STOCHASTIC MODELLING AND INFERENCE IN ELECTRONIC HOSPITAL DATABASES FOR THE SPREAD OF INFECTIONS: *CLOSTRIDIUM DIFFICILE* TRANSMISSION IN OXFORDSHIRE HOSPITALS 2007–2010[1]

BY MADELEINE CULE AND PETER DONNELLY[2]

University of Oxford

The combination of genetic information with electronic patient records promises to provide a powerful new resource for understanding human disease and its treatment. Here we develop and apply a novel stochastic compartmental model to a large dataset on *Clostridium difficile* infection (CDI) in three Oxfordshire hospitals over a 2.5 year period which combines genetic information on 858 confirmed cases of CDI with a database of 750,000 patient records. *C. difficile* is a major cause of healthcare-associated diarrhoea and is responsible for substantial mortality and morbidity, with relatively little known about its biology or its transmission epidemiology. Bayesian analysis of our model, via Markov chain Monte Carlo, provides new information about the biology of CDI, including genetic heterogeneity in infectiousness across different sequence types, and evidence for ward contamination as a significant mode of transmission, and allows inferences about the contribution of particular individuals, wards or hospitals to transmission of the bacterium, and assessment of changes in these over time following changes in hospital practice. Our work demonstrates the value of using statistical modelling and computational inference on large-scale hospital patient databases and genetic data.

1. Introduction. The increasingly widespread linkage of electronic patient records, which document aspects of clinical care and of patient response, offers an unprecedented opportunity for *in silico* studies of human health and disease on very large scales. Analyses in these biomedical "big data" settings are also challenging on several levels, not least in the need to develop scalable statistical or stochastic models which capture central features of the application and yet are amenable to inference.

One particular opportunity is for the study of infectious disease transmission within hospitals. In spite of stringent hospital anti-infection protocols, hospital-acquired infections of bacteria such as methicillin-resistant *S*taphylococcus aureus (MRSA) and *Clostridium difficile* represent major public health challenges, are sources of considerable public anxiety, and incur very substantial costs to health-care systems. The combination of hospital records on patient location and move-ment within hospitals and their infection status, together with genetic information on the infecting bacteria, offer an opportunity to better understand both the basic biology of the pathogens and the epidemiology of infection transmission within hospitals.

Infection events can never be observed directly, and so must be inferred from circumstantial evidence, including patient proximity and similarities between the genetic types of their infecting pathogens. Our approach is to introduce a novel stochastic compartmental model for infection within, and transmission from, an in-dividual, and then to fit this model to multiple individuals in the available hospital record data. The modelling challenge is to capture enough of the features of the real infection and transmission process whilst still allowing tractable inference on large datasets. A major advantage of this framework is that it not only allows in-ference about "biological" parameters such as infectious rates, incubation periods and recovery times, but it also allows probabilistic inference about transmission events. Individually, these transmission inferences can potentially resolve partic-ular events, but collectively they allow assessment of the major modes of hospi-tal transmission. They also have a second important benefit. Changes in hospital practices aim to reduce hospital transmission of infections, but their efficacy is currently not easy to assess exactly because transmission events are unobserved. Approaches such as the one we develop here, which allow (probabilistic) infer-ence of transmission events, provide a much more direct route to assessing the consequences of changes in hospital policy and practice.

In this paper we focus on a particular infectious bacterium, *Clostridium difficile*, in the three main hospitals in Oxfordshire, UK. *C. difficile* is a Gram-negative obli-gate anaerobe carried in the gut of between 2% and 7% of adults and up to 30% of infants [McFarland et al. (1989), Rolfe (1988)]. Since its recognition as a human pathogen in the late 20th century [Bartlett et al. (1978)], healthcare-associated out-breaks of *C. difficile* infection (CDI) have been reported worldwide [Health Protec-tion Agency (2010)]. Attributable mortality is estimated at 8% [Karas, Enoch and Aliyu (2010)]. It is thought that symptoms, which range from diarrhoea to severe pseudomembranous colitis, result from the exposure to antibiotics of an individual carrying *C. difficile*: the antibiotics disrupt healthy gut flora and allow *C. difficile* to proliferate; the bacteria produce toxins, and when the size of the *C. difficile* pop-ulation in the gut increases, so does the amount of toxin, which in turn induces the symptoms. As well as antibiotic use, risk factors for developing CDI include ad-vanced age and extended periods of hospitalisation (which may partly be explained

by contact with other infectious CDI patients) [Loo et al. (2011)], with asymptomatic carriage suggested as a possible protective factor [Loo et al. (2011)]. In 2003, a highly pathogenic strain (known as variously as Ribotype 027/NAP1/ST1) emerged in Canada and spread rapidly [McDonald et al. (2005)], accounting for 55% of cases in the United Kingdom in 2007/8. Although the UK incidence has declined from its 2005 peak of 50,000 reported cases to approximately 23,000 cases in 2010 [Health Protection Agency (2009)] (with the prevalence of ST1 declining to 12%), CDI remains a significant problem, and identifying effective measures to control its spread remains a priority. Devising, implementing and assessing such strategies is challenging, in part because its biology and transmission epidemiology is not fully understood, and in part because transmission events are not directly observed.

Oxford University Hospitals (OUH) NHS Trust provides more than 90% of all inpatient care, including all acute services, for approximately 600,000 people in Oxfordshire, UK. Briefly, there are 3 main hospital sites with, respectively, 77,000, 14,000 and 13,000 admissions/year. Further details are given in Table 1. During the study period of 1 September 2007 to 31 March 2010, the routine microbiology laboratory tested all samples sent for *C. difficile* diagnostics from OUH, local doctors' practices and smaller specialist hospitals in the county. Local policy was that any OUH inpatient with diarrhoea should have samples sent for testing for CDI using an enzyme immunoassay (EIA) for *C. difficile* toxins A and B (Meridian Bioscience Inc., Cincinatti, Ohio). Most (96%) EIA positive samples were retrieved for confirmatory testing by culture, a gold-standard method for diagnosis [Planche et al. (2008)]. If culture established the presence of *C. difficile*,

TABLE 1

Details of the 3 Oxfordshire hospitals included in this study, which ran from September 2007 *to March* 2010 *inclusive* (31 *months*). *The Oxford University Hospitals* (*OUH*) *NHS Trust provides all acute services for* 600,000 *people in Oxfordshire. We included in this study all patients with at least one OUH admission. Of these*, 858 *had at least one C. difficile infection* (*CDI*) [*defined as an enzyme immunoassay* (*EIA*) *positive, culture positive faecal sample*]. *All CDI were strain typed using MLST. There were an additional* 73 *CDI cases who had no overnight admissions to the John Radcliffe*, *Churchill or Horton hospitals in OUH during the study period. These were excluded from the analysis*

Hospital	Wards	Overnight admissions	Patients	CDI cases
John Radcliffe	100	193,641	99,149	639
Churchill	48	36,978	19,098	277
Horton	21	33,800	24,692	257
Miscellaneous[1]	36	2514	2439	0
Total	163	269,419	131,597	858
OUH catchment	–	–	~600,000	931

[1]Including "take" rounds where patients are admitted to hospital.

genetic information on the isolated bacteria was collected by genotyping them using multi-locus sequence typing (MLST), a robust and moderately discriminating genetic typing system which in effect reads the DNA sequence in seven prespecified small regions of the *C. difficile* genome [Griffiths et al. (2010)]. The result of this genotyping is the classification of the bacteria into so-called *s*equence types (ST). As of 2016/12/18, there are 360 MLSTs of *C. difficile* recorded in the *C. difficile* MLST database http://pubmlst.org/cdifficile/. In our dataset there are few instances of mixed infections, where bacteria from more than one MLST sequence type are recovered from an individual, and, as described below, these had limited impact on inference. A total of 858 individuals who had at least one overnight hospital stay were confirmed culture-positive during the study period.

Admission records and microbiology laboratory records were electronically linked and anonymized [Finney et al. (2011)] within the Infections in Oxfordshire Research Database (IORD) approved by the Oxford Research Ethics Committee (09/H0606/85) and the National Information Governance Board (5-07(a)/2009). Amongst other information, the database records all admissions of each patient to the hospital together with their subsequent movement around the hospital between different wards, and contains a total of around 750,000 patient records which cover the study period.

Figure 5 aims to give a sense of the complexity and structure in the dataset by illustrating a subset of those parts of the database which are relevant to our study. (The figure also represents transmission events inferred under our model; see below.) For each patient the database records the periods of time for which they were in the hospital system and the specific wards in which they were housed during each stay. For individuals who tested positive for CDI, the date of the positive test, and the associated MLST, is also recorded. The figure only shows hospital stay information for those individuals (each of which is represented horizontally) who tested positive with a particular subset of the MLST types during the study period. [See Figure 1 of Cule and Donnelly (2017) for all such individuals.] For simplicity, ward information is not shown in the figure. The full dataset can be thought of as a much larger set of timelines like those in the figures, with one for each of the 131,597 patients who visited at least one of the hospitals over the 31 month study period, augmented by information as to which ward they are in at each time. Although none of these individuals is represented in the figure, we note below that knowledge of transmission events which could have, but in fact did not, occur is informative about infection rates and epidemiology.

