**Main Manuscript for**

Option 1: Quantifying nosocomial transmission of pathogenic bacteria in hospital settings using patient records and culture data

Option 2: Hospital traffic and surveillance determine nosocomial transmission and likelihood of detection of pathogenic bacteria in hospital settings in New York City

Jaime Cascante Vega1, Rami Yaari1, Tal Robin1, 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, J.C.V., S.P. and J.S. 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

Figure 1. Data figure. A) observed data. B) Timeseries for hospitalizations, admissions and discharges. C) Ward level weekly occupancy (hospitalizations normalized by ward size). D) Heatmap with ward-level number of tests. E) Heatmap with ward-level number of admissions.

Figure 2. Posterior joint inferences on simulated data. A) \gamma=25%, B) \gamma=50%.

Figure 3. Posterior joint inferences on observed data.

Figure 4. Simulation with parameter estimates to assess fit at the hospital. A) Hospital level fit. B) Reliability plots.                                     evel.

Tables 1 to 2

             Table 1. Values of \gamma used for sensitivity - ranges reported.

             Table 2. Posterior parameters estimates and 95% CI for all micro-organisms

**This Supplementary Information PDF file includes:**

           SII Figures S1 to S14

* Figure S1. Transfer matrix(c-es) at ward and building level.
* Figure S2. Ward size distribution **A)** Hospital level. **B)** Building level.
* Figure S3. Hospital traffic at building level.
* Figure S4. Hospital traffic at ward level.
* Figure S5. Relationship between hospitalizations-admissions.
* Figure S6. Discharge rate distributions at hospital and building level (length of stay).
* Figure S7. Convergence plots of inferences with \gamma equal to 25 and 50%.
* Figure S8. **A)** Goodness-of-fit vs oddness. **B)** Goodness-of-fit benchmarked vs oddness normalized
* Figure S9. Hospital level simulation of synthetic tests and calibration. **A.1)** Hosp sim with gamma=25% **A.2)** Hosp sim with gamma=50% **B.1)** Hosp calibration with gamma=25% **B.2)** Hosp calibration with gamma=50%

If including ‘pop’ level fit:

* Figure S10.  Posterior joint inferences on simulated data and ‘population’ level observational model. A) \gamma=25%, B) \gamma=50%.
* Figure S11. Convergence plots of inferences on simulated data and ‘population’ level observational model.

if not:

* Figure S10. Another way of viz. Figure 3. But as a means with error bars.
* Figure S11. Simulation with posterior estimates at building level and calibration for each pathogen.
* Figure S12. Number of clinical cultures in each hospital and its buildings.
* Figure S13. Number of clinical cultures per ward.

**Abstract (269/250 words)**

Pathogenic bacteria are a major threat to patients' health in hospitals. These micro-organisms increase mortality, hospital length of stay, and in consequence healthcare-associated costs. Information on bacterial carriage rates and transmission 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 nosocomial transmission for eight pathogens. We develop an agent-based model informed by patient hospitalization records to simulate importation from the community, nosocomial transmission, and patient spontaneous decolonization of bacteria. The model is coupled with a Bayesian inference algorithm to estimate the likelihood of detection upon testing and nosocomial transmission rates. We evaluate parameter identifiability for this model-inference system and found individual-level observational model yields more accurate estimates. We apply the framework to estimate both quantities for eight prevalent bacterial pathogens: *Escherichia coli, Klebsiella pneumoniae, Pseudomonas aeruginosa, Staphylococcus aureus* (both sensitive, MSSA, and resistant, MRSA, phenotypes), *Staphylococcus epidermidis*, *Enterococcus faecium* and *Enterococcus faecalis*. We find that nosocomial transmission for *E. coli* is negligible, and *MSSA* has a lower nosocomial transmission rate than *MRSA.* We show that while bacterial pathogens have different levels of importation rates nosocomial transmission rates were similar suggesting similar modes of transmission for all. We also found effective sensitivities are similar for all pathogens and proportional their abundance. This work highlights how fine-scale patient data can support simulation-based inference of epidemiological properties of micro-organisms and how hospital traffic, patient contact and surveillance determine epidemiological features. Evaluation of the surveillance 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**

Therapeutic bacterial pathogen control is done with antibiotics.

