Stochastic Models

for

Tuberculosis Transmission and Control

Maria Xiridou

Department of Statistical Science

University College London

Thesis submitted for the degree of Doctor of Philosophy

Faculty of Science, University of London

February 2001

Acknowledgements

First of all, I wish to express my gratitude to my supervisor, Valerie Isham, for her help and guidance. Her understanding, encouragement, and support for all the problems that arose during the development of this thesis are greatly appreciated. It has been a privilege and a pleasure working with her on this project.

I would further like to thank Rex Galbraith, Chris Dye, Daryl Daley, and Phil Pollett for their useful suggestions, insights, and comments on various aspects of this thesis. My thanks are also due to Julian Herbert for his advice and the many helpful and enlightening discussions. I also want to thank Paul Northorp and Richard Chandler for providing me with algorithms for random number generators for the simulations. I am grateful to the students and staff in the Department of Statistical Science at University College London for their assistance and advice on statistical and computing issues. Also I am indebted to the State Scholarships Foundation of Greece for their financial support.

There are also a number of other people whose contribution to this thesis, although indirect, is invaluable. First of all, Anastasios Skiadas and Christos Langaris for introducing me into the magical world of mathematics and stochastic processes, respectively. Angeliki Pollali for being the *αγκαλιτσας* always ready to give support, encouragement, and compassion. Also, a big thanks to all my friends and family, scattered around the four corners of the planet, for dripping a few drops of themselves inside me and hence being always with me wherever I go.

A special word of thanks is for my partner, Marc Peters, for his help and support. His advice and assistance contributed a lot to the writing of this thesis and his cookies, kicks, and taps on the shoulder gave me the support and encouragement I needed. *Dank je wel, monk,* for sharing with me the drive to this top and the view ahead.

Finally, my greatest gratitude is to my parents and my sister for their support, trust, faith, and understanding that simply keep me going. I have no words to thank them; I only hope that when they make their *σουμα* I will bring them a deep smile...

στο μπακι, το μακι,
και το ντουλι . . .

Abstract

The natural evolution of tuberculosis in the absence of any medical interventions and its evolution in populations where control measures are implemented, are studied using various mathematical techniques and especially those of stochastic models.

In developed countries the numbers of tuberculosis cases increased and then declined, even before the introduction of effective therapy. Although now curable, tuberculosis remains endemic in developing countries and among infectious diseases it is the leading cause of death worldwide. Serious questions have been raised with respect to the efficacy of the control measures currently available and the reasons for their failure to control the spread of tuberculosis in some areas. This thesis investigates the spread of tuberculosis in the absence and in the presence of medical interventions and addresses questions about the endemicity of the disease and the efficacy of the controls, via stochastic models describing the dynamics of the infection.

In particular, the probability of extinction of the disease, the time until extinction, the size of an individual epidemic, and the distributions of the numbers of infected and infectious individuals are considered. Special attention is given to epidemiological indices, such as prevalence, risk of infection, incidence, and mortality, which are used by public health authorities to assess the severity of an epidemic. Approximating methods, including the use of deterministic models, are investigated and their results are compared with those from numerical simulations of the stochastic models being studied. The effect of chemotherapy in controlling an epidemic is assessed by the percentage reduction in the epidemiological indices for various levels of detection and cure rates. The effect of BCG vaccination is studied separately for various coverage and protective levels.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the 18th century tuberculosis was promoted to the rank of "captain of all these men of death". Major epidemics arose in Europe and North America in the 17th century and then declined throughout the 19th century, even before the introduction of chemotherapy (Bloom & Murray 1992). When effective therapy became available, in the late 1940's, tuberculosis (TB) began to fade from existence in most developed countries (Bloom & Murray 1992). Nevertheless, TB epidemics still remain at tragic levels in many developing countries, despite effective control programs (Dye, Scheele, Dolin, Pathania & Raviglione 1999). In addition, a resurgence of TB has been witnessed in developed countries since 1985 (Enarson & Rouillon 1998). The "captain" still keeps his place as the leading cause of death among infectious diseases worldwide (Bloom & Murray 1992). Various explanations have been proposed for this "rise and fall and rise" of TB, but the causes have not been completely determined.

Are subsequent outbreaks of TB going to follow in developed countries and when? Is the epidemic wave in developing countries on its way up or down? What is the impact of control measures and what can ensure their success? Answers to epidemiological questions like these are hidden within the characteristics and dynamics of *Mycobacterium tuberculosis* (the causative agent of TB) and the current state of tuberculosis infections. Answers to such questions can be extracted if the appropriate techniques are developed. Mathematical modelling is one of the techniques that have substantially contributed to the understanding of observed epidemiological trends and predictions for future trends for other infectious diseases (see, e.g., Bailey 1975, Anderson & May 1991).

The first effective modelling of TB was that of Waaler, Geser & Andersen (1962) and subsequent modellers have mainly been concerned with deterministic and simulation models (see, e.g., ReVelle, Feldmann & Lynn 1969, Blower, McLean, Porco, Small, Hopewell, Sanchez & Moss 1995, Vynnycky 1996, Brewer, Heymann, Colditz, Wilson, Auerbach, Kane & Fineberg 1996, Dye, Garnett, Sleeman & Williams 1998, Aparicio, Capurro & Castillo-Chavez 2001). Stochastic models, in contrast to deterministic ones, allow for statistical fluctuations and, in many respects, the probability element is essential in the study of epidemic phenomena. In this thesis we consider stochastic models that describe the natural evolution of TB and the evolution after the introduction of a control policy. The structure of this thesis is as follows.

Tuberculosis and epidemic modelling:
Chapter 2 discusses the basic concepts and principles of the epidemiology of TB. A brief review of the literature concerned with epidemic modelling is given in Chapter 3.

Models for the natural evolution of TB:
In Chapter 4 we present a simple closed model for the natural evolution of TB; this is a three-class model for a population with constant size. A similar model for a population with varying size is studied in Chapter 5; here we allow for immigration of susceptibles and death (both natural and caused by TB). In Chapter 6 we present a more detailed model (called Zeus) which accounts for some subtle features of TB (such as reinfection of individuals with an old infection), which for simplicity were not included in the previous two models.

Models for the effects of chemotherapy and vaccination:
Model Zeus (Chapter 6) is extended to allow for the fact that some of the TB cases receive treatment and then further extended to allow for the possibility of vaccination of newborns. In Chapter 7 we present the model for treatment, called Clio. By comparing numerical results from models Zeus and Clio (Chapters 6 and 7), we assess the effectiveness of chemotherapy and the levels of detection and cure rates necessary to achieve certain levels of reduction in the severity of TB epidemics. In Chapter 8 we present the model for treatment and vaccination, called Erato[1]. Numerical results are compared

[1] The names of the models Zeus, Clio, and Erato reflect their relationship and the purpose for their development. In Greek mythology, Zeus was the supreme deity, considered as the "father" of all gods. Clio and Erato were two of the nine Muses, daughters of Zeus. Model Zeus (for the natural evolution of TB) was developed with a view to extending it to models that account for medical treatment (such as Clio and Erato), so that model Zeus can be considered as the 'parent' model of Clio and Erato.

with results from Chapter 7 in order to assess the additional effectiveness of the vaccine (over and above that of chemotherapy).

Discussion:

In Chapter 9 we conclude with a summary of the main results presented in this thesis and discuss possible modifications and extensions.

Chapter 2

Tuberculosis

2.1 History

Tuberculosis (TB) has a very long history, since its occurrence has been traced back to 4000 BC, through Egyptian mummies and Stone Age skeletons. Nevertheless, major epidemics did not arise in Europe until the early 1600's and somewhat later in North America. The incidence of TB increased to pandemic proportions in the following centuries and TB was referred to as "phthisis", "consumption", the "white death", and the "white plague".

Even until the second half of 19th century, TB was attributed to causes such as heredity, evil spirits, and demons. In 1882 Robert Koch announced that TB is caused by the bacillus *Mycobacterium tuberculosis.* This discovery contributed to the decline of the incidence of TB, which was witnessed in most developed countries, from at least the beginning of this century, although effective therapy had not been introduced then. Segregation of the infectious people in sanatoria, higher standards of hygiene, higher living standards, and better nutrition are some of the possible explanations for this decline. In the later half of the 20th century, the introduction of chemotherapy accelerated this decline in most industrial countries and it was hoped this would be the end of TB.

However, celebrations proved to be premature. In many developing countries, there has been virtually no decline in the incidence of TB, which remains at tragic levels. And even in developed countries there has been a resurgence of TB since 1985, attributed mainly to the rise of the HIV/AIDS epidemic, the emergence of multidrug-resistant TB, immigration, increased poverty, homelessness, unsanitary living conditions, poor nutrition, and substance misuse. In the rise of the new millennium, TB remains a problem worldwide: the World Health Organisation (WHO) estimates that about one third of the world population is infected with TB and approximately 2–3 million people die of TB annually.

References: Blower et al. (1995), Cohen & Durham (1995), Dye et al. (1999), Kanai (1990), LaScolea & Rangoonwala (1996), May (1995), Murray, Styblo & Rouillon (1993).

2.2 Diagnosis

The diagnostic tools used for the detection of tuberculous infection and disease also provide a means for the classification of those who have been infected, according to the degree of infectivity, stage of the disease, and site of the infection. The most important diagnostic tools are the following:
Tuberculin Skin Test (intradermal injection of tuberculin that causes an induration in 48–72 hours); it is neither 100% sensitive (which means that there are false "negative" reactions) nor 100% specific (false "positive" reactions). It does not distinguish between recent and remote infection or between infection and disease. False "positive" reactions may be caused by the presence of other mycobacteria and former BCG vaccination.
Chest Radiography (x-ray); it is high in sensitivity, but low in specificity. It needs experienced x-ray interpreters, because several pulmonary diseases may have radiographic abnormalities similar to those of TB and pulmonary TB may result in almost any kind of radiographic abnormality. Also, most persons with severe immunosuppression present with atypical radiographic findings.
Sputum Microscopy; this is the most important tool to detect highly infectious TB cases, because the patients whose sputum contains sufficient bacilli to be detected by microscopy are the most infectious. These are referred to as smear-positive and they make up approximately half of the TB cases. However, smear examination does not exclude TB infection or infectiousness and a positive test may be caused by organisms other than *M. tuberculosis.*
Sputum Culture; cultures take 4–6 weeks to give results, but they allow the identification of the organisms and drug susceptibility tests to be done. Patients who are smear-negative and culture-positive are 9/10th less infectious than those who are smear-positive.

References: American Thoracic Society (1990), Cohen, Harriman & Madsen (1995), De Cock, Binkin, Zuber, Tappero & Castro (1996), LaScolea & Rangoonwala (1996), Murray et al. (1993), Pan American Health Organization (1986), Styblo (1991).

2.3 Transmission, infection, and development of disease

Tubercle bacilli include several members of the genus *Mycobacterium*, including *M. tuberculosis, M. bovis, M. africanum,* and *M. microti.* TB is most commonly caused by *M. tuberculosis.* The second most frequent source of infection is raw milk containing *M. bovis* from cows infected with this bacillus, which causes a disease clinically similar to TB. Other means of infection are very rare. The most common site of TB is pulmonary, but almost any organ can be infected; these cases are referred to as extra-pulmonary TB, in contrast to pulmonary TB.

The most important route of transmission of TB is through inhalation of infected droplet nuclei. These nuclei are created by a person suffering from infectious pulmonary TB through forced exhalations, such as coughing, sneezing, yelling, singing, and loud talking. The very smallest of them may remain airborne for several hours (so that a room may remain infectious for a while, even in the absence of the infectious person) and when inhaled, infection may become established.

There are several factors that influence whether or not tuberculous infection occurs relating to the host (nutritional status, immune competence, presence of other diseases), the organism (virulence of strain, concentrations of *M. tuberculosis* in droplet nuclei, size and number of aerosolised droplets), the environment (crowding, unclean living conditions, fresh air ventilation), the contact with the source patient (duration and closeness), and the source patient (site of TB, positive or negative sputum smear, TB medication status). For example, smear-negative patients are far less infectious than smear-positive and non-pulmonary TB patients are virtually non-infectious.

Apart from all these factors, the host's defence system is the ultimate determinant for the establishment of infection and its further development. When a person inhales droplet nuclei containing tubercle bacilli, only the very smallest of them can penetrate into the respiratory system and implant on the alveolar surface. If that happens, alveolar macrophages (one of the most critical kind of cells in the human defence system against pathogens) ingest the tubercle bacilli and can kill or remove them, in which case infection does not occur. However, if the microbiocidal capacity of macrophages is exceeded (depending on the number and virulence of the tubercle bacilli and characteristics of the alveolar macrophages) then the surviving organisms will multiply and establish tuberculous infection.

When the infection occurs, the immunological defences are stimulated, but they only become effective after 3–8 weeks. In the mean time, the host has no specific defence against *M. tuberculosis*; the bacilli proliferate and, transported within the macrophages, they enter into the bloodstream; then they can be seeded throughout the body, creating potential sites for extra-pulmonary tuberculosis development.

The stimulation of the defence system initiates a series of responses which result in the formation of granulomas, a type of histological pattern, that contain any viable organisms. The tubercle bacilli can survive within these granulomas indefinitely, but they can not proliferate.

This favourable result occurs at the expense of local host tissue; the granuloma formation results in tissue necrosis, fibrosis, encapsulation, and scar formation. Caseous sites of necrotic tissue (so named because of their "cheesy" consistency) are produced. At that point, the host is asymptomatic and the infection is latent. The tuberculin skin test is usually positive and a chest x-ray may show some abnormalities (for example, the granulomas or the caseous sites). For the majority of those infected, there is no further development; they never develop disease and are immune to tubercle bacilli, usually for life.

The crucial point is whether or not caseous necrosis undergoes liquefaction, because that will determine whether or not the infection develops to disease. If liquefaction does occur, the liquefied material is expelled and a cavity forms. Caseous *M. tuberculosis* becomes widespread and the tubercle bacilli multiply. Patients with cavitary tuberculosis typically exhibit coughing and systematic symptoms and the infectiousness of the host is increased. The relation between the stage of the disease and the extent of infectiousness is still not clear. Some pointers for assessing the infectiousness of a TB patient are the presence or absence of cavities on x-ray, radiological extent of the disease, bacteriological status, and cough frequency.

Risk of developing disease (after infection). The risk varies with age, with a peak at the very young and the very old ages; it is higher for men than women and higher for non-whites than whites. It is enhanced by immunosuppression, malnutrition, alcoholism, etc. It is estimated that approximately 80–90% of those infected will never develop disease; of the remaining, half will develop disease in the first few years (primary TB) and half much later (secondary TB) even several decades after infection.

Reactivation vs. Reinfection. There is a strong controversy about secondary TB. The first theory maintains that tubercle bacilli cannot survive forever, immunity wanes, and inhalation of tubercle bacilli by persons who have been infected, say five years previously, increases the risk of development of TB after this reinfection. According to a second theory, tubercle bacilli can remain alive within their host during his/her lifetime and at any time they may start multiplying and cause the development of disease; immunity can remain intact, protecting the host against reinfection, usually for life. The truth probably lies between the two theories. In areas with low risk of infection, secondary TB is mostly due to reactivation of an old infection, while in areas with high risk of infection it is mainly due to reinfection.

TB and HIV infection. When a person is infected with both HIV and TB, the development of TB follows the same pattern, but more rapidly. HIV infection increases the risk of developing disease after infection; between 5% and 10% of persons co-infected with *M. tuberculosis* and HIV will develop TB each year, compared with less than 0.2% of persons infected with *M. tuberculosis* but not HIV. Those with both TB and HIV have shorter incubation and infectious periods, higher mortality rate, they are more susceptible to reinfection and reinfection with multidrug-resistant TB. Also, they are more likely to develop extra-pulmonary TB, than TB patients who are HIV negative.

The fate of TB patients without treatment. After the onset of disease, about 10–15% of smear negative patients will become smear positive. Approximately 10% of all TB patients 0–14 years old and 50% of all patients older than 15 are smear positive. A patient remains infectious for two years, on the average. 50–60% of patients (60–70% for smear-positive, 40–50% for smear-negative) will die within the first five years. The majority of those who remain alive eight years after diagnosis have quiescent TB (they are naturally cured, but they may relapse later) and the remaining are chronic excretors of tubercle bacilli (sporadically infectious and ill). There is a 30% chance of spontaneous cure, but among those who recover 4.4% relapse each year for the first five years and 1.6% annually for the second five years.

References: Clancy (1990), De Cock et al. (1996), Dolin, Raviglione & Kochi (1994), Grzybowski & Enarson (1978), Kanai (1990), LaScolea & Rangoonwala (1996), Murray et al. (1993), Pan American Health Organization (1986), Shekleton (1995), Styblo (1991).

2.4 Symptoms

In most cases, the clinical presentation of TB involves only a gradual development of insidious and vague symptoms. The most common symptoms are cough, hemoptysis, sputum production, fatigue, anorexia, chilly sensations, night sweats, low grade fever, chest pain, and dyspnoea. These symptoms may or may not be present and with variable severity. There are patients who are truly asymptomatic. Others may have only nonspecific symptoms (such as anorexia, fatigue, weight loss) which are often attributed to overwork or emotional stress. Also, the symptoms may be attributed to other diseases (such as influenza, pneumonia, asthma, lung cancer). Actually, several studies have reported that the percentage of missed diagnoses is about 40–50% at the time of hospital admission and about 5% of cases are only discovered by autopsy. The presentation of TB in HIV patients is usually atypical, with symptoms such as malaise and weight loss, which are also seen as part of the HIV infection. In extra-pulmonary TB the symptoms are related to the organ system affected and there may be nonspecific symptoms, as well.

References: American Thoracic Society (1990), Cohen et al. (1995), Crofton, Horne & Miller (1992), LaScolea & Rangoonwala (1996), Stead & Dutt (1988).

2.5 Treatment

The first anti-tubercular drug, Streptomycin, was discovered in 1944, by Selman Waksman. Since then, a number of anti-TB drugs have been developed, among which Isoniazid, Rifampin, Pyrazinamide, and Ethambutol are the most frequently used. Their main effect is that they reduce bacteria counts, cough frequency, and excretion of tubercle bacilli, thus rendering the patient much less infectious. Therefore, the benefits of chemotherapy are not only direct (for the patient treated), but also indirect (reduced transmission of the disease).

Indeed, high cure rates have been achieved with regimens consisting of the drugs aforementioned. Under ideal conditions of 100% compliance, 80–90% of smear-positive cases will have converted to smear-negative after two months of treatment and the remaining in the following two months. Even among patients who discontinue chemotherapy at two months, only 40% will be smear-positive (or dead) after two years. The indirect benefits of chemotherapy were clearly depicted by the rapid decline of the incidence of TB in most developed countries between 1950 and 1985.

However, despite these potential high cure rates, most chemotherapy programs in developing countries failed to achieve the WHO target of 85% cure rate. The primary reason for this is failure to ensure patients' compliance (due to financial constraints, for instance). Nevertheless, cure rates of 80–90% have been achieved (for instance in Malawi, Mozambique, Tanzania, and China) with intensive control programs, that provided for close supervision of treatment, bacteriological examinations, etc. It is, therefore, organisational/administrative, as well as technical, factors that are the most important determinants for the success of a control program.

Multidrug-resistant TB. Soon after the introduction of chemotherapy, it was discovered that drug-resistant organisms would begin to appear in the sputum by the fourth week of therapy. Even from the 1960's there are studies reporting that the frequency of Isoniazid-resistant bacilli was 1.5% in Canada in 1964 and 13.6% in Taiwan in 1963 (and rose to 27% by 1968). Combination therapy with at least two drugs was then introduced, as the "miracle solution", but again TB proved to be an ingenious opponent; strains of *M. tuberculosis* were becoming resistant to multiple drugs, as a result of a sequence of strains resistant to individual drugs. For instance, a strain resistant to drug A, say, infects a person who is then treated with drugs A and B. Basically, this is monotherapy, likely to cause resistance to drug B. The result is resistance to both drugs.

In general, the possible ways of drug-resistance development are: (a) a drug-sensitive patient develops drug-resistance during treatment with regimens that are poorly conceived or poorly complied with, (b) drug-resistant patients infect susceptibles (not previously infected) who become drug-resistant, (c) there is exogenous reinfection with a new multidrug-resistant strain of TB, during or after therapy for drug-sensitive TB.

Several studies indicate that there is an increasing trend in the prevalence of multidrug-resistant TB; for instance in the United States it increased from 2% to 9% in the last 30 years. Moreover, there is a concern that the HIV epidemic is likely to increase the risk of developing drug resistance.

Currently, the basic regimens are for six months and include four drugs. For TB and HIV infected patients, extra-pulmonary TB patients, the elderly, and infants the therapy is often extended to at least nine months. If multidrug-resistant TB is suspected or proved the regimen includes at least five drugs.

References: Chan & Yew (1998), China Tuberculosis Control Collaboration (1996), Clancy (1990), Dye et al. (1998), Grzybowski & Enarson (1978), Kanai (1990), Kochi, Vareldzis & Styblo (1993), LaScolea & Rangoonwala (1996), Murray et al. (1993), Murray, DeJonghe, Chum, Nyangulu, Salomao & Styblo (1991), Pan American Health Organization (1986), Small, Shafer, Hopewell, Singh, Murphy, Desmond, Sierra & Schoolnik (1993), Stead & Dutt (1988), World Health Organization (1993).

2.6 The BCG vaccine and preventive therapy

The purpose of the BCG (bacille Calmette-Guérin) vaccine is to prevent the development of disease after infection. Its protection lasts for 10–15 years and it is not clear, yet, whether revaccination has any significant effect. With pulmonary TB, the efficacy of BCG varies from 0% to 80%. Several factors account for this large variation and for particular ages or countries the variation is smaller (for instance 40–70% for children 0–14 years old if given at birth and 20–30% in India). It is suggested that it should be given at birth or as early in age as possible. There are still questions about vaccination at older ages, as well as about people who are vaccinated and then infected with HIV. But for advanced HIV and AIDS patients it is contra-indicated, because of the high risk of disseminated BCG infection. In most developed countries BCG is not routinely recommended, except in specific situations, due to the variation of its efficacy and its invalidation of the tuberculin skin test.

Preventive therapy (also referred as chemoprophylaxis) is used in preventing the establishment of infection and the development of infection to disease. In general, mass chemoprophylaxis is not recommended, but only indicated for specific situations (such as suckling babies of infectious mothers, close contacts of TB patients, HIV/AIDS patients, and newly infected persons who have not developed disease). Several other factors may determine the applicability of chemoprophylaxis (for instance, age, degree of exposure to TB, presence of other infections and/or diseases) and still there are many questions regarding its applicability and effectiveness. Reductions by 60–90% of the risk to develop disease have been reported and the protective effect may last for several years. Chemoprophylaxis is also effective in reducing TB incidence in HIV-infected individuals, especially those with a positive tuberculin skin test.

References: De Cock et al. (1996), Horne (1990), LaScolea & Rangoonwala (1996), Murray et al. (1993), Pan American Health Organization (1986), Smith & Fine (1998), Styblo (1991).

2.7 Current epidemiology of TB

TB Incidence (number of new TB cases reported during a year per 100,000 general population): The global incidence of TB remains at tragic levels, as it is estimated that about 7.5–8 million new TB cases occur per year worldwide. Particularly in developing countries, about 120–260 cases are notified per 100,000 population annually (Dye et al. 1999, Murray et al. 1993).

Annual Risk of TB infection (percentage of population that is infected or reinfected during a year): The annual risk of infection in developing countries is between 0.5% and 2.5%; in the absence of HIV, it is stable or decreasing by 1–2% per year (a rate slightly less than the population growth in these countries). HIV infection is believed to increase the risk of infection. In Eastern Europe and the former USSR the risk was about 0.05–0.35% in 1994 and in developed countries less than 0.1%. It is estimated that an undiagnosed and untreated TB patient infects about 10–14 susceptibles each year and is infectious for almost two years; a smear-positive case infects 2–5 persons (2–3 in developed countries and 4–5 in developing countries) before his/her detection. Again, all the numbers above may vary depending on age, sex, and other factors (Bloom & Murray 1992, Murray et al. 1993, Murray et al. 1991, Raviglione, Rieder, Styblo, Khomenko, Esteves & Kochi 1994, Styblo 1991).

TB prevalence (number of registered TB cases per 100,000 general population): In 1997 the global prevalence was around 277 cases per 100,000 population. The percentage of smear-positive cases among all TB cases is about 10% for those 0–14 years old and about 50% for those older than 15. Smear-positive TB is rare in children; about 80% of smear-positive cases occur between the ages 15–59 (Dye et al. 1999, Murray et al. 1993, Styblo 1991).

Prevalence of TB infection (proportion of infected individuals in a given popula-

tion): In 1997 approximately 32% of the world's population was infected with *M. tuberculosis.* In developed countries 20% of those infected were less than 50 years old, while in developing countries the same percentage is about 75% (Cohen 1995, Dye et al. 1999).

Mortality Rate (number of deaths from TB in a given population): Despite the implementation of chemotherapy programs, TB continues to exact a terrible toll. In 1997 an estimated 1.87 million people died of TB. About 10–20% of these deaths were in children. The developing world bears the heaviest burden accounting for 98% of TB deaths (Cohen 1995, Dye et al. 1999, Murray et al. 1993).

TB Fatality (Lethality) Rate (number of TB deaths per 100 cases of TB): Without treatment, approximately 50–60% of TB patients will die (60–70% of smear-positive and 40–50% of all other TB cases). In 1997 the global case fatality rate was estimated to be 23%, but more than 50% in some African countries. In the pre-chemotherapy era, the average time from diagnosis to death (over all age groups) was 13–14 months, ranging from 18 months for those 15–44 years old and down to 3 months for those older than 65. With good chemotherapy programs case fatality can be reduced to 10–15% (China Tuberculosis Control Collaboration 1996, Dye et al. 1999, Grzybowski & Enarson 1978, Murray et al. 1993, Styblo 1991).

Chapter 3

Review of epidemic modelling

3.1 The roots of mathematical modelling of epidemics

> "We share the world with the smallest living things: bacteria and viruses... A hundred million virus particles could live very comfortably in an area the size of the period at the end of this sentence. This planet in many respects, is ruled not by the macrobes, but by the microbes — which have a far greater power to kill." (LaScolea & Rangoonwala 1996).

These powerful "planet-mates" of ours cause the major epidemics that have scourged our planet for thousands of years and accounted for tremendous numbers of human lives lost. It is therefore not surprising that these formidable epidemics attracted the interest of many scholars and scientists of various disciplines, from the very early years.

Records of epidemics date back to the ancient Greeks (e.g. the "Epidemics" by Hippocrates, 458–377 BC) and medical statistics to the 17th Century (e.g. J. Grant, 1620–1674 and W. Rethy 1623–1687) (Bailey 1975, Anderson & May 1991). In 1760 Daniel Bernoulli used a mathematical method to evaluate the effectiveness of variolation (a technique of inoculation) against smallpox, with a view to influencing public health policy (Bailey 1975, Dietz & Schenzle 1985). More mathematical studies were developed in the 19th century that used patterns of cases to examine the spread of diseases (e.g. J. Snow, 1855 and W. Budd, 1873) and investigated the use of curve-fitting to data on various diseases (e.g. W. Farr, 1840; J. Brownlee, 1906) (Bailey 1975).

Progress in biomedical sciences lead to the modern scientific achievements of the 20th century in this field and enhanced the flourishing of medical statistics as well as of the mathematical theory of epidemics. From the beginning of this century there was a definite upsurge in epidemic modelling. Hamer (1906) and Ross (1908) were the first who formulated mathematically certain hypotheses about the mechanisms of infectious diseases. Hamer, in particular, was the first who considered that the course of an epidemic must depend on the number of susceptibles and infectives. Soper (1929) worked on the ideas of Hamer and Ross and deduced important results about the periodicity of some epidemics.

3.2 The Kermack-McKendrick and Reed-Frost models

McKendrick (1926) published the first purely stochastic model of epidemics. He assumed that an individual is infectious from the moment he receives infection until he recovers or dies or is isolated and that the probability of one new infection occurring in a short interval is proportional to the length of the interval and the numbers of susceptible and infectious individuals.

Kermack & McKendrick (1927) developed deterministic models, whose structure forms the basis of the *general stochastic model*, one of the now classical epidemic models. They considered a population divided into three classes, those who are susceptible to the disease, those who are infectious, and those who have recovered or been removed (so that they are considered as non-infectious and immune). The population is subject to homogeneous mixing, which means that the contact rate is the same for each individual (and thus, at any instant each susceptible has the same probability of being infected and each infective has the same chance of infecting any susceptible).

The number of new infections in a very short time interval is proportional to the length of the interval and the current numbers of susceptible and infectious individuals, $X(t)$ and $Y(t)$, respectively. The latent period is assumed negligible and the infectious period is exponentially distributed. They considered both variable and constant infection and removal rates. In the case of constant rates, the probability of one new infection occurring in the interval $[t, t+dt]$ is $\beta X(t)Y(t)dt+o(dt)$ and the probability of a recovery or a removal in the same interval is $\gamma Y(t)dt + o(dt)$.

One of the most important results deduced in this paper is the celebrated Threshold Theorem. According to this theorem, no epidemic can occur unless the density of susceptibles exceeds the threshold value ρ/n (where n is the total population size and

$\rho = \gamma/\beta$). A similar result is deduced for the general case entailing variable infectivity and recovery rates.

Around 1928 L. J. Reed and W. H. Frost developed a different kind of probabilistic model, which was illustrated for teaching purposes, but was not published until much later; also, Greenwood (1931) developed a slightly different variation of this model (see, e.g., Bailey 1975, Chapter 14). The Reed-Frost model (also referred as the *chain-binomial model*) assumes that the latent and the incubation periods can be regarded as constant and the infectious period is reduced to a single point. Starting with one infective (or with several simultaneously infective persons) the process will continue in a series of stages, separated by intervals equal to the latent period. At each stage, the susceptibles will yield a number of new cases at the next stage, which under certain conditions, will be distributed in a binomial series, depending on the numbers of susceptibles and infectives at the previous stage. We thus have a chain of binomial distributions.

3.3 Models for macroparasitic infections and other population processes

The models described in the previous section have the common characteristic that the population is divided into several classes (such as susceptibles, infectives, immune and/or removed) but there is no distinction as to the severity of the infection (i.e. the abundance of the parasite within the host); an individual simply either has or does not have the disease. This formulation is generally adequate for microparasites (viruses, bacteria, and protozoa) but not for macroparasites (helminths and arthropods). The former multiply within the host at high rates, but the latter generally do not have direct reproduction within the host and they accumulate only via reinfection. The factors characterising the development of the infection within a host (such as transmissibility of the infection, presence and severity of symptoms, immunity of the host) depend on the number of parasites harboured in the host. Therefore, models for macroparasitic infections take into account the distribution of parasites among hosts (see, e.g., Bailey 1975, Dietz & Schenzle 1985, Anderson & May 1991).

Another characteristic of the models for microparasitic infections is that after a successful contact between an infective and a susceptible, the number of susceptibles is decreased by one, and that of infectives is increased by one. This is not always the case for other population processes, such as the predator-prey processes, where after a "successful" contact the number of prey decreases by one, but the number of predators remains the same (see, e.g., Bailey 1964, Hitchcock 1986).

M. tuberculosis is a microparasite and hence the following review concentrates on the literature for microparasitic infections.

3.4 Development of epidemic modelling

During the last five decades variants of the Kermack-McKendrick and Reed-Frost models have been developed and many new results have been published. Most of the relevant work up to 1974 is covered in Bailey (1975). Here we can find a detailed description of the most important epidemic models, including discussion of probabilistic analyses of these models, simulation studies, estimation of parameters, and applications to household data. Recent reviews of the relevant literature and theory have also been published (see, e.g., Lefèvre 1990, Dietz & Schenzle 1985, Isham 1993, Mollison, Isham & Grenfell 1994, Daley & Gani 1999, Renshaw 1993, Anderson & May 1991).

The Kermack-McKendrick model is concerned with diseases of the SIR type, which entails a closed, homogeneously mixing population divided into three classes: susceptible to the infection (S), infected and infectious (I), and removed, recovered, or dead (R). Other formats are:

- the SI model: the population is divided into two classes, susceptible (S) and infectives (I), and the only possible transitions are infections (S → I)
- the SIS model: again there are only susceptibles and infectives, but after infection immunity may wane, so that an infective may become susceptible. Hence, the possible transitions are infections (S → I) and loss of immunity (I → S).
- the SIRS model: the population is divided into three classes, susceptibles (S), infectives (I), and removed or recovered (R). The possible transitions are infections (S → I), removals or recovery (I → R) and loss of immunity after recovery (R → S).

Several other types of epidemics have been examined and described by models whose structure is based on that of the Kermack-McKendrick model. Some of these are the following:

Recurrent epidemics; Soper (1929) was the first to investigate the periodicity of recurrent outbreaks of measles. Other researchers followed and Bartlett, in a series of papers and his books (see, e.g., Bartlett 1956, 1957, 1960*a*, 1960*b*), made a considerable contribution by the formulation and study of stochastic models, primarily concerned with measles and smallpox.

Carrier models; these involve the presence of carriers, i.e. individuals who are infected and infectious but appear outwardly healthy (see, e.g., Downton 1967).

Competition between epidemics; here there are two types of infectious agents and hence two types of infectious individuals, who may recover or be removed, forming two classes of "removals". Susceptibles can be infected by either type of infection (see, e.g., Kendall & Saunders 1983).

Multistate models; the underlying mechanisms of some diseases are too complicated to be described with three or four classes of individuals. The available control measures for these diseases resulted in models involving a large number of states. Most of these models are deterministic or simulation models (e.g. Blower, Small & Hopewell 1996, Brewer et al. 1996).

Most of these types of models can be further classified according to whether or not they entail immigration into and/or migration out of the population; some models account for recruitment of susceptibles and/or infectives, deaths due to natural causes and due to the disease, and the total population size may be constant or variable (see, e.g., Bartlett 1960*b*, Lefèvre 1990, Jacquez & O'Neill 1991, Isham 1993).

Advances made in stochastic processes in the 1940's enhanced the use of more advanced (mathematically) techniques and formulations: branching process formulations (e.g. Bartoszyński 1967), coupling methods (e.g. Ball 1995), point processes (e.g. Lefèvre 1990), modelling on random graphs (e.g. Barbour & Mollison 1990), and many others. There is also a growing interest in the development of methods of statistical inference for the analysis of infectious disease data (e.g. Becker 1989).

Approximating stochastic systems and asymptotic approximations have also been used and proved a very helpful tool in the study of stochastic models. For instance, Tan & Hsu (1989) developed an approximating system by assuming the number of susceptibles to be a deterministic function of time (under the assumption that this number is always large) and keeping all the other probabilistic elements of the model. Kurtz (1970, 1971, 1981) proved that certain stochastic systems can be approximated by Gaussian diffusion processes, if the initial numbers of susceptibles and infectives are large.

The variability of parameters and stages of the models is a problem that has received a lot of attention and prompted the development of more variants of the Kermack-McKendrick and the Reed-Frost models (see, e.g., Lefèvre 1990, Mollison 1995, Mollison et al. 1994). Some of the problems considered are: the existence of several types of infectives (with different distributions for the period of infectiousness), the variability in susceptibility and/or infectiousness, variable transition rates, and several stages of the infectious and/or incubation period. One of the basic assumptions of the first stochastic models of epidemics was that of the homogeneous mixing, i.e. that at any given instant, any susceptible has the same probability of being infected by each infective. In reality, this is not always the case, both because of differences between individuals (e.g. differences in susceptibility and behavioural differences) and because of heterogeneity of mixing (due to the geographical distribution of cases, for instance). Several attempts have been made in order to deal with this problem, for example the development of spatial and multi-population models (see, e.g., Bailey 1975, Lefèvre 1990).

3.5 Important statistics in epidemic theory

One of the important statistics in epidemic theory is the final size of an epidemic, i.e. the total number of individuals infected during the epidemic, not counting the initial infectives (or equivalently, for closed populations, the total number of susceptibles uninfected at the end of the epidemic). Both the exact distribution and the asymptotic behaviour of the total size have been investigated. For instance, Kermack & McKendrick's (1927) deterministic treatment deduced the Threshold Theorem and from that the approximate result that when the density of susceptibles exceeds the threshold value, then the size of the epidemic will be twice the excess, and thus, at the end of the epidemic, the density of susceptibles will be just as far below the threshold density, as initially it was above it. Fundamental results were published later (see, e.g., Whittle 1955, Kendall 1956) about the J-shaped and U-shaped form of the probability distribution of the ultimate number of individuals infected during the epidemic, together with recursive techniques to obtain this distribution. Ridler-Rowe (1967) worked on a model involving immigration of new susceptibles and new infectives and deduced results for the probability of extinction of the infectives (and hence of the infection) and the mean time until extinction. See also

Daniels (1967), Sellke (1983), Abramov (1994), Ball (1983).

Another statistic of interest is the maximum number of infectives present at any time during an epidemic. Kendall (1956) in his deterministic treatment of the Kermack-McKendrick model derived exact solutions of the equations for the ultimate number of infectives and removals and from these deduced the maximum number of infectives, over the course of the epidemic. Asymptotic results for the distribution of this maximum have also been obtained (see, e.g., Daniels 1974, Abramov 1994).

Mention should also be made of the basic reproduction ratio (usually denoted by $\mathcal{R}_0$). This is effectively defined as the number of cases generated by one infective over the period of infectiousness, when this infective is introduced into a large population of susceptibles (see, e.g., Diekmann & Heesterbeek 2000, Jacquez & O'Neill 1991). Formally $\mathcal{R}_0$ can be calculated from the following definition by Diekmann & Heesterbeek (2000) (see also Heesterbeek 1992, Diekmann, Heesterbeek & Metz 1990):

Definition 3.1 *Assume that the infected individuals could be in a finite (say k) number of different states and the number of individuals in each state is $x_1, x_2, \ldots, x_k$. State transitions occur according to a rate matrix S and death occurs according to a diagonal rate matrix D. Let T be the matrix whose (i, j) element is the rate at which an infected individual with state j produces secondary cases with state i. Then the vector $\mathbf{x} = (x_1, \ldots, x_k)$ satisfies the differential equation*

$$\frac{d\mathbf{x}}{dt} = (\mathsf{T} + \mathsf{S} - \mathsf{D})\mathbf{x},$$

and $\mathcal{R}_0$ is the dominant eigenvalue of the matrix $\mathsf{K} = -\mathsf{T}(\mathsf{S} - \mathsf{D})^{-1}$ (dominant in the sense that $\mathcal{R}_0 \geq |\lambda|$ for any other eigenvalue λ of K).

