Basic reproduction number for pandemic *Escherichia coli* clones is comparable to typical pandemic viruses

Fanni Ojala[a,1], Henri Pesonen[b,1], Rebecca A. Gladstone[b], Tommi Mäklin[c], Gerry Tonkin-Hill[b,d,e,f], Pekka Marttinen[a,2], and Jukka Corander[b, c, g, h, 2, 3]

This manuscript was compiled on May 9, 2024

Extra-intestinal pathogenic *Escherichia coli* (ExPEC) ubiquitously colonize the human gut and are clinically the most significant bacterial species causing urinary tract infections and bacteremia in addition to contributing to meningitis in neonates. During the last two decades, new *E. coli* multi-drug resistant (MDR) clones such as ST131, particularly clades C2 and C1, have spread globally, as has their less resistant counterpart ST131-A and phylodynamic modeling has indicated exponential growth in their populations during the early 2000s. However, it remains unknown how their transmission dynamics compare to viral epidemics and pandemics in terms of key epidemiological quantities such as the basic reproduction number (R_0). Estimation of R_0 for opportunistic pathogenic bacteria poses a difficult challenge compared to viruses causing acute infections, since data on *E. coli* infections accumulate with a much longer delay, even in the most advanced public health reporting systems. Here, we develop a compartmental model for asymptomatic gut colonization and onward transmission coupled with a stochastic epidemiological observation model for bacteremia and fit the model to Norwegian national *E. coli* disease surveillance and bacterial population genomics data for 2002-2017. Approximate Bayesian Computation (ABC) leveraged by the ELFI (Engine for Likelihood-Free Inference) Python package is used to infer R_0 for the pandemic ST131-A and MDR ST131-C2 clades. The resulting estimates for ST131-A and ST131-C2 are 2.00 and 1.35, respectively, where the ST131-A estimate is comparable to the basic reproduction number of typical pandemic influenza. In summary our results provide a fundamental advance in understanding the effective transmission rate of novel *E. coli* clones.

Bacterial transmission | Basic reproduction number | *Escherichia coli* |

Significance Statement

Similar to viruses, novel clones of pathogenic bacteria have the potential for pandemic spread and may impose a considerable public health burden of infections. Unlike viruses, opportunistic pathogenic bacteria are mostly carried asymptomatically and cause delayed infections making it difficult to infer their transmission rates in the general population. We estimated a key epidemiological quantity, the basic reproduction number, for two clades of the pandemic *Escherichia coli* ST131 clone and show that their transmission rates differ markedly and are comparable to typical pandemic viruses, such as influenza. This provides a significant advance in understanding how rapidly new clones of *E. coli* can establish themselves, highlighting the need to identify the genetic factors that drive transmission efficiency.

As evidenced by extensive research on emerging pathogens over the last couple of decades, such as Ebola, H1N1 influenza, SARS, MERS and SARS-CoV-2, epidemiological modeling of their transmission rates provides a key piece of information that serves many purposes, including the forecasting of future spread and design of interventions. For a critical overview of different methods for inferring R_0, and their reliability see (1–5). Different compartmental models can be used to model the spread of an infectious disease in a population, out of which the most basic one is the Susceptible-Infected-Recovered (SIR) model (for details see the above references). For emerging pathogenic viruses the incubation period tends to be short, typically a matter of days, which translates to an opportunity to trace and count cases by reporting symptomatic individuals in a population on a given day or week of an ongoing epidemic. These data represent the most fundamental information needed in modeling the spread of the pathogen in such a setting, irrespective of whether deterministic or stochastic models are considered. For bacteria the situation is much more complicated and consequently R_0 is rarely inferred. The notable exceptions where R_0 has been estimated are highly virulent bacteria causing local outbreaks, such as *Neisseria meningitidis* (6) and TB (7), or bacterial strains in more closed ecological settings such as animal production facilities where their spread can be traced for example using seroconversion data (8). Outside particularly vulnerable populations, e.g. hospital patients, common opportunistic pathogenic bacteria are frequently carried asymptomatically, potentially for months or years, and reports of infections caused by them can occur with a considerable time lag after the colonization event. Therefore determining the length of time between the acquisition of the pathogen to an infection or transmission is not feasible even in the most advanced and systematic public health surveillance systems.

Author affiliations: [a]Department of Computer Science, Aalto University, Espoo, Finland; [b]Department of Biostatistics, University of Oslo, Oslo, Norway; [c]Helsinki Institute for Information Technology HIIT, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland; [d]Peter MacCallum Cancer Centre, Melbourne, Australia; [e]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Australia; [f]Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Australia; [g]Parasites and Microbe, Wellcome Sanger Institute, Cambridgeshire, UK; [h]Department of Genetics, University of Cambridge, Cambridge, UK

Please provide details of author contributions here.