Antimicrobial resistance microorganisms (AMRO) 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 AMR in 2019 globally, and mortality caused by ABR is projected to reach 10 million by 2050 2. 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 AMR pathogens, and microorganisms in general, within hospital settings, is critical for effective control planning and design of testing/culture protocols. Quantification of these characteristics remains challenging due to limited observation of micro-organisms carriage, difficulty assessing interventions in real-world hospital settings and incomplete understanding of the underlying data generation processes 3–5.

To circumvent these difficulties, mathematical models have been applied to study pathogen transmission in hospital settings, to quantify and understand the relative roles of different routes of transmission 4, and to characterize the condition of the hospital settings to sustain transmission of both resistant and sensitive strains 6. In the context of AMRO, theory has been used to understand the emergence of resistance and its interplay with community-acquired infections 7,8, to evaluate antibiotic treatment protocols 9, to assess control measures to reduce nosocomial transmission 6,10–13. More recently, models have been used in pair with empirical observations to assess the role of competition of different strains at the between-host level and the role of within-host microbiome pathogen interactions 14.

Most existing modeling studies focus on general theoretical frameworks of AMR organisms 6–9,14–18 providing insights on possible mechanisms behind transmission, co-existence and other processes in the hospital settings. Process-based models need to be interrogated with data before being used as reliable tools for public health 19. Investigations using simulation-based inference 20 tools usually focused on a single pathogen of interest such as (MRSA) 10,13,21,22, *P. aeuruginosa* 4*, C. difficil* 23*,* and *Vancomycin-resistant enterococci* (VRE)24using carriage detection 10,13,22,24, intensive care unit (ICU) clinical culture data 4,22 or sequencing data 23. Studies on the epidemiological characteristics of multiple co-circulating organisms in a single hospital network supported by real-world data are absent.

Understanding the transmission dynamics of healthcare-associated (HA) infections is challenging. The epidemiology and transmission HA infections differ from community-acquired (CA) 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). Variations are explained by heterogeneous patient traffic at the building and ward scale (Results). In contrast, communities are relatively 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 7,25. 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 S1 for the patient transfer matrix at both spatial scales) 26,27. 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 its epidemiological features. Lastly, the emergence of strains with different genetic backgrounds that were once confined to hospital circulation such as CA-MRSA clone-USA300 in the United States with enhanced features compared to HA-MRSA makes epidemiology in communities and healthcare settings different 8,25.

In this study, we use a process agent-based model (ABM) and patient clinical culture data for seven prevalent pathogenic bacterial 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, decolonization via host clearance and patient transfer across hospital wards. 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 fix the importation rates by surveying the literature and estimate nosocomial transmission and the likelihood of detection given carriage upon testing. We couple the ABM with a Bayesian inference algorithm and explore the identifiability of the system against simulated data. We estimate the two epidemiological features for eight pathogenic bacteria, which cause substantial mortality associated with AMR worldwide 2. 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. To simulate the transmission of different microorganisms using the same modeling framework, we simplify the underlying biological details that are unique to each microparasitic infection. Our aim is, in consequence, to understand the general similarities and quantify epidemiological properties of communities of circulating bacteria among patients in hospitals. We show that while bacterial pathogens have different levels of importation rates, nosocomial transmission rates were similar suggesting similar level of transmission for all. We also found likelihood of detection is similar for all pathogens and is proportional to the prevalence of bacteria.

**Results**

### *Empirical patterns of hospital traffic and heterogeneity of nosocomial infections burden*

In Figure 1A, we plot weekly and monthly incidence using faded and solid lines respectively, color-coded by building for *E. coli* (total positives cultures, n=2890), *K. pneumoniae* (n=1139), *P. aeruginosa* (n=809), MSSA (n=773), MRSA (n=486), *S. epidermidis* (n=694), *E. faecalis* (n=596) and *E. faecium (n=*263*)*. We deduplicated multiple positive clinical culture results during a patient visit, and consistently deduplicated them 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 S2A) 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. Average ward size was 9 considering all the wards and 20 excluding wards with ward size equal to 1.

