Manuscript Number:**EPIDEMICS-D-24-00069**  
  
Manuscript title: **Estimating nosocomial transmission of micro-organisms in hospital settings using patient records and culture data**

Authors: **Jaime Cascante Vega, Rami Yaari, Tal Robin, Lingsheng Wen, Jason Zucker, Anne-Catrin Uhlemann, Sen Pei, Jeffrey Shaman**

**Editor's comments**  
**Editor 1**

I concur with the reviewers that bypassing HCWs in such modelling is quite a strong assumption... How to interpretet the estimated transmission parameters in a hospital hygiene perspective is therefore difficult. At best, the model would be changed, but at least, some discussion regarding the consequences is warranted.

Answer. Thank you for the comment. We agree that our inference of nosocomial transmission rate embeds substantial uncertainty as it models person-to-person transmission, vectored transmission by HCWs, and also fomite transmission using the same parameter. However, in the absence of data to parametrize the movement of HCWs or on the presence of environmental reservoirs we limit ourselves to available data on patient movement. In the section *Understanding nosocomial transmission,* we discuss these limitations and suggest that future models explicitly and separately represent those modes of transmission.

To further elaborate on the consequences of our assumption we have present this issue in the discussion: ‘Fomite transmission could contribute to nosocomial acquisition through both patient contact, for example, via the handling of uncleaned devices, or by supporting vectorial transmission via HCWs. If those modes of transmission contribute linearly to the estimated nosocomial transmission rate, then a fraction of our estimate would be attributable to that specific mode. If the contribution is non-linear, then it would be necessary to investigate the frequency of interaction and the fraction of HCWs colonized on a ward at a given time. The effects of such interactions on nosocomial transmission are difficult to resolve without more data, specifically measurements of environmental and HCW colonization.

**Editor 2**

A second point is I did not understand how the ABM formulation was essential to your results? It seems that the transmission model is Markovian and could have been simulated otherwise. In other words, you could have simulated the hospital with standard compartmental models, and I don't think that the reslts would be changed. It would be good to show whether the ABM is really of importance, or if it is just a choice for simulation.

Answer. Thanks for the comment as it allows us to further describe our model-inference approach. The preference for an ABM versus a compartmental stochastic model, which is in part choice, is described below.

1. Both the observational model, as well as the process model, use individual-level test and movement data. Together, these data resolve the specific agents who have been tested—and where they are at that time. The ABM form thus enables use of a dynamical observational model that represents and allows assimilation of individual-level testing sensitivity.
2. During the study period, few wards are consistently populated, and thus using, for example, a stochastic compartmental model requires aggregation of the wards to a coarser spatial scale. This also requires a different parametrization of the observational model, for example by assuming a Binomial observational model with the number of trials equal to the number of tests and the probability of testing positive proportional to the fraction of carriers inside a spatial unit.
3. The movement of patients in the ABM is defined by the individual-level data on patient location in the hospital. The movement data are thus suited to be represented by an ABM framework, or some system with explicit spatial simulation, e.g. a metapopulation model. However, the use of a coarser aggregation form would require design of an additional movement model.
4. Finally, even when another movement model of patients on a coarse spatial scale is considered it is likely that the parametrization of this system would not capture the heterogeneity in admissions and discharge rates that are time-varying. Additionally, patient length of stay was also heterogeneous across the hospital network (supplementary material Figure S6), and aggregations at any spatial scale are unlikely to represent this heterogeneity.

**Reviewers' comments**  
**Reviewer #1**  
  
**Comment 1.** A justification is missing in the methods section or supplementary material for the assumption of a Gaussian observation process.  
Answer. Thanks for this comment. We did assume that the observations/assimilated data are normally distributed, as the method used for data assimilation, the ensemble adjustment Kalman filter (EAKF), models both the likelihood and prior as Gaussian, so the posterior is also Gaussian. As we only have one observation point, the Gaussian approximation is used because of the mathematical and computational convenience of the data assimilation method. Indeed, the EAKF has been widely used for forecasting and inference of infectious disease systems. However, our observational process in the ABM simulation directly observes the patient status in each ward, and it is not assumed that the number of detected carriers is drawn from a normal distribution.

