Modeling the Transmission of Vancomycin-Resistant Enterococcus (VRE) in Hospitals: A Case Study

A. R. Ortiz¹, H. T. Banks², C. Castillo-Chavez¹, G. Chowell¹, C. Torres-Viera³ and X. Wang¹

¹ School of Human Evolution and Social Change
   Arizona State University
   Tempe, Arizona 85287-2402
² Center for Research in Scientific Computation
   North Carolina State University
   Raleigh, North Carolina 27695-8212
³ South Florida Infectious Diseases
   8700 Tyndall Drive, Suite 100
   Miami, Fl 33176

February 24, 2010
Abstract

Nosocomial, or hospital-acquired infections, the fourth cause of death in the US, are evidence that hospitals provide not only medical care but also harbor pathogens that pose serious, often fatal, risks of infection, particularly to the young, the elderly, and immune-compromised individuals. Infection-control measures aimed at reducing their impact are being implemented with various degrees of efficiency at US hospitals. Data from general and oncology hospital units on Vancomycin-resistant Enterococcus (VRE), one of the most prevalent and dangerous pathogens involved, are used to highlight the importance of modeling nosocomial infection dynamics as a prelude to the testing and evaluation of control measures.

New mathematical models of the transmission dynamics of VRE in hospitals are introduced in order to identify and quantitatively assess the time evolution of nosocomial infections. Ordinary differential equation (ODE) and discrete delay differential (DDD) models in conjunction with statistical methods are used to estimate key population-level nosocomial transmission parameters. This framework is tested using unpublished surveillance data from two types of hospital units. The population is divided into uncolonized, VRE colonized, and VRE colonized-in-isolation categories and the use of constant and variable rates of isolation admitted with VRE or VRE-colonized during hospital stays is evaluated in models including health care workers’ hand-hygiene compliance. The process of model calibration detected irregularities in the available surveillance data; these irregularities that are most likely the result of the data recording-process. Efforts to fit data within our highly flexible dynamic-modeling framework suggest that clinical-trial level surveillance data is needed.

The usefulness of our new flexible modeling framework for the transmission dynamics of nosocomial infections like VRE was first evaluated using synthetic noisy data and then tested against the available data. Parameters whose estimates are required for the testing and evaluation of competing or integrated intervention/control measures via mathematical models could not be accurately estimated from the available data. It would be extremely difficult to take advantage of transmission dynamic models in the fight against nosocomial infections unless systemic and careful efforts to collect data are put in place at hospitals.
1 Introduction

People go to the hospital to be treated for their health problems, believing they will be discharged in better health than when they were admitted. Hospitals and health care professionals are committed to improve the health of their patients. However, there are risks associated with the provision of health care with one of the most important being the acquisition of infections at hospitals. The Centers for Disease Control and Prevention (CDC) estimates that 5% to 10% of patients, or more than two million patients each year will get an infection while in a United States hospital with about 90,000 of them dying from such infections [19]. These hospitals-acquired infections or nosocomial infections are infections not present or incubating in a patient at the time of admission to a hospital or health care facility. If symptoms first appear 48 hours or more after a hospital admission or within 30 days after discharge, they are considered nosocomial infections. Studies reveal that about 70% of bacteria that cause nosocomial infections are resistant to at least one antibiotic commonly used to treat them [40]. Three decades ago infection-control measures were put in place to control antibiotic-resistant nosocomial infections and yet these infections have continued to increase. Multidrug-resistant pathogens have become increasingly problematic, especially in the critical care setting, according to the CDC’s National Nosocomial Infection Surveillance System (NNIS System).

Nosocomial infections have been classified as: urinary-tract infections, surgical-incision infections, pneumonia infections, blood-stream infections, skin infections, gastrointestinal-tract infections, and central-nervous-system infections. The urinary-tract infections are the most prevalent type of nosocomial infection as they account for about 35% of all nosocomial infections [40]. Studies have demonstrated that these infections occur after urinary catheterization. The surgical-incision infections account for about 20% [40] of nosocomial infections. Pneumonia infections account for about 15% of nosocomial infections [40] in which the most susceptible individuals are those with chronic obstructive pulmonary disease or those utilizing mechanical ventilation. Blood-stream infections account for another 15% of nosocomial infections [40].

Nosocomial infections may cause severe morbidity in patients leading to extended hospital stays. The average number of extra days a patient has to spend in the hospital varies depending on the type of nosocomial infection: 1 to 4 extra days for urinary-tract infections; 7 to 8 extra days for surgical-incision infections; 7 to 21 extra days for blood-stream infections; and 7 to 30 extra days for pneumonia infections [18]. Further, deaths due to nosocomial infections are the fourth leading cause of death following heart disease, cancer, and stroke [1].

While death and illness are of primary concern, there are also additional consequences associated with the growth of nosocomial infections. The CDC estimated a $5 billion additional cost to US health care in 2000 where due to nosocomial infections. Estimates of this number range from $4.5 billion to $11 billion and up [18]. Nosocomial infections constitute a substantial medical and socioeconomic problem. Finding way of reducing them is of critical importance.
1.1 VRE as an antibiotic resistant pathogen

Most of the nosocomial infections are primarily caused by antibiotic resistant pathogens, such as Vancomycin-resistant Enterococcus (VRE). VRE is the group of bacterial species of the genus enterococcus that is resistant to the antibiotic vancomycin and it can be found in the digestive/gastrointestinal, urinary-tracts, surgical-incision, and bloodstream sites. The CDC during 2006 and 2007 reported that enterococci caused about 1 of every 8 infections in hospitals and about 30% of these are VRE [21].

The bacteria responsible for VRE can be a member of the normal, usually commensal bacterial flora that becomes pathogenic when they multiply in normally sterile sites. Currently there are six different types of vancomycin resistance shown by enterococcus: Van-A, Van-B, Van-C, Van-D, Van-E and Van-F. Of these, only Van-A, Van-B and Van-C have been seen in general clinical practice so far. The significance is that Van-A VRE is resistant to both vancomycin and teicoplanin antibiotics, Van-B VRE is resistant to the antibiotic vancomycin but sensitive to antibiotic teicoplanin, and Van-C is only partly resistant to the antibiotic vancomycin, and sensitive to antibiotic teicoplanin. In addition, VRE has an enhanced ability to pass resistant genes to other bacteria.

1.1.1 Colonization and transmission.

There is a distinction between VRE colonized individuals and VRE infected individuals. The former means that the organism is present in or on the body but is not causing illness while the latter means that the VRE is present and causing illness. Colonization is a VRE carrier state preceding potential infections. This distinction is important in VRE screening [17] since VRE colony counts are similar in the stools of colonized and infected patients. A hospital facility may be adequately reporting its infection rate if its VRE rate is based solely on clinical cultures of VRE infected patients, but it may be underestimating the true burden (and therefore potential transmissibility) of VRE. Screening for patients colonized by VRE provides information about potential sources of illness. The goal of screening is to identify as many VRE colonized patients as possible so that infection control measures that decrease transmission and reduce the number of VRE infected patients can be put in place.

The duration of colonization could last from weeks to months. A study has shown that patients in a university hospital had a mean length of VRE colonization of 204 days (29 weeks) ranging from 4 to 709 days [14]. The factors most associated in predisposing VRE colonization to patients includes: a compromised immune system or nutritional status, the use of catheters (such as urinary or central venous), co-morbidities (e.g., diabetes, renal insufficiency, cancer), length of stay in the hospital, inadequate infection control practice among health care workers (HCW), and prolonged antibiotic used (> 10 days). Hence VRE patients admitted in hospital units such as intensive care and oncology have a greater colonization risk.

Transmission of VRE can occur through contact with colonized or infected individuals (although, there are cases in which VRE acquisition may arise from the patient’s own gut flora). The most frequent form of transmission is by contact, categorized as
direct-contact transmission or indirect-contact transmission. Direct-contact transmission involves direct physical contact (mostly hands) that results in the physical transfer of microorganisms between a susceptible host and a colonized agent such as a patient who is infected or carrying the organism. Indirect-contact transmission involves contact between a susceptible host and a contaminated institutional environment, that includes health care workers (human vectors).

### 1.1.2 Treatment and interventions

People who are colonized with VRE do not usually need treatment. Most VRE infections can be treated with antibiotics other than vancomycin. Laboratory testing of the VRE can determine which antibiotics will work.

A number of interventions have been proposed by CDC Hospital Infection Control Program to try to break the chain of transmission of nosocomial infections, [20]. The CDC Hospital Infection Control Program encourages hospitals to develop their own institution-specific plans particularly when it comes down to VRE, that should stress: prudent vancomycin use by clinicians, hospital staff education regarding vancomycin resistance, early detection and prompt reporting of vancomycin resistance in enterococci by the hospital microbiology laboratory, and immediate implementation of appropriate infection control measures to prevent person-to-person VRE transmission (such as isolation). Isolation procedures consist mostly of frequently hand washing which is considered as the single most important measure needed to reduce the risks of transmitting microorganisms from one person to another or from one site to another on the same patient. Although hand washing may seem like a simple process, it is often performed incorrectly. In addition to hand washing, the systematic use of gloves and gowns play an important role in reducing the risks of transmission of microorganisms. Gloves must be changed between patient contacts and hands should be washed after gloves are removed. Wearing gloves does not replace the need for hand washing, because gloves may have small, non-apparent defects or may be torn during use, that is, hands can become contaminated during the removal of gloves. Failure to change gloves between patient contacts is an infection control hazard.

### 1.2 The role of mathematical and statistical modeling

Mathematical and statistical models have made substantial contributions to our understanding of the epidemiological dynamic of infections [12, 2, 32]. Hence, they have been valuable tools to predict and explain the epidemiology of nosocomial infections. Many of the models developed to describe the transmission of nosocomial infection in a health care setting have been based on the Ross-Macdonald model [42] where the transmission of pathogens in health care settings considers health care workers as vectors and patients as hosts [35, 36, 46, 24, 3]. Also, models for nosocomial infections have established a fundamental distinction between hospital-acquired infections and community-acquired infections. This is because, in general, the average patient stay in about a week. The hospital population turns over rapidly, individuals bring in bacteria from outside. Patient
discharge brings bacteria back into the community.

The utility of these models have had as a goal to explain the spread of infections, specifically by studying the impact of infection control measures such as patient isolation, hand-washing, and bacterial-control among others. Lipsitch, et al., [35, 36] developed a mathematical model of the transmission and spread of antibiotic-resistance bacteria. His model considers colonization and infection by antibiotic-sensitive and resistant bacteria in a hospital setting, in which he tries to explain the rapid rate of change in response to interventions, the efficacy of control measures, and the use of one drug as an individual risk factor for the acquisition of resistance to other drugs. Also, Webb et al. [46] has proposed a model that describes transmission and spread of antibiotic-resistant bacteria by connecting two environmental levels: bacteria level in infected host where non-resistant and resistant strains are produced in the bodies of individual patients and patients’ level where susceptible patients are cross-infected by health care workers (who become contaminated by contact with infected patients). With this model Webb also tries to explain the efficacy of therapy regimens and hospital infection control measures. Recently, Chow, et al., [16], developed a mathematical model that looks at different strategies for curbing the prevalence of antibiotic resistance in nosocomial infections. Their model suggests that antimicrobial cycling and patient isolation may be effective approaches when patients are harboring dual-resistant bacteria. In fact, isolation of patients dramatically reduces the persistence of dual resistance. However, it was difficult to control antimicrobial resistance through the exclusive use of integrated microbial management approaches that focus entirely on the prescription of antibiotics. These researchers, found that isolating individuals harboring multi-drug resistance infectious could be quite effective.

1.3 Objectives

Mathematical and statistical models are valuable tools to predict and explain the epidemiology of nosocomial infections. The overall objectives of this research are to develop mathematical and statistical models with compatible methodology to improve the understanding of the transmission of Vancomycin-Resistant Enterococcus (VRE) in hospitals. The development of plausible models is based on the epidemiological knowledge of VRE in a setting that allows for the implementation of infection control measures in hospitals. We focus in the connection of these models with unpublished VRE surveillance data from a hospital in order to estimate some of the parameters that govern the underlying transmission infection dynamics. The idea is to use these models as the foundation of a statistical model that can be used to probe the value of surveillance data and quantify the levels of VRE transmission.

In Section 2 we review the inverse problem methodology used to estimate parameters. In Section 3 we develop a simple VRE epidemic model that plays a crucial role in the effort to control the growing threat posed in a hospital unit by this antibiotic-resistant bacteria. Given the small scale of patients population in hospital units we first propose a stochastic continuous time Markov Chain (MC) model and show how it can be converted to an equivalent continuous time ordinary differential equation (ODE) model.
We provide a discussion of the qualitative features of the ODE model. We calibrate this ODE model using the VRE surveillance data and the inverse problem methodology described in Section 2. In Section 4 we present simulations in which the MC and the ODE models are compared and discrepancies analyzed. In order to assess the impact of infection control measures, the effect that health care worker hand-hygiene compliance has on the basic reproductive number (i.e., average number of secondary VRE colonized patients generated by a primary case of VRE colonized patient in a VRE-free hospital unit) is studied. In Section 5 a discrete VRE model with delay that incorporates more details about the infection control procedures that is employed in hospitals units is introduced. We calibrate this model using the available VRE surveillance data. We present in Section 6 the methodology used to estimate parameters when the surveillance data contains missing data. Finally, in Section 7 we collect the conclusions of this effort.

2 Inverse Problem Methodology

Closely tied to the formulation of mathematical models is the need to estimate the parameters (and initial conditions) involved. These parameters often describe the stability and control behavior of the system. If these were to be known, we could investigate what is called a “forward” or simulation problem. Unfortunately not all parameters are directly measurable and finding these unknown parameters using data is an “inverse” or parameter estimation problems. Solutions to the inverse problem is one of the ways used to gain a deeper understanding of the system’s characteristics. Estimation of these parameters from experimental data of the system is thus an important step in the analysis of dynamical systems. In this section we will be reviewing a well known inverse problem methodology for deterministic dynamical systems. The methodology is described for the “scalar case” in which only one type of data is used. Further, whether or not we can estimate these parameters with a specific set of experimental data is certainly of utmost relevance. Given an experimental data set, a mathematical model may be more sensitive to some parameters than others, the dependence between the parameters can impact the well-posedness of an inverse problem. Therefore, we limit the analysis to the subset of parameters for which the mathematical model is most sensitive. A priori analysis needed to identify the type of inverse problem formulation or the subset of parameters to be estimated from a given data set is based on the work done in [23], and is reviewed in Section 2.6.

2.1 Mathematical and statistical models

The mathematical model is described by

\[
\frac{dx(t)}{dt} = g(t, x(t; \theta), \theta)
\]

\[x(t_0) = x_0,\]

(1)
with parameter vector \( \theta \in \mathbb{R}^p \), \( x(t) = (x_1(t), ..., x_N(t))^T \in \mathbb{R}^N \), and an observational process \( y(t_j) = Cx(t_j; \theta) \in \mathbb{R}^m \) for \( j = 1, ..., n \), where \( C \) is an \( m \times N \) matrix. The mathematical model is assumed to be well-posed (i.e., the existence of unique solution that depends smoothly on the parameters and initial data).

Let \( Y_j \) for \( j = 1, ..., n \) be longitudinal data observations corresponding to the experimental data of the observational process. Since in general \( Y_j \) is not assumed to be free of error (i.e., error in data collection process), \( Y_j \) will not be exactly \( y(t_j) \). We can thus envision experimental data as generally consisting of observations from a “perfect” model plus an error component represented by

\[
Y_j = f(t_j; \theta_0) + \varepsilon_j \quad \text{for} \quad j = 1, ..., n, \tag{2}
\]

where \( \theta_0 \) corresponds to the “true” parameter that generate the observations \( Y_j \) for \( j = 1, ..., n \). The function \( f(t_j, \theta) \) corresponds to the observation process of Model (1) and depends in the parameters \( \theta \) in a nonlinear fashion. If the \( \varepsilon_j 's \) are generated from a probability distribution \( P \), then the statistical model has the following assumptions:

1. The measurement errors \( \varepsilon_j \) for \( j = 1, ..., n \) have mean zero, i.e., \( E(\varepsilon_j) = 0 \);
2. The measurement errors \( \varepsilon_j \) for \( j = 1, ..., n \) are independent, i.e., \( P(\varepsilon_1, ..., \varepsilon_n) = \prod_{j=1}^n P(\varepsilon_j) \);
3. The measurement errors \( \varepsilon_j \) for \( j = 1, ..., n \) have same variance, i.e., \( \text{var}(\varepsilon_j) = \sigma_0^2 < \infty \), and are identically distributed random variables for all \( t_j \).

Assumption (1) ensures that the function \( f(t, \theta) \) for mean response is correctly specified. In other words, that Model (1) is a correct description of the process being modeled. Note that \( Y_j \) in Model (2) is a random process implying that the solutions to the corresponding estimates are also random variables and are obtained with \( \{Y_j\} \) replaced by a given realization.

### 2.2 Maximum likelihood estimation (MLE)

Under the case that the error process is known and is normally distributed, i.e., \( \varepsilon_j \sim \mathcal{N}(0, \sigma_0^2) \) and hence \( Y_j \sim \mathcal{N}(f(t_j, \theta_0), \sigma_0^2) \). We can then determine \( \theta_0 \) and \( \sigma_0^2 \) by seeking the maximum of the likelihood function for \( \varepsilon_j = Y_j - f(t_j, \theta) \) defined by

\[
L(\theta, \sigma^2|Y) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} [Y_j - f(t_j, \theta)]^2 \right\}
\]

or equivalent

\[
\log L(\theta, \sigma^2|Y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^n [Y_j - f(t_j, \theta)]^2. \tag{3}
\]
By differentiating (3) with respect to $\theta$ (with $\sigma^2$ fixed), setting the resulting equation equal to zero, and solving for $\theta$ we have that

$$
\sum_{j=1}^{n} [Y_j - f(t_j, \theta)] \nabla f(t_j, \theta) = 0. \tag{4}
$$

Solving (4) is the same as the least square optimization

$$
\hat{\theta}_{MLE}(Y) = \arg \min_{\theta \in \Theta} \sum_{j=1}^{n} [Y_j - f(t_j; \theta)]^2. \tag{5}
$$

Similarly, by differentiating (3) with respect to $\sigma^2$ (with $\theta = \hat{\theta}_{MLE}$ fixed) we have that

$$
\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{j=1}^{n} [Y_j - f(t_j, \hat{\theta}_{MLE})]^2. \tag{6}
$$

Finally, a second derivative test verifies that the expression above for $\hat{\theta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ maximize (3).