Hospital-level variation in hospital demographics, admission and discharge rate, is explained mostly by the heterogeneity at the building scale in the hospital network with some ranging from 0-50% (See building traffic for Allen, Harkness Pavilion and Milstein hospitals in Supplementary Information Figure S1), and others with faster patient replacement between 80% to 150%, a consequence of admitting mostly outpatients (See Presbyterian hospital and Rest in SI Figure S3). This heterogeneity at the building scale is in turn explained by the variation across the ward traffic composing each building (See SI Figure S4 ward traffic). Discharges or outflux of patients at ward facility scale follows similar patterns with a few wards per building admitting the majority of patients per week and with stable patient traffic (See Figure 1C and SI Figure S4). This heterogeneity in admissions and hospitalizations is principally dictated by ward size that was also variable at the building scale (See SI Figure S2B for ward size distribution and SI Figure S5 for the linear relationship). 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 hospitalized patients (Figures 1D and 1E) was linear and depended on ward size (see SI Figure S5). 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). In SI Figure S6A we plotted the distribution of discharge rates at hospital and building level and in SI Figure S6B the distribution of time in hospital obtained from the patient hospitalization records, in average a person spent 3.85 days in the hospital, and 3.17, 4.18, 6.22, 5.05, 1.25 and 1.85 days in Allen Hospital, Harkness Pavilion, Milstein Hospital, Mschony, Presbyterian Hospital and Rest respectively.

### *Model framework and seting: microoganism prevalence*

### We searched the literature to inform the importation rate of each microorganism as a proxy of the community prevalence. Our search terms include 'prevalence', 'colonization', 'carriage rate', and use reviews for some of the microparasites that report pooled estimates across different geographical locations. In the Supplementary Information section *Prevalence estimates* we include the different values, sources, and a brief description of the study, with the geographical location and population of interest. We report the major results in Table 1. In the following analyses, we fixed importation rates of microorganisms using the estimated community prevalence.

### *Identifiability: synthetic simulations and inference*

To investigate the identifiability of the model-inference system, we explore inference on simulated trajectories with known parameters. We investigate if model inference is capable of recovering parameters for a variety of parameter combinations: fix the importation rate to 25% and 50% and vary the likelihood of detection upon testing and the nosocomial transmission rate . The 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). In Figure 2 we show first how the model inference was able to explore different regions of the prior range (range of each axis), and second how the posterior captures the true parameter values. Inferences with are more biased but resulted in a broader distribution than inferences with . Estimates with both importation rates generally captured the truth, with only two instances that were substantially biased, see scenarios 2 and 4 in Figure 2. We visually inspect if the marginal posterior asymptotically converged to the true parameter values in the inference algorithm (see SI Figure S7A and S7B).

We investigate how statistical uncertainty and Monte Carlo error impact the inference (see *Making sense of bias* in Methods). Guided by numerical results, SI Figure S8A shows the goodness-of-fit can be linearly predicted by the oddness of the stochastic realization and SI Figure S8B shows the relation when controlling both by the statistical and Monte Carlo error.

We run model simulations using parameters drawn from posterior distributions and compare them with simulations generated using the true parameters. We highlight the observation used to conduct inference and the mean trajectory across the simulations based on true parameters, defined as the true ensemble simulations (See SI Figure S9A.1 and S9B.1, for respectively). We find that, qualitatively, even biased inferences (scenarios 2 and 4 in Figure 2) reproduce well both the true ensemble simulations and the assimilated data. We quantify the calibration displaying the reliability plot (see *Calibration of inference* in Methods) of both the inferred posterior and truth ensemble simulations (see SI Figure S9A.2 and S9B.2, for respectively). We find that the coverage of confidence intervals for both the posterior inference and true ensemble simulations are usually over the expected values (above the diagonal line), indicating the uncertainty in these simulations is typically higher than that in the observation data. Calibration of the inference and true ensemble simulations are similar, but calibration is better for the true ensemble simulations, as expected. We find it depends on the magnitude of the nosocomial data: lower values of weekly incident simulated data (low and low ) result in less calibrated ensembles.

The last approach shows the performance of the model inference to solve the inverse problem in different regions of the parameter range. We study its ability in parameter combination, pairs , whose simulated data reproduce the observed data aggregated at the hospital level (aggregation of buildings time series shown in Figure 1A) for each pathogenic bacteria. We use the middle prevalence of each bacterial microorganism (Table 1, and Methods). We find the model-inference system 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.

### *Individual vs 'population'-level observational model*

We design the observational model at the patient level making use of the richness of the dataset available. However nosocomial infection data was assimilated at the building scale. We investigate the benefits of this approach compared to a more common approach (see *The population observational model* in Methods). We use the same settings of simulated nosocomial infections detections described in the previous section. Note the meaning of the parameter in the observational model and in the individual and population level observational model respectively is different (Methods). SI Figure S10 shows the joint posterior parameter estimate (same as Figure 2 but for the population-level observational model). We find inferences are in general more bias and much less sharp compared to ones obtained in Figure 2, with the individual-level observational model. We also visually inspect the convergence plots (SI Figure S11) and find that the marginal posteriors are not consistently getting sharper as the algorithm advance (Methods). This lack of identifiability is magnified when compared to the convergence plots of the individual observational model (SI Figure S7).