**Comment 2.** The observed data is assumed to be normally distributed, should this not be a lognormal distribution? I assume you are constraining to positive values.

Answer. Thanks for the comment. The EAKF assumes the observed/assimilated data are normally distributed so the posterior can be updated with a deterministic mathematical formula. With this approach, we only parametrize the observed data using an observational error variance (OEV, as indicated in the Supplementary Material section **Simulation-based inference framework**) that we model proportional to the data, that is to the number of positive cultures. There's no need to constrain it to positive values as the OEV enters the update deterministically. It is possible, however, that the posterior of either the parameters or the state variables fall outside the prior range or even become negative. To prevent this, we constrain the parameters inside the specified prior range and the state variables to be positive.  
  
**Comment 3.** Figures are missing axis labels and alphabetic labels (e.g. A, B which are used in the figure caption) throughout the manuscript and supplementary materials. For example, in Figure 1 and Figure 4. Please add these to ensure the figures are clear.

Answer. Thanks for this. We forgot to add the labels in the last update of the manuscript. In the current submission, we have included them.  
  
**Comment 4.** Figure 4 caption refers to a legend which isn’t present. Please add and check other figures.

Answer. Thanks, in the revision we have included the caption.   
  
**Comment 5.** Figure S12 refers to ‘institutions’, this terminology is not used anywhere else in the manuscript. The distinction between institution and hospital is not clear to me, please clarify.

Answer. Thank for making noting this oversight. We have changed it to 'hospital'.

**Comment 6.** The abstract would benefit from stating that an ensemble adjustment Kalman filter is used.

Answer. We now include this information in the abstract.

**Comment 7.** A rationale is needed for using the ensemble adjustment Kalman filter method over a method more often used for inference of transmission rate parameters, such as ABC/SMC/MCMC.

Answer. The inference algorithm is a choice. Both SMC and EAKF have been used widely for simulation-based inference of epidemiological dynamics, see for example: (1–6). Indeed, some comparisons between those are available (7–10). Due to the high dimensional state space of the problem corresponding to the number of patients during the study period, the EAKF is preferred as it allows efficient update of the prior distribution. As a consequence, fewer Monte Carlo samples are needed when compared to the other Bayesian approaches mentioned by the reviewer. Indeed, the EAKF has been used for data assimilation in other ABMs, including network models with millions of individuals (11).

**Comment 8.** The study period coinciding with peak COVID-19 cases in the city is mentioned as a limitation. Further discussion on how this might affect the study results would be beneficial. Did you consider fitting multiple transmission rates for different time periods, perhaps when different levels of restrictions were in places?

Answer. We thank the reviewer for this comment, which is similar to **Comment 4** of **Reviewer 3**. We studied the possible effect of the pandemic by conducting inferences, ignoring the data during this period. Inferring multiple transmission rates for different time periods is challenging because the study period is relatively short; however, the impact of the data reported for a specific period on the parameter estimate can be investigated. We point the reviewer to **Comment 4** of **Reviewer 3** for details on our experiments and the findings.  
  
**Comment 9.** A more detailed discussion is needed on the types of infectious pressure that the ABM model is not capturing, particularly the lack of hospital staff in the model. These are briefly mentioned in *Understanding nosocomial transmission*. Ideally, when modelling nosocomial transmission an ABM would either explicitly model HCWs and hospital staff as agents in the model, or implicitly include them within the force of infection. Moreover, infection dynamics could be disentangled further if other sources were considered, such as: staff infectious pressure, spatial infectious pressure from other wards (patients/staff) that may be connected, and a background infection pressure. I assume staff data was unavailable for this study?

Answer. Thank you for these comments as they allow us to further strengthen our discussion around the estimated parameters. As the reviewer intuits, HCW data were not available for this study; we implicitly represent transmission via HCWs as part of the nosocomial transmission rate parameter. This comment is similar to the comment of **Editor 1,** which the reviewer can see at the beginning of the document. We have extended our discussion in the section *Understanding nosocomial transmission* to further clarify how our nosocomial transmission rate embeds person-to-person, vectored transmission via HCWs, and possible transmission by contact via environmental reservoirs. We also discuss the possible contributions of these various modes to the nosocomial transmission rate.  
  