### 2.3 Ordinary least squares (OLS) estimation

If the error distribution is unknown, an OLS optimization procedure is often employed. This method can be viewed as minimizing the distance between the data and the model were all observations are treated as of equal importance. The OLS method defines “best” as when the norm square of the residuals is a minimum

$$
\theta_{OLS} = \theta_{OLS}^n = \arg \min_{\theta \in \Theta} \sum_{j=1}^{n} [Y_j - f(t_j, \theta)]^2. \tag{7}
$$

This corresponds to solving for $\theta$ in

$$
\sum_{j=1}^{n} [Y_j - f(t_j, \theta)] \nabla f(t_j, \theta) = 0.
$$

Since we do not know the distribution, under asymptotic theory we have

$$
\theta_{OLS} = \theta_{OLS}^n \sim N_\mu(\theta_0, \Sigma_0^n) \tag{8}
$$

where the covariance matrix $\Sigma_0^n$ is defined by

$$
\Sigma_0^n \equiv \sigma_0^2 |n\Omega|^{-1}
$$

with

$$
\Omega_0 \equiv \lim_{n \to \infty} \frac{1}{n} \chi_n^2(\theta_0)^T \chi_n^2(\theta_0)
$$
where $\chi^n(\theta) = \{\chi_{jk}\}$ is the sensitivity matrix with

$$\chi_{jk}(\theta) = \frac{\partial f(t_j, \theta)}{\partial \theta_k} \quad j = 1, \ldots, n \quad \text{and} \quad k = 1, \ldots, p.$$  

The error variance $\sigma_0^2$ is approximated by

$$\hat{\sigma}_{OLS}^2 = \frac{1}{n - p} \sum_{j=1}^{n} [Y_j - f(t_j, \hat{\theta}_{OLS})]^2 \quad (9)$$

as the bias adjusted estimate for $\sigma_0^2$. The covariance matrix $\Sigma^n_0$ is approximated by

$$\hat{\Sigma}_{OLS}^n = \hat{\sigma}_{OLS}^2 [\chi^T(\hat{\theta}_{OLS})\chi(\hat{\theta}_{OLS})]^{-1}. \quad (10)$$

Therefore

$$\hat{\theta}_{OLS} \sim N_p(\theta_0, \Sigma^n_0) \approx N_p(\hat{\theta}_{OLS}, \hat{\Sigma}_{OLS}^n). \quad (11)$$

The standard errors for $\hat{\theta}_{OLS}$ can be calculated by taking the square roots of the diagonal elements of the covariance matrix $\hat{\Sigma}_{OLS}^n$, i.e. $SE(\hat{\theta}_{OLS}) = \sqrt{(\hat{\Sigma}_{OLS}^n)_{kk}}$ for $k = 1, \ldots, p$. The sensitivity matrix can be calculated by solving the following sensitivity equations:

$$\frac{d}{dt} \frac{\partial x}{\partial \theta} = \frac{\partial g}{\partial x} \frac{\partial x}{\partial \theta} + \frac{\partial g}{\partial \theta} \quad (12)$$

where $\partial g/\partial x$ is $N \times N$ and both, $\partial x/\partial \theta$ and $\partial g/\partial \theta$, are $N \times p$. Then the elements of $\chi(\hat{\theta}_{OLS})$ are computed as

$$\chi_{jk}(\hat{\theta}_{OLS}) = \frac{\partial f(t_j, \hat{\theta}_{OLS})}{\partial \theta_k} = C \frac{\partial x(t_j, \hat{\theta}_{OLS})}{\partial \theta_k}. \quad (13)$$

### 2.4 Generalized least squares (GLS) estimation

If the error distribution is unknown and the assumption of constant variance of the error (3) in the longitudinal data does not hold, a generalized least square (GLS) optimization procedure should be employed. In this case we need to formulate a new statistical model to take into consideration the non-constant error variability. If we can assume that the size of the error depends on the size of the observed quantity, the statistical model (i.e., relative error model) is given by

$$Y_j = f(t_j, \theta_0)(1 + \varepsilon_j) \quad \text{for} \quad j = 1, \ldots, n. \quad (14)$$

where $Y_j \sim N(f(t_j, \theta_0), \sigma_0^2 f^2(t_j, \theta_0))$ which derives from the assumptions (1) and (3). In this case, GLS can be view as minimizing the distance between the data and the model while taking into account unequal quality of the observations. The GLS method defines
“best” as \( \hat{\theta}_{GLS} \) obtained after solving

\[
\sum_{j=1}^{n} f^{-2}(t_j, \theta_{GLS})[Y_j - f(t_j, \theta_{GLS})] \nabla f(t_j, \theta_{GLS}) = 0.
\] (15)

The idea is to assign to each observation a weight that reflects the uncertainty of the measurement. Under asymptotic theory we find

\[
\hat{\theta}_{GLS} = \hat{\theta}_{GLS}^n \sim N_p(\theta_0, \Sigma_0^n)
\] (16)

where

\[
\Sigma_0^n \approx \sigma_0^2[F^T(\theta_0)W(\theta_0)F(\theta_0)]^{-1}
\]

with

\[
F(\theta) = \begin{bmatrix}
\frac{\partial f(t_1, \theta)}{\partial \theta_1} & \cdots & \frac{\partial f(t_1, \theta)}{\partial \theta_p} \\
\vdots & & \vdots \\
\frac{\partial f(t_n, \theta)}{\partial \theta_1} & \cdots & \frac{\partial f(t_n, \theta)}{\partial \theta_p}
\end{bmatrix}
\]

and \( W^{-1}(\theta) = diag(f^2(t_1, \theta), \ldots, f^2(t_n, \theta)) \). Using the estimates we have the covariance matrix approximation

\[
\hat{\Sigma}_{GLS}^n = \hat{\Sigma}_{GLS}^n \approx \hat{\Sigma}_{GLS}^n [F^T(\hat{\theta}_{GLS})W(\hat{\theta}_{GLS})F(\hat{\theta}_{GLS})]^{-1}
\] (17)

and the error variance approximation

\[
\hat{\sigma}_{GLS}^2 = \frac{1}{n-p} \sum_{j=1}^{n} \frac{1}{f^2(t_j, \hat{\theta}_{GLS})}[Y_j - f(t_j, \hat{\theta}_{GLS})]^2.
\] (18)

Therefore

\[
\theta_{GLS} \sim N_p(\theta_0, \Sigma_0^n) \approx N_p(\hat{\theta}_{GLS}, \hat{\Sigma}_{GLS}^n).
\] (19)

We can also calculate the standard errors for \( \hat{\theta}_{GLS} \) by taking the square roots of the diagonal elements of the covariance matrix \( \hat{\Sigma}_{GLS}^n \). The sensitivity matrix \( \chi(\hat{\theta}_{GLS}) = \{\chi_{jk}\} \) can be calculated using the sensitivity equations in (12). The estimate \( \hat{\theta}_{GLS} \) can be solved directly or iteratively using following procedure:

1. Set \( k = 0 \). Estimate the initial \( \hat{\theta}_{GLS}^{(k)} \) by using the OLS estimate in (7);
2. Form the weights \( \hat{w}_j = f^{-2}(t_j, \hat{\theta}_{GLS}^{(k)}) \);
3. Estimate \( \hat{\theta}_{GLS}^{(k+1)} \) by solving

\[
\sum_{j=1}^{n} \hat{w}_j(t_j, \theta_{GLS})[Y_j - f(t_j, \theta_{GLS})] \nabla f(t_j, \theta_{GLS}) = 0;
\] (20)
4. Set $k = k + 1$ and return to 2. Terminate the process when two successive estimates for $\hat{\theta}_{GLS}$ are “close” to one another.

2.5 Residual analysis

Since there is no way of knowing beforehand which optimization procedure (OLS or GLS) needs to be employed, a residual analysis can be conducted to check which assumptions are satisfied. A plot of the residuals $r_j = y_j - f(t_j, \hat{\theta})$ versus time $t_j$, where $y_j$ is a realization of $Y_j$, should appear random for the assumption of independency of error (2) to be satisfied. The constant variance assumption (3) of the errors can be checked by a random pattern in the plot of the residuals $r_j = y_j - f(t_j, \hat{\theta})$ versus model $f(t_j, \hat{\theta})$. Confidence intervals can be misleading if using OLS procedure with nonconstant variance as well as if using GLS procedure with constant variance. This is because in the OLS estimator (7) the sum of square deviations between the experimental data and the observational process receive equal weight in the minimization, which implies that all experimental observations are subject to the same degree of uncertainty. On the other hand, for the GLS estimator (15) each deviation is weighted in inverse proportion to the magnitude of the uncertainty. Therefore, if one uses OLS when GLS is indicated, one has that all deviations are weighted equally, resulting in measurements of lower quality being given too much importance in determining the fit.

2.6 Subset selection methodology

In order to identify the subset of parameters that has a high sensitivity to the mathematical model, we use the identifiability analysis suggested in [23]. The parameter selection or parameter identifiability consists of the following two criteria:

1. Select the parameter vectors that has a full rank sensitivity matrix, $\chi^n(\hat{\theta})$. Its degree of sensitivity is measured in the form of its condition number $k(\chi^n(\hat{\theta}))$ defined below in (2.28);

2. For each parameter vector selected in the first criteria, estimate its degree of uncertainty. Its degree of uncertainty is measured in the form of the parameter selection score $\psi(\hat{\theta})$ defined by (2.29).

Motivation behind the first criteria is as follows. If $\theta_0$ is the true parameter, then $\Delta \theta = \theta - \theta_0$ denotes a local perturbation from $\theta_0$. This gives rise to local perturbation in the output of a model $\Delta y = y(t, \theta) - y(t, \theta_0)$. Then, by a first order Taylor approximation we obtain the approximate relationship

$$\Delta y \approx \chi \Delta \theta. \quad (21)$$

A parameter vector is sensitivity identifiable (locally) if the above equation can be solved uniquely for $\Delta \theta$. This is the case if $\text{rank}(\chi) = p$, or equivalently, if and only if the Fisher
information matrix, \( F = \chi^T(\hat{\theta})\chi(\hat{\theta}) \) is nonsingular or
\[
\det(\chi^T\chi) \neq 0. \tag{22}
\]
The Fisher information matrix measures the amount of information that an observation process carries about an unknown parameter \( \theta \). If near-singularity of \( F \) is present then the approximation of the covariance matrix and consequently the calculation of standard error and confidence intervals for the estimated parameters are affected.

By focusing on properties of the sensitivity matrix \( \chi(\theta) \) rather than the Fisher information matrix, a singular value decomposition (SVD) of the sensitivity matrix plays a crucial role in uncertainty quantification. The SVD of the sensitivity matrix is denoted as:
\[
\chi(\theta) = U \begin{bmatrix} \Lambda \\ 0 \end{bmatrix} V^T \tag{23}
\]
where \( U = [U_1 \ U_2] \) and \( V \) are \( n \times n \) and \( p \times p \) orthogonal matrixes, with \( U_1 \) containing the first \( p \) columns of \( U \) and \( U_2 \) containing the last \( n - p \) columns. \( \Lambda \) is a \( p \times p \) diagonal matrix defined as \( \Lambda = diag(s_1, ..., s_p) \) with \( s_1 \geq s_2 \geq ... \geq s_p \geq 0 \), and \( 0 \) denotes an \( (n - p) \times p \) matrix of zeros.

Suppose that \( f(t, \theta) \) is well approximated by its linear Taylor expansion around \( \theta_0 \) as
\[
f(t, \theta) \approx f(t, \theta_0) + \chi(\theta_0)(\theta - \theta_0). \tag{24}
\]
By the statistical model \( Y(t) = f(t, \theta_0) + \epsilon \) and the above equation, we have that
\[
Y(t) - f(t, \theta) = -\chi(\theta_0)(\theta - \theta_0) + \epsilon. \tag{25}
\]
Then, we can define the estimator \( \hat{\theta}_{OLS} \) that minimizes \( |Y(t) - f(t, \theta)|^2 \) and using the invariance property of the Euclidean norm (i.e., \( |w|^2 = w^T w = w^T I w = w^T U U^T w = |U^T w|^2 \)) we have that
\[
|Y(t) - f(t, \theta)|^2 = |-\chi(\theta_0)(\theta - \theta_0) + \epsilon|^2
= \left| U^T \left( -U \begin{bmatrix} \Lambda \\ 0 \end{bmatrix} V^T (\theta - \theta_0) + \epsilon \right) \right|^2
= \left| -\begin{bmatrix} \Lambda \\ 0 \end{bmatrix} V^T (\theta - \theta_0) + \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \epsilon \right|^2. \tag{26}
\]
Solving \( | -\Lambda V^T (\theta - \theta_0) + U_1^T \epsilon |^2 = 0 \) for \( \theta \) we have
\[
0 = -\Lambda V^T (\theta - \theta_0) + U_1^T \epsilon,
\theta - \theta_0 = (\Lambda V^T)^{-1} U_1^T \epsilon.
\]
This implies

\[
\hat{\theta}_{OLS} = \theta_0 + V\Lambda^{-1}U^T_1\epsilon
\]

\[
= \theta_0 + \sum_{i=1}^p \frac{1}{s_i} v_i u_i^T \epsilon.
\]

(27)

Note that if \( s_i \to 0 \), the estimator is particular sensitive to \( \epsilon \).

If \( \chi(\theta) \in \mathbb{R}^{n \times p} \) is a full rank sensitivity matrix (i.e, \( \text{rank}(\chi(\theta)) = p \)) its \textit{condition number} \( k \) is defined as the ratio of the largest to smallest singular value given by

\[
k(\chi(\theta)) = \frac{s_1}{s_p}.
\]

(28)

which provides a degree of singularity due to perturbations and hence a criteria for parameter identifiability. If the columns of \( \chi(\theta) \) are nearly dependent then (28) is large.

Motivation of the second criteria is the uncertainty in the parameters of a particular subset combination that can be quantified using the standard errors, \( SE(\theta) \). In order to compare the degree of variation from one parameter to another the coefficient of variation, \( CV = SE(\theta)/\theta \in \mathbb{R}^p \) is used. The \( CV \) allows one to compare the parameters even if the parameter estimates are substantially different in units and scales. Hence, a second criteria can be established by considering the parameter \textit{selection score}

\[
v(\theta) = |CV(\theta)|.
\]

(29)

In (29) a value near zero indicates lower uncertainty possibilities in the estimation while large values suggest a possibility of a wide uncertainty in at least some of the estimates.

In general, the algorithm that searches all possible parameter combinations and selects the ones satisfying criteria (1) and (2) is the following:

1. \textbf{Combinatorial search}. For a fixed \( j = 1, ..., p \), calculate the set

\[
S_p = \{ \theta = (q_1, ..., q_j) \in \mathbb{R}^p | q_k \in \mathbb{R}, q_k \neq q_l \ \forall k, l = 1, ..., j \}
\]

where the set \( S_p \) collects all the parameter vectors obtained from the combinatorial search;

2. \textbf{Full rank test}. Calculate the set of feasible parameters \( \Theta_p \) as

\[
\Theta_p = \{ \theta | \theta \in S_p \subset \mathbb{R}^p, \text{rank}(\chi(\theta) = p) \}.
\]

Calculate the condition number defined by

\[
k(\chi(\theta)) = \frac{s_1}{s_p};
\]

3. \textbf{Standard error test}. For every \( \theta \in \Theta_p \), calculate a vector of coefficients of variation \( CV(\theta) \in \mathbb{R}^p \) by

\[
CV_i = \frac{\sqrt{\sum(\theta)_{ii}}}{\hat{\theta}_i},
\]
for $i = 1, \ldots, p$ and $\sum(\theta) = \sigma^2_i \chi(\theta)^T \chi(\theta)^{-1} \in \mathbb{R}^{p \times p}$. Calculate the parameter selection score as $v(\theta) = |CV(\theta)|$.

3 General VRE Epidemic Model and Surveillance Data

In this Section 3.1 we present a description of the VRE surveillance data used in this study. In Section 3.2 we provide a detailed description of model assumptions. In Section 3.3 and 3.4 a simple stochastic model and its analogous deterministic version of the transmission dynamic of VRE in a hospital unit are developed. We calibrate the deterministic model to the VRE surveillance data by estimating directly and indirectly the parameters underlying the model. In Section 3.5 we collect the parameters that were estimated directly from the VRE surveillance data. In Section 3.6 we describe the formulation of the inverse problem used to estimate the parameters that were difficult to estimate in the previous section.

3.1 VRE data description

The data under study is from the VRE Infection Control database of the Department of Quality Improvement Support Service of Yale-New Haven Hospital. This hospital has a surveillance program for VRE in three medical wards: MICU (medical intensive care unit), EP 9-5 (general medical unit) and 9-WEST (medical oncology unit). Data reports the number of VRE cases occurred on admission (includes patients transferred), patients’ length of stay, number of isolated patients due to VRE colonization, admission swab culture compliance, and health care worker contacts precautions compliance. Data collection occurred during the period of October 2004 to December 2005 for the EP 9-5 unit, January 2005 to January 2007 for the MICU, and January 2005 to January 2007 for the 9-WEST unit. The mean number of in-patients for each ward per day was 28 for the EP 9-5 unit (total of 29 beds), 13 patients for the MICU (total of 15 beds), and 31 patients for the 9-West unit (total of 37 beds).

3.1.1 VRE surveillance protocol

Wards protocol requires rectal swabbing all patients on admission, and once a week for VRE surveillance. Compliance is not 100%, as the mean percentage of swab cultures taken on admitted patients per day was 66% for the EP 9-5 unit, 64% for the MICU, and 77% for the 9-WEST unit. Swab-test results were returned usually 48 hours after admission. If a patient tested VRE positive, he/she would be isolated. The isolation procedure consisted of contact precautions by the use of gowns, gloves, and the location of a patient in a single room or in a room with another patient who is also VRE positive (the three medical wards are single-bed rooms). There was a difference in the isolation procedure for the MICU ward. This unit implements pre-emptive isolation which consists on isolating every patient on admission while waiting for the result of the rectal swab performed on admission. If negative, the isolation for that patient gets removed; if positive then that patient remains isolated. For all three wards, if a readmission patient
has a positive VRE culture in the past, he/she does not get the rectal swab on admission and is isolated immediately. The isolation report is performed on weekdays (no weekends or holidays), providing a count of VRE isolations. The mean number of isolated VRE colonized patients per day was 8.98 (std=2.96) patients for the EP 9-5 unit, 4.98 (std=1.98) patients for the MICU and 9.39 (std=2.90) patients for the 9-WEST unit.

### 3.2 General VRE epidemic model

We consider a simple compartmental model in which patients in a hospital unit are classified as either uncolonized $U(t)$, VRE colonized $C(t)$ or VRE colonized in isolation $J(t)$, as depicted in the compartmental schematic of Figure 1. A description of the variables and parameters used in our model are given in Table 1. The model assumptions are listed in the following itemization.

1. Compartments are homogeneous that is, each patient is considered equally likely to be in contact with a health care worker in any time interval, equally likely to be VRE colonized, and, if VRE colonized at a given time, equally likely to transmit the pathogen at a given time.

2. The transition from one compartment to another follows an exponential distribution and the expected mean duration within a compartment is given by the inverse of the parameters of the exponential distribution.

3. Patients are admitted to the hospital unit at a rate $\Lambda$ per day and some fraction $m$ are already VRE colonized. Therefore VRE colonized patients can enter the hospital at a rate $m\Lambda$ and remain for an average duration of time $1/(\mu_2 + \alpha)$.

4. New uncolonized patients enter the hospital at a rate $(1 - m)\Lambda$.

5. Isolated patients remain admitted for an average duration of time $1/\mu_2$.

6. It is assumed that an average patient in the population makes $\beta N$ effective contacts (i.e., contact sufficient to lead to VRE colonization) with other patients per unit time through health care workers, where $N$ is the total population size. This assumption of a rate of contact per infective proportional to the population size $N$ is called mass action incidence.

7. The hand-hygiene policy applied to health care workers on isolated VRE colonized patients reduces infectivity by a factor of $\gamma$ ($0 < \gamma < 1$). This assumption means that isolated VRE colonized patients make fewer contacts than regular patients, so transmission of the bacteria by these isolated members has an infectivity factor $(1 - \gamma)$.

8. Uncolonized patients become VRE colonized at a rate proportional to the prevalence of patients carrying the bacteria. We use assumptions (6) and (7) to compute the rate. Since the probability is $U/N$ that a random contact by a VRE colonized patient is with an uncolonized patient, the number of new colonization in unit
time per infective is \((\beta N)(U/N)\). This yields a rate of new VRE colonization \((\beta N)(U/N) [C + (1 - \gamma)J] = \beta U[C + (1 - \gamma)J]\).

9. It is assumed VRE colonization periods last from weeks to months. However, because spontaneous decolonization occurs infrequently, clearance of the bacteria is not considered in the model.

10. VRE colonized patients are not treated for VRE.

11. It is assumed that all patients on admission are swab tested for VRE.

12. The waiting time for the results of the swab-test cultures is assumed to be the same for all patients in any time period. After results are returned, VRE colonized patients are moved into isolation at a rate \(\alpha\).

13. As a simplification, we assume that the total number of patients remains fixed (i.e., overall admission rate equals overall discharge rate, \(\Lambda = \mu_1 U + \mu_2 (C + J)\) and VRE colonization confers no additional mortality. Then the total population of patients can be written as \(N = U + C + J\).

### 3.3 The VRE stochastic model

We model the dynamic of the VRE colonization of patients in a hospital unit as a continuous time Markov Chain (MC) with discrete state space embedded in \(\mathbb{R}^3\). Therefore, the population of patients is considered discrete (i.e., VRE colonization occurs in units of whole individuals) and the timing of events is a probabilistic process. The state of the MC at time \(t\) is denoted by \(\{U(t) = i, C(t) = j, J(t) = k\}, t \geq 0\) and \(i, j, k \in \{0, 1, \ldots N\}\). Then, the probability that during a small time interval, \(dt\), of transiting from one state
Table 1: Model Parameters

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U(t)$</td>
<td>Number of uncolonized patients</td>
<td>Individuals</td>
</tr>
<tr>
<td>$C(t)$</td>
<td>Number of VRE colonized patients</td>
<td>Individuals</td>
</tr>
<tr>
<td>$J(t)$</td>
<td>Number of VRE colonized patients in isolation</td>
<td>Individuals</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Lambda$</td>
<td>Patients admission rate</td>
<td>Individuals/day</td>
</tr>
<tr>
<td>$m$</td>
<td>VRE colonized patients on admission rate</td>
<td>Dimensionless</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Effective contact rate</td>
<td>1/day</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>HCW hand hygiene compliance rate</td>
<td>Dimensionless</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Patient Isolation rate</td>
<td>1/day</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>Uncolonized patients discharged rate</td>
<td>1/day</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>VRE colonized patients discharged rate</td>
<td>1/day</td>
</tr>
</tbody>
</table>

We present next a model description underlying these equations. In the VRE epidemic model a constant population is assumed in which the hospital remains full for all $t$ (Assumption (13)). Hence, the admission of a patient in either compartments $U$ or $C$ are dependent events on the discharged in either compartment $U$ or $C$ or $J$ (or vice versa). In our MC model it is assumed that when a patient is discharged from the hospital, he/she is immediately replaced by an admission into the compartment $U$ with probability $(1 - m)$ or into the compartment $C$ with probability $m$. Equation (3.1) is
the probability of entering compartment $C$ by either an admission (due to a discharge in compartment $U$) or by effective colonization. Equation (3.2) is the probability of entering compartment $C$ by an admission due to a discharge in $J$. Equation (3.3) is the probability of admission to compartment $U$ by a discharge in $C$. Equation (3.4) is the probability of admission into compartment $U$ by a discharge in $J$. Equation (3.5) is the probability of moving a VRE colonized patient into isolation. Finally, Equation (3.6) is the probability that none of the states changes due to: an uncolonized patient being discharged and replaced back into the $U$ compartment, or a VRE colonized patient in $C$ being discharged and replaced back into the $U$ compartment, or a VRE colonized patient in $C$ being discharged and replaced back into the $C$ compartment, or no event occurs.