### *Inference using real data*

We apply the model-inference system to estimate the epidemiological quantities of interest for eight different microbial bacterial pathogens. We compare joint posterior estimates of the effective sensitivity and nosocomial transmission rate in Figure 3A. Posterior for the different values of are color-coded and indicated in the legend of each subplot (Table 1). Species are sorted from the most abundant (*E. coli*) to the least (*E. faecium*) from left to right (see previous section *Empirical patterns and heterogeneity of microorganism burden* for the total numbers). We plot the posterior estimates in the main plot to show how for most species (except *MRSA*) the system is able to localize the posterior in the same region of the prior range (range of each axis). We also plot a zoomed version inside each subplot to highlight that the posterior inferences are sensitive to , i.e. both the mean (intersection of dashed lines) and the posterior estimate change (note that zoomed plots have different ranges in each axis). We find *E. coli* estimated mean nosocomial transmission rates was the lowest across the bacterial species and the highest mean likelihoods of detection of 18.53%, 18.63%, and 17.31% for community prevalence of 55%, 63% and 70% respectively (see Table 2 for confidence intervals). *E. coli* is followed by *MSSA* with mean nosocomial transmission rates of 0.141, 0.121, and 0.0994 and likelihoods of detection of 1.63%, 1.65%, and 1.58% for prevalence of 25%, 29% and 35% respectively (Table 2 for 95% CI). For the rest of the microorganisms, except MRSA and *E. faecalis*, we find that mean nosocomial transmission rate estimates are consistently between 0.17 and 0.19, and likelihoods of detection upon testing between 0.5% (*E. faecium)* to 2.5% (*K. pneumoniae*). Effective sensitivity match in order abundance of bacteria but nosocomial transmission does not (see SI Figure S10A for marginal estimate). For MRSA*,* we found that the model inference system found two solutions as a function of the value of . For equal to 5 and 10%, inference estimated high and low , ; for the lowest community prevalence of 3.9%, inference estimated that the nosocomial transmission rate was 0.0017 and likelihood of detection upon testing was 17.15%. We compile the mean posterior estimates and 95% CI for the different values of in Table 2.

To assess the goodness-of-fit of the modeled nosocomial infection to the observed data, we simulated the dynamics using the posterior estimates of parameters and found that simulated detected nosocomial infections span the observed numbers at the hospital and building level for all micro-organism (Figure 4A and SI Figure S11). We produce reliability plots to examine whether the uncertainty in the observed data can be reproduced by model simulations (Methods). We found that the coverage of simulated CIs is slightly below expected values (under the diagonal line), indicating an uncertainty level biased low (Fig. 4B). The model simulations allow estimation of the relative contributions of importation and nosocomial transmission to the overall burden of different bacterial pathogens as well as the prevalence in the hospital.

**Discussion**

Transmission is a fundamental property that governs epidemiological dynamics and is also a step in the life cycle of bacterial pathogens 29. It is also one of the most challenging processes to understand and quantify. Estimating this property could improve understanding of the mechanism behind the risk of contagion and ultimately support improved control in healthcare systems. Nosocomial infection data have been the usual measure used to approximate and infer transmission; however it represents multiple factors of the surveillance system. Surveillance of microorganism circulation in clinics is a product of patients that are being screened (P(culture | non-infected)) at the discretion of the clinicians and patients who present symptoms due to infection (P(culture | infected)).

In this study, we used a simulation-based inference method to estimate nosocomial transmission rate representing the transmission process and effective sensitivity quantifying surveillance in the hospital. We built a process-based model to simulate patient states at daily time scale and informed the contact network using patient hospitalization records. We parametrized the risk of acquiring bacteria proportional to a nosocomial transmission rate , and the fraction of carriers an individual was contacted during the day. We designed an individual-level observational model to quantify the likelihood of detection upon testing . We assumed community-acquired carriage was at a steady state and in consequence parametrized the probability of importing a microorganism from the community or . Epidemiological observational studies show consistency in the range of human prevalence reported for each bacterial microorganism (Table 1 and SI section *prevalence estimates*), other studies have also pointed at this stability in prevalence for resistant strains of *S. pneumoniae*, and *S. aureus* 30. Ecological theory present mechanistic principles should determine population density (cite), we argue that those mechanisms can also be thought to determine population density in communities of microorganisms inhabiting different body sites in humans. Armed with these experimental and theoretical arguments we set for each microorganism and studied the sensitivity of the inference to different values of this parameter taken from the observational studies.