**Comment 10.** It is unclear to me why the force of infection for each ward only considers the infection/colonization status of patients in the same ward. Did you consider adding a ‘background’ transmission rate parameter which scaled with the number of infections in the hospital at that time?

Answer. Thanks for this comment. A background transmission rate might model, for example, fomite transmission, which could be represented as proportional to prevalence in the hospital. We initially attempted to estimate two transmission rates: a person-to-person transmission rate and a second additional environmental transmission rate parameter. We were unable to accurately identify both parameters using available data, i.e. the data used to inform the inference are sparse and represent only a small fraction of the prevalence and thus do not support inference of multiple modes of transmission. As a consequence, we decided to combine transmission into a single parameter. This challenge magnifies as we are considering a range of pathogenic bacteria whose natural history is poorly understood.  
  
**Comment 11.** A daily contact network was used rather than a finer-scale (e.g. hourly) to capture patient movements, was this for computational benefit? It is not clear to me if a patient was admitted into A&E and subsequently to an inpatient ward on the same day, whether they would only appear in the inpatient ward of the contact network? If so, the model may not be capturing transmission in these staging areas.

Answer. We decided to use a daily time step, both for simplicity and computational benefit.

**Comment 12.** The final three sentences of the introduction seem unnecessary and should be covered in the abstract/results and discussion.

Answer. We have deleted these sentences.  
  
**Reviewer #2**

This is an interesting paper and the methods are well explained. However, by focussing on carriage, particularly in the case of *E. coli*, I worry that the conclusions are not helpful or informative and this needs to be addressed and discussed in more detail. The model does not distinguish between commensal bacteria and pathogenic or resistant bacteria and the message that *E. coli* is not commonly transmitted in hospital settings is likely not true, and is in fact an artefact of the model structure that only allows for a single colonization event with *E. coli*. While all models are abstractions, this issue cannot be ignored.  
  
Here is a list of comments, please also see annotations in the PDF document attached (if you can't get the attached file, inquire with the journal).

**Comment 1.** There is no definition of what is classed as a nosocomial case in the underlying data. Many cases that are considered community acquired based on days in hospital before infection is detected are in fact potentially related to a previous hospital event. Is this included in the definition of nosocomial used here?

Answer. Thanks for these comments. In the first paragraph of the discussion, we wrote *"The patient-level clinical culture data mostly represent infection. Clinical culture samples were targeted at one body site and were typically obtained on patients presenting symptoms, suggestive of infection at a specific site and a preponderate use of clinical cultures for diagnosis rather than surveillance."* In addition, in the section *Understanding effective sensitivity* we mention that surveillance is mostly targeted to infected patients: "*Surveillance of micro-organism infection and colonization in patients is primarily the product of clinician-directed collection of cultures for infection-compatible clinical symptoms (i.e. fever, dysuria, or cough).*"

The study hospital system is a clinical setting in which most carriage is detected by diagnosis and not by screening. To clarify this further, we now include a sentence in the Methods section *Data: Clinical cultures*: "Cultures were collected mostly at the discretion of clinicians as part of diagnosis in order to confirm infection ".

**Comment 2.** Are some sites more likely to be sampled than others, therefore effecting the estimated prevalence in the underlying data?

Answer. There are some wards in which patients were cultured more often, see Figure 1D and SI Figure S12. Our observational model accounts for the effect of different sampling across wards in the hospital network. Indeed, as our observational model uses the number of clinical cultures at the individual level we are accounting for the likelihood of diagnosis given infection, which is differential across wards. It is possible, however, that wards have different surveillance settings: i.e. they differ on the amount of screening and diagnostic. Those differences are not modeled and could bias the estimate of the effective sensitivity . However, most of the nosocomial data used to inform the inference were typically clinician ordered for patients suspected to be infected. In the SM we show that under a scenario in which screening is negligible, high values of could only be obtained with high values of the ratio between diagnosis given infection and probability of testing (i.e. high values of , see SM Fig S11) combined with high values of the infection to clearance ratio . This ratio, , if high, indicates that infection rates given carriage are much faster than carriage clearance, which is unlikely. It also suggests that even when and the infection-to-clearance ratio varies within wards, it is unlikely that this will result in substantially different effective sensitivity values.