When dividing these probabilities by $dt$ and taking the limit when $dt$ tends to $0+$, we obtain the rates of transitions that are given in Table 2. Since this is a stochastic model, the rates represent the mean number of transitions that can be expected in a given period, with the actual numbers distributed about these means. Hence, the rates determine the frequencies of occurrence through time for the transitions or events.

Letting $R_i(t)$ for $i = 1, \ldots, 6$, be the number of times the $i^{th}$ transition has occurred by time $t$. Then, the state of the system at time $t$ can be written as

\[
U(t) = U(0) - R_1(t) - R_4(t) - R_5(t) + (1 - m)(R_1(t) + R_2(t) + R_3(t))
\]
\[
C(t) = C(0) - R_2(t) + R_4(t) + R_5(t) - R_6(t) + m(R_1(t) + R_2(t) + R_3(t))
\]
\[
J(t) = J(0) - R_3(t) + R_6(t),
\]

where $R_i(t)$ is a counting process with intensity $\lambda_i(U(t), C(t), J(t))$ given by

\[
R_i(t) = Y_i \left( \int_0^t \lambda_i(u(s), c(s), j(s)) ds \right), \quad i = 1, \ldots, 6,
\]

with $Y_i$ as independent unit Poisson processes. Note that the state of the system is $\{U(s), C(s), J(s)\}$ and hence each $\lambda_i(U(s), C(s), J(s))$ is constant between transition times. Also, note that sample paths $r_i(t)$ of $R_i(t)$ are given in terms of sample paths $\{u(t), c(t), j(t)\}$ of $\{U(t), C(t), J(t)\}$ by

\[
r_i(t) = Y_i \left( \int_0^t \lambda_i(u(s), c(s), j(s)) ds \right), \quad i = 1, \ldots, 6.
\]

### 3.4 The VRE deterministic model

We are also interested in the deterministic approximation of the continuous time discrete state Markov Chain model already described when the population size $N$ is large. The deterministic approximation is based on Kurtz’s approximation in mean theory [29].

Let $X^N(t) = U(t)/N$, $Y^N(t) = C(t)/N$, $Z^N(t) = J(t)/N$ be the patients units per system size or the proportion in the stochastic process with sample paths $\{x^N(t), y^N(t), z^N(t)\}$. 

19
Table 2: Transition rates

<table>
<thead>
<tr>
<th>Event</th>
<th>Effect</th>
<th>Transition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discharge of uncolonized patient</td>
<td>(U,C,J)=(i-1,j,k)</td>
<td>$\lambda_1 = \mu_1 U$</td>
</tr>
<tr>
<td>Discharge of VRE colonized patient</td>
<td>(U,C,J)=(i,j-1,k)</td>
<td>$\lambda_2 = \mu_2 C$</td>
</tr>
<tr>
<td>Discharge of VRE colonized patient in isolation</td>
<td>(U,C,J)=(i,j,k-1)</td>
<td>$\lambda_3 = \mu_2 J$</td>
</tr>
<tr>
<td>Patient colonization due to VRE colonized patients</td>
<td>(U,C,J)=(i-1,j+1,k)</td>
<td>$\lambda_4 = \beta UC$</td>
</tr>
<tr>
<td>Patient colonization due to VRE colonized patients in isolation</td>
<td>(U,C,J)=(i-1,j+1,k)</td>
<td>$\lambda_5 = \beta (1-\gamma) UJ$</td>
</tr>
<tr>
<td>Isolation of VRE colonized patient</td>
<td>(U,C,J)=(i,j-1,k+1)</td>
<td>$\lambda_6 = \alpha C$</td>
</tr>
<tr>
<td>Admission of uncolonized patient</td>
<td>$U=i+1$</td>
<td>$(1-m)(\lambda_1 + \lambda_2 + \lambda_3)$</td>
</tr>
<tr>
<td>Admission of VRE colonized patient</td>
<td>$C=j+1$</td>
<td>$m(\lambda_1 + \lambda_2 + \lambda_3)$</td>
</tr>
</tbody>
</table>

We rescale the rates $\lambda_i$ for $i = 1, ..., 6$ as follows:

$$
\lambda_1 = \mu_1 u(t) = N\mu_1 x(t), \quad \lambda_4 = \beta u(t)c(t) = N^2 \beta x(t)y(t), \\
\lambda_2 = \mu_2 c(t) = N\mu_2 y(t), \quad \lambda_5 = \beta (1-\gamma) u(t)j(t) = N^2 \beta (1-\gamma) x(t)z(t), \\
\lambda_3 = \mu_2 j(t) = N\mu_2 z(t), \quad \lambda_6 = \alpha c(t) = N\alpha y(t).
$$

Using these rates we can obtain an approximation for the sample paths $r_i(t)$ of (38) by applying the SLLN for the Poisson Process (i.e., $Y(N\mu)/N \approx \mu$). Therefore, we find

$$
\begin{align*}
    r_1^N(t) &= \frac{r_1(t)}{N} = \frac{1}{N} Y_1 \left( \int_0^t \lambda_1(u(s))ds \right) \\
             &= \frac{1}{N} Y_1 \left( \int_0^t \mu_1 x(s)ds \right) \\
             &\approx \int_0^t \mu_1 x(s)ds.
\end{align*}
$$

The approximations $r_i^N(t)$ for $i = 2, ..., 6$, can be obtained similarly. When dividing both sides of each equation in (36) by $N$ and applying the approximations for $r_i(t)$, we can write the system of integral equations (i.e., rate equations) that approximate the
Stochastic equations (36) via the SLLN are given by

\[
x^N(t) = x(0) - r_1^N(t) - r_2^N(t) - r_3^N(t) + (1 - m)(r_1^N(t) + r_2^N(t) + r_3^N(t)) \\
\approx x(0) - \int_0^t \mu_1 x(s)ds - \int_0^t N\beta x(s)y(s)ds - \int_0^t N\beta(1 - \gamma)x(s)y(s)ds \\
+ \int_0^t (1 - m)(\mu_1 x(s) + \mu_2 y(s) + z(s)))ds \\
y^N(t) = y(0) - r_2^N(t) + r_4^N(t) + r_5^N(t) - m(r_1^N(t) + r_2^N(t) + r_3^N(t)) \\
\approx y(0) - \int_0^t \mu_2 y(s)ds + \int_0^t N\beta x(s)y(s)ds + \int_0^t N\beta(1 - \gamma)x(s)z(s)ds \\
- \int_0^t \alpha y(s)ds + \int_0^t m(\mu_1 x(s) + \mu_2 y(s) + z(s)))ds \\
z^N(t) = z(0) - r_3^N(t) + r_6^N(t) \\
\approx z(0) - \int_0^t \mu_2 z(s)ds + \int_0^t \alpha y(s)ds. \tag{40}
\]

Upon approximating \(\{x^N(t), y^N(t), z^N(t)\}\) by \(\{x(t), y(t), z(t)\}\) and differentiating the above equations we obtain the ordinary differential equations

\[
\frac{dx(t)}{dt} = -\mu_1 x(t) - \beta N x(t)y(t) - \beta N(1 - \gamma)x(t)z(t) + (1 - m)(\mu_1 x + \mu_2 y + z)) \\
\frac{dy(t)}{dt} = -\mu_2 y(t) + \beta N x(t)y(t) + \beta N(1 - \gamma)x(t)z(t) - \alpha y(t) + m(\mu_1 x + \mu_2 y + z)) \\
\frac{dz(t)}{dt} = -\mu_2 z(t) + \alpha y(t), \tag{41}
\]

with initial conditions \(x(0) = x_0, y(0) = y_0, \) and \(z(0) = z_0.\)

To facilitate comparison with the MC model, we let \(U(t) = N x(t), \) \(C(t) = N y(t),\) and \(J(t) = N z(t).\) Then, the system of ordinary differential equations which provides an approximation to averages over sample paths of \(\{U(t), C(t), J(t)\}\) is described by

\[
\frac{dU(t)}{dt} = (1 - m)[\mu_1 U(t) + \mu_2 (C(t) + J(t))] - \beta U(t)[C(t) + (1 - \gamma)J(t)] - \mu_1 U(t) \\
\frac{dC(t)}{dt} = m[\mu_1 U(t) + \mu_2 (C(t) + J(t))] + \beta U(t)[C(t) + (1 - \gamma)J(t)] - (\alpha + \mu_2) C(t) \\
\frac{dJ(t)}{dt} = \alpha C(t) - \mu_2 J(t), \tag{42}
\]

with initial conditions \(U(0) = U_0, C(0) = C_0, \) and \(J(0) = J_0.\)
3.4.1 Steady states and the basic reproductive number

In the absence of VRE colonization on admission (i.e., $m = 0$) there are two steady states. The one indicating the absence of VRE colonization or the VRE-free equilibrium denoted by $E^0$, and the other indicating the presence of VRE colonization denoted by $E^e$. The former is used to calculate the basic reproductive number $R_0$, known in epidemiological models as the average number of secondary cases caused by an infected individual during his/her infectious period when introduced in a completely disease-free population.

Since we have a constant population, for convenience the Model \((42)\) can be reduced to a two dimension system given by

\[
\begin{align*}
\frac{dC(t)}{dt} &= \beta U(t)[C(t) + (1 - \gamma)J(t)] - (\alpha + \mu_2)C(t) \\
\frac{dJ(t)}{dt} &= \alpha C(t) - \mu_2 J(t) \\
U(t) &= N - C(t) - J(t) \\
C(t_0) &= C_0 \\
J(t_0) &= J_0.
\end{align*}
\]

(43)

The basic reproductive number

The basic reproductive number $R_0$ can be defined as the average number of secondary VRE colonized patients generated by a primary case VRE colonized patient in a VRE-free hospital unit. Hence, this quantity plays an important role in determining the possible limiting behaviors of the model, i.e., to what extent VRE colonizations become or remain endemic in a hospital unit population. Assuming that $C$ and $J$ are infective classes and using methodology in [27], we have

\[
\mathcal{F} = \begin{bmatrix} \beta U[C + (1 - \gamma)J] \\ 0 \end{bmatrix}, \quad \mathcal{V} = \begin{bmatrix} (\alpha + \mu_2)C \\ -\alpha C + \mu_2 \end{bmatrix},
\]

where the vector $\mathcal{F}$ represents all new colonizations and the vector $\mathcal{V}$ represents the transitions out of each compartment. Note that progression from $C$ to $J$ is not considered to be new colonization but rather the progression of a VRE colonized patient to the isolation compartment. Since the VRE-free equilibrium is $E^0 = (C^0, J^0) = (0, 0)^T$ we have

\[
\begin{align*}
\mathcal{F} &= \begin{bmatrix} \beta N \\ \alpha \end{bmatrix}, \quad \mathcal{V} = \begin{bmatrix} \alpha + \mu_2 & 0 \\ -\alpha & \mu_2 \end{bmatrix},
\end{align*}
\]

giving
\[ FV^{-1} = \begin{bmatrix} \frac{\beta N}{\alpha + \mu_2} & \frac{\beta N(1-\gamma)}{(\alpha + \mu_2)\mu_2} & \frac{\beta N(1-\gamma)}{\mu_2} \\ \frac{\alpha}{\alpha + \mu_2} & 0 & \frac{\alpha}{\mu_2} \\ \frac{\alpha}{\alpha + \mu_2} & \frac{\alpha}{\mu_2} & 0 \end{bmatrix} \].

Therefore \( R_0 \) is given by the spectral radius of the matrix \( FV^{-1} \) (i.e., \( R_0 = \rho(FV^{-1}) \)):

\[ R_0 = \frac{\beta N}{\alpha + \mu_2} + \frac{\alpha}{\alpha + \mu_2} \frac{\beta N(1-\gamma)}{\mu_2}. \]  \( (44) \)

Each term in \( R_0 \) has an epidemiological interpretation. We may argue that a VRE colonized patient in a totally susceptible hospital unit population causes \( \beta N \) new colonizations in unit time. The quantity \( \frac{1}{\alpha + \mu_2} \) is the average time that a VRE colonized patient spends in the compartment \( C \), and this multiplied by \( \beta N \) are the average number of individuals recruited in this class. This indeed, roughly speaking, is the basic reproductive number for a VRE compartmental model including only the first infective class \( C \), i.e., \( R_0(U \rightarrow C) \). The quantity \( \frac{\alpha}{\alpha + \mu_2} \) is the fraction of VRE colonized patients that are isolated. While in the isolation compartment \( J \), the number of new VRE colonizations caused in unit time is \( \beta N(1-\gamma) \) and the mean time in isolation compartment is \( \mu_2 \). Therefore, the second term in (44) represents the average number of secondary VRE colonizations, patients recruited to \( C \), by individuals who progressed to the isolation compartment \( J \).

The fact that \( R_0 \) is dependent on \( \gamma \) leads us to conclude that health HCW hand-hygiene encouragement has an effect controlling the epidemic of VRE. Note that if we have the case \( \gamma = 1 \), a 100% HCW hand hygiene compliance, this term cancels out and does not contribute to \( R_0 \). Hence, with a high compliance rate the VRE incidence can be attributed more to the first infective class \( C \) than to the second infective class \( J \).

The VRE-free equilibrium. It is given by

\[ E^0 = (C^0, J^0)^T = (0, 0)^T. \]  \( (45) \)

**Proposition 3.1.** Let \( E^0 = (0, 0) \) be the VRE-free equilibrium of (43), then it is locally asymptotically stable if and only if \( R_0 < 1 \).

**Proof.** The Jacobian given from the linearization at the VRE-free equilibrium point is:

\[ J(0, 0) = \begin{bmatrix} \beta N - (\alpha + \mu_2) & \beta N(1-\gamma) \\ \alpha & -\mu_2 \end{bmatrix}. \]

Then, \( det(J) = -\mu_2(\alpha + \mu_2)(R_0 - 1) > 0 \Leftrightarrow R_0 < 1 \) and this implies that \( tr(J) < 0 \). \( \square \)
Proposition 3.2. Let $E^0 = (0,0)$ be the VRE-free equilibrium of (43), then it is globally stable if $\beta N < \mu_2$.

Proof. Let $V = C + J$ be the Lyapunov function (i.e., a continuously differentiable real valued function) and $x^*$ be the VRE-free equilibrium point. We have

1. $V(x) > 0$ for all $x \neq x^*$, and $V(x^*) = 0$.
2. If $\beta N < \mu_2$, then $\frac{dV}{dt} < 0$ for all $x \neq x^*$,

$$
\frac{dV}{dt} = \frac{dC}{dt} + \frac{dJ}{dt} = \beta U[C + (1 - \gamma)J] - \mu_2(C + J) \leq \beta U(C + J) - \mu_2(C + J) \leq (\beta N - \mu_2)(C + J), \text{ since } U \leq N.
$$

Thus, $x^*$ is globally stable for all initial conditions, and $x(t) \to x^*$ as $t \to \infty$. \qed

The endemic equilibrium. It is given by $E^e = (C^e, J^e)$, where

$$
C^e = \frac{\mu_2^2}{\beta[\alpha(1 - \gamma) + \mu_2][R_0 - 1]}, \tag{46}
$$

$$
J^e = \frac{\alpha \mu_2}{\beta[\alpha(1 - \gamma) + \mu_2][R_0 - 1]}. \tag{47}
$$

The existence and local stability of the endemic equilibrium is conditioned on $R_0 > 1$.

Proposition 3.3. Let $E^e = (C^e, J^e)$ be the VRE-endemic equilibrium of (43), then it is locally asymptotically stable if and only if $R_0 > 1$.

Proof. The Jacobian given from the linearization at the VRE-endemic equilibrium point is:

$$
J(C^e, J^e) = \begin{bmatrix}
-\frac{(2\mu_2(R_0 - 1) - \beta N + \alpha + \mu_2)}{\alpha} & -\frac{(2\mu_2(R_0 - 1) - \beta N(1 - \gamma))}{-\mu_2} \\
\end{bmatrix}.
$$

Then $\det(J) > 0$ is equivalent to $R_0 > 1$ and this implies that $\text{tr}(J) < 0$. \qed
3.5 Parameters estimated directly from the VRE surveillance data

In this section we estimate the parameters from the VRE model discussed previously using the VRE surveillance data corresponding to the oncology unit and the general unit. The isolation procedure in these units corresponds to the one assumed in the general VRE epidemic model.

In an attempt to estimate the fraction \((m)\) of patients that are colonized on admission, in both units we found inconsistencies in the reported prevalence of VRE on admission (the summaries of admitted patients did not match the actual data). Also, in an attempt to estimate the initial conditions \((S_0, C_0, J_0)\) from the data reported on the first day of data collection (January 3, 2005 for the oncology unit and October 1, 2004 for the general unit), we found that only the number of VRE colonized patients in isolation was reported. Hence, the initial conditions for \(S_0\) and \(C_0\) cannot be easily estimated. Another parameter that is of interest and can not be estimated directly from the data is the VRE transmission rate \((\beta)\). As a result, the fraction of patients that are colonized on admission, the initial conditions, and the transmission rate are estimated using inverse problem methodology discussed in the following section. In Table 3 we present the assumed initial values of these parameters needed for inverse problem purposes.

Colonized isolation rate \((\alpha)\): VRE colonized patients were put into isolation as soon as the admission swab result were known to be positive. It was told that the test will be back in 2 days but it is possible that this test will be back after 5 days. Therefore, we set \(1/\alpha = (2 + 5)/2 = 3.5\), giving the isolation rate \(\alpha = 0.29\).

Health care worker hand-hygiene compliance \((\gamma)\): Infection control measures were implemented in the form of health care worker hand-hygiene before and after patients contact by the use of gloves and gowns, and washing the hands. In Figure 2 we present the level of compliance at three months interval. For this study we are going to consider the health care worker before patient contact compliance as a better estimator for the parameter \(\gamma\). The mean compliance was 50.63\% for the general unit, and 57.56\% for the oncology unit.

Discharge rate: We do not consider same day discharges. In the oncology unit VRE colonized patients had a mean length of stay of 13.15 days (std=18.28) compared with 6.27 (std=6.80) for the uncolonized patients. In the general ward unit, VRE-colonized patients had a mean stay of 9 (std=13.05) compared with 5 (std=6.89) for the uncolonized patients. In both units, the means between VRE colonized and uncolonized patients were statistically significant different suggesting to us the consideration of different discharge rates. For the oncology unit, we take \(1/\mu_1 = 6.27\) and \(1/\mu_2 = 13.15\) giving \(\mu_1 = 0.16\) and \(\mu_2 = 0.08\). On the other hand, for the general unit we take \(1/\mu_1 = 5\) and \(1/\mu_2 = 9\) giving \(\mu_1 = 0.20\) and \(\mu_2 = 0.11\).

Figure 3 depicts the distribution of length of stay for the oncology unit. It shows a highly skewed distribution where the data initially peak after the first week of stay in the hospital indicating that the majority of patients leave the hospital within this time period. There is a very long gradual tail to the right of the distribution where there is a steady decrease in the number of patients who leave the hospital after longer stays. This gradual tail in the distribution is a contribution by a very small number of patients.
Figure 2: Hand-hygiene compliance comparison for before and after patient contact in the oncology and general unit. HHB = Hand-hygiene before patient contact, HHA = Hand-Hygiene after patient contact.
staying in the hospital for a considerable amount of time, some of whom are present for at least 7 weeks (50 days). This is expected, as the needs of patients suffering from cancer are quite complex in nature, possibly requiring additional rehabilitation and care. The length of hospital stay distribution fits an exponential distribution (Anderson-Darling p-value = 0.2043). Similar results are found for the general unit (Anderson-Darling p-value = 0.2627).