A published informal analysis of this data [Walker et al. (2011)] suggested that most *C. difficile* disease cannot be explained by contact with symptomatic individuals within the hospital, challenging conventional beliefs that most disease is transmitted within the hospital. The informal analysis ignored the information in much of the data and was unable to investigate heterogeneity between individuals, wards and bacterial strains. Interpretation is complicated by the fact that there

was no underlying statistical model and the consequences of various underpinning assumptions are difficult to assess.

Here we present a reanalysis of this data by introducing and fitting a stochastic model for CDI to the electronic patient records. In addition to learning about the biology and epidemiology of *C. difficile* infection, and assessing changes in hospital practice, both of which have important public health benefits, we see our approach more generally as a test case for the potential of combining genetic information and electronic patient records through principled statistical analyses.

In the next section we describe and motivate our stochastic compartmental model for *C. difficile* infection. Section 3 describes the fitting of the model including various diagnostic checks (though mathematical details are deferred to the Appendix). In Section 4 we give the results of our analysis, first for the biology of *C. difficile*, and then for the epidemiology and transmission dynamics of CDI. The final section provides a discussion of our approach and results, and potential further work.

Stochastic compartmental models have been applied before to hospital transmission of infectious disease, typically in simulation studies, or only been on relatively small datasets (one or two wards), with inference performed in both maximum likelihood and fully Bayesian frameworks [Cooper et al. (2008, 2012), Kypraios et al. (2010), Lanzas et al. (2011), Starr and Campbell (2001), Starr et al. (2009)].

2. Stochastic compartmental model for *C. difficile* infection. The starting point for our analysis is a novel stochastic compartmental model designed to capture the main features of CDI and its spread. The model is illustrated by the diagram in Figure 1.

Informally speaking, individual patients at a point in time are modelled as susceptible (not carrying *C. difficile*), colonized (carrying *C. difficile* but not yet either symptomatic or infectious), symptomatic and infectious, or removed from the model. Infectious patients may transmit *C. difficile* to susceptible individuals at a rate which depends on the nature of the interaction between them. Transmission may occur due to direct contact (patients are in the same ward at the same time), inter-ward contact when the patients are in different wards at the same time, or via ward contamination after the infectious patient has left the ward [Best et al. (2010)]. When such a transmission occurs, the *C. difficile* in the newly infected individual will have the same MLST as that in the individual responsible for the transmission event. We also allow susceptible individuals to acquire *C. difficile* from sources other than the known CDI cases, and refer to these as *background transmissions*. The MLST of background transmissions is drawn at random from a distribution (which we learn from the data).

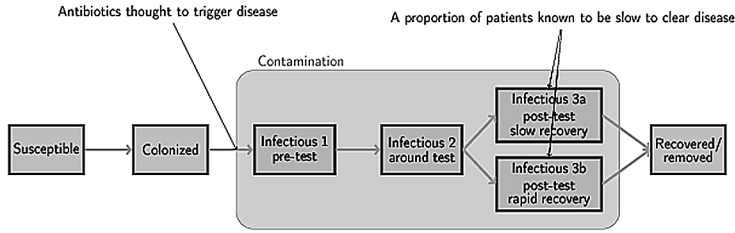We now give formal description of the model and its parameters.

FIG. 1. *Diagram showing model for* C. difficile *infection* (*CDI*). *Patients enter in a susceptible state and are colonized at a rate which depends on* (1) *whether the patient is inside or outside the hospital*; (2) *the number of infectious patients in the same ward*; (3) *the number of infectious patients within the same hospital*; (4) *the number of infectious patients who have been discharged from the same ward* (*post-discharge contamination*). *Once colonized, they become infectious at a rate which depends on whether they are in or outside the hospital. Once infectious, they are tested, and then recover at one of two rates corresponding to different rates of recovery.*

2.1. *States of the stochastic compartmental model.* In this section we describe in detail the states and transitions of the compartmental model of *C. difficile* infection.

At the start of the study, all individuals are assumed to be *susceptible*. We ignore the possibility that some individuals are infected at the start of the study; however, given the length of the study, impact on parameter estimates and inferred transition events is small.

Upon exposure to *C. difficile* (via an infected patient or environmental contamination), susceptible patients may become *colonized*. Colonized patients carry *C. difficile* but are not infectious. We assume all patients who become colonized eventually (within the timescale of the study) become infectious. Our use of a separate colonised state between the susceptible and infectious states allows us to model (albeit imperfectly) the period between becoming colonized with *C. difficile* and the triggering of CDI, for example, after administration of antibiotics. We were unable to incorporate antibiotic use explicitly, as this information is not present in our database. As the level of antibiotic use is much higher inside the hospital than outside, we allow the rate of movement from colonized to infected states to differ according to whether the patient is inside or outside the hospital.

Once colonized, patients become *infectious* at a rate which differs depending on whether they are inside or outside the hospital. The infectious state is further subdivided into Infectious 1, Infectious 2 and Infectious 3. For convenience, Infectious 2 period is fixed at a window (in days) around the diagnosis time T. Infectious 1 is a period before diagnosis in which the patient is infectious. Infectious 3 is a period after the test, where the patient may well be isolated but may continue to infect other patients. For our main analysis, we fixed $a = 2$ and $b = 1$. These numbers were chosen through consultation with infectious disease physicians, taking into account the typical timeline from onset of symptoms to test result and treatment. A sensitivity analysis is given in Section 3.1 of Cule and Donnelly (2017).

Note that the period of infectiousness before the patient is tested could cover both asymptomatic and symptomatic transmissions, and our model has no method to distinguish between them.

When two patients are visiting the same ward, we assume that an infected patient is equally likely to infect any other patient within the ward. They are also equally likely to infect any other patient within the same hospital site (but on a different ward), at a different rate from those in the same ward.

Some patients are observed to have recurrent relapses of disease for several months following initial infection [Johnson (2009)]. We therefore allow in our model two rates of recovery for the patients, one corresponding to a relatively rapid recovery and one to a longer period of infectiousness.

Sensitivity analysis assuming similarly persisting contamination after a patient left "Infectious 3" while on a ward produced similar results (data not shown).

C. difficile can form highly resilient spores which are known to persist in the environment for many months after the initial infection. Asymptomatic colonization of healthcare workers is another possible cause of post-discharge transmission. We therefore model post-discharge contamination as follows. If a patient leaves the ward in one of the Infectious" states, they may continue to infect other patients who visit that ward.

Where one patient infects another (either within the same ward, within the same hospital or via post-discharge contamination), we assume that the newly infected patient has the same MLST as the infecting patient.

2.2. Transitions between states of the stochastic compartmental model.

We model transitions between these states as a (nonconstant) Poisson process, whose waiting times are governed by parameters to be estimated.

We assume a susceptible patient becomes colonized at a time-dependent rate, which depends on (1) whether the patient is inside or outside the hospital, and (2) if inside, other infectious individuals to whom the patient was exposed. There are two basic background rates, while the patient is in the hospital and while the patient is outside the hospital. Transmission from infectious individuals occurs at a rate which depends on which state (Infectious 1, 2 or 3) the infectious individual is in and whether the contact is direct, hospital-wide or via post-discharge contamination. In summary,

$$
(1) \quad \beta(t) =
\begin{cases}
\beta_0 + \beta_1 I_1^w(t) + \beta_2 I_2^w(t) + \beta_3 I_3^w(t) \\
\quad + \psi\left(\beta_1 D_1^w(t) + \beta_2 D_2^w(t) + \beta_3 D_3^w(t)\right) & \text{inside the hospital,} \\
\quad + \eta\left(\beta_1 I_1^h(t) + \beta_2 I_2^h(t) + \beta_3 I_3^h(t)\right) \\
\quad + \psi\eta\left(\beta_1 D_1^h(t) + \beta_2 D_2^h(t) + \beta_3 D_3^h(t)\right) \\
\beta_{-1} & \text{outside the hospital.}
\end{cases}
$$

In this expression,

- $I_i^w(t)$ is the number of patients in infectious state i in the same ward as the susceptible patient at time t.
- $I_i^h(t)$ is the number of patients in infectious state i in the same hospital (different ward) as the susceptible patient at time t.
- $D_i^{\{w\}}(t)$ [resp., $D_i^{\{h\}}(t)$] is the number of patients causing ward contamination who were discharged in state i from the same ward (resp., a different ward in the same hospital building).
- β_0 is background pressure not associated with any particular infected patient.
- β_1 is the rate of infection of patients in the same ward, before testing.
- β_2 is the rate of infection of patients in the same ward within a window $[-2, 1]$ of the test.
- β_3 is the rate of infection of patients in the same ward after testing.
- β_{-1} is the background pressure when between hospital visits.
- ψ is a "contamination multiplier" such that the rate of infection for patients who have been discharged in state I_j is $\psi \beta_j$ of these states.
- η is a "hospital-wide multiplier" such that the rate of infection for patients in the same hospital, but different ward, due to a patient in state I_j is $\eta \beta_j$.