**Comment 3.** The parameters underlying the model are drawn from literature outside of the region being simulated. While I appreciate that this is a limitation that cannot be avoided when there are no available data, a full sensitivity analysis should be performed on these parameters to determine the impact that they have. For example some of the *E. coli* estimates for community carriage are very high and this reduces the impact of nosocomial transmission in this model framework where patients already carrying *E. coli* cannot be infected again. A sensitivity analysis is essential here as it is highly unlikely these results hold true for antibiotic-resistant *E. coli* which likely have a much lower prevalence in the community. I also feel that this point should be drawn out in the discussion. The introduction talks about ABRO, but this is a very different problem to what is being modelled in this paper.

Answer. Thanks for this comment, as it allows us to refine the relationship between community prevalence and estimated nosocomial transmission. In our previous analysis, we selected 3 values for the importation rate spanning values found in the literature worldwide. The inferences with these parametrizations were presented in Figure 3. We now present inferences using only the highest value reported in the literature , and fractions of this value (). For all bacterial pathogens we varied between 1 and 0.3, with steps of 0.05. The results of this work are presented in Figure R1 below. In each subplot, we highlight the previous ranges of importation rates (from lowest to highest reported in the literature) with green dashed lines. We find that within the previously considered range our results are consistent with the findings presented: all bacterial pathogens, except *E. coli* and MSSA*,* span similar ranges of both the nosocomial transmission rate and the effective sensitivity. In Table R1 we present the coefficient of determination to investigate if the changes in the mean parameter estimates are accounted for by the set value of .

The nosocomial transmission rate estimates for *E. coli* are consistent until values drop below importation rates of 40%, a drop of more than 50% from the initial highest value . We find a similar pattern for the estimated effective sensitivity of this pathogen. We find that the importation rate explains inferred changes of both the nosocomial transmission rate () and the effective sensitivity (). However, the relation is non-linear with importation rate values below 40% causing an increase in and a decrease in . For higher importation values the estimates remain unchanged. This result is of interest when contrasted with the relation obtained for the other microbial bacterial pathogens. For all the other microbial pathogens we find only one of the estimated parameter changes as a function of the importation rate. Specifically, the nosocomial transmission rate was found to be explained by changes in the importation rate for *E. faecalis* () and *E. faecium* (). For the enterococcus species the decay in is well explained by an increase in . For the rest of the bacteria, *K. pneumoniae ()*, *P. aeruginosa* (), and MSSA () the effective sensitivity changes as a function of the importation rate but the nosocomial transmission rate does not. The relationship is non-linear and displays a threshold pattern with values below 50% of the highest importation rate for *P. aeruginosa* and *K. pneumoniae*. For MSSA however, the decay of is well explained by an increase in , thus displaying a linear relationship.

These results together show a robust estimate of the nosocomial transmission rate , and the effective sensitivity on a broader range of importation rates than presented in the main text, which we found in the literature. Biases in the parameter estimates occur for importation levels not reported in the literature (Figure R1), and such biases are typically nonlinear.

**Comment 4.** limitations could be addressed by a global sensitivity analysis as suggested.

Answer. Thanks for the comment. In the inferences we presented in the previous submission, we first investigated the effect of readmission. For this, we tracked the patient's state even when discharged from the hospital and presented the inference results in the Supplementary Material Figure S13. The discussion of these results was presented in the paragraph *Limitations: heterogeneity in carriage duration and readmission*.

We conducted synthetic tests: parameter estimates on simulated data with two levels of community prevalence . The posterior joint density was presented in Figure 2, and calibration of the fit compared to the truth was shown in SM Figure S9. These results showed we were able to estimate both , and the calibration to the observed data was as good as the one obtained with simulations with the true parameters. However, some of the microbial pathogens have community prevalences of between 15 and 35% (*K. pneumoniae*), or between 5 to 15% (MRSA). These ranges are considered the lowest to highest community prevalences we found in studies worldwide.