3.6 Inverse problem

As a result of the previous section, it is of interest to estimate the initial conditions, the fraction of patients that are colonized on admission, and the VRE transmission rate, i.e., $\theta = (J_0, C_0, m, \beta)$. The ODE model (42) along with the surveillance data collected for the oncology and general units are used to estimate the parameters via the nonlinear least squares optimization methods described in Section 2.

The VRE model (51) can be rewritten in the general vector form (1) with $x(t) = (U(t), C(t), J(t))^T$ and observational process as

$$y(t_j) = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} U(t_j) \\ C(t_j) \\ J(t_j) \end{bmatrix} = J(t_j) = f(t_j, \theta) \text{ for } j = 1, \ldots, n. \quad (48)$$
Table 3: Parameters values (some values are assumed for optimization purposes)

<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Oncology Unit (N=37)</th>
<th>General Unit (N=29)</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>(U(t_0))</td>
<td>29</td>
<td>23</td>
<td>Assumed</td>
</tr>
<tr>
<td>(C(t_0))</td>
<td>4</td>
<td>3</td>
<td>Assumed</td>
</tr>
<tr>
<td>(J(t_0))</td>
<td>4</td>
<td>3</td>
<td>data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>(\mu_1U(t) + \mu_2(C(t) + J(t)))</th>
<th>(\mu_1U(t) + \mu_2(C(t) + J(t)))</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Lambda)</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(m)</td>
<td>0.04</td>
<td>0.04</td>
<td>Assumed</td>
</tr>
<tr>
<td>(\beta)</td>
<td>0.001</td>
<td>0.001</td>
<td>Assumed</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>0.58</td>
<td>0.51</td>
<td>data</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>0.29</td>
<td>0.29</td>
<td>data</td>
</tr>
<tr>
<td>(\mu_1)</td>
<td>0.16</td>
<td>0.20</td>
<td>data</td>
</tr>
<tr>
<td>(\mu_2)</td>
<td>0.08</td>
<td>0.11</td>
<td>data</td>
</tr>
</tbody>
</table>

Note that the relationship between number of patients in the isolation compartment \((J)\) and time is described by a nonlinear function in its parameters. The algorithm that searches all possible parameter combinations in our problem is:

1. **Combinatorial search.** For a fixed \(j = 1, ..., 8\), and hence fixed \(p\), calculate the set
   \[S_p = \{\theta = (q_1, ..., q_j) \in \mathbb{R}^p | q_k \in \mathbb{R}, q_k \neq q_l \forall k, l = 1, ..., j\}\]
   where \(q_k = \{\alpha, \gamma, \mu_1, \mu_2, J_0, C_0, m, \beta\}\) and the set \(S_p\) collects all the parameter vectors obtained from the combinatorial search;

2. **Full rank test.** Calculate the set of feasible parameters \(\Theta_p\) as
   \[\Theta_p = \{\theta | \theta \in S_p \subset \mathbb{R}^p, \text{rank}(\chi(\theta) = p)\}\]. Calculate the condition number defined by
   \[k(\chi(\theta)) = \frac{s_1}{s_p}\]

3. **Standard error test.** For every \(\theta \in \Theta_p\), calculate a vector of coefficients of variation \(CV(\theta) \in \mathbb{R}^p\) by
   \[CV_i = \frac{\sqrt{\sum(\theta)^2}}{\theta_i},\]
   for \(i = 1, ..., p\) and \(\sum(\theta) = \sigma_0^2 [\chi(\theta)^T \chi(\theta)]^{-1} \in \mathbb{R}^{p \times p}\). Calculate the parameter selection score as \(v(\theta) = |CV(\theta)|\).

### 3.6.1 Optimization algorithm testing with synthetic data

Before using the VRE surveillance data we test the optimization algorithm to investigate the convergence of the parameters estimates \(\hat{\theta}\) to the known values \(\theta_0\). In order to do this, we construct a synthetic data set \(\{y_j\}\) for \(j = 1, ..., n\), by adding variability to the
corresponding model solution, \( f(t_j, \theta_0) = J(t_j, \theta_0) \). The statistical model that captures the variability is taken as

\[
y_j = f(t_j, \theta_0) + \sigma z_j
\]

(49)

where \( z_j \) is a standard normal variable (i.e., \( z_j \sim N(0,1) \)) and \( \sigma \) is the constant variability. The magnitude of \( \sigma \) determines how much noise is added. A low value indicates that the data points tend to be very close to the same value (the mean), while high values indicates that the data are "spread out" over a large range of values. Therefore, we can expect that 95% of the time, numbers generated from this distribution will fall in the interval \([-1.96\sigma, 1.96\sigma]\). To this end, we look at the standard error as one indication of the ability of the algorithm to estimate the parameters using the synthetic data set.

The OLS and GLS optimization were solved with MATLAB routine \textit{lsqnonlin} for \( n=500 \). Parameter upper bounds are taken as \((0.5, 1, 1, 1, 1, 1)\) and lower bounds are set to zero. Note that the upper bound for \( \alpha \) is 0.5 because the method for VRE detection takes at least 2 days. The model solutions \( f(t_j; \theta_0) = J(t_j; \theta_0) \) are generated with initial conditions and parameter values for the oncology unit as described in Table 3 (which are assumed to be the true values). By introducing variability levels such as \( \sigma = 0, \sigma = 0.01, \sigma = 0.05, \) and \( \sigma = 0.1 \) in the model solutions the reliability of the algorithm and hence that of estimates are explored. Note that for this purpose, there is no need to test the algorithm using the general unit values. Also, even though we are adding constant variability to the synthetic data, the GLS optimization algorithm is tested with this data. This is because we are investigating how the noise affects the standard deviation and not how meaningful they are.

In Tables 4, 5, and 6 we summarize the results for the inverse problems for \( \theta = (J_0, C_0, \beta) \), \( \theta = (J_0, C_0, m, \beta) \), and \( \theta = (\alpha, J_0, C_0, m, \beta) \) using an OLS and a GLS optimization formulation. Results indicates that both algorithms appear to be reliable for the estimation of the parameters since the estimated values are close to their true values. Note that as \( \sigma \) increases the corresponding standard errors increases. This indicates that the reliability of both algorithms in estimating the parameters may depend on the observational error in the data. Similar results were obtained for the other types of inverse problems formulations.

### 3.6.2 Subset selection results using the oncology unit surveillance data

To carry out the subset selection algorithm with the oncology unit surveillance data we assumed parameter values described in Table 3. Since we are interested in estimating the initial conditions, transmission rate, and the fraction of patients that are already colonized on admission, when \( p = 4 \) the only parameter combination considered is that of \( \theta = (J_0, C_0, m, \beta) \). When \( p = 1, 2, 3 \) the only parameters considered are \( \theta = (\beta) \), \( \theta = (m, \beta) \), and \( \theta = (J_0, C_0, \beta) \). In Table 7 we present the resulting selection score \( v(\theta) \)
Table 4: OLS and GLS optimization algorithm testing for $\theta = (J_0, C_0, \beta)$ using synthetic data. The model was fit to the synthetic data with levels of noise: $\sigma = 0, 0.01, 0.05,$ and $0.1$. Subscripts in $\theta_\sigma$ denote the level of noise in the synthetic data.

<table>
<thead>
<tr>
<th></th>
<th>$J_0$</th>
<th>$C_0$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_0^{OLS}$</td>
<td>4.000e+00</td>
<td>4.000e+00</td>
<td>1.000e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_0^{OLS})$</td>
<td>2.301e-13</td>
<td>3.097e-13</td>
<td>1.691e-17</td>
</tr>
<tr>
<td>$\hat{\theta}_{0.01}^{OLS}$</td>
<td>4.007e+00</td>
<td>3.998e+00</td>
<td>1.006e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{0.01}^{OLS})$</td>
<td>2.162e-03</td>
<td>2.907e-03</td>
<td>1.584e-07</td>
</tr>
<tr>
<td>$\hat{\theta}_{0.05}^{OLS}$</td>
<td>4.022e+00</td>
<td>3.995e+00</td>
<td>1.032e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{0.05}^{OLS})$</td>
<td>1.077e-02</td>
<td>1.444e-02</td>
<td>7.793e-07</td>
</tr>
<tr>
<td>$\hat{\theta}_{0.1}^{OLS}$</td>
<td>4.074e+00</td>
<td>3.954e+00</td>
<td>1.063e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{0.1}^{OLS})$</td>
<td>2.222e-02</td>
<td>2.971e-02</td>
<td>1.585e-06</td>
</tr>
<tr>
<td>$\hat{\theta}_0^{GLS}$</td>
<td>4.000e+00</td>
<td>4.000e+00</td>
<td>1.000e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_0^{GLS})$</td>
<td>3.956e-15</td>
<td>5.346e-15</td>
<td>4.358e-19</td>
</tr>
<tr>
<td>$\hat{\theta}_{0.01}^{GLS}$</td>
<td>4.002e+00</td>
<td>4.000e+00</td>
<td>1.006e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{0.01}^{GLS})$</td>
<td>4.150e-05</td>
<td>5.604e-05</td>
<td>4.553e-09</td>
</tr>
<tr>
<td>$\hat{\theta}_{0.05}^{GLS}$</td>
<td>4.040e+00</td>
<td>3.973e+00</td>
<td>1.032e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{0.05}^{GLS})$</td>
<td>4.119e-04</td>
<td>2.847e-04</td>
<td>2.290e-08</td>
</tr>
<tr>
<td>$\hat{\theta}_{0.1}^{GLS}$</td>
<td>4.016e+00</td>
<td>4.015e+00</td>
<td>1.067e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{0.1}^{GLS})$</td>
<td>4.119e-04</td>
<td>5.523e-04</td>
<td>4.343e-08</td>
</tr>
</tbody>
</table>
Table 5: OLS and GLS optimization algorithm testing for $\theta = (J_0, C_0, m, \beta)$ using synthetic data. The model was fit to the synthetic data with levels of noise: $\sigma = 0, 0.01, 0.05, \text{and } 0.1$. Subscripts in $\theta_\sigma$ denote the level of noise in the synthetic data.

<table>
<thead>
<tr>
<th></th>
<th>$J_0$</th>
<th>$C_0$</th>
<th>$m$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $\theta$</td>
<td>4</td>
<td>4</td>
<td>0.04</td>
<td>0.001</td>
</tr>
<tr>
<td>Initial $\theta$</td>
<td>3</td>
<td>5</td>
<td>0.05</td>
<td>0.002</td>
</tr>
<tr>
<td>$\hat{\theta}^{OLS}_0$</td>
<td>4.000e+00</td>
<td>4.000e+00</td>
<td>4.000e-02</td>
<td>1.000e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{OLS}_0)$</td>
<td>8.620e-12</td>
<td>1.557e-11</td>
<td>1.611e-13</td>
<td>1.352e-14</td>
</tr>
<tr>
<td>$\hat{\theta}^{OLS}_{0.01}$</td>
<td>4.004e+00</td>
<td>4.008e+00</td>
<td>4.013e-02</td>
<td>9.955e-04</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{OLS}_{0.01})$</td>
<td>2.372e-03</td>
<td>4.287e-03</td>
<td>4.457e-05</td>
<td>3.735e-06</td>
</tr>
<tr>
<td>$\hat{\theta}^{OLS}_{0.05}$</td>
<td>4.029e+00</td>
<td>3.992e+00</td>
<td>4.032e-02</td>
<td>1.004e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{OLS}_{0.05})$</td>
<td>1.102e-02</td>
<td>1.990e-02</td>
<td>2.091e-04</td>
<td>1.741e-05</td>
</tr>
<tr>
<td>$\hat{\theta}^{OLS}_{0.1}$</td>
<td>4.074e+00</td>
<td>3.945e+00</td>
<td>3.987e-02</td>
<td>1.074e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{OLS}_{0.1})$</td>
<td>2.271e-02</td>
<td>4.067e-02</td>
<td>4.273e-04</td>
<td>3.529e-05</td>
</tr>
<tr>
<td>$\hat{\theta}^{GLS}_0$</td>
<td>4.000e+00</td>
<td>4.000e+00</td>
<td>4.000e-02</td>
<td>1.000e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{GLS}_0)$</td>
<td>1.496e-13</td>
<td>2.696e-13</td>
<td>3.145e-15</td>
<td>2.626e-16</td>
</tr>
<tr>
<td>$\hat{\theta}^{GLS}_{0.01}$</td>
<td>4.003e+00</td>
<td>4.001e+00</td>
<td>4.003e-02</td>
<td>1.004e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{GLS}_{0.01})$</td>
<td>4.636e-05</td>
<td>8.350e-05</td>
<td>9.762e-07</td>
<td>8.137e-08</td>
</tr>
<tr>
<td>$\hat{\theta}^{GLS}_{0.05}$</td>
<td>4.009e+00</td>
<td>4.013e+00</td>
<td>4.022e-02</td>
<td>1.014e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{GLS}_{0.05})$</td>
<td>2.343e-04</td>
<td>4.214e-04</td>
<td>4.971e-06</td>
<td>4.118e-07</td>
</tr>
<tr>
<td>$\hat{\theta}^{GLS}_{0.1}$</td>
<td>4.050e+00</td>
<td>4.011e+00</td>
<td>4.046e-02</td>
<td>1.025e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}^{GLS}_{0.1})$</td>
<td>4.434e-04</td>
<td>7.976e-04</td>
<td>9.545e-06</td>
<td>7.845e-07</td>
</tr>
</tbody>
</table>
Table 6: OLS and GLS optimization algorithm testing for $\theta = (\alpha, J_0, C_0, m, \beta)$ using synthetic data. The model was fit to the synthetic data with levels of noise: $\sigma = 0, 0.01, 0.05,$ and 0.1. Subscripts in $\theta_\sigma$ denote the level of noise in the synthetic data.

<table>
<thead>
<tr>
<th>$\hat{\theta}_{OLS}$</th>
<th>$\alpha$</th>
<th>$J_0$</th>
<th>$C_0$</th>
<th>$m$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_{OLS}$</td>
<td>2.890e-01</td>
<td>4.003e+00</td>
<td>4.136e+00</td>
<td>4.451e-02</td>
<td>1.856e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{OLS})$</td>
<td>2.145e-04</td>
<td>5.126e-04</td>
<td>1.674e-03</td>
<td>2.009e-05</td>
<td>1.220e-06</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS,0.01}$</td>
<td>2.895e-01</td>
<td>4.010e+00</td>
<td>4.120e+00</td>
<td>4.454e-02</td>
<td>1.872e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{OLS,0.01})$</td>
<td>1.094e-03</td>
<td>2.620e-03</td>
<td>8.517e-03</td>
<td>1.025e-04</td>
<td>6.264e-06</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS,0.05}$</td>
<td>2.811e-01</td>
<td>4.023e+00</td>
<td>4.269e+00</td>
<td>5.080e-02</td>
<td>1.670e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{OLS,0.05})$</td>
<td>5.755e-03</td>
<td>1.337e-02</td>
<td>4.696e-02</td>
<td>6.381e-04</td>
<td>3.168e-05</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS,0.1}$</td>
<td>2.899e-01</td>
<td>4.051e+00</td>
<td>4.109e+00</td>
<td>4.541e-02</td>
<td>1.880e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{OLS,0.1})$</td>
<td>1.071e-02</td>
<td>2.572e-02</td>
<td>8.329e-02</td>
<td>1.033e-03</td>
<td>6.263e-05</td>
</tr>
<tr>
<td>$\hat{\theta}_{GLS}$</td>
<td>2.901e-01</td>
<td>4.000e+00</td>
<td>4.095e+00</td>
<td>4.325e-02</td>
<td>1.538e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{GLS})$</td>
<td>3.606e-06</td>
<td>8.920e-06</td>
<td>2.854e-05</td>
<td>3.686e-07</td>
<td>2.277e-08</td>
</tr>
<tr>
<td>$\hat{\theta}_{GLS,0.01}$</td>
<td>2.894e-01</td>
<td>4.009e+00</td>
<td>4.130e+00</td>
<td>4.300e-02</td>
<td>1.869e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{GLS,0.01})$</td>
<td>2.282e-05</td>
<td>5.581e-05</td>
<td>1.836e-04</td>
<td>2.373e-06</td>
<td>1.525e-07</td>
</tr>
<tr>
<td>$\hat{\theta}_{GLS,0.05}$</td>
<td>2.787e-01</td>
<td>4.030e+00</td>
<td>4.348e+00</td>
<td>2.360e-02</td>
<td>1.860e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{GLS,0.05})$</td>
<td>1.068e-04</td>
<td>2.581e-04</td>
<td>8.901e-04</td>
<td>5.955e-06</td>
<td>5.713e-07</td>
</tr>
<tr>
<td>$\hat{\theta}_{GLS,0.1}$</td>
<td>2.900e-01</td>
<td>4.035e+00</td>
<td>4.189e+00</td>
<td>4.739e-02</td>
<td>1.882e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta}_{GLS,0.1})$</td>
<td>2.147e-04</td>
<td>5.191e-04</td>
<td>1.740e-03</td>
<td>2.506e-05</td>
<td>1.502e-06</td>
</tr>
</tbody>
</table>
and condition number \( k(\chi(\theta)) \) for each subset of parameters when \( p = 1, \ldots, 8 \). Values that fall in the smallest selection score with the relative small condition number are considered the most feasible subset of parameters. Results indicate that the subsets of parameters \( \theta = (J_0, C_0, m, \beta) \) have small condition numbers and relatively small selection scores indicating that these subsets might provide low uncertainty in the parameter estimates. In Table 8 we summarize the results of 4 inverse problems corresponding to the subsets with the lowest selection scores and small condition numbers. These subsets of parameters are:

\[
\theta = (\gamma, J_0, C_0, m, \beta) \\
\theta = (J_0, C_0, m, \beta) \\
\theta = (J_0, C_0, \beta) \\
\theta = (m, \beta)
\]

We analyze the results using the coefficient of variation (CV) by looking at the effect that the inclusion or exclusion of parameters has on the vector \( \theta = (J_0, C_0, m, \beta) \). In this subset, the standard errors for \( J_0 \) is about 0.4% of the estimate, for \( C_0 \) it is about 0.8% of the estimate, for \( m \) it is about 1.6% of the estimate, and for \( \beta \) it is 0.3% of the estimate. When including \( \gamma \) (i.e., \( \theta = (\gamma, J_0, C_0, m, \beta) \)), the CV increases for almost all parameters. On the other hand, when \( m \) is dropped or when the initial conditions are dropped, there is a reduction in the CV. Since this reduction is not significant, we can conclude that the subset \( \theta = (J_0, C_0, m, \beta) \) is a good choice to be estimated from the oncology surveillance data since it provides estimates with low uncertainty.

Residual plots for all subsets of parameters combinations suggested that the assumptions of the Statistical Model (14) corresponding to the GLS procedure are satisfied. In particular, the residual analysis for \( \theta = (J_0, C_0, m, \beta) \) is presented in Figure 4. The OLS residual plots (a) and (b) in Figure 4 reveal a fan shaped error structure which indicates the nonconstant variance assumption is suspect. When GLS optimization is used instead, the residual plots (c) and (d) in Figure 4 reveals a more random error structure, suggesting that the GLS procedure was correctly used. Finally, a best fit of the model solution to the oncology surveillance data is shown in Figure 5.

### 3.6.3 Subset selection results using the general unit surveillance data

To carry out the subset selection algorithm with the general unit surveillance data, it is assumed nominal parameter values as described in Table 3. We also estimate the initial conditions, transmission rate, and the fraction of patients that are colonized on admission, i.e., \( \theta = (J_0, C_0, m, \beta) \). In Table 9 we present the results of the subset of parameters combinations with the corresponding selection score \( v(\theta) \) and condition number \( k(\chi(\theta)) \). From this table we can conclude that \( \theta = (J_0, C_0, m, \beta) \) is not a good option to be estimated. Its selection score is high suggesting high uncertainty for at least one of the parameters. Since the parameter combinations for \( p = 5, 6, 7, 8 \) also suggest additional possible sources of uncertainty, the exclusion of either the initial conditions or the fraction of patients that are colonized on admission is considered. In Table 10...
Table 7: Subset parameter selected as a result of the selection algorithm for $p = 4, \ldots, 8$ using the oncology unit surveillance data with nominal parameter values described in Table 3 using the GLS optimization.