The progress of the epidemic and the endemic steady state of the system (and whether this state is achieved) depend on $\mathcal{R}_0$ and other parameters. For instance, for the general stochastic model described in Section 3.2, with infection and removal rates β and γ, respectively, the basic reproduction ratio is $\mathcal{R}_0 = n/\rho$, where n is the total population size and $\rho = \gamma/\beta$. If $\mathcal{R}_0$ is greater than one, then with a deterministic model there will always be an epidemic (i.e. a major outbreak), while with a stochastic model the probability of an epidemic is $1 - (1/\mathcal{R}_0)^{\alpha}$, where α is the initial number of infectives (Kermack & McKendrick 1927, Whittle 1955). Nevertheless, things become more complicated in other complex models, for example, when the population is divided into subpopulations or when the period of infectiousness comprises several different stages (see, e.g., Mollison et al. 1994).

3.6 Mathematical modelling of tuberculosis

The first model for TB was developed by Waaler, Geser, and Andersen in 1962. This is a deterministic three-class model involving susceptibles, infectious, and non-infectious cases subject to homogeneous mixing, recruitment of susceptibles, natural death, and excess death caused by TB. Non-infectious cases may become infectious (as the disease progresses) and infectious cases may heal and become non-infectious. The number of new infections occurring in a short interval is assumed to be proportional only to the length of the interval and the number of infectious cases. Waaler and Piot carried out extensive and remarkable research on various epidemiological aspects of TB and the effectiveness and cost-benefits of control measures mainly using simulation models (see, e.g., Waaler & Piot 1969, Waaler & Piot 1970).

Another breakthrough in TB modelling came through the models developed by ReVelle, Lynn, and Feldmann in the late 1960's. The assumption that the rate of spread of TB infection depends on the numbers of both the susceptible and the infectious individuals was introduced and the effects of BCG vaccination, chemoprophylaxis, and chemotherapy were incorporated in a deterministic multi-state model (see, e.g., ReVelle et al. 1969).

The structure of this model and that of Waaler, Geser, and Andersen influenced the development of subsequent models for TB, although several other issues have been considered in the recent literature. For instance, the variability of epidemiological factors (such as the infection, recovery, and relapse rates) by age has been considered by several authors (see, e.g., Rusu 1973*a*, Schulzer, Radhamani, Grzybowski, Mak & Fitzgerald 1994, Vynnycky 1996, Vynnycky & Fine 1997, Dye et al. 1998); the technique employed in these cases is that of dividing the population into specific age-groups as well as clinical states. Other authors have considered the variability of the epidemiological factors in time (e.g. Trefny & Hejdova 1982) and the effect of clustering (i.e. populations divided into clusters of close contacts, such as the home, the work place, and/or the school — see, e.g., Aparicio et al. (2001)).

The effect of the available control measurements (BCG vaccination, chemoprophylaxis, and chemotherapy) has attracted the interest of many authors (see, e.g., Vynnycky 1996, Vynnycky & Fine 1997, Blower et al. 1996, Castillo-Chavez & Feng 1997, Dye et al. 1998, Goh & Fam 1981, Azuma 1975, Joesoef, Remington & Tjiptoherijanto 1989, Chorba & Sanders 1971, Trefny & Hejdova 1982). Also, the effect of multidrug-resistance (e.g. Blower et al. 1996, Castillo-Chavez & Feng 1997) and of the HIV infection (e.g. Schulzer et al. 1994, Brewer et al. 1996) have been examined.

All of these models are deterministic, hybrid (mixed deterministic-stochastic), or simulation models (e.g. Blower et al. 1996, Schulzer et al. 1994, Chorba & Sanders 1971, Brewer et al. 1996). Operational models have been developed, especially for the study of the cost-effectiveness of the various control measures (see, e.g., Rusu 1973*b*, Joesoef et al. 1989, Chorba & Sanders 1971). Also, statistical models have been studied for parameter estimation (see, e.g., Vynnycky 1996, Vynnycky & Fine 1997, Schulzer, Enarson, Grzybowski, Hong, Kim & Lin 1987).

Finally, mention should also be made of models for the spread of bovine TB (caused by *M. bovis*) in animal populations (see, e.g., Barlow 1993, Bentil & Murray 1993).

Chapter 4

The first model: a simple closed model for TB

4.1 Introduction

We start modelling TB by considering first only the natural evolution of the disease. Infectivity and immunity are the most important determinants for the spread of TB within a population. Therefore, as a first approach, we consider a fixed population of size n subject to homogeneous mixing. The assumption of homogeneous mixing implies that the contact rate is the same for each individual, so that at any instant each susceptible has the same probability of being infected and each infective has the same chance of infecting any susceptible. The population is divided into three classes:

(a) those who are susceptible to TB (they are neither infectious nor immune)
(b) those who have developed clinical disease and are infectious
(c) those who have been infected, but are not clinically diseased (either because the infection is still latent, or because they developed TB in the past and recovered spontaneously); they are non-infectious and immune (temporarily or permanently).

Individuals in each of the above classes will be referred to as susceptibles, infectious cases (or infectives), and inactive cases, respectively (for simplicity of terminology, sometimes we will abuse the adjectives susceptible and inactive as nouns). The sizes of these classes at time t will be denoted by $X(t)$, $Y(t)$, $Z(t)$, respectively.

As was explained in Chapter 2, the duration of the latent period for TB can be very short or very long — less than one year and up to several decades or even lifelong

(Shekleton 1995). Therefore we assume that it is possible to have transitions from the susceptible to both the infectious and the inactive classes. The rates of these transitions are proportional to the number of susceptibles and the number of infectives. The possible transitions and their rates are illustrated in Figure 4.1.

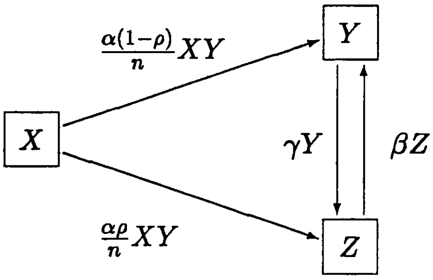


Figure 4.1: The first model for the spread of tuberculosis

In particular, we assume that c is the rate at which each individual in the population contacts others so that $cX(t)/n$ is the rate at which each infective in the population contacts susceptibles, assuming homogeneous mixing. Now, if q is the probability of transmission per contact between a susceptible and an infective, then the probability that an infective will infect a susceptible in the interval $[t, t + dt]$ is $qcX(t)dt/n + o(dt)$. If $\alpha = qc$, then the probability of one new infection occurring in the interval $[t, t + dt]$ is $\alpha X(t)Y(t)dt/n + o(dt)$. Let $1 - \rho$ denote the probability that the infected will develop disease soon (and the latent period is negligible). Then the probability of a transition from the susceptible to the infectious class in the interval $[t, t + dt]$ is $(1 - \rho)\alpha X(t)Y(t)dt/n + o(dt)$ while the probability of a transition from the susceptible to the inactive class in the same interval is $\rho\alpha X(t)Y(t)dt/n + o(dt)$.

Those who are inactive may develop disease at some point and become infectious either because of reactivation of an old infection or because of relapse after recovery. The reactivation and relapse rates are not necessarily exactly the same, but as a first approximation we will assume that they are equal; let β denote this common rate. Then transitions from the inactive to the infectious class occur at a rate βZ. An infectious individual may recover spontaneously and become inactive. The per capita recovery rate is γ, which means that transitions from the infectious to the inactive class occur at a rate γY. Therefore the transitions from state (X, Y, Z) that can occur in the interval

$[t, t+dt]$ and the respective probabilities are:

$$\begin{aligned}
&\mathrm{P}[(X,Y,Z) \to (X-1,Y+1,Z)] = (1-\rho)\frac{\alpha}{n}XYdt + o(dt)\\
&\mathrm{P}[(X,Y,Z) \to (X-1,Y,Z+1)] = \rho\frac{\alpha}{n}XYdt + o(dt)\\
&\mathrm{P}[(X,Y,Z) \to (X,Y-1,Z+1)] = \gamma Ydt + o(dt)\\
&\mathrm{P}[(X,Y,Z) \to (X,Y+1,Z-1)] = \beta Zdt + o(dt).
\end{aligned}$$

The total population size is $n = X(t) + Y(t) + Z(t)$, which is constant in time. Therefore, the number of inactive cases can be completely determined by the numbers of susceptibles and infectives and we have a two-dimensional stochastic process $\{(X(t), Y(t)), t \geq 0\}$.

4.2 The deterministic model

For the corresponding deterministic model, let $x(t), y(t)$, and $z(t)$ denote the number of susceptibles, infectives, and inactive cases, respectively, at time t. The differential equations for x, y, z are:

$$\begin{aligned}
\frac{dx}{dt} &= -\frac{\alpha}{n}xy\\
\frac{dy}{dt} &= (1-\rho)\frac{\alpha}{n}xy - \gamma y + \beta z\\
\frac{dz}{dt} &= \rho\frac{\alpha}{n}xy + \gamma y - \beta z,
\end{aligned} \tag{4.1}$$

where x, y, and z are non-negative continuous variables. Initially there are x_0 susceptibles, y_0 infectives, and z_0 inactive cases, where x_0, y_0, z_0 are non-negative integers such that $0 < y_0 + z_0 < n$ and $x_0 + y_0 + z_0 = n$. Since $z(t) = n - x(t) - y(t)$, the system (4.1) reduces to

$$\begin{aligned}
\frac{dx}{dt} &= -\frac{\alpha}{n}xy\\
\frac{dy}{dt} &= (1-\rho)\frac{\alpha}{n}xy - \beta x - (\beta+\gamma)y + \beta n.
\end{aligned} \tag{4.2}$$

The function $M_d(\phi_1, \phi_2; t) = \exp\{\phi_1 \frac{x}{n} + \phi_2 \frac{y}{n}\}$, of the deterministic proportions x/n and y/n, corresponds to the moment generating function of the random variables X/n, Y/n when the variables X, Y take the values x, y, respectively, with probability one. Then M_d satisfies the following differential equation:

$$\frac{\partial M_d}{\partial t} = \beta\phi_2\left(M_d - \frac{\partial M_d}{\partial \phi_1}\right) - (\beta+\gamma)\,\phi_2\frac{\partial M_d}{\partial \phi_2} + [(1-\rho)\,\alpha\phi_2 - \alpha\phi_1]\,\frac{\partial^2 M_d}{\partial\phi_1\partial\phi_2}, \tag{4.3}$$

with initial condition $M_d(\phi_1, \phi_2; 0) = \exp\{\phi_1 \frac{x_0}{n} + \phi_2 \frac{y_0}{n}\}$.

The equilibrium of the deterministic model

The first of equations (4.2) shows that $x(t)$ is a non-increasing function of t. As long as y is not zero the value of $x(t)$ will be decreasing until it becomes zero. If $y(t)$ is zero at some point t, then the second equation of (4.2) shows that the derivative of y will be positive (since $x(t)$ is always strictly less than n). Hence y will increase to some positive value and then x will begin decreasing once more. Intuitively, this shows that the only possible equilibrium value for x is zero. Formally, solving the system (4.2) with $dx/dt = dy/dt = 0$, it follows that the system (4.2) admits two possible equilibria $\mathbf{v}^* = (0, \beta n/(\beta+\gamma))$ and $\mathbf{v}^{**} = (n, 0)$. Since $x(0) = x_0 < n$ and $x(t)$ is a non-increasing function of t, $\mathbf{v}^{**}$ is not a possible equilibrium for the system (4.2) with initial condition $x(0) < n$ and thus $\mathbf{v}^*$ is the only possible equilibrium.

In order to study the stability of $\mathbf{v}^*$, write the system (4.2) in the form

$$\frac{d\mathbf{v}}{dt} = F(\mathbf{v}),$$

where $\mathbf{v}(t) = (x(t), y(t))$ for $t \geq 0$ and F is a mapping from $\mathbb{R}^2_+$ into $\mathbb{R}^2$ with coordinates $f_i(x,y)$, $i = 1, 2$, defined by

$$\begin{aligned} f_1(x,y) &= -\frac{\alpha}{n}xy \\ f_2(x,y) &= \frac{\alpha(1-\rho)}{n}xy - \beta x - (\beta+\gamma)y + \beta n. \end{aligned}$$

Let $DF(\mathbf{v}^*)$ be the Jacobian matrix of F at the point $\mathbf{v}^*$, i.e. the matrix whose $(i,\ j)$ element is $\partial f_i(\mathbf{v}^*)/\partial j$ for $i = 1, 2$ and $j = x, y$. Then

$$DF(\mathbf{v}^*) = \begin{bmatrix} -\frac{\alpha\beta}{\beta+\gamma} & 0 \\ \beta\left(\frac{\alpha(1-\rho)}{\beta+\gamma} - 1\right) & -(\beta+\gamma) \end{bmatrix}.$$

If both eigenvalues of $DF(\mathbf{v}^*)$ have negative real parts, then $\mathbf{v}^*$ is uniformly asymptotically stable (see, e.g., Reinhard 1986, Chapters 2, 3). The eigenvalues of $DF(\mathbf{v}^*)$ are $\lambda_1 = -\alpha\beta/(\beta+\gamma)$ and $\lambda_2 = -(\beta+\gamma)$, which are both strictly negative (if both α and β are positive) and hence $\mathbf{v}^* = (0, \beta n/(\beta+\gamma))$ is stable.

4.3 The stochastic model

4.3.1 The probability generating function

Assume that initially there are x_0 susceptibles, y_0 infectives, and z_0 inactive cases, where x_0, y_0, z_0 are non-negative integers such that $0 < y_0 + z_0 < n$ and $x_0 + y_0 + z_0 = n$. The only possible transitions to and from the class of susceptibles X are infections, which each decreases the size of X by one. Therefore the number of susceptibles can only decrease from its initial value, x_0, and hence $X(t) \leq x_0$ for all $t \geq 0$ and the state space $\mathcal{S}$ of the process $\{(X(t), Y(t))\}$ is

$$\mathcal{S} = \left\{(r,s) \in \mathbb{Z}_+^2 : \ 0 \leq r \leq x_0, \ 0 \leq s \leq n, \ 0 \leq r+s \leq n\right\},$$

where $\mathbb{Z}_+^m$ denotes the set of all the vectors $(x_1, \dots, x_m)$ with non-negative integer entries, such that $x_i = 0, 1, \dots$ for all $i = 1, \dots, m$.

Let $p_{rs}(t)$ be the probability that there are r susceptibles and s infectives in the population at time t. Then the $p_{rs}(t)$ satisfy the equations

$$\begin{aligned}\frac{dp_{rs}}{dt} &= \gamma(s+1)p_{r,s+1} + \beta(n-r-s+1)p_{r,s-1} \\ &+ \frac{(1-\rho)\alpha}{n}(r+1)(s-1)p_{r+1,s-1} \\ &+ \frac{\rho\alpha}{n}(r+1)sp_{r+1,s} - \left[\frac{\alpha}{n}rs + \beta n - \beta r + (\gamma-\beta)s\right]p_{rs},\end{aligned} \tag{4.4}$$

for $(r,s) \in \mathcal{S}$, and $p_{rs}(t) = 0$ for all other values of (r,s). The initial conditions are $p_{x_0y_0}(0) = 1$ and $p_{rs}(0) = 0$ for any other $(r,s) \neq (x_0, y_0)$.

The joint probability generating function (PGF) of $X(t)$ and $Y(t)$ is defined as

$$\mathcal{P}(x,y;t) = \mathrm{E}\left[x^{X(t)}y^{Y(t)}\right] = \sum_{(r,s)\in\mathcal{S}} p_{rs}(t)x^r y^s.$$

Using (4.4) it can be shown that $\mathcal{P}(x,y;t)$ satisfies the differential equation

$$\begin{aligned}\frac{\partial\mathcal{P}}{\partial t} &= \beta n(y-1)\mathcal{P} - \beta x(y-1)\frac{\partial\mathcal{P}}{\partial x} - (y-1)(\beta y+\gamma)\frac{\partial\mathcal{P}}{\partial y} \\ &+ \frac{\alpha}{n}y[(1-\rho)y + \rho - x]\frac{\partial^2\mathcal{P}}{\partial x\partial y},\end{aligned} \tag{4.5}$$

with initial condition $\mathcal{P}(x,y;0) = x^{x_0}y^{y_0}$.

The PGF can be written in the form

$$\mathcal{P}(x,y;t) = \sum_{r=0}^{x_0} x^r f_r(y;t), \tag{4.6}$$

where

$$f_r(y;t) = \sum_{s=0}^{n-r} y^s p_{rs}(t), \quad r = 0, 1, \dots, x_0, \tag{4.7}$$

a method suggested by Gani (1965) and Siskind (1965). The functions f_r are defined only for $r = 0, 1, \dots, x_0$, since $p_{rs}(t) = 0$ whenever $r > x_0$, and hence $f_r \equiv 0$ for any

$r > x_0$. Substituting (4.6) in (4.5) and equating the coefficients of x^r, the following equations are deduced:

$$\begin{aligned}\frac{\partial f_r}{\partial t} &= \beta(y-1)(n-r)f_r - \left[(y-1)(\beta y+\gamma) + \frac{\alpha}{n}yr\right]\frac{\partial f_r}{\partial y} \\ &\quad + \frac{\alpha}{n}y[(1-\rho)y+\rho](r+1)\frac{\partial f_{r+1}}{\partial y},\end{aligned} \tag{4.8}$$

for $r = 0, 1, ..., x_0$ and $f_r \equiv 0$ for $r = x_0+1, ..., n$.

Theoretically speaking it is possible to solve these equations recursively. Gani (1965) and Siskind (1965) suggested this approach for the SIR model. Siskind solved the set of equations for the f_r, while Gani solved the respective equations for the Laplace transforms of f_r. In both cases the algebra required was quite cumbersome and the results that Gani and Siskind deduced for the probabilities $p_{rs}(t)$ involve highly complicated formulae. In order to get some insight into the feasibility of these recursive solutions for the model presented in this chapter, we consider a simple case of a population with two susceptibles and one infectious case at time $t = 0$ ($x_0 = 2$, $y_0 = 1$, $z_0 = 0$, and $n = 3$). According to (4.8), $f_3 \equiv 0$ and f_2 satisfies a first-order partial differential equation whose solution is $f_2(y;t) = p_{20}(t) + p_{21}(t)y$, where

$$\begin{aligned}p_{20}(t) &= \frac{c_0c_1}{2\sqrt{E}}\left[\exp\{-(1+c_0)\beta t\} - \exp\{-(1+c_1)\beta t\}\right] \\ p_{21}(t) &= \frac{1}{2\sqrt{E}}\left[c_0\exp\{-(1+c_0)\beta t\} - c_1\exp\{-(1+c_1)\beta t\}\right],\end{aligned} \tag{4.9}$$

and $c_0 = D+\sqrt{E}$, $c_1 = D-\sqrt{E}$, $D = (\gamma-\beta+2\alpha/3)/(2\beta)$, and $E = D^2+\gamma/\beta \geq 0$. Since $\partial f_2(y;t)/\partial y = p_{21}(t)$, equation (4.8) for $r = 1$ gives

$$\frac{\partial f_1}{\partial t} = 2\beta(y-1)f_1 - \left[(y-1)(\beta y+\gamma) + \frac{\alpha}{3}y\right]\frac{\partial f_1}{\partial y} + \frac{2\alpha}{3}y\left[(1-\rho)y+\rho\right]p_{21}(t). \tag{4.10}$$

The method of separation of variables, applied to both (4.10) and to its equivalent for the Laplace transform of f_1, does not prove to be very effective in obtaining a solution. Substituting for f_1 from (4.7) and equating the coefficients of y^s in (4.10), we obtain a system of three differential equations with four unknowns, the $p_{10}(t)$, $p_{11}(t)$, $p_{12}(t)$, $p_{21}(t)$. Substituting for $p_{21}(t)$ from (4.9), the system will also include expressions of t. It is clear that this recursive technique will be difficult to implement in practice, with a population of even a moderate size, and so we have not continued further with this approach.

A similar approach for solving (4.5) was suggested by Dietz (1967). For this, write the PGF in the form

$$\mathcal{P}(x,y;t) = \sum_{j=0}^{n} \binom{n}{j} (x-1)^j f_j(y;t), \tag{4.11}$$

where the functions f_j are to be determined. Substituting (4.11) in (4.5) and equating the coefficients of $(x-1)^j$, a system of differential equations for the f_j's is obtained, which can be solved iteratively. Again the amount of algebra required makes this methodology difficult to implement in practice.

Also the method of separation of variables is not applicable for equation (4.5); the time dependence can be separated, but not that of x and y, since the variables $X(t)$ and $Y(t)$ are not independent.

4.3.2 The moment generating function and the moments of X and Y

Let $U(t) = X(t)/n$ and $V(t) = Y(t)/n$ denote the proportions of susceptibles and infectives, respectively, in the population at time t. From the differential equation for the PGF, it is easily shown that the moment generating function (MGF), $M_p(\phi_1, \phi_2; t) = \mathrm{E}[\exp\{\phi_1 U(t) + \phi_2 V(t)\}]$, for the proportions U and V satisfies the equation

$$\begin{aligned}\frac{\partial M_p}{\partial t} &= \beta n \left(e^{\phi_2/n} - 1\right)\left(M_p - \frac{\partial M_p}{\partial \phi_1}\right)\\ &\quad + n\left[\gamma e^{-\phi_2/n} + (\beta-\gamma) - \beta e^{\phi_2/n}\right]\frac{\partial M_p}{\partial \phi_2}\\ &\quad + n^2\left[\frac{(1-\rho)\alpha}{n} e^{(\phi_2-\phi_1)/n} + \frac{\rho\alpha}{n} e^{-\phi_1/n} - \frac{\alpha}{n}\right]\frac{\partial^2 M_p}{\partial\phi_1\partial\phi_2}.\end{aligned} \tag{4.12}$$

Substituting the series expansions for the exponentials in (4.12) and keeping only the terms of order n^0, we obtain the same differential equation (4.3) that is satisfied by the MGF for the deterministic proportions. Including the terms of order n^{-1} in the expansion of equation (4.12), we obtain:

$$\begin{aligned}\frac{\partial M_{pn}}{\partial t} &= \beta\left(\phi_2 + \frac{\phi_2^2}{2n}\right)\left(M_{pn} - \frac{\partial M_{pn}}{\partial \phi_1}\right)\\ &\quad - \left[(\beta+\gamma)\phi_2 + (\beta-\gamma)\frac{\phi_2^2}{2n}\right]\frac{\partial M_{pn}}{\partial \phi_2}\\ &\quad + \alpha\left[(1-\rho)\left(\phi_2 - \phi_1 + \frac{(\phi_2-\phi_1)^2}{2n}\right) + \rho\left(-\phi_1 + \frac{\phi_1^2}{2n}\right)\right]\frac{\partial^2 M_{pn}}{\partial\phi_1\partial\phi_2},\end{aligned} \tag{4.13}$$

from which a normal approximation to the distribution of U, V (and of X, Y as well) can be deduced. For, a bivariate Gaussian distribution with mean $\boldsymbol{\xi} = (\xi_1, \xi_2)$ and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$, $i, j = 1, 2$, has moment generating function of the form

$$M_{pn}(\phi_1, \phi_2; t) = \exp\left\{\phi_1\xi_1 + \phi_2\xi_2 + \frac{1}{2n}(\phi_1^2\sigma_{11} + 2\phi_1\phi_2\sigma_{12} + \phi_2^2\sigma_{22})\right\}. \tag{4.14}$$

Substituting (4.14) in both sides of (4.13) and equating the coefficients of ϕ_1, ϕ_2, the following system for the approximate moments of X and Y is deduced:

$$\begin{aligned}
\frac{d\mu_X}{dt} &= -\frac{\alpha}{n}(\sigma_{XY} + \mu_X\mu_Y) \\
\frac{d\mu_Y}{dt} &= \beta n - \beta\mu_X - (\beta+\gamma)\mu_Y + \frac{\alpha(1-\rho)}{n}(\sigma_{XY} + \mu_X\mu_Y) \\
\frac{d\sigma_{XX}}{dt} &= -\frac{2\alpha}{n}(\mu_X\sigma_{XY} + \mu_Y\sigma_{XX}) + \frac{\alpha}{n}(\sigma_{XY} + \mu_X\mu_Y) \\
\frac{d\sigma_{XY}}{dt} &= -\beta\sigma_{XX} - (\beta+\gamma)\sigma_{XY} + \frac{\alpha}{n}(\mu_Y\sigma_{XY} + \mu_X\sigma_{YY}) \\
&\quad + \frac{\alpha(1-\rho)}{n}(\mu_X\sigma_{XY} + \mu_Y\sigma_{XX} - \sigma_{XY} - \mu_X\mu_Y) \\
\frac{d\sigma_{YY}}{dt} &= \beta(n-\mu_X) - (\beta-\gamma)\mu_Y - 2\beta\sigma_{XY} - 2(\beta+\gamma)\sigma_{YY} \\
&\quad + \frac{2\alpha(1-\rho)}{n}(\mu_Y\sigma_{XY} + \mu_X\sigma_{YY}) + \frac{\alpha(1-\rho)}{n}(\sigma_{XY} + \mu_X\mu_Y),
\end{aligned} \tag{4.15}$$

where μ_X, μ_Y and σ_{XX}, σ_{YY} are the means and variances of X, Y, respectively, and σ_{XY} is the covariance of X, Y.

The system (4.15) can also be deduced directly from the corresponding system for the exact first and second moments of X and Y, by using the appropriate moment relationships for the bivariate normal distribution. The exact moments are obtained from the differential equation for the MGF, or that for the PGF, or from the forward equations for the expectations of X, Y, X^2, Y^2, and XY. The equations for the means are the same as in (4.15). The equations for the variances and the covariance of X and Y (given in the Appendix, equation (A.1)) involve third order moments of X and Y. Hence the system of equations for the first and second moments involves higher-order moments (a result of the non-linearity of the transition rates), so that it is an open system and cannot be solved. Assuming normality, the third moments of X and Y can be expressed in terms of the first and second moments, thus yielding the closed system (4.15). Alternatively, the system (4.15) can be deduced from the cumulant-generating function by setting all the cumulants of order greater than two to be zero.

Further approximations can be obtained by incorporating cumulants of order greater than two. For instance, replacing $M_p(\phi_1, \phi_2; t)$ by $\exp\{K(\phi_1, \phi_2; t)\}$, (where $K(\phi_1, \phi_2; t)$ is the cumulant-generating function) in (4.12), expanding both sides of the resulting equation in powers of ϕ_1 and ϕ_2, and equating coefficients, yields an open system of equations for the cumulants (the system for the cumulants of order up to m includes cumulants of order $m+1$). If we assume that all the cumulants of order greater than m are zero, then we get a closed system. In this case it is preferable to work with the cumulant-generating function, since that gives directly a system for the cumulants and the order of the approximation (i.e. the highest order of the cumulants included in the system) can be changed easily (see, e.g., Matis & Kiffe 2000 and references). Other approximation techniques, such as the saddlepoint approximation (see, e.g., Renshaw 1998, 2001) and the linear approximation (see, e.g., Herbert 1998 and Sections 5.4 and 6.3.8) have also been developed. In this chapter only the normal approximation has been investigated and its results are compared with results from simulations and from the deterministic model in Section 4.4.

The idea for the normal approximation follows from a suggestion of Whittle (1957) and its validity has been established for a more general class of Markov processes (see, e.g., Kurtz 1970, 1971, 1981) by limiting results showing that such processes can be approximated by Gaussian diffusion processes as the total population size tends to infinity appropriately. The relevant theorem is as follows:

Theorem 4.1 *(Kurtz 1970, Theorem 3.1, and Kurtz 1971, Theorem 3.1) Let* $\mathbf{X}_N(t)$ *with* $N = 1, 2, \ldots$ *be a one-parameter family of continuous time Markov chains with state space* $\mathbf{E}_N \subset \mathbb{Z}^k$ *and let* $\mathbf{V}_N(t) = N^{-1}\mathbf{X}_N(t)$. *If* q_{ij}^N *are the infinitesimal rates of* $\mathbf{X}_N$, *we assume that there exists a continuous function* $f : \mathbb{R}^k \times \mathbb{Z}^k \to \mathbb{R}$ *that satisfies*

$$q_{\mathbf{x},\mathbf{x}+\mathbf{h}} = Nf\left(\frac{1}{N}\mathbf{x}, \mathbf{h}\right),$$

for all positive integers N, $\mathbf{x} \in \mathbf{E}_N$, *and increments* $\mathbf{h} = (h_1, \ldots, h_k) \in \mathbb{Z}^k$. *Define the functions* $F : \mathbb{R}^k \to \mathbb{R}^k$ *and* $g_{ij} : \mathbb{R}^k \to \mathbb{R}$ *for* $i, j = 1, 2, \ldots, k$ *by*

$$F(\mathbf{v}) = \sum_{\mathbf{h}} \mathbf{h} f(\mathbf{v}, \mathbf{h})$$

$$g_{ij}(\mathbf{v}) = \sum_{\mathbf{h}} h_i h_j f(\mathbf{v}, \mathbf{h}),$$

for $\mathbf{v} \in \mathbb{R}^k$. *Let* $G(\mathbf{v})$ *be the matrix whose* (i,j) *element is* $g_{ij}(\mathbf{v})$. *Suppose that there exists an open set* $\mathbf{E} \subset \mathbb{R}^k$ *such that*

(a) there exists a constant M *such that*

$$|F(\mathbf{v}_1) - F(\mathbf{v}_2)| \leq M|\mathbf{v}_1 - \mathbf{v}_2|, \quad \textit{for all } \mathbf{v}_1,\ \mathbf{v}_2 \textit{ in } \mathbf{E}$$

(b) $\sup_{\mathbf{v}\in\mathbf{E}} \sum_{\mathbf{h}} |\mathbf{h}| f(\mathbf{v},\mathbf{h}) < \infty$

(c) $\lim_{d\to\infty} \sup_{\mathbf{v}\in\mathbf{E}} \sum_{\mathbf{h}:|\mathbf{h}|>d} |\mathbf{h}| f(\mathbf{v},\mathbf{h}) = 0$

(d) the solution $\mathbf{Z}(t,\mathbf{z}_0)$ *of the deterministic initial value problem:*

$$\left\{\frac{d\mathbf{Z}}{dt} = F(\mathbf{Z}),\ \mathbf{Z}(0) = \mathbf{z}_0\right\},$$

is such that $\mathbf{Z}(t,\mathbf{z}_0) \in \mathbf{E}$ *for all* $t \leq T$ *and* $\lim_{N\to\infty} \frac{1}{N}\mathbf{X}_N(0) = \mathbf{z}_0$.

Then for every $\epsilon > 0$

$$\lim_{N\to\infty} \mathrm{P}\left[\sup_{t\leq T}\left|\frac{1}{N}\mathbf{X}_N(t) - \mathbf{Z}(t,\mathbf{z}_0)\right| > \epsilon\right] = 0.$$

Let

$$\mathbf{Z}_N(t) = \frac{1}{N}\mathbf{X}_N(t) - \frac{1}{N}\mathbf{X}_N(0) - \int_0^t F(\mathbf{V}(s))ds$$

and $\mathbf{W}_N(t) = \sqrt{N}\mathbf{Z}_N(t)$. *If* $G(\mathbf{v})$ *is bounded and uniformly continuous in* $\mathbf{E}$ *and*

$$\lim_{d\to\infty} \sup_{\mathbf{v}\in\mathbf{E}} \sum_{\mathbf{h}:|\mathbf{h}|>d} |\mathbf{h}|^2 f(\mathbf{v},\mathbf{h}) = 0,$$

then, as $N \to \infty$, $\mathbf{W}_N(t)$ *converges weakly to the diffusion process* $\mathbf{W}(t)$ *with characteristic function*

$$\mathrm{E}[\exp\{i\theta\mathbf{W}(t)\}] = \exp\left\{-\frac{1}{2}\sum_{i,j}\theta_i\theta_j\int_0^t g_{ij}(\mathbf{Z}(s,\mathbf{z}_0))ds\right\}.$$

Kurtz (1981) defines N as "a parameter which has the same order of magnitude as the total population size". For the model considered in this chapter, N can be taken equal to the actual population size n. It then follows from this theorem that for large populations (and as long as the initial numbers $\mathbf{X}_0$ scale with N so that the proportions $\mathbf{V}_0 = \mathbf{X}_0/N$ are kept fixed) the process can be approximated by a Gaussian diffusion process about its expected value.

For our model the increments $\mathbf{h} \in \mathbb{R}^2$ are

$$\mathbf{h}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix},\quad \mathbf{h}_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix},\quad \mathbf{h}_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix},\quad \mathbf{h}_4 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The functions $f : \mathbb{R}^2 \times \mathbb{Z}^2 \to \mathbb{R}$ and $F : \mathbb{R}^2 \to \mathbb{R}^2$ are defined by

$$f(\mathbf{v}, \mathbf{h}_i) = \begin{cases} \alpha(1-\rho)v_1v_2 & i = 1 \\ \alpha\rho v_1 v_2 & i = 2 \\ \gamma v_2 & i = 3 \\ \beta(1 - v_1 - v_2) & i = 4 \end{cases}$$

$$F(\mathbf{v}) = \begin{bmatrix} -\alpha v_1 v_2 \\ \alpha(1-\rho)v_1v_2 - \beta v_1 - (\beta+\gamma)v_2 + \beta \end{bmatrix},$$

for $\mathbf{v} = (v_1, v_2)$. It can easily be proved that the conditions of Kurtz's theorem hold and hence the process can be approximated by a Gaussian diffusion process as n tends to infinity.

4.3.3 The equilibrium state of the process

The process described by (4.4) is a continuous time Markov chain with finite state space

$$\mathcal{S} = \left\{(r,s) \in \mathbb{Z}_+^2 : \ 0 \le r \le x_0, \ 0 \le s \le n, \ 0 \le r+s \le n\right\}.$$

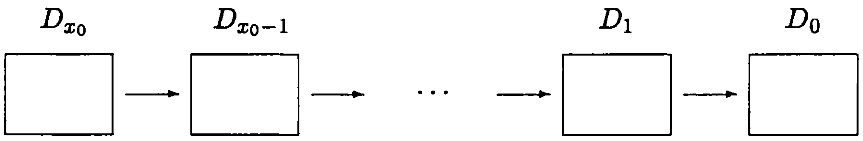
The state space $\mathcal{S}$ can be partitioned into the sets D_0, $D_1, \dots$, D_{x_0} where

$$D_r = \{(r,s) \in \mathcal{S} : \ 0 \le s \le n-r\}, \quad \text{for } r = 0, 1, \dots, x_0.$$

All the states within each of the sets D_0, $D_1, \dots$, D_{x_0} communicate with each other and hence each of the sets D_0, $D_1, \dots$, D_{x_0} is an irreducible class. There is a positive probability for transitions from D_r to $D_{r'}$ for any $r = 1, \dots, x_0$ and $r' < r$, but there cannot be any transitions from any D_r to any $D_{r'}$ with $r' > r$. This means that the classes $D_1, \dots, D_{x_0}$ are open, while D_0 is closed. Hence if the chain reaches one of the states in D_0 then it will remain within D_0. From Markov chain theory it is known that open irreducible classes are transient, while finite closed irreducible classes are positive recurrent (see, e.g., Wolff 1989, Chapters 3, 4). Therefore D_0 is positive recurrent and $D_1, \dots, D_{x_0}$ are transient. The irreducible classes of $\mathcal{S}$ are illustrated in Figure 4.2.

Let $\mathbf{W}(t) = (X(t), Y(t))$ and $P_{ij}(t) = \mathrm{P}[\mathbf{W}(t) = j | \mathbf{W}(0) = i]$, for i, j in $\mathcal{S}$ and $t \ge 0$. For finite-state Markov chains the pointwise limits $\lim_{t\to\infty} P_{ij}(t)$ always exist and they are equal to zero if j is transient. Therefore

$$\begin{aligned} &\lim_{t\to\infty} P_{ij}(t) = 0 && \text{if } j \notin D_0 \\ &\lim_{t\to\infty} \sum_{j \in D_0} P_{ij}(t) = 1, \end{aligned}$$



$$D_r = \{(r,s) \in \mathbb{Z}_+^2 : \; 0 \leq s \leq n-r\}, \text{ for } r = 0, 1, \ldots, x_0$$

Figure 4.2: The irreducible classes of $\mathcal{S}$

which means that the chain will ultimately be absorbed in D_0 and there will be no susceptibles in the population.

Intuitively this result should be expected since there is no replenishment of the susceptible population. As long as there exists at least one infective in the population, he may infect any existing susceptibles and thus reduce that population. If at some point there are no infectives in the population then there will be $n - X(t) > 0$ inactive cases (since $X(t)$ is always strictly less than n); the system will remain in this state until a latent or recovered individual develops disease. Then there will be one infective who may infect any existing susceptibles and the population of susceptibles may be reduced again.

Therefore ultimately all the susceptibles will get infected and the population will consist of infectious and inactive cases only. After the last susceptible has been infected, there will still be transitions between the infectious and the inactive class (due to recovery, relapse and reactivation), so that the numbers of infectious and inactive cases will still vary with time. At that point the number of infectives can be described by a birth and death process, which will reach an equilibrium state, depending on the parameters of the process.

