. It has been reported that resistance generates a decrease in fitness with respect to the sensitive strain, making it less transmissible 1,16,23 consistent with our findings here. Our model-inference system estimated *P. aeruginosa* to be the most transmissible microbial species, consistent with reports that classify this organism as an opportunistic pathogen able to easily colonize non-healthy patients more likely to experience lengthy stays in hospital 4,36,37. We also found that while *E. coli* has the highest importation rate, it has the lowest nosocomial transmission rate suggesting that most carriage can be associated with infections of host commensal strains that do not contribute to transmission, which has been reported empirically 38. These results remained unchanged in response to the underlying observational network across body sites and to exclusion of outpatient wards (See Figures S7-S10).

Here, the goal of using detailed mathematical models to infer transmission properties is to provide quantitative evidence disentangling the force of transmission, which might be used to inform interventions in hospital networks. We show that importation and nosocomial transmission play different roles across organisms. This information itself could potentially help with the design of screening strategies in hospitals. Species with importation considerably greater than nosocomial transmission (e.g. *E. coli*) could be actively screened upon admission to reduce subsequent transmission within the hospital. On the other hand, further research to understand the sources of nosocomial transmission should be undertaken. Our model considers a ward-level force of infection that might be due to not only patient-to-patient transmission but also transmission vectored by healthcare workers or environmental contamination 4,39. It might be possible to test interventions in the first days of admission that potentially disrupt transmission chains (e.g. improved sanitation, vaccines) for pathogens with high nosocomial transmission rates (e.g. *P. aeuruginosa*) 4,36,37. Finally, our model-inference sensitivity analysis, in which we dropped observations from specific body sites, showed that, for quantifying importation and nosocomial rates, it might be enough to consider only a portion of body sites. The selection of body sites could be investigated more rigorously to prevent possible biases in detection and further inform differential testing strategies across body sites.

Mathematical models of infectious diseases are simplifications of real-world transmission dynamics. The ABM used here is not an exception. While the ABM captures multiple sources of heterogeneity (e.g., patient length of stay, contact patterns, individual observational model, etc.) that otherwise cannot be represented by compartmental models, it has several limitations. First, we used an abstract nosocomial transmission rate to represent the collective effects of several processes that may contribute to the dissemination of microbial species, including direct patient-to-patient contact, healthcare worker-mediated contact, environment-mediated contact, etc. Other mechanisms potentially occurring within host (such as horizontal gene transfer) were also not represented in the current model form. The inferred transmission rate therefore cannot distinguish the relative contributions of these processes.

Second, we treated species with resistant and sensitive phenotypes separately, as we were limited by the classification available from the culture data used to detect carriage; however, it has been suggested that resistant and sensitive strains could compete (although it is possible that for some species this is not true). Further research is needed to test this competition hypothesis for species for which data from both resistant and sensitive strains are available. In addition, bacterial species do not simply compete with their own species. As many species are commensals with the human host, a variety of ecological interactions may occur. It may be possible to extend the proposed model to study these interactions. Other limitations of the current framework include the absence of information on demographics and age profiles of patients admitted to the hospital network, which we therefore do not model. Finally, we estimated the effective sensitivity of testing using empirical data on community prevalence for *S. aureus* and validated these estimates with *E. coli* community prevalence (see Methods); however, this approach assumes a 100% specificity and that the effective sensitivity of the clinical culture does not vary across hospital or microbial species. Further clinical and empirical information on these issues should be used to better inform the observational model in the future.

The implications of these findings for disease control are linked to the spatial resolution of the analysis. Previous research has highlighted the important contribution of community prevalence to hospital prevalence, a direct effect of host population exchange between healthcare systems and the community at various rates 44–47. However, as shown in Figures 1 D-E and Figure S3, patient admissions are heterogeneous and have linear relationships with in-hospital patients that are dictated by ward size, and ward category (Figure S2); therefore, community importation should be assessed differently across wards. The use of antibiotics and different schemes for using them (e.g. cycling, mixing at ward level, alternating antibiotics) have been tested using mathematical models and found less effective in model simulations 17. However, those conclusions were mostly drawn from research performed using compartmental models with assumed rates of movement within hospital networks and at the hospital level. Analyses using data-driven agent-based models, such as the one developed here, should also be used to study different antibiotic usage schema and the introduction of new antibiotics. These analyses might be performed at the ward level to study how different scenarios of interventions affect prevalence at both ward and hospital levels.Further, the current work exploits the availability of individual-level patient records to estimate the transmission properties of pathogenic microorganisms. Individual-level data and models can be employed in future research to understand the impact of individual-level interventions on disease control. An example might be to test counterfactual scenarios of individual-level interventions that replace infections caused by resistant strains with infections caused by sensitive strains in order to quantify the impact of resistance at the hospital level 1.