<table>
<thead>
<tr>
<th>Parameter vector, $q$</th>
<th>Selection score, $v(q)$</th>
<th>Condition number, $\kappa(\chi(q))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(\beta)$</td>
<td>1.975e-05</td>
<td>1.000e+00</td>
</tr>
<tr>
<td>$(m, \beta)$</td>
<td>2.358e-03</td>
<td>8.070e+02</td>
</tr>
<tr>
<td>$(Jo, Co, \beta)$</td>
<td>7.134e-03</td>
<td>8.236e+04</td>
</tr>
<tr>
<td>$(Jo, Co, m, \beta)$</td>
<td>1.815e-02</td>
<td>9.946e+04</td>
</tr>
<tr>
<td>$(\gamma, Jo, Co, m, \beta)$</td>
<td>1.539e-01</td>
<td>2.253e+05</td>
</tr>
<tr>
<td>$(\alpha, Jo, Co, m, \beta)$</td>
<td>1.597e-01</td>
<td>1.063e+06</td>
</tr>
<tr>
<td>$(\mu_1, Jo, Co, m, \beta)$</td>
<td>1.715e+01</td>
<td>1.308e+08</td>
</tr>
<tr>
<td>$(\mu_2, Jo, Co, m, \beta)$</td>
<td>6.123e+03</td>
<td>3.695e+05</td>
</tr>
<tr>
<td>$(\alpha, \mu_1, Jo, Co, m, \beta)$</td>
<td>6.225e+00</td>
<td>5.522e+06</td>
</tr>
<tr>
<td>$(\gamma, \mu_1, Jo, Co, m, \beta)$</td>
<td>1.741e+01</td>
<td>1.127e+08</td>
</tr>
<tr>
<td>$(\alpha, \mu_2, Jo, Co, m, \beta)$</td>
<td>6.315e+01</td>
<td>2.453e+05</td>
</tr>
<tr>
<td>$(\gamma, \mu_2, Jo, Co, m, \beta)$</td>
<td>8.472e+02</td>
<td>7.122e+05</td>
</tr>
<tr>
<td>$(\alpha, \gamma, Jo, Co, m, \beta)$</td>
<td>2.297e+03</td>
<td>2.852e+06</td>
</tr>
<tr>
<td>$(\mu_1, \mu_2, Jo, Co, m, \beta)$</td>
<td>7.475e+04</td>
<td>2.091e+05</td>
</tr>
<tr>
<td>$(\alpha, \gamma, \mu_1, Jo, Co, m, \beta)$</td>
<td>8.413e+02</td>
<td>2.074e+09</td>
</tr>
<tr>
<td>$(\alpha, \mu_1, \mu_2, Jo, Co, m, \beta)$</td>
<td>1.929e+03</td>
<td>3.760e+05</td>
</tr>
<tr>
<td>$(\alpha, \gamma, \mu_2, Jo, Co, m, \beta)$</td>
<td>3.447e+04</td>
<td>4.305e+06</td>
</tr>
<tr>
<td>$(\gamma, \mu_1, \mu_2, Jo, Co, m, \beta)$</td>
<td>4.589e+04</td>
<td>4.311e+07</td>
</tr>
<tr>
<td>$(\alpha, \gamma, \mu_1, \mu_2, Jo, Co, m, \beta)$</td>
<td>1.469e+04</td>
<td>1.967e+09</td>
</tr>
</tbody>
</table>
Table 8: Results of 4 inverse formulations solved with nominal values in table 3 via GLS optimization for the oncology unit surveillance data.

<table>
<thead>
<tr>
<th></th>
<th>$\gamma$</th>
<th>$J_0$</th>
<th>$C_0$</th>
<th>$m$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}$</td>
<td>6.392e-01</td>
<td>4.004e+00</td>
<td>1.092e+00</td>
<td>5.277e-02</td>
<td>4.770e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta})$</td>
<td>2.680e-02</td>
<td>1.811e-02</td>
<td>4.985e-02</td>
<td>6.007e-03</td>
<td>3.955e-04</td>
</tr>
<tr>
<td>$CV(\hat{\theta})$</td>
<td>4.192e-02</td>
<td>4.524e-03</td>
<td>4.567e-02</td>
<td>1.139e-01</td>
<td>8.291e-02</td>
</tr>
<tr>
<td>$\hat{\theta}$</td>
<td>-</td>
<td>3.706e+00</td>
<td>1.966e+00</td>
<td>3.608e-02</td>
<td>4.865e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta})$</td>
<td>-</td>
<td>1.499e-02</td>
<td>1.560e-02</td>
<td>5.616e-04</td>
<td>1.675e-05</td>
</tr>
<tr>
<td>$CV(\hat{\theta})$</td>
<td>-</td>
<td>4.044e-03</td>
<td>7.934e-03</td>
<td>1.556e-02</td>
<td>3.443e-03</td>
</tr>
<tr>
<td>$\hat{\theta}$</td>
<td>-</td>
<td>3.706e+00</td>
<td>1.966e+00</td>
<td>-</td>
<td>4.865e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta})$</td>
<td>-</td>
<td>1.419e-02</td>
<td>1.184e-02</td>
<td>-</td>
<td>9.945e-08</td>
</tr>
<tr>
<td>$CV(\hat{\theta})$</td>
<td>-</td>
<td>3.829e-03</td>
<td>6.020e-03</td>
<td>-</td>
<td>2.044e-05</td>
</tr>
<tr>
<td>$\hat{\theta}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.070e-02</td>
<td>4.725e-03</td>
</tr>
<tr>
<td>$SE(\hat{\theta})$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9.290e-05</td>
<td>2.802e-06</td>
</tr>
<tr>
<td>$CV(\hat{\theta})$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.282e-03</td>
<td>5.931e-04</td>
</tr>
</tbody>
</table>
Figure 4: Residual analysis for the OLS and GLS optimization of $\theta = (J_0, C_0, m, \beta)$ using the oncology unit surveillance data.
we summarize the results of 5 inverse problems that correspond to the subsets with the lowest selection score and relative condition number:

\[\theta = (J_0, C_0, m, \beta)\]
\[\theta = (\alpha, J_0, C_0, \beta)\]
\[\theta = (J_0, C_0, \beta)\]
\[\theta = (m, \beta)\]
\[\theta = (\beta)\]

As expected for \(\theta = (J_0, C_0, m, \beta)\), the coefficient of variation for the parameter \(m\) and \(C_0\) indicates an extremely high uncertainty in the estimation. This subset is not good for the inverse problem formulation. When the parameter \(m\) is dropped (i.e., \(\theta = (J_0, C_0, \beta)\)), the standard error for \(C_0\) reduces to about 15 times the estimate. When the initial conditions are dropped (i.e., \(\theta = (m, \beta)\)), the standard error for the parameter \(m\) reduces but is still considered extremely high. It is concluded that the subset considering only \(\beta\) is the subset for which the mathematical model responds best. The standard error for \(\beta\) is 1.2% the estimate.

The residual analysis for \(\theta = (\beta)\) is presented in Figure 6. This analysis indicates that the OLS residual plots reveal no noticeable difference when compared with the GLS residual plots. On the other hand, for other inverse problem formulations, the GLE residuals versus time seems to reveal an inverted fan shape. Therefore, it was assumed that OLS procedure was a better choice for this data. A best fit of the model solution to the general unit surveillance data is shown in Figure 7.
Table 9: Subset parameter selected as a result of the selection algorithm for \( p = 1, \ldots, 5 \) using the general unit surveillance data with nominal parameter values described in Table 3 using OLS optimization.

<table>
<thead>
<tr>
<th>Parameter vector, ( q )</th>
<th>Selection score, ( v(q) )</th>
<th>Condition number, ( \kappa(\chi(q)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\beta))</td>
<td>3.259e-03</td>
<td>1.000e+00</td>
</tr>
<tr>
<td>((m, \beta))</td>
<td>1.232e+06</td>
<td>8.189e+02</td>
</tr>
<tr>
<td>((Jo, Co, \beta))</td>
<td>7.258e+01</td>
<td>2.858e+04</td>
</tr>
<tr>
<td>((\alpha, Jo, Co, \beta))</td>
<td>4.572e+01</td>
<td>3.680e+04</td>
</tr>
<tr>
<td>((m, Jo, Co, \beta))</td>
<td>1.475e+04</td>
<td>4.640e+04</td>
</tr>
<tr>
<td>((\mu_2, Jo, Co, \beta))</td>
<td>5.253e+06</td>
<td>1.157e+05</td>
</tr>
<tr>
<td>((\gamma, Jo, Co, \beta))</td>
<td>3.459e+07</td>
<td>7.108e+04</td>
</tr>
<tr>
<td>((\mu_1, Jo, Co, \beta))</td>
<td>8.499e+07</td>
<td>4.225e+03</td>
</tr>
<tr>
<td>((\mu_1, Jo, Co, m, \beta))</td>
<td>1.224e+06</td>
<td>5.727e+03</td>
</tr>
<tr>
<td>((\mu_2, Jo, Co, m, \beta))</td>
<td>2.785e+07</td>
<td>2.904e+05</td>
</tr>
<tr>
<td>((\gamma, Jo, Co, m, \beta))</td>
<td>4.101e+11</td>
<td>1.990e+05</td>
</tr>
<tr>
<td>((\alpha, Jo, Co, m, \beta))</td>
<td>6.992e+12</td>
<td>1.203e+05</td>
</tr>
</tbody>
</table>
Table 10: Results of 5 inverse formulations solved with nominal values in table 3 via OLS optimization for the general unit surveillance data.

<table>
<thead>
<tr>
<th></th>
<th>( \alpha )</th>
<th>( J_0 )</th>
<th>( C_0 )</th>
<th>( m )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta} )</td>
<td>-</td>
<td>3.180e+00</td>
<td>3.349e-02</td>
<td>5.143e-06</td>
<td>1.074e-02</td>
</tr>
<tr>
<td>( SE(\hat{\theta}) )</td>
<td>-</td>
<td>6.647e-01</td>
<td>1.016e+00</td>
<td>7.608e-02</td>
<td>2.718e-03</td>
</tr>
<tr>
<td>( CV(\hat{\theta}) )</td>
<td>-</td>
<td>2.090e-01</td>
<td>3.034e+01</td>
<td>1.479e+04</td>
<td>2.531e-01</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>5.000e-01</td>
<td>3.011e+00</td>
<td>1.551e-02</td>
<td>-</td>
<td>9.047e-03</td>
</tr>
<tr>
<td>( SE(\hat{\theta}) )</td>
<td>2.412e-01</td>
<td>7.616e-01</td>
<td>7.092e-01</td>
<td>-</td>
<td>1.546e-04</td>
</tr>
<tr>
<td>( CV(\hat{\theta}) )</td>
<td>4.824e-01</td>
<td>2.530e-01</td>
<td>4.571e+01</td>
<td>-</td>
<td>1.709e-02</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>-</td>
<td>4.008e+00</td>
<td>8.314e-03</td>
<td>-</td>
<td>9.283e-03</td>
</tr>
<tr>
<td>( SE(\hat{\theta}) )</td>
<td>-</td>
<td>6.420e-01</td>
<td>6.035e-01</td>
<td>-</td>
<td>3.092e-05</td>
</tr>
<tr>
<td>( CV(\hat{\theta}) )</td>
<td>-</td>
<td>1.570e-01</td>
<td>7.258e+01</td>
<td>-</td>
<td>3.331e-03</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.686e-08</td>
</tr>
<tr>
<td>( SE(\hat{\theta}) )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.077e-02</td>
</tr>
<tr>
<td>( CV(\hat{\theta}) )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.232e+06</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( SE(\hat{\theta}) )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( CV(\hat{\theta}) )</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 6: Residual analysis for OLS and GLS optimization of $\theta = (\beta)$ using the general unit surveillance data.
In this section, we have derived a simple model of the dynamic of VRE colonization of patients in hospitals. Because of the small number of beds in a hospital unit, we first considered a continuous time MC model with discrete state space. We are interested in estimating some of the epidemiological parameters underlying this system but carrying out the parameter estimation using this type of stochastic model turned out to be a difficult task. Therefore, we take an alternative approach which involve the estimation of parameters using the corresponding deterministic approximation (using Kurtz’ theory) to the MC model. The deterministic approximation is built in terms of large sample size average over sample paths.

The deterministic approximation is used to compute $R_0$ and we showed that the VRE-free equilibrium $E^0$ is locally stable when $R_0 < 1$. Whenever $R_0 > 1$ the VRE-free equilibrium becomes unstable and an endemic stable equilibrium is established in the hospital population. In other words, when $R_0 > 1$ the VRE colonization persist in our hospital population. In Proposition (3.4.2) we established the global stability (i.e., all solutions approaching a given point regardless the initial conditions) of $E^0$ conditioned
on $\beta N < \mu_2$. This proposition implies

$$
\mathcal{R}_0 < \frac{\mu_2}{\alpha + \mu_2} + \frac{\alpha (1 - \gamma)}{\alpha + \mu_2} \\
= 1 - \frac{\alpha \gamma}{\alpha + \mu_2} \\
= 1 - \mathcal{R}_c < 1.
$$

(50)

Hence, the global stability of $E^0$ corresponds to the region $[0, \mathcal{R}_c)$. Future work should consider to study the stability of $E^0$ in the region $[\mathcal{R}_c, 1)$.

We calibrated the ODE model to two hospital units: the oncology and general ward, via a nonlinear least square procedure. For the oncology unit, we were able to estimate the parameters of interest which are not typically measurable or reported in the literature. All standard errors for these parameters provide confidence in the values obtained. On the other hand, results obtained when using the general unit surveillance data suggest that this data does not enable us to reliably estimate the initial conditions $(C_0, J_0)$ and the fraction $(m)$ of VRE colonized in admission. We conclude that the solution of the inverse problems involves much more than simple curve fitting. It is essential to know the conditions under which the unknown parameters are identifiable and whether the data is sufficient to show this problem.

4 Simulations

We carry out numerous simulations to compare the results of the stochastic and deterministic models. The deterministic system is numerically solved using 45 solver in Matlab. Both deterministic and stochastic results are generated using the same parameter values and initial conditions. A particular case that we consider in this analysis is having the fraction of already VRE colonized patients on admission equal to zero (i.e., $m = 0$). In this case, the effect of the health care worker hand hygiene compliance rate $\gamma$ has on the basic reproductive number $\mathcal{R}_0$, is studied.

4.1 Stochastic simulation algorithm

For stochastic models, the stochastic simulation algorithm proposed by Gillespie [30] is standard for discrete state continuous time Markov Chain models. The algorithm is the following:

1. **Initialize** the state of the system;

2. For a given state of the system calculate the transition rates, $\lambda_i$, for $i = 1, \ldots, n$, where $n$ is the total types of transitions;

3. Calculate the sum of all transition rates $\lambda = \sum_{i=1}^{n} \lambda_i$;

4. **Monte Carlo Step**: Simulate the time until the next transition, $\tau$, by drawing from an exponential distribution with mean $1/\lambda$;
5. **Monte Carlo Step**: Simulate the transition type by drawing from the discrete distribution with \( P(\text{transition} = i) = \lambda_i / \lambda \). Generate an uniformly distributed random number \( r_2 \). For \( 0 < r_2 < \lambda_1 / \lambda \) transition 1 is chosen, for \( \lambda_1 / \lambda < r_2 < (\lambda_1 + \lambda_2) / \lambda \) transition 2 is chosen, and so on:

6. **Update** the new time \( t = t + \tau \) and the new system state;

7. **Iterate** steps 2-6 until \( t > t_{stop} \).

The basis of this method is the computing two values: time of next transition to take place, and which transition will take place. In any state, the tentative times until entering a particular state are all exponentially distributed with a mean reciprocal to the relevant transition rate. The number of transitions in any time interval has a Poisson distribution with rate \( \lambda \), where it is assumed that at most one event occurs in the time period \( dt \). In our case, this event can be either a colonization, or a discharge, or an isolation, which depends only on the values of the state variables at the current time. Since the hospital remains full for all \( t \), the discharge of a patient in either compartments \( U \), or \( C \), or \( J \), will lead to an admission in either compartments \( U \) or \( C \). Therefore, if an event of discharge occurs in step 5, the corresponding admission is chosen as follows: for \( mx > m \) and \( mx \leq 1 \) admission in \( U \) is chosen; otherwise admission in \( C \) is chosen, where \( mx \) is an uniformly distributed random number.

### 4.2 Simulation results for the oncology unit

Figure 8 depicts a simulation of the stochastic model for a sample of 5 stochastic realizations \((N = 37, t_{stop} = 500)\) plotted in comparison to the deterministic numerical solution for the three compartments. Note that some of the stochastic realizations hit zero but they do not go to extinction. This is due to the fact that we have VRE colonized patients entering to the system \( (m = 0.04) \) for all \( t \). We also note that the stochastic realizations exhibit very large differences. However, when carrying out the simulations for larger values of \( N \), the variation between the stochastic realizations decreases as the value of \( N \) increases and are closer to the numerical solution of the deterministic model, as seen in Figure 9. To quantitatively analyze how the variability of the stochastic realizations decreases as \( N \) increases, we calculated the coefficient of variation (CV) in the range \( t = [300, 400] \) using 100 stochastic realizations and there are also given in the cap from Figure 9.

Taking all the final system states into consideration for a vast number of stochastic realizations, we obtain an estimate of the Probability Density Function (PDF) for the system being in a certain state at time \( t_{stop} \). Figure 10 contains plots of the resulting probability distributions (with a fitted normal curve) for the system at \( t_{stop} = 500, N = 37 \), and 1,000 realizations. The distribution for the number of uncolonized patients has a mean of 25.30 patients (median = 25) and standard deviation of 4.36. The distribution for the number of VRE colonized patients has as mean 2.55 patients (median = 2) and standard deviation of 1.82. Finally, the distribution for the number of VRE isolated patients has a mean of 9.15 patients (median = 9) and standard deviation of 3.65.
Figure 8: Sample of 5 stochastic realizations in comparison to the numerical solution of the deterministic model; $N = 37$ patients, $t_{stop} = 500$.

Figure 11 depicts the effect of the health care worker hand hygiene compliance for different values of $\gamma$. Numerical solutions indicate that hand-hygiene compliance controls the VRE colonization. In the case of the absence of infection control measure ($\gamma = 0$), the number of VRE colonized patients stabilize around five patients. When $\gamma = 0.58$ this number stabilize around nearly three patients. In the extreme case of having 100% compliance, the number of VRE colonization stabilize around one patient.

4.2.1 Case: No VRE colonization on admission

In the case of having no colonization on admission, Figure 12 depicts a sample of 5 stochastic realizations plotted in comparison to the deterministic numerical solution for the three compartment. In this case we have all realizations but 5 going to extinction. After 1,000 stochastic realizations with $N = 37$ and $t = 500$ the probability of extinction is 0.98 for VRE colonized patients with a mean time of extinction of 116.93 days. For VRE isolated patients the mean time of extinction is 123.88 days. Considering the case $\gamma = 0$ (absence of hand-hygiene infection control) the probability of extinction is 0.04 for VRE colonized patients and 0.02 for VRE isolated patients. On the other hand, when $\gamma = 1$ (100% hand-hygiene compliance) the probability of extinction was one for both VRE colonized and VRE isolated patients with a mean time of extinction of 5.17 and 31.67 days. In Figure 14-(a) we depict how $\gamma$ affects the number of cases of VRE colonized and the number of VRE isolated patients (actual endemic state). With $\gamma = 0.72$ these numbers drop to zero.
Figure 9: Sample of 5 stochastic realizations for each compartment in comparison to the numerical solution of the deterministic model for $N = 137, 537, 937, 2037$, $t_{stop} = 500$. The coefficient of variation (CV) for $[U, C, J]$ in the range $t = [300, 400]$ using a sample of 100 stochastic realizations is: (a)$[0.064, 0.060, 0.027]$, (b)$[0.036, 0.027, 0.008]$, (c)$[0.031, 0.021, 0.006]$, (d)$[0.029, 0.014, 0.004]$
Figure 10: Probability Distributions of the system being in a certain state at time $t_{\text{stop}} = 500$ days; $N = 37$ patients; 1000 realizations, for the oncology unit.

Figure 13 contains the resulting corresponding probability distributions for the system at $t_{\text{stop}} = 500$, $N = 37$, and 1000 realizations. The resulting distributions for the number of colonized and colonized isolated patients exhibit a high frequency in zero cases that can be explained by the high probability of extinction. The distribution for the number of uncolonized patients has a mean of 36.80 (median of 37) patients and standard deviation of 1.34. The distribution for the number of VRE colonized patients has a mean of 0.03 (median of 0) patients and standard deviation of 0.25. To conclude, the distribution for the number of VRE colonized patients in isolation has a mean of 0.17 (median of 0) patients and standard deviation of 1.14.