Please declare any competing interests here.

[1,2]F.O.contributed equally to this work with H.P. P.M. contributed equally to this work with J.C. .

[3]To whom correspondence should be addressed. E-mail: jukka.corandermedisin.uio.no

DRAFT

Bacteria causing "hidden" pandemics can be broadly categorized into two distinct types, those which are strongly associated with nosocomial infections and are mainly spread through healthcare networks and those which predominantly spread in the community as part of the commensal microbiota of the gut, skin or the nasopharynx (9, 10). Extra-intestinal pathogenic *Escherichia coli* (ExPEC) represent the latter category and are among the clinically most significant causes of human bacterial infections, such as bacteremia and urinary tract infections (11). During the last decades several so-called pandemic clones of ExPEC have emerged and these are typically characterized by both high levels of virulence and antibiotic resistance (12, 13). For two particularly famous clades belonging to the pandemic ST131 clone, ST131-A and ST131-C2, previous work has inferred the timing and rates of their expansions using a Bayesian phylodynamic approach which provides an estimate of the effective population over time (14). Using systematic genomic surveillance of *E. coli* bloodstream infections (BSIs) in Norway over 2002-2017, it was shown that these clones likely expanded in parallel in the host population around early 2000s but also that ST131-A appeared to have markedly higher dissemination success than ST131-C2. However, since systematic surveillance of asymptomatic colonization rates of *E. coli* clones has been missing, it has not been possible to turn the disease surveillance data into similar epidemiological insight as captured by R_0. Recently, an advance in this direction was made by Mäklin et al. who inferred strain-level colonization competition dynamics from shotgun metagenomics data for a large neonatal birth cohort in the UK and for several bacterial pathogens, including *E. coli* (15). By combining these data with temporally matched genomic *E. coli* disease surveillance from the UK, it was possible to infer the relative success of colonization vs infection for the two clades ST131-A and ST131-C2 (16). Further, this information can be merged with the comprehensive population level bloodstream infection statistics available for Norway and the annual incidences of ST131-A and ST131-C2 in Norway based on the published genomic surveillance results. Here, we harness the unique opportunity presented to us by the combination of these data, and develop a compartmental model for the asymptomatic colonization and transmission of these two *E. coli* clades of ST131 in Norway and couple this with a stochastic observation model for bloodstream infections, to estimate R_0 using simulation-based inference. Comparison of their values provides useful epidemiological insight and allows relating them more directly to other pandemic pathogens in terms of the rate of their spread.

Results

In order to link the asymptomatic spread of *E. coli* in the host population with the reported bloodstream infections from the national surveillance in Norway, we developed a hybrid model with both a deterministic and a stochastic component. The model structure is illustrated in Figure 1. Our model consists of an unobserved compartmental SIR-like colonization model for the host population at the national level and an observation model which links the colonization with observed annual bloodstream infection incidence for clades A and C2. The observation model further uses estimated odds ratios for each clade. These compare the frequency of asymptomatic colonisation to the frequency of bloodstream infection in a temporally and geographically matched host population (UK) from which data based on representative sampling of hosts were available (15),(16). Full details of the model are provided in the Materials and Methods section.

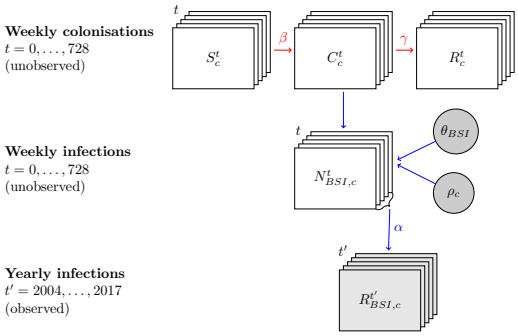

Fig. 1. Model structure. The first row shows the latent population colonization model, implemented by simulating outcomes from a compartmental SIR-like model. The compartments S_c^t, C_c^t and R_c^t are the numbers of susceptible, colonized, and removed individuals in week t for clade c. The transmission coefficient β and the recovery rate γ describe the flow between the compartments. The observation model is visualized below the colonization model, comprising the weekly simulated BSI counts $N_{\text{BSI,c}}^t$ with the aggregated yearly BSI infection counts $R_{\text{BSI,c}}^{t'}$, for which observed values are available from the national surveillance system. In the observation part, parameter ρ_c is the estimated relative invasiveness of clade c, θ_{BSI} is the fraction of the population experiencing an *E. coli* BSI caused by any clade, and α is a convolution parameter reflecting the expected temporal lag between colonisation and BSI events.