To further study the sensitivity of the importation rate parameter we conducted additional synthetic tests with importation rates of . We used 3 different levels of detection rates, varying the effective sensitivity , and 3 different levels of nosocomial transmission , note that the values found for the effective sensitivity span those inferred for the microbial pathogens. To show that the inference is robust to low transmission regimes we selected values of lower than those inferred for the pathogenic bacteria. In Figure R2 A, B and C we present the posterior joint distribution for each community prevalence. We find a relatively consistent pattern regardless of the community prevalence: the simulation-based inference method accurately identifies the parameters when the nosocomial transmission rate and the effective sensitivity are not at their lowest value. However, even when or are at the lowest value selected for the synthetic testing, i.e. or , we find the inference can locate the region of the truth value, though the posterior does not capture the truth (yellow crosses). Thus, though some bias remains in some instances, the solution is localized to the correct region of parameter space. To further highlight this localization around the truth in Figure R3 we present the convergence plots of the synthetic inferences. We find that across iterations of the Iterated Filtering, the marginal posterior estimates asymptotically gravitate toward the true parameter values, though not as well for high values of or .

Second, we investigate the inference of the parameters by using lower values of the importation rate for each microbial pathogen as indicated in ***comment 3*** of the reviewer. Lastly, we investigated how the COVID-19 epidemic might affect the results we presented. This work is presented and discussed in response to **Comment 4** of **Reviewer 3**.

**Comment 5.** The layout of the manuscript is excellent and the subheadings are very clear and make the manuscript very easy to navigate. However the writing in this manuscript feels rushed and is confusing in places. The manuscript would benefit from a thorough readthrough and edit by someone less familiar with it's content.  
Answer. We thank the reviewer for the comment. We have conducted another round of editing for style, syntax, and grammar.

**Reviewer #3**

I feel the article applies interesting methodology and a lot of technical interesting work has been performed, but still, I wonder how much added value the manuscript has to the existing literature. The biological conclusions, that *E. coli* spreads not very well and MRSA reasonably well in hospital environments is not new and I do not trust the numerical values too much.  
  
**Comment 1.** My impression is that fixing the importation rate is quite a big assumption. Especially when the amount of data is limited, as is the case as primarily infection data are being used, differentiation between importation and transmission is difficult as both are positively related to a high prevalence. A choice of the importation rate based on the literature, from many different countries, can have quite some impact on the transmission parameter as is seen in the sensitivity analysis.

Answer. We agree with the reviewer that using importation from other locations is a big assumption. For that reason, we selected values that spanned the lowest, median, and highest values we found in the literature across the globe. We discussed this in detail in the section *Heterogeneity and steady state of community prevalence*, and further discussed some consistencies in the values reported worldwide. Without additional data to parametrize the model-inference system with information from New York City, specifically northern Manhattan, it is suspect to pick one value. Despite importation rates spanning a range of reported values worldwide, we find consistency in the estimated values for both the nosocomial transmission rate and the effective sensitivity (see **Comment 3** of **Reviewer 2**).   
  
**Comment 2.** Also, I find the assumption of a constant transmission parameter in all wards and all hospitals rather unrealistic? I expect that there will be at least a three-fold difference between wards (with ICUs among the wards with the highest transmission parameters). How should we interpret the average transmission parameter which is calculated in this study?

Answer. We use a single parameter to model the transmission rate from infected to susceptible patients. However, by creating a daily contact network we are capturing heterogeneity in contact in time. In a typical epidemiological model, the force of infection is proportional to a transmission rate that embeds the probability of transmission upon contact, times the number of contacts. Our modeling approach, by using the individual-level patient data, is discriminating the number of contacts in time. However, as the reviewer suggests control measures are different between wards. Our estimate presents a weighted average across the hospital network, but we are controlling by the time-varying nature of the contacts. In future studies consideration of differential control measures across the hospital network might reveal differential spread across the hospital. However, without additional data to support the parametrization of control measures the system would be unidentifiable.

**Comment 3.** Why are the analyses to determine identifiability of the parameters only done with importation rates of 25% and 50%? A lower value of 5% or 10% seems to reflect the situation for MRSA and Pseudomonas better.

Answer. Thank you for the suggestion as it allows us to further investigate the identifiability of the simulation-based inference system we present. Please see our response to **Comment 4** of **Reviewer 2**;we investigated the identifiability of the parameters using simulated data with importation rate values of 5, 10, and 15% as the reviewer suggests (Figures R2).