4.2.2 Effect of health care worker hand-hygiene compliance rate on the basic reproductive number

Given the estimated values in Section 3, the estimated basic reproductive number for the oncology units is $R_0 = 1.24$. In Figure 14-(b) the effect of $\gamma$ on $R_0$ is graphed. We see that as $\gamma$ increases $R_0$ decreases. Considering the cases $\gamma = 0$ (absence of hand-hygiene infection control) and $\gamma = 1$ (100% hand-hygiene compliance), we find the estimated basic reproductive numbers are $R_0 = 2.27$ and $R_0 = 0.49$. Note that for $R_0$ to be less than one, a health care worker hand hygiene compliance rate of at least 72% is required.
Figure 11: Numerical solution of the deterministic model for $N = 37$ and $t_{\text{stop}} = 500$. Effect of the HCW hand-hygiene compliance rate ($\gamma$) for the oncology unit.
Figure 12: Sample of 5 stochastic realizations in comparison to the numerical solution of the deterministic model when considering when considering no VRE colonization on admission; \( N = 37 \) patients, \( t = 500 \), for the oncology unit

Figure 13: Probability Distributions of the system being in a certain state at time \( t = 500 \) days; \( N = 37 \) patients; 1000 realizations for the oncology unit when considering no VRE colonization on admission
Figure 14: Oncology unit: (a) Effect of health care worker hand-hygiene compliance rate $\gamma$ on the number of VRE colonized and VRE isolated patients, $N = 37$ patients. (b) Effect of health care worker hand-hygiene compliance rate $\gamma$ on $R_0$, $N = 37$ patients.
4.3 Simulation results for the general unit

Similar results were carried out for the general unit. Numerical solutions also indicate that hand-hygiene compliance controls the VRE colonization. In the case of the absence of infection control measure ($\gamma = 0$), the number of VRE colonized patients stabilizes to around five patients. When $\gamma = 0.58$ this number stabilizes to around three to four patients. In the extreme case of having 100% compliance, the number of VRE colonization stabilizes to around one patient, as seen in Figure 15.

Figure 16 depicts the resulting probability distributions for the system at $t_{\text{stop}} = 320$, $N = 29$, and 1000 realizations. The distribution for the number of uncolonized patients has a mean of 17.40 (median of 17) patients and standard deviation of 4.22. The distribution for the number of VRE colonized patients has a mean of 3.30 (median of 3) patients and standard deviation of 2.00. The distribution for the number of VRE colonized patients in isolation has a mean of 8.30 (median of 8) patients and standard deviation of 3.37.

4.3.1 Case: No VRE colonization on admission

Figure 17 contains a sample of 5 stochastic realizations plotted in comparison to the deterministic numerical solution for the three compartments when considering no VRE colonization on admission. In this case we have all but 3 realizations going to extinction. After 1,000 stochastic realizations with $N = 29$ and $t = 320$ the probability of extinction is 0.83 for VRE colonized patients with a mean time of extinction of 100.50 days. For VRE isolated patients the mean time of extinction is 106.32 days. Considering the case $\gamma = 0$ (absence of hand-hygiene infection control) we find the probability of extinction is 0.04 for VRE colonized patients and 0.03 for VRE isolated patients. On the other hand, when $\gamma = 1$ (100% hand-hygiene compliance) the probability of extinction is one for both VRE colonized and VRE isolated patients with a mean time of extinction of 6.93 and 26.65 days.

Figure 18 depicts the resulting probability distributions for the system at $t_{\text{stop}} = 320$, $N = 29$, and 1000 realizations. The distribution for the number of uncolonized patients has a mean of 27.11 (median of 29) patients and standard deviation of 4.14. The distribution for the number of VRE colonized patients has a mean of 0.50 (median of 0) patients and standard deviation of 1.32. The distribution for the number of VRE colonized patients in isolation has a mean of 1.39 (median of 0) patients and standard deviation of 3.07.

4.3.2 Effect of health care worker hand-hygiene compliance rate on the basic reproductive number

Given the estimated values in Section 3, the estimated basic reproductive number for the general unit is $R_0 = 1.53$. In Figure 19 we present the effect of $\gamma$ on $R_0$. Considering the cases $\gamma = 0$ (absence of hand-hygiene infection control) and $\gamma = 1$ (100% hand-hygiene compliance) we obtain the estimated basic reproductive numbers $R_0 = 2.43$ and
Figure 15: Numerical solution of the deterministic model for $N = 29$ and $t_{\text{stop}} = 320$. Effect of the HCW hand-hygiene compliance rate $\gamma$ in the general unit.
Figure 16: Probability Distributions of the system being in a certain state at time $t=320$ days; $N = 29$ patients; 1000 realizations, for the general unit.

Figure 17: Sample of 5 stochastic realizations in comparison to the numerical solution of the deterministic model when considering no VRE colonization on admission; $N = 29$ patients, $t = 320$ in the general unit.
Uncolonized Patients: mean= 27.112, std= 4.1396, median= 29

VRE Colonized Patients: mean= 0.496, std= 1.3205, median= 0

VRE Isolated Patients: mean= 1.392, std= 3.0714, median= 0

Figure 18: Probability Distributions of the system being in a certain state at time $t=320$ days when considering no VRE colonization on admission; $N = 29$ patients; 1000 realizations, for the general unit when considering no VRE colonization on admission

$R_0 = 0.67$. Note that for $R_0$ to be less that one, a health care worker hand hygiene compliance rate of at least 83% is needed.

4.4 Conclusions

As expected for large $N$, we find that the solutions for the stochastic and deterministic models provide similar results, with the realizations fluctuating around the solution of the deterministic model for the averages as predicted by the approximation theory discussed in Section 3.4. The solutions for the deterministic model approach the steady state rather quickly. All compartments remain stable thereafter suggesting that the steady states are asymptotically stable (at least locally). The only simulations that result in the extinction of the VRE epidemic arose when considering the absence of VRE-colonized patients on admission ($m = 0$). When $m > 0$ the VRE always remains naturally endemic, regardless of the value of $R_0$. We estimated a value of $R_0 > 1$ for both units, indicating an endemic level of VRE present.

5 Discrete VRE Model with Delay

In Section 3, the VRE general epidemic model assumes the same rate of isolation for VRE colonized patients tested on admission and VRE colonized patients tested on weekly
Figure 19: General Unit: (a) Effect of health care worker hand-hygiene compliance rate ($\gamma$) on the number of VRE colonized and VRE isolated patients, $N = 29$ patients. (b) Effect of health care worker hand-hygiene compliance rate ($\gamma$) on $R_0$, $N = 29$ patients.
swab cultures. Also, the rate of change of VRE colonized patients in isolation at time \( t \) depends instantaneously on the number of VRE colonized patients at time \( t \). In reality, there is a difference on how VRE colonized patients tested on admission and VRE colonized patients tested on weekly swab cultures are moved into isolation.

VRE colonized patients are tested on admission and will be moved into isolation after the test results are back, which takes at least 2 days from the admission day. Therefore, the isolation rate of VRE colonized patients tested on admission depends in the past history of these patients, i.e., a delay. On the other hand, new VRE colonizations acquired during the hospital stay identified by the swab cultures administered once a week (i.e., every Tuesday) will be moved into isolation two days later (i.e., at least every Thursday). In this case, there is no delay involved but rather a jump discontinuity every isolation day. Consequently, a reformulation of the ODE model (42) is considered in order to account for these details (see the next section for more details on the model derivation).

Let \( U(t) \) represent the number of uncolonized patients, \( C_1(t) \) the number of VRE colonized patients through admission, \( C_2(t) \) the number of new VRE colonizations during the hospital stay, and \( J(t) \) the number of VRE colonized patients currently in isolation. A diagram of the new compartmental model under consideration is presented in Figure 20 and the corresponding system of equations are

\[
\begin{align*}
\frac{dC_1(t)}{dt} &= m\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t) + C_2(t) + J(t)]\} \\
&\quad - me^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t-2) + C_2(t-2) + J(t-2)]\} \\
&\quad - \mu_2 C_1(t) \\
\frac{dC_2(t)}{dt} &= \beta\{N - [C_1(t) + C_2(t) + J(t)]\}[C_1(t) + C_2(t) + (1-\gamma)J(t)] \\
&\quad - \mu_2 C_2(t), \quad \text{for } t_i < t \leq t_{i+1} \\
C_2(t_{i+1}^+) &= C_2(t_{i+1}^-) - C_2(t_{i+1}^- - 2)e^{-2\mu_2} \\
\frac{dJ(t)}{dt} &= me^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t-2) + C_2(t-2) + J(t-2)]\} \\
&\quad - \mu_2 J(t), \quad \text{for } t_i < t \leq t_{i+1} \\
J(t_{i+1}^+) &= J(t_{i+1}^-) + C_2(t_{i+1}^- - 2)e^{-2\mu_2} \quad (51)
\end{align*}
\]

with initial conditions: \( C_1(0) = C_{01}, C_2(0) = C_{02}, J(0) = J_0 \), and a trajectory of the solution in the past: \( C_1(\theta) = \Gamma(\theta), C_2(\theta) = \Psi(\theta), J(\theta) = \Omega(\theta) \) for \( \theta \in [-2, 0) \).

Equivalently, by assuming a fixed delay \( \tau = 2 \) for each patient, we can describe the number of VRE colonized individuals through admission to be isolated as a function of time. In other words, exactly two units of time after a patient is admitted, he/she is isolated. Assuming a Heaviside distribution with unit jump at \( \tau = 2 \) (that corresponds to a Dirac delta density) the system can be written as
\[ \frac{dC_1(t)}{dt} = m\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t) + C_2(t) + J(t)]\} \]
\[ - me^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)\int_{-\infty}^{0} [C_1(t + \tau) + C_2(t + \tau) + J(t + \tau)]\delta_{-\tau}(\tau)d\tau\} \]
\[ - \mu_2 C_1(t) \]
\[ \frac{dC_2(t)}{dt} = \beta\{N - [C_1(t) + C_2(t) + J(t)]\}[C_1(t) + C_2(t) + (1 - \gamma)J(t)] \]
\[ - \mu_2 C_2(t), \text{ for } t_i < t \leq t_{i+1} \]
\[ C_2(t_{i+1}) = C_2(t_{i+1}^-) - C_2(t_{i+1}^- - 2)e^{-2\mu_2} \]
\[ \frac{dJ(t)}{dt} = me^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)\int_{-\infty}^{0} [C_1(t + \tau) + C_2(t + \tau) + J(t + \tau)]\delta_{-\tau}(\tau)d\tau\} \]
\[ - \mu_2 J(t), \text{ for } t_i < t \leq t_{i+1} \]
\[ J(t_{i+1}) = J(t_{i+1}^-) + C_2(t_{i+1}^- - 2)e^{-2\mu_2}. \]

5.1 Derivation of the model

It is assumed that swab cultures are administered to every patient that is admitted (100% swab compliance) and that it takes two days for the test results to be returned. Therefore, two days after a colonized patient has been admitted, he/she will be isolated
if he/she is not discharged over the waiting period of two days. Therefore, the isolation rate of VRE colonized patients tested on admission depends on the past history of these patients.

The rate of change in the total population of $C_1$’s can be modeled by considering the rate of admission to this compartment minus the rate at which $C_1$’s are isolated and minus the rate at which $C_1$’s are discharged before isolation. Since $m\Lambda(t)$ is the admission rate of VRE colonized patients, then $m\Lambda(t - 2)$ is the rate at which the $C_1$’s were admitted at time $t - 2$. Some of those admitted at time $t - 2$ can get discharged at a rate $\mu_2$ before being isolated over the period of two days. Thus, we need to find the fraction of those admitted at time $t - 2$ that are still admitted two days after.

We consider the “cohort” of patients who were all admitted at $t = 0$, denoted by $C_1(0)$. Let $C_1(2)$ denote the number of these who are still in $C_1$ class 2 days later. If patients leave $C_1$ class at the rate $\mu_2$ per day, then

$$\frac{dC_1(t)}{dt} = -\mu_2C_1(t)$$

with initial condition $C_1(0) = C_{01}$. Hence

$$\frac{C_1(2)}{C_1(0)} = e^{-2\mu_2}$$

denotes the fraction of individuals who were admitted at time $t = 0$ and who are still admitted at $t = 2$, or in probabilistic language this denotes the probability of a patient still being admitted at time $t = 2$ given that he/she was admitted at time $t = 0$. Thus, the rate of isolation of $C_1$’s at time $t$ by factoring in discharges over the period of two days is $m\Lambda(t - 2)e^{-2\mu_2}$. Therefore, the rate of change of $C_1$’s can be modeled as

$$\frac{dC_1(t)}{dt} = m\Lambda(t) - m\Lambda(t - 2)e^{-2\mu_2} - \mu_2C_1(t)$$

with initial condition $C_1(0) = C_{01}$.

Consequently, the rate of change of isolated VRE colonized patients ($J$) at time $t$ can be modeled as

$$\frac{dJ(t)}{dt} = m\Lambda(t - 2)e^{-2\mu_2} - \mu_2J(t),$$

which is the rate of $C_1$’s entering to $J$ class minus the discharge rate in unit time and initial condition $J(0) = J_0$.

Let us consider now the influence of swab cultures administered once a week in isolating VRE colonizations during hospital stays, i.e., $C_2 \rightarrow J$. Swab cultures once a week were administered every Tuesday and assuming that every patient is tested (100% compliance) with two days for the test results to be returned (sometime before health care worker’s night shift), we can assume that VRE colonized patients in $C_2$ class will be moved into isolation every Thursday night. Hence, there is not a delay involved here
Table 11: Description of the days for the process $C_2(t)$

<table>
<thead>
<tr>
<th>Description</th>
<th>Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thursday night - Friday night</td>
<td>0/7</td>
</tr>
<tr>
<td>Friday night - Saturday night</td>
<td>1</td>
</tr>
<tr>
<td>Saturday night - Sunday night</td>
<td>2</td>
</tr>
<tr>
<td>Sunday night - Monday night</td>
<td>3</td>
</tr>
<tr>
<td>Monday night - Tuesday night</td>
<td>4</td>
</tr>
<tr>
<td>Tuesday night - Wednesday night</td>
<td>5</td>
</tr>
<tr>
<td>Wednesday night - Thursday night</td>
<td>6</td>
</tr>
</tbody>
</table>

but rather a jump discontinuity every Thursday night.

We assume that our process starts on Thursday night (with the days defined in Table 11) and we let $t_{i+1}$ be the next Thursday night or the next time in which $C_2$’s are isolated. We can then expect a jump discontinuity in the number of patients in $C_2$ class as well as in the number of patients in $J$ class at time $t_{i+1}$. If $C_2(t_{i+1}^+)$ represents the number of patients in $C_2$ after isolation, then

$$C_2(t_{i+1}^+) = C_2(t_{i+1}^-) - C_2(t_{i+1} - 2)e^{-2\mu_2}.$$  \hspace{1cm} (57)

Here, $C_2(t_{i+1} - 2)$ represents the number of patients in $C_2$ class that were tested on Tuesday and $e^{-2\mu_2}$ is the fraction of those $C_2(t_{i+1} - 2)$ that were not discharged over the period of two days before isolation (it follows from the same derivations as in (54)).

As we can see, this jump discontinuity in $C_2$ influences a jump in $J$ defined as

$$J(t_{i+1}^+) - J(t_{i+1}^-) = C_2(t_{i+1} - 2)e^{-2\mu_2}.$$ \hspace{1cm} (58)

The number of isolated patients after isolation, $J(t_{i+1}^+)$, can be determined by the number of patients in $J$ class before isolation plus the number of patients in $C_2$ class that were isolated at time $t_{i+1}$. Therefore, the dynamics for compartment $C_2$ and $J$ at time $t$ for $t_i < t \leq t_{i+1}$ are modeled by

$$\frac{dC_2(t)}{dt} = \beta U(t) [C_1(t) + C_2(t) + (1 - \gamma)J(t)] - \mu_2 C_2(t)$$ \hspace{1cm} (59)

$$\frac{dJ(t)}{dt} = m\Lambda(t - 2)e^{-2\mu_2} - \mu_2 J(t),$$ \hspace{1cm} (60)

with initial condition $C_2(0) = C_{02}, J(0) = J_0$.

Finally, we assume the overall admission rate equal to the overall discharge rate, i.e., $\Lambda(t) = \mu_1 U(t) + \mu_2 [C_1(t) + C_2(t) + J(t)]$. In order to keep the overall rate of admission equal to the overall rate of discharge, we replace $U(t) = N - [C_1(t) + C_2(t) + J(t)]$ giving the system of equations defined in (51).
5.2 Numerical implementations

The solutions to the System (51) can be simulated using an algorithm developed by Banks and Kappel [7] for nonlinear systems (see also [8, 10] for applications of this algorithm). The idea behind the algorithm is to first convert the system to an abstract evolution equation (AEE) and consider a space spanned by piecewise linear splines. Then, an approximation to the solution of System (51) is obtained by calculating numerically the generalized Fourier coefficients of the approximate solutions corresponding to the splines. With these coefficients we can recover an approximation to the solution of the system. This approach provides global existence and uniqueness of the solution (see [8] for corresponding proofs).

5.2.1 Abstract evolution equation formulation

Let \( x(t) \) represent the state of the System (51) at time \( t \)

\[
x(t) = (C_1(t), C_2(t), J(t))^T,
\]

and the state of the system due to the delay as

\[
x_t(\tau) = x(t + \tau), \quad -2 \leq \tau \leq 0.
\]

Then our System (51) can be written as

\[
\frac{dx(t)}{dt} = g + L(x(t), x_t) + f(x(t)) \quad \text{for} \quad t_i < t \leq t_{i+1}
\]

\[
(x(0), x_0) = (\Phi(0), \Phi) \in Z,
\]

(61)

where \( t > 0 \) and \( \Phi \in C(-2, 0; \mathbb{R}^3) \) is the trajectory history function of the system defined in \([-2, 0]\). We define \( Z \equiv \mathbb{R}^3 \times L^2(-2, 0; \mathbb{R}^3) \) as an infinite dimensional Hilbert space with norm

\[
\| (\eta, \phi) \|_Z = \left( |\eta|^2 + \int_{-2}^{0} |\phi(\tau)|^2 d\tau \right)^{1/2},
\]

where \((\eta, \phi) \in Z \) and \( L^2 \) is the space of square integrable functions, and inner product

\[
\langle (\eta, \phi), (\zeta, \psi) \rangle_Z = \eta^T \zeta + \int_{-2}^{0} \phi(\tau)^T \psi(\tau) d\tau
\]

with \((\eta, \phi), (\zeta, \psi) \in Z \).

In Model (61), \( g \) is the state independent part of the system, \( L(x(t), x_t) \) is the linear part of the system, and \( f(x(t)) \) is the nonlinear part of the system. If we let \( \delta(\tau) \) be the Dirac delta ‘density’ (by assuming a Heaviside distribution with a unit jump at \( \tau = -2 \)), we have
With respect to the nonlinear terms in $f(\eta)$, a more realistic model requires that these terms be bounded in the limit (i.e., saturation should be considered in the nonlinear terms so that in the limit it is at least affine in $x_1$ or $x_2$ or $x_3$). For well posedness considerations the nonlinear terms can be replaced by a function such as:

$$
\beta_i(x_i) = \begin{cases} 
0 & x_i < 0 \\
\beta x_i & 0 \leq x_i \leq \bar{x}_i \\
\beta \bar{x}_i & \bar{x}_i < x_i
\end{cases}
$$

with $\bar{x}_i \in \mathbb{R}^+$ as finite upper bounds and $i = 1, 2, 3$. Then $f(\eta)$ can be replaced by

$$
\tilde{f}(\eta) = \begin{bmatrix} 
0 & 0 & 0 \\
-\beta(\eta_1 + 2\eta_2 - (2 - \gamma)\eta_3) & -\beta(\eta_2 + (2 - \gamma)\eta_3) & -\beta(1 - \gamma)\eta_3
\end{bmatrix} \eta,
$$

which yields

$$
\frac{dx(t)}{dt} = g + L(x(t), x_t) + \tilde{f}(x(t)) \quad \text{for} \quad t_i < t \leq t_{i+1}
$$

$$
(x(0), x_0) = (\Phi(0), \Phi) \in Z.
$$
5.2.2 Abstract evolution equation implementation

We define a nonlinear operator \( A : D(A) \subset Z \rightarrow Z \) by

\[
A(\phi(0), \phi) = \left( L(\phi(0), \phi) + \hat{f}(\phi(0)), \frac{d}{dt} \phi \right)
\]

with domain defined as

\[
D(A) = \{(\phi(0), \phi) \in Z| \phi \in H^1(-2,0; \mathbb{R}^3)\}.
\]

where neither the nonlinear operator \( A \) nor the domain \( D(A) \) depends on \( t \). If we let \( z(t) = (x(t), x_t) \in Z \), then the delay System (63) can be formulated as

\[
\frac{dz(t)}{dt} = (g, 0) + Az(t)
\]

\[
z(0) = z_0.
\]  

(63)

Define \( Z^M \) to be the approximating piecewise linear spline subspace on \( Z \), \( P^M \) as the orthogonal projection of \( Z \) onto \( Z^M \), and \( A^M \) as the approximating operator of \( A \) given by \( A^M = P^M A P^M \). Then, the System (63) can be approximate by the finite dimensional problem

\[
\frac{dz^M(t)}{dt} = P^M (g, 0) + A^M z^M(t)
\]

\[
z^M(0) = P^M z_0.
\]  

(64)

Let \( Z^M_1 \) be a subspace of \( Z^M \). By partitioning the interval \([-2,0]\) with \( t^M_j = -j(2/M) \) for \( j = 0, ..., n \), we can define a basis \( \hat{\beta}^M \) by

\[
\hat{\beta}^M = (\beta^M(0), \beta^M) \quad \text{where} \quad \beta^M = (\beta^M_0, \beta^M_1, ..., \beta^M_M) \otimes I_3,
\]

where an element in \( Z^M_1 \) can be written as

\[
z^M = \hat{\beta}^M \alpha^M = \sum_{j=0}^{M} (\beta^M_j(0), \beta^M_j)a_j^M,
\]

with \( a_j^M \in \mathbb{R}^M \) and the \( \beta^M_j \)'s are piecewise linear functions defined by

\[
e^M_0(t) = \begin{cases} 
\frac{t - t^M_1}{2/M} & t^M_1 \leq t \leq 0 \\
0 & \text{otherwise}
\end{cases}
\]

\[
e^M_M(t) = \begin{cases} 
1 - \frac{t + 2}{2/M} & -2 \leq t \leq t^M_{M-1} \\
0 & \text{otherwise}
\end{cases}
\]
and, for \( j = 1, \ldots, M - 1 \),
\[
e^M_j(t) = \begin{cases} \frac{t - t^{M}_{j+1}}{2/M}, & t^{M}_{j+1} \leq t \leq t^{M}_j \\ 1 - \frac{t - t^{M}_j}{2/M}, & t^{M}_j \leq t \leq t^{M}_{j-1} \\ 0 & \text{otherwise} \end{cases}
\]
with \( e^M_j(t^{M}_i) = \delta_{ij} \) for \( i, j = 0, \ldots, M \).