We assumed the decolonization rate was constant for all microorganisms. However, it is known for example that *S. pneumoniae* bacterial clearances are strain dependent 30. Host clearance rates for the bacteria studied in this work are reported between 5 months to 2 years 10,24 (cite Cooper, Sen, K pneumoniae and *E coli*). In SI section *The ordinary differential equation* we show the basic reproductive number is non-sensitive to clearance rate given fast replacement of patients in the hospital network.

We derived a mathematical expression based on the mean-field approximation of the system and showed that the system is not sensitive to the value for the decolonization rate given the fast replacement of patients in the hospital system. Specifically, we computed the basic reproductive number , and found is the product between the nosocomial transmission rate and the average duration of a carrier inside the hospital, the inverse of the discharge rate plus the clearance rate (SI section *The ordinary differential equation*). There is at least one order of magnitude of difference between the discharge rate and the clearance rate and therefore .

The record of observed nosocomial infection (Figure 1A) is a stochastic realization among an ensemble of possible outcomes of an underlying stochastic process. In order to evaluate the ability of the system to infer key epidemiological parameters, we studied how deviation of the specific realization from the mean trajectory impacts the estimate. We demonstrated in the previous experiments it has an effect on the posterior inferences of epidemiological parameters but in general good inferences are made. We applied the model inferences to estimate the epidemiological properties of eight co-circulating bacterial pathogens. We found that hospital traffic, patient contact network and surveillance of microorganism in hospitals appears to dictate the epidemiological features of most of the co-circulating bacteria. Mean nosocomial transmission rates of all pathogens except *E. coli* and *MSSA* were consistently between 0.15 and 0.19 and the likelihood of detection given carriage from 0.57% to 2.51%.

We found nosocomial transmission rates of *S. aureus* phenotypes were different, suggesting a difference in the fitness of the two strains. It has been assumed principally in modeling studies that resistance generates a decrease in fitness with respect to the sensitive strain, making it less transmissible 1,5,6,8,15. The emergence of CA-MRSA with an equal level of resistance as HA-MRSA without an observed compromise in fitness 25 violates this assumption. We found *MSSA* has a lower nosocomial transmission rate than *MRSA*. A simple model of competition between resistant and sensitive strains would predict that as the ratio between nosocomial transmissions between strains the resistant strain should outcompete the sensitive strains ('competitive exclusion') 18,32 that is not the case observed empirically (Figure 1A). Split into resistant and sensitive phenotypes could possibly be the most important categorization clinically 33 however is agnostic to the number of resistant genes carried by each. Within-host dynamics between these two 5,33 including an accurate representation of the levels of resistance in each, loss and gain of resistance genes through means of HGT 14, and interactions between CA-MRSA and HA-MRSA strains 8,25 intertwined by the fast dynamical nature of hospital traffic could help to understand estimated differences in the nosocomial transmission rates.

Ultimately considering and modeling all these processes could lead to more specified models and provide an accurate understanding of the processes shaping the co-existence of *S. aureus* strains in the hospital. Finally, between species competition could also be important, *S. aureus* and *S. epidermis* inhabit the nasal nares as commensal, these overlap in their ecological niche produce negative feedback between the population dynamics within the host of both species ('competitive release') (cite) and possibly ramify up to hospital-level dynamics of these two. Polymicrobial infections with *P. aeruginosa* and *S. aureus* are considered to be harmful to their host resulting in worst healthcare outcomes. In turn, *in-vitro* experiments have shown that *P. aeruginosa* outcompetes *S. aureus* possibly modulating the individual risk of contagion as well as hospital-level observed carriage of each 34.

We used an individual-based computational model to enrich the dynamical representation of patient movement across the hospital network using patient records in a major New York City hospital network. We were able to accurately reproduce the observed data across the hospital and building levels. Our model absorbs both direct and indirect modes of transmission in the same parameter . Fast dynamics of patient traffic, from daily to weekly timescales, points to transmission possibly dominated by indirect modes of transmission, that do not depend on host-to-host contact, especially environmental or fomite as well as HCW-mediated transmission 4,35–40. In fact, the Center for Disease Control and Prevention (CDC) healthcare infection control webpage 41,42 highlights the environment, surfaces and devices, as a common reservoir of germs in healthcare places and hand washing, cleaning. How does fomite transmission impact the force of infection? (linearly or non-linearly) and the timescales of bacterial clearance rates (survival times) in the environment will also be relevant questions. Other limitations of the current study include the absence of information on the demographics, age profiles and co-morbidities of patients in the hospital network that are known to have differential susceptibility to bacterial acquisition and infection 1(cite), which we therefore do not model.