Table 1. Mean, median and 95% credible intervals for R_0, β and γ

Parameter	Mean	Median	95% CI
Clade A			
R_0	2.00	1.75	(1.38,4.50)
β	0.070	0.068	(0.053, 0.089)
γ	0.039	0.039	(0.011, 0.063)
Clade C2			
R_0	1.35	1.18	(1.10, 3.30)
β	0.13	0.13	(0.059, 0.18)
γ	0.10	0.11	(0.013, 0.16)

Figure 2 shows the marginal and joint posterior distribution of the transmission coefficient β and the recovery rate γ for the two *E. coli* ST131 clades. Posterior distributions of the corresponding R_0 parameters are correspondingly visualized in the leftmost panel of Figure 2. Posterior means, medians and 95% posterior credible intervals for the three parameters are presented in Table 1. The posterior distributions indicate much higher R_0 for clade A, translating to markedly better transmission efficiency in the underlying population.

The model fit to data and the 50/95% credible intervals for predicted BSI counts are visualized in Figure 3, showing that the hybrid model structure adequately captures the major differences between the two clades. Note that the variation during the early observation period is ignored due to the one-peak structure of the SIR-like model, which

DRAFT

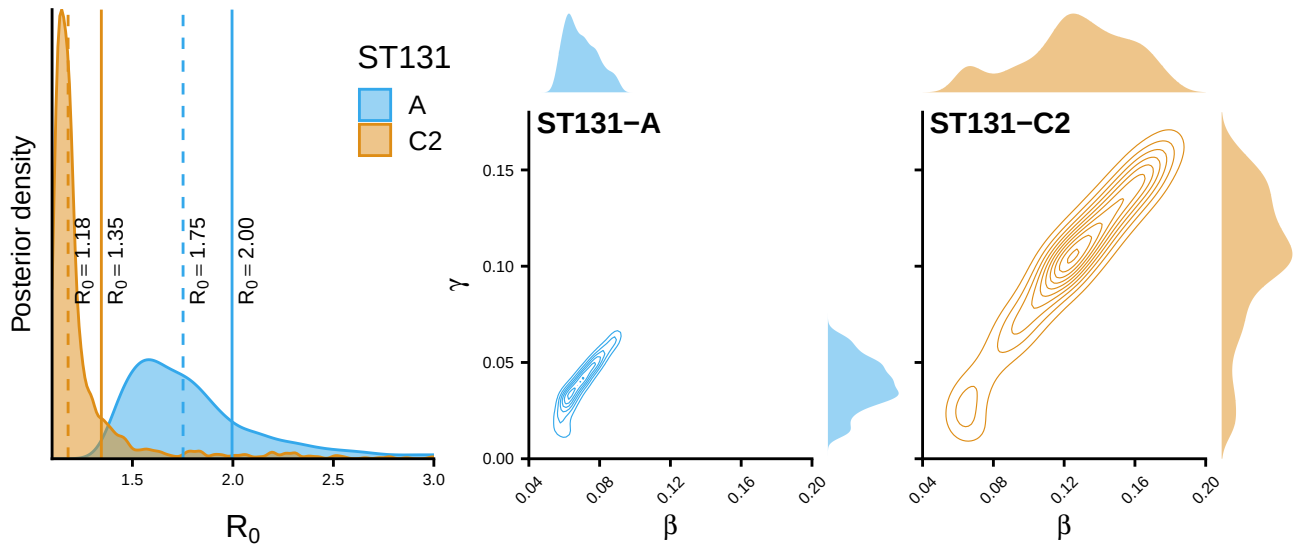

Fig. 2. The leftmost panel shows the posterior distributions of R_0 for ST131-A and ST131-C2, with median point estimate of R_0 highlighted by a dashed line and the mean by a solid line. The two other panels display the joint and marginal posterior distributions of β (x-axis) and γ (y-axis) parameters for both clades.