The transition from colonized to infectious occurs at rate λ_{out} outside the hospital and λ_{in} inside the hospital. We distinguish between the two possibilities in order to capture the heterogeneity in risk for *C. difficile* infection within and outside the hospital system. While infectious, the patient infects others in the same ward at rate β_1 and in different wards in the same hospital at rate $\eta \beta_1$.

The transition from state infectious 1 to infectious 2 is assumed to occur at rate λ_3 regardless of whether the patient is in or outside the hospital. The patient remains in the infectious 2 (acute symptomatic) state for 3 days, with the test assumed to occur on day 2.

After this, the patient remains infectious. With probability θ, the patient progresses to the removed state at rate μ_{1a}; otherwise, the patient progresses at rate μ_{1b}. This distinction allows for the empirical observation that a small number of patients experience recurrent disease (which we do not model explicitly). Identifiability is ensured by the prior assumption that $\mu_{1b} < \mu_{1a}$. Note that this does not necessarily imply that the patients have been symptomatic for the entire period.

2.2.1. *Modelling post-discharge contamination.* *C. difficile* is known to form highly resilient spores which can persist in the hospital environment for a long time. These are assumed to occur at a rate $\psi \beta_i$ if an infectious patient is discharged from a ward in state I_i. These spores persist for an $\text{Exp}(\mu_2)$ time.

3. Fitting the model. We adopted a Bayesian framework in order to fit this model to the available data using Markov chain Monte Carlo methods to sample from the posterior distribution. Encouragingly, even for our nontrivial stochastic model and a hospital database of hundreds of thousands of patients, inference was

TABLE 2
Parameters of the model and their priors, with references for the values where appropriate

Parameter	Description (rate of exponential distribution, unless otherwise specified)	Prior	Prior mean	Prior std. dev.
β_0	Transmission from background	Gamma	0.001	0.001
β_1	Transmission from pre-test infectious CDI cases	Gamma	0.001	0.001
β_2	Transmission from near test infectious CDI cases	Gamma	0.001	0.001
β_3	Transmission from post-test infectious CDI cases	Gamma	0.001	0.001
β_{-1}	Transmission from CDI cases outside the hospital	Gamma	0.001	0.001
η	Hospital-wide multiplier on rate of transmission from infectious CDI cases	Gamma	1	1
ψ	Contamination multiplier on rate of transmission from infectious CDI cases	Gamma	1	1
λ_1	C to I_1 outside hospital	Gamma	0.23	0.15[1]
λ_2	C to I_1 inside hospital	Gamma	0.23	0.15[1]
λ_3	I_1 to I_2	Gamma	0.23	0.15[1]
μ_0	I_3 to R (long)	Gamma	0.14	0.12[2]
μ_1	I_3 to R (short)	Gamma	0.14	0.12[2]
μ_2	Spore duration	Gamma	0.14	0.12
θ	Proportion of CDI cases with fast recovery (fast loss of infectivity)	Beta	0.5	0.29

[1] Health Protection Agency (2010).

[2] Bobulsky et al. (2008).

computationally tractable (e.g., a run of 100,000 iterations of the MCMC took 7 days on a 3 GHz processor).

3.1. *Prior distributions for parameters.* Information from the literature about parameters in the model, including transmission rates and incubation periods, was incorporated via diffuse prior distributions (Table 2). Sensitivity analyses subsequently showed that conclusions were robust to details of these prior assumptions, which is unsurprising given the amount of data available (see Appendix).

3.2. *MCMC diagnostics and model assessment.* A common concern in inference for complex models is assessing the mixing of the Markov chain used to explore the parameter space, that is, whether the space has been fully explored. Since we update the (unknown) source and time of transmission, a particular concern is that we could get stuck in one particular transmission direction and never explore the other direction of "donor" and "recipient."

3.2.1. *Assessing mixing of the Markov chain.* We used several standard techniques to assess mixing of the Markov chain [Gilks, Richardson and Spiegelhalter (1998)] including the following:

- Running the chains for 300,000 iterations, storing every 10th to reduce autocorrelation in the stored samples.
- Discarding the first half of each run to ensure the Markov chain has converged to its stationary distribution.
- Visual inspection of trace plots, shown in Figure 3 of Cule and Donnelly (2017).
- Starting multiple chains from 3 dispersed starting points, and pooling the samples for inference. The marginal posterior distribution of each parameter is shown in Figure 4 of Cule and Donnelly (2017).
- The Brooks–Gelman–Rubin statistic, shown in Figure 5 of Cule and Donnelly (2017).

3.2.2. *Assessing sensitivity to prior information.* A priori plausible priors were chosen for the epidemiological parameters by reference to the literature on *C. difficile* transmission (Table 2). However, sensitivity to prior information is always a concern in complex models. Figure 6 of Cule and Donnelly (2017) shows the posterior parameter distributions when the prior means are respectively halved and doubled (keeping the prior variances the same as in the main model). For most of the parameters, the posterior distributions are quite robust to such changes in prior information. The exception is the infectivity prior to testing, which is strongly influenced by the prior. The short duration of the period, revealed by the lack of transmissions occurring before diagnosis is confirmed, means that there is little information on the rate to be obtained from the data, and thus the posterior value is strongly influenced by the prior. The remaining parameters are more strongly influenced by the data (equivalently, the likelihood).

3.2.3. *Assessing sensitivity of results to the presence of mixed infections.* Multiple strains have been isolated from a small proportion of cases (∼5%). Since these were only identified if there were morphologically distinct colonies (in which case multiple colonies were typed, and did not always yield distinct strains), we did not explicitly model this possibility. Rather, we tried two approaches:

1. Pick one of the types at random (used in the main analysis).
2. Split into two psuedo-patients behaving as independent infections.

Both yielded extremely similar results.

3.2.4. *Assessing robustness to missing samples.* 4% of EIA-positive samples were not retrieved for culture. Two approaches have been used to account for missing EIA positive samples: First, assuming they are negative and, second, assuming they are positive with unknown MLST (used in the main analysis). Neither of these extreme assumptions led to a significant change in conclusions, from which we conclude that our results are robust to a small amount of missing data.

3.3. *Simulation study.* To investigate the sensitivity of our approach to various modelling assumptions, we undertook several simulation studies. We first simulated data according to the generative model described above, in each case using the same hospital visit schedule as the original dataset (since admissions and discharges from hospital are not modelled explicitly).

In order to make our simulation realistic, we set each parameter to its posterior mean for this simulation. We then additionally simulated data under several different scenarios by modifying the parameter values. These scenarios are described in detail below. For each simulated dataset, we re-fit the model using the procedure described above.

3.3.1. *Removing routes of transmission.* Our first set of simulations eliminated one of the routes of transmission in the model. We simulated three new datasets with the following characteristics:

1. No person-to-person transmission within hospitals.
2. No post-discharge contamination.
3. No between-ward transmissions.

The posterior distributions for key parameters are shown in Figure 2. The inferred parameters under the posterior mean simulation are shown by a solid line. For the data with no person-to-person transmission within hospitals (shown in a dashed line), the inferred transmission parameters were smaller before, near, and after the test. The dataset generated without post-discharge contamination (dotted line) shows a much smaller estimate for the relevant multiplier (labelled contamination multipler). The data generated with no between-ward contamination, shown in a dot-dashed line, shows a much smaller value for between-ward transmission. The unmodified parameters for background transmission show relatively similar values for all four datasets.

3.3.2. *One recovery rate.* Our model allows for both "slow" and "fast" recovery. We simulated data with one recovery rate only. The posterior distributions for the transmission rate parameters are illustrated in Figure 3. Note that the probability of a fast recovery is inferred to be close to 0 in this case.

Neither dataset shows a fast recovery time distribution which is significantly different from the prior distribution, indicating that this parameter may not be identifiable from the data.

3.3.3. *Much higher rates of transmission.* Finally, we simulated a dataset with between-person transmission 10 times higher than the baseline dataset. Figure 4 shows the posterior distribution for key parameters.