We found that while *E. coli* has the highest likelihood of detection (mean estimates from 17.31% to 18.53%), it has the lowest nosocomial transmission rate almost negligible suggesting that most nosocomial infections can be associated with infections with host commensal strains that do not contribute to transmission, which has been reported empirically 31.

Further, the current work exploits the availability of individual-level patient records to estimate the transmission properties of pathogenic bacteria in hospital settings. Individual-level data and models can be employed in future research to understand the impact of individual-level interventions on disease control, as disinfection of shared medical devices to diminish the risk of contagion (CDC recommendations). Another 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 bacterial pathogens, we developed an ABM to simulate the dynamics of these organisms in hospital settings. The model was informed by patient hospitalization and clinical 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 nosocomial infection data, validated the ability of this ABM inference system to identify likelihood of detection and nosocomial transmission rates. The process model, as is an ABM, tracks the state of 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 clinical culture. We parametrize the patient observational model with a likelihood of detection given carriage upon testing given the effective sensitivity or (See *The individual observational model* inMethods). Data assimilation is conducted at the building level therefore we consistently aggregate simulated patient detections at this scale. We then assimilated the clinical culture data and estimated the effective sensitivity and nosocomial transmission rates for eight co-circulating bacterial pathogens. Using these estimated parameters and the ABM, we were able to reproduce the time series of positive cases for five major buildings in the hospital network and the aggregation of all the other wards in a fictitious unit 'Rest'. The estimated effective sensitivity and nosocomial transmission rates were compared for the eight pathogens.

### *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 221 wards of different types including emergency, infusion, cardiology, pediatrics, etc. Hospitalization and clinical culture 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. The dates and wards in which each patient stayed during hospitalization were used to construct a time-evolving contact network. Clinical culture records to confirm nosocomial infection were available for the eight most prevalent organisms in the hospital system: *Eschericia coli*, *K. pneumoniae* *Pseudomonas aeruginosa*, methicillin-susceptible *S. aureus* (MSSA), MRSA, *Staphylococcus epidermidis*, *Enterococcus faecalis* and *Enterococcus faecium*. 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 and it's buildings (SI Figure S12) and in the wards in each hospital (SI Figure S13), which show a similar order in cultures across the hospital but substantial heterogeneity inside the hospital at the ward scale.

### *The transmission model*

We use a process ABM to simulate transmission in the study hospital system 10. 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, care from the same hospital workers who might facilitate transmission of microbial pathogens and share the same environment. 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 days 12. Colonization in hospitals is 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 but keep track of patient states on consecutive days. However, if patients were admitted on 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 define a frequency-dependent transmission rate per ward as , where is the average ward daily occupancy (See *Empirical patterns of hospital traffic and heterogeneity of nosocomial infections burden* in Results). 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 shown below by, where is the decolonization period of patient .

### *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 variables 1,6. 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 apply 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. 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).

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. For instance, blood cultures of a patient colonized by *E. coli* in the urinary tract will likely test negative, and this patient may not necessarily present symptoms (Seth Blumberg told us - Sen and I - this, but I did not find any reference. Should we just leave the guess there?). This compounded with differential detection of screening and diagnosis makes the parametrization challenging. To define our observational model, we therefore estimated the ‘effective sensitivity’ of detecting colonization upon testing defined by the culture data, 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 that colonization or infection exists anywhere on a patient. The defined observational model indicates the likelihood that a colonized patient is detected given the effective sensitivity or .

### *The population-level observational model*

The process model keeps track of the patient states that can be aggregated at any scale available in the patient records. We modeled the ward-level detections using a Binomial observational model. The number of trials is equal to the number of clinical cultures in each ward at day , and the probability of sucess is assumed to be proportional to the fraction of colonized people times the ward-level effective sensitivity , as shown below. We aggregate at the building scale for inference.