The limiting distribution can be deduced in terms of the dominant eigenvalue of the infinitesimal matrix, Q, of the process (since for Markov processes with finite state space, $\mathcal{S}$, the matrix $P(t) = \{P_{ij}(t), i, j \in \mathcal{S}\}$ can be expressed as $P(t) = e^{Qt}$), or from the differential equation (4.5) for the probability generating function. Let

$$q_{rs} = \lim_{t \to \infty} p_{rs}(t) \quad \text{and} \quad Q(x,y) = \sum_{r,s} q_{rs} x^r y^s.$$

Taking the limits as $t \to \infty$ and setting $y = 1$ and $\partial Q / \partial t = 0$ in (4.5) gives:

$$0 = \frac{\alpha}{n}(1-x) \left[\frac{\partial^2 Q}{\partial x \partial y} \right]_{y=1} = \frac{\alpha}{n}(1-x) \sum_{r=1}^{x_0} \sum_{s=1}^{n-r} r s x^{r-1} q_{rs}.$$

Equating the coefficients of x^r, we find that:

$$\sum_{s=1}^{n-r} sq_{rs} = 0, \quad \text{for all } r = 1, 2, \dots, x_0,$$

and hence

$$q_{rs} = 0, \quad \text{for any } r > 0 \text{ and } s > 0. \tag{4.16}$$

Taking the limits as $t \to \infty$ in (4.4) and setting $s = 0$, equation (4.4) reduces to

$$0 = \gamma q_{r1} - \beta(n - r)q_{r0}, \quad \text{for } r > 0.$$

Use of (4.16) leads to

$$q_{r0} = 0, \quad \text{for any } r > 0. \tag{4.17}$$

If X_e is the random variable whose distribution is defined by

$$\mathrm{P}[X_e = r] = \lim_{t\to\infty} \mathrm{P}[X(t) = r],$$

and similarly for Y_e and Z_e, then combining (4.16) and (4.17) we deduce that

$$1 = \sum_{(r,s)\in S} q_{rs} = \sum_{s=0}^{n} q_{0s} = \mathrm{P}[X_e = 0], \tag{4.18}$$

which shows that the population of susceptibles will ultimately be exhausted with probability one and hence

$$\mathrm{E}[X_e] = \sum_{r=1}^{x_0}\sum_{s=0}^{n-r} rq_{rs} = 0.$$

Since $X_e + Y_e + Z_e = n$, from (4.18) also follows that the probability of ultimate extinction of TB, $\mathrm{P}[Y_e = Z_e = 0]$, is zero. The limiting distribution of the number of infectives is easily obtained from the probability generating function; using (4.16) and (4.17), $Q(x, y)$ becomes a function of y only:

$$Q(x, y) = \sum_{r=1}^{x_0}\sum_{s=0}^{n-r} x^r y^s q_{rs} = \sum_{s=0}^{n} y^s q_{0s} = Q(y).$$

Therefore the derivatives of Q with respect to x are zero and (4.5) reduces to a homogeneous first-order partial differential equation, whose solution is

$$Q(y) = \left(\frac{\gamma}{\beta+\gamma} + \frac{\beta}{\beta+\gamma}y\right)^n.$$

This is the probability generating function of the binomial distribution with parameters n and $\beta/(\beta+\gamma)$, so that the distribution of the number of infectives at equilibrium is

$$q_{0s} = \binom{n}{s}\frac{\beta^s\gamma^{n-s}}{(\beta+\gamma)^n}, \quad \text{for } s = 0, 1, \dots, n,$$

with mean and variance

$$\mathrm{E}[Y_e] = \frac{\beta n}{\beta+\gamma} \qquad \mathrm{Var}[Y_e] = \frac{\beta\gamma n}{(\beta+\gamma)^2}.$$

Also, since $\mathrm{E}[Z_e] = n - \mathrm{E}[X_e] - \mathrm{E}[Y_e]$, it follows that $\mathrm{E}[Z_e] = \gamma n/(\beta+\gamma)$.

It is possible though that the size of the susceptible population becomes zero before the process of infectives reaches this equilibrium state. Then the distribution of the numbers of infectives and inactive cases may still vary in time. Supposing this is true, let T be the time that the number of susceptibles becomes zero and write the time point t as $t = T + \tau$ (so that τ counts the time after T).

The mean and variance of X, the covariance of X and Y, and their derivatives are zero after T. Hence the system for the first and second moments of X and Y reduces to a system for the mean and variance of Y only, whose solution is

$$\begin{aligned}\mathrm{E}[Y(T+\tau)] &= \frac{\beta n}{\beta+\gamma}[1 - e^{-(\beta+\gamma)\tau}] + \mathrm{m}_0^e e^{-(\beta+\gamma)\tau} \\ \mathrm{Var}[Y(T+\tau)] &= \frac{\beta\gamma n}{(\beta+\gamma)^2} + [\beta n - (\beta+\gamma)\mathrm{m}_0^e]\frac{\beta-\gamma}{(\beta+\gamma)^2}e^{-(\beta+\gamma)\tau} \\ &+ \left[\mathrm{v}_0^e - \frac{\beta^2 n}{(\beta+\gamma)^2} + \frac{(\beta-\gamma)\mathrm{m}_0^e}{\beta+\gamma}\right]e^{-2(\beta+\gamma)\tau},\end{aligned} \tag{4.19}$$

where $\mathrm{m}_0^e = \mathrm{E}[Y(T)]$ and $\mathrm{v}_0^e = \mathrm{Var}[Y(T)]$. The mean and variance of Z and the covariance of Y, Z after time T are obtained from (4.19), since $Z(t) = n - Y(t)$, for $t \geq T$.

After the extinction of the susceptible population, the process of infectives can be described by a birth and death process with birth and death rates, respectively,

$$\lambda_s = \begin{cases} \beta(n-s) & s = 0, 1, \dots, n-1 \\ 0 & \text{otherwise} \end{cases} \qquad \mu_s = \begin{cases} \gamma s & s = 1, 2, \dots, n \\ 0 & \text{otherwise.} \end{cases}$$

If $p_s(\tau) = \mathrm{P}[Y(T+\tau) = s]$ is the probability that there are s infectives at time $T+\tau$ then the probability generating function, $G(y;\tau) = \sum_{s=0}^{n} y^s p_s(\tau)$, satisfies the equation

$$\frac{\partial G}{\partial \tau} = \beta n(y-1)G + [\gamma + (\beta-\gamma)y - \beta y^2]\frac{\partial G}{\partial y}. \tag{4.20}$$

This equation can also be deduced from the differential equation (4.5) for the probability generating function, by setting $\partial\mathcal{P}/\partial x = 0$ and $\mathcal{P}(x, y; T + \tau) \equiv G(y; \tau)$. The solution of (4.20) is of the form

$$G(y;\tau) = \sum_{j=0}^{n} c_j e^{j(\beta+\gamma)\tau}(y-1)^j \left(y + \frac{\gamma}{\beta}\right)^{n-j},$$

where the coefficients c_j are determined by solving the initial conditions $G(y; 0) = \sum_{s=0}^{n} y^s \mathrm{P}[Y(T) = s]$.

Differentiating (4.20) with respect to y and taking the values for $y = 1$, the differential equation for $\mathrm{E}[Y(T+\tau)]$ is obtained, whose solution is given in (4.19). Similarly, the expression in (4.19) for the variance of $Y(T + \tau)$ can be deduced.

4.3.4 The time until the extinction of susceptibles

The distribution of the time, T, until the extinction of the susceptibles is given by

$$\mathrm{P}[T \le t] = \mathrm{P}[X(t) = 0] = \sum_{s=0}^{n} p_{0s}(t) = \mathcal{P}(0, 1; t), \tag{4.21}$$

and, in principle at least, this can be determined using an iterative scheme, as discussed in Section 4.3.1. Another possible way to derive the distribution of T is by considering the times between successive infections. Let τ_1 be the time until the first infection and τ_r the time between the $(r-1)$th and the rth infection, for $r = 2, 3, \dots, x_0$. Then

$$T = \sum_{r=1}^{x_0} \tau_r,$$

and the distribution of T can be deduced from that of τ_r. However it is not straightforward to determine the distribution of τ_r, because this depends on the number of infectives present during the interval τ_r, which may change during the interval.

To this end, we consider the embedded random walk: the successive states of the process are described by the points (r, s) on the plane, where $r = 0, \dots, x_0$, $s = 0, \dots, n$, and $0 \le r + s \le n$. The transition probabilities are:

$$\begin{aligned}
\mathrm{P}[(r,s) \to (r-1, s+1)] &= \frac{\alpha(1-\rho)rs/n}{\alpha_{rs}}\\
\mathrm{P}[(r,s) \to (r-1, s)] &= \frac{\alpha\rho rs/n}{\alpha_{rs}}\\
\mathrm{P}[(r,s) \to (r, s-1)] &= \frac{\gamma s}{\alpha_{rs}}\\
\mathrm{P}[(r,s) \to (r, s+1)] &= \frac{\beta(n-r-s)}{\alpha_{rs}},
\end{aligned} \tag{4.22}$$

where

$$\alpha_{rs} = \frac{\alpha}{n} rs + \gamma s + \beta(n - r - s),$$

for $r = 0, 1, \ldots, x_0$, $s = 0, 1, \ldots, n$, and $0 \leq r + s \leq n$. The process starts from the point $(x_0,\ y_0)$ and ends at some point $(0,\ s)$, $s = 0, \ldots, n$. The line $r = 0$ is an absorbing barrier. From (4.22) it follows that the time from the epoch when the process enters the point $(r,\ s)$ until the first transition out of $(r,\ s)$ is exponentially distributed with parameter α_{rs}; the probability that this first step from $(r,\ s)$ is an infection is $\alpha n^{-1} rs/\alpha_{rs}$.

Let E_{uv} denote the expected number of steps until absorption for a random walk starting from $(u,\ v)$. Then E_{uv} is equal to $E_{u-1,v+1} + 1$ if the first step is "infection that leads to disease", $E_{u-1,v} + 1$ if the first step is "infection followed by non-zero latent period", and $E_{u,v-1} + 1$ or $E_{u,v+1} + 1$ if the first step is "recovery" or "development of disease", respectively. Therefore, using (4.22), E_{uv} can be written as

$$\begin{aligned} E_{uv} = {} & \frac{\alpha(1-\rho)uv}{n\alpha_{uv}} E_{u-1,v+1} + \frac{\alpha\rho uv}{n\alpha_{uv}} E_{u-1,v} \\ & + \frac{\gamma v}{\alpha_{uv}} E_{u,v-1} + \frac{\beta(n-u-v)}{\alpha_{uv}} E_{u,v+1} + 1, \end{aligned} \tag{4.23}$$

with initial condition $E_{0v} = 0$, for any value of v. The simplest case is when there is just one susceptible left. Setting $u = 1$, (4.23) reduces to

$$E_{1v} = \frac{\gamma v}{\alpha_{1v}} E_{1,v-1} + \frac{\beta(n-1-v)}{\alpha_{1v}} E_{1,v+1} + 1, \quad \text{for } v = 0, 1, \ldots, n-1. \tag{4.24}$$

The general solution of the homogeneous equation corresponding to (4.24) is

$$E_{1v}^{Hom} = A_1 \phi_1^v + A_2 \phi_2^v,$$

where

$$\phi_1(v) = \frac{1 + \sqrt{1 - 4\epsilon(v)\delta(v)}}{2\epsilon(v)} \qquad \phi_2(v) = \frac{1 - \sqrt{1 - 4\epsilon(v)\delta(v)}}{2\epsilon(v)},$$

A_1, A_2 are constants, $\delta(v) = \gamma v/\alpha_{1v}$, $\epsilon(v) = \beta(n-1-v)/\alpha_{1v}$, and $1 - 4\epsilon(v)\delta(v) \geq 0$ for any $v \geq 0$. Now, if ϕ_0 is a solution of (4.24), then its general solution is

$$E_{1v} = \phi_0 + A_1 \phi_1^v + A_2 \phi_2^v.$$

The values of the constants A_1, A_2 are determined by the initial condition, E_{10}, which is unknown. One possibility is to obtain approximations to the general solutions E_{1v} by using numerical results to approximate the value of E_{10}, but solution by this approach will not be pursued further here. Equations similar to (4.23) can be deduced for the expected time until absorption and the probability generating function for the number of steps until absorption, but similar problems in determining solutions arise.

4.4 Numerical Results

In this section results are presented for an epidemic in a population of size $n = 1000$ starting with the introduction of ten infective cases at time $t = 0$, so that $x_0 = 990$ and $y_0 = 10$. The values of the parameters were chosen to be representative of those for TB (see references in Sections 2.3 and 2.7):

$$\begin{aligned} \alpha &= 10 & \beta &= 0.0022 \\ \rho &= 0.9725 & \gamma &= 0.066. \end{aligned} \tag{4.25}$$

The values of the deterministic x, y, z and the stochastic means of X, Y, Z as obtained from the normal approximation and from 10000 simulations are shown, as functions of time, in Figures 4.3, 4.4, and 4.5. Table 4.1 shows the respective values for X and Y for the first 15 years, while Figures 4.7 and 4.8 show the values of $X(t)$ and $Y(t)$ from an individual realisation of the stochastic model. The distribution of the time until the extinction of susceptibles was also calculated from the simulations (Figure 4.6). The results for the deterministic model and the normal approximation were obtained by solving numerically the systems (4.1) and (4.15), respectively. Details for the implementation of the simulations can be found in the Appendix (Section A.1.2).

The results for the X means are presented in Figure 4.3. The three curves are for the deterministic x, the mean of X based on the simulations, and the mean of X from the normal approximation. Initially the three curves are almost the same. During the interval between 5 and 20 years the deterministic x decreases more rapidly and falls below the values of the stochastic means. TB spreads quickly among the susceptibles and by time $t = 50$ both the deterministic x and the stochastic means have become zero. In the single realization presented in Figure 4.7 the value of $X(t)$ became zero during the 27th year.

The distribution of T, the time that the last susceptible gets infected, is given by

$$F_T(t) = \mathrm{P}[T \leq t] = \mathrm{P}[X(t) = 0].$$

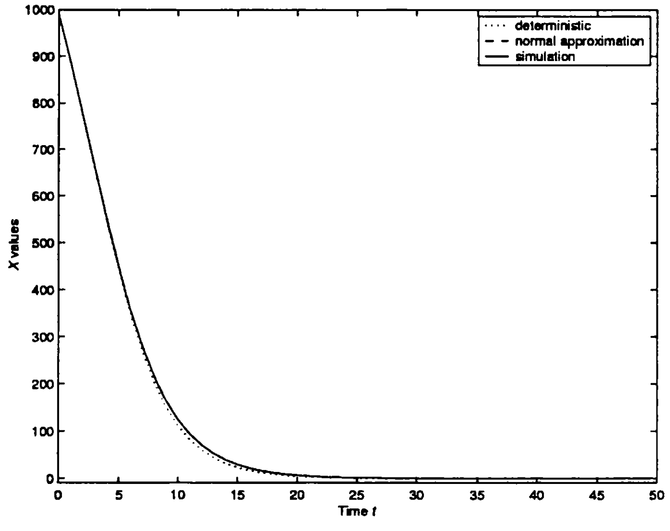


Figure 4.3: The deterministic $x(t)$ and the stochastic means of $X(t)$ as obtained from the normal approximation and from simulations. The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$. Time is measured in years.

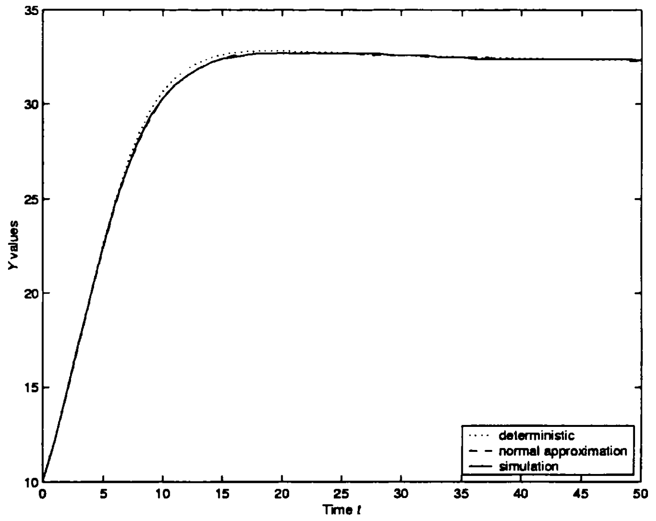


Figure 4.4: The deterministic $y(t)$ and the stochastic means of $Y(t)$ as obtained from the normal approximation and from simulations. The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$. Time is measured in years.

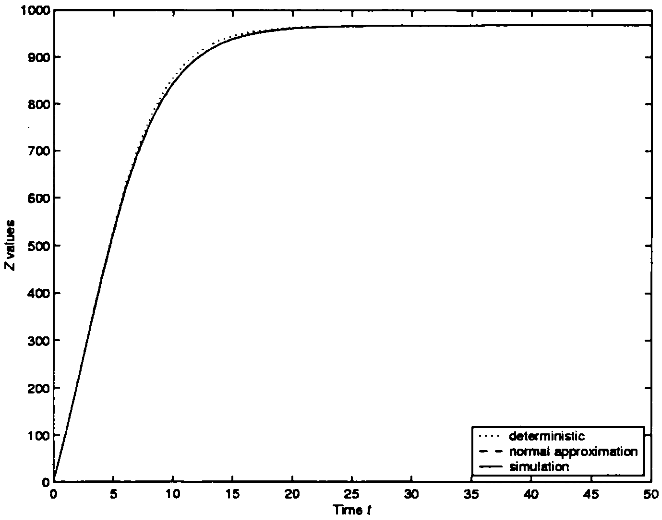


Figure 4.5: The deterministic $z(t)$ and the stochastic means of $Z(t)$ as obtained from the normal approximation and from simulations. The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$. Time is measured in years.

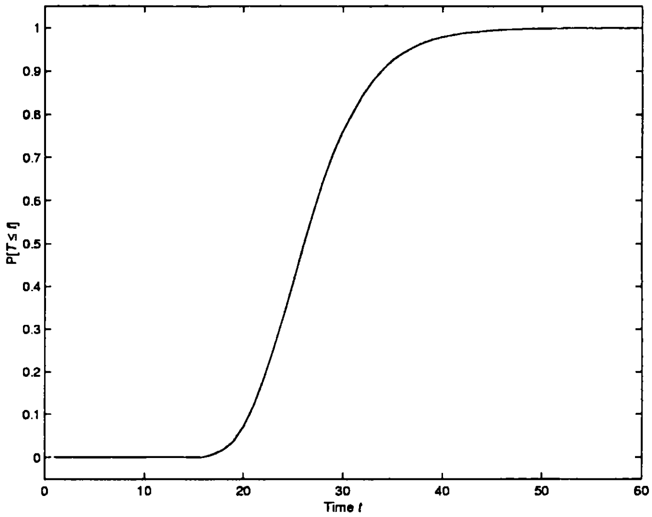


Figure 4.6: Estimates of the distribution of the time T until the extinction of susceptibles. The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$. Time is measured in years.

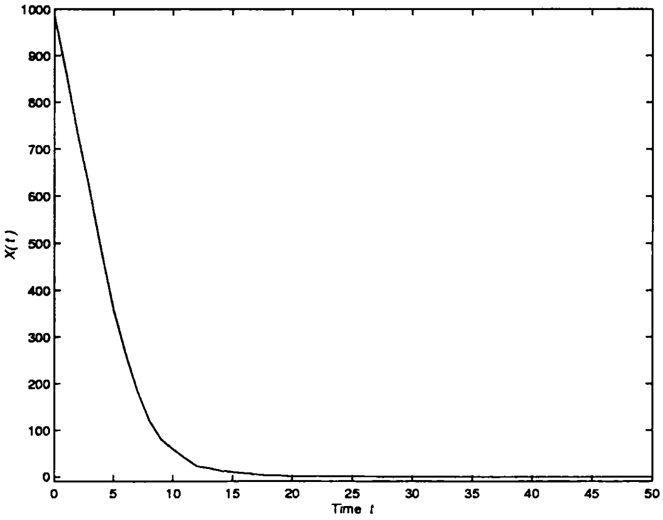


Figure 4.7: The value of $X(t)$ as obtained from an individual realisation of the stochastic model. The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$. Time is measured in years.

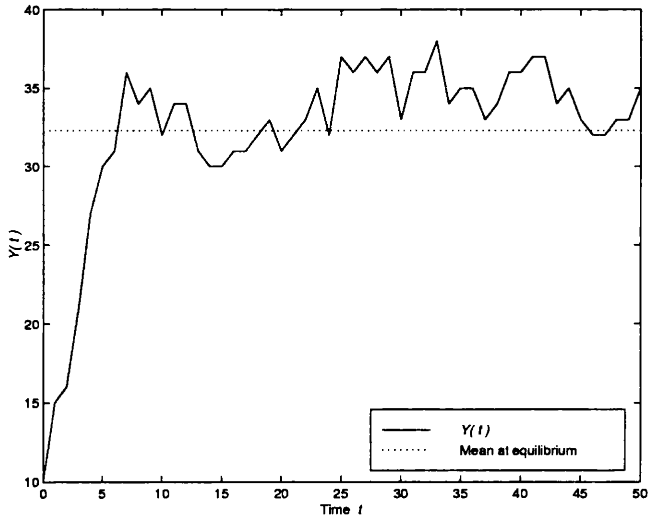


Figure 4.8: The value of $Y(t)$ as obtained from an individual realisation of the stochastic model. The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$. Time is measured in years.

Year	x	$\mathrm{E}_n[X]$	$\mathrm{E}_s[X]$	y	$\mathrm{E}_n[Y]$	$\mathrm{E}_s[Y]$
1	886.0	886.2	886.2	12.2	12.2	12.2
2	774.3	775.0	774.7	14.7	14.7	14.8
3	659.1	661.2	660.6	17.4	17.4	17.5
4	546.1	550.5	549.5	20.1	20.0	20.1
5	440.7	448.1	446.6	22.7	22.5	22.6
6	347.3	357.4	355.8	24.9	24.7	24.8
7	268.1	280.3	278.9	26.9	26.6	26.6
8	203.1	216.8	215.5	28.4	28.1	28.2
9	151.8	165.8	164.8	29.7	29.3	29.4
10	112.1	125.7	125.0	30.6	30.3	30.3
11	82.2	94.6	94.2	31.3	31.0	31.0
12	59.9	70.9	70.7	31.8	31.5	31.5
13	43.5	52.9	52.9	32.2	31.9	31.9
14	31.4	39.4	39.5	32.4	32.2	32.2
15	22.7	29.2	29.4	32.6	32.4	32.4

Table 4.1: The deterministic x, y and the stochastic means of X, Y as obtained from the normal approximation ($\mathrm{E}_n[X]$, $\mathrm{E}_n[Y]$) and from simulations ($\mathrm{E}_s[X]$, $\mathrm{E}_s[Y]$). The parameter values are as shown in (4.25) and $n = 1000$, $x_0 = 990$, $y_0 = 10$.

The results for $F_T(t)$ as obtained from the simulations are presented in Figure 4.6. After 20 years the probability $\mathrm{P}[T \leq t]$ increases very rapidly and by time $t = 62$ it becomes equal to one. At that point there were no susceptibles remaining uninfected in any of the 10000 simulations.

Figure 4.4 shows the results for the Y means. Initially the three curves are almost identical. In the interval between 5 and 25 years the value of the deterministic y deviates slightly from those of the stochastic means, which remain very close. By time $t = 50$, the three values are almost equal to the equilibrium mean $\beta n(\beta+\gamma)^{-1} = 32.3$ (which is equal to the deterministic equilibrium). From the single realisation presented in Figure 4.8 it can be observed that the value of $Y(t)$ increases rapidly during the first 7 years and then fluctuates around the equilibrium mean $\beta n(\beta + \gamma)^{-1}$.

The results for the Z means (Figure 4.5) are deduced from those for the X and the Y means. Initially the deterministic z deviates from the means of Z. After $t = 20$ the three curves remain very close and by time $t = 50$ they are around the deterministic equilibrium level $\gamma n(\beta + \gamma)^{-1} = 967.7$.

Chapter 5

The second model: a simple open model for TB

5.1 Introduction

This model is an extension of the model presented in the previous chapter. The population is divided into three classes: susceptibles, infectives, and inactive cases. The sizes of these classes are $X(t)$, $Y(t)$, $Z(t)$, respectively. The transitions between the classes and the relevant rates of these transitions are as in the first model: infections occur at a rate $\alpha XY/n$ (where n is the initial total population size); ρ is the probability that an infection is followed by a non-zero latent period; the common reactivation and relapse rate is β; and γ is the per capita recovery rate.

The difference here is that we also have immigration of new susceptibles (at a constant rate b), death from normal causes (at a rate μ per capita), and excess death of the infectives from TB (at a rate δ per capita). The total population size is $N(0) = n$ initially and $N(t) = X(t) + Y(t) + Z(t)$ at time t (not constant in time). At some points the special case $b = \mu n$ will be studied. The possible transitions and their rates are illustrated in Figure 5.1.

5.2 The Deterministic Model

For the corresponding deterministic model, let $x(t)$, $y(t)$, and $z(t)$ denote the number of susceptibles, infectives, and inactives, respectively, at time t. The differential equations

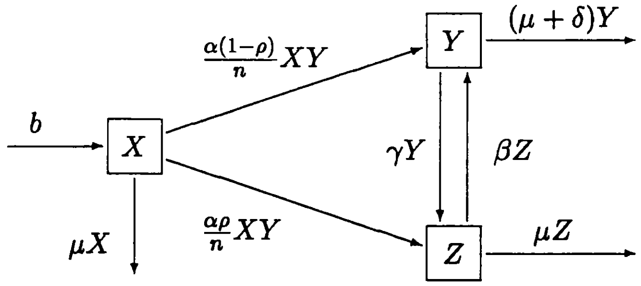


Figure 5.1: The second model for the spread of tuberculosis

for x, y, z are

$$\begin{aligned}\frac{dx}{dt} &= -\frac{\alpha}{n}xy - \mu x + b \\ \frac{dy}{dt} &= (1-\rho)\frac{\alpha}{n}xy - (\gamma+\mu+\delta)y + \beta z \\ \frac{dz}{dt} &= \rho\frac{\alpha}{n}xy + \gamma y - (\beta+\mu)z.\end{aligned} \tag{5.1}$$

Here x, y, and z are non-negative continuous variables. The initial conditions are $x(0) = x_0$, $y(0) = y_0$, $z(0) = z_0$, with $(x_0, y_0, z_0) \in \mathcal{S}_0$ where

$$\mathcal{S}_0 = \{(x,y,z) \in \mathbb{Z}^3_+ : 1 \le x \le n-1, x+y+z = n\}. \tag{5.2}$$

The infected individuals can be in two different states, infectious Y and non-infectious Z. Therefore, following the notation in Definition 3.1, the basic reproduction ratio, $\mathcal{R}_0$, is the dominant eigenvalue of the matrix $\mathsf{K} = -\mathsf{T}(\mathsf{S}-\mathsf{D})^{-1}$, where

$$\mathsf{S} = \begin{bmatrix} -\gamma & \beta \\ \gamma & -\beta \end{bmatrix}, \qquad \mathsf{D} = \begin{bmatrix} \mu+\delta & 0 \\ 0 & \mu \end{bmatrix}, \qquad \mathsf{T} = \begin{bmatrix} \alpha(1-\rho) & 0 \\ \alpha\rho & 0 \end{bmatrix},$$

and hence

$$\mathcal{R}_0 = \alpha\frac{\beta+(1-\rho)\mu}{(\beta+\mu)(\gamma+\delta+\mu)-\beta\gamma}. \tag{5.3}$$

The total population size is $N(t) = x(t) + y(t) + z(t)$. By adding the equations of system (5.1) the differential equation for $N(t)$ is deduced

$$\frac{dN(t)}{dt} = b - \mu N(t) - \delta y(t). \tag{5.4}$$

With integration, (5.4) gives

$$N(t) = \frac{b}{\mu} + e^{-\mu t}\left(n - \frac{b}{\mu}\right) - \delta e^{-\mu t}\int_0^t e^{\mu s}y(s)ds,$$

for $t \geq 0$. From this equation it follows that $N(t)$, and hence $x(t)$, $y(t)$, and $z(t)$ as well, are always bounded above by n if $b \leq \mu n$ and by b/μ if $b > \mu n$. If $b = \mu n$ then, in the absence of excess death due to TB, the population size remains constant in time and $N(t) = n$. This is the case applying, at least approximately, in many populations, so for the rest of this section we will assume that $b = \mu n$ (for simplicity of notation, sometimes the term b will still be used), so that whatever fluctuations in the population size are observed are caused by the excess TB death rate.

Solving the system (5.1) with $dx/dt = dy/dt = dz/dt = 0$, it follows that the system (5.1) admits two possible equilibria, $\mathbf{e}_1 = (n, 0, 0)$ and $\mathbf{e}_2 = (x_{\mathbf{e}_2}, y_{\mathbf{e}_2}, z_{\mathbf{e}_2})$, where

$$x_{\mathbf{e}_2} = \frac{n}{\alpha} \frac{(\beta + \mu)(\gamma + \delta + \mu) - \beta\gamma}{\beta + (1 - \rho)\mu} \tag{5.5a}$$

$$y_{\mathbf{e}_2} = \frac{n}{\alpha}\left(\frac{b}{x_{\mathbf{e}_2}} - \mu\right) \tag{5.5b}$$

$$z_{\mathbf{e}_2} = \frac{1}{\beta}\left[\gamma + \delta + \mu - \frac{\alpha(1-\rho)}{n} x_{\mathbf{e}_2}\right] y_{\mathbf{e}_2}. \tag{5.5c}$$

The two points $\mathbf{e}_1$ and $\mathbf{e}_2$ are equal if and only if $\mathcal{R}_0 = 1$.

Feasibility of $\mathbf{e}_1$ and $\mathbf{e}_2$

If N_e is the total population size at equilibrium, then (5.4) gives

$$0 = b - \mu N_e - \delta y_e,$$

which, for $N_e = x_e + y_e + z_e$, can be written as

$$b = \mu x_e + (\mu + \delta) y_e + \mu z_e. \tag{5.6}$$

It can easily be proved that both $\mathbf{e}_1$ and $\mathbf{e}_2$ satisfy (5.6). Since $b = \mu n$, from (5.6) it follows that N_e, and hence x_e, y_e, and z_e as well, are bounded above by n. In addition, these numbers must be non-negative. Therefore, we will call a critical point (x_e, y_e, z_e) feasible if $0 \leq x_e, y_e, z_e \leq n$.

Clearly $\mathbf{e}_1$ is feasible. That is not always the case for $\mathbf{e}_2$. From (5.5a), it follows that $x_{\mathbf{e}_2}$ is always well-defined and positive if the parameter values (α, β, γ, δ, μ, ρ, and $1 - \rho$) are not zero. Also $x_{\mathbf{e}_2} \leq n$ if $\mathcal{R}_0 \geq 1$, and then $y_{\mathbf{e}_2} \geq 0$. From (5.5b), it results that $y_{\mathbf{e}_2} \leq n$ if

$$x_{\mathbf{e}_2} \geq \frac{b}{\alpha + \mu}. \tag{5.7}$$

Finally, from (5.5c), it follows that, when $y_{\mathbf{e}_2} \geq 0$, $z_{\mathbf{e}_2}$ is non-negative if

$$\gamma + \delta + \mu - \frac{\alpha(1-\rho)}{n} x_{\mathbf{e}_2} \geq 0, \tag{5.8}$$

and, substituting $y_{\mathbf{e}_2}$ from (5.5b), that $z_{\mathbf{e}_2} \leq n$ if

$$\frac{1}{\alpha\beta}\left[\gamma + \delta + \mu - \frac{\alpha(1-\rho)}{n} x_{\mathbf{e}_2}\right]\left(\frac{b}{x_{\mathbf{e}_2}} - \mu\right) \leq 1. \tag{5.9}$$

After some calculations it can be shown that the conditions (5.7)–(5.9) hold for any set of positive parameters (α, β, γ, δ, μ, ρ, $1-\rho$) and therefore $\mathbf{e}_2$ is feasible if and only if $\mathcal{R}_0 \geq 1$.

Stability of $\mathbf{e}_1$ and $\mathbf{e}_2$

The system (5.1) can be written in the form

$$\frac{d\mathbf{v}}{dt} = \mathcal{F}(\mathbf{v}), \tag{5.10}$$

where $\mathbf{v}(t) = (x(t), y(t), z(t))$ for $t \geq 0$ and $\mathcal{F}$ is a mapping from $\mathbb{R}^3_+$ into $\mathbb{R}^3$ with coordinates $f_i(x, y, z)$, $i = 1, 2, 3$ given by

$$\begin{aligned} f_1(x,y,z) &= -\frac{\alpha}{n}xy - \mu x + b \\ f_2(x,y,z) &= \frac{\alpha(1-\rho)}{n}xy - (\gamma+\delta+\mu)y + \beta z \\ f_3(x,y,z) &= \frac{\alpha\rho}{n}xy + \gamma y - (\beta+\mu)z. \end{aligned}$$

Let $\mathbf{v}^*$ be an equilibrium point of (5.10), so that $\mathcal{F}(\mathbf{v}^*) = 0$. Let $D\mathcal{F}(\mathbf{v}^*)$ be the Jacobian matrix of $\mathcal{F}$ at the point $\mathbf{v}^*$, i.e. the matrix whose (i, j) element is $\partial f_i(\mathbf{v}^*)/\partial j$, for $i = 1, 2, 3$ and $j = x, y, z$. If all eigenvalues of $D\mathcal{F}(\mathbf{v}^*)$ have negative real parts, then $\mathbf{v}^*$ is uniformly asymptotically stable (see, e.g., Reinhard 1986, Chapters 2, 3). For $\mathbf{v}^* = \mathbf{e}_1$, the Jacobian matrix of $\mathcal{F}$ at $\mathbf{e}_1$ is

$$D\mathcal{F}(\mathbf{e}_1) = \begin{bmatrix} -\mu & -\alpha & 0 \\ 0 & \alpha(1-\rho) - (\gamma+\delta+\mu) & \beta \\ 0 & \alpha\rho + \gamma & -(\beta+\mu) \end{bmatrix}.$$

One can easily compute the eigenvalues of $D\mathcal{F}(\mathbf{e}_1)$ and prove that if $\mathcal{R}_0 > 1$ then at least one of them is positive, but if $\mathcal{R}_0 < 1$ then they are all negative (where $\mathcal{R}_0$ is as defined in (5.3)). Therefore $\mathbf{e}_1$ is stable if $\mathcal{R}_0 < 1$ and unstable if $\mathcal{R}_0 > 1$. In order to study the stability of $\mathbf{e}_2$, we will use the following criterion:

Theorem 5.1 *(Routh-Hurwitz criterion)*

Suppose that the vector $\mathbf{v}(t) = (v_1(t), \dots, v_m(t))$ defined in $\mathbb{R}^m_+$ for $t \geq 0$ satisfies the system

$$\frac{d\mathbf{v}}{dt} = \mathcal{F}(\mathbf{v}), \tag{5.11}$$

where $\mathcal{F}$ is a mapping from $\mathbb{R}^m_+$ into $\mathbb{R}^m$ with coordinates $f_i(\mathbf{v}) = f_i(v_1, \dots, v_m) = dv_i/dt$ for $i = 1, \dots, m$. Suppose that $\mathbf{v}^*$ is an equilibrium point of (5.11), so that $\mathcal{F}(\mathbf{v}^*) = 0$. Let $D\mathcal{F}(\mathbf{v}^*)$ be the Jacobian matrix of $\mathcal{F}$ at the point $\mathbf{v}^*$, i.e. the matrix whose (i,j) element is $\partial f_i(\mathbf{v}^*)/\partial v_j$ for $i, j = 1, \dots, m$. Let $P_{\mathbf{v}^*}(\tau) = \det(D\mathcal{F}(\mathbf{v}^*) - \tau\mathbf{I})$ denote the characteristic polynomial of $D\mathcal{F}(\mathbf{v}^*)$

$$P_{\mathbf{v}^*}(\tau) = \alpha_0\tau^m + \alpha_1\tau^{m-1} + \dots + \alpha_{m-1}\tau + \alpha_m,$$

and define a matrix H as follows

$$H = \begin{bmatrix} \alpha_1 & \alpha_3 & \alpha_5 & \cdots & \cdots & \alpha_{2m-1} \\ \alpha_0 & \alpha_2 & \alpha_4 & \cdots & \cdots & \alpha_{2m-2} \\ 0 & \alpha_1 & \alpha_3 & & & \vdots \\ 0 & \alpha_0 & \alpha_2 & \ddots & & \\ 0 & 0 & \alpha_1 & & & \\ 0 & 0 & \alpha_0 & & & \\ 0 & 0 & 0 & & & \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \cdots & \alpha_m \end{bmatrix},$$

where $\alpha_0, \alpha_1, \dots, \alpha_m$ are the coefficients of $P_{\mathbf{v}^*}(\tau)$ and $\alpha_{m+j} = 0$ for $j = 1, \dots, m-1$. If $\alpha_0 > 0$ then all the eigenvalues of $D\mathcal{F}(\mathbf{v}^*)$ have negative real parts if and only if all the principal minors of H are strictly positive.

For $\mathbf{v}^* = \mathbf{e}_2$ the Jacobian matrix of $\mathcal{F}$ at $\mathbf{e}_2$ is

$$D\mathcal{F}(\mathbf{e}_2) = \begin{bmatrix} -\frac{b}{x_{\mathbf{e}_2}} & -\frac{\alpha}{n}x_{\mathbf{e}_2} & 0 \\ (1-\rho)\left(\frac{b}{x_{\mathbf{e}_2}} - \mu\right) & \frac{\alpha(1-\rho)}{n}x_{\mathbf{e}_2} - (\gamma + \delta + \mu) & \beta \\ \rho\left(\frac{b}{x_{\mathbf{e}_2}} - \mu\right) & \frac{\alpha\rho}{n}x_{\mathbf{e}_2} + \gamma & -(\beta + \mu) \end{bmatrix},$$

with characteristic polynomial $P_{\mathbf{e}_2}(\tau) = -(\alpha_0\tau^3 + \alpha_1\tau^2 + \alpha_2\tau + \alpha_3)$, where $\alpha_0 = 1$ and

$$\begin{aligned}
\alpha_1 &= \beta + \gamma + \delta + 2\mu + \frac{b}{x_{\mathbf{e}_2}} - \frac{\alpha(1-\rho)}{n}x_{\mathbf{e}_2} \\
\alpha_2 &= -\frac{\alpha}{n}x_{\mathbf{e}_2}[\beta + 2(1-\rho)\mu] + [-\beta\gamma + (\beta+\mu)(\gamma+\delta+\mu)] + \frac{b}{x_{\mathbf{e}_2}}(\beta+\gamma+\delta+2\mu) \\
\alpha_3 &= \frac{b}{x_{\mathbf{e}_2}}[-\beta\gamma + (\beta+\mu)(\gamma+\delta+\mu)] - \frac{\alpha}{n}(\beta + (1-\rho)\mu)\mu x_{\mathbf{e}_2}.
\end{aligned}$$

The matrix H in this case is

$$H = \begin{bmatrix} \alpha_1 & \alpha_3 & 0 \\ \alpha_0 & \alpha_2 & 0 \\ 0 & \alpha_1 & \alpha_3 \end{bmatrix},$$

and its principal minors are $D_1 = \alpha_1$, $D_2 = \alpha_1\alpha_2 - \alpha_3$, and $D_3 = \alpha_3 D_2$. After some calculations it can be shown that if $\mathcal{R}_0 > 1$ and all the parameters (α, β, γ, δ, μ, ρ, $1-\rho$) are positive then D_1, D_2, D_3 are positive and hence every eigenvalue of $D\mathcal{F}(\mathbf{e}_2)$ has negative real part, which implies that $\mathbf{e}_2$ is stable.

Summarising the results above, the system (5.1) admits two equilibria, the disease-free equilibrium $\mathbf{e}_1$ and an endemic equilibrium $\mathbf{e}_2$. If $\mathcal{R}_0 < 1$ then only $\mathbf{e}_1$ is feasible and it is asymptotically stable, so that the epidemic eventually dies out. If $\mathcal{R}_0 > 1$ then both $\mathbf{e}_1$ and $\mathbf{e}_2$ are feasible, but $\mathbf{e}_1$ is unstable and $\mathbf{e}_2$ stable, so the epidemic settles down at an endemic level.