As well as having higher estimated transmission rates, the correspondingly larger number of transmissions in the resultant synthetic dataset leads to narrower posterior distributions for other parameters (not shown).
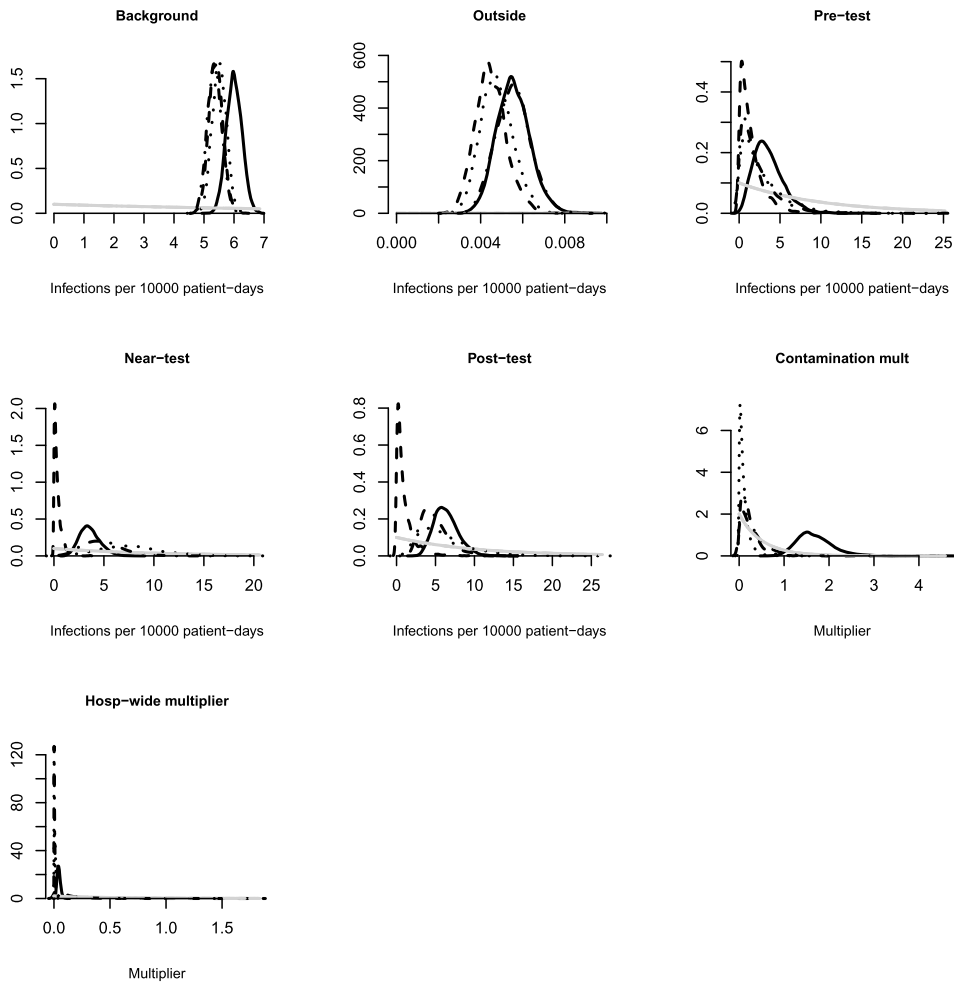
FIG. 2. *Figure showing distribution of key parameters from simulated data under the posterior mean (solid), with no person-to-person transmission (dashed), with no post-discharge contamination (dotted) and with no between-ward transmission (dot-dash). The prior distribution is shown in grey.*

4. Results. The Bayesian approach allows inferences to be drawn from the data about parameters in the model, and also about unobserved events (such as whether patient A transmitted *C. difficile* to patient B) and times (e.g., for how long did patient C incubate the disease before their test). In each case these inferences are in terms of posterior probabilities (of events) or distributions (for times and parameters).

Before turning to detailed results, we first illustrate some of the inferences made by the model (Figure 5) for a subset of the patients [Figure 1 of Cule and Donnelly (2017) for all other CDI patients]. These figures depict the sequence of hospital
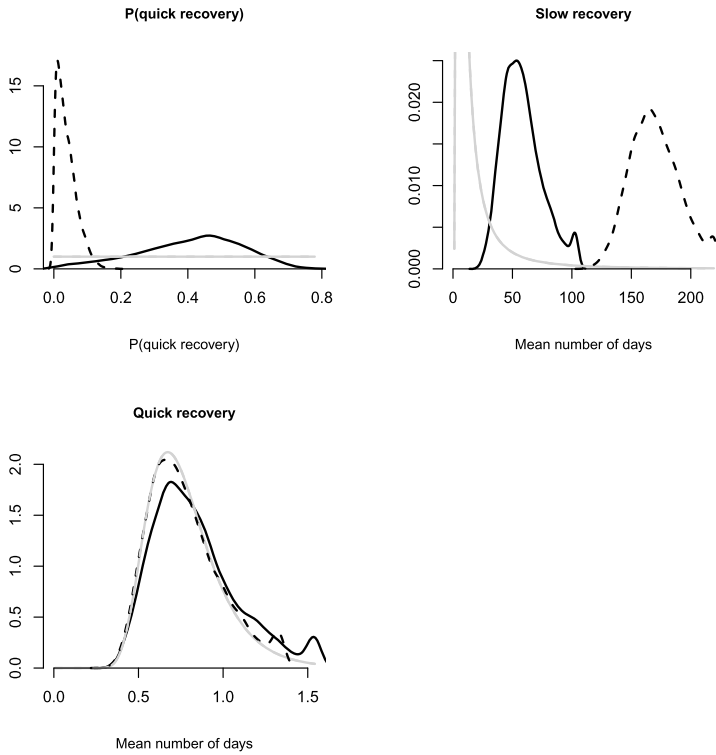
FIG. 3. *Figure showing distribution of key parameters from simulated data under the posterior mean (solid) and a simulation with one recovery rate only (dashed).*

visits by a subset of patients who went on to test positive for *C. difficile*, the time of their positive test, and illustrate the inferred transmission events and their posterior probabilities (above a certain posterior probability threshold).

4.1. *Biology of* C. difficile *disease*.   The posterior distributions of the parameters in the model are shown in Figure 7. Note that in all cases these posteriors differ substantially from the prior distributions for the parameter, reflecting the substantial information in the data. In this section, we focus on parameters related to the biology and epidemiology of *C. difficile*.

Our model allowed for the possibility of different rates of transmission in the period before, around or after the positive test for *C. difficile*, which could reflect biology or possibly changes in hospital practice, such as isolation of the affected patient after the clear onset of symptoms. An infectious individual transmits *C. difficile* to a particular susceptible individual in the same ward at a mean rate of 4.4, 3.6 and 7.9 infections per 10,000 bed-days before, around and after their positive test, respectively [see Figure 6(d) for full posterior distributions]. While there is some evidence that transmission rates may be higher after the test than in the pe-
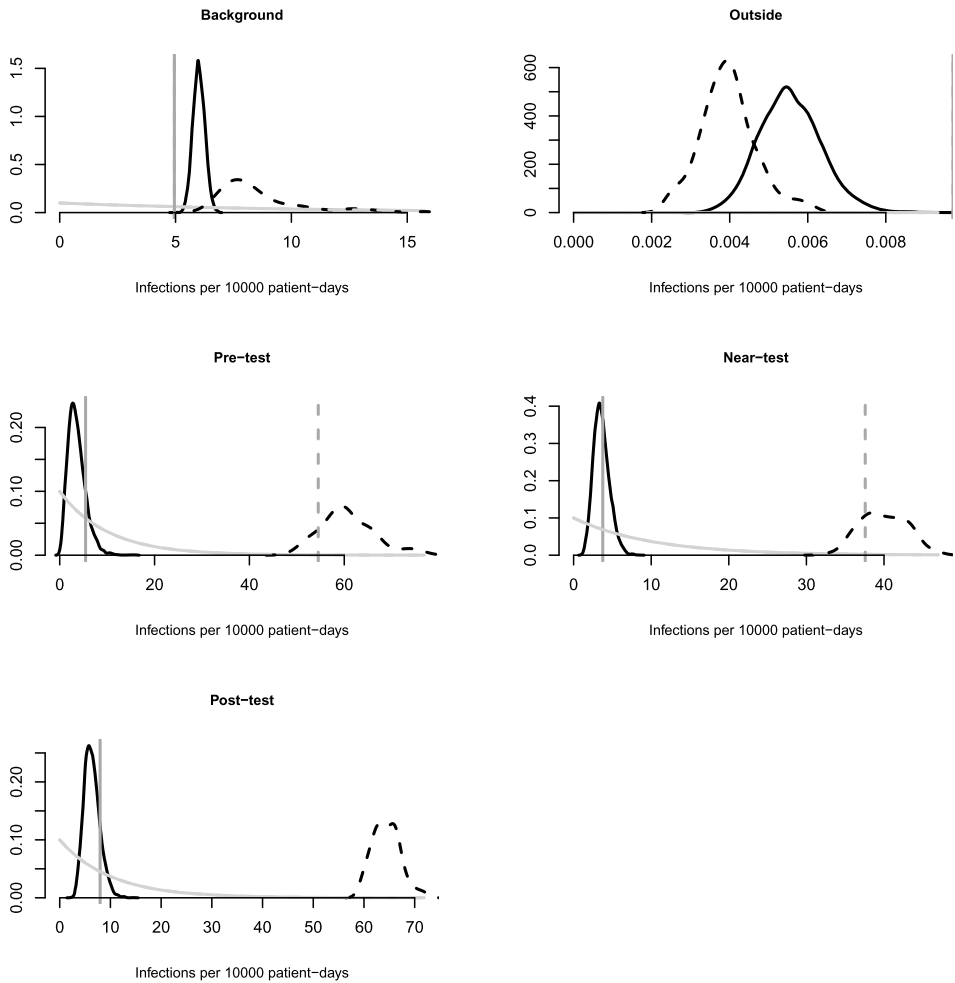
FIG. 4.   *Figure showing distribution of key parameters from simulated data under the posterior mean (*solid*) and a simulation with* 10 *times the between-person transmission rate within hospitals (*dashed*). As well as higher point estimates, note the narrower posterior distributions due to the larger amount of data in the high-transmission simulation. The grey vertical lines in each case show the simulated value.*

riod around the test, the data are also consistent with no change in transmission rates in these different periods. Notwithstanding possible differences in rates, it is noteworthy that most (75%) transmissions from infected patients actually occurred after the diagnosis of CDI, even though the study hospitals had in place aggressive infection control measures throughout the study period, which included isolation of infected individuals.