Define \( A^M_1 \) as a matrix representation of \( A^M \) restricted to the subspace \( Z^M_1 \) of \( Z^M \) and let \( w^M(t) \) and \( K^M \) be defined such that \( z^M(t) = \hat{\beta}^M w^M(t) \) and \( P^M(g, 0) = \hat{\beta}^M K \). Then, solving for \( z^M(t) \) in the finite dimensional System (64) is equivalent to solving for \( w^M(t) \) in the system
\[
\frac{dw^M(t)}{dt} = K^M + A^M_1 w^M(t)
\]
\( w^M(0) = w^M_0, \quad (65) \)

where \( \hat{\beta}^M w^M_0 = P^M z_0 \). Note that having obtained \( w^M \), the product \( \hat{\beta}^M w^M(t) \) converges uniformly in \( t \) to the solution of (63).

In order to approximate \( P^M(\eta, \phi) \) for any \( (\eta, \phi) \in Z \), where \( P^M(\eta, \phi) \) is the orthogonal projection of \( (\eta, \phi) \in Z \) onto \( Z^M \), assume \( P^M(\eta, \phi) = \hat{\beta}^M u^M \) where \( u^M \in \mathbb{R}^3 \) then
\[
0 = \langle \hat{\beta}^M u^M - (\eta, \phi) \rangle_Z
\]
which implies that
\[
\langle \hat{\beta}^M, \hat{\beta}^M \rangle u^M = \langle \hat{\beta}^M, (\eta, \phi) \rangle_Z.
\]  
(66)
The orthogonal projection \( P^M \) is uniquely determined by solving (66) for \( u^M \) and implies that
\[
K^M = \langle \langle \hat{\beta}^M, \hat{\beta}^M \rangle_Z \rangle^{-1} \langle \hat{\beta}^M, (g, 0) \rangle_Z
\]
and
\[
u^M_0 = \langle \langle \hat{\beta}^M, \hat{\beta}^M \rangle_Z \rangle^{-1} \langle \hat{\beta}^M, (x(0), x_0) \rangle_Z,
\]
where the inner products are given by
\[
\langle \hat{\beta}^M, \hat{\beta}^M \rangle_Z = \begin{bmatrix} 1 + \frac{2}{3M} & \frac{1}{3M} & 0 & \cdots & \cdots & 0 \\ \frac{1}{3M} & \frac{4}{3M} & \frac{1}{3M} & \cdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \frac{1}{3M} & \frac{4}{3M} & \frac{1}{3M} & \frac{2}{3M} & \frac{3}{3M} & \frac{3}{3M} \end{bmatrix}_{(M+1) \times (M+1)} \otimes I_3,
\]
\[ \langle \hat{\beta}^M, (g(t), 0) \rangle_\mathbb{Z} = \begin{bmatrix} g(t) \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}_{3(M+1)\times 1}, \]

and

\[ \langle \hat{\beta}^M, (x(0), x_0) \rangle_\mathbb{Z} = \begin{bmatrix} x(0) \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}_{3(M+1)\times 1} + \int_{-2}^{0} \beta^{MT}(\tau)x_0(\tau)d\tau. \]

Similarly, in order to approximate \( A_1^M \alpha^M \) for any \( \alpha^M \in \mathbb{R}^3 \), we note

\[ A^M \hat{\beta}^M \alpha^M = P^M(A \hat{\beta}^M \alpha^M) = \hat{\beta}^M A_1^M \alpha^M \]

and

\[ P^M(A \hat{\beta}^M \alpha^M) = P^M(L(\beta^M(0)\alpha^M, \beta^M \alpha^M) + \hat{f}(\beta^M(0)\alpha^M), \frac{d}{d\tau}(\beta^M \alpha^M)). \]

Thus

\[ 0 = \hat{\beta}^M A_1^M \alpha^M - P^M(L(\beta^M(0)\alpha^M, \beta^M \alpha^M) + \hat{f}(\beta^M(0)\alpha^M), \frac{d}{d\tau}(\beta^M \alpha^M)) \]

and

\[ 0 = \langle \hat{\beta}^M, \hat{\beta}^M A_1^M \alpha^M - (L(\beta^M(0)\alpha^M, \beta^M \alpha^M) + \hat{f}(\beta^M(0)\alpha^M), \frac{d}{d\tau}(\beta^M \alpha^M)) \rangle_\mathbb{Z}, \]

which implies that

\[ \langle \hat{\beta}^M, \hat{\beta}^M \rangle_\mathbb{Z}(A_1^M \alpha^M) = \langle \hat{\beta}^M, (L(\beta^M(0)\alpha^M, \beta^M \alpha^M) + \hat{f}(\beta^M(0)\alpha^M), \frac{d}{d\tau}(\beta^M \alpha^M)) \rangle_\mathbb{Z}. \]

Therefore, \( A_1^M \alpha^M \) is uniquely defined by solving (67) and implies that

\[ A_1^M w^M(t) = ((\hat{\beta}^M, \hat{\beta}^M)_\mathbb{Z})^{-1} Q^M w^M(t) \]

where

\[ Q^M = \langle \beta^M(0), L(\beta^M(0), \beta^M) + \hat{f}(\beta^M(0)) \rangle + \langle \beta^M, \hat{\beta}^M \rangle = Q_1^M + Q_2^M \]

63
and the inner products are defined by

\[
Q_1^M = \begin{bmatrix}
L_0^M + E_0^M + F_0^M & E_1^M & \cdots & E_M^M \\
0 & 0 & 0 & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & 0
\end{bmatrix}_{3(M+1)\times 3(M+1)}
\]

with

\[
L_0^M = \begin{bmatrix}
m(\mu_2 - \mu_1) - \mu_2 & m(\mu_2 - \mu_1) & m(\mu_2 - \mu_1) \\
\beta N & \beta N - \mu_2 & \beta N(1 - \gamma) \\
0 & 0 & -\mu_2
\end{bmatrix}
\]

\[
E_j^M = me^{-2\mu_2(\mu_2 - \mu_1)} \begin{bmatrix}
-1 & -1 & -1 \\
0 & 0 & 0 \\
1 & 1 & 1
\end{bmatrix} \int_{-2}^{0} e_j^M(\tau)\delta_{-2}(\tau) d\tau, \ \text{for} \ j = 0, \ldots, M,
\]

(note that since \(\delta_{-2}(\tau)\) is a Dirac distribution, \(E_j^M = 0\) for \(j = 0, \ldots, M-1,\))

\[
F_0^M = \begin{bmatrix}
0 & 0 & 0 \\
0 & -\beta_1(w_1) - 2\beta_2(w_2) - (2 - \gamma)\beta_3(w_3) & -\beta_2(w_2) - (2 - \gamma)\beta_3(w_3) - (1 - \gamma)\beta_3(w_3) \\
0 & 0 & 0
\end{bmatrix}
\]

and

\[
Q_2^M = \begin{bmatrix}
\frac{1}{2} & -\frac{1}{2} & 0 & \cdots & \cdots & 0 \\
\frac{1}{2} & 0 & -\frac{1}{2} & \ddots & \vdots \\
0 & \cdots & \cdots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \cdots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & 0 & \frac{1}{2} & -\frac{1}{2}
\end{bmatrix}_{(n+1)\times (n+1)} \otimes I_3.
\]

5.2.3 Convergence of solutions

Before performing an inverse problem using the discrete model with delay (51) we need to know how many partitions \(M\) to take in the interval \([-2, 0]\) so that the solutions of (64) converge to the solution of (63) as \(M \to \infty\). In order to do this, we carried out the forward problem for increasing values of \(M\) with the parameter values presented on Table 12. Note that the parameter values chosen relates to the inverse problem formulation
Table 12: Parameters values used in forward type problem to determine converge of solutions

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Oncology Unit (N=37)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Lambda )</td>
<td>( 0.16U(t) + 0.08(C(t) + J(t)) )</td>
</tr>
<tr>
<td>( m )</td>
<td>0.04</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.0049</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>0.58</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.29</td>
</tr>
<tr>
<td>( \mu_1 )</td>
<td>0.16</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 13: Computational time required to solved the discrete model with delay (51) for increasing number of partitions, \( M \), using oncology unit parameters.

<table>
<thead>
<tr>
<th>( M/\text{partitions} )</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1.99</td>
</tr>
<tr>
<td>24</td>
<td>18.17</td>
</tr>
<tr>
<td>48</td>
<td>194.44</td>
</tr>
<tr>
<td>96</td>
<td>2,402.10</td>
</tr>
</tbody>
</table>

results from Section 3 corresponding to the oncology unit.

We used Matlab 45 to solve System (65) with \( M = 12, 24, 48, 96 \). In Figure 21 we present the solutions for each \( M \) partition using the oncology unit parameters. Note that solutions considering \( M = 24, 48, 96 \) partitions are very close. However, the computational time required in solving the system in the time interval \( t = [0, 100] \) (see Table 13) suggests that any \( M \) between 24 and 48 seems to be a reasonable choice. This is because the computational time increases by a factor of 12 when we increase the partitions from \( M = 48 \) to \( M = 96 \). As a result, we chose \( M = 30 \).

5.2.4 Inverse problem results

Following the inverse problem methodology in Section 2, in Sections 2.1-2.5 we present the results of estimating \( \theta = (\beta) \) and \( \theta = (m, \beta) \) using Model (51). See Section 5.4 for more details about the analysis of numerical accuracy of our model.

As we mentioned before, the VRE surveillance data does not report the number of patients in isolation during weekends and holidays. Not surprising, when trying to fit the
Model (51) to the data, we found that there were additional observations missing that did not correspond to weekends and holidays. Hence, we identified a period of about three months (January 17, 06 to April 13, 06) that only had weekends as missing values to carry out the inverse problem. Since our model solves through weekends, in order to carry out the fitting we omitted the weekend values from the cost functional.

Results from estimating $\theta = (\beta)$ are shown in Figure 22 and Figure 23. Results from estimating $\theta = (m, \beta)$ are shown in Figure 24 and Figure 25. Overall, from the residual analysis we can conclude that whether we use OLS or GLS, the errors are independent (residuals vs time are random). On the other hand, the residuals versus model plot for OLS compared to the residuals/model versus model for GLS have the same pattern indicating no difference in whether we use absolute error model or relative error model.

5.3 Model refinement

The inverse problem results from the previous section using Model (51) suggests that the model fit does not agree particularly well with the data. Therefore, we put more details in the model and compared it again to the data to see if there were possible improvements in the model fit.

Details that we can include in the model are in relation to patients that are VRE colonized on admission. The epidemiology laboratory is closed on weekends and this results in two consequences. First, pending test-results from patients admitted on a Thursday or a Friday will be back by Monday or Tuesday. Second, swab-tests taken
Figure 22: Residual analysis for the OLS and GLS optimization as a result of estimating $\theta = (\beta)$ using oncology unit surveillance data.
on patients admitted on a Saturday and a Sunday will be sent to the epidemiology laboratory on Monday, then these patients will be isolated by Wednesday. For simplicity, we are going to focus on the first detail to see if this provides any improvement in the model fit.

We assume that for patients that are admitted on Thursdays and Fridays, the test-results are back on Tuesdays. These patients will be moved into isolation by Tuesday night. As a result, we also have a jump discontinuity every Tuesday. If we let $\hat{t}_j$ be a Tuesday and $\hat{t}_{j+1}$ be the next Tuesday then

$$C_1(\hat{t}_{j+1}^+) = C_1(\hat{t}_{j+1}^-) - C_1(\hat{t}_{j+1} - 4)e^{-\mu_2}$$

represents the number of patients in $C_1$ after isolation on Tuesday. This is basically the number of patients in compartment $C_1$ before isolation minus the number of patients that were isolated at time $\hat{t}_{j+1}^-$. Hence, $C_1(\hat{t}_{j+1} - 4)$ represents the number of patients in $C_1$ on a Friday (this includes the ones that where admitted on a Thursday and have not been discharged by Friday) and $e^{-\mu_2}$ is the fraction of those $C_1(\hat{t}_{j+1} - 4)$ that were not discharged over a period of 4 days before isolation (this follows the same derivation as in (5.4)).

Consequently, this jump discontinuity in $C_1$ influences a jump in $J$ defined as

$$J(\hat{t}_{j+1}^+) = J(\hat{t}_{j+1}^-) + C_1(\hat{t}_{j+1} - 4)e^{-\mu_2},$$

Figure 23: Best fit model solutions to oncology unit surveillance data via OLS optimization, $\beta = 0.0039$. At each jump, data is fitted using the model value after isolation.
Figure 24: Residual analysis for the OLS and GLS optimization as a result of estimating $\theta = (m, \beta)$ using oncology unit surveillance data.
Figure 25: Best fit model solutions to oncology unit surveillance data via OLS optimization, \((\hat{m}, \hat{\beta}) = (0.1332, 0.008)\). At each jump, data is fitted using the model value after isolation.

where \(J(t_{j+1})\) represent the total number of patients in isolation after that the isolation on Tuesday takes place. The dynamics for compartment \(C_1\) at time \(t\) are modeled by

\[
\frac{dC_1(t)}{dt} = m\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t) + C_2(t) + J(t)]\}
- m e^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t-2) + C_2(t-2) + J(t-2)]\}
- \mu_2 C_1(t) \quad \text{for } \hat{t}_j < t < t_{Sat}
\]

(70)

\[
\frac{dC_1(t)}{dt} = m\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t) + C_2(t) + J(t)]\}
- \mu_2 C_1(t) \quad \text{for } t_F < t \leq \hat{t}_{j+1}
\]

(71)

with \(t_F\) as day Friday and \(t_{Sat}\) as day Saturday. Equation (70) models the rate of change of \(C_1\)'s from Wednesday through Friday in which isolation takes place by the corresponding rate. On the other hand, Equation (71) models the rate of change of \(C_1\)'s from Saturday through Tuesday in which isolation is not employed. Tuesday is included on the interval because it corresponds to the number of \(C_1\)'s on Tuesday before isolation.
takes place. Finally, the complete model that includes the new detail is given by

\[
\begin{align*}
\frac{dC_1(t)}{dt} &= m\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t) + C_2(t) + J(t)]\} \\
&\quad - me^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t - 2) + C_2(t - 2) + J(t - 2)]\} \\
&\quad - \mu_2 C_1(t) \quad \text{for } \tilde{t}_j < t < t_{Sat} \\
\frac{dC_1(t)}{dt} &= m\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t) + C_2(t) + J(t)]\} \\
&\quad - \mu_2 C_1(t) \quad \text{for } t_F < t \leq \tilde{t}_{j+1} \\
C_1(\tilde{t}_{j+1}^-) &= C_1(\tilde{t}_{j+1}^-) - C_1(\tilde{t}_{j+1}^- - 4)e^{-4\mu_2} \\
\frac{dC_2(t)}{dt} &= \beta\{N - [C_1(t) + C_2(t) + J(t)]\}[C_1(t) + C_2(t) + (1 - \gamma)J(t)] \\
&\quad - \mu_2 C_2(t), \quad \text{for } t_i < t \leq t_{i+1} \\
C_2(t_{i+1}^+) &= C_2(t_{i+1}^-) - C_2(t_{i+1}^- - 2)e^{-2\mu_2} \\
\frac{dJ(t)}{dt} &= me^{-2\mu_2}\{\mu_1 N + (\mu_2 - \mu_1)[C_1(t - 2) + C_2(t - 2) + J(t - 2)]\} \\
&\quad - \mu_2 J(t), \quad \text{for } \tilde{t}_{j+1} < t < t_{Sat} \\
\frac{dJ(t)}{dt} &= -\mu_2 J(t), \quad \text{for } t_F < t < \tilde{t}_{j+1} \\
J(\tilde{t}_{j+1}^+) &= J(\tilde{t}_{j+1}^-) + C_1(\tilde{t}_{j+1}^- - 4)e^{-4\mu_2} \\
J(t_{i+1}^-) &= J(t_{i+1}^-) + C_2(t_{i+1}^- - 2)e^{-2\mu_2}. \\
\end{align*}
\]

with \(t_F\) as day Friday, \(t_{Sat}\) as day Saturday, \(\tilde{t}_j\) as isolation-day on Tuesday, \(t_i\) as isolation-day on Thursday, initial conditions: \(C_1(0) = C_{01}, C_2(0) = C_{02}, J(0) = J_0\), and a trajectory of the solution in the past: \(C_1(\theta) = \Gamma(\theta), C_2(\theta) = \Psi(\theta), J(\theta) = \Omega(\theta)\) for \(\theta \in [-2, 0]\).

### 5.3.3 Inverse problem results

In this section we present the results of estimating \(\theta = (\beta)\) using the refined Model (72). We also verify the numerical accuracy of the refined model. See Section 5.4 for more details.

Figure 26 depicts the residual analysis and Figure 27 depicts the best fit. The results indicate that there is not improvement in the model fit to the data. That is, the additional detail in the model corresponding to jumps on Tuesdays does not have a significant effect in the fit to the data.

### 5.4 Comparison between models using forward analysis

In this section we want to compare Model (51) with Model (72) as well as to show the numerical accuracy of these models when carrying out an inverse problem.

In order to compare both models, we simulated a solution with the parameters values
Figure 26: Refined model: Residual analysis for the OLS and GLS optimization as a result of estimating $\theta = (\beta)$ using oncology unit surveillance data.
Figure 27: Refined model: Best fit model solutions to oncology unit surveillance data via OLS optimization, $\beta = 0.0039$. Note jumps every 5 days corresponding to every Tuesdays and jumps every 7 days corresponding to Thursdays. At each jump, data is fitted using the model value after isolation.
recorded in Table 14. Figure 28 contains a plot of the forward solutions of the models in comparison to the data. The plots indicates that the dynamic of Model (51) in comparison to Model (72) are essentially the same. In the first 14 days we see a little difference but after that there is no qualitative difference but the Tuesdays’ jump in Model (72). We can conclude that with the addition of more details to Model (51) there seems to be no significant improvement in what Model (51) do.

In order to test the numerical accuracy of the models, we added noise (constant variance error) to the forward solution and carried out an inverse problem using the noise-added solution as data to attempt to recover the original parameters. In other words, a synthetic data \( \{y_j\} \) for \( j = 1, \ldots, n \) is constructed by adding variability to the model solution, \( f(t_j, \theta_0) = J(t_j, \theta_0) \). The statistical model that captures the variability is

\[
y_j = f(t_j, \theta_0) + \sigma z_j
\]

where \( z_j \) is a standard normal variable (i.e., \( z_j \sim N(0, 1) \)) and \( \sigma \) is the constant variability. The magnitude of \( \sigma \) determines how much noise is added.