**Materials and Methods**

### *Overview*

To estimate key epidemiological characteristics of AMROs, we developed an ABM to simulate the dynamics of these organisms in hospital settings. The model was informed by patient hospitalization and culture data from electronic healthcare records collected between February 1 2020 and February 28 2021. We coupled the ABM with a Bayesian inference algorithm and, using simulated outbreaks, validated the ability of this ABM-inference system to identify importation and nosocomial transmission rates. We then assimilated the real-world lab-confirmed positive case data and estimated the importation and nosocomial transmission rates for eight co-circulating organisms. Using these estimated parameters and the ABM, we were able to reproduce the time series of positive cases for six clusters of wards in the hospital system. The estimated importation rates and nosocomial transmission rates were compared for the eight organisms.

### *Data*

Data for this study derives from 3 hospitals of a Northern Manhattan Hospital System, including a quaternary care center, pediatric hospital, and community hospital. The hospitals contain 224 wards of different types including emergency, infusion, cardiology, pediatrics, etc. Hospitalization and microorganism testing data were collected during the study period from February 1 2020 to February 28 2021. The hospitalization data include admission, discharge, and transfer of patients within the hospital system (see Figure 6 for schematic). The dates and wards in which each patient stayed during hospitalization were used to construct a time-evolving contact network. Laboratory test results were available for the eight most prevalent organisms in the hospital system: *Eschericia coli*, *K. pneumoniae* *Pseudomonas aeruginosa*, methicillin-susceptible *S. aureus* (MSSA), *Candida albicans*, MRSA, *Staphylococcus epidermidis* and *Enterococcus faecalis*.The date and patient associated with each test were used to inform the observation model (see Methods below). We plotted the number of tests across the 3 hospitals (Figure S11) and in the wards in each hospital (Figure S12), which show a similar order in cultures across the hospital but substantial heterogeneity inside the hospital at the ward scale.

### *The transmission model*

We used an ABM to simulate transmission in the study hospital system 18. For the ABM, the patient-to-patient daily contact networks were constructed using hospitalization records. Two patients staying in the same ward on the same day were connected in the contact network as they are expected to have close contact and care from the same hospital workers who might facilitate transmission of microbial species. Due to patient movement (admission, discharge, and transfer), the contact network is time-varying and was updated daily. We assumed each patient is either susceptible (S) or colonized (C). Patients in contact on a given day (those who shared the same ward) can be and converted to carriers proportional to the number of patients in the ward carrying a particular microorganism. We model the force of infection in ward , , following the law of mass action, as described below. We assumed colonized patients can spontaneously decolonize and become susceptible after a decolonization period, sampled from a uniform distribution days 20. Colonization in hospital can be attributed to two mechanisms: importation from the community and nosocomial transmission. Specifically, we defined these two processes as follows:

1. **Importation from the community:** Using hospital admission records, we assumed a newly admitted patient is colonized with a particular pathogen with an importation probability . The number of patients admitted from the community is shown in Figure 2B. We treated all admissions independently and did not track changes in importation rates due to re-admission. However, if patients were admitted in consecutive days we assumed their colonization status did not change. As a consequence, the number of admitted, colonized patients on day , , among all admitted patients on day , , can be computed as:
2. **Nosocomial contact transmission:** We defined a force of infection for each ward as . The force of infection is proportional to the number of individuals carrying the pathogen in a given ward on day , denoted by . We defined a frequency-dependent transmission rate per ward as , where is the average ward daily occupancy (Density distribution shown in Supplementary Information Figure S1). The force of infection per ward is computed as:

Thus, the force of infection per ward, or probability of colonization for susceptible individuals staying in ward at time , is given by , and the transition equations governing the change of state for a patient, , residing in ward during time are given by:

where is the decolonization period of patient and is a random variable with uniform distribution as indicated previously.

### *The individual observational model*

Colonized patients may develop clinical infections due to the invasion of typically sterile body sites such as blood. At other body sites (i.e. respiratory and urinary tract) the presence of pathogens may indicate both infection or colonization, depending on additional clinical variables1,16. Other individuals carrying pathogens may be discovered through routine screening or cultures ordered discretionarily by clinicians for patients without clinical manifestations. In the ABM, we did not explicitly distinguish between colonized and clinically infected patients. Instead, we applied an observational model to detect carriers. This observational model represents the detection probability for clinical cultures taken from individuals in the hospital network during the study period; it captures the heterogeneous observation of microorganisms across wards and body sites. Weekly number of cultures across hospital wards and weekly number of admitted patients have a strong positive correlation (cultures are proportional to admissions) (Figure 1D-E). Ward size distribution is shown in Supplementary Information Figure S1, and we further stratified by ward type, based on the procedure conducted in each ward (Figure S2). The proportional relationship between the number of admissions, in-hospital days, and ward size is shown in Figure S3.