It should be noted that this study was carried out with the assumption that all the parameters of the model (namely, α, β, γ, δ, μ, ρ, and $1-\rho$) are non-zero. The reason for making this assumption is the fact that for each of the quantities whose signs needed to be strictly positive or negative (for instance, the principal minors of H), a different subset of the set of parameters (α, β, γ, δ, μ, ρ, and $1-\rho$) must have strictly positive entries. In order to avoid complicated assumptions and, instead, have an assumption that applies throughout the whole study of this model (both the deterministic and the stochastic) we exclude the possibility of having any of the parameters equal to zero. Simpler special cases of this model resulting by setting some parameters equal to zero (such as the SIR model, for $\beta = \delta = \mu = \rho = 0$), can be investigated separately or by taking the limits of the full solutions as the respective parameters tend to zero, and some special cases have already been extensively studied in the literature (see, e.g., Bailey 1975, Lefèvre 1990).

5.3 The Stochastic Model

5.3.1 The transient behaviour

Let $p_{rsv}(t)$ be the probability that there are r susceptibles, s infectives, and v inactive cases in the system at time $t \geq 0$. Initially there are x_0 susceptibles, y_0 infectives, and z_0 inactive cases, where $(x_0, y_0, z_0) \in \mathcal{S}_0$ as defined in (5.2), so that $p_{x_0y_0z_0}(0) = 1$ and $p_{rsv}(0) = 0$ for any $(r, s, v) \neq (x_0, y_0, z_0)$. The differential equation for the probabilities $p_{rsv}(t)$ is

$$\begin{aligned}\frac{dp_{rsv}(t)}{dt} &= \beta[(v+1)p_{r,s-1,v+1} - vp_{rsv}] + \gamma[(s+1)p_{r,s+1,v-1} - sp_{rsv}] \\ &\quad + \frac{\alpha}{n}(1-\rho)(r+1)(s-1)p_{r+1,s-1,v} + \frac{\alpha}{n}\rho(r+1)sp_{r+1,s,v-1} \\ &\quad - \frac{\alpha}{n}rsp_{rsv} + b(p_{r-1,s,v} - p_{rsv}) + \mu((r+1)p_{r+1,s,v} - rp_{rsv}) \\ &\quad + (\mu+\delta)((s+1)p_{r,s+1,v} - sp_{rsv}) + \mu((v+1)p_{r,s,v+1} - vp_{rsv}),\end{aligned} \tag{5.12}$$

for $(r, s, v) \in \mathbb{Z}_+^3$ and $p_{rsv}(t) = 0$ otherwise. The probability generating function, defined as $\mathcal{P}(x, y, z; t) = \mathrm{E}[x^{X(t)}y^{Y(t)}z^{Z(t)}]$, satisfies the equation

$$\begin{aligned}\frac{\partial\mathcal{P}}{\partial t} &= b(x-1)\mathcal{P} - \mu(x-1)\frac{\partial\mathcal{P}}{\partial x} + [\gamma(z-y) + (\mu+\delta)(1-y)]\frac{\partial\mathcal{P}}{\partial y} \\ &\quad + [\beta(y-z) + \mu(1-z)]\frac{\partial\mathcal{P}}{\partial z} + \frac{\alpha}{n}y[(1-\rho)y + \rho z - x]\frac{\partial^2\mathcal{P}}{\partial x\partial y},\end{aligned} \tag{5.13}$$

with initial condition $\mathcal{P}(x, y, z; 0) = x^{x_0}y^{y_0}z^{z_0}$.

From (5.13) a system of differential equations for the first and second moments of X, Y, and Z can be deduced; the equations for the means are the following:

$$\begin{aligned}\frac{dm_X}{dt} &= -\frac{\alpha}{n}(\sigma_{XY} + m_Xm_Y) - \mu m_X + b \\ \frac{dm_Y}{dt} &= \frac{\alpha}{n}(1-\rho)(\sigma_{XY} + m_Xm_Y) - (\gamma+\mu+\delta)m_Y + \beta m_Z \\ \frac{dm_Z}{dt} &= \frac{\alpha}{n}\rho(\sigma_{XY} + m_Xm_Y) + \gamma m_Y - (\beta+\mu)m_Z,\end{aligned} \tag{5.14}$$

where $m_W(t)$ is the expected value of $W(t)$, for $W = X, Y, Z$, and $\sigma_{XY}(t)$ is the covariance of $X(t)$ and $Y(t)$. The equations for the variances and covariances are given in the Appendix (Section A.2.1).

The system of differential equations for the first and second moments of X, Y, Z involves third order moments and hence it is open and cannot be solved. One way of overcoming this problem is to express the third order moments in terms of the first and second moments. For example, if $(X, Y, Z)'$ has a multivariate normal distribution, then

$$\mathrm{E}[XYZ] = \mathrm{E}[X]\mathrm{E}[Y]\mathrm{E}[Z] + \mathrm{E}[X]\mathrm{Cov}[Y, Z] + \mathrm{E}[Y]\mathrm{Cov}[X, Z] + \mathrm{E}[Z]\mathrm{Cov}[X, Y], \quad (5.15)$$

with similar expressions for $\mathrm{E}[X^2Y]$, $\mathrm{E}[XY^2]$ etc. Substituting for the third order moments from these expressions, makes the system for the first and second moments closed and hence it can be solved. Distributions other than the normal can also be used, for instance the Negative Binomial (see, e.g., Herbert 1998, Herbert & Isham 2000). In addition, it has been observed that there may be situations where it is unreasonable to assume that the vector $(X, Y, Z)'$ has a multivariate normal distribution, and yet the moments of X, Y, Z can be very well approximated by those of a multivariate normal using the formulae like (5.15) (see, e.g., Herbert 1998). When the actual distribution is not known, one way of assessing which type of approximation is more appropriate is via simulations of the stochastic model; for instance, if the distribution has a single spike at the mean, then the normal approximation could be more appropriate, while if the distribution is highly skewed then the negative binomial might be a more appropriate choice.

The validity of the normal approximation can be established by Kurtz's theory (see Theorem 4.1). If the excess death rate δ due to TB is zero then the mean population size remains constant and equal to n. Therefore if δ has a small positive value then there will be only small fluctuations of the total population size $N(t)$ around n (at least in the relatively short term) and hence we can assume that n and $N(t)$ will be of the same order of magnitude. In this case the parameter N in Kurtz's notation can be taken as the initial total population size n.

With the notation in Theorem 4.1, the function $F : \mathbb{R}^3_+ \to \mathbb{R}^3$ is defined by

$$F(\mathbf{V}) = \begin{bmatrix} -\alpha V_1 V_2 + \mu - \mu V_1 \\ \alpha(1-\rho) V_1 V_2 - (\gamma + \delta + \mu) V_2 + \beta V_3 \\ \alpha\rho V_1 V_2 + \gamma V_2 - (\beta + \mu) V_3, \end{bmatrix}$$

where $\mathbf{V} = (V_1, V_2, V_3)$. It can easily be proved that the conditions of Kurtz's theorem hold and hence the process can be approximated by a Gaussian diffusion process if n is large.

From system (5.14) the differential equation for the expected value $m_N(t)$ of the total population size at time t is deduced

$$\frac{dm_N}{dt} = b - \mu m_N - \delta m_Y = \mu(n - m_N) - \delta m_Y, \tag{5.16}$$

which shows that initially the population decreases, since $m_N(0) = n$. With integration, (5.16) gives

$$m_N(t) = n - \delta e^{-\mu t} \int_0^t m_Y(w) e^{\mu w} dw.$$

This means that $m_N(t) \leq n$, for $t \geq 0$, and therefore $m_X(t)$, $m_Y(t)$, and $m_Z(t)$ are always less than or equal to n, too.

5.3.2 The equilibrium state of the process

The process described in this chapter is a Markov process in continuous time with countable state space $\mathcal{S} = \mathbb{Z}_+^3$. Let $\mathcal{A}$ denote the subset of $\mathcal{S}$ that contains the states of the form $(x, 0, 0)$ for $x \geq 0$, and $\mathcal{D}$ the remaining set of states

$$\mathcal{A} = \{(x, 0, 0) \in \mathbb{Z}_+^3\}$$

$$\mathcal{D} = \mathcal{S} - \mathcal{A} = \{(x, y, z) \in \mathbb{Z}_+^3 : (y, z) \neq (0, 0)\}.$$

It can easily be seen that all states in $\mathcal{A}$ communicate with each other, all states in $\mathcal{D}$ communicate with each other, but there can be no transitions from the set $\mathcal{A}$ to the set $\mathcal{D}$; once the chain reaches one of the states in $\mathcal{A}$ there are no infected individuals in the population and hence there cannot be any more infections and the chain will remain within the set $\mathcal{A}$. On the other hand there is always a positive probability for transitions from $\mathcal{D}$ to $\mathcal{A}$. Therefore the sets $\mathcal{A}$ and $\mathcal{D}$ form two irreducible classes, the former is a closed absorbing class and the latter is an open and hence transient class.

The fact that a subset of the state space $\mathcal{S}$ is absorbing does not in general mean that the chain will be absorbed in this subset. Nevertheless we will show that this is the case for this process using the theorems of Reuter (1961), who is concerned with a particular class of Markov processes which he calls *competition processes.* These are processes in continuous time with a countable set of states $\mathcal{S} = \mathbb{Z}_+^2$, having the property that jumps from a state $(m, n) \in \mathcal{S}$ always lead to one of the following adjacent states: $(m \pm 1, n)$, $(m, n \pm 1)$, $(m - 1, n + 1)$, $(m + 1, n - 1)$. The boundaries $m = 0$ and $n = 0$ of the positive (m, n)-quadrant cannot be crossed, and all the states $(m, 0)$ and $(0, n)$

are absorbing. For such a process, Reuter gives sufficient conditions for the chain to be absorbed and, if that happens with probability one, conditions for the expected time until absorption to be finite.

In the following we define a class of Markov processes with state space $\mathbb{Z}_+^m$ for $m \geq 2$ and absorbing states $(x, 0, \dots, 0)$ and Theorem 5.3 gives an extended form of Reuter's results for these processes.

Definition 5.2 *Consider a time-homogeneous Markov process* $\{\mathbf{X}(t), t \geq 0\}$*, with continuous time parameter* t *and state space* $\mathcal{S} = \{(x_1, x_2, \dots, x_m) \in \mathbb{Z}_+^m\}$*, for* $m \geq 2$*. For* $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ *define the transition matrix* $P(t) = \{p(\mathbf{x}, \mathbf{y}; t), \mathbf{x}, \mathbf{y} \in \mathcal{S}\}$*, where*

$$p(\mathbf{x}, \mathbf{y}; t) = \mathrm{P}[\mathbf{X}(t) = \mathbf{y} | \mathbf{X}(0) = \mathbf{x}].$$

The transition rates are $q(\mathbf{x}, \mathbf{y}) = p'(\mathbf{x}, \mathbf{y}; 0)$*, subject to the conditions*

$$\begin{gathered} q(\mathbf{x}, \mathbf{y}) \geq 0 \quad \textit{for } \mathbf{x} \neq \mathbf{y} \\ 0 \leq -q(\mathbf{x}, \mathbf{x}) \equiv q(\mathbf{x}) = \sum_{\mathbf{y} \neq \mathbf{x}} q(\mathbf{x}, \mathbf{y}) < \infty. \end{gathered}$$

Define the matrix $Q = \{q(\mathbf{x}, \mathbf{y}), \mathbf{x}, \mathbf{y} \in \mathcal{S}\}$*. At least one such matrix exists; if there is exactly one, the matrix* Q *is called regular (thus regularity means that the matrix* Q *defines the process uniquely - see also Reuter 1961, Wolff 1989). For* $\mathbf{x} = (x_1, \dots, x_m)$ *and* $\mathbf{y} = (y_1, \dots, y_m)$ *in* $\mathcal{S}$*, the* $q(\mathbf{x}, \mathbf{y})$ *are defined as follows:*

$$q(\mathbf{x}, \mathbf{y}) = \begin{cases} a_i(\mathbf{x}) & \textit{if } \mathbf{y} = \mathbf{x} + \mathbf{e}_i & i = 1, \dots, m \\ d_i(\mathbf{x}) & \textit{if } \mathbf{y} = \mathbf{x} - \mathbf{e}_i & i = 1, \dots, m \\ e_{ij}(\mathbf{x}) & \textit{if } \mathbf{y} = \mathbf{x} - \mathbf{e}_i + \mathbf{e}_j & i, j = 1, \dots, m, j \neq i \\ 0 & \textit{any other } \mathbf{y} \neq \mathbf{x}, & \end{cases}$$

where $\mathbf{e}_i$ *is the vector of* $\mathbb{Z}_+^m$ *whose* i*-th coordinate is equal to one and all the other coordinates are zero. The functions* a_i*,* d_i*,* e_{ij} *are such that* $a_i(\mathbf{x}) \geq 0$*,* $d_i(\mathbf{x}) \geq 0$*,* $e_{ij}(\mathbf{x}) \geq 0$*, for all* $i, j = 1, \dots, m$ *with* $j \neq i$ *and* $\mathbf{x} \in \mathcal{S}$*, and they are equal to zero if* $\mathbf{x} \notin \mathcal{S}$*. Finally the* $q(\mathbf{x}, \mathbf{x})$ *are defined by*

$$-q(\mathbf{x}, \mathbf{x}) = q(\mathbf{x}) = \sum_{i=1}^{m} [a_i(\mathbf{x}) + d_i(\mathbf{x})] + \sum_{i=1}^{m} \sum_{j \neq i} e_{ij}(\mathbf{x}), \mathbf{x} \in \mathcal{S}.$$

Since there are no states with negative coordinates, we must also have

$$\begin{aligned}
d_1(0,x_2,x_3,\dots,x_m) &= e_{1j}(0,x_2,x_3,\dots,x_m)=0 \qquad \forall j\neq 1\\
d_2(x_1,0,x_3,\dots,x_m) &= e_{2j}(x_1,0,x_3,\dots,x_m)=0 \qquad \forall j\neq 2\\
&\vdots\\
d_m(x_1,x_2,x_3,\dots,0) &= e_{mj}(x_1,x_2,x_3,\dots,0)=0 \qquad \forall j\neq m,
\end{aligned} \tag{5.17}$$

for all $x_1,x_2,\dots,x_m=0,1,\dots$ *Therefore jumps from the state* $\mathbf{x}=(x_1,\dots,x_m)\in\mathcal{S}$ *always lead to one of the adjacent states* $\mathbf{x}\pm\mathbf{e}_i$, $\mathbf{x}-\mathbf{e}_i+\mathbf{e}_j$, *for* $i,j=1,\dots,m$ *and* $j\neq i$, *but the boundaries* $x_1=0,x_2=0,\dots,x_m=0$ *cannot be crossed.*

Assume that the states $(x_1,0,\dots,0)$ *are absorbing, so that* $a_i(x_1,0,\dots,0)=0$, $d_i(x_1,0,\dots,0)=0$, $e_{ij}(x_1,0,\dots,0)=0$, *for all* $x_1\in\mathbb{Z}_+$, $i,j=1,\dots,m$, *and* $j\neq i$, *but*

$$\sum_{i=1}^{m} a_i(\mathbf{x})>0 \quad \textit{and} \quad \sum_{i=1}^{m} d_i(\mathbf{x})>0,$$

whenever $\mathbf{x}\in\mathcal{D}=\mathcal{S}-\mathcal{A}$, *where*

$$\begin{aligned}
&\mathcal{A}=\{(x,0,\dots,0)\in\mathbb{Z}_+^m\}\\
&\mathcal{D}=\mathcal{S}-\mathcal{A}=\{(x_1,x_2,\dots,x_m)\in\mathbb{Z}_+^m:(x_2,\dots,x_m)\neq(0,\dots,0)\}.
\end{aligned}$$

Finally, for every $k=1,2,\dots$ *define*

$$\begin{aligned}
\mathcal{A}_k &= \{(x_1,\dots,x_m)\in\mathcal{D}:x_1+\cdots+x_m=k\}\\
r_k &= \max_{\mathbf{x}\in\mathcal{A}_k}\sum_{i=1}^{m} a_i(\mathbf{x})\\
s_k &= \min_{\mathbf{x}\in\mathcal{A}_k}\sum_{i=1}^{m} d_i(\mathbf{x}).
\end{aligned}$$

Theorem 5.3 *Consider a Markov process as defined in Definition 5.2. The following three statements hold:*

(a) A sufficient condition for regularity is

$$S_1=\frac{1}{r_2}+\sum_{k=3}^{\infty}\left(\frac{1}{r_k}+\frac{s_k}{r_kr_{k-1}}+\cdots+\frac{s_k\dots s_3}{r_k\dots r_2}\right)=\infty.$$

(b) For a regular process let $\pi(\mathbf{x},\mathbf{y})=lim_{t\to\infty}p(\mathbf{x},\mathbf{y};t)$ *for* $\mathbf{x},\mathbf{y}\in\mathcal{S}$. *Then* $\pi(\mathbf{x},\mathbf{y})=1$ *if* $\mathbf{x}=\mathbf{y}\in\mathcal{A}$, $\pi(\mathbf{x},\mathbf{y})=0$ *if* $\mathbf{y}\in\mathcal{D}$, *and* $\alpha_{\mathbf{x}}=\sum_{\mathbf{y}}\pi(\mathbf{x},\mathbf{y})$ *is the probability that some absorbing state is reached from* $\mathbf{x}\in\mathcal{D}$. *Also,* $\alpha_{\mathbf{x}}=1$ *for all* $\mathbf{x}\in\mathcal{D}$ *if*

$$S_2=\sum_{k=1}^{\infty}\frac{s_1\dots s_k}{r_1\dots r_k}=\infty.$$

(c) If in statement (b) $\alpha_{\mathbf{x}} = 1$ for all $\mathbf{x} \in \mathcal{D}$, let $\tau_{\mathbf{x}}$ be the mean time to reach $\mathcal{A}$, starting at $\mathbf{x} \in \mathcal{D}$. Then $\tau_{\mathbf{x}} < \infty$ for all $\mathbf{x} \in \mathcal{D}$ if

$$S_3 = \frac{1}{s_1} + \sum_{k=2}^{\infty} \frac{r_1 r_2 \ldots r_{k-1}}{s_1 s_2 \ldots s_k} < \infty.$$

The proof of this theorem is almost identical to Reuter's proof (see Reuter 1961) so we present only a brief outline here. To prove statement (a) we use the following criterion (see, e.g., Wolff 1989, Chapter 4):

(A) *$Q = \{q(\mathbf{x}, \mathbf{y})\}$ is regular if, for each $\lambda > 0$, the equations*

$$\lambda w_{\mathbf{x}} = \sum_{\mathbf{y}} q(\mathbf{x}, \mathbf{y}) w_{\mathbf{y}},$$

where $0 \leq w_{\mathbf{x}} \leq 1$, have only the trivial solution $\{w_{\mathbf{x}} = 0$, for all $\mathbf{x}\}$.

Suppose that a non-trivial solution $\{w_{\mathbf{x}}\}$ exists with $0 \leq w_{\mathbf{x}} \leq 1$. For $k = 1, 2, \ldots$ let W_k be the maximum of $w_{\mathbf{x}}$ over all $\mathbf{x} = (x_1, \ldots, x_m)$ such that $x_1 + \cdots + x_m = k$ and $(x_2, \ldots, x_m) \neq (0, \ldots, 0)$. Then, it can be shown that if $S_1 = \infty$ the series $\sum_{k=1}^{\infty} (W_{k+1} - W_k)$ diverges and hence W_k tends to infinity as $k \to \infty$, which contradicts the initial assumption that $0 \leq w_{\mathbf{x}} \leq 1$ for all $\mathbf{x}$. Hence $w_{\mathbf{x}} \equiv 0$ and the process is regular.

To prove (b) we will use the following criterion (see Reuter (1961) for a proof):

(B) *Suppose Q is regular. Then $\sum_{\mathbf{y}} \pi(\mathbf{x}, \mathbf{y}) = 1$ for all $\mathbf{x}$, if there exist $u_{\mathbf{x}} \geq 0$ such that $u_{\mathbf{x}} \to \infty$ as $\mathbf{x} \to \infty$ and*

$$\sum_{\mathbf{y}} q(\mathbf{x}, \mathbf{y}) u_{\mathbf{y}} \leq 0, \quad \textit{for all } \mathbf{x}. \tag{5.18}$$

Define $U_0 = 0$, $U_1 = 1$, and $U_{k+1} = U_k + (s_1 s_2 \ldots s_k)/(r_1 r_2 \ldots r_k)$, for $k = 1, 2, \ldots$. Clearly, $U_k \geq 0$ for all $k = 0, 1, \ldots$ From the assumption that $S_2 = \infty$, it follows that $\lim_{k \to \infty} U_k = \infty$. Finally, it can be shown that the sequence $\{u_{\mathbf{x}}\}$ with $u_{\mathbf{x}} = U_{x_1 + \cdots + x_m}$ for $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{S}$ satisfies (5.18). Therefore, from criterion (B) it follows that $\sum_{\mathbf{y}} \pi(\mathbf{x}, \mathbf{y}) = 1$ for all $\mathbf{x}$.

For statement (c) we use the following criterion (see Reuter (1961) for a proof):

(C) *Suppose that Q is regular. Let $\mathcal{A}$ and $\mathcal{D}$ denote the sets of absorbing and non-absorbing states, respectively. Let $\tau_{\mathbf{x}}$ be the expected time to reach $\mathcal{A}$, starting from the state $\mathbf{x}$. If there exist finite $u_{\mathbf{y}} \geq 0$ such that*

$$\sum_{\mathbf{y}} q(\mathbf{x}, \mathbf{y}) u_{\mathbf{y}} + 1 \leq 0, \quad \textit{for all } \mathbf{x} \textit{ in } \mathcal{D}, \tag{5.19}$$

then $\tau_{\mathbf{x}} \leq u_{\mathbf{x}} < \infty$ *for all* $\mathbf{x}$ *in* $\mathcal{D}$.

Let $\{V_k\}$ be the sequence defined by

$$V_k = \frac{s_1 s_2 \dots s_k}{r_1 r_2 \dots r_k}\left[V_0 - \left(\frac{1}{s_1} + \frac{r_1}{s_1 s_2} + \dots + \frac{r_1 r_2 \dots r_{k-1}}{s_1 s_2 \dots s_k}\right)\right],$$

for $k = 2, 3, \dots$ and $V_1 = (s_1/r_1)(V_0 - 1/s_1)$. Since $S_3 < \infty$, the sequence

$$\alpha_k = \frac{1}{s_1} + \frac{r_1}{s_1 s_2} + \dots + \frac{r_1 r_2 \dots r_{k-1}}{s_1 s_2 \dots s_k}$$

is bounded. Hence, we can choose V_0 finite and positive but sufficiently large, such that $V_k \geq 0$ for any $k = 1, 2, \dots$. Now, let $\{u_{\mathbf{x}}\}$ be the sequence defined by $u_{\mathbf{x}} = U_k$ for $\mathbf{x} = (x_1, \dots, x_m) \in \mathcal{D}$, $k = x_1 + \dots + x_m$ and $U_{k+1} = U_k + V_k$ for $k = 0, 1, \dots$. Choosing $U_0 \geq 0$, it can be shown that $\{u_{\mathbf{x}}\}$ satisfies (5.19) and thus $\tau_{\mathbf{x}} < \infty$ for all $\mathbf{x} \in \mathcal{D}$.

Criterion (C) also provides a means to find upper bounds for the expected time until absorption, which we summarise in the following lemma.

Lemma 5.4 *Consider a Markov process as defined in Definition 5.2. Let* α_k *be the sequence*

$$\alpha_k = \frac{1}{s_1} + \frac{r_1}{s_1 s_2} + \dots + \frac{r_1 r_2 \dots r_{k-1}}{s_1 s_2 \dots s_k}, \quad \textit{for } k \geq 2,$$

with $\alpha_1 = 1/s_1$. *If*

$$S_3 = \frac{1}{s_1} + \sum_{k=2}^{\infty} \frac{r_1 r_2 \dots r_{k-1}}{s_1 s_2 \dots s_k} < \infty,$$

then there exists $M > 0$ *such that* $0 < \alpha_k < M < \infty$ *for all* $k \geq 1$. *Choose* $U_0 \geq 0$ *and* $V_0 \geq M > 0$ *and define*

$$V_k = \frac{s_1 s_2 \dots s_k}{r_1 r_2 \dots r_k}(V_0 - \alpha_k), \quad k \geq 1,$$

and $U_{k+1} = V_k + U_k$, *for* $k \geq 0$. *Then the expected time until absorption,* $\tau_{\mathbf{x}}$, *starting from the state* $\mathbf{x} = (x_1, \dots, x_m)$, *with* $(x_2, \dots, x_m) \neq (0, \dots, 0)$ *and* $x_1 + \dots + x_m = k \geq 1$ *is bounded above by* $U_k : \tau_{\mathbf{x}} \leq U_k$.

For the process described in this chapter, the functions a_j, d_j, e_{ij} are:

$$\begin{aligned}
a_1(x,y,z) &= b & e_{12}(x,y,z) &= \alpha(1-\rho)xy/n \\
a_2(x,y,z) &= 0 & e_{13}(x,y,z) &= \alpha\rho xy/n \\
a_3(x,y,z) &= 0 & e_{21}(x,y,z) &= 0 \\
d_1(x,y,z) &= \mu x & e_{23}(x,y,z) &= \gamma y \\
d_2(x,y,z) &= (\mu+\delta)y & e_{31}(x,y,z) &= 0 \\
d_3(x,y,z) &= \mu z & e_{32}(x,y,z) &= \beta z,
\end{aligned}$$

for $(x, y, z) \in \mathcal{S} = \mathbb{Z}_+^3$, which satisfy the conditions (5.17). Also it should be noted that the states $(x, 0, 0)$ are not absorbing; but the set $\mathcal{A} = \{(x, 0, 0) : x = 0, 1, \dots\}$ is absorbing, in the sense that once (y, z) becomes $(0, 0)$ then the infection dies out and the population of susceptibles grows as a simple birth-and-death process. The device of "freezing" the states $(x, 0, 0)$ can be adopted in this case, since the evolution of the process after the extinction of the infection does not affect the evolution before the extinction. Therefore, we will assume that $a_1(x, 0, 0) = d_1(x, 0, 0) = 0$, for any $x = 0, 1, \dots$ and thus the states $(x, 0, 0)$ are absorbing.

Under these assumptions (and using the notation in Definition 5.2) one easily computes that $r_k = b$ and $s_k = \mu k$ for $k \geq 1$. Hence from the statements (a), (b), and (c) of Theorem 5.3 it follows that:

- $S_1 = \frac{1}{b} + \sum_{k=3}^{\infty} \left(\frac{1}{b} + \frac{k\mu}{b^2} + \frac{k(k-1)\mu^2}{b^3} + \cdots + \frac{k(k-1)\dots 3\mu^{k-2}}{b^{k-1}} \right) = \infty,$
  and hence the process is regular.
- $S_2 = \sum_{k=1}^{\infty} k! \left(\frac{\mu}{b} \right)^k = \infty,$
  hence, if $\mathbf{X}(t)$ is the vector $(X(t), Y(t), Z(t))$ and
  $$\pi(\mathbf{x}, \mathbf{y}) = \lim_{t \to \infty} p(\mathbf{x}, \mathbf{y}; t) = \lim_{t \to \infty} \mathrm{P}[\mathbf{X}(t) = \mathbf{y} | \mathbf{X}(0) = \mathbf{x}], \quad \mathbf{x}, \mathbf{y} \text{ in } \mathcal{S},$$
  then $\pi(\mathbf{x}, \mathbf{y}) = 0$ for all $\mathbf{y}$ in $\mathcal{D}$. Also, if $\alpha_{\mathbf{x}} = \sum_{\mathbf{y}} \pi(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} \in \mathcal{D}$ is the probability that the chain will be absorbed in $\mathcal{A}$ starting at $\mathbf{x} \in \mathcal{D}$, then $\alpha_{\mathbf{x}} = 1$ for all $\mathbf{x} \in \mathcal{D}$.
- $S_3 = \frac{1}{b} \sum_{k=1}^{\infty} \frac{(b/\mu)^k}{k!} = \frac{1}{b} \left(e^{b/\mu} - 1 \right) < \infty,$
  therefore the mean time $\tau_{\mathbf{x}}$ to reach $\mathcal{A}$, starting at $\mathbf{x} \in \mathcal{D}$, is finite for all $\mathbf{x} \in \mathcal{D}$.

The upper bounds for the average extinction time deduced from Lemma 5.4 for some specific sets of parameter values and initial conditions will be presented in Section 5.3.6.

Results for the limiting distribution of $(X(t), Y(t), Z(t))$ can also be obtained from the differential equation (5.13) for the probability generating function $\mathcal{P}$. Define

$$q_{rsv} = \lim_{t \to \infty} p_{rsv}(t), \quad \text{for } (r, s, v) \in \mathcal{S}$$

and

$$Q(x, y, z) = \sum_{(r,s,v) \in \mathcal{S}} x^r y^s z^v q_{rsv}. \tag{5.20}$$

Setting $\partial Q/\partial t = 0$ and $x = 1$, (5.13) becomes

$$0 = [\gamma(z-y) + (\mu+\delta)(1-y)]\left[\frac{\partial Q}{\partial y}\right]_{x=1} + [\beta(y-z) + \mu(1-z)]\left[\frac{\partial Q}{\partial z}\right]_{x=1}$$
$$+ \frac{\alpha}{n}y[(1-\rho)y + \rho z - 1]\left[\frac{\partial^2 Q}{\partial x \partial y}\right]_{x=1},$$

and substituting for the derivatives of Q from (5.20) this gives

$$0 = [\gamma(z-y) + (\mu+\delta)(1-y)]\sum_{r,s,v} s y^{s-1} z^v q_{rsv}$$
$$+ [\beta(y-z) + \mu(1-z)]\sum_{r,s,v} v y^s z^{v-1} q_{rsv} + \frac{\alpha}{n}y[(1-\rho)y + \rho z - 1]\sum_{r,s,v} r s y^{s-1} z^v q_{rsv}.$$

Equating the coefficients of y^s and z^v we obtain systems of equations from which we deduce successively that

$$\begin{aligned} \sum_{r=0}^{\infty} q_{rsv} &= 0 \qquad \text{for all } v \geq 0, s > 0 \\ \sum_{r=0}^{\infty} q_{r0v} &= 0 \qquad \text{for all } v > 0. \end{aligned} \tag{5.21}$$

If Y_e is the random variable whose distribution is given by

$$\mathrm{P}[Y_e = s] = \lim_{t\to\infty} \mathrm{P}[Y(t) = s], \quad \text{for } s = 0, 1, 2, \dots,$$

and similarly for X_e and Z_e, then (5.21) implies that Y_e cannot have any finite positive value with probability one. This, along with the results from Theorem 5.3, implies that

$$\mathrm{P}[Y_e = s] = 0 \quad \text{for all } s > 0, \tag{5.22}$$

and

$$\mathrm{P}[Y_e = 0] = \mathrm{P}[Y_e = Z_e = 0] = 1 \tag{5.23}$$

Equations (5.22) and (5.23) imply that $Q(x, y, z)$ is a function of x only. Hence its derivatives with respect to y and z are zero and equation (5.13) reduces to a homogeneous first-order partial differential equation whose solution is

$$Q(x) = e^{n(x-1)},$$

which is the probability generating function of the Poisson distribution with mean $n = b/\mu$. Thus

$$\begin{aligned} q_{r00} &= \frac{e^{-n} n^r}{r!} \qquad \text{for } r = 0, 1, \dots \\ q_{rsv} &= 0 \qquad \text{for } (s, v) \neq (0, 0). \end{aligned}$$

Summarising, the results obtained so far are

$$\pi(\mathbf{x},\mathbf{y}) = \lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = \mathbf{y}|\mathbf{X}(0) = \mathbf{x}] = 0, \text{ for all } \mathbf{y} \text{ in } \mathcal{D}$$

$$\alpha_{\mathbf{x}} = \sum_{\mathbf{y}\in\mathcal{A}} \pi(\mathbf{x},\mathbf{y}) = 1, \text{ for all } \mathbf{x} \text{ in } \mathcal{D} \tag{5.24}$$

$$q_{\mathbf{y}} = \lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = \mathbf{y}] = \begin{cases} 0 & \mathbf{y} \in \mathcal{D} \\ \frac{e^{-n}n^r}{r!} & \mathbf{y} = (r,0,0) \in \mathcal{A}, \end{cases}$$

where $\mathcal{A} = \{(r,0,0) : r = 0,1,\dots\}$ and $\mathcal{D} = \mathcal{S} - \mathcal{A}$. The probability of extinction of the disease is one and the expected time until extinction is finite.

Finally let m_W^e denote the mean of W_e, for $W_e = X_e, Y_e, Z_e, N_e = X_e + Y_e + Z_e$. From equation (5.16) it follows that the m_X^e, m_Y^e, m_Z^e must satisfy the equation

$$b = \mu m_N^e + \delta m_Y^e = \mu m_X^e + (\mu+\delta) m_Y^e + \mu m_Z^e, \tag{5.25}$$

which is exactly the equation (5.6) that the deterministic x_e, y_e, z_e satisfy. Also (5.23) implies that $m_Y^e = m_Z^e = 0$ and therefore (5.25) gives $m_X^e = m_N^e = n$.

5.3.3 The deterministic values and the stochastic means

The study presented so far shows that there is a difference between the behaviour of the stochastic model and that of its deterministic counterpart. For the deterministic model the disease dies out only if $\mathcal{R}_0 < 1$ (and otherwise stabilises at an endemic equilibrium), while for the stochastic model it always dies out with probability one. This difference has been observed in other models as well, for instance in Bartlett (1956) and Stirzaker (1975) for the open SI model and in Jacquez & Simon (1993) for the closed SIS and the open SI models.

Since the deterministic model can be viewed as an approximation to the stochastic (the system for the moments of X, Y, and Z reduces to the deterministic system (5.1) if the random variables X, Y, Z take the values x, y, z with probability one), at least in this sense no fundamental differences between their behaviours is to be expected. Nevertheless there is a basic difference (at least mathematically) between the two models that can cause such fundamental differences.

The deterministic model is described by the system (5.1) in which the functions $x(t)$, $y(t)$, and $z(t)$ are continuous functions. But in real life these numbers are discrete since they represent the numbers of susceptibles, infectives, and inactives, respectively. The stochastic model does account for this discreteness; it is described by the differential equations for the transition probabilities $p_{rsv}(t)$ and in this case the stochastic variables $X(t)$, $Y(t)$, and $Z(t)$ can take only integer values. The relation between a stochastic model and its deterministic counterpart has been explored in the literature of epidemic modelling where it has been shown that the stochastic elements introduced into the description of epidemiological processes (by accounting for the fact that humans exist in integer units and not as fractions) can affect the behaviour of the process in various ways (see, e.g., Bartlett 1957, Bartlett 1960*a*, Anderson & May 1991, Section 6.5).

With respect to the equilibrium state of the process for the deterministic model if there exists even a small fraction of infectivity in the population (for example if $y(t)$ has a positive value less than one) then this *can* cause a new infection and thus prevent the infection from extinction. With the stochastic formulation though that can not happen; if the number of infectives is equal to one at some point t, then it will either remain equal to one or decrease to zero or increase to two. There is always a positive probability that it will decrease to zero and Theorem 5.3 proves that eventually this *will* happen with probability one and in finite time. The same holds for the number of inactive cases, $Z(t)$, and hence ultimately the vector $(Y(t), Z(t))$ becomes equal to $(0, 0)$, which means that the infection dies out for any set of positive parameters. But for the deterministic model this happens only for certain sets of parameters (that satisfy $\mathcal{R}_0 < 1$) and otherwise the epidemic settles down at an endemic equilibrium.