We estimate that, per unit exposure, post-discharge contamination, that is, colonisation of one patient with the bacteria carried by another patient in a par-
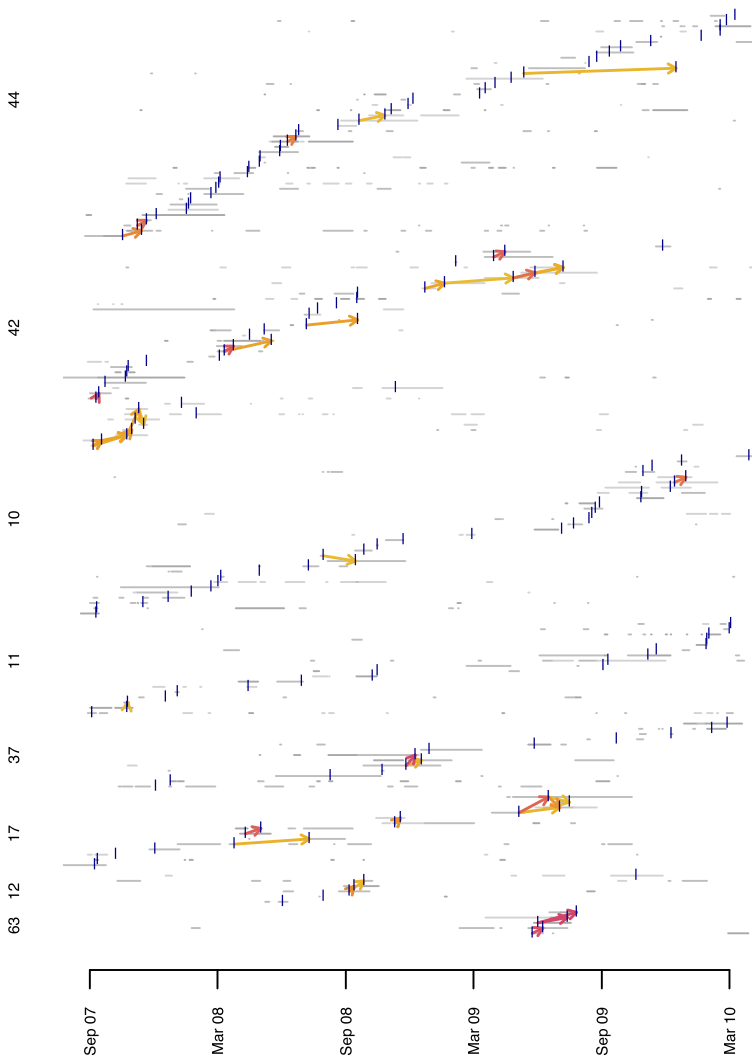
FIG. 5. *Data and inferred transmissions for* 183 *of the* 931 *infected patients, chosen from* 8 *example multi-locus sequence types. Periods during hospital admissions are shown in grey. First positive tests are shown with a blue vertical tick mark. Inferred direct transmissions by any route are shown with arrows, with darker red corresponding to a higher posterior probability (only transmissions with P > 0.2 are shown). Overall, there are relatively few direct transmissions within the hospital.*

ticular ward after the infected patient has left the ward, causes infection at a rate of roughly 70% [interquartile range (IQR) 56%–85%] of that of direct contact. Further, the time period after discharge from the ward over which transmission via contamination can occur is not small: the median duration is 14 days (IQR 5.9–29.9 days).
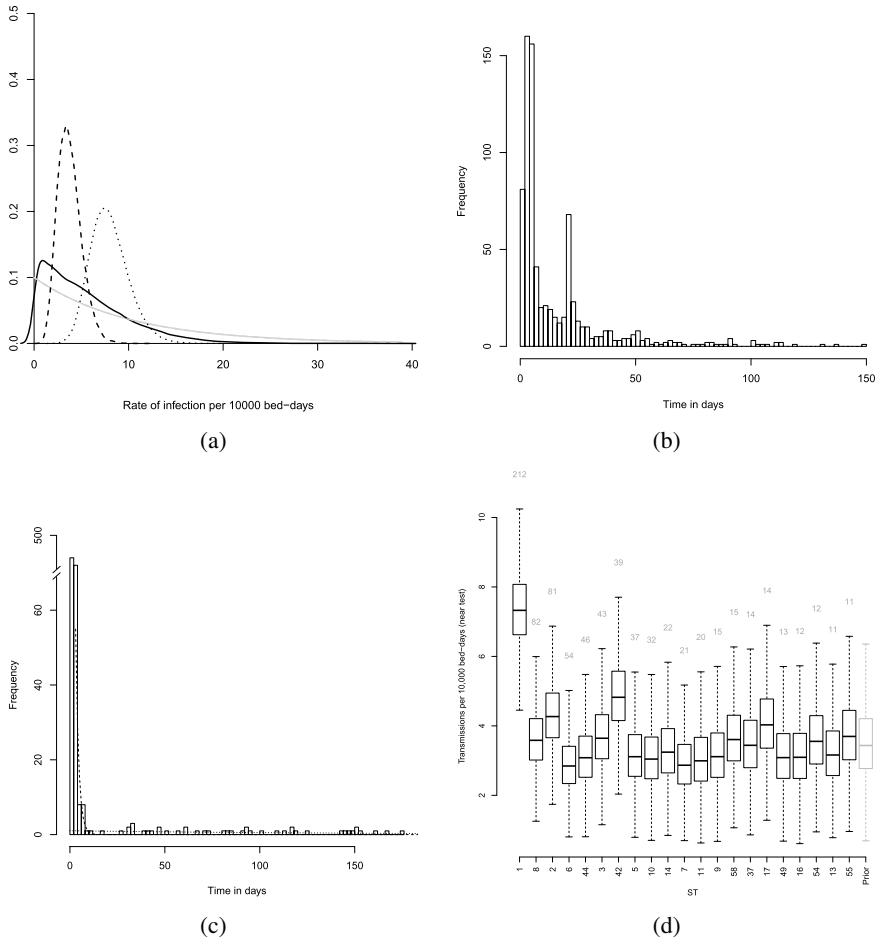
(a)

(b)

(c)

(d)