### *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 used the building aggregation to sum the number of nosocomial infection data and perform inference. SI Figure S12 shows the number of clinical cultures per building, we merge buildings composed mostly by outpatients and with few clinical cultures into a fictitious unit presented as 'Rest'. The number of transfers between wards is shown in SI Figure S1. Wards within each identified building 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 43,44. This framework has been applied to many infectious diseases, including influenza, dengue, malaria, cholera, Ebola, enterovirus D68, and COVID-19 45–54, typically 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 solve the inverse problem without solving the filtering problem and therefore only estimate parameters, i.e., infer the effective sensitivity and the nosocomial transmission rate , without re-adjusting the state space.

We use the ensemble adjustment Kalman filter (EAKF 43), 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 (cite). We assimilated weekly observations of total numbers nosocomial infection detections for microorganisms 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 55,56 ,but has also been proposed in the context of inverse problems using Kalman filters 28,57. We base most of our implementation in the algorithms proposed in the epidemiological context but grab some ideas from the algorithms using Kalman filters. The implementation of the IF-EAKF in Python was made using the packages NumPy and SciPy 58,59. 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 data (i.e., the ability of the framework to infer the parameters when they are known). We generate synthetic observations of incident microorganism colonization using the ABM and assigned effective sensitivity and nosocomial transmission rates . The synthetic observations were then assimilated into the full model-inference system to assess the system ability to accurately estimate the parameters.

*Identifiability: Making sense of the bias*

In the synthetic inferences the true parameters **P** are known and the simulator allows repetition of the experiment. We define the oddness of a stochastic realization as the average Continuous Ranked Probability Score (CRPS) over data assimilation times between an ensemble and a single stochastic realization. Additionally, the parameters of the simulator are known to impact the spread of the simulated data (statistical uncertainty) 51,60 (Bolker and Grenfell). In consequence, we define the ensemble spread using the CRPS of ensemble simulations and its mean. The CRPS is defined as shown below, F is the cumulative density distribution computed from the ensembles and y is one trajectory 61,62. is the indicator function where if or otherwise. The CRPS is in fact a standard proper score to evaluate epidemic forecasts 61.

In synthetic tests, inferences about the mean trajectory could be made. The resulting inferred parameter density provides a benchmark of the method given the noise-less signal possible, the mean trajectory. Quality of inferred parameter densities conducted on single stochastic realizations can be weighted by the goodness-of-fit of the benchmark . We measure the goodness-of-fit of a posterior density given **P** as proposed in 57. Specifically, the inferred posterior can be parametrized with a vector of means , and a covariance matrix and the goodness-of-fit assessed with the L2 norm scaled by , as shown below.

We first studied how the oddness of the observation explain the goodness-of-fit of the inferred parameters (SI Figure S7A). In this controlled experiment the true parameters and the ensemble spread are known. Monte Carlo errors 1 are controlled by normalizing by and removing statistical uncertainty . In SI Figure S7B we show the contolled goodness-of-fit of against the (Figure S7B).

### *Inference using real-world data*

We next used actual weekly observations for the six buildings to infer the epidemiological parameters for all eight bacterial pathogens. The model-inference framework was applied to each bacteria separately. 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.

### *Calibration of inferences*

We measure the consistency of the modeled nosocomial detection and the observed ones statistically. We produced reliability plots 64. 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.

**Acknowledgments**

This study was supported by funding from the Centers for Disease Control and Prevention U01CK000592 and 75D30122C14289. We thank Matteo Perini, the Shaman lab 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**

**A picture containing text, line, plot, font

Description automatically generated**

**Figure 1. Empirical colonization of microbial organisms, hospital admissions and testing.** **A)** Nosocomial infection data for pathogenic bacteria studied, faded dots are the weekly incident and solid lines the monthly incident detections, from left to right and top to bottom: *E. coli*, *K. pneumoniae*, *P. aeruginosa*, *MSSA*, , *MRSA*, *S. epidermis,* *E. faecali* and *E. faecium*. **B)** Numbers of hospitalized 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 5 most populated wards, blue lines the 5 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.

**A picture containing text, screenshot, line, number

Description automatically generated**

**Figure 2. Identifiability, posterior estimates.** Joint posterior estimates for **A)** importation rate =25% and **B)** =50%. The posterior estimate is highlighted with a density plot (darker means more probable), and posterior ensemble members are shown as purple dots. In each subplot the true value used for simulating the stochastic trajectory to infer is highlighted in the title of each scenario, with a yellow cross and the intersection of the two black dashed lines. x-axis shows the likelihood of detection upon testing (%) and y-axis the nosocomial transmission rate . Note that in both **A)** and **B)** the prior range is the limits of each axis, increments from upper to lower plots and from left to right.