This behaviour of the deterministic model is shown in Figure 5.2. The system (5.1) for the deterministic $x(t)$, $y(t)$, and $z(t)$ was solved numerically for two sets of positive parameters, the first one with $\mathcal{R}_0 = 0.591 < 1$ and the second one with $\mathcal{R}_0 = 5.914 > 1$. The values used in the second case are representative for TB:

$$\begin{aligned} &\alpha = 10 && \rho = 0.9725 \\ &\beta = 0.0022 && \gamma = 0.066 \\ &\mu = 0.02 && \delta = 0.13 \\ &b = \mu n, \end{aligned} \tag{5.26}$$

while the values used in the first case are the same as in (5.26) except for α which was taken equal to 1. For the same sets of parameters, the means of $X(t)$, $Y(t)$, and $Z(t)$ as obtained from the normal approximation and from 10000 simulations are also

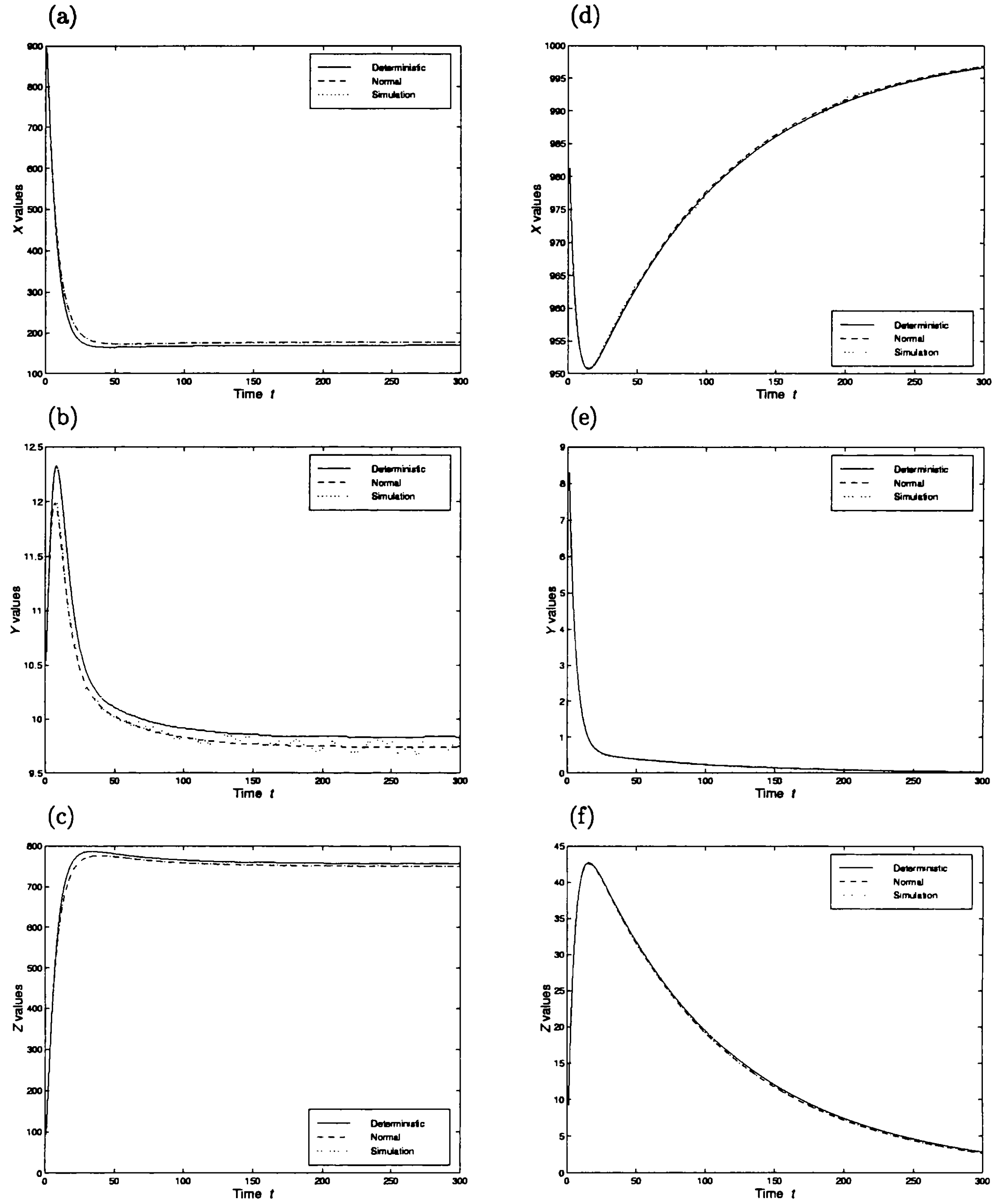


Figure 5.2: (a), (b), (c) The means of X, Y, and Z, respectively, with $\mathcal{R}_0 = 5.914$. The parameter values are as shown in (5.26). (d), (e), (f) The means of X, Y, and Z, respectively, with $\mathcal{R}_0 = 0.591$. The parameter values are as shown in (5.26), except that $\alpha = 1$. In each graph there are three curves, one for the deterministic value, one for the mean from the normal approximation, and one for the mean from simulations. The initial conditions are $n = 1000$, $x_0 = 990$, $y_0 = 10$, $z_0 = 0$. Time t is measured in years.

presented in Figure 5.2 (details of the implementation of the simulations can be found in the Appendix, Section A.2.2). For all these cases the initial conditions are $n = 1000$, $x_0 = 990$, $y_0 = 10$, $z_0 = 0$. In all cases the three curves (for the deterministic, the normal approximation, and the simulations) are very close, and in some cases they can hardly be distinguished.

What is most surprising from these results is the fact that when $\mathcal{R}_0 > 1$ the stochastic means of $Y(t)$ and $Z(t)$ (both from the normal approximation and from the simulations) do not tend to zero, but seem to stabilise at a level that is quite close to the deterministic equilibrium. At first this result seems odd since the limit of the probability of extinction of TB as time tends to infinity is equal to one.

For the solutions of the system for the means under the normal assumption, it is not so unreasonable that they exhibit a behaviour that is closer to the behaviour of the deterministic model than that predicted for the stochastic model; according to Kurtz's theory (see Theorem 4.1), when the initial population size n is large the deterministic model provides a good approximation to the stochastic model. Despite the fact that the discreteness of X, Y, Z (accounted for in the stochastic model but not in the deterministic), as explained above, may give rise to various differences between the two models, its effects are in a way diminished in the solution of the system of differential equations for the moments, because the means are continuous functions. If at some point the mean of $Y(t)$ has a positive value less than one, this means that there is positive probability that the value of $Y(t)$ is positive and thus there could still be new infections.

Also it should be noted that for a general random variable $Y(t)$ which has a limiting distribution with mean $\mathrm{E}[Y]$, this need not necessarily be the same as the limit, $\lim_{t\to\infty} \mathrm{E}[Y(t)]$, of the expected value of $Y(t)$, since the lim and E operators are not always interchangeable. One example of this is the birth and death process with birth and death rates λ and μ per capita, respectively, with $\lambda = \mu$. For this model it has been proved (see, e.g., Cox & Miller 1965, Section 4.3) that if $N(t)$ is the size of the population at time t then

$$\lim_{t\to\infty} \mathrm{P}[N(t) = 0] = 1,$$

and hence the expected value of this limiting distribution is zero, but

$$\lim_{t\to\infty} \mathrm{E}[N(t)] = \mathrm{E}[N(0)],$$

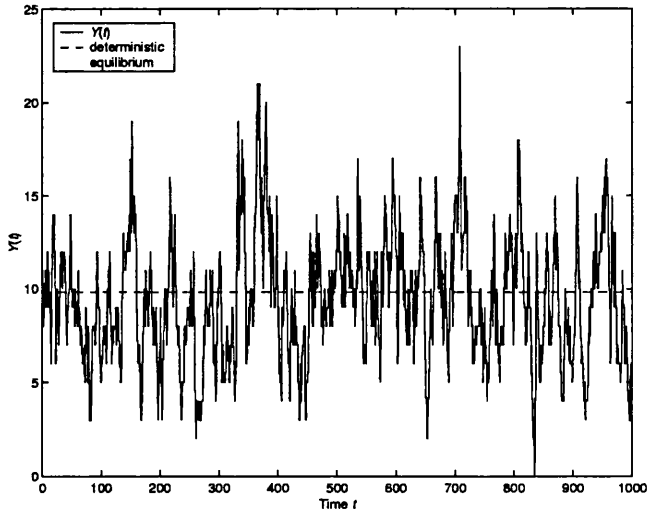


Figure 5.3: The value of $Y(t)$ as obtained from an individual realisation of the stochastic model. The initial conditions are $n = 1000$, $x_0 = 990$, $y_0 = 10$, and $z_0 = 0$. The parameter values are as shown in (5.26). Time is measured in years.

and that is why, as Cox & Miller (1965) comment, when $\lambda = \mu$ the equation for the expected value of $N(t)$ "does not give a good idea of the behaviour of individual realisations". Thus the fact that the values of $\mathrm{E}[Y(t)]$ do not seem to tend to zero as t increases (which is what we observe from the simulations for large t) does not prove that the expected value of the limiting distribution of $Y(t)$ is not zero.

In addition, there is another argument that explains why the stochastic means do not seem to tend to zero. Figure 5.3 shows the results for $Y(t)$ from an individual realisation of an epidemic in a population of initial size $n = 1000$ starting at $t = 0$ with $x_0 = 990$, $y_0 = 10$, $z_0 = 0$ (the results were obtained from numerical simulation, using the parameter values shown in (5.26)). The value of $Y(t)$ fluctuates around the deterministic endemic equilibrium y_{e_2} and it does not exhibit any decreasing tendency to justify the fact that ultimately it will become zero (as predicted theoretically by the analysis in the previous section). Around $t = 820$, Y finally becomes equal to zero, but the value of Z is positive at that point. Soon afterwards an inactive case becomes infectious and the value of Y increases to positive values once more and continues to fluctuate around y_{e_2} until $t = 1000$ when the simulation terminates.

This phenomenon is quite common in many stochastic processes and not only those for epidemics (see, e.g., Oppenheim, Shuler & Weiss 1977 for chemical reactions, Hitchcock 1986 for predator-prey processes): ultimate extinction of a class of the population is certain, but the process exhibits an apparent stationarity and it does not seem to die out over any reasonable length of time.

One well-known and extensively studied example of this phenomenon is the open SI model (see, e.g., Bartlett 1956, Stirzaker 1975, Ridler-Rowe 1967). In a series of papers Bartlett (1956, 1957, 1960*a*) studied this model and tried to explain the sustained oscillations that the number of infectives exhibit, despite the fact that it was proven that extinction of the infection is certain. It has been proposed (see, e.g., Pollett & Stewart 1994) that the explanation for this *apparent* stationarity is that the underlying process has a quasi-stationary (or limiting-conditional) distribution. This may well be the case for the model presented in this chapter, and we discuss this possibility in the following sections.

5.3.4 Quasi-stationary distributions

In the literature there are several forms and definitions for quasi-stationary distributions and limiting-conditional distributions. We will focus our discussion on the following distributions.

Definition 5.5 *Let $\{\mathbf{X}(t), t \geq 0\}$ be a Markov process in continuous time with denumerable state space $\mathcal{S}$. Let $\mathcal{A}$ denote the set of absorbing states and $\mathcal{D}$ the remaining set of states where $\mathcal{D}$ is an irreducible class. Suppose that the probability of ultimate absorption in $\mathcal{A}$ is 1 and the expected time until absorption is finite for any initial conditions. Finally let $p_{ij}(t) = \mathrm{P}[\mathbf{X}(t) = j | \mathbf{X}(0) = i]$ for i, j in $\mathcal{S}$ and $\mathbf{P}(t) = \{p_{ij}(t)\}$ be the matrix of transition probabilities.*

A distribution $\mathbf{m} = (m_j, j \in \mathcal{D})$ over $\mathcal{D}$ is a quasi-stationary distribution (QSD) if, whenever $\mathrm{P}[\mathbf{X}(0) = j] = m_j$ for j in $\mathcal{D}$,

$$\mathrm{P}[\mathbf{X}(t) = j | \mathbf{X}(t) \notin \mathcal{A}] = m_j, \quad j \text{ in } \mathcal{D},\ t \geq 0,$$

which means that if the initial distribution of the process is $\mathbf{m}$ then the distribution of $\mathbf{X}(t)$ conditioned on non-absorption by t is $\mathbf{m}$.

A distribution $\mathbf{u} = (u_j, j \in \mathcal{D})$ over $\mathcal{D}$ is called a limiting-conditional distribution (LCD) if

$$\lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = j | \mathbf{X}(0) = i, \mathbf{X}(t) \notin \mathcal{A}] = u_j, \quad i, j \text{ in } \mathcal{D}. \qquad (5.27)$$

A distribution $\mathbf{v} = (v_j, j \in \mathcal{D})$ *over* $\mathcal{D}$ *is called a doubly limiting-conditional distribution (DLCD) if*

$$\lim_{t \to \infty} \lim_{s \to \infty} \mathrm{P}[\mathbf{X}(t) = j | \mathbf{X}(0) = i, \mathbf{X}(t+s) \notin \mathcal{A}] = v_j, \quad i, j \text{ in } \mathcal{D}. \tag{5.28}$$

Kingman (1963) proved that to each irreducible class $\mathcal{D}$ there corresponds a number $\lambda \geq 0$, called the decay parameter of $\mathcal{D}$, such that

$$\lim_{t \to \infty} \frac{1}{t} \log p_{ij}(t) = -\lambda, \quad i, j \text{ in } \mathcal{D}. \tag{5.29}$$

If $\lambda > 0$ then (5.29) implies that $p_{ij}(t) = O(e^{-\lambda t})$ as $t \to \infty$ which means that the probabilities $p_{ij}(t)$ decrease exponentially to zero and λ is the common rate of decrease.

Using the decay parameter λ, an irreducible class $\mathcal{D}$ can be classified as λ-recurrent or λ-transient according as $\int_0^\infty p_{ii}(t) e^{\lambda t} dt$ diverges or converges for all i in $\mathcal{D}$. Further, a λ-recurrent class is called λ-positive recurrent if $\lim_{t \to \infty} p_{ij}(t) e^{\lambda t} > 0$ for all i, j in $\mathcal{D}$ and λ-null recurrent otherwise. Kingman (1963) showed that λ-transience and λ-positivity are "solidarity" properties, so that either all or none of the transition probabilities exhibit the behaviour of the specified type.

Therefore apart from the standard classification of the states (as positive recurrent, null recurrent, or transient) according to the limit behaviour of the probabilities $p_{ij}(t)$, there is a further classification according to the behaviour of $p_{ij}(t) e^{\lambda t}$, where λ is the common rate of decrease of the probabilities $p_{ij}(t)$ to zero. Hence even a transient class may be λ-positive recurrent. Moreover there is a relationship between λ-positive recurrence and existence of QSD's and LCD's similar to the relationship between positive recurrence and stationary distributions.

Theorem 5.6 *(Kingman 1963, Theorem 4) If* $\mathcal{D}$ *is a* λ*-recurrent class then there exist vectors* $\mathbf{x} = (x_j, j \in \mathcal{D})$, $\mathbf{y} = (y_j, j \in \mathcal{D})$ *unique (up to constant multiples) such that:*

$$\sum_{j \in \mathcal{D}} p_{ij}(t) x_j = e^{-\lambda t} x_i, \quad i \in \mathcal{D}$$

$$\sum_{i \in \mathcal{D}} y_i p_{ij}(t) = e^{-\lambda t} y_j, \quad j \in \mathcal{D},$$

which means that if $\mathbf{P}_1(t)$ *is the restriction of* $\mathbf{P}(t)$ *to the class* $\mathcal{D}$ *then* $\mathbf{x}$ *and* $\mathbf{y}$ *are the right and left eigenvectors, respectively, of* $\mathbf{P}_1(t)$ *corresponding to the eigenvalue* $e^{-\lambda t}$*:*

$\mathbf{P}_1(t)\mathbf{x} = e^{-\lambda t}\mathbf{x}$ *and* $\mathbf{y}'\mathbf{P}_1(t) = e^{-\lambda t}\mathbf{y}'$. *The class* $\mathcal{D}$ *is* λ*-positive recurrent if and only if* $\sum_{k\in\mathcal{D}} x_k y_k < \infty$ *and then*

$$\lim_{t\to\infty} p_{ij}(t)e^{\lambda t} = \frac{x_i y_j}{\sum_{k\in\mathcal{D}} x_k y_k}, \quad i,j \in \mathcal{D}.$$

From this theorem it follows that (see, e.g., Vere-Jones 1969, Flaspohler 1974) if $\mathcal{D}$ is λ-positive recurrent then the limits in (5.27) and (5.28) (LCD and DLCD, respectively) exist, are independent of the initial state i, and are given by

$$\lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = j|\mathbf{X}(0) = i, \mathbf{X}(t) \notin \mathcal{A}] = \frac{y_j}{\sum_{k\in\mathcal{D}} y_k}, \quad i,j \text{ in } \mathcal{D}$$

$$\lim_{t\to\infty}\lim_{s\to\infty} \mathrm{P}[\mathbf{X}(t) = j|\mathbf{X}(0) = i, \mathbf{X}(t+s) \notin \mathcal{A}] = \frac{x_j y_j}{\sum_{k\in\mathcal{D}} x_k y_k}, \quad i,j \text{ in } \mathcal{D},$$

where $\mathbf{x}$ and $\mathbf{y}$ are the unique eigenvectors defined by Theorem 5.6 and the limit (5.27) is equal to zero if $\sum_{k\in\mathcal{D}} y_k = \infty$.

Nair & Pollett (1993) proved that a proper distribution $\mathbf{y}$ over $\mathcal{D}$ is a QSD on $\mathcal{D}$ if and only if $\mathbf{y}$ is a left eigenvector for $\mathbf{P}_1(t)$ corresponding to an eigenvalue $e^{-\mu t}$ for some $\mu > 0$. This result combined with Theorem 5.6 proves the existence of QSD's for λ-recurrent classes. Therefore λ-recurrence implies the existence of a QSD and λ-positive recurrence the existence of LCD and DLCD. Nevertheless, in contrast to the theory for stationary distributions, if a class is not λ-positive recurrent the limiting-conditional distribution may still exist (see, e.g., Seneta 1966). Similar conditions for the existence of LCD's can be stated that depend on the reversed process or on the infinitesimal parameters q_{ij} of the process (see, e.g., Flaspohler 1974, Pollett 1988).

Also, the relationship between the eigenvectors of the $\mathbf{P}_1$ matrix and those of the $\mathbf{Q}_1$ matrix (the restriction of the q-matrix to the class $\mathcal{D}$) has been investigated since for practical purposes the q-matrix is easier to manipulate. If the $\mathbf{P}$ matrix of the process is the minimal transition function[1] (which is the case for regular processes) then any non-negative left eigenvector of $\mathbf{P}_1$ corresponding to an eigenvalue $e^{-\mu t}$ is a left eigenvector

[1] The matrix $P(t) = \{p_{ij}(t), i,j \in \mathcal{S}\}$ is the minimal transition function (see, e.g., Reuter 1957) if the probabilities $p_{ij}(t)$ are the minimal solution to the backward Kolmogorov equations

$$p'_{ij}(t) = \sum_{k\in\mathcal{S}} q_{ik}p_{kj}(t), \qquad t > 0,$$

and $p'_{ij}(0) = q_{ij}$. The minimal transition function also satisfies the forward Kolmogorov equations

$$p'_{ij}(t) = \sum_{k\in\mathcal{S}} p_{ik}(t)q_{kj}, \qquad t > 0.$$

of $\mathbf{Q}_1$ with the eigenvalue $-\mu$ for $\mu > 0$ (Vere-Jones 1969). If in addition the minimal transition function $\mathbf{P}(t)$ is honest (i.e. $\sum_{j \in \mathcal{S}} p_{ij}(t) = 1$, for all $i \in \mathcal{S}$, $t \geq 0$) and $\mathbf{y}$ is a convergent left eigenvector of $\mathbf{Q}_1$ for the eigenvalue $-\mu$ satisfying

$$\sum_{i \in \mathcal{D}} y_i q_{i\mathcal{A}} = \mu \sum_{i \in \mathcal{D}} y_i,$$

then $\mathbf{y}$ is a left eigenvector for $\mathbf{P}_1(t)$ corresponding to the eigenvalue $e^{-\mu t}$ (Pollett & Vere-Jones 1992).

Another interesting point that should be made is about the extremal character of the decay parameter λ. Vere-Jones (1969) showed that if the matrix $\mathbf{P}$ is strictly stochastic, the non-absorbing states form an irreducible class $\mathcal{D}$ with decay parameter λ and the equations

$$\sum_{i \in \mathcal{D}} v_i p_{ij}(t) = v_j e^{-\mu t}, \quad t \geq 0,\ j \in \mathcal{D}$$

hold for some non-negative, non-zero eigenvector $\mathbf{v} = (v_j, j \in \mathcal{D})$ then $0 < \mu \leq \lambda$.

For Markov processes with finite state spaces there exist unique positive left and right eigenvectors for this extremal eigenvalue $e^{-\lambda t}$ and therefore there is only one QSD which is the LCD (see Darroch & Seneta 1965, 1967). If the state space is infinite then QSD's may or may not exist, there may be more than one QSD and even if QSD's do exist an LCD may not exist.

The existence of LCD's was established for branching processes by Yaglom (1947), for random walks by Seneta (1966) and Pakes (1973) and for birth and death processes by Good (1968) and van Doorn (1991). Unfortunately for multidimensional processes with infinite state spaces there have not been many significant advances and even for the open SI model mentioned above the proof for the existence of QSD's and LCD's remains an open problem.

The existence of limiting-conditional distributions results in the *apparent* stationarity that these processes exhibit before absorption. Ultimately the process will be absorbed in the absorbing class $\mathcal{A}$, which is the set $\{(x, 0, 0) : x = 0, 1, 2, \ldots\}$ in our case. But since the distribution of the process conditioned on non-absorption,

$$\mathrm{P}[\mathbf{X}(t) = j | \mathbf{X}(0) = i, \mathbf{X}(t) \notin \mathcal{A}], \quad i, j \in \mathcal{D} = \mathcal{S} - \mathcal{A}, \tag{5.30}$$

has a stationary (limiting, as $t \to \infty$) distribution, it may go through this stationary phase before it gets absorbed in $\mathcal{A}$; the process will remain there for some time and then eventually will be absorbed. "Eventually" may be a very long time (as the numerical results in Figure 5.2 suggest) so it makes sense to study the conditional distribution (5.30).

Let $\mathrm{P}_{\mathcal{A}}(t)$ be the probability of extinction of TB, that is

$$\mathrm{P}_{\mathcal{A}}(t) = \mathrm{P}[Y(t) = Z(t) = 0] = \sum_{x=0}^{\infty} p_{x00}(t), \quad t \geq 0.$$

The distribution of the process conditioned on non-absorption is given by

$$q_{xyz}(t) = \mathrm{P}[(X(t), Y(t), Z(t)) = (x, y, z) | (Y(t), Z(t)) \neq (0,0)] = \frac{p_{xyz}(t)}{1 - \mathrm{P}_{\mathcal{A}}(t)},$$

for (x, y, z) in $\mathcal{D}$ and $q_{xyz}(t) = 0$ otherwise, where $\mathcal{D} = \{(x, y, z) \in \mathbb{Z}_+^3 : (y, z) \neq (0,0)\}$. If we define $c(\mathbf{v}, \mathbf{w})$ the coefficient of $p_{\mathbf{w}}(t)$ in the equation (5.12) for $dp_{\mathbf{v}}(t)/dt$ (where $p_{\mathbf{v}}(t) = p_{xyz}(t)$ for $\mathbf{v} = (x, y, z) \in \mathcal{S}$), then the equation (5.12) can be written as

$$\frac{dp_{\mathbf{v}}(t)}{dt} = \sum_{\mathbf{w}} c(\mathbf{v}, \mathbf{w}) p_{\mathbf{w}}(t), \quad \mathbf{v} \in \mathcal{D}.$$

From this equation it follows that

$$\frac{dq_{\mathbf{v}}(t)}{dt} = \sum_{\mathbf{w}} c(\mathbf{v}, \mathbf{w}) q_{\mathbf{w}}(t) + q_{\mathbf{v}}(t) Q(t), \quad \mathbf{v} \in \mathcal{D}, \tag{5.31}$$

where

$$Q(t) = (\mu + \delta) \sum_{x=0}^{\infty} q_{x10}(t) + \mu \sum_{x=0}^{\infty} q_{x01}(t), \quad t \geq 0.$$

The limiting-conditional distribution can be deduced by solving equation (5.31) with the left-hand side set equal to zero. The last term in (5.31) makes the equation nonlinear so that an analytical solution for the q_{xyz}'s can not easily be deduced from (5.31). A computational procedure for evaluating the limiting-conditional distributions is described by Pollett & Stewart (1994). This method, which is an 'iterative version' of Arnoldi's algorithm (see Pollett & Stewart 1994 for references), is based on the idea of truncating the q-matrix (restricted to the non-absorbing states) to an $m \times m$ matrix and constructing a sequence of vectors $\{x_m\}$, such that x_m is the left eigenvector of the truncated matrix corresponding to the eigenvalue with maximum real part. Then the quasi-stationary distribution can be estimated by taking successively larger truncations until the difference in the normalised eigenvectors is as small as desired. Pollett & Stewart (1994) describe a method for calculating the eigenvectors x_m and illustrate their algorithm with reference to the open SI model studied by Ridler-Rowe (1967).

5.3.5 The marginal distributions

The marginal distributions of X, Y, and Z have been calculated from numerical simulations of the stochastic model (details of the implementation of the simulations can be found in the Appendix, Section A.2.2). Figure 5.4 shows the marginal distributions of X and Z with $n = 50$, $y_0 = 1$, $x_0 = n - y_0$ and the parameter values shown in (5.26), for which $\mathcal{R}_0 = 5.914$.

The probability mass distribution soon splits into two distributions: one has mass distributed around the stochastic equilibrium ($m_x^e = n$, $m_y^e = 0$, $m_z^e = 0$, which corresponds to the extinction of TB) and the other has mass distributed around an endemic level, which is close to the deterministic endemic equilibrium. During the first two years the distribution of Z is basically unimodal with one peak between the stochastic equilibrium 0 and the deterministic equilibrium $z_{e_2} = 37.86$. As time increases this peak moves closer to z_{e_2} and becomes smaller, with a second peak developing at $z = 0$. By time $t = 15$ the distribution has become bimodal, with a very high peak around the deterministic equilibrium. After $t = 40$ the mass around z_{e_2} starts decreasing and "moves" towards the point $z = 0$ (the peak around z_{e_2} becomes more flat, more mass appears between the points $z = 0$ and $z = z_{e_2}$, and the probability at $z = 0$ increases). The simulation was carried out up to time $t = 500$ (results not shown here) and at that point the peak around z_{e_2} is almost flat and the probability $\mathrm{P}[Z(500) = 0]$ is 0.5.

The same observations can be made from the distribution of X (here we show the results up to time $t = 500$). The distribution is again bimodal, with one mode around the stochastic equilibrium $x = 50$ and one around an endemic level close to the deterministic equilibrium $x_{e_2} = 8.45$. Slowly in time the mode around x_{e_2} reduces in size and the mode around $x = 50$ increases, showing the tendency towards extinction.

Figure 5.5 shows the marginal distribution of Z for two other cases with (a) $n = 50$, $y_0 = 1$, $x_0 = n - y_0$ and the parameter values shown in (5.26) except $\alpha = 20$, for which $\mathcal{R}_0 = 11.828$ and (b) $n = 50$, $y_0 = 5$, $x_0 = n - y_0$ and the parameter values shown in (5.26), for which $\mathcal{R}_0 = 5.914$. Comparing the first of these with that for the distribution of Z in Figure 5.4 shows the effect of increasing the value of $\mathcal{R}_0$. Again during the first 2–3 years the distribution is unimodal with one peak between $z = 0$ and $z = z_{e_2} = 41.71$. However, this mode moves more quickly towards $z = z_{e_2}$ than when $\alpha = 10$ (Figure 5.4) and during the first 20 years the distribution remains unimodal with

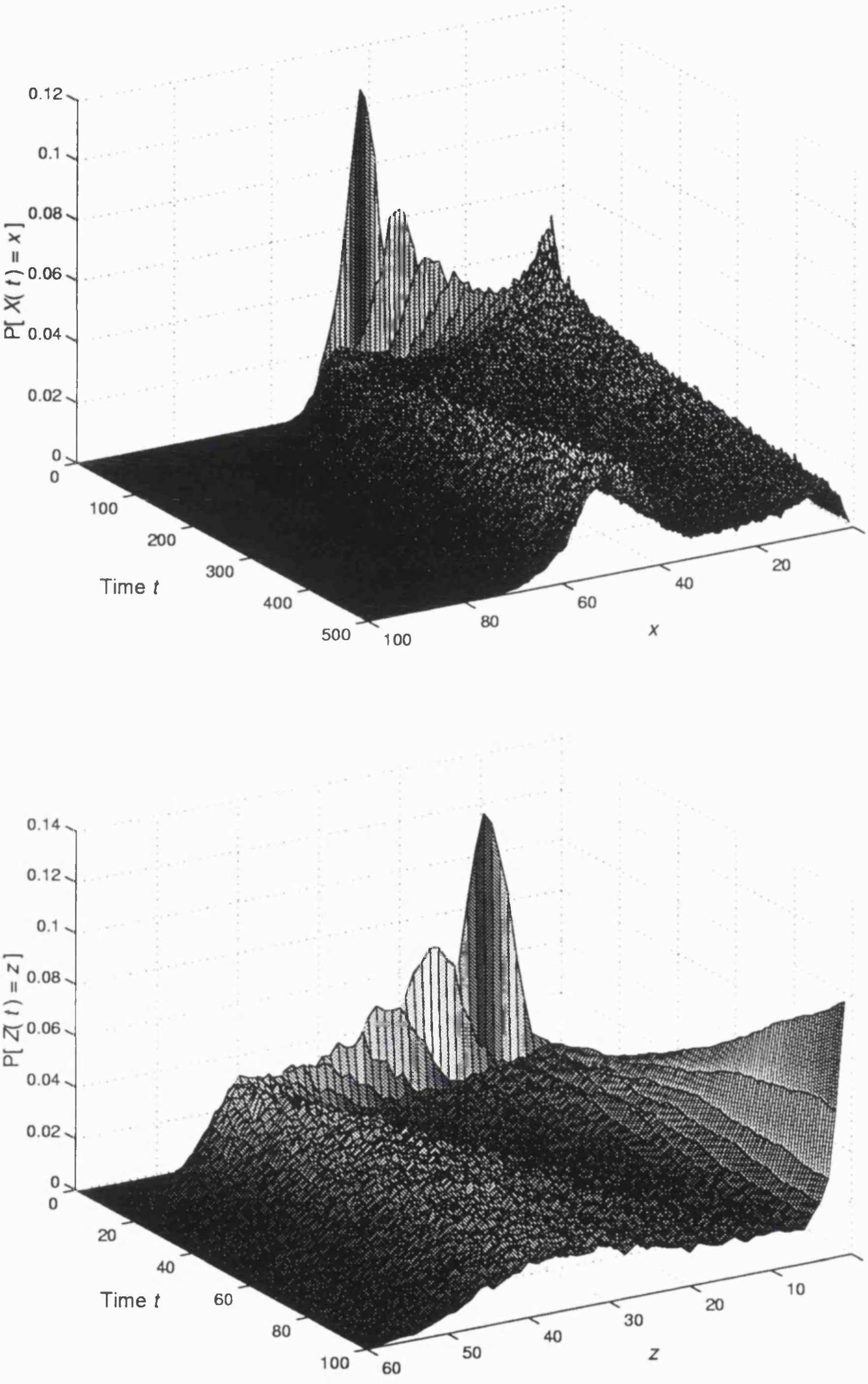


Figure 5.4: The marginal distributions of X and Z. The initial conditions are $n = 50$, $y_0 = 1$, $x_0 = n - y_0$. The parameter values are as shown in (5.26), for which $\mathcal{R}_0 = 5.914$. The deterministic endemic equilibrium is $\mathbf{e}_2 = (8.45, 0.49, 37.86)$. Time is measured in years.

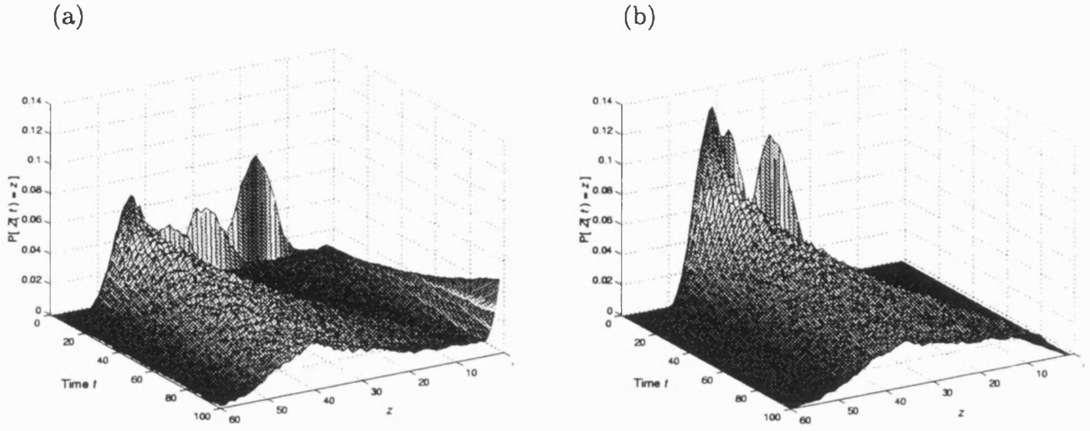


Figure 5.5: The marginal distribution of Z with $n = 50$, $x_0 = n - y_0$ and (a) $y_0 = 1$ and parameter values as in (5.26) except $\alpha = 20$, for which $\mathcal{R}_0 = 11.828$. The deterministic endemic equilibrium in this case is $\mathbf{e}_2 = (4.23, 0.54, 41.71)$. (b) $y_0 = 5$ and parameter values as in (5.26), for which $\mathcal{R}_0 = 5.914$. The deterministic endemic equilibrium for this case is $\mathbf{e}_2 = (8.45, 0.49, 37.86)$. Time is measured in years.

one peak around $z_{\mathbf{e}_2}$. At time $t = 100$ the distribution of Z is bimodal (one mode around $z = 0$ and one around $z = z_{\mathbf{e}_2}$), but the splitting into these two modes takes longer time (than when $\alpha = 10$), less mass appears around $z = 0$ and more around $z = z_{\mathbf{e}_2}$.

The same effect was observed by increasing the initial number of infectives present at the beginning of the epidemic (Figures 5.4.b and 5.5.b), although to a greater extreme. With $y_0 = 5$, in the beginning all the probability mass is centred around $z = z_{\mathbf{e}_2}$ (forming a mode around $z_{\mathbf{e}_2}$ which is much bigger in size than when $y_0 = 1$). As time increases this mode reduces in size and more mass appears between the points $z = 0$ and $z = z_{\mathbf{e}_2}$. Still though, even up to time $t = 100$ the distribution of Z is unimodal and the probability $\mathrm{P}[Z(100) = 0]$ is almost equal to zero.

Therefore it seems that if the process starts with more infected individuals then it is more likely that the infection will persist for a long time. Intuitively, if it starts with the introduction of a very small number of infectious cases then the infection may die out before this number increases. For instance if the epidemic begins with only one infectious case, this one infective may die before infecting any susceptibles. Clearly, the larger the y_0, the less likely it is that all of these y_0 infectives will die before infecting any susceptibles and hence the mass around the stochastic equilibrium (corresponding to extinction) will be more when $y_0 = 1$ than when $y_0 = 5$.

Finally, Figure 5.6 shows the marginal distribution of Y for two cases with $n = 1000$ and $n = 100$ (in both cases $y_0 = 1$, $x_0 = n - y_0$ and the parameter values are as

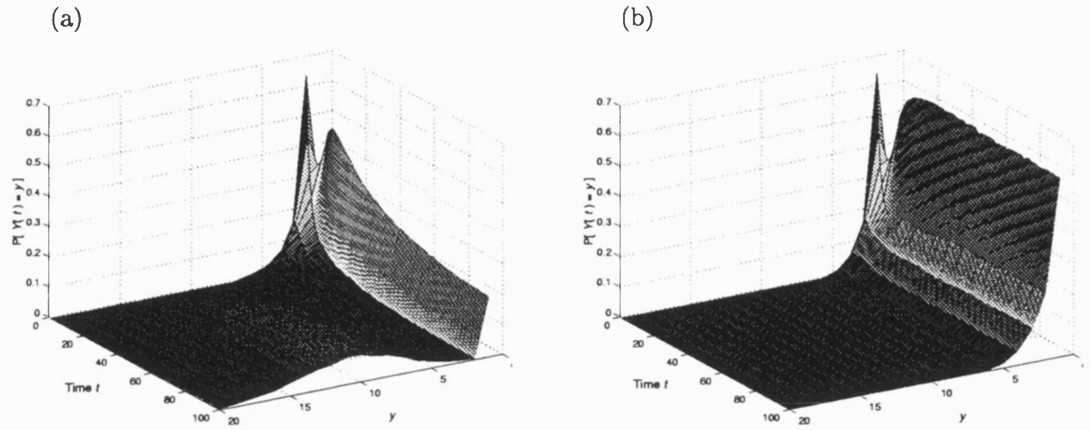


Figure 5.6: The marginal distribution of Y with (a) $n = 1000$ and (b) $n = 100$. In both cases $y_0 = 1$, $x_0 = n - y_0$, and the parameter values are as in (5.26), for which $\mathcal{R}_0 = 5.914$. The deterministic endemic equilibrium is (a) $\mathbf{e}_2 = (169.09, 9.83, 757.20)$ and (b) $\mathbf{e}_2 = (16.91, 0.98, 75.72)$. Time is measured in years.

in (5.26)). With $n = 1000$ the distribution is bimodal, with one mode around $y = 0$ and one around $y = y_{\mathbf{e}_2} = 9.83$. With $n = 100$ the distribution seems to be unimodal since the deterministic equilibrium is very close to zero ($y_{\mathbf{e}_2} = 0.98$). It is clear though from this figure how the initial size of the population affects the modes of the distribution: the probability $\mathrm{P}[Y(t) = 0]$ is not only much smaller with the larger value of n, but it even seems to decrease as time increases (which means that it is not only that we cannot observe any tendency towards extinction, but rather the opposite: a tendency towards the endemic equilibrium is observed, and extinction becomes less likely).

Summarising, these results suggest that the marginal distributions of X, Y, and Z are bimodal. The probability mass distribution splits into two parts, one centred around the stochastic equilibrium and the other around an endemic level close to the deterministic equilibrium. As n, $\mathcal{R}_0$, or the initial number of infected individuals increases, the mass around the stochastic equilibrium decreases and the mass around the deterministic equilibrium increases. With small values of n, $\mathcal{R}_0$, and y_0, the mode around the deterministic equilibrium reduces in size as time increases until the distribution becomes essentially unimodal (with one mode around the stochastic equilibrium) which shows the ultimate extinction of the infection (similar observations for the form of the marginal distributions have been made for the open SI and the closed SIS model (Jacquez & Simon 1993) and the open and closed SIS model in discrete time (Allen & Burgin 2000)). The simulations have to be carried out for a very long time in order to demonstrate this behaviour so that our results here are given only for very small values of n. It appears that the time until extinction can be very long and also that it is more likely to be long for large values of n and $\mathcal{R}_0$. The following section presents some more information about how long the extinction time can be.

5.3.6 The time until extinction and the size of the epidemic

Lemma 5.4 provides an upper bound for the expected time until extinction (which is finite, as shown in Section 5.3.2). Using the notation of the lemma for our process $r_k = b$ and $s_k = \mu k$ for $k \geq 1$. Therefore the sequence α_k defined by

$$\alpha_k = \frac{1}{s_1} + \frac{r_1}{s_1 s_2} + \cdots + \frac{r_1 r_2 \ldots r_{k-1}}{s_1 s_2 \ldots s_k}, \quad k \geq 2,$$

and $\alpha_1 = 1/s_1$ is

$$\alpha_k = \frac{1}{b} \sum_{v=1}^{k} \frac{(b/\mu)^v}{v!}, \quad k \geq 1.$$

Since $S_3 < \infty$, where

$$S_3 = \frac{1}{s_1} + \sum_{k=2}^{\infty} \frac{r_1 r_2 \ldots r_{k-1}}{s_1 s_2 \ldots s_k} = \frac{1}{b}\left(e^{b/\mu} - 1\right),$$

we can choose $U_0 = 0$ and $V_0 = e^{b/\mu}/b$ and define the sequences

$$\begin{aligned} V_k &= \frac{s_k \ldots s_1}{r_k \ldots r_1}(V_0 - \alpha_k) = \frac{k!}{(b/\mu)^k} \frac{1}{b}\left[e^{b/\mu} - \sum_{v=1}^{k} \frac{(b/\mu)^v}{v!}\right] && k \geq 1 \\ U_k &= U_{k-1} + V_{k-1} && k \geq 1. \end{aligned}$$

Then the expected time, $\tau(x, y, z)$, until extinction starting from the non-absorbing state (x, y, z) with $x+y+z = k \geq 1$ is $\tau(x, y, z) \leq U_k$. Some numerical results for these upper bounds are given later in this section.