FIG. 6.   *Distributions of quantities related to biology of CDI.* (a) *Posterior distribution on transmission rates for pre-, around- and post-test (curves in solid, dashed and dotted, respectively). The grey curve indicates the prior distribution on these rates. There is limited data on pre-test rates because the inferred infectious time prior to the test is short.* (b) *Distribution across individuals of their posterior median time from infection to when the individual becomes infectious for individuals inferred to be infected during a hospital stay. Only individuals who become infected after at least one hospital visit are included.* (c) *Distribution across individuals of their posterior median recovery times. This distribution was modelled as a mixture of two recovery-time distributions, one for "quick recoverers" and one for "slow-recoverers." The posterior distribution of recovery times for the "quick recoverers" is shown as a dashed line, and that for slow recoverers" as a dotted line. Around* 70% *of individuals are "quick recoverers."* (d) *Posterior distributions of transmission rates (per* 10,000 *days exposure) by sequence type, shown only for sequence types with more than* 10 *cases. The thick horizontal line marks the median of the posterior, the vertical box shows its inter-quartile range, and the dashed vertical lines delimit its* 5th *and* 95th *percentiles. The number of cases for each sequence type is shown above the representation of the posterior.*
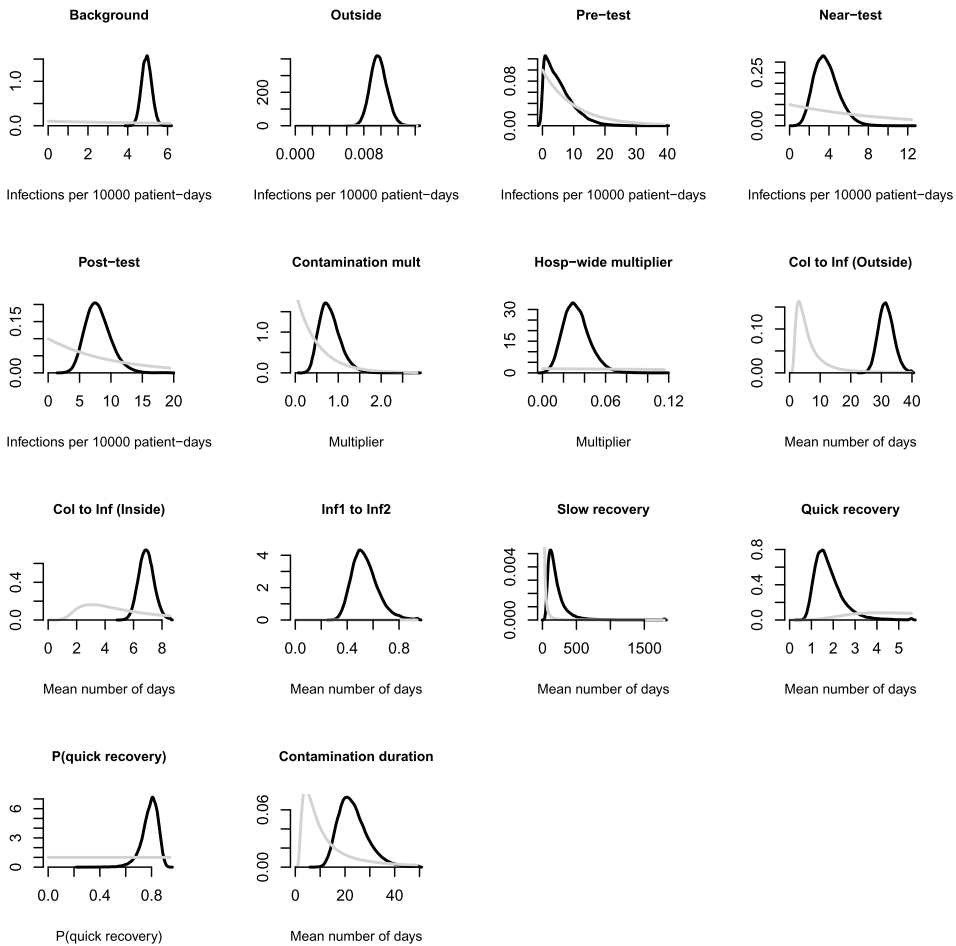
FIG. 7. *Marginal posterior distribution of each parameter from the main analysis. Prior distributions are shown in grey. Distributions are obtained by combining* 15,000 *saved draws from each of* 3 *runs of the MCMC after burn-in from different starting points and thinned to every* 10*th iteration. From top left, across successive rows, the panels relate to the following*: colonization rate within the hospital; colonization rate outside hospital; rates of transmission from infectious patients before, near to and after diagnosis; multipliers for contamination and hospital-wide transmission; rate of progressing from colonized to infectious state inside and outside the hospital; rate of progressing to near-test state; rate of recovery (slow recovery); rate of recovery (quick recovery); proportion of patients recovering quickly; duration of contamination.*

Inter-ward transmission occurs at a rate of only 3% that of direct contact (IQR 2.5%–4.0%) per pair of infective and susceptible individuals, but because of the much larger number of potential infective-susceptible pairs between wards, we will see below that the overall contribution of inter-ward transmissions is significant,

totalling around half that of each of direct contact within wards and post-discharge ward contamination.

Figure 6(b) shows the posterior distribution of the time from colonization until the start of the infectious period, and illustrates that some patients are infected a long time before becoming infectious. This is probably because, unlike many other infections, CDI (and the accompanying dramatic increase in infectiousness) is primarily precipitated by antibiotic use.

The data support two different paths to recovery from CDI (and with it removal from our model). Most patients (70%, IQR 64%–75%) recover relatively quickly (posterior median of parameter for mean recovery time 2.4 days, IQR of mean 1.8–3.1 days), while the remainder can remain infected for substantial periods (1–6 months) [Figure 6(c)]. Although most patients recover relatively quickly, most transmissions (73%, IQR 68%–77%) are attributed to those who recover more slowly. This has consequences for infection control since it implies that careful management of CDI cases with ongoing or recurrent infection could have the greatest impact on reducing transmission.

From the perspective of a particular susceptible individual, averaged over their likely ward movements and composition and with no direct contact with infectious patients, colonization which leads to CDI in hospital is inferred to occur at a rate of 4.86 (IQR 4.69–5.03) per 10,000 bed-days inside the hospital, and 0.00970 (IQR 0.00906–0.01040) per 10,000 days outside the hospital. The estimated overall risk of colonization from spending an extra day in the hospital is slightly higher than the risk specifically due to spending a day in the same ward as an infectious patient (5:3 ratio).

To investigate potential heterogeneity between STs, we refitted the model to allow transmissibility to vary. We see evidence of heterogeneity in the infectiousness of different MLSTs, with ST1 (Ribotype 027/NAP1) being more transmissible [Figure 2(d)]. It is responsible for 50% (IQR 49%–52%) of inferred transmissions, but only 13% (IQR 12.6%–13.9%) of new introductions (that is, instances of CDI not involving transmission from another patient). There is some evidence in our data that ST42 may also be more transmissible than others sequence types, and we note that this strain (Ribotype 106) has dominated in other areas of the UK [Health Protection Agency (2010)].

4.2. *Differences between transmission routes, individuals, wards and hospitals.* One advantage of our approach is that it provides probabilistic assignment of likely transmission routes for each transmission leading to CDI. These assignments can be aggregated to compare different routes of transmission, and differences between individuals, wards or hospitals.

There has been considerable interest in understanding the routes by which *C. difficile* transmission occurs, not least so as to develop and refine hospital management and infection control plans for reducing CDI. Our analyses show that the
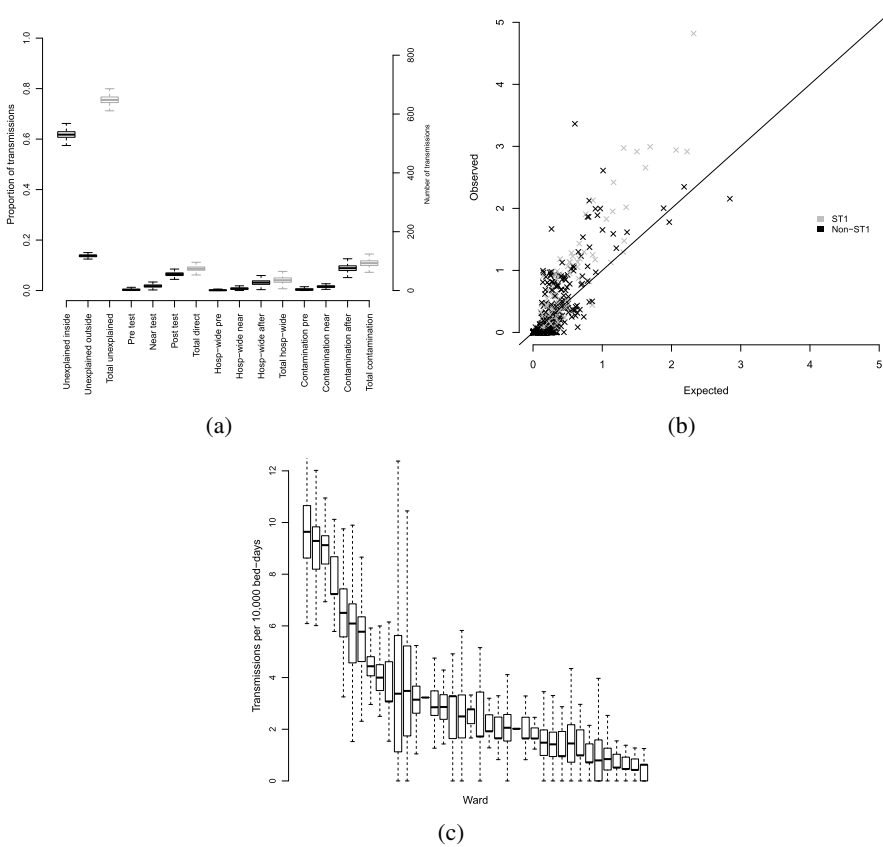
(a)

(b)

(c)

FIG. 8. *Posterior distributions on the number of transmissions attributable to routes, patients and wards. (a) Posterior distribution represented as in Figure 6(d) of the number of infections by different routes. (b) Observed versus expected (i.e., mean number averaging over parameter values in the posterior distribution) number of transmissions attributed to each infectious patient. (c) Posterior distribution represented as in Figure 2(d) of the rates of direct transmission per 10,000 days exposure across different wards in the study (data shown only for top 38 wards).*

majority (∼75%) of infections cannot be explained by contact with symptomatic patients in the hospital [Figure 8(a)].