A key challenge in colonization detection is that cultures collected from one body site may miss colonization of other sites even though the patient carries the bacterium or fungus. For instance, blood cultures of a patient colonized by *E. coli* in the urinary tract will likely test negative. To define our observational model, we therefore estimated the ‘effective sensitivity’ of detecting colonization, i.e. the probability of identifying carriage given that the patient is carrying a particular microorganism. This effective sensitivity not only represents culture test sensitivity, in the strict sense, but also the likelihood a specimen will be taken from a colonized site (and capture bacteria) given colonization or infection exists anywhere on a patient.

To estimate this effective sensitivity, denoted by , we used the empirical estimate of *S. aureus* prevalence (30%) in the community of northern Manhattan, site of the study hospital network 9,10, as a proxy for the community importation rate. This empirical estimate indicates that 30% of new admissions from the community carry *S. aureus* (both sensitive and resistant phenotypes). We then performed a grid search for the nosocomial transmission rate and the effective sensitivity that best fit the observed *S. aureus* carriage in the hospital system (see Supplementary Information Figure S13 - Estimating culture sensitivity for further details). We used the resulting estimate, , as the observational model across all other microbial species. For instance, using , we estimated the community prevalence of *E. coli* is around 70% (see Results section), which generally agrees with the prevalence of *E. coli* in the human population and provides cross-validation of the observational model 35. We also performed additional sensitivity analysis using alternative values of , and the findings remain qualitatively similar.

The defined observational model indicates the likelihood that a colonized patient is detected given the effective sensitivity is . Assuming 100% specificity for the cultures (i.e., no false positives), the number of false negatives is , where TP is the number of true positives (observed positives). Combining the observational model and the observed positives, we can estimate the total positives in hospital, , by summing the observed positives and false negatives, . This estimate of total positives adjusts for the under-detection of microorganism carriage and was used for parameter inference with the ABM.

### *Hospital-level Observational model*

Due to substantial testing heterogeneity, observations of positive cultures were dominated by a few wards with a disproportionately large number of tests. To facilitate use of more granular observations at the sub-hospital level while avoiding excessive noise, we partitioned all wards (n=224) into six clusters and used the aggregated number of positive tests in those clusters to perform inference. The clusters were defined based on patient movement across wards within the hospital system and were identified using a network community detection algorithm 40. The number of transfers between wards is shown in Figure 2A. Wards within each identified cluster have more frequent within-cluster transfers than cross-cluster transfers.

### *Inference*

Inference for dynamic and latent variables and parameters is often treated as a filtering problem, in which the state space is sequentially estimated as observations become available 41,42. This framework has been applied to many infectious diseases, including influenza, dengue, malaria, cholera, Ebola, enterovirus D68, and COVID-19 43–52, using compartmental models that represent transmission dynamics with a set of ordinary differential equations (ODEs). However, there are fewer instances in which filters have been used in conjunction with ABMs, possibly due to the considerably higher dimension of the state space for these models 17. For this ABM system, we simplify the inference problem and only estimate two parameters, i.e., the importation rate and the nosocomial contact rate .

We use the ensemble adjustment Kalman filter (EAKF41 or EnKF), which assumes both the prior and the observations are normally distributed. Compared to other data assimilation techniques, the EAKF is amenable for use with high-dimensional models such as numerical weather models. We assimilated weekly observations of total numbers carriage detections for microorganisms 41 using an iterated filtering framework (IF). Iterated filtering for dynamical systems was proposed in the context of epidemiology using particle filters or sequential Monte Carlo 53,54 (we based our implementation on those algorithms), but have also been proposed in the context of inverse problems using Kalman filters (cite). We substitute the usual ODE model form with the ABM. The implementation of the IF-EAKF in Python was made using the packages NumPy and SciPy 55,56. Further details and hyperparameters for this implementation are available in the Supplementary Information.

### *Identifiability: Synthetic simulations and inference*

To verify that the model-inference system is able to accurately estimate the two key epidemiological parameters, we first investigated the identifiability of the system using simulated nosocomial detections (i.e., the ability of the framework to infer the parameters when they are known). We generated synthetic observations of incident microorganism colonization using the ABM and assigned importation and nosocomial transmission rates. The synthetic observations were then assimilated into the full model-inference system to assess system ability to accurately estimate the parameters. In total, ten synthetic time series of observations were generated with parameters spanning a broad range of values (see Table 1 for the parameter combinations). We chose these scenarios such that different combinations of importation and nosocomial transmission rates could be tested.