Lower bounds for the expected extinction time can be deduced from Lemma 5 of Ridler-Rowe (1967) which provides lower bounds for the mean absorption time for Markov processes with state space $S = \{(x_1, \ldots, x_m) \in \mathbb{Z}_+^m\}$ where all the states with $x_m = 0$ are absorbing. This lemma can be modified to account for processes with absorbing states $(x_1, 0, \ldots 0)$:

Lemma 5.7 *Consider a Markov process in continuous time with state space $\mathcal{S} = \mathbb{Z}_+^m$ for $m \geq 2$. Let $q_{\mathbf{xy}}$ be the infinitesimal parameters of the process with $q_{\mathbf{xy}} \geq 0$ for all $\mathbf{x} \neq \mathbf{y}$ and $0 \leq -q_{\mathbf{xx}} \equiv q_{\mathbf{x}} = \sum_{\mathbf{y} \neq \mathbf{x}} q_{\mathbf{xy}} < \infty$. Let $\mathcal{A} = \{(x_1, 0, \ldots, 0) \in \mathcal{S}\}$ be the set* *of absorbing states and $\mathcal{D} = \mathcal{S} - \mathcal{A}$ the remaining set of states. Assume that the process almost surely reaches $\mathcal{A}$ from any initial state. Let $\tau(\mathbf{y})$ be the mean absorption time starting at $\mathbf{y}$ and define $\tilde{\mathbf{x}} = x_2 + x_3 + \cdots + x_m$ for $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathcal{S}$. If*

$$\begin{aligned} q_{\mathbf{xy}} &= 0, \quad \textit{whenever} \;\; \tilde{\mathbf{y}} < \tilde{\mathbf{x}} - 1 \\ Q_v &= \sup_{\mathbf{x}:\tilde{\mathbf{x}}=v} \sum_{\mathbf{y}:\tilde{\mathbf{y}}=\tilde{\mathbf{x}}-1} q_{\mathbf{xy}}, \end{aligned} \tag{5.32}$$

then

$$\tau(\mathbf{x}) \geq \sum_{i=1}^{\tilde{\mathbf{x}}} \frac{1}{Q_i}, \quad \textit{for all} \;\; \mathbf{x} \in \mathcal{D}.$$

The proof is similar to Ridler-Rowe's proof, so we will present only an outline here. Let $p_{\mathbf{xy}}(t)$ denote the transition probabilities of the process and $p^n_{\mathbf{xy}}(t)$ the n-step transition probabilities in the Feller minimal process (see Reuter 1957). Define $\phi_{\mathbf{xy}}(\theta)$ and $\phi^n_{\mathbf{xy}}(\theta)$ to be the Laplace transforms of $p_{\mathbf{xy}}(t)$ and $p^n_{\mathbf{xy}}(t)$, respectively. Then it is known (see Reuter 1957) that $\phi^0_{\mathbf{xy}}(\theta) = 0$ for all $\theta > 0$ and $\mathbf{x}$, $\mathbf{y}$ in $\mathcal{S}$ and

$$(\theta + q_{\mathbf{x}})\phi^{n+1}_{\mathbf{xy}}(\theta) = \delta_{\mathbf{xy}} + \sum_{\mathbf{z}\neq\mathbf{x}} q_{\mathbf{xz}}\phi^n_{\mathbf{zy}}(\theta), \quad \mathbf{x}, \mathbf{y} \in \mathcal{S}, \; \theta > 0, \; n \geq 0,$$

(where $\delta_{\mathbf{xy}}$ is the Kronecker delta) and for a regular process $\phi^n_{\mathbf{xy}}(\theta) \uparrow \phi_{\mathbf{xy}}(\theta)$ as $n \to \infty$. Let the absorbing states be classed as one state, say $\mathcal{A}$. Define

$$\psi_n(\theta) = \frac{1}{\theta} \prod_{k=1}^{n} \frac{Q_k}{\theta + Q_k}, \quad \theta > 0, \; n = 1, 2, \ldots,$$

and $\psi_0(\theta) = 1/\theta$ for $\theta > 0$. Clearly $\phi^0_{\mathbf{x}\mathcal{A}}(\theta) \leq \psi_{\tilde{\mathbf{x}}}(\theta)$ for any $\theta > 0$, $\mathbf{x} \in \mathcal{S}$ and $\phi^k_{\mathcal{AA}}(\theta) \leq \psi_{\tilde{\mathcal{A}}}(\theta)$, for any $\theta > 0$, $k \geq 0$. We will show that $\phi^k_{\mathbf{x}\mathcal{A}}(\theta) \leq \psi_{\tilde{\mathbf{x}}}(\theta)$ for any $\mathbf{x}$ in $\mathcal{S}$ and $k \geq 0$. For $\mathbf{x} \neq \mathcal{A}$ suppose that $\phi^k_{\mathbf{x}\mathcal{A}}(\theta) \leq \psi_{\tilde{\mathbf{x}}}(\theta)$ for some $k \geq 0$ and let $\tilde{\mathbf{x}} = n > 0$. Then

$$(\theta + q_{\mathbf{x}})\phi^{k+1}_{\mathbf{x}\mathcal{A}}(\theta) = \delta_{\mathbf{x}\mathcal{A}} + \sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{xy}}\phi^k_{\mathbf{y}\mathcal{A}}(\theta) \leq \sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{xy}}\psi_{\tilde{\mathbf{y}}}(\theta) \tag{5.33}$$

$$\leq \sum_{\mathbf{y}:\tilde{\mathbf{y}}<\tilde{\mathbf{x}}} q_{\mathbf{xy}}\psi_n(\theta)\left(1 + \frac{\theta}{Q_n}\right) + \sum_{\mathbf{y}:\tilde{\mathbf{y}}\geq\tilde{\mathbf{x}},\mathbf{y}\neq\mathbf{x}} q_{\mathbf{xy}}\psi_n(\theta) \tag{5.34}$$

$$= \psi_n(\theta)\left[q_{\mathbf{x}} + \theta\frac{1}{Q_n}\sum_{\mathbf{y}:\tilde{\mathbf{y}}<\tilde{\mathbf{x}}} q_{\mathbf{xy}}\right] \leq \psi_n(\theta)(\theta + q_{\mathbf{x}}),$$

where the inequality (5.34) follows from the fact that the sequence ψ_n is a decreasing function of n and $\psi_{n-1}(\theta) = \psi_n(\theta)(1 + \theta/Q_n)$ for $n \geq 1$. Therefore $\phi^k_{\mathbf{x}\mathcal{A}}(\theta) \leq \psi_{\tilde{\mathbf{x}}}(\theta)$ for any $\mathbf{x}$ in $\mathcal{S}$, $k \geq 0$, $\theta > 0$. Letting $k \to \infty$, follows that $\phi_{\mathbf{x}\mathcal{A}}(\theta) \leq \psi_{\tilde{\mathbf{x}}}(\theta)$ for all $\mathbf{x} \in \mathcal{S}$. From Reuter (1961) we know that

$$\tau(\mathbf{x}) = \lim_{\theta \downarrow 0} \frac{1 - \theta\phi_{\mathbf{x}\mathcal{A}}(\theta)}{\theta},$$

and hence for all $\mathbf{x} \neq \mathcal{A}$

$$\tau(\mathbf{x}) \geq \lim_{\theta \downarrow 0} \frac{1 - \theta\psi_{\tilde{\mathbf{x}}}(\theta)}{\theta} = \sum_{i=1}^{\tilde{\mathbf{x}}} \frac{1}{Q_i}.$$

Lemma 5.7 can be modified to give a tighter bound for $\tau(\mathbf{x})$. If $||\mathbf{x}||_1$ denotes the l_1 norm on $\mathbb{Z}^m$ defined by $||\mathbf{x}||_1 = |x_1| + |x_2| + \cdots + |x_m|$ and the conditions (5.32) are substituted by

$$q_{\mathbf{xy}} = 0, \quad \text{whenever } ||\mathbf{y}||_1 < ||\mathbf{x}||_1 - 1$$

$$Q'_v = \sup_{\mathbf{x}:||\mathbf{x}||_1 = v} \left\{ \sum_{\mathbf{y}:||\mathbf{y}||_1 = ||\mathbf{x}||_1 - 1} q_{\mathbf{xy}} \right\}, \quad v = 1, 2, \ldots,$$

then

$$\tau(\mathbf{x}) \geq \sum_{i=1}^{||\mathbf{x}||_1} \frac{1}{Q'_i}, \quad \text{for all } \mathbf{x} \neq \mathcal{A} \text{ with } q_{\mathbf{x}\mathcal{A}} = 0. \tag{5.35}$$

The proof is similar to the proof for Lemma 5.7 and is omitted here. The states $\mathbf{x} \neq \mathcal{A}$ with $q_{\mathbf{x}\mathcal{A}} \neq 0$ have to be excluded in this case because for these states the inequality (5.33) does not necessarily hold.

For our process

$$Q_v = \sup_{y+z=v} \{(\mu + \delta)y + \mu z\} = (\mu + \delta)v, \quad v = 1, 2, \ldots$$

$$Q'_v = \sup_{x+y+z=v} \{\mu x + (\mu + \delta)y + \mu z\} = (\mu + \delta)v, \quad v = 1, 2, \ldots.$$

Therefore if the process begins from the state $(x, y, z) \in \mathcal{S}$ then the mean time $\tau(x, y, z)$ until extinction is bounded below by

$$\mathcal{L}_1 \equiv \frac{1}{\mu + \delta} \sum_{i=1}^{y+z} \frac{1}{i}, \qquad \text{for } (x, y, z) \in \mathcal{S} \text{ with } y + z \geq 1$$

$$\mathcal{L}_2 \equiv \frac{1}{\mu + \delta} \sum_{i=1}^{x+y+z} \frac{1}{i}, \qquad \text{for } (x, y, z) \in \mathcal{S} \text{ with } y + z \geq 2,$$

where $\mathcal{L}_1$, $\mathcal{L}_2$ are the two lower bounds from Lemma 5.7 and (5.35), respectively.

Initial population size	Lower bound $\mathcal{L}_2$	Upper bound
10	19.5	$1.245 \cdot 10^5$
25	25.4	$1.503 \cdot 10^{11}$
50	29.9	$5.293 \cdot 10^{21}$
100	34.6	$1.358 \cdot 10^{43}$

Table 5.1: Bounds for the expected extinction time. $\mathcal{L}_2$ is the bound from inequality (5.35), for processes starting with $y_0 + z_0 \geq 2$. The upper bound (obtained from Lemma 5.4) is for processes with $y_0 + z_0 \geq 1$. Both bounds were calculated with the parameter values shown in (5.26).

$y_0 + z_0$	1	2	5	10	20	100	200	1000
$\mathcal{L}_1$	6.67	10.00	15.22	19.53	23.98	34.58	39.19	49.90

Table 5.2: The lower bound $\mathcal{L}_1$ from Lemma 5.7 for the expected extinction time (for processes starting with $y_0 + z_0 \geq 1$) and for $\mu = 0.022$, $\delta = 0.139$.

Tables 5.1 and 5.2 show the values of these lower bounds and the upper bound from Lemma 5.4 for several initial conditions using the parameter values shown in (5.26). Unfortunately the intervals defined by the lower and upper bounds are very wide because these bounds account for extreme cases. For instance the lower bound $\mathcal{L}_1$ is essentially the expected extinction time for a death process that starts at level $y_0 + z_0$ and has death rate $Q_v = (\mu + \delta)v$ when the size of the population is v. Therefore $\mathcal{L}_1$ basically counts only the time until all initial infectives and inactives have died assuming that no other event occurs. It has to be noted though that both lower bounds tend to infinity as the initial sizes $y + z$ or $x + y + z$ tend to infinity.

Processes starting with one infectious and zero inactive cases

First we consider only epidemics that begin with the introduction of one infectious case into a wholly susceptible population, so that the initial conditions are $x_0 = n - 1$, $y_0 = 1$, $z_0 = 0$. In order to determine the effect of n and $\mathcal{R}_0$, the process was simulated for the following values of n and α: $n = 50, 100, 200, 400$ and $\alpha = 2, 3, 4$ (recall from (5.3) that $\mathcal{R}_0$ is a multiple of α, so that by increasing α, the value of $\mathcal{R}_0$ increases proportionately). The other parameter values are as shown in (5.26). The simulations were repeated 10^4 times and terminated at a time point large enough such that the epidemic had died out by that point in all 10^4 runs. Also, the size of the epidemic (defined as the total number of susceptibles infected from time $t = 0$ until the epidemic dies out) was calculated from

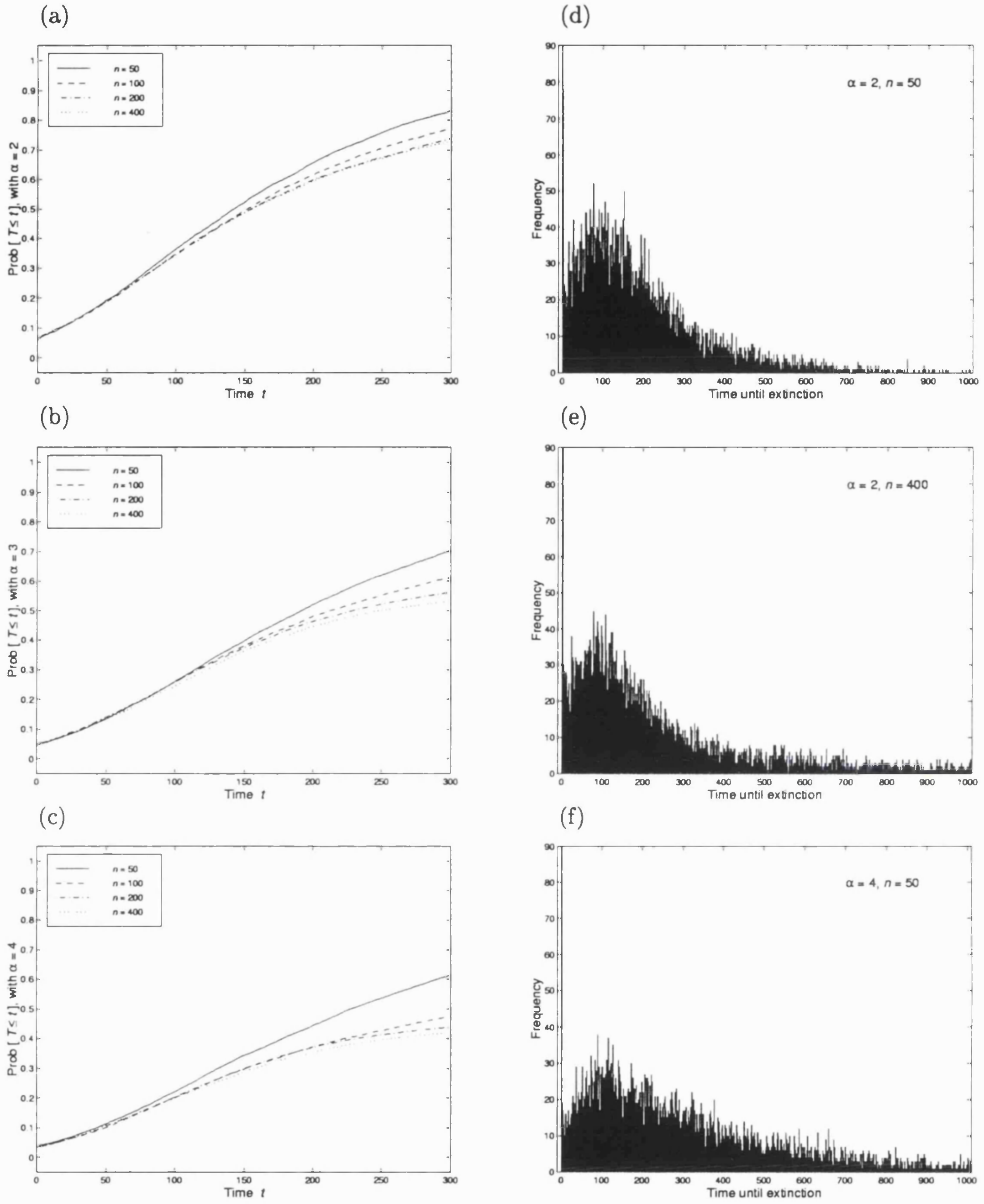


Figure 5.7: The distribution of the extinction time, $\mathcal{T}$, for processes that start with one infectious case: $y_0 = 1$ and $x_0 = n-1$. The parameter values are as shown in (5.26) except for α, which has the value indicated in each graph. With these values and α equal to 2, 3, and 4, the value of $\mathcal{R}_0$ is 1.183, 1.774, and 2.366, respectively. Time is measured in years. (a), (b), (c) The cumulative distribution, $\mathrm{P}[\mathcal{T} \leq t]$, with $\alpha = 2, 3, 4$, respectively. In each graph there are four curves one for each of the following values of n: 50, 100, 200, and 400. (d), (e), (f) Histograms of $\mathcal{T}$ for the three cases: $\alpha = 2$, $n = 50$; $\alpha = 2$, $n = 400$; and $\alpha = 4$, $n = 50$. 10^4 simulation runs of the stochastic model were carried out for a time long enough such that all runs ended with extinction of the infection, thus yielding a sample of size 10^4 from the distribution of $\mathcal{T}$. These 10^4 values are plotted here in a histogram, where the width of each box is equal to 1 (year) and the height is equal to the number of simulation runs in which the epidemic died out during that year, so that roughly these graphs show the distribution of $\mathcal{T}$ as $\mathrm{P}[t-1 \leq \mathcal{T} \leq t]$, counting t in years.

		$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\alpha = 2$	Min:	$2.079{\cdot}10^{-4}$	$5.069{\cdot}10^{-4}$	$3.563{\cdot}10^{-4}$	$6.659{\cdot}10^{-4}$
	Q_1:	$6.815{\cdot}10^{+1}$	$7.056{\cdot}10^{+1}$	$6.989{\cdot}10^{+1}$	$7.000{\cdot}10^{+1}$
	Q_2:	$1.418{\cdot}10^{+2}$	$1.517{\cdot}10^{+2}$	$1.546{\cdot}10^{+2}$	$1.552{\cdot}10^{+2}$
	Mean:	$1.771{\cdot}10^{+2}$	$2.058{\cdot}10^{+2}$	$2.402{\cdot}10^{+2}$	$2.856{\cdot}10^{+2}$
	Q_3:	$2.453{\cdot}10^{+2}$	$2.812{\cdot}10^{+2}$	$3.127{\cdot}10^{+2}$	$3.264{\cdot}10^{+2}$
	Max:	$1.441{\cdot}10^{+3}$	$1.701{\cdot}10^{+3}$	$2.462{\cdot}10^{+3}$	$3.928{\cdot}10^{+3}$
	SD:	$1.531{\cdot}10^{+2}$	$1.959{\cdot}10^{+2}$	$2.649{\cdot}10^{+2}$	$3.870{\cdot}10^{+2}$
	Corr:	$8.521{\cdot}10^{-1}$	$8.809{\cdot}10^{-1}$	$9.004{\cdot}10^{-1}$	$9.287{\cdot}10^{-1}$
$\alpha = 3$	Min:	$2.363{\cdot}10^{-4}$	$7.713{\cdot}10^{-4}$	$4.233{\cdot}10^{-4}$	$1.543{\cdot}10^{-4}$
	Q_1:	$9.627{\cdot}10^{+1}$	$9.675{\cdot}10^{+1}$	$9.697{\cdot}10^{+1}$	$1.022{\cdot}10^{+2}$
	Q_2:	$1.916{\cdot}10^{+2}$	$2.133{\cdot}10^{+2}$	$2.290{\cdot}10^{+2}$	$2.486{\cdot}10^{+2}$
	Mean:	$2.459{\cdot}10^{+2}$	$3.453{\cdot}10^{+2}$	$6.827{\cdot}10^{+2}$	$3.497{\cdot}10^{+3}$
	Q_3:	$3.334{\cdot}10^{+2}$	$4.707{\cdot}10^{+2}$	$8.867{\cdot}10^{+2}$	$4.258{\cdot}10^{+3}$
	Max:	$3.142{\cdot}10^{+3}$	$4.541{\cdot}10^{+3}$	$1.071{\cdot}10^{+4}$	$6.266{\cdot}10^{+4}$
	SD:	$2.179{\cdot}10^{+2}$	$3.783{\cdot}10^{+2}$	$1.004{\cdot}10^{+3}$	$6.474{\cdot}10^{+3}$
	Corr:	$9.153{\cdot}10^{-1}$	$9.546{\cdot}10^{-1}$	$9.857{\cdot}10^{-1}$	$9.988{\cdot}10^{-1}$
$\alpha = 4$	Min:	$9.474{\cdot}10^{-4}$	$1.040{\cdot}10^{-3}$	$4.815{\cdot}10^{-4}$	$1.074{\cdot}10^{-3}$
	Q_1:	$1.122{\cdot}10^{+2}$	$1.242{\cdot}10^{+2}$	$1.239{\cdot}10^{+2}$	$1.280{\cdot}10^{+2}$
	Q_2:	$2.279{\cdot}10^{+2}$	$3.349{\cdot}10^{+2}$	$6.771{\cdot}10^{+2}$	$1.819{\cdot}10^{+4}$
	Mean:	$3.029{\cdot}10^{+2}$	$6.007{\cdot}10^{+2}$	$2.597{\cdot}10^{+3}$	$9.976{\cdot}10^{+4}$
	Q_3:	$4.151{\cdot}10^{+2}$	$8.390{\cdot}10^{+2}$	$3.672{\cdot}10^{+3}$	$1.402{\cdot}10^{+5}$
	Max:	$2.423{\cdot}10^{+3}$	$6.322{\cdot}10^{+3}$	$4.261{\cdot}10^{+4}$	$1.540{\cdot}10^{+6}$
	SD:	$2.725{\cdot}10^{+2}$	$6.936{\cdot}10^{+2}$	$3.978{\cdot}10^{+3}$	$1.612{\cdot}10^{+5}$
	Corr:	$9.430{\cdot}10^{-1}$	$9.823{\cdot}10^{-1}$	$9.984{\cdot}10^{-1}$	$9.999{\cdot}10^{-1}$

Table 5.3: Summary statistics of the extinction time for processes that begin with one infectious case ($y_0 = 1$, $z_0 = 0$). The parameter values are as shown in (5.26) except for α which has the value indicated in the table. With these values and α equal to 2, 3, and 4, the value of $\mathcal{R}_0$ is equal to 1.183, 1.774, and 2.366, respectively. Time is measured in years. Q_1, Q_2, Q_3 are the first, second, and third quartiles, respectively. "Corr" is the correlation coefficient for the extinction time and the size of the epidemic.

the simulations (details of the implementation of the simulations can be found in the Appendix, Section A.2.2).

Table 5.3 shows summary statistics for the extinction time $\mathcal{T}$ and the distribution of $\mathcal{T}$ is shown in Figure 5.7. As n and/or $\mathcal{R}_0$ increase, the probabilities $\mathrm{P}[\mathcal{T} \leq t]$ decrease for each $t \geq 0$, more so if both $\mathcal{R}_0$ and n increase. For instance with $\alpha = 2$ and $n = 50$ more than 80% of the simulation runs ended with extinction at time $t = 300$, while with $\alpha = 4$ and $n = 400$ this proportion is less than 45%. Also the mean and standard deviation, the first, second, and third quartiles, and the maximum of $\mathcal{T}$ increase. For large values of n, the increase (as α increases) is more significant than for small values of n and similarly for large values of α the moments increase more (as n increases) than they do for small values of α.

The statistics in Table 5.3 and the graphs in Figure 5.7 also show that the distribution of $\mathcal{T}$ is highly skewed to the right, with a very high peak at the very beginning

		$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\alpha = 2$	Min:	0.0	0.0	0.0	0.0
	Q_1:	3.0	3.0	3.0	3.0
	Q_2:	10.0	11.0	11.0	11.0
	Mean:	25.5	45.0	82.7	166.4
	Q_3:	33.0	48.0	62.0	69.0
	Max:	449.0	699.0	2203.0	5421.0
	SD:	38.0	78.9	178.8	444.7
$\alpha = 3$	Min:	0.0	0.0	0.0	0.0
	Q_1:	5.0	5.0	5.0	5.0
	Q_2:	22.0	26.0	30.0	36.0
	Mean:	54.9	151.5	725.8	9846.2
	Q_3:	77.0	199.0	853.0	11700.0
	Max:	1304.0	2954.0	16560.0	188600.0
	SD:	78.3	256.3	1400.1	19348.8
$\alpha = 4$	Min:	0.0	0.0	0.0	0.0
	Q_1:	7.0	8.0	8.0	8.0
	Q_2:	40.0	112.0	752.0	76190.0
	Mean:	87.8	411.8	4710.1	425900.3
	Q_3:	126.0	596.0	6688.0	598000.0
	Max:	1043.0	5430.0	84620.0	6591000.0
	SD:	117.3	620.0	7769.8	689720.6

Table 5.4: Summary statistics of the size of the epidemic for processes that begin with one infectious case ($y_0 = 1$, $z_0 = 0$). The parameter values are as shown in (5.26) except for α which has the value indicated in the Table. With these values and α equal to 2, 3, and 4, the value of $\mathcal{R}_0$ is equal to 1.183, 1.774, and 2.366, respectively. Time is measured in years. Q_1, Q_2, Q_3 are the first, second, and third quartiles, respectively.

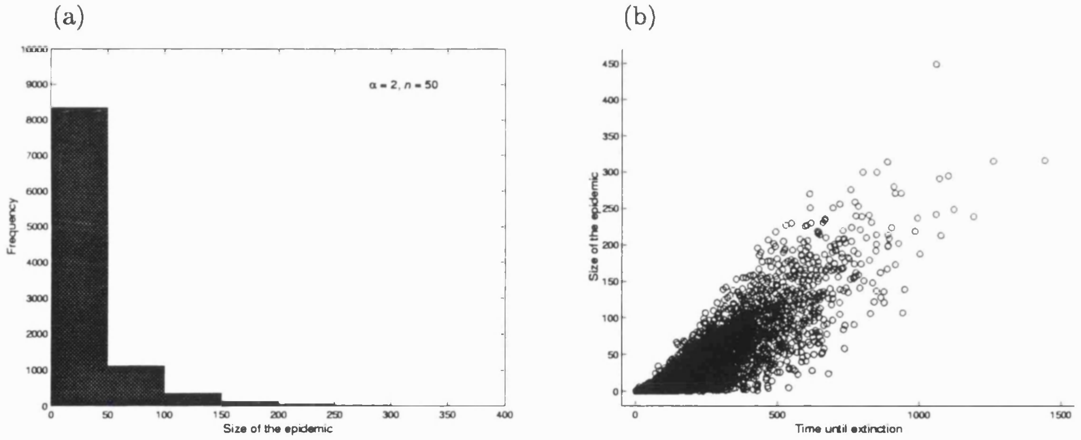


Figure 5.8: (a) Histogram of the size of the epidemic. (b) Scatter plot of the size of the epidemic against the extinction time $\mathcal{T}$. In both cases the parameter values are as in (5.26) except $\alpha = 2$ ($\mathcal{R}_0 = 1.183$). The initial conditions are $n = 50$ and $y_0 = 1$, $x_0 = n - 1$. Time is measured in years. 10^4 simulation runs of the stochastic model were carried out for a time long enough such that all runs ended with extinction of the infection. Hence the $R = 10^4$ runs yielded a sample of size R from the distribution of $\mathcal{T}$ and a sample of size R from the distribution of the size of the epidemic. In (a) the R values of the size are plotted in a histogram, where the width of each box is equal to 50. In (b) each circle represents one of the R pairs (τ_i, s_i), where τ_i, s_i are the extinction time and the size of the epidemic, respectively, in the i-th simulation run.

(first 1–2 years). As n and/or $\mathcal{R}_0$ increase, the distribution becomes less skewed, the peak in the beginning reduces in size and the tail of the distribution becomes longer and has more mass.

The same trend was observed for the size of the epidemic. Table 5.4 shows summary statistics for the size and Figure 5.8 shows the distribution of the size and a scatter plot of the size against the extinction time for the case with $\alpha = 2$ and $n = 50$. The other cases were qualitatively the same and are not shown here. The distribution of the size is highly skewed to the right. As n and/or $\mathcal{R}_0$ increase, the distribution becomes less skewed and it is more likely that the size of the epidemic will be large. The scatter plot in Figure 5.8 and the correlation coefficient of the size and the extinction time (Table 5.4) show that there is a very strong positive correlation between the two variables.

Processes starting with more than one infected individual

In this part we consider epidemics that begin with more than one infected individual, so that $y_0 + z_0 > 1$. The process was simulated for the following three cases: $\alpha = 2$ and $n = 50$; $\alpha = 2$ and $n = 100$; $\alpha = 4$ and $n = 50$. For each of these three cases, four different sets of initial conditions were used:

$$\begin{array}{lll} y_0 = 10\% \text{ of } n & \text{and} & z_0 = 0 \\ y_0 = 20\% \text{ of } n & \text{and} & z_0 = 0 \\ y_0 = 0 & \text{and} & z_0 = 10\% \text{ of } n \\ y_0 = 0 & \text{and} & z_0 = 20\% \text{ of } n. \end{array} \tag{5.36}$$

For each of these twelve combinations of parameter values and initial conditions, the simulations were repeated $R = 10^4$ times and terminated at a time point large enough such that the epidemic had died out in all runs. From each individual run, the extinction time and the size of the epidemic were obtained, thus yielding a sample $\tau = \{\tau_1, \ldots, \tau_R\}$ from the distribution of the extinction time $\mathcal{T}$ and a sample $s = \{s_1, \ldots, s_R\}$ from the distribution of the size of the epidemic. These two samples were used to calculate the statistics of $\mathcal{T}$ and those of the size, shown in Table 5.5. Also their distribution was calculated and is shown in Figures 5.9 and 5.10 for some of the cases (the other cases were qualitatively the same and are not presented here).

As y_0 or z_0 increase the statistics (mean and standard deviation, quartiles, min-

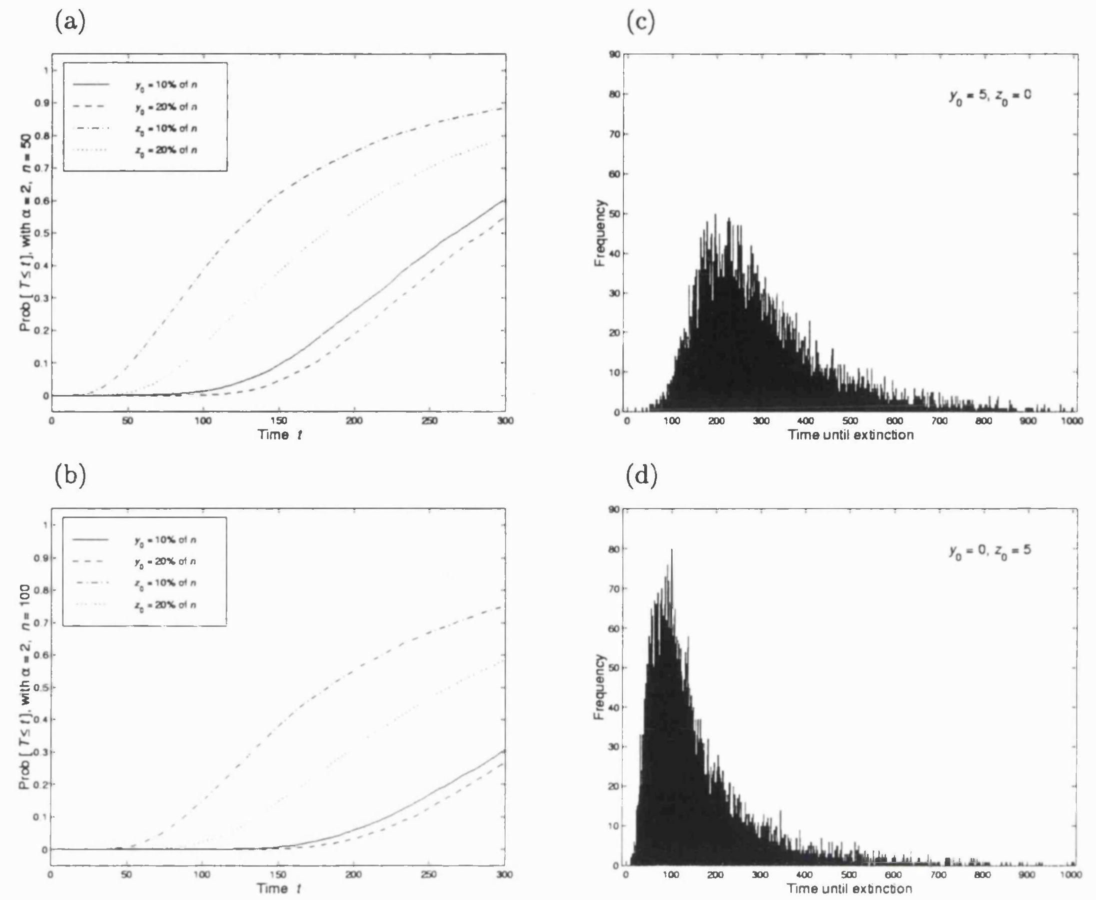


Figure 5.9: The distribution of the extinction time $\mathcal{T}$ for processes with $y_0 + z_0 > 1$. The parameter values are as shown in (5.26) except $\alpha = 2$ ($\mathcal{R}_0 = 1.183$). Time is measured in years. (a), (b) The cumulative distribution, $\mathrm{P}[\mathcal{T} \leq t]$, of $\mathcal{T}$ with $n = 50$ and $n = 100$, respectively. In each graph there are four curves, one for each of the sets of initial conditions shown in (5.36). (c), (d) Histograms of $\mathcal{T}$ with $(y_0, z_0) = (5, 0)$ and $(y_0, z_0) = (0, 5)$, respectively. In both cases $n = 50$. In these histograms the width of each box is equal to 1 (year) and the height of each box is equal to the number of simulation runs in which the epidemic died out during that year, so that roughly these graphs show the distribution of $\mathcal{T}$ as $\mathrm{P}[t-1 \leq \mathcal{T} \leq t]$, counting t in years.

imum and maximum) of $\mathcal{T}$ increase and the probability $\mathrm{P}[\mathcal{T} \leq t]$ decreases, more so when y_0 increases than when z_0 increases. In particular, the moments when y_0 is 10% (or 20%) of n and $z_0 = 0$ are higher than the moments with $y_0 = 0$ and z_0 equal to 10% (or 20%) of n and similarly the probability $\mathrm{P}[\mathcal{T} \leq t]$ is smaller.

Comparing Figures 5.7 and 5.9, it can be observed that the distribution of $\mathcal{T}$ is less skewed when $y_0 + z_0 > 1$. In this case there is no peak at $t = 1$ and the tail of the distribution is longer and has more mass. This difference between the cases with $y_0 = 1$ and $y_0 + z_0 > 1$ is more evident as y_0 increases than as z_0 increases (for instance, with $y_0 = 5$ the distribution is less skewed than with $z_0 = 5$).

These results agree with the results in Section 5.3.5 for the effect of increasing

	$\alpha = 2,\ n = 50$							
	$(y_0, z_0) = (5, 0)$		$(y_0, z_0) = (10, 0)$		$(y_0, z_0) = (0, 5)$		$(y_0, z_0) = (0, 10)$	
	$\mathcal{T}$	Size	$\mathcal{T}$	Size	$\mathcal{T}$	Size	$\mathcal{T}$	Size
Min:	17.4	1.0	82.7	10.0	9.8	0.0	17.0	0.0
Q_1:	196.5	36.0	216.9	47.0	77.5	0.0	120.0	0.0
Q_2:	266.3	54.0	284.9	64.0	120.7	0.0	179.9	4.0
Mean:	300.3	66.4	319.7	76.8	159.0	11.5	217.0	20.4
Q_3:	365.9	84.0	382.2	93.0	199.3	7.0	273.2	26.0
Max:	1415.0	419.0	1821.0	504.0	1217.0	297.0	1616.0	455.0
SD:	148.7	45.6	148.8	43.6	124.1	28.4	139.0	35.6
Corr:	0.8247		0.8202		0.8119		0.8208	

	$\alpha = 2,\ n = 100$							
	$(y_0, z_0) = (10, 0)$		$(y_0, z_0) = (20, 0)$		$(y_0, z_0) = (0, 10)$		$(y_0, z_0) = (0, 20)$	
	$\mathcal{T}$	Size	$\mathcal{T}$	Size	$\mathcal{T}$	Size	$\mathcal{T}$	Size
Min:	87.8	11.0	108.9	49.0	29.6	0.0	52.4	0.0
Q_1:	281.0	97.0	293.7	117.0	121.7	0.0	181.3	4.0
Q_2:	370.3	141.5	383.6	158.0	182.7	4.0	263.2	26.0
Mean:	421.3	169.3	435.7	189.6	242.0	38.5	320.5	65.4
Q_3:	507.0	211.0	521.7	230.0	298.3	36.0	403.0	90.0
Max:	1932.0	1007.0	1971.0	1016.0	2919.0	1266.0	2068.0	1058.0
SD:	202.9	104.5	202.9	104.6	186.2	80.6	200.0	95.2
Corr:	0.8599		0.8676		0.8810		0.8759	

	$\alpha = 4,\ n = 50$							
	$(y_0, z_0) = (5, 0)$		$(y_0, z_0) = (10, 0)$		$(y_0, z_0) = (0, 5)$		$(y_0, z_0) = (0, 10)$	
	$\mathcal{T}$	Size	$\mathcal{T}$	Size	$\mathcal{T}$	Size	$\mathcal{T}$	Size
Min:	54.8	6.0	97.9	26.0	7.3	0.0	20.5	0.0
Q_1:	259.0	70.0	261.9	71.0	79.0	0.0	127.6	0.0
Q_2:	370.4	121.0	368.6	118.0	129.3	0.0	209.5	10.0
Mean:	451.1	159.1	446.4	157.7	212.9	39.2	301.9	67.1
Q_3:	561.7	207.0	552.0	201.0	251.7	24.0	388.2	95.0
Max:	2555.0	1106.0	2932.0	1362.0	2606.0	1111.0	2392.0	1015.0
SD:	274.9	125.7	266.9	122.0	224.7	90.5	259.6	112.1
Corr:	0.9367		0.9359		0.9350		0.9392	

Table 5.5: Summary statistics of the extinction time, $\mathcal{T}$, and the size of the epidemic for processes with $y_0 + z_0 > 1$. The parameter values are as shown in (5.26) except for α which has the value indicated in the table. With these values and α equal to 2 and 4, the value of $\mathcal{R}_0$ is 1.183 and 2.366, respectively. Time is measured in years. Q_i is the i-th quartile, for $i = 1, 2, 3$. "Corr" is the correlation coefficient between the two variables.

y_0 on the marginal distributions of X, Y, and Z and the explanation for this effect is the same as for the marginal distributions. If the epidemic starts with a small number of infectives the infection may die out before this number increases. The greater the value of y_0, the less likely it is that all of these y_0 infectives will die within a year, and hence the minimum of $\mathcal{T}$ is larger, the probabilities $\mathrm{P}[\mathcal{T} \leq t]$ are smaller, and the histogram of $\mathcal{T}$ does not have the peak at $t = 1$ that it has in the cases with $y_0 = 1$. The more infectious cases are present when the epidemic starts, the more likely it is that the infection will persist and it will take a longer time until it dies out. The same holds if there are many inactive cases in the beginning but the effect is not so strong because

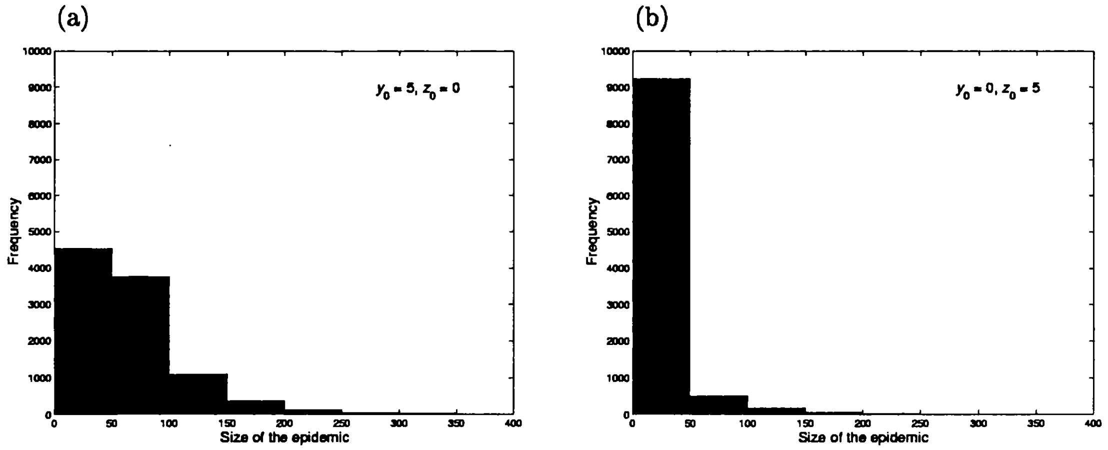


Figure 5.10: Histograms for the size of the epidemic with (a) $(y_0, z_0) = (5, 0)$ and (b) $(y_0, z_0) = (0, 5)$. In both cases $n = 50$, $\alpha = 2$ ($\mathcal{R}_0 = 1.183$), and the other parameter values are as shown in (5.26). In these histograms the width of each box is equal to 50.

the inactive cases cannot infect any susceptibles and spread the infection (for instance, starting with ten inactive cases boosts the persistence of the infection more than starting with five inactives, but still not as much as starting with ten infectious cases).