**Comment 4.** I find the choice of the time period a bit strange, a relatively short time period is chosen to ensure that the admission prevalence is more or less constant, but the time period does overlaps with the COVID-19 epidemic which had a big impact on processes in the hospital. Moreover, from Figure 1 I conclude that the number of hospitalized patient is zero at the start of the study? It seems that only data are used from the starting date onwards, but this means that patients already present in the hospital at the starting data are not part of the estimation process, which means that in the first weeks/months, the prevalence will be relatively low. This can also be seen form the empirical colonization data. I think only the data should be used from the moment the hospital size is in its stable distribution as the initial data do not represent a stable situation.

Answer. Thank you for this comment as it allows us to understand the impact of different time periods on the parameter estimates. We investigated the impact of the low number of patients at the start of the study period, as well as the overlap of the COVID-19 peak cases in March 2020, by ignoring data from the start of the study period. We conducted parameter inference by not assimilating data prior to May 24 of 2020, removing as many as 16 weeks. We examined the results when omitting the first 0, 2, 4,…, 16 weeks. This period includes the peak of COVID-19 hospitalizations in New York City (NYC). In Figure R4, we present the clinical culture data and a schematic to highlight the removed weeks and, thus, the data used to inform the new inference testing. We include the hospitalization due to COVID-19 as well as the hospitalizations, discharges and admission to the hospital network for visual comparison of the dynamics of SARS-CoV2 transmission and the hospital traffic.

To investigate the impact of the importation rate we conducted inferences using the 3 values: the highest (red), the median (blue), and the lowest (purple) values we found in the literature and present these estimates in Figures R5 A, B and C respectively. To inspect the difference with the *original* estimate (the inference using the entire time series) we highlight the mean and the 95% CI of this inference with continuous and dashed lines, respectively. For all bacterial pathogens, the sensitivity to the weeks ignored depends on the level of importation rate . In the first row of each Figure in Figure R5 we first present the results for the changes in the nosocomial transmission rate and then those of the effective sensitivity .

**References**

1. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. Nature. 2008 Aug;454(7206):877–80.

2. Shaman J, Kohn M. Absolute humidity modulates influenza survival, transmission, and seasonality. Proc Natl Acad Sci USA. 2009 Mar 3;106(9):3243–8.

3. Shaman J, Kandula S, Yang W, Karspeck A. The use of ambient humidity conditions to improve influenza forecast. Pascual M, editor. PLoS Comput Biol. 2017 Nov 16;13(11):e1005844.

4. Pei S, Morone F, Liljeros F, Makse H, Shaman JL. Inference and control of the nosocomial transmission of methicillin-resistant Staphylococcus aureus. eLife. 2018 Dec 18;7:e40977.

5. Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science. 2020 May 1;368(6490):489–93.

6. Baracchini T, King AA, Bouma MJ, Rodó X, Bertuzzo E, Pascual M. Seasonality in cholera dynamics: A rainfall-driven model explains the wide range of patterns in endemic areas. Advances in Water Resources. 2017 Oct;108:357–66.

7. Ionides EL, Nguyen D, Atchadé Y, Stoev S, King AA. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. Proc Natl Acad Sci USA. 2015 Jan 20;112(3):719–24.

8. Ionides EL, Ning N, Wheeler J. An iterated block particle filter for inference on coupled dynamic systems with shared and unit-specific parameters [Internet]. arXiv; 2022 [cited 2024 Sep 4]. Available from: http://arxiv.org/abs/2206.03837

9. Ionides EL, Asfaw K, Park J, King AA. Bagged Filters for Partially Observed Interacting Systems. Journal of the American Statistical Association. 2023 Apr 3;118(542):1078–89.

10. Katzfuss M, Stroud JR, Wikle CK. Ensemble Kalman Methods for High-Dimensional Hierarchical Dynamic Space-Time Models. Journal of the American Statistical Association. 2020 Apr 2;115(530):866–85.

11. Schneider T, Dunbar ORA, Wu J, Böttcher L, Burov D, Garbuno-Inigo A, et al. Epidemic management and control through risk-dependent individual contact interventions. Moreno Y, editor. PLoS Comput Biol. 2022 Jun 23;18(6):e1010171.

12. Ionides EL, Breto C, King AA. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences. 2006 Dec 5;103(49):18438–43.