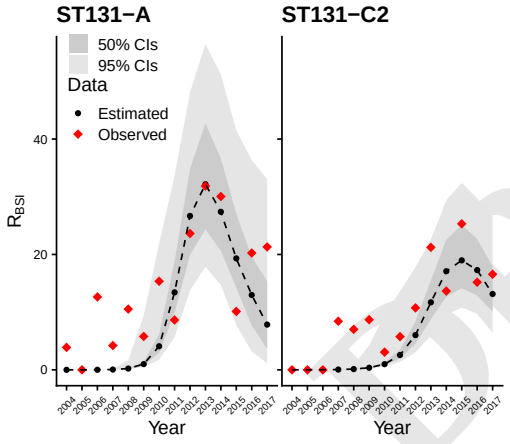

Fig. 3. Predictive fit of the model to the observed data. The red diamonds represent observed BSI counts and the black dots connected by a dashed line represent the median fit for each year.

does not take the initial randomness in an epidemic into account. The corresponding estimated colonization curves are visualized in the Supporting Information Figure S1 for clades A and C2.

Discussion

While R_0 represents a key epidemiological characteristic that is typically rapidly inferred for emerging pathogenic viruses, its quantification has remained far more elusive for pathogenic bacteria causing opportunistic infections. The primary reasons behind this difficulty are the lack of systematic high-resolution surveillance of asymptomatic colonisation rates for almost any bacterial species, the temporal lag with which infection cases tend to occur, and the need to have a system in place for whole-genome sequencing of infection isolates from sufficiently representative surveillance. Here we were able to fill this knowledge gap by combining systematic population genomics based surveillance of *E. coli* bloodstream infection isolates in both Norway and the UK, with demographics and annual national-level mandatory reporting statistics of these infections in Norway. Furthermore, we had the opportunity to use recently estimated odds ratios for each considered clade of *E. coli* (16), which compare the frequency of asymptomatic colonisation to the frequency of bloodstream infection in a temporally matched host population in the UK.

By using all the above-mentioned data sources, we built a hybrid deterministic-stochastic model that first simulates the asymptomatic spread of a novel bacterial clone in a population and then feeds this information into an observation model that generates infection cases at the desired temporal resolution. Unlike the basic epidemiological models, our approach results in an intractable likelihood, calling for simulator-based inference methods, such as ABC (17), to calibrate the model parameters. Recently, Pesonen et al. demonstrated the extended possibilities offered by ABC inference for simulator-based modelling of R_0 when the transmission process exhibits latent heterogeneity (18), using a model derived for the Ebola virus by Britton and Tomba (19). Despite our focus on *E. coli*, we believe that the current study can serve as a proof-of-concept for modelling expansions of other bacteria with pathogenic potential and thus extend the reach of infectious disease epidemiology to further novel applications.

Compartmental models for transmission studies are generally built from assumptions that might not hold in reality, such as homogeneous mixing of host population, mass action incidence to determine rate of new infections, effective contacts (contact always leads to successful transmission)

DRAFT

and unchanging demographics on the relevant timescale (20). Despite these simplifying assumptions, even basic compartmental models have been widely popular and successfully used for a large number of studies in infectious disease epidemiology.

An alternative compartmental model for bacterial colonisation of the type we consider here would be the SIS model (20). Here, re-colonisation of the same individual is possible and there is no resistance acquired against colonisation, leading to exclusion of the recovered compartment. However, since we are operating on relatively short timescales of a few years at a national level, using an SIR-like colonisation model is more realistic. We did not include any explicit age structure in our model, although its use could be motivated by the higher incidence of BSI in the elderly population. Nevertheless, it is justifiable to consider that the expanding clades are equally capable of colonising hosts irrespective of their age, translating to model fits where the relative difference between the clades is robustly quantified. To better capture the variability in incidence at the beginning of the clade expansion, a stochastic branching process based on contact networks could be used in addition to the deterministic compartmental model (20). In this regard, the deterministic nature of the SIR-like model for colonisation is visible in our results as a lack of fit in the earliest phase of expansion, while the shape of the peaking incidence is still well covered (Figure 3). Such extensions of our model could be worth considering in future work.

Our results indicate a much higher R_0 for clade A (posterior mean/median 2.00/1.75) compared with clade C2 (posterior mean/median 1.35/1.18), which translates to a markedly higher transmission efficiency of A, at least during the early phase of their global spread. This is well aligned with the earlier Bayesian phylodynamic inference based finding of significantly higher increase in the effective population size of A clade during the mid-2000s (14). Importantly, these two clades of the ST131 clone have diverged very recently and are still closely related in terms of their core genomes (14, 21, 22). It is thus remarkable how different they are in terms of their estimated ability to colonise new hosts, which highlights the need for further research on whether this difference can be explained by genetic diversity in their core genomes, such as in metabolic capacity, or by variable accessory genome content (or both). These topics have been a subject of intensive recent research efforts in *E. coli* (21–27), but thus far no clear conclusions have been reached concerning main genetic drivers underlying phenotypic variation at this level of relatedness. Improved *in vitro* and *in vivo* experiments combined with high-resolution comparative population genomics and further modelling would offer a promising strategy towards resolving the likely polygenic basis of the difference in transmission efficiency.