We conducted the inverse problem to estimate \( \theta = (\beta) \) and \( \theta = (m, \beta) \) via ordinary least squares (OLS) with different noise levels: \( \sigma = 0, 0.10, 0.30, 0.50 \) and \( 0.70 \). In Tables 15 and 16 we summarize the results in relation to the estimation of \( \theta = (\beta) \) and \( \theta = (m, \beta) \) corresponding to Model (51) and Model (72). Fitting results are presented in Figures 29 and 30 for Model (51), and Figures 31 and 32 for Model (72). We can conclude that the estimation procedure was a success using both models.

### 5.5 Conclusions

In this section we introduce a discrete VRE model with delay that incorporates specific details with respect to the isolation procedure employed to patients in a hospital unit. We attempt to estimate some of the epidemiological parameters such as the transmission

---

**Table 14: Parameters values used in the forward type problem to compare models**

<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Oncology Unit (N=37)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x(0) )</td>
<td>([2;4;11])</td>
</tr>
<tr>
<td>( x_0 )</td>
<td>([4,2;5,3;9,10])</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Lambda )</td>
<td>( 0.16U(t) + 0.08(C(t) + J(t)) )</td>
</tr>
<tr>
<td>( m )</td>
<td>0.04</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.0039</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>0.58</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.29</td>
</tr>
<tr>
<td>( \mu_1 )</td>
<td>0.16</td>
</tr>
<tr>
<td>( \mu_2 )</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Figure 28: Solutions of the models using forward type problem in comparison to the oncology unit surveillance data.
Table 15: OLS optimization testing using Model (51). The model was fit to generate data with $\sigma = 0, 0.10, 0.30, 0.50, 0.70$ noise where subscripts in $\theta_\sigma$ denote the level of noise in the synthetic data.

<table>
<thead>
<tr>
<th></th>
<th>$m$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $\theta$</td>
<td>0.04</td>
<td>0.0039</td>
</tr>
<tr>
<td>Initial $\theta$</td>
<td>0.03</td>
<td>0.0029</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0}$</td>
<td>-</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.10}$</td>
<td>-</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.30}$</td>
<td>-</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.50}$</td>
<td>-</td>
<td>0.0041</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.70}$</td>
<td>-</td>
<td>0.0041</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.04}$</td>
<td>0.04</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.10}$</td>
<td>0.0392</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.30}$</td>
<td>0.0386</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.50}$</td>
<td>0.0364</td>
<td>0.0042</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.70}$</td>
<td>0.0334</td>
<td>0.0043</td>
</tr>
</tbody>
</table>
Figure 29: Model (51) fit to the synthetic data using OLS optimization procedure to estimate $\theta = (\beta)$. 
Figure 30: Model (51) fit to the synthetic data using OLS optimization procedure to estimate θ = (m, β).
Table 16: OLS optimization testing using Model (72). The model was fit to generate data with $\sigma = 0, 0.10, 0.30, 0.50, 0.70$ noise where subscripts in $\theta_\sigma$ denote the level of noise in the synthetic data.

<table>
<thead>
<tr>
<th></th>
<th>$m$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $\theta$</td>
<td>0.04</td>
<td>0.0039</td>
</tr>
<tr>
<td>Initial $\theta$</td>
<td>0.03</td>
<td>0.0029</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0}$</td>
<td>-</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.10}$</td>
<td>-</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.30}$</td>
<td>-</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.50}$</td>
<td>-</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.70}$</td>
<td>-</td>
<td>0.0041</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0}$</td>
<td>0.04</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.10}$</td>
<td>0.0407</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.30}$</td>
<td>0.0404</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.50}$</td>
<td>0.0424</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.70}$</td>
<td>0.0449</td>
<td>0.0040</td>
</tr>
</tbody>
</table>
Figure 31: Model (72) fit to the synthetic data using OLS optimization procedure to estimate $\theta = (\beta)$. 

(a) Level of noise: $\sigma=0.10$

(b) Level of noise: $\sigma=0.30$

(c) Level of noise: $\sigma=0.50$

(d) Level of noise: $\sigma=0.70$
Figure 32: Model (72) fit to the synthetic data using OLS optimization procedure to estimate $\theta = (m, \beta)$
rate \((\beta)\) and the fraction \((m)\) of patients VRE colonized on admission. Results suggested that there was not much success in the model fit to the data, especially when trying to estimate \(\theta = (m, \beta)\). We followed with a model refinement that includes additional detail to see if an improvement in the model fit to the data is achieved. Results show no improvement in the model fit to the data. However, we used simulated data (produced by the model with noise) to validate our methodology and our ability to successfully estimate parameters in the model. Thus we demonstrated that if we have data that looks like the model process, we can successfully carry out the inverse problem, even with very noisy data.

We conclude that the VRE surveillance data appears to be very irregular as it does not seem to follow any standard process. One reason that the data has little to do with the process could be due to the fact that there is not a true parameter \(\theta_0\) in the statistical model, \(Y_j = f(t_j, \theta_0) + \epsilon_j\), that can generate the data we have from this hospital unit. The assumption of existence of such a \(\theta_0\) is an essential foundation of much of the statistical and mathematical methodology for inverse problems (see \([9]\)). Another possible reason could be that it is likely that the underlying assumption for statistical model should be modified to \(Y_j = f(t_j, \theta_0) + g(t_j, \theta_0)\epsilon_j\), where we do not know \(g(t_j, \theta_0)\). Another possibility is that there is no underlying model \(f(t_j, \theta_0)\) that describes this data. In other words, this data is just so irregular due to inherent irregularities under data reporting.

The models describe a data collection process as described to us by the health workers. Thus we suggest that there are irregularities in reporting or in their descriptions of what they actually do. In Section 3.5 we described inconsistencies in the data reporting. In addition, we found glaring discrepancies in the data corresponding to the general unit.

For the month of October of 2004 the dates of the reported data did not match the 2004 Calendar. Even though the rest of the data seems to be fine, we were suspicious that this error affected the rest of the data. This error also puts in question the regularity of the oncology data. Such discrepancies are not uncommon in data reporting depending on a diverse population of reporting individuals.

6 Dealing with Missing Data

The EM algorithm is a popular tool in statistical estimation problems involving incomplete data. By incomplete data we mean that values at some time point are missing. This iterative procedure computes the Maximum Likelihood (ML) estimate of the model parameters in the presence of missing data. Each iteration consists of two processes: the E-step and the M-step. In the E-step the missing data are estimated by computing an expectation of the complete-data log likelihood given the observed data and the parameter estimates

\[
Q(\theta|\hat{\theta}^{(a)}) = E[\log L(\theta|Y_{obs}, Y_{mis})| Y_{obs}, \hat{\theta}^{(a)}].
\]  

Then, given a complete-data log likelihood, the M-step finds the parameter estimates to maximize the complete-data log likelihood from the E-step,

\[
\theta^{(a+1)} = \arg \min_{\theta} Q(\theta|\hat{\theta}^{(a)}).
\]
The two steps are iterated until the iterations converge. It is assumed convergence eventually occurs since the algorithm is guaranteed to increase the likelihood at each iteration.

6.1 EM algorithm derivation

In this section we review a basic theory supporting the use of maximum likelihood estimate and describe how it is implemented in the incomplete data setting.

We assume that a complete data exists, where we can partition the complete data into observed (or incomplete) and missing components, i.e. $Y = (Y_{obs}, Y_{mis})$. Then, we can write a joint density function corresponding to the complete data as

$$P(Y|\theta) = P(Y_{obs}, Y_{mis}|\theta) = P(Y_{obs}|\theta)P(Y_{mis}|Y_{obs}, \theta),$$

where $P(Y_{obs}|\theta)$ is the density of the observed data and $P(Y_{mis}|Y_{obs}, \theta)$ is the density of the missing data given the observed data. The decomposition of the likelihood of the complete data is

$$\log L(Y|\theta) = \log L(Y_{obs}|Y_{mis}) = \log L(Y_{obs}|\theta) + \log P(Y_{mis}|Y_{obs}, \theta).$$

We want to estimate $\theta$ by maximixing the observed data loglikelihood with respect to $\theta$ for fixed $Y_{obs}$ given by

$$L(\theta|Y_{obs}) = \int P(Y|\theta)dY_{mis},$$

however this can be difficult. We write

$$\log L(\theta|Y_{obs}) = \log L(\theta|Y) - \log P(Y_{mis}|Y_{obs}, \theta).$$

The main idea is that it is much easier to optimize $\log L(\theta|Y)$ if we had known the values for $Y_{mis}$. Then, taking the conditional expectation given $Y_{obs}$ of both sides of (79), we have

$$\log P(\theta|Y_{obs}) = Q(\theta|\theta^{(a)}) - H(\theta|\theta^{(a)})$$

where

$$Q(\theta|\theta^{(a)}) = E_{\theta^{(a)}}[\log L(\theta|Y)|Y_{obs}]$$

$$= \int \log L(\theta|Y_{obs}, Y_{mis})L(Y_{mis}|Y_{obs}, \theta^{(a)})dY_{miss}$$

(81)
and

\[
H(\theta|\theta^{(a)}) = E_{\theta^{(a)}}[\log L(Y_{\text{mis}}|Y_{\text{obs}}, \theta)|Y_{\text{obs}}] = \int \log L(Y_{\text{miss}}|Y_{\text{obs}}, \theta)L(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(a)})dY_{\text{miss}}.
\]  

(82)

Note that the quantity \(Q(\theta^{(a+1)}|\theta^{(a)}) - Q(\theta^{(a)}|\theta^{(a)})\) is non-negative because \(\theta^{(a+1)}\) has been chosen to satisfy

\[
Q(\theta^{(a+1)}|\theta^{(a)}) \geq Q(\theta|\theta^{(a)}) \text{ for all } \theta.
\]  

(83)

Also, by using Jensen’s inequality one can show that \(H(\theta|\theta^{(a)}) \leq H(\theta^{(a)}|\theta^{(a)})\).

As a result, the algorithm that involves iterating between the following 2 steps and should in practice converge to the value of \(H(\theta|\theta^{(a)})\) is the following:

1. E-step: for current iterate \(\theta^{(a)}\), calculate the conditional expectation \(Q(\theta|\theta^{(a)})\);
2. M-step: Maximize \(Q(\theta|\theta^{(a)})\) in \(\theta\) to obtain \(\theta^{(a+1)}\). Set \(a = a + 1\) and return to the E-step.

6.2 EM algorithm application to our problem

An important part in using an EM algorithm is to understand the mechanism that creates the missing data. The missing values in our problem correspond to the missing counts at particular time points excluding the weekends. In our case, we can assume that the fact of omission of the data is independent of the missing segment values. In other words, the probability that a particular segment of \(Y\) is missing does not depend on the values in the segment. Therefore, we can assume that the missing data is missing at random (MAR). In order to carry out the EM algorithm, we identified a period of about three months that contains missing values not including weekends. Figure 33 depicts a plot of the data for the time period January 6, 2005 to April 7 2005. This data contains ten missing data points.

Assuming that a complete data exists and is governed by the Statistical Model (2), we can write the full data loglikelihood function as

\[
\log L(\theta, \sigma^2|Y) = -n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{j=1}^{n} [Y(t_j) - f(t_j, \theta)]^2.
\]  

(84)

The main idea is that it is much easier to optimize \(\log L(\theta, \sigma^2|Y)\) if we had known the values for \(Y_{\text{mis}}\). Unfortunately, we do not know them and that is why we need the E-step. If we rearrange the data such that the first \(n_0\) are the observed data (this does not imply that last observations are always missing). Then the likelihood function for
the full data can be written as

$$L(\theta, \sigma^2 \mid Y) = \prod_{j=1}^{n_0} \frac{1}{\sigma} \phi \left( \frac{Y(t_j) - f(t_j, \theta)}{\sigma} \right) \prod_{j=n_0+1}^{n} \frac{1}{\sigma} \phi \left( \frac{Y(t_j) - f(t_j, \theta)}{\sigma} \right)$$  \hspace{1cm} (85)$$

and the loglikelihood function as

$$\log L(\theta, \sigma^2 \mid Y) = - \frac{1}{2\sigma^2} \sum_{j=1}^{n_0} [Y(t_j) - f(t_j, \theta)]^2 - \frac{1}{2\sigma^2} \sum_{j=n_0+1}^{n} [Y(t_j) - f(t_j, \theta)]^2 - n \log(\sigma)$$ \hspace{1cm} (86)$$

Thus, for the E-step and current estimates of the parameters, the algorithm calculates

1. \(E[\sum_{j=1}^{n} Y(t_j) \mid \theta^{(a)}, Y_{obs}] = \sum_{j=1}^{n_0} Y(t_j) + \sum_{j=n_0+1}^{n} f(t_j, \theta^{(a)})\);

2. \(E[\sum_{j=1}^{n} Y^2(t_j) \mid \theta^{(a)}, Y_{obs}] = \sum_{j=1}^{n_0} Y^2(t_j) + \sum_{j=n_0+1}^{n} [f^2(t_j, \theta^{(a)}) + \sigma_{(a)}^2]\).
Note that maximizing $Q(\theta^{(a+1)}|\theta^{(a)})$ corresponds roughly to solving for $\theta$ in

$$
\sum_{j=1}^{n}[Y_i - f(t_j, \theta)]\nabla f(t_j, \theta) = 0
$$

which is equivalent to maximizing for $\theta$ in

$$
\sum_{j=1}^{n}|Y(t_j) - f(t_j, \theta)|^2.
$$

The EM algorithm is the following:

1. Initialize: Set $\tilde{Y}_j = \tilde{Y}_{obs}$ for $n_0 + 1,...n$ and obtain $\theta^{(0)}$ by OLS. Let $\sigma^2_{(0)} = n^{-1} \sum_{j=1}^{n}[\tilde{Y}_j - f(t_j, \theta^{(0)})]^2$. Set $a = 0$;

2. E-step (expectation): Compute $Q(\theta^{(a+1)}|\theta^{(a)})$;

3. M-step (maximization): Estimate $\theta^{(a+1)}$ and $\sigma^2_{(a+1)}$ by

$$
\arg \min_{\theta} \sum_{j=1}^{n}[\tilde{Y}(t_j) - f(t_j, \theta)]^2;
$$

4. Iteratively improve the estimate of $\theta$ by alternating between steps 2 and 3;

5. Stop when the likelihood converges; that is, when $|\theta^{(a+1)} - \theta^{(a)}|$ is very small.

### 6.3 EM algorithm testing

Before using the EM algorithm with the oncology data, we are going to test the algorithm using synthetic data. We used the synthetic data generated in Section 5.4 corresponding to Model (51) using noise levels: $\sigma = 0, 0.10, 0.30, 0.50, 0.70$. Then we created an incomplete data by deleting units independently with probability 0.50, i.e., $Pr(Y(t_j) \text{ is observed}) = 0.50$.

In Table 17 we summarize the results corresponding to the estimation of $\theta = (\beta)$ and $\theta = (m, \beta)$. Figures 34 and 35 depict the resulting fit plotted against the complete set of synthetic data. These results demonstrate the ability of the EM algorithm in estimating the parameters. Note that as the level of noise increases the estimates are more affected although the estimates are not significantly different from the true values.

### 6.4 EM algorithm results using the oncology unit surveillance data

In this section we present the results of estimating $\theta = (\beta)$ and $\theta = (m, \beta)$ using Model (51) and data from the time period of January 6, 2005 to April 7 2005. Figure 36 and
Table 17: EM algorithm testing using Model (51). The model was fit to generate data with $\sigma = 0, 0.10, 0.30, 0.50, 0.70$ noise where subscripts in $\theta$ denote the level of noise in the synthetic data. N=35. Total of 8 observations were deleted independently with probability 0.50.

<table>
<thead>
<tr>
<th></th>
<th>$m$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True $\theta$</td>
<td>0.04</td>
<td>0.0039</td>
</tr>
<tr>
<td>Initial $\theta$</td>
<td>0.03</td>
<td>0.0029</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0}$</td>
<td>-</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.10}$</td>
<td>-</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.30}$</td>
<td>-</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.50}$</td>
<td>-</td>
<td>0.0040</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.70}$</td>
<td>-</td>
<td>0.0041</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.04}$</td>
<td>0.04</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.10}$</td>
<td>0.0414</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.30}$</td>
<td>0.0432</td>
<td>0.0039</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.50}$</td>
<td>0.0498</td>
<td>0.0038</td>
</tr>
<tr>
<td>$\hat{\theta}_{OLS}^{0.70}$</td>
<td>0.0514</td>
<td>0.0037</td>
</tr>
</tbody>
</table>
Figure 34: Model (51) fit to the complete-synthetic data using EM algorithm to estimate \( \theta = (\beta) \).
Figure 35: Model (51) fit to the complete-synthetic data using EM algorithm to estimate \( \theta = (m, \beta) \)
Figure 37 depict the residual analysis and the best fit as a result of estimating \( \theta = (\beta) \). Figures 38 and 39 illustrate the residual analysis and best fit as result of estimating \( \theta = (m, \beta) \). These results indicate that there was little success in estimating \( \theta = (m, \beta) \).

6.5 Conclusions

In order to carry out the EM algorithm, in the Statistical Model (2) we have assumed that the errors \( \epsilon_i \) have a normal distribution, implying normality for the data \( Y_i \) for \( i = 1, \ldots, n \). Even though we presented a factorization of the complete data likelihood function in terms of the observed and missing, the parameter \( \theta \) in the factorization are not distinct. Thus, maximizing the factors separately will not maximize the likelihood. Thus, the EM algorithm is essential in this case.

The objective of applying the EM algorithm in our problem was also to show that our methodology works in the presence of missing data. We were able to conclude that there was not any significant difference between the results of the EM algorithm and the inverse problem from Section 5.2.4. In both cases, we were able to show that the inability to estimate the parameters is not because of the methodology we used.

7 Conclusions

In this report we develop models that assure our understanding of the transmission dynamics of Vancomycin-Resistant Enterococcus in a hospital unit. Using these models in conjunction with surveillance data we try to estimate some of the epidemiological parameters that are difficult to estimate directly from data and not found in literature. As a result, this report offers new contributions regarding the modeling of the transmission of VRE in a hospital unit. We develop and explore models that have not been described previously by other research studies. The models developed take into consideration specific details regarding the isolation procedure employed in most hospitals.

In section 3, we attempt to investigate estimation of parameters using a parameter selection methodology. This methodology allow us to determine the conditions under which the unknown parameters are estimable and whether the data are sufficient for inverse problem solution. We are able to estimate at least the transmission rate for both units. On the other hand, the resulting model fit to the data suggests that we cannot capture the variability that appears to be present in the data. As a result, in Section 5 we introduce a new mathematical model to account for this. Given the complexity that arises in deriving the sensitivity equations for the discrete VRE model with delay developed in Section 5, we are unable to conduct readily the parameter selection methodology using this more detailed model.

The ODE model in Section 3 considers a population of patients divided as uncolonized, VRE colonized, and VRE colonized in isolation. While this model considers the same per-capita rate of isolation for patients that are admitted with VRE and that are newly colonized during hospital stay, the discrete VRE model with delay takes into account the differences that exist in isolating these patients. Both models take into ac-
Figure 36: Residual analysis for estimating $\theta = (\beta)$ using oncology unit surveillance data (time period: January 6, 2005 to April 7, 2005) via EM algorithm.
count the role that health care workers hand-hygiene compliance has in controlling the transmission of VRE.

The results in Section 5 have raised questions as to the accuracy and regularity of the surveillance data under study. The data seems to be very irregular with errors that do not appear to follow any standard processes. On the other hand, we were able to carry out simulations to suggest that the poor model fit to the data is more likely because of the data and not due to the methodology we used. We showed that with data that looks like the process, we could successfully carry out the inverse problem. The need is apparent for more accurate (regular) surveillance data in order to estimate parameters which are not measurable but on which disease dynamics are highly dependent. Any future efforts should depend heavily upon the availability of new VRE surveillance data (preferably generated from a designed experiment) to support the discrete VRE model with delay introduced in this dissertation. This model accounts for what appears to be essential details and hence it is the most realistic of the models introduced.

8 Acknowledgments

This research was supported in part by the Alfred P. Sloan Foundation Graduate Scholarship (sponsored by the Sloan Foundation and NACME), the Louis Stokes Alliance for Minority Participation (LSAMP) Fellowship (sponsored by NSF/WAESO), and the More Graduate Education at Mountain State Alliance (MGE@MSA) Scholarship.
Figure 38: Residual analysis for estimating $\theta = (m, \beta)$ using oncology unit surveillance data (time period: January 6, 2005 to April 7, 2005) via EM algorithm.
Figure 39: Best fit model solutions to oncology unit surveillance data (time period: January 6, 2005 to April 7, 2005) via EM algorithm, $(\hat{m}, \hat{\beta}) = (0.1506, 0.0000)$.

References