### *Inference using real-world data*

We next used actual weekly observations for the six ward clusters to infer community importation and nosocomial transmission rates for all eight microbial species. The model-inference framework was applied to each species separately with the assumption that no ecological interaction exists among species. The primary observational model uses culture data from all body sites; however, we also performed sensitivity analyses in which we restricted the observation of positives to particular body sites. Specifically, we estimated parameters using positive observations excluding one body site at a time. A detailed description of the number of specimens used and the number of cultures across ward clusters is shown in the Supplementary Information. We also performed sensitivity analysis on the Observational Error Variance (OEV) used in the IF-EAKF to investigate its impact on inferred parameter outcomes (See Figure S14). Finally, we performed a sensitivity analysis that excluded wards only receiving outpatients (See SI for further details, Figures S9-S10).

**Acknowledgments**

This study was supported by funding from the Centers for Disease Control and Prevention U01CK000592 and 75D30122C14289. We thank Matteo Perini and the CDC Modeling Infectious Diseases in Healthcare Network (MInD – Healthcare) for comments and discussions.

**References**

Automatic citation updates are disabled. To see the bibliography, click Refresh in the Zotero tab.

**Figures and Tables**

**Figure 1. Empirical colonization of microbial organisms, hospital admissions and testing.** **A)** Incident observations for microorganisms of interest, from left to right and top to bottom: *E. coli*, *K. pneumoniae*, *P. aeruginosa*, *MSSA*, *S. epidermis*, *Candida albicans*, *MRSA*, *E. faecalis*. **B)** Numbers of in-hospital patients (red) and admitted patients (blue) during the study period at daily resolution (weekly variability is evident). **C)** Numbers of in-hospital patients normalized by ward size (average occupancy per day during the study period); blue lines show the 10 most populated wards, green lines the 10 least populated, and the remaining wards are shown in gray in the background **D)** Heatmap showing the number of weekly cultures in each ward during the study period. **E)** Heatmap plot showing the number of patients admitted weekly to each ward during the study period.

**Figure 2. A)** Adjacency matrix of transfers between wards during the study period. **B)** Adjacency matrix of transfers between clusters (aggregation of wards) during the study period. Color bar is in Log10 scale; darker colors indicate greater movement of individuals between each pair of wards or ward clusters. **C)** Incident colonization for the 8 study species; line color designates the cluster. **D)** Heatmap plot showing the number of weekly cultures identified in each ward cluster during the study period. **E)** Heatmap showing the number of patients admitted weekly to each ward cluster during the study period.

**Figure 3. Posterior parameter estimates.** Estimates for the **A)** importation rate and **B)** nosocomial transmission rate. Violin plots show the posterior distribution an estimates for the last iteration of the IF-EAKF (see Methods section) for 3 different runs of the model-inference system. Red dots show the value used in the synthetic simulation (i.e. the truth).Hospital-level simulations ofthe number of **C)** imported, **D)** nosocomial, and **E)** detected colonizations. Each column represents one scenario used for studying the identifiability of the system. Light and dark ribbons show the 95% and 50% uncertainty, respectively, constructed from 300 simulations of the posterior, and dots show the hospital-level observation of the synthetic simulation. Note, hospital level data were not used to optimize the model.

**Figure 4. Posterior parameter estimates (A-B):** Violin plots show the posterior distribution and point estimates from the last iteration of the IF-EAKF (Methods section) for each of the microbial species (Data section). **C)** Hospital level fit: Light and dark ribbons show the 95% and 50% quantiles, respectively, constructed from 300 simulations with the posterior estimates of parameters. Observed carriage is plotted with dots at weekly time scale. D) Calibration plot. Hospital-level fit with 4 different confidence intervals (25%, 50, 75%, 95%). The black dotted line is the reference perfect calibration.

**Figure 5. Importation and nosocomial contribution.** Weekly incident colonization of microbial species. Light and dark ribbons show the 95% and 50% quantiles, respectively, constructed from 300 simulations with the posterior estimates of parameters. Salmon/red color shows importation; light blue shows nosocomial transmission (left and right axis, respectively).

**Table 1.** Parameters used for synthetic scenarios and investigation of ABM identifiability.

|  |  |  |
| --- | --- | --- |
|  | **Importation probability** | **Nosocomial contact transmission rate** |
| Scenario 1 | 0.10 | 0.04 |
| Scenario 2 | 0.05 | 0.04 |
| Scenario 3 | 0.07 | 0.04 |
| Scenario 4 | 0.10 | 0.01 |
| Scenario 5 | 0.03 | 0.02 |
| Scenario 6 | 0.03 | 0.025 |
| Scenario 7 | 0.015 | 0.02 |
| Scenario 8 | 0.15 | 0.04 |
| Scenario 9 | 0.20 | 0.03 |
| Scenario 10 | 0.30 | 0.035 |

**Table 2.** **Posterior estimates**. Mean posterior estimates for the eight species studied; both

importation and nosocomial transmission rates, γ and β, are presented as rates per day.