Materials and Methods

Model and inference overview. Our modelling framework, illustrated in Figure 1, consists of two components: a *colonisation model*, represented by the standard compartmental SIR (susceptible-infected-recovered) model, and an *observation model* to connect the simulated colonisations to the incidence of bloodstream infections (BSI). The specific *E. coli* clade whose dynamics is being considered is denoted by c. In the SIR-like model, the I compartment is replaced by C_c (colonisation) that represents the number of individuals colonised by a clade c. The parameters of interest in the colonisation model are the transmission coefficient β and the recovery rate γ. The reproduction number R_0 characterising an expansion of a clade is defined as the ratio of β and γ:

$$R_0 = \frac{\beta}{\gamma}. \qquad [1]$$

The challenge in inferring parameters of the colonisation model is that colonisation is seldom under surveillance with sufficient genomic resolution. Consequently, we need an observation model to connect the unobserved colonisation cases that can be simulated from the colonisation model to the number of BSI cases, which are assumed to be observed in systematic surveillance. Together the colonisation and observations models then allow us to quantify the number of BSI cases that would be expected under different combinations of β and γ. We can thus infer the posterior distributions of these key parameters using ABC by identifying the β and γ such that the simulated BSI counts reproduce the observed BSI counts as accurately as possible. The model is fitted separately for the two ST131 clades of interest, A and C2, to compare their respective parameters.

Colonisation model. The SIR-like model for the colonisation by clade c is defined by

$$\begin{aligned} dS_c/dt &= -\beta S_c C_c, \\ dC_c/dt &= \beta S_c C_c - \gamma C_c, \\ dR_c/dt &= \gamma C_c. \end{aligned} \qquad [2]$$

The S_c compartment is the number of individuals at risk of being colonized by the clade c, while the C_c compartment is the number of individuals colonized by the clade c. The structure of the colonization model is visualized on the first row of Figure 1, where the horizontal arrows represent the flow of individuals from one compartment to the other. In practice, we discretize time and simulate transitions between compartments in steps of one week. This results in weekly counts of individuals in the different compartments, S_c^t, C_c^t, and R_c^t. Here, t indexes the week, with $t = 1, \ldots, 728$, corresponding to an observation period of 14 years ($\approx 52 \times 14$ weeks), from 2004 to 2017.

Observation model. The observation model connects the simulated, unobserved colonisation to observed bloodstream infections. To define the observation model, let ρ_c denote the invasiveness of the clade of interest, c, relative to the other clades:

$$\rho_c = \frac{N_{\mathrm{BSI},c}}{C_c} / \frac{N_{\mathrm{BSI},0}}{C_0}, \qquad [3]$$

where $N_{\mathrm{BSI},c}$ is the number of BSI cases caused by clade c, $N_{\mathrm{BSI},0}$ is the number of BSI cases not associated with clade c, and, similarly, C_c, C_0 denote the numbers of colonization by clade c or any other clade, respectively. In practice the quantities in Eq. (3) are defined for each week t, but the index t is here suppressed for clarity. The relative invasiveness value ρ_c is obtained from literature (15) (see below for details), and $N_{\mathrm{BSI},0}$ is calculated as

$$N_{\mathrm{BSI},0} = N\theta_{\mathrm{BSI}} - N_{\mathrm{BSI},c}, \qquad [4]$$

where N is the population size, and θ_{BSI} is the estimated proportion of the population with a BSI infection in the given week. As an estimate of θ_{BSI} we use the weekly average number of BSI cases across the whole study period:

$$\theta_{\mathrm{BSI}} = \frac{n}{728}, \qquad [5]$$

where n is the total number of BSI infections in 728 weeks, obtained from the annual Norwegian national bacterial infection surveillance reports for the current study. Substituting Equations 4 and 5 into Eq. (3) and solving for $N_{\mathrm{BSI},c}$ yields

$$N_{\mathrm{BSI},c} = \frac{\rho_c C_c}{(C - C_c) + \rho_c C_c} N\theta_{\mathrm{BSI}}, \qquad [6]$$

which shows how the number of colonised individuals, C_c, is connected with the number of BSI cases, $N_{BSI,c}$, under the

DRAFT

observation model. Some intuition for this connection can be gained by considering the special case of clade c being as invasive as other clades, in which case $\rho_c = 1$, and Eq. (6) reduces to

$$N_{\mathrm{BSI},c} = \frac{C_c}{C} N\theta_{\mathrm{BSI}}. \quad [7]$$

In this special case the proportion of BSI infections caused by clade c is simply equal to the proportion of colonisation by clade c among all *E. coli* colonisations. Eq. (6) adjusts this proportion to take into account the invasiveness of the clade c, such the proportion of infections can be larger or smaller than what is implied by the proportion of colonisation. In practise we use $C = C_0 + C_c$ and assume that $C = N$, i.e., everybody in the population is colonised by some *E. coli* which is justified by considerable evidence in the literature (11).