We infer that about 10% of colonisation events occur through direct transmission from an infected to a susceptible patient within the same ward, with a further 10% occurring through ward contamination. The remaining 5% of colonisation events are inferred to be between patients in different wards, plausibly through carriage of bacteria on hospital personnel or equipment. In terms of absolute numbers, estimated transmissions through within-ward contact, ward contamination and inter-ward transmission were 82, 89 and 40, respectively. Although there can be considerable uncertainty over the transmission route for an individual colonisation event, these overall estimates average over the uncertainty across all CDI cases, and are themselves reasonably precise (Figure 8).

Single individuals have played critical roles in some historical outbreaks of infectious disease. Referred to as "superspreaders," the most famous is probably so-called "typhoid Mary," an asymptomatic cook who was reputed to have infected over 50 people with typhoid in early 20th century New York [Bourdain (2001), Lloyd-Smith et al. (2005)]. We see no evidence for individuals responsible for a large number of transmissions in our data: Figure 8(b) compares the observed and expected numbers of direct transmissions per individual, with no single individual responsible for more than five transmission events, and none greatly exceeding the number of transmissions expected for their period of infectiousness. (To be conservative, expected numbers of transmissions are calculated assuming all MLST types have the same transmission rate, and so the excess of observed to expected for individuals carrying ST1 is to be expected.)

Comparison across wards reveals considerable heterogeneity in the number of transmission events [Figure 8(c)]. However, it is important to note when interpreting these findings that no attempt was made to control for case mix (e.g., some wards, such as gerontology, containing patients more likely to develop CDI) or other factors which may be responsible for the differences between wards. While in principle such covariates could be explicitly modelled, we leave that as the subject of future studies.

We also observe differences in transmission rates between the three hospitals in our study (Figure 9). Again, care is needed in their interpretation because of differences in case mix between hospitals. Our approach also allows comparisons between time periods within hospitals, for example, to investigate the consequences of changed infection control protocols. These within-hospital comparisons will largely control for case mix (which does not greatly change over time), making them easier to interpret. In all three hospitals transmission rates dropped markedly, as did inter-hospital differences, between the first and second half of our study period (Figure 9).

5. Discussion. We have developed a stochastic model for *C. difficile* infection and applied it to a large database of electronic patient records, augmented by genetic information in the form of MLST. Stochastic modelling and inference is a powerful tool for improving our understanding of key questions relating to biology, epidemiology and healthcare management which were not previously tractable because of the complex nature of the disease process. Crucially, the model makes probabilistic assessments of likely transmission events. Aggregation across the 858 CDI cases with at least one overnight hospital stay in our data provides considerable information on questions of interest, and natural measures of uncertainty in estimated quantities, even when there are relatively few transmission events whose provenance can be assigned with high confidence.

It is encouraging that the combination of MLST data and patient records proves so powerful in studying the disease process. As higher resolution whole-genome data becomes widely available, we anticipate a greatly improved understanding of
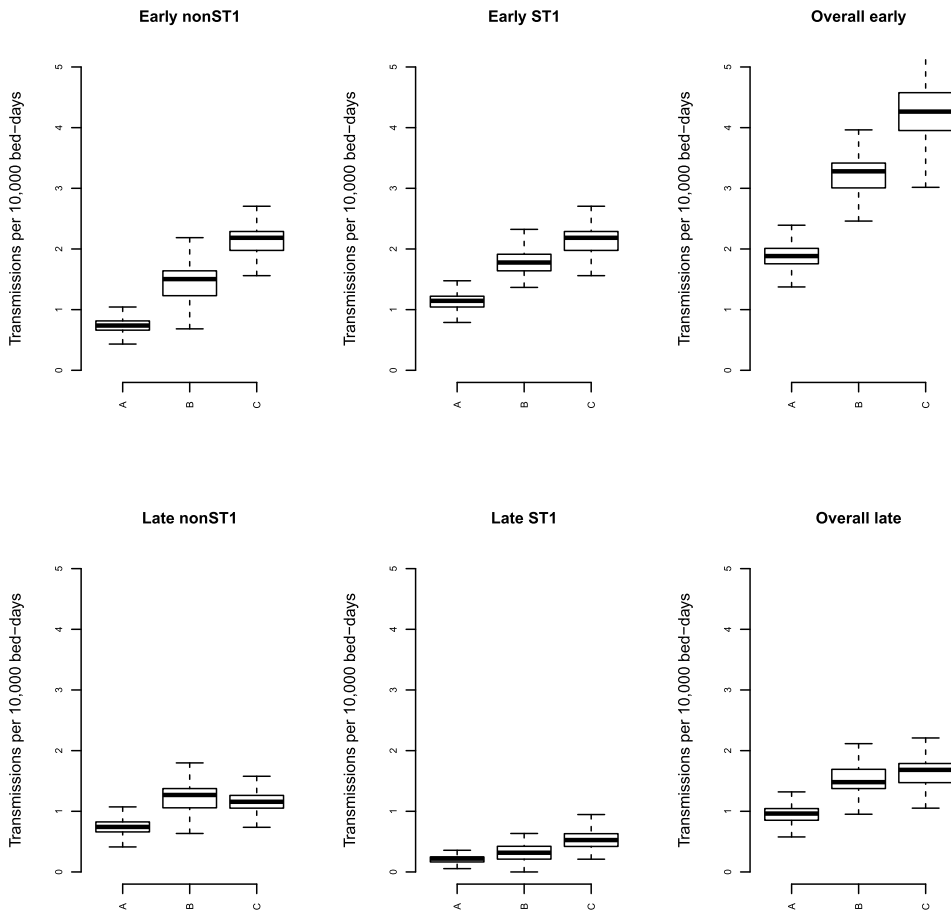
FIG. 9. *Posterior distribution represented as in Figure* 8(c) *of number of transmissions attributed to infectious patients, broken down by hospital and into early and late periods (first and second half of study), and ST*1 *only, and all non-ST*1 *MLSTs.*

infectious disease transmission. However, we have established that even the relatively course level of resolution afforded my MLST, combined with electronic patient records, allows inferences of transmission, and of other aspects of the epidemiology of hospital-based infectious disease, which are quite informative.

Our analyses established or confirmed several key features of *C. difficile* transmission and disease. We estimated transmission rates and showed that these are similar before, around or after an individual"s positive test. There can be a significant time lag between acquisition of *C. difficile* from a symptomatic patient and the onset of symptoms, presumably at least partly explained by the interaction with antibiotic use. Whilst most patients recover quickly from CDI, a sizeable minority (30%) can take months to recover. For these "slow recoverers," there is the po-

tential for onward transmission to occur a long time (weeks or months) after the initial positive test, and we found that they are responsible for 73% of transmission events from symptomatic patients. In spite of the infection control procedures in place in the hospitals, many of which only come into play when an individual displays symptoms of CDI, we found that the majority (70%) of transmissions from symptomatic patients occur after the diagnosis of CDI has been confirmed. We also saw clear evidence for a higher rate of transmission for a particular MLST sequence type, the known hypervirulent strain ST1/NAP1/RT027, with a median transmission rate roughly double that for most other sequence types.

The statistical approach taken here has several additional advantages. It allows us to compare patients, wards and hospitals in a sensible way, automatically taking into account the amount of exposure. We demonstrate the absence of *C. difficile* "superspreaders" (a small number of patients responsible for a large number of transmissions) more convincingly than simply looking for clusters of disease, which does not take into account the relative background frequency of the different sequence types and the different levels of exposure to susceptible patients.

More generally, our approach allows inferences about transmission events themselves. Transmission is the key endpoint for many important questions, including the consequences of changes in infection control procedures, but it cannot be observed directly so that previous approaches have necessarily focussed on surrogates for transmission, such as disease incidence. We provided an example of the value of learning directly about transmission events in comparing hospitals between different time periods.

In the hospitals studied, only 25% of CDI cases were explained by contact with symptomatic patients, leaving the majority unexplained, in agreement with the original published analysis of a subset of the data [Walker et al. (2011)]. We refer to that paper for discussion and implications of the finding.

In contrast to the earlier report, our analysis revealed an important role for transmission after patients had left a ward, in what we have referred to as ward contamination. We inferred slightly higher numbers of transmissions overall via this route than from direct transmission within wards, with the rate of infection per unit of exposure due to contamination around 70% of that for patients in the same ward. The time period after discharge from the ward over which transmission via contamination can occur has a median duration of 14 days, with 25% lasting more than 28 days. *C. difficile* is known to produce highly resilient spores [Donskey (2010), Wilcox and Fawley (2000)], and so, given the relative importance of this transmission route, it will be important to better understand this mechanism.