Increasing y_0 or z_0 had the same effect on the size of the epidemic. As y_0 or z_0 increases the moments and the minimum and maximum value of the size increase and its distribution becomes much less skewed, although with a longer tail. This result should be expected since for epidemics that begin with a large number of infected individuals it is more likely that they will last long and therefore more susceptibles will be infected. It has to be noted though that the effect of increasing z_0 as described above is not always the same if we compare cases with $(y_0, z_0) = (1, 0)$ with cases $(y_0, z_0) = (0, m)$, for some positive integer m. For instance the distribution of the size when $(y_0, z_0) = (0, 5)$ is slightly more skewed than when $(y_0, z_0) = (1, 0)$ and also its statistics are smaller. This result at first may seem odd, but it is probably due to the fact that with $\alpha = 2$ the value of $\mathcal{R}_0$ is just above one (1.183) so that the dynamics of the infection are not that strong and the epidemic dies out in some simulation runs before any of these 5 inactive cases become infectious (or before they infect any susceptibles, if they do become infectious).

5.3.7 Discussion and conclusions

When $\mathcal{R}_0 < 1$ the dynamics of the infection are weak and the infection dies out regardless of the population size for both the deterministic and the stochastic models. When $\mathcal{R}_0 > 1$ the dynamics of the infection are stronger and the deterministic model always stabilises at an endemic equilibrium. For the stochastic model though things are different; the infection eventually dies out with probability one but it may take a very long time for this event to occur.

If the population size is small or $\mathcal{R}_0$ is not much larger than one then the disease dies out in a relatively short time. If the population size is large and $\mathcal{R}_0$ is much greater than one, then after the initial phase of almost exponential growth of the number of infected individuals, the marginal distributions of X, Y, and Z split into two distributions, one centred around the stochastic equilibrium and the other around an endemic level which is close to the deterministic endemic equilibrium. The larger the initial population size and/or $\mathcal{R}_0$ the more mass appears around the endemic level. In these cases the process exhibits a quasi-stationary behaviour centred around the conditional means which are close to the deterministic endemic equilibrium. The simulations show that the process settles down there for some time. Nevertheless this is a temporary phenomenon and the infection eventually dies out. This can be observed in Figure 5.4 where the mode around the endemic level reduces (and the mode around the disease-free equilibrium increases) in size as time increases and the distribution tends to become unimodal. Unfortunately, from the perspective of demonstration via simulation, "eventually" may be a very long time and some simulations cannot show that. The larger the initial size of the population, the greater the possibility that this quasi-stationary behaviour will appear and last a long time.

For instance, Figure 5.3 shows the results for $Y(t)$ from an individual realisation of the stochastic model with $n = 1000$ and $y_0 = 10$, $z_0 = 0$. For 800 years the value of Y fluctuates around the deterministic endemic equilibrium ($y_{e_2} \approx 10$) and does not show any tendency to become zero. Around $t = 830$, finally Y becomes zero, but Z is positive at that point and hence Y starts increasing again and continues to fluctuate around y_{e_2} up to $t = 1000$ when the simulation was terminated.

This also explains why the mean of $Y(t)$ remains positive for finite t (and does not tend to zero, as for the case with $\mathcal{R}_0 = 5.914$ in Figure 5.2). The expected value of $Y(t)$ is a weighted average of the results with and without extinction and hence the mean of $Y(t)$ tends to zero as the probability of extinction increases. As was explained above, for large values of n it is likely that the $Y(t)$ will remain at the quasi-stationary level for a long time. Therefore during this time the probability of extinction increases

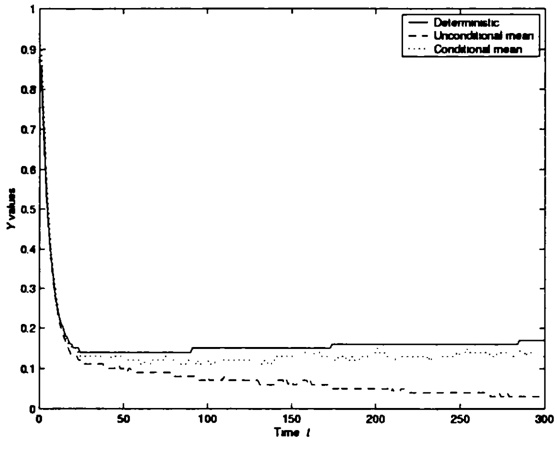


Figure 5.11: The stochastic means $\mathrm{E}[Y(t)]$ and $\mathrm{E}[Y(t)|(Y(s),Z(s)) \neq (0,0), s \leq t]$ (as obtained from numerical simulations) and the deterministic $y(t)$. The parameter values are as shown in (5.26) except $\alpha = 2$ ($\mathcal{R}_0 = 1.183$). The initial conditions are $n = 100$, $y_0 = 1$, $z_0 = 0$. Time is measured in years.

very slowly and the mean of $Y(t)$ remains around the endemic level.

For instance, in the results shown in Figure 5.2 with $\mathcal{R}_0 = 5.914$, there was no simulation that ended with $Y = Z = 0$ and the expected value of $Y(t)$ remained very close to y_{e_2}. In contrast, we present the results from another simulation with $n = 100$ and $\mathcal{R}_0 = 1.183$ (Figure 5.11) where we also calculated the expected value of $Y(t)$ conditioned on non-extinction of TB by time t. The proportion of simulations in which the infection died out soon is very small so that the conditional and unconditional means are very close initially. As the probability of extinction increases (see Figure 5.7, the curve for $\alpha = 2$, $n = 100$), the difference between the conditional and unconditional means increases: the conditional mean gets closer to the deterministic and the unconditional falls more and more below these two values, tending to zero. Therefore we have this seeming contradiction that the mean of $Y(t)$ remains positive over long periods (and sometimes does not even seem to tend to zero, as in Figure 5.2) although the expected value of the limiting distribution of Y is zero.

The same analysis holds for the number of inactive cases and this probably makes the time until the extinction of the disease even longer (than if, for instance, there were no class of inactive cases in the population) because the disease dies out only when both $Y(t)$ and $Z(t)$ become zero. But each of the events $\{Y(t) = 0\}$ and $\{Z(t) = 0\}$ is quite unlikely to occur, at any particular time point t. Actually in some realisations (for instance the one shown in Figure 5.3) the value of $Y(t)$ became zero at some point, fluctuated between the values 0 and 1 for a while, but $Z(t)$ had a positive value during this period and finally $Y(t)$ started increasing and "ended" up (when the simulation was terminated) fluctuating around the deterministic y_{e_2}.

Finally, it should also be mentioned that according to Anderson & May (1991), the presence of the latent (inactive) class makes the extinction of the disease less likely. Because it acts like a "reservoir" of the infection (latents are not infectious but they may become infectious; also there is no excess death for the inactive class) thus boosting the persistence of the infection.

In conclusion, when $\mathcal{R}_0 < 1$ the disease dies out, but when $\mathcal{R}_0 > 1$ it tends to stabilise at an endemic equilibrium; in the deterministic formulation it always does but in the stochastic only when the population is large and even then this is a temporary phenomenon. The stochastic fluctuations will eventually interrupt this quasi-stationarity and extinguish the infection. If the population size is small the infection may die out even without exhibiting this quasi-stationary behaviour. Nevertheless, the numerical results presented in this chapter suggest that the time until extinction of the infection may be very long, especially for large values of n and $\mathcal{R}_0$, and hence for any practical purposes it is only the endemic steady state that will be observed. For further research it might be interesting to modify this model by introducing immigration of infected individuals, so that the whole state space will become a single irreducible class of recurrent states and then the endemic steady state observed may become a genuine equilibrium distribution.

5.4 Linear Approximation

5.4.1 Formulation and analytical results

The first epidemic models that appeared in the literature were relatively simple and could be studied analytically quite extensively (see, e.g., Bailey 1975). In time modellers have turned to more complicated models resulting in quite complicated stochastic models that are not manageable and hence analytical results cannot easily be deduced.

One of the elements of epidemic models that makes them so complicated and not "solvable" is the non-linear term that is used in most cases for the rate at which new infections occur (defined as a fraction of the product of the number of susceptibles and the number of infectives). Obviously this non-linearity makes the model more complicated mathematically. For instance if all the rates (at which the various events occur) were linear, the differential equation for the probability generating function would be of first order and the system for the moments would be closed (with the non-linear terms they are always open, containing higher order moments, and cannot be solved).

One of the methods that have been used in order to make the stochastic models simpler is the linear approximation: one of the variables in the non-linear terms is taken as deterministic and the resulting model is linear (see, e.g., Tan & Hsu 1989, Isham 1991, Herbert 1998). Depending on the structure of the original stochastic model and the purposes of the approximation, more than one variable can be taken as deterministic and not only those involved in the non-linear terms. It is expected that the resulting linear model will give reliable approximations if the values of the variables that are taken as deterministic are large enough, since it has been shown that in these cases the deterministic values approximate the stochastic means quite well (see, e.g., Bailey 1975, Chapter 5).

One of the main advantages of this method is that it makes the stochastic model simpler and more manageable, while keeping the structure of the original model (for instance, without reducing the number of variables involved or changing the rates). Of course this happens with the cost of "losing" some of the randomness of the original stochastic model, which means losing all the information for the variation of the variables that are taken as deterministic. This can further affect the information for the remaining stochastic variables as well. However, a model of this short still accounts for more variation than for instance the corresponding deterministic model.

Overall there are indications that this approach can be helpful at least in some cases (for instance approximating the moments). In the remaining part of this section we present a linear model that approximates the stochastic model studied in this chapter and at the end we discuss the advantages and disadvantages of this approach. The linear approximation is used again for the model presented in the following chapter and some further remarks can be found in the corresponding Section 6.3.8.

The transient behaviour

Consider the model defined in Section 5.1 (Figure 5.1) and assume that X and Z evolve deterministically. The number of infectious cases, $Y_\ell(t)$, is a random variable for $t \geq 0$. Note that the numerical results presented in the previous sections indicate that the sizes of both X and Z are quite large in general, so that the model discussed in this section may be a good approximation to the original stochastic model.

Define $p^\ell_y(t) = \mathrm{P}[Y_\ell(t) = y]$ for $t \geq 0$ and $y = 0, 1, \ldots$. The $p^\ell_y(t)$ satisfy the differential equations

$$\begin{aligned}\frac{dp^\ell_y(t)}{dt} = &\left[(1-\rho)\frac{\alpha}{n}x(t)(y-1) + \beta z(t)\right] p^\ell_{y-1}(t) + (\gamma+\mu+\delta)(y+1)p^\ell_{y+1}(t) \\ &- \left[(1-\rho)\frac{\alpha}{n}x(t)y + \beta z(t) + (\gamma+\mu+\delta)y\right] p^\ell_y(t),\end{aligned} \tag{5.37}$$

for $y = 0, 1, \ldots$ and $p^\ell_y(t) = 0$ for any other y. Initially there are x_0 susceptibles, y_0 infectives, and z_0 inactive cases, so that $p^\ell_{y_0}(0) = 1$ and $p^\ell_y(0) = 0$ for any $y \neq y_0$. The probability generating function $\mathcal{P}_\ell(\theta; t) = \mathrm{E}[\theta^{Y_\ell(t)}]$ satisfies the equation

$$\frac{\partial \mathcal{P}_\ell}{\partial t} = \beta z(t)(\theta - 1)\mathcal{P}_\ell + \left[(1-\rho)\frac{\alpha}{n}x(t)\theta - (\gamma+\mu+\delta)\right](\theta-1)\frac{\partial \mathcal{P}_\ell}{\partial \theta}, \tag{5.38}$$

with initial condition $\mathcal{P}_\ell(\theta; 0) = \theta^{y_0}$.

From (5.38) the following equations for the mean and variance of Y_ℓ are deduced:

$$\frac{d\mathrm{E}[Y_\ell(t)]}{dt} = \left[(1-\rho)\frac{\alpha}{n}x(t) - (\gamma+\mu+\delta)\right]\mathrm{E}[Y_\ell(t)] + \beta z(t) \tag{5.39}$$

$$\begin{aligned}\frac{d\mathrm{Var}[Y_\ell(t)]}{dt} = &\beta z(t) + \left[(1-\rho)\frac{\alpha}{n}x(t) + \gamma+\mu+\delta\right]\mathrm{E}[Y_\ell(t)] \\ &+ 2\left[(1-\rho)\frac{\alpha}{n}x(t) - (\gamma+\mu+\delta)\right]\mathrm{Var}[Y_\ell(t)],\end{aligned} \tag{5.40}$$

whose solutions are given by

$$\mathrm{E}[Y_\ell(t)] = y_0 e^{F(t)} + e^{F(t)}\int_0^t \beta z(w)e^{-F(w)}dw \tag{5.41}$$

$$\mathrm{Var}[Y_\ell(t)] = e^{2F(t)}\int_0^t h(w)e^{-2F(w)}dw, \tag{5.42}$$

where

$$F(t) = (1-\rho)\frac{\alpha}{n}\int_0^t x(w)dw - (\gamma+\mu+\delta)t$$

$$h(t) = \beta z(t) + \left[(1-\rho)\frac{\alpha}{n}x(t) + \gamma+\mu+\delta\right]\mathrm{E}[Y_\ell(t)].$$

The equation (5.39) for the mean of Y_ℓ is the same as the equation for the deterministic y (system (5.1)). Since the solution for $\mathrm{E}[Y_\ell(t)]$ depends also on the values of the deterministic $x(t)$ and $z(t)$ (equations (5.39) and (5.41)), the value of the mean of Y_ℓ from the linear approximation is the same as the value of the deterministic y. Numerical results for the moments of Y_ℓ are shown at the end of this section.

Equilibrium

Define x_e, z_e the values of $x(t)$, $z(t)$ at equilibrium

$$x_e = \lim_{t\to\infty} x(t), \qquad z_e = \lim_{t\to\infty} z(t),$$

and $Y_{\ell e}$ the random variable defined by the limiting distribution of $Y_\ell(t)$

$$\lim_{t\to\infty} \mathrm{P}[Y_\ell(t) = y] = \mathrm{P}[Y_{\ell e} = y] = q_y,$$

for $y = 0, 1, \dots$. Let $\mathcal{P}_{\ell e}(\theta)$ be the probability generating function of $Y_{\ell e}$, defined by $\mathcal{P}_{\ell e}(\theta) = \mathrm{E}[\theta^{Y_{\ell e}}]$. Taking the limits as t tends to infinity in (5.38), it follows that $\mathcal{P}_{\ell e}$ satisfies the equation

$$0 = \beta z_e \mathcal{P}_{\ell e} + \left[(1-\rho)\frac{\alpha}{n}x_e\theta - (\gamma + \mu + \delta)\right] \frac{\partial \mathcal{P}_{\ell e}}{\partial \theta},$$

whose solution is given by

$$\mathcal{P}_{\ell e}(\theta) = \left[\frac{1}{Q_1 + (1 - Q_1)\theta}\right]^{\mathcal{K}}, \tag{5.43}$$

where

$$Q_1 = \frac{\gamma + \delta + \mu}{\gamma + \delta + \mu - \alpha(1-\rho)x_e/n} \quad \text{and} \quad \mathcal{K} = \frac{\beta z_e}{\alpha(1-\rho)x_e/n}. \tag{5.44}$$

If $z_e = 0$ then $\mathcal{P}_{\ell e}(\theta) = 1$ for any θ, hence for $\theta = 0$ as well. Therefore $q_0 = 1$ and $q_y = 0$ for any $y \neq 0$. If $z_e \neq 0$ then equation (5.43) gives the probability generating function of the negative binomial distribution with parameters $\mathcal{K}$ and $p_{\ell e} = 1/Q_1$ (where the $\mathcal{K}$ and Q_1 are non-negative because x_e and z_e are non-negative and from the second equation of (5.1) at equilibrium it follows that $\gamma + \delta + \mu - \alpha(1-\rho)x_e/n \geq 0$; also $Q_1 \geq 1$ so that $0 \leq p_{\ell e} \leq 1$). Thus the limiting distribution of $Y_\ell(t)$ is given by

$$q_y = \binom{\mathcal{K} + y - 1}{\mathcal{K} - 1} p_{\ell e}^{\mathcal{K}} (1 - p_{\ell e})^y, \quad y = 0, 1, \dots . \tag{5.45}$$

The limiting distribution (5.45) can also be deduced directly from the equation (5.37), taking the limits as $t \to \infty$. Finally, from the mean and variance of the negative binomial distribution it follows that

$$\mathrm{E}[Y_{\ell e}] = \mathcal{K}(Q_1 - 1) = \frac{\beta z_e}{\gamma + \delta + \mu - \alpha(1-\rho)x_e/n} \tag{5.46}$$

$$\mathrm{Var}[Y_{\ell e}] = \mathcal{K}(Q_1 - 1)Q_1 = \frac{(\gamma + \delta + \mu)\beta z_e}{[\gamma + \delta + \mu - \alpha(1-\rho)x_e/n]^2}. \tag{5.47}$$

In Section 5.2 it was shown that the deterministic model has two equilibria $\mathbf{e}_1 = (n, 0, 0)$ and $\mathbf{e}_2 = (x_{\mathbf{e}_2}, y_{\mathbf{e}_2}, z_{\mathbf{e}_2})$, given by (5.5a)–(5.5c). If $\mathcal{R}_0 < 1$ then $\mathbf{e}_1$ is stable and $\mathbf{e}_2$ infeasible; if $\mathcal{R}_0 > 1$ then $\mathbf{e}_1$ is unstable and $\mathbf{e}_2$ stable. From (5.5c) it follows that

$$y_{\mathbf{e}_2} = \frac{\beta z_{\mathbf{e}_2}}{\gamma + \delta + \mu - \alpha(1-\rho)x_{\mathbf{e}_2}/n}, \tag{5.48}$$

which also gives the limit of $\mathrm{E}[Y_\ell(t)]$ as t tends to infinity, since the value of $\mathrm{E}[Y_\ell(t)]$ is equal to the deterministic $y(t)$. The right hand sides of (5.46) and (5.48) are the same if $x_e = x_{\mathbf{e}_2}$ and $z_e = z_{\mathbf{e}_2}$. Therefore the mean of the limiting distribution and the limit of the mean of $\mathrm{E}[Y_\ell]$ are the same as the deterministic equilibrium y_e: equal to 0 if $\mathcal{R}_0 < 1$ and positive if $\mathcal{R}_0 > 1$. Finally, the probability of extinction of the disease is given by

$$q_0 = \begin{cases} 1, & \text{if } z_e = 0 \\ \dfrac{1}{Q_1^{\mathcal{K}}}, & \text{if } z_e \neq 0, \end{cases} \tag{5.49}$$

where the Q_1 and $\mathcal{K}$ are as defined in (5.44).

5.4.2 Numerical results

In Section 5.3.1 the system of differential equations (5.14) was deduced for the means of X, Y, Z. This system involves second order moments and similarly the system for the variances and covariances involves third order moments. Therefore the system for the first and second moments is open and cannot be solved. This is a problem encountered in most epidemic models, resulting from the non-linear term XY for the rate at which new infections occur.

One way of overcoming this problem is to express the higher order moments in terms of the first and second moments (see, e.g., Isham 1991, Herbert 1998 for moment closure methods). For instance, if $(X, Y, Z)'$ has a multivariate normal distribution then

$$\mathrm{E}[XYZ] = \mathrm{E}[X]\mathrm{E}[Y]\mathrm{E}[Z] + \mathrm{E}[X]\mathrm{Cov}[Y,Z] + \mathrm{E}[Y]\mathrm{Cov}[X,Z] + \mathrm{E}[Z]\mathrm{Cov}[X,Y], \tag{5.50}$$

with similar expressions for $\mathrm{E}[X^2Y]$, $\mathrm{E}[XY^2]$ etc. Substituting for the third order moments from these expressions, makes the system for the first and second moments closed and hence it can be solved at least numerically (see Section 5.3.1 for the validity of the normal approximation). The idea of using the normal distribution to approximate the distribution of $(X, Y, Z)'$ is based on a suggestion by Whittle (1957). Distributions other than the normal can also be used (for instance the Negative Binomial, see, e.g., Herbert 1998, Herbert & Isham 2000). Moreover it has been observed that there may be situations where it is unreasonable to assume that the vector $(X, Y, Z)'$ has a multivariate normal distribution, but the moments of X, Y, Z can be very well approximated by those of a multivariate normal using the formulae like (5.50) (see, e.g., Herbert 1998).

The linear approximation is another way to overcome this problem since it makes the stochastic model linear and hence the resulting system for the moments is closed (see, e.g., Tan & Hsu 1989 and Isham 1991 for applications of the linear approximation). In this section we present numerical results from both the linear and the normal approximation and compare them with results from the simulations. For the linear approximation, the system of equations (5.1) and (5.40) was solved numerically, since the mean of Y from the linear approximation is simply the deterministic y. For the normal approximation the third order moments were expressed in terms of the first and second moments and the resulting system was solved numerically.

For the results discussed in this section, we used the parameter values shown in (5.26), $n = 500$ and $n = 1000$, and for each of these two values of n the following sets of initial conditions:

$$\begin{aligned} y_0 &= 1\% \text{ of } n, \quad z_0 = 0 \\ y_0 &= 1\% \text{ of } n, \quad z_0 = 1\% \text{ of } n \end{aligned} \tag{5.51a}$$

$$\begin{aligned} y_0 &= 10\% \text{ of } n, \quad z_0 = 0 \\ y_0 &= 10\% \text{ of } n, \quad z_0 = 10\% \text{ of } n. \end{aligned} \tag{5.51b}$$

Also the case with $n = 500$ and $y_0 = 10\%$ of n, $z_0 = 0$ was examined with other sets of parameter values: (a) the values shown in (5.26) except $\alpha = 8$ and (b) the values shown in (5.26) except $\beta = 0.0012$. Figures 5.12 and 5.13 show the means of X, Y, Z and the standard deviation of Y for the cases with $n = 500$, $z_0 = 0$, and (a) $y_0 = 1\%$ of n, (b) $y_0 = 10\%$ of n, respectively. For the other cases the results were qualitatively similar (as

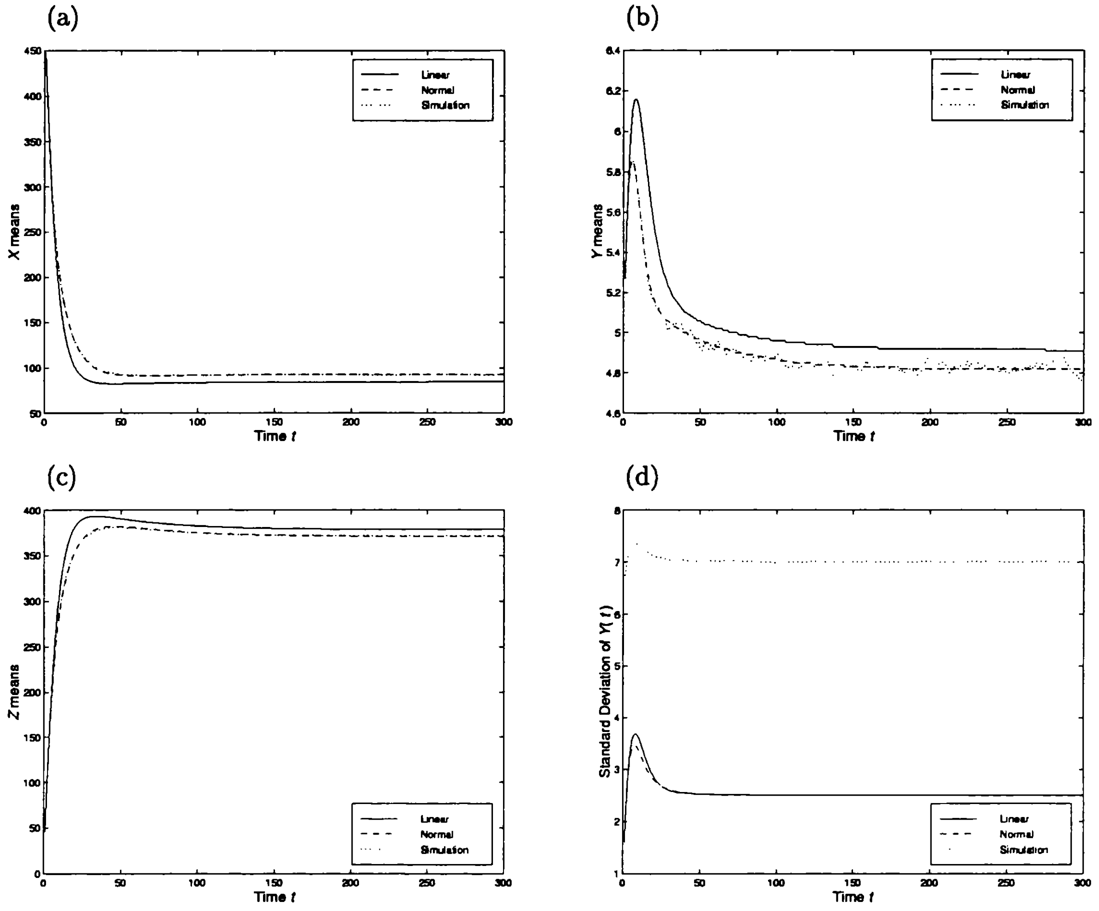


Figure 5.12: Results from the linear model, the normal approximation, and numerical simulations for (a), (b), (c) the mean of X, Y, Z, respectively, and (d) the standard deviation of Y. In each graph there are three curves, one for each of the methods mentioned above (linear, normal, simulation). The parameter values are as in (5.26). The initial conditions are $n = 500$, $y_0 = 5$, $z_0 = 0$. Time t is measured in years.

it will be explained later) and are not shown. The three analyses were carried out up to time $t = 300$ so that the results discussed here are for the time interval $I_t = [1, 300]$.

For the means the three analyses seem to agree quite well and in some cases the three curves can be hardly distinguished. In most cases, though, the linear approximation gives slightly different results, while there is closer agreement between the normal approximation and the simulations.

For the standard deviations there were two kinds of behaviour observed:

(i) in the cases with initial conditions (5.51a) (and for both $n = 500$ and $n = 1000$) the standard deviation from the simulations was quite a bit larger than that from the linear and the normal approximations, throughout the interval I_t. All three curves have a peak during the first 10–20 years, and after $t = 40$ they settle down.

(ii) in the cases with initial conditions (5.51b) (and for both $n = 500$ and $n = 1000$) the

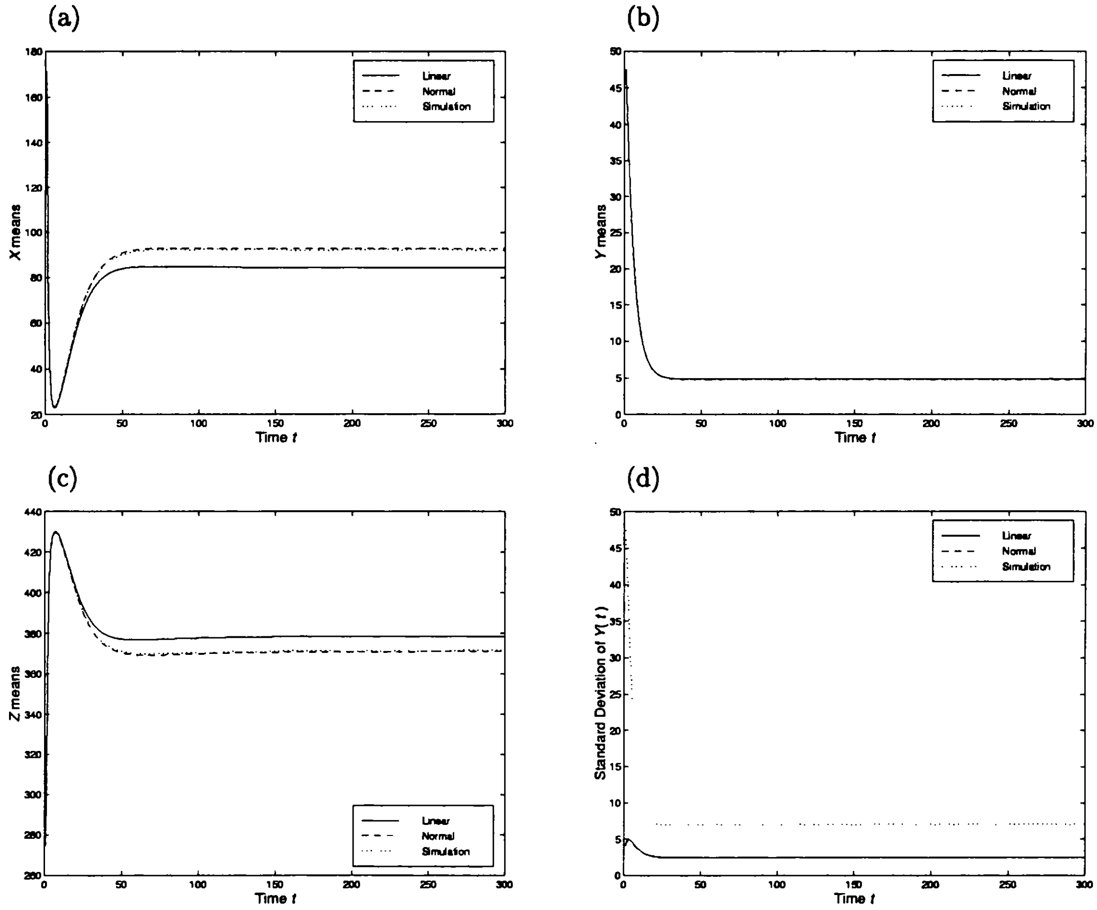


Figure 5.13: Results from the linear model, the normal approximation, and numerical simulations for (a), (b), (c) the mean of X, Y, Z, respectively, and (d) the standard deviation of Y. The parameter values are as in (5.26) and $n = 500$, $y_0 = 50$, $z_0 = 0$. Time t is measured in years.

behaviour is the same as for the cases with (5.51a), but in the beginning the standard deviation from the simulations is very much larger than that from the normal and linear approximations. The peak in the beginning (for all three curves) appears at an earlier time point (the first 5–10 years) and they settle down earlier (after $t = 25$).

It has to be noted though, that for each value of n the value of the standard deviation to which the curves finally settled down was the same with all initial conditions (5.51a) and (5.51b) (for instance with $n = 500$ and all four sets of initial conditions, the standard deviation from the simulations was around 7 at $t = 300$ and around 2.5 from the normal and the linear approximations).

The large standard deviation from the simulations (compared to that from the other two methods) in the beginning can be explained from the difference between the initial value y_0 and the final value y_e (at $t = 300$) of Y. For instance with $n = 500$ the value of the endemic equilibrium y_e (i.e. the deterministic equilibrium y_{e_2} and the mean of the quasi-stationary distribution of Y) is $y_e = 4.9$. With $y_0 = 5$ (1% of $n = 500$) in the beginning of the epidemic the value of the stochastic Y slightly increases (to a maximum around $y \approx 6$) from its initial value y_0 and then drops to $y_e = 4.9$. With $y_0 = 50$ (10% of n) the stochastic Y drops steeply to y_e, but the difference from $y_0 = 50$ to $y_e = 4.9$ is already quite large (and larger than the difference between the maximum $y \approx 6$ and $y_e = 4.9$ for the case with $y_0 = 5$). Therefore there will be more variation (between realisations) in the actual value of Y during the interval that it drops to y_e and the time it takes to "cover" this difference and hence a greater standard deviation during the beginning of the epidemic. Similarly for the cases with $n = 1000$, for which the value of y_e is around 9.8.

Nevertheless this variation cannot be "measured" in full by the linear approximation because in the linear model only the Y variable is stochastic, so that some of the randomness of the full stochastic model (which is reflected in the value of the standard deviation of Y) is not accounted for in the linear model, resulting in smaller standard deviation for Y. Also, the normal approximation seems to fail in these situations, since clearly it does not account for this variation in the value of Y (in the beginning of the epidemic) and agrees more with the linear approximation than with the simulations.

After the first 50 years, the value of the standard deviation has settled down in all cases. Its value is underestimated by the normal and the linear approximations until the end of the interval I_t. For the linear approximation this can be expected since, as was explained above, the linear model does not account for all the randomness of the stochastic model, and that will be reflected in the variation of Y (the only stochastic variable in the linear model) as well. For the normal approximation, it was expected that it would agree more with the simulations than with the linear approximation, but this is not the case here. One possible explanation for this is the fact that for the assumption for normality to be valid, it is not only n that has to be large, but also the proportions X/n, Y/n, Z/n. The proportions Y/n were around 1% in all cases examined here.

5.4.3 Discussion and conclusions

The results presented in this section suggest that the method of linearising a model like the one presented in this chapter can be helpful in deducing analytical as well as numerical results. For the numerical results, it appears that in some situations the linear approximation can give equally good estimates of the moments, as the normal approximation or even the numerical simulations. However, there is the disadvantage of a loss of information for the variation of the variables that are taken as deterministic, which can further affect the variation of the remaining stochastic variables (for instance, in the numerical results presented here, the linear model underestimates the standard deviation of the stochastic variable).

Also, care needs to be taken as to whether and how the analytical results from the linear model can be extended to the stochastic one. For example, for this model with the linear formulation, the epidemic ultimately dies out with probability one only when $\mathcal{R}_0 < 1$, but when $\mathcal{R}_0 > 1$ the probability of ultimate extinction is $q_0 < 1$ (see equation (5.49)). On the other hand, with the stochastic formulation the epidemic always dies out with probability one. For the simple underlying model discussed in this chapter, this substantial qualitative difference may not be so significant for any practical purposes, because in most cases of interest for TB, it is not the actual stochastic equilibrium (extinction) that will be observed, but the endemic quasi-stationary distribution; and there the two models agree. Nevertheless, in other situations differences like this may be significant and one should be cautious in interpreting the results from the linear model.

The linear approximation is used again for the model presented in the following chapter. Further remarks about the advantages and disadvantages of this method and the merits of each approximation can be found in the corresponding Section 6.3.8.

Chapter 6

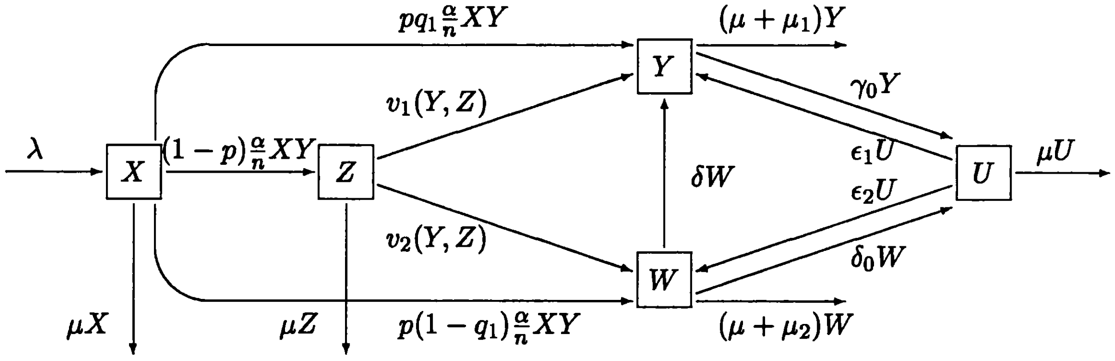
Model Zeus: a detailed model for TB

6.1 Introduction

The two models presented in Chapters 4 and 5 have accounted for the most important determinants of the spread of tuberculosis within a population and they provide a means of understanding the very basic intrinsic dynamics of the infection. Nevertheless there are some further interesting features of TB that, for simplicity, were not taken into account (for instance the possibility of reinfection of individuals with an old latent infection) which we are going to investigate in this chapter.