The observed data on BSI incidence are available as yearly aggregates, but the discretized colonization model produces weekly colonization counts. To compare the simulation and observations, we consequently aggregate the weekly BSI counts by clade c, $N^t_{\mathrm{BSI},c}$, where the index of the week $t = 1, \ldots, 728$, into a yearly count $R^{t'}_{\mathrm{BSI},c}$:

$$R^{t'}_{\mathrm{BSI},c} = \sum_{t \in T_{t'}} N^t_{\mathrm{BSI},c}, \quad [8]$$

where $T_{t'}$ is the set of all weeks in the t'-th year. This summation is repeated for each year $t' \in \{2004, \ldots, 2017\}$.

To increase the flexibility of our model, we allow a time delay between the unobserved increase in colonisation frequency and the observed infection cases. This added flexibility permits the model to better adapt to the underlying generative process where the majority of infection cases do not appear immediately upon the peak in colonisation but after some amount of time and in particular after spreading among the elderly. To allow for such delays to by represented by the observation model, we use exponential smoothing (a.k.a exponentially weighted moving average) (28, 29) for $R^{t'}_{\mathrm{BSI},c}$ with a convolution parameter α:

$$\tilde{R}^{t'}_{\mathrm{BSI},c} = \alpha R^{t'}_{\mathrm{BSI},c} + (1-\alpha)\tilde{R}^{(t'-1)}_{\mathrm{BSI},c}, \; t' = 2005, \ldots, 2017, \quad [9]$$

where $\tilde{R}^{t'}_{\mathrm{BSI},c}$ denotes the smoothed number of BSI cases, $\tilde{R}^{2004}_{\mathrm{BSI},c} = R^{2004}_{\mathrm{BSI},c}$, and $0 < \alpha < 1$. Figure S2 demonstrates the impact of the smoothing for different values of α. During inference the posterior distribution of α is integrated over to account for the uncertainty about its value.

Data. An unbiased longitudinal genomic survey of BSI isolates in Norway collected between 2002-2017 were used to estimate the incidence of ST131 clades in BSI per 1,000,000 individuals in Norway between years from 2004 to 2017 (14). The incidence/1,000,000 for total *E. coli* BSI was available from the Norwegian Surveillence of antimicrobial resistance (NORM) reports (30). This incidence is used to calculate the value of θ_{BSI} in Eq. (5). We previously estimated the relative invasiveness of different *E. coli* ST131 clades, specified with odds ratios, by comparing the frequency of each clade in geographically matched 2014-2017 neonatal gut colonization genomic data and longitudinal BSI 2003-2017 genomic data from the UK, after the emergence of ST131 and ST69 when a stable population equilibrium was reached (16), (15). These relative invasiveness values of different *E. coli* clades were used to estimate ρ_c for the observation model.

Statistical analysis. The likelihood function is intractable for the defined hybrid deterministic-stochastic model. However, since it can be implemented efficiently as a forward simulation, we use the Approximate Bayesian Computation (ABC) for likelihood-free inference to calibrate the model parameters in the light of data (17). The forward simulation is sufficiently fast for the use of the standard rejection sampling ABC algorithm, to accurately approximate the posterior distribution of the model parameters. The prior distributions for the transmission coefficient β, the recovery rate γ and the convolution parameter α are specified as follows. First, the recovery rate γ is the fraction of colonised individuals leaving the colonised compartment (20) and consequently the quantity $\frac{1}{\gamma}$ is the mean colonization duration (31), which was used to construct an informative Gamma prior based on the average duration of *E. coli* colonization of travellers returning home, estimated to be 30 days (32). Secondly, the R_0 value should be greater than 1 for a population expansion to occur in the first place and it should further have a biologically plausible upper bound. We assume R_0 to be between 1.01 and 6.01. Consequently, by taking into account the relationship between β, γ, and R_0 in Eq. (1), we define the distribution of β conditionally, such that the joint prior for (γ, β) becomes