We also established a significant role for transmission between individuals in different wards. Even though individual-to-individual transmission rates across wards are low, leading to less emphasis in infection control policies, collectively they account for about 20% of the transmission events from other established CDI cases, demonstrating the clinical importance of these routes overall.

Our model does not allow the possibility of individuals carrying and transmitting *C. difficile* whilst remaining asymptomatic [Loo et al. (2011)]. (We do allow asymptomatic carriage, but assume that all such individuals will eventually become symptomatic, for example, after antibiotic use, potentially after a long incubation time.) Whilst this could, in principle, be handled in the modelling framework by treating *C. difficile* carriage status as missing data, inference would be computationally intractable because of the need to average over the unobserved states in hundreds of thousands of individuals. The impact on transmission of asymptomatic carriage remains an important unanswered question, best addressed through a longitudinal study of both asymptomatic carriage and transmission in the same group of patients.

The current analysis is computationally tractable, but a 7-day turnaround may be insufficient for clinical application. We leave as a potential topic for further research the development of new computational methodology for improving this. A particularly intriguing possibility is the application of online or sequential methods to update parameter estimates online, as new data becomes available, in real time during an outbreak.

In addition to information about the transmission and epidemiology of *C. difficile* infection, we see our analysis as a proof of principle of the value in using stochastic epidemic models to study linked electronic hospital records and data documenting genetic variation of bacterial pathogens. With the collection of genetic data on pathogens and access to anonymised hospital records both likely to become widespread, we see considerable potential for these approaches more generally.

SUPPLEMENTARY MATERIAL

Supplement to "Stochastic modelling and inference in electronic hospital databases for the spread of infections: *Clostridium difficile* transmission in Oxfordshire hospitals 2007–2010" (DOI: 10.1214/16-AOAS1011SUPP; .pdf). Full likelihood derivation, full results and sensitivity analysis to accompany "Stochastic modelling and inference in electronic hospital databases for the spread of infections: *Clostridium difficile* transmission in Oxfordshire hospitals 2007–2010."

REFERENCES

BARTLETT, J. G., CHANG, T. E. W., GURWITH, M., GORBACH, S. L. and ONDERDONK, A. B. (1978). Antibiotic-associated pseudomembranous colitis due to toxin-producing clostridia. *N. Engl. J. Med.* **298** 531–534.

BEST, E. L., FAWLEY, W. N., PARNELL, P. and WILCOX, M. H. (2010). The potential for airborne dispersal of *Clostridium difficile* from symptomatic patients. *Clin. Infect. Dis.* **50** 1450–1457.

BOBULSKY, G. S., AL-NASSIR, W. N., RIGGS, M. M., SETHI, A. K. and DONSKEY, C. J. (2008). *Clostridium difficile* skin contamination in patients with *C. difficile*-associated disease. *Clin. Infect. Dis.* **46** 447–450.

BOURDAIN, A. (2001). *Typhoid Mary*: *An Urban Historical*. Bloomsbury, New York.

COOPER, B. S., MEDLEY, G. F., BRADLEY, S. J. and SCOTT, G. M. (2008). An augmented data method for the analysis of nosocomial infection data. *Am. J. Epidemiol.* **168** 548–557.

COOPER, B. S., KYPRAIOS, T., BATRA, R., WYNCOLL, D., TOSAS, O. and EDGEWORTH, J. D. (2012). Quantifying type-specific reproduction numbers for nosocomial pathogens: Evidence for heightened transmission of an Asian sequence type 239 MRSA clone. *PLoS Comput. Biol.* **8** e1002454.

CULE, M. and DONNELLY, P. (2017). Supplement to "Stochastic modelling and inference in electronic hospital databases for the spread of infections: *Clostridium difficile* transmission in Oxfordshire hospitals 2007–2010." DOI:10.1214/16-AOAS1011SUPP.

DONSKEY, C. J. (2010). Preventing transmission of *Clostridium difficile*: Is the answer blowing in the wind? *Clin. Infect. Dis.* **50** 1458–1461.

FINNEY, J. M., WALKER, A. S., PETO, T. E. A. and WYLLIE, D. H. (2011). An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Med. Inform. Decis. Mak.* **11** 7.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J. (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

GRIFFITHS, D., FAWLEY, W., KACHRIMANIDOU, M., BOWDEN, R., CROOK, D. W., FUNG, R., GOLUBCHIK, T., HARDING, R. M., JEFFERY, K. J. M., JOLLEY, K. A., KIRTON, R., PETO, T. E. A., REES, G., STOESSER, N., VAUGHAN, A., WALKER, A. S., YOUNG, B. C., WILCOX, M. and DINGLE, K. E. (2010). Multilocus sequence typing of *Clostridium difficile*. *J. Clin. Microbiol.* **48** 770–778.

HEALTH PROTECTION AGENCY (2009). Voluntary surveillance of *Clostridium difficile* in England, Wales and Northern Ireland, 2008.

HEALTH PROTECTION AGENCY (2010). *Clostridium difficile* Ribotyping Network (CDRN) for England and Northern Ireland Annual Report 2010/2011.

JOHNSON, S. (2009). Recurrent *Clostridium difficile* infection: A review of risk factors, treatments, and outcomes. *J. Infect.* **58** 403–410.

KARAS, J. A., ENOCH, D. A. and ALIYU, S. H. (2010). A review of mortality due to *Clostridium difficile* infection. *J. Infect.* **61** 1–8.

KYPRAIOS, T., NEILL, P. D. O., HUANG, S. S., RIFAS-SHIMAN, S. L. and COOPER, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing methicillin-resistant *Staphylococcus aureus* transmission in intensive care units. *BMC Infect. Dis.*

LANZAS, C., DUBBERKE, E. R., LU, Z., RESKE, K. A. and GRÖHN, Y. T. (2011). Epidemiological model for *Clostridium difficile* transmission in healthcare settings. *Infect. Control Hosp. Epidemiol.* **32** 553–561.

LLOYD-SMITH, J. O., SCHREIBER, S. J., KOPP, P. E. and GETZ, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature* **438** 355–359.

LOO, V. G., BOURGAULT, A.-M., POIRIER, L., LAMOTHE, F., MICHAUD, S., TURGEON, N., TOYE, B., BEAUDOIN, A., FROST, E. H., GILCA, R., BRASSARD, P., DENDUKURI, N.,

BÉLIVEAU, C., OUGHTON, M., BRUKNER, I. and DASCAL, A. (2011). Host and pathogen factors for *Clostridium difficile* infection and colonization. *N. Engl. J. Med.* **365** 1693–1703.

MCDONALD, L. L. C., KILLGORE, G. E., THOMPSON, A., OWENS JR, R. C., KAZAKOVA, S. V., SAMBOL, S. P., JOHNSON, S. and GERDING, D. N. (2005). An epidemic, toxin gene-variant strain of *Clostridium difficile*. *N. Engl. J. Med.* **353** 2433–2441.

MCFARLAND, L. V., MULLIGAN, M. E., KWOK, R. Y. Y. and STAMM, W. E. (1989). Nosocomial acquisition of *Clostridium difficile* infection. *N. Engl. J. Med.* **320** 204–210.

PLANCHE, T., AGHAIZU, A., HOLLIMAN, R., RILEY, P., POLONIECKI, J., BREATHNACH, A. and KRISHNA, S. (2008). Diagnosis of *Clostridium difficile* infection by toxin detection kits: A systematic review. *Lancet, Infect. Dis.* **8** 777–784.

ROLFE, R. D. (1988). Asymptomatic intestinal colonization by *Clostridium difficile*. In *Clostridium Difficile*: *Its Role in Intestinal Disease* 201–225. Academic Press, San Diego, CA.

STARR, J. M. and CAMPBELL, A. (2001). Mathematical modeling of *Clostridium difficile* infection. *Clin. Microbiol. Infect.* **7** 432–437.

STARR, J. M., CAMPBELL, A., RENSHAW, E., POXTON, I. R. and GIBSON, G. J. (2009). Spatio-temporal stochastic modelling of *Clostridium difficile*. *J. Hosp. Infect.* **71** 49–56.

WALKER, A. S., EYRE, D. W., WYLLIE, D. H., DINGLE, K. E., HARDING, R. M., O'CONNOR, L., GRIFFITHS, D., VAUGHAN, A., FINNEY, J. M., WILCOX, M. H., CROOK, D. W., PETO, T. E. A. and WALKER, A. S. (2011). Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. *PLoS Med.* **9** e1001172.

WILCOX, M. and FAWLEY, W. (2000). Hospital disinfectants and spore formation by *Clostridium difficile*. *Lancet* **356** 1324–1324.

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
1 SOUTH PARKS ROAD
OXFORD OX1 3TG
UNITED KINGDOM
E-MAIL: mcule@cantab.net

WELLCOME TRUST CENTRE
  FOR HUMAN GENETICS
ROOSEVELT DRIVE
OXFORD OX3 7BN
UNITED KINGDOM
E-MAIL: donnelly@well.ox.ac.uk