**A picture containing screenshot

Description automatically generated**

**Figure 3. Posterior parameter estimates.** Joint estimates for the likelihood of detection upon testing (%) (x-axis) and the nosocomial transmission rate (y-axis). Importation rate values for sensitvity analysis are coloxr-coded with high, medium and low prevalences, and indicated in each subplot. Color-coded dashed lines show the mean estimates, and posterior is shown with both a density plot (darker indicates more probable) and with the posterior ensemble members plotted as dots.

**A collage of graphs

Description automatically generated with low confidence**

**Figure 4. Hospital level fit (not used for data assimilation):** **A)** Ensemble simulation of modeled nosocomial data with estimated parameters, solid lines show the mean and ribbons the 95% CI. Importation rate is highlighted in the legend and color-coded for high, medium and low prevalences in all plots. Weekly nosocomial infection data is displayed with red dots. **B)** Reliability plot (Methods). Hospital-level fit with 4 different confidence intervals (25%, 50%, 75%, 95%). Importation rate is color-coded. The black dotted line is the reference perfect calibration.

**Table 1.** Human prevalence range from the literature, we present the 3 values used in the inferences (SI *Prevalence estimates)*.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Importation rate, (%)** | **Nosocomial transmission rate**  **Mean (95% CI)** | **Likelihood of detection (%)**  **Mean (95% CI)** |
| *E. coli* | 55 | 2.54E-03 (1.72E-03, 3.58E-03) | 18.53 (17.30, 19.43) |
|  | 63 | 8.66E-04 (3.78E-04, 1.58E-03) | 18.63 (17.85, 19.23) |
|  | 70 | 9.62E-04 (4.06E-04, 1.73E-03) | 17.31 (16.33, 18.02) |
| *K. pneumoniae* | 15 | 0.187 (0.180, 0.194) | 2.51 (2.36, 2.65) |
|  | 23 | 0.189 (0.180, 0.195) | 2.14 (1.98, 2.29) |
|  | 35 | 0.181 (0.165, 0.193) | 2.05 (1.92, 2.17) |
| *P. aeruginosa* | 11.6 | 0.189 (0.183, 0.194) | 1.93 (1.78, 2.07) |
|  | 18.8 | 0.190 (0.182, 0.195) | 1.61 (1.47, 1.75) |
|  | 25. | 0.187 (0.178, 0.194) | 1.55 (1.43, 1.67) |
| *MSSA* | 25 | 0.141 (0.122, 0.164) | 1.63 (1.47, 1.79) |
|  | 29 | 0.121 (0.103, 0.140) | 1.65 (1.49, 1.81) |
|  | 35 | 9.94E-02 (8.43E-02, 0.116) | 1.58 (1.40, 1.73) |
| *MRSA* | 10 | 0.170 (0.136, 0.184) | 1.38 (1.21, 1.60) |
|  | 3.9 | 1.70E-02 (9.16E-03, 2.74E-02) | 17.15 (12.48, 18.96) |
|  | 5 | 0.182 (0.168, 0.190) | 1.96 (1.72, 2.19) |
| *S. epidermidis* | 58 | 0.171 (0.117, 0.190) | 1.24 (1.12, 1.35) |
|  | 75 | 0.174 (0.134, 0.190) | 1.21 (1.10, 1.33) |
|  | 90 | 0.173 (0.107, 0.189) | 1.23 (1.12, 1.36) |
| *E. faecalis* | 36 | 0.167 (0.121, 0.186) | 1.05 (0.96, 1.17) |
|  | 47.6 | 0.149 (9.69E-02, 0.182) | 1.07 (0.96, 1.19) |
|  | 55.0 | 0.159 (0.101, 0.185) | 1.09 (0.96, 1.21) |
| *E. faecium* | 36.8 | 0.182 (0.163, 0.192) | 0.58 (0.54, 0.63) |
|  | 40.6 | 0.179 (0.147, 0.191) | 0.57 (0.54, 0.65) |
|  | 50.0 | 0.178 (0.123, 0.193) | 0.58 (0.54, 0.66) |

**Table 2.** **Parameter estimates**. Posterior estimates for the eight bacterial pathogens; the value for the importation rate γ is shown in the second column. Nosocomial transmission rates β is presented as rate per day in the third column and likelihoods of detection upon testing in the fourth column.