$$\gamma \sim \Gamma(1.5, 6), \quad [10]$$

$$\beta|\gamma \sim \mathcal{U}(1.01\gamma, 6.01\gamma), \quad [11]$$

where $\Gamma(\theta_s, \theta_r)$ is the gamma distbribution with shape θ_s and rate θ_r, and $\mathcal{U}(\theta_l, \theta_u)$ is the uniform distribution on the interval (θ_l, θ_u). Finally, we assign a non-informative prior on the convolution parameter $\alpha \in (0, 1)$ in Eq. (9) as

$$\alpha \sim \mathcal{U}(0, 1). \quad [12]$$

To implement the algorithm, we need to introduce 1) summary statistics to reduce the dimension of the time-series of the yearly BSI observations, 2) a discrepancy function that allows us to compare the simulated synthetic data summaries with the summaries of the observed data, and 3) a threshold parameter that reflects tolerable error between the summarized synthetic and observed data, to accept the parameter that generated the simulation as a draw from the posterior. We constructed four summary statistics, $S_1, \ldots, S_4$, to compare how the observed and simulated BSI incidences evolve in time. S_1 is defined as the maximum BSI incidence over the years within the interval $[2004, 2017]$. S_2 is the number of years from the beginning of the simulation (year 2004) to the maximum BSI incidence. S_3 is the BSI incidence one year before the maximum BSI incidence occurs. If the maximum is achieved in the year 2004, we set $S_3 = S_1$. S_4 is the BSI incidence one year after the maximum BSI incidence. If the maximum is achieved in the year 2017, we set or $S_3 = S_1$. We apply $\log(x+1)$-transformation on the summaries S_1, S_3 and S_4 because the scale of these BSI count -based summaries, which can vary between 0 and 1,000,000 (population size used in the simulation), is very different from the scale of the peaking time summary S_2 (ranging between 1 and 14 years). The Euclidean distance was used to compare the observed and simulated data summaries.

For the rejection ABC we generated 2,000,000 joint random samples from the prior distribution, and obtained 10,000 samples as an approximation to the posterior distribution. During post-processing we observed that the discrepancies were almost constant for simulations where the expansion did not occur at all, and accordingly we lowered the acceptance threshold to prevent parameters of these simulations from entering the sampled posterior. The model was fitted separately for clades A and C2. The simulator model and the ABC inference pipeline was implemented in open-source likelihood-free inference software package, Engine for Likelihood-free Inference (ELFI) (33). The complete analysis pipeline written in Python is available in GitHub (https://github.com/fanoja/Ecoli-R0/tree/master).

ACKNOWLEDGMENTS. Research in this publication was supported by the Trond Mohn Foundation (BATTALION grant) and Academy of Finland (EuroHPC grant, Flagship programme: Finnish Center for Artificial Intelligence FCAI, and grants 358246, 352986) and EU (H2020 grant 101016775 and NextGenerationEU).

1. K Dietz, The estimation of the basic reproduction number for infectious diseases. *http://dx.doi.org/10.1177/096228029300200103* **2**, 23–41 (1993).

DRAFT

2. JM Heffernan, RJ Smith, LM Wahl, Perspectives on the basic reproductive ratio. *J. The Royal Soc. Interface* **2**, 281–293 (2005).
3. M O'Driscoll, C Harry, CA Donnelly, A Cori, I Dorigatti, A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics, with implications for the current coronavirus disease 2019 (covid-19) pandemic. *Clin. Infect. Dis.* **73**, e215–e223 (2021).
4. W Ali, CE Overton, RR Wilkinson, KJ Sharkey, Deterministic epidemic models overestimate the basic reproduction number of observed outbreaks. *Infect. Dis. Model.* (2024).
5. PL Delamater, EJ Street, TF Leslie, YT Yang, KH Jacobsen, Complexity of the basic reproduction number (r0). *Emerg. Infect. Dis.* **25**, 1 (2019).
6. AL Presti, et al., Estimates of the reproductive numbers and demographic reconstruction of outbreak associated with c:p1.5–1,10–8:f3–6:st–11(cc11) neisseria meningitidis strains. *Infect. Genet. Evol.* **84**, 104360 (2020).
7. Y Ma, CR Horsburgh, LF White, HE Jenkins, Quantifying tb transmission: a systematic review of reproduction number and serial interval estimates for tuberculosis. *Epidemiol. & Infect.* **146**, 1478–1494 (2018).
8. WW Laegreid, JE Keen, Estimation of the basic reproduction ratio (r0) for shiga toxin-producing escherichia coli o157:h7 (stec o157) in beef calves. *Epidemiol. & Infect.* **132**, 291–295 (2004).
9. A Henderson, GR Nimmo, Control of healthcare- and community-associated mrsa: recent progress and persisting challenges. *Br. Med. Bull.* **125**, 25–41 (2018).
10. SJ van Hal, et al., The interplay between community and hospital enterococcus faecium clones within health-care settings: a genomic analysis. *The Lancet Microbe* **3**, e133–e141 (2022).
11. M Donnenberg, *Escherichia coli: Pathotypes and Principles of Pathogenesis, Edition 2* ed. M Donnenberg. (Academic Press), 2 edition, (2013).
12. LW Riley, Pandemic lineages of extraintestinal pathogenic escherichia coli. *Clin. microbiology infection : official publication Eur. Soc. Clin. Microbiol. Infect. Dis.* **20**, 380–390 (2014).
13. JD Pitout, G Peirano, L Chen, R DeVinney, Y Matsumura, Escherichia coli st1193: Following in the footsteps of e. coli st131. *Antimicrob. agents chemotherapy* **66** (2022).
14. RA Gladstone, et al., Emergence and dissemination of antimicrobial resistance in escherichia coli causing bloodstream infections in norway in 2002–17: a nationwide, longitudinal, microbial population genomic study. *The Lancet Microbe* **2**, e331–e341 (2021).
15. T Mäklin, et al., Strong pathogen competition in neonatal gut colonisation. *Nat. Commun. 2022 13:1* **13**, 1–13 (2022).
16. AK Pöntinen, et al., Modulation of multidrug-resistant clone success in escherichia coli populations: a longitudinal, multi-country, genomic and antibiotic usage cohort study. *The Lancet Microbe* **5**, e142–e150 (2024).
17. J Lintusaari, MU Gutmann, R Dutta, S Kaski, J Corander, Fundamentals and Recent Developments in Approximate Bayesian Computation. *Syst. Biol.* **66**, e66–e82 (2017).
18. H Pesonen, et al., Abc of the future. *Int. Stat. Rev.* **91**, 243–268 (2023).
19. T Britton, GS Tomba, Estimation in emerging epidemics: biases and remedies. *J. Royal Soc. Interface* **16** (2019).
20. F Brauer, C Castillo-Chavez, Z Feng, *Mathematical Models in Epidemiology*. (Springer Science+Business Media, LLC), (2019).
21. A McNally, et al., Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* **12** (2016).
22. A McNally, et al., Diversification of colonization factors in a multidrug-resistant escherichia coli lineage evolving under negative frequency- dependent selection. *mBio* **10** (2019).
23. CH Connor, et al., Multidrug-resistant e. coli encoding high genetic diversity in carbohydrate metabolism genes displace commensal e. coli from the intestinal tract. *PLoS Biol.* **21** (2023).
24. RJ Hall, et al., Gene-gene relationships in an escherichia coli accessory genome are linked to function and mobility. *Microb. Genomics* **7** (2021).
25. EA Cummins, RJ Hall, C Connor, JO McInerney, A McNally, Distinct evolutionary trajectories in the escherichia coli pangenome occur within sequence types. *Microb. Genomics* **8**, 000903 (2022).
26. F Rousset, et al., The impact of genetic diversity on gene essentiality within the escherichia coli species. *Nat. Microbiol. 2021 6:3* **6**, 301–312 (2021).
27. M Micaletto, S Fleurier, S Dion, E Denamur, I Matic, The protein carboxymethyltransferase–dependent aspartate salvage pathway plays a crucial role in the intricate metabolic network of escherichia coli. *Sci. Adv.* **10**, 767 (2024).
28. RG Brown, RF Meyer, DA D'esopo, The fundamental theorem of exponential smoothing. *Source: Oper. Res.* **9**, 673–687 (1961).
29. JS Hunter, The exponentially weighted moving average. *J. Qual. Technol.* **18**, 203–210 (1986).
30. U Nord-Norge, Norm - norsk overvåkingssystem for antibiotikaresistens hos mikrober (year?).
31. P van den Driessche, Reproduction numbers of infectious disease models. *Infect. Dis. Model.* **2**, 288 (2017).
32. MS Arcilla, et al., Import and spread of extended-spectrum β-lactamase-producing enterobacteriaceae by international travellers (combat study): a prospective, multicentre cohort study. *The Lancet Infect. Dis.* **17**, 78–85 (2017).
33. J Lintusaari, et al., Elfi: Engine for likelihood-free inference. *J. Mach. Learn. Res.* **19**, 1–7 (2018).

DRAFT