Also, our ultimate goal is to study the effects of the various control measures currently available (vaccination, chemoprophylaxis, and chemotherapy). With this goal in mind the model presented in Chapter 5 will now be modified to account for some features of TB that previously were either omitted or simplified. For instance in Chapter 5 we assumed that the population is divided into three classes: susceptibles, infectives, and inactives. Chemoprophylaxis is recommended only for those with a latent infection and close contacts of infectious cases. Since homogeneous mixing is assumed we will assume that chemoprophylaxis is used only for latents and therefore these individuals must be in a separate class in the model from the other inactive cases.

An interesting question from a public-health point of view is with respect to the effect of treating non-infectious cases (which constitute about half of the TB cases – see Chapter 2). In order to answer this question the number of non-infectious cases must



$$v_1(Y,Z) = q_2\beta Z + q_3\tfrac{\alpha_2}{n}YZ \qquad v_2(Y,Z) = (1-q_2)\beta Z + (1-q_3)\tfrac{\alpha_2}{n}YZ$$

Figure 6.1: A detailed model for TB: Model Zeus

be known.

Therefore now we consider a population subject to homogeneous mixing and divided into five classes:

(a) *uninfected:* individuals who have never been infected with TB

(b) *latents:* individuals who have been infected (at least once) but the infection has remained latent (they are non-infectious and healthy)

(c) *infectious TB cases:* individuals who have developed clinical disease and are infectious (smear-positive TB cases)

(d) *non-infectious cases:* individuals who have developed clinical disease but they are not infectious (this class includes all the smear-negative TB cases: patients with non-pulmonary TB and those with smear-negative pulmonary TB)

(e) *recovered:* individuals who have developed clinical TB and recovered spontaneously without treatment

The sizes of these classes at time t will be denoted by $X(t)$, $Z(t)$, $Y(t)$, $W(t)$, $U(t)$, respectively, and the size of the population by $N(t) = X(t) + Y(t) + Z(t) + W(t) + U(t)$ (for simplicity these classes will be sometimes referred to as the X class, the Z class, and so on). The initial sizes of the five classes are $X(0) = x_0$, $Y(0) = y_0$, $Z(0) = z_0$, $W(0) = w_0$, $U(0) = u_0$, and $N(0) = x_0 + y_0 + z_0 + w_0 + u_0 = n$, where $1 \le x_0 \le n-1$, $y_0 \ge 0$, $z_0 \ge 0$, $w_0 \ge 0$, $u_0 \ge 0$, and $n \ge 2$. Occasionally we will use the notation $\mathbf{X}(t)$ and $\mathbf{x}$ for the vectors $(X(t), Y(t), Z(t), W(t), U(t))$ and (x, y, z, w, u), respectively.

If at time t there are $X(t)$ uninfected and $Y(t)$ infectious cases in the population then the probability of one new infection occurring in the interval $[t, t + dt]$ is $\alpha X(t)Y(t)dt/n + o(dt)$ where α is the effective contact rate (as was explained in Section 4.1). Among those who get infected a proportion p develop clinical TB within a year after infection (primary TB) and the remaining proportion, $1 - p$, become latents; those who develop TB are infectious or non-infectious with probabilities q_1 and $1 - q_1$, respectively. The difference between primary and secondary TB (i.e. whether an infected develops TB within a year or later) could be modelled with a time-delay model, but in this chapter we will not investigate this possibility and we keep the structure of a basic Markov process.

Latents may develop clinical disease at some point as a result of endogenous reactivation of an old infection or exogenous reinfection (acquiring a new infection). There have been doubts about whether exogenous reinfection is possible for TB (see Section 2.3) but in the recent literature it has become more certain that reinfection is possible and should be taken into account, especially in areas with high risk of infection (see, e.g., Vynnycky 1996, Dye et al. 1998). Therefore, the possibility of exogenous reinfection for the latents is included in this model. On the other hand, we assume that reinfection is possible only for the latents and not for the non-infectious and recovered cases, since the relapse rates from these two classes (to the infectious class) are very high, so that the effect of reinfection is negligible for these cases.

The reactivation rate is denoted by β so that the probability of an endogenous reactivation occurring in $[t, t+dt]$ is $\beta Z(t)dt+o(dt)$. After reactivation the individual has infectious or non-infectious TB with probability q_2 and $1-q_2$, respectively. For exogenous reactivation we assume that the effective contact rate between latents and infectious cases is $p_r\alpha$, where $0 \leq p_r \leq 1$. If $p_r = 0$ then reinfection is not possible; if $p_r = 1$ this means that past infections confer no immunity at all, so that latents are equally likely to get infected as the uninfected. This is not the case with TB, since infection does provide immunity (at least partial and/or temporal) and therefore p_r must be strictly less than 1. A more realistic approach would be to assume that p_r is an increasing function of the time since infection since most results (see, e.g., Styblo 1991, Dye et al. 1998) suggest that immunity conferred by an old infection wanes in time. This approach though would increase substantially the complexity of the model (since that implies keeping track of the time since infection for each infected individual) and therefore for simplicity we assume that p_r is constant. After reinfection an individual develops clinical disease within a year (primary TB) or remains latent with probabilities p_3, $1 - p_3$, respectively; those who develop TB are infectious or non-infectious with probabilities q_3, $1-q_3$, respectively. For simplicity we will denote $\alpha_2 = p_3 p_r \alpha$ so that the probability of a reinfection leading to primary TB occurring in $[t, t+dt]$ is $\alpha_2 Z(t)Y(t)dt/n + o(dt)$. We assume that additional infections do not change the reactivation rate β or the effective contact rate α_2.

Non-infectious TB cases become infectious at a rate δW. Infectious and non-infectious cases recover spontaneously at rates γ_0 and δ_0 per capita, respectively, and those who have recovered may relapse later and become infectious or non-infectious cases at rates $\epsilon_1 U$ and $\epsilon_2 U$, respectively.

Finally, there is immigration of susceptibles at a constant rate λ, normal death at rate μ per capita, and excess death due to TB at rates μ_1 and μ_2 (per capita) for the infectious and non-infectious cases, respectively. Individuals with latent infection and those who have recovered are healthy and hence there is no excess death for these two classes. At some points the special case $\lambda = \mu n$ will be investigated.

The definitions of the variables and parameters used for this model are summarised in Table 6.1. The possible transitions and their rates are illustrated in Figure 6.1.

It has to be noted that this formulation assumes that the values of q_1, q_2, q_3 may be different in general. This means that when an individual develops clinical disease the probability that the form of disease will be infectious or non-infectious depends on

(a) whether the individual had an old infection or not (q_2, q_3 for the former, q_1 for the latter)

(b) for an individual who had an old infection, whether the current incidence of disease is a result of the old infection (endogenous reactivation) or of a new additional infection (exogenous reinfection); the probabilities are q_2 and q_3, respectively.

In the literature there is not enough evidence to prove either that the q_1, q_2, q_3 are equal or that they are not, and therefore modellers have taken either line (for instance, Dye et al. (1998) assumed that $q_1 = q_2 = q_3$; Blower et al. (1995) assumed that q_1 and q_2 are not equal and $q_3 = 0$). In this chapter we have used the three different parameters q_1, q_2, q_3, since that allows for both approaches to be adapted, and in some cases we investigate the situation $q_1 = q_2 = q_3$ as a special case.

$X(t)$	Number of uninfected individuals at time t
$Y(t)$	Number of infectious TB cases at time t
$Z(t)$	Number of latents at time t
$W(t)$	Number of non-infectious TB cases at time t
$U(t)$	Number of naturally recovered patients at time t
λ	Immigration of uninfected individuals
μ	Normal death rate (per capita)
μ_1	Excess death rate due to TB for infectious cases (per capita)
μ_2	Excess death rate due to TB for non-infectious cases (per capita)
α	The effective contact rate between uninfected and infectious cases
p	Probability of developing primary TB (after first infection)
q_1	Probability of developing infectious TB for those with primary TB (after the first infection)
β	Reactivation rate for the latents
q_2	Probability that reactivation leads to infectious TB
$p_r\alpha$	The effective contact rate between latents and infectious cases
p_3	Probability of developing primary TB (after reinfection)
α_2	$\alpha_2 = p_3 p_r \alpha$
q_3	Probability of developing infectious TB for those with primary TB (after reinfection)
δ	Rate at which non-infectious cases become infectious
γ_0	Natural recovery rate for infectious cases
δ_0	Natural recovery rate for non-infectious cases
ϵ_1	Relapse rate to the infectious class (for those naturally recovered)
ϵ_2	Relapse rate to the non-infectious class (for those naturally recovered)
n	Initial total population size

Table 6.1: Variables and parameters used in model Zeus

The same problem arises for p and p_3. Assuming that they are not equal means that the probability that an individual develops primary TB after infection depends on whether this is the first infection or a superinfection. The mathematics here are carried out with the two distinct parameters and the situation $p = p_3$ is studied as a special case.

6.2 The deterministic model

For the corresponding deterministic model, let $x(t)$, $y(t)$, $z(t)$, $w(t)$, and $u(t)$ denote the number of uninfected, infectious cases, latents, non-infectious cases, and recovered, respectively, at time t. The differential equations for x, y, z, w, and u are:

$$\begin{aligned}\frac{dx}{dt} &= -\frac{\alpha}{n}xy - \mu x + \lambda\\ \frac{dy}{dt} &= pq_1\frac{\alpha}{n}xy + q_3\frac{\alpha_2}{n}yz - \Gamma_s y + q_2\beta z + \delta w + \epsilon_1 u\\ \frac{dz}{dt} &= (1-p)\frac{\alpha}{n}xy - \frac{\alpha_2}{n}yz - \Phi_s z\\ \frac{dw}{dt} &= p(1-q_1)\frac{\alpha}{n}xy + (1-q_3)\frac{\alpha_2}{n}yz + (1-q_2)\beta z - \Delta_s w + \epsilon_2 u\\ \frac{du}{dt} &= \gamma_0 y + \delta_0 w - \mathrm{E}_s u,\end{aligned} \tag{6.1}$$

where, for simplicity, the parameters have been grouped as

$$\begin{aligned}\Gamma_s &= \gamma_0 + \mu + \mu_1\\ \Delta_s &= \delta + \delta_0 + \mu + \mu_2\\ \mathrm{E}_s &= \epsilon_1 + \epsilon_2 + \mu\\ \Phi_s &= \beta + \mu.\end{aligned} \tag{6.2}$$

Here x, y, z, w, and u are non-negative continuous functions. The initial conditions are $(x(0), y(0), z(0), w(0), u(0)) = (x_0, y_0, z_0, w_0, u_0) \in \mathcal{S}_0$, where

$$\mathcal{S}_0 = \{\mathbf{x} = (x, y, z, w, u) \in \mathbb{Z}_+^5 : 1 \le x \le n-1, x+y+z+w+u = n\}, \tag{6.3}$$

where $n \ge 2$ is the initial total population size: $n = x_0 + y_0 + z_0 + w_0 + u_0$. From the system (6.1) it follows that the total population size, $N(t) = x(t)+y(t)+z(t)+w(t)+u(t)$, satisfies the following equation:

$$\frac{dN(t)}{dt} = \lambda - \mu N(t) - \mu_1 y(t) - \mu_2 w(t). \tag{6.4}$$

With integration this gives

$$N(t) = \frac{\lambda}{\mu} + e^{-\mu t}\left(n - \frac{\lambda}{\mu}\right) - e^{-\mu t}\int_0^t e^{\mu s}[\mu_1 y(s) + \mu_2 w(s)]ds, \tag{6.5}$$

for $t \ge 0$. From this equation it follows that $N(t)$, and hence $x(t)$, $y(t)$, $z(t)$, $w(t)$, and $u(t)$ as well, are always bounded above by n if $\lambda \le \mu n$ and by λ/μ if $\lambda > \mu n$.

Solving the system (6.1) with the derivatives on the left-hand side equal to zero, we see that the system (6.1) admits three possible equilibria, $\mathbf{e}_i = (x_i^e, y_i^e, z_i^e, w_i^e, u_i^e)$, for $i = 1, 2, 3$. The first of these points, $\mathbf{e}_1$, corresponds to the extinction of the infection: $\mathbf{e}_1 = (\lambda/\mu, 0, 0, 0, 0)$, while the other two points, $\mathbf{e}_2$ and $\mathbf{e}_3$, are defined by

$$x_2^e = \frac{-\varphi_2 + \sqrt{\mathcal{D}}}{2\varphi_1} \text{ and } x_3^e = \frac{-\varphi_2 - \sqrt{\mathcal{D}}}{2\varphi_1}, \tag{6.6}$$

where $\mathcal{D} = \varphi_2^2 - 4\varphi_1\varphi_3$, and for $i = 2, 3$

$$\begin{aligned} y_i^e &= \frac{\lambda - \mu x_i^e}{\alpha x_i^e/n} \\ z_i^e &= \frac{(1-p)\frac{\alpha}{n}x_i^e y_i^e}{\frac{\alpha_2}{n}y_i^e + \beta + \mu} \\ w_i^e &= \frac{\mathrm{E}_s}{\Delta_s\mathrm{E}_s - \delta_0\epsilon_2}\left[p(1-q_1)\frac{\alpha}{n}x_i^e y_i^e + (1-q_3)\frac{\alpha_2}{n}y_i^e z_i^e + \frac{\gamma_0\epsilon_2}{\mathrm{E}_s}y_i^e + (1-q_2)\beta z_i^e\right] \\ u_i^e &= \frac{1}{\mathrm{E}_s}(\gamma_0 y_i^e + \delta_0 w_i^e), \end{aligned} \tag{6.7}$$

where

$$\begin{aligned} \varphi_1 &= \frac{\alpha}{n}(\delta\mathrm{E}_s + \delta_0\epsilon_1)^{-1}\left\{\frac{\alpha}{n}(1-p)(\beta d_2 - \Phi_s d_3) + (\frac{\alpha}{n}\Phi_s - \mu\frac{\alpha_2}{n})[pd_1 + (1-p)d_3]\right\} \\ \varphi_2 &= \frac{\lambda\frac{\alpha}{n}\frac{\alpha_2}{n}[pd_1 + (1-p)d_3] - (\frac{\alpha}{n}\Phi_s - \mu\frac{\alpha_2}{n})[\Gamma_s(\Delta_s\mathrm{E}_s - \delta_0\epsilon_2) - \gamma_0(\delta\epsilon_2 + \epsilon_1\Delta_s)]}{\delta\mathrm{E}_s + \delta_0\epsilon_1} \\ \varphi_3 &= -\lambda\frac{\alpha_2}{n}\frac{[\Gamma_s(\Delta_s\mathrm{E}_s - \delta_0\epsilon_2) - \gamma_0(\delta\epsilon_2 + \epsilon_1\Delta_s)]}{\delta\mathrm{E}_s + \delta_0\epsilon_1}, \end{aligned} \tag{6.8}$$

and

$$d_j = q_j(\Delta_s\mathrm{E}_s - \delta_0\epsilon_2) + (1-q_j)(\delta\mathrm{E}_s + \delta_0\epsilon_1), \;\; j = 1, 2, 3.$$

Depending on the parameter values, the coordinates of $\mathbf{e}_2$ and $\mathbf{e}_3$ may be positive, negative, or complex. So we will call an equilibrium point $\mathbf{e}$ feasible if all its coordinates are non-negative. $\mathbf{e}_1$ is always feasible. Table 6.2 shows when $\mathbf{e}_2$ and $\mathbf{e}_3$ are feasible, where $\mathcal{R}_0$ is the basic reproduction ratio, defined by

$$\mathcal{R}_0 = \frac{\alpha}{n}\frac{\lambda/\mu}{\Phi_s}\frac{p\Phi_s d_1 + (1-p)\beta d_2}{[\Gamma_s(\Delta_s\mathrm{E}_s - \delta_0\epsilon_2) - \gamma_0(\delta\epsilon_2 + \epsilon_1\Delta_s)]}, \tag{6.9}$$

(see Section A.3.1 for the calculation of $\mathcal{R}_0$) and θ_1 is defined (whenever $\mathcal{R}_0 < 1$) as

$$\theta_1 = \frac{[\Gamma_s(\Delta_s\mathrm{E}_s - \delta_0\epsilon_2) - \gamma_0(\delta\epsilon_2 + \epsilon_1\Delta_s)]\left[\sqrt{\mu\frac{\alpha_2}{n}} + \sqrt{\frac{\alpha}{n}\Phi_s(1-\mathcal{R}_0)}\right]^2}{(\delta\mathrm{E}_s + \delta_0\epsilon_1)\lambda/\mu}. \tag{6.10}$$

When $\mathcal{R}_0 > 1$ the system has only two possible equilibria, the disease-free equilibrium $\mathbf{e}_1$ and the endemic equilibrium $\mathbf{e}_2$, as is the case with most epidemic models. When $\mathcal{R}_0 < 1$ though, we have a situation that is not that common in epidemic modelling: there is a subset of the parameter space where the system admits three feasible equilibria: the disease-free one and two endemic. In the remaining of this section we will study the stability of these critical points for the cases with $\mathcal{R}_0 \neq 1$ (when $\mathcal{R}_0 = 1$ the study of the stability is more complicated and since $\mathcal{R}_0$ is a function of 15 parameters,

Conditions on $\mathcal{R}_0$	Other conditions[a]	Equilibrium points
$\mathcal{R}_0 > 1$		$\mathbf{e}_1$, $\mathbf{e}_2$
$\mathcal{R}_0 < 1$	$\varphi_1 < -\theta_1$	$\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$
	$\varphi_1 = -\theta_1$	$\mathbf{e}_1$, $\mathbf{e}_2$
	$\varphi_1 > -\theta_1$	$\mathbf{e}_1$
$\mathcal{R}_0 = 1$	$\varphi_1 > 0$, $\mathcal{C}_\varphi > 0$	$\mathbf{e}_1$, $\mathbf{e}_3$
	$\varphi_1 > 0$, $\mathcal{C}_\varphi < 0$	$\mathbf{e}_1$, $\mathbf{e}_2$
	$\varphi_1 > 0$, $\mathcal{C}_\varphi = 0$	$\mathbf{e}_1$
	$\varphi_1 < 0$, $\mathcal{C}_\varphi < 0$	$\mathbf{e}_1$, $\mathbf{e}_2$
	$\varphi_1 < 0$, $\mathcal{C}_\varphi \geq 0$	$\mathbf{e}_1$
	$\varphi_1 = 0$	$\mathbf{e}_1$

[a] φ_1 and θ_1 are defined in (6.8) and (6.10), respectively, and $\mathcal{C}_\varphi$ is defined by $\mathcal{C}_\varphi = \varphi_2 + 2\varphi_1\lambda/\mu$.

Table 6.2: The distinct and feasible equilibria of the deterministic model

it is highly unlikely that it can be exactly equal to 1; therefore the cases with $\mathcal{R}_0 = 1$ have not been studied here).

The system (6.1) can be written in the form

$$\frac{d\mathbf{v}}{dt} = F(\mathbf{v}),$$

where $\mathbf{v}(t) = (x(t), y(t), z(t), w(t), u(t))$ for $t \geq 0$ and F is a mapping from $\mathbb{R}^5_+$ into $\mathbb{R}^5$ with coordinates $f_i(x, y, z, w, u)$, $i = 1, 2, \ldots, 5$ given by $f_1(\mathbf{v}) = dx/dt$, $f_2(\mathbf{v}) = dy/dt$, $f_3(\mathbf{v}) = dz/dt$, $f_4(\mathbf{v}) = dw/dt$, $f_5(\mathbf{v}) = du/dt$ and the derivatives of x, y, z, w, and u as defined in (6.1).

For the equilibrium points $\mathbf{e}_k$, $k = 1, 2, 3$, let $DF(\mathbf{e}_k)$ denote the Jacobian matrix of F at the point $\mathbf{e}_k$, i.e. the matrix whose (i, j) element is $\partial f_i(\mathbf{e}_k)/\partial j$ for $i = 1, \ldots, 5$ and $j = x, y, z, w, u$. Then $DF(\mathbf{e}_k)$ is defined by

$$\begin{bmatrix} -\mu - \frac{\alpha}{n}y^e_k & -\frac{\alpha}{n}x^e_k & 0 & 0 & 0 \\ pq_1\frac{\alpha}{n}y^e_k & pq_1\frac{\alpha}{n}x^e_k + q_3\frac{\alpha_2}{n}z^e_k - \Gamma_s & q_3\frac{\alpha_2}{n}y^e_k + q_2\beta & \delta & \epsilon_1 \\ (1-p)\frac{\alpha}{n}y^e_k & (1-p)\frac{\alpha}{n}x^e_k - \frac{\alpha_2}{n}z^e_k & -\frac{\alpha_2}{n}y^e_k - \Phi_s & 0 & 0 \\ p(1-q_1)\frac{\alpha}{n}y^e_k & p(1-q_1)\frac{\alpha}{n}x^e_k + (1-q_3)\frac{\alpha_2}{n}z^e_k & (1-q_3)\frac{\alpha_2}{n}y^e_k + (1-q_2)\beta & -\Delta_s & \epsilon_2 \\ 0 & \gamma_0 & 0 & \delta_0 & -\mathrm{E}_s \end{bmatrix},$$

where $\mathbf{e}_k = (x^e_k, y^e_k, z^e_k, w^e_k, u^e_k)$ for $k = 1, 2, 3$ as defined in (6.6) and (6.7).

The characteristic polynomial $P_1(\tau)$ of $DF(\mathbf{e}_1)$ is

$$P_1(\tau) = -(\mu + \tau)(\tau^4 + M_{11}\tau^3 + M_{12}\tau^2 + M_{13}\tau + M_{14}), \tag{6.11}$$

where the M_{11}, M_{12}, M_{13}, M_{14} are functions of the parameters (see Section A.3.1). According to the Routh-Hurwitz criterion (see Theorem 5.1) $\mathbf{e}_1$ is stable if the following four quantities are positive:

$$
\begin{aligned}
D_1 &= M_{11} & \qquad D_3 &= M_{13}D_2 - M_{11}^2 M_{14} \\
D_2 &= M_{11}M_{12} - M_{13} & \qquad D_4 &= M_{14}D_3.
\end{aligned}
$$

After some calculations it can be shown that if $\mathcal{R}_0 < 1$ then $D_i > 0$ for all $i = 1, 2, 3, 4$ so that $\mathbf{e}_1$ is stable, but if $\mathcal{R}_0 > 1$ then at least one of the D_i is negative and $\mathbf{e}_1$ is unstable. If $\mathcal{R}_0 = 1$ then $D_4 = 0$ and the criterion does not apply.

The characteristic polynomials $P_2(\tau)$ and $P_3(\tau)$ of $DF(\mathbf{e}_2)$ and $DF(\mathbf{e}_3)$, respectively, are

$$
P_i(\tau) = -(\tau^5 + M_{i1}\tau^4 + M_{i2}\tau^3 + M_{i3}\tau^2 + M_{i4}\tau + M_{i5}), \;\; i = 2, 3, \tag{6.12}
$$

where the $M_{i1}, \ldots, M_{i5}$ are functions of the parameters (see Section A.3.1). According to the Routh-Hurwitz criterion (Theorem 5.1) $\mathbf{e}_i$ is stable if the following five quantities are positive:

$$
\begin{aligned}
D_{i1} &= M_{i1} \\
D_{i2} &= M_{i1}M_{i2} - M_{i3} \\
D_{i3} &= M_{i3}D_{i2} - M_{i1}(M_{i4}M_{i1} - M_{i5}) \\
D_{i4} &= M_{i4}D_{i3} - M_{i5}[M_{i1}(M_{i2}^2 - M_{i4}) - M_{i2}M_{i3} + M_{i5}] \\
D_{i5} &= M_{i5}D_{i4}.
\end{aligned}
$$

The coefficient M_{35} can be written as

$$
M_{35} = -(\delta \mathrm{E}_s + \delta_0 \epsilon_1)\frac{y_3^e}{x_3^e}[-\varphi_1(x_3^e)^2 + \varphi_3],
$$

and from that it can be shown that if $\mathcal{R}_0 < 1$ and $\varphi_1 < -\theta_1$ then $M_{35} < 0$ and therefore at least one of the D_{34}, D_{35} is negative and hence $\mathbf{e}_3$ is unstable. Therefore in the subspace of the parameter space where $\mathbf{e}_3$ is feasible (when $\mathcal{R}_0 < 1$ and $\varphi_1 < -\theta_1$) it is also unstable. If we define

$$
\begin{aligned}
\mathcal{D} &= \varphi_2^2 - 4\varphi_1\varphi_3 \\
\mathcal{R}_1 &= \frac{2\varphi_1\lambda/\mu}{-\varphi_2 + \sqrt{\mathcal{D}}},
\end{aligned}
\tag{6.13}
$$

then the second equilibrium point $\mathbf{e}_2$ is $\mathbf{e}_2 = (x_2^e, y_2^e, z_2^e, w_2^e, u_2^e)$ where

$$x_2^e = \frac{\lambda/\mu}{\mathcal{R}_1} \tag{6.14}$$

$$y_2^e = \frac{\mu}{\alpha/n}(\mathcal{R}_1 - 1), \tag{6.15}$$

and z_2^e, w_2^e, u_2^e as defined in (6.7). Also it can be shown that the conditions for the feasibility of $\mathbf{e}_2$ (shown in Table 6.2) are equivalent to the condition $\mathcal{R}_1 > 1$: if $\mathcal{R}_1 > 1$ then $\mathbf{e}_2$ is feasible; if $\mathcal{R}_1 = 1$ then $\mathbf{e}_2 = \mathbf{e}_1$; otherwise (if $\mathcal{R}_1 < 1$ or $\mathcal{R}_1$ is not real) then $\mathbf{e}_2$ is not feasible. It can be shown that if $\mathcal{R}_1 > 1$ then $D_{21} > 0$ and $M_{25} > 0$. So it only remains to show whether (or when) D_{22}, D_{23}, and D_{24} are positive. Unfortunately the algebra involved in these calculations does not allow us to investigate the signs of these quantities. Further results can possibly be deduced using a computer algebra package, although here we will try a numerical approach instead.

First of all it has to be noted that the expression (6.14) for x_2^e is quite common in epidemic models (see, e.g., Jacquez & Simon 1993), where instead of $\mathcal{R}_1$ it is usually $\mathcal{R}_0$, the basic reproduction ratio, that appears in the denominator of (6.14). When $\lambda = \mu n$ equation (6.14) gives an approximation to the proportion of uninfected individuals

$$\frac{x_2^e}{n} = \frac{1}{\mathcal{R}_1}.$$

Because of the form of $\mathbf{e}_2$ and the way it is defined we suspect that $\mathbf{e}_2$ is stable if $\mathcal{R}_1 > 1$ and unstable if $\mathcal{R}_1 < 1$. One way to assess this numerically is by minimising the quantities D_{22}, D_{23}, D_{24} over the whole parameter space and under the condition $\mathcal{R}_1 > 1$. If the minimum values of D_{22}, D_{23}, D_{24} are positive in this subspace (where $\mathcal{R}_1 > 1$) then D_{22}, D_{23}, D_{24} are positive whenever $\mathcal{R}_1 > 1$ and hence $\mathbf{e}_2$ is stable. Details for the implementation of the numerical minimisation can be found in the Appendix (Section A.3.1).

The minimum values for D_{22}, D_{23}, D_{24} were found to be positive, but floating underflow occurred during the implementation of the method (i.e. some of the quantities calculated during the procedure were smaller than the smallest number that the machine recognises). This fact does not necessarily invalidate the final result (that the minimum values are positive) but certainly raises serious questions for its validity. Also the fact that the numerical minimisation did not give negative values for the D_{22}, D_{23}, D_{24} means that at least for some combination of parameters with $\mathcal{R}_1 > 1$, $\mathbf{e}_2$ is indeed stable. We will investigate this point in a little more detail.

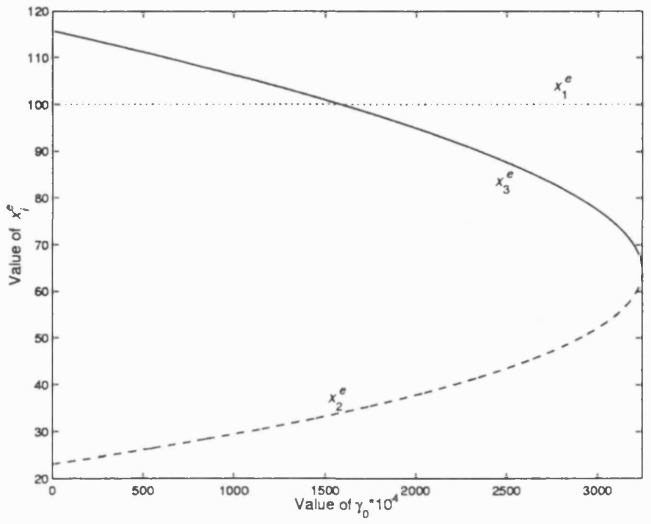


Figure 6.2: The values of x_1^e, x_2^e, x_3^e for a particular set of parameter values: $n = 100$, $\mu = 0.0222$, $\mu_1 = 0.3$, $\mu_2 = 0.21$, $\lambda = \mu n$, $q_1 = 0.55$, $q_2 = 0.55$, $q_3 = 0.055$, $p = 0.1$, $p_r = 0.6$, $\alpha = 6$, $\alpha_2 = p_r\alpha$, $\beta = 0.0001$, $\delta = 0.015$, $\delta_0 = 0.2$, $\epsilon_1 = 0.03$, $\epsilon_2 = 0.03$, and γ_0 varies over $\gamma_0 = 0.0001, 0.0002, \ldots, 0.3243, 0.3244$. With $\gamma_0 \geq 0.3245$, x_2^e and x_3^e are not real and $x_1^e = 100$. Table 6.3 shows when the points x_1^e, x_2^e, and x_3^e are feasible and stable.

The quantities $D_{21}, \ldots, D_{25}$, $\mathcal{R}_0$, and $\mathcal{R}_1$ were iteratively calculated with $\gamma_0 = 0.0001, 0.0002, \ldots, 0.9999, 1.0000$ and the other parameters equal to the following values

$$\begin{array}{cccc}
\mu = 0.0222 & q_1 = 0.55 & p = 0.1 & \delta = 0.015 \\
\lambda = \mu n & q_2 = 0.55 & p_r = 0.6 & \epsilon_1 = 0.03 \\
\mu_1 = 0.3 & q_3 = 0.055 & \alpha_2 = p_r\alpha & \epsilon_2 = 0.03 \\
\mu_2 = 0.21 & \alpha = 6 & \beta = 0.0001 & \delta_0 = 0.2
\end{array} \tag{6.16}$$

(and $n = 100$). These parameter values are not all representative of TB (see Table 6.8 for the values that are representative for TB), but they are used here in order to show that $\mathbf{e}_2$ can be stable even when $\mathcal{R}_0 < 1$. The following results were deduced:

(a) For $\gamma_0 = 0.0001, 0.0002, \ldots, 0.1587$: both $\mathcal{R}_0$ and $\mathcal{R}_1$ are greater than 1; x_3^e is greater than $100 = \lambda/\mu$ and hence $\mathbf{e}_3$ is not feasible; x_2^e is between 0 and 100 and $\mathbf{e}_2$ is feasible (the values of y_2^e, z_2^e, w_2^e, u_2^e are not shown here); the values of $D_{21}, \ldots, D_{25}$ are all positive and hence $\mathbf{e}_2$ is stable.

(b) For $\gamma_0 = 0.1588, 0.1589, \ldots, 0.3244$: $\mathcal{R}_0 < 1$ and $\mathcal{R}_1 > 1$ ($\varphi_1 < -\theta_1$); both $\mathbf{e}_2$ and $\mathbf{e}_3$ are feasible; D_{35} is negative, so $\mathbf{e}_3$ is unstable; all $D_{21}, \ldots, D_{25}$ are positive and hence $\mathbf{e}_2$ is stable.

(c) For $\gamma_0 = 0.3245, 0.3246, \ldots, 1.0000$: $\mathcal{R}_0 < 1$ and $\mathcal{R}_1$ is not real; both x_2^e and x_3^e are

Value of γ_0	Stability
$0.0001, \ldots, 0.1587$	$\mathcal{R}_0 > 1$, $\mathcal{R}_1 > 1$ $\mathbf{e}_1$ unstable $\mathbf{e}_2$ feasible, stable $\mathbf{e}_3$ infeasible
$0.1588, \ldots, 0.3244$	$\mathcal{R}_0 < 1$, $\mathcal{R}_1 > 1$ $\mathbf{e}_1$ stable $\mathbf{e}_2$ feasible, stable $\mathbf{e}_3$ feasible, unstable
$0.3245, \ldots, 1$	$\mathcal{R}_0 < 1$, $\mathcal{R}_1$ not real $\mathbf{e}_1$ stable $\mathbf{e}_2$ infeasible $\mathbf{e}_3$ infeasible

Table 6.3: Feasibility and stability of the deterministic equilibria in a subspace of the parameter space: grid for γ_0

not real and hence $\mathbf{e}_2$ and $\mathbf{e}_3$ are not feasible.

Figure 6.2 shows the values of x_1^e, x_2^e, x_3^e as γ_0 varies from 0.0001 to 0.3244 (when $\gamma_0 \geq 0.3245$ the values of x_2^e, x_3^e are not real).

These observations (which are summarised in Table 6.3) show that indeed there is a subspace of the parameter space where $\mathcal{R}_0$ is less than one but still the endemic equilibrium is stable. Of course this is local stability since the disease-free equilibrium is also stable in the area where $\mathcal{R}_0 < 1$ and $\mathcal{R}_1 > 1$. We investigated further this behaviour of $\mathbf{e}_2$ by taking another grid of the parameter space, this time centred around the values which are representative for TB. For this grid we used $n = 100$, $\mu = 0.02$, $\lambda = \mu n$, $\alpha_2 = p_r \alpha$ and for the other parameters the range of values shown in Table 6.4. All the combinations of parameters in these ranges were tried and for each combination the values of $D_{21}, \ldots, D_{25}$, $\mathcal{R}_0$, and $\mathcal{R}_1$ were calculated. The result deduced was the same as above: whenever $\mathcal{R}_1 > 1$ (even if $\mathcal{R}_0 < 1$) $\mathbf{e}_2$ is stable.

It goes without saying that the results aforementioned do not constitute proof of our assertion that $\mathbf{e}_2$ is stable whenever $\mathcal{R}_1 > 1$ (and not only when $\mathcal{R}_0 > 1$), but they do give positive indications that this could be the case, and most of all they do prove that the endemic equilibrium $\mathbf{e}_2$ can be stable even when $\mathcal{R}_0 < 1$. This behaviour is not common in epidemic models, although in the recent years it has appeared in the literature (see, e.g., Castillo-Chavez & Feng 1997, Kribs-Zaleta & Velasco-Hernández 2000, Kribs-Zaleta 2001, and their references), partly because modellers have turned to

Parameter	Range of values
$q_1 = q_2 = q_3$	$0.50, 0.51, \ldots, 0.60$
p	$0.05, 0.06, \ldots, 0.20$
$\gamma_0 = \delta_0$	$0.055, 0.056, \ldots, 0.075$
$\epsilon_1 = \epsilon_2$	$0.010, 0.011, \ldots, 0.020$
p_r	$0.30, 0.31, \ldots, 0.60$
β	$0.0020, 0.0021, \ldots, 0.0040$
δ	$0.010, 0.011, \ldots, 0.020$
μ_1	$0.10, 0.11, \ldots, 0.20$
μ_2	$0.08, 0.09, \ldots, 0.15$
α	$8, 9, \ldots, 13$

Table 6.4: Grid of the parameter space centred around values representative for TB

more complicated models, which in some cases yield non-trivial behaviours.

The question that naturally arises from this behaviour is what happens when both $\mathbf{e}_1$ and $\mathbf{e}_2$ are stable. When two equilibrium points are locally stable, each one has a domain of attraction, say $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, so that if the vector of initial conditions $\mathbf{x}_0 = (x_0, y_0, z_0, w_0, u_0)$ belongs to $\mathcal{A}_1$ then the vector $\mathbf{x}(t)$ tends to $\mathbf{e}_1$ as t tends to infinity, while if $\mathbf{x}_0 \in \mathcal{A}_2$ then $\mathbf{x}(t)$ tends to $\mathbf{e}_2$.

The deterministic system (6.1) was solved numerically with various initial conditions $\mathbf{x}_0$ and with the parameter values shown in (6.16) and $\gamma_0 = 0.2$. For these parameter values $\mathcal{R}_0 = 0.9475$, $\mathcal{R}_1 = 2.6481$, and both $\mathbf{e}_1$ and $\mathbf{e}_2$ are stable. With $z_0 = 1$, $x_0 = n-1$ the system tends to extinction, while with $y_0 = 1$, $x_0 = n-1$; $w_0 = 1$, $x_0 = n-1$; and $u_0 = 1$, $x_0 = n-1$ it tends to the endemic equilibrium. These four points $\mathbf{x}_0$ are the ones that are "closest" to the disease-free equilibrium, since we are only interested in integer values for the initial conditions $\mathbf{x}_0$ (see (6.3) for the definition of the initial values $\mathbf{x}_0$). Therefore it seems that for any practical purposes, the point $\mathbf{x}_0^1 = (n-1, 0, 1, 0, 0)$ is the only point $\mathbf{x}_0 \neq \mathbf{e}_1$ from which the disease-free equilibrium can be reached, and for all the other possible $\mathbf{x}_0$ the system tends to the endemic equilibrium.

We tried different parameter values from the ones shown in (6.16), changing one of the parameters at a time (and keeping $\gamma_0 = 0.2$, since we know the behaviour when γ_0 varies) for the two points

$$\mathbf{x}_0^1 = (n-1, 0, 1, 0, 0) \quad \text{and} \quad \mathbf{x}_0^2 = (n-1, 1, 0, 0, 0).$$

In all cases the same behaviour was observed: starting from $\mathbf{x}_0^1$ the system tends to $\mathbf{e}_1$ and starting from $\mathbf{x}_0^2$ it tends to $\mathbf{e}_2$. The only exceptions were the following:

(α) Close to the point where $\mathcal{R}_0 - 1$ changes sign, the deterministic system tends to $\mathbf{e}_2$ (as long as $\mathcal{R}_1 > 1$) with both initial conditions $\mathbf{x}_0^1$, $\mathbf{x}_0^2$. For instance with $\gamma_0 = 0.1587$ ($\mathcal{R}_0 > 1$) and $\gamma_0 = 0.1588$ ($\mathcal{R}_0 < 1$) the system tends to $\mathbf{e}_2$ with both $\mathbf{x}_0^1$ and $\mathbf{x}_0^2$. The same happens with $q_1 = 0.6$. So we calculated the values of $D_{21}, \ldots, D_{25}$, $\mathcal{R}_0$, and $\mathcal{R}_1$ for those parameter values (as in (6.16) and $\gamma_0 = 0.2$) and for $q_1 = 0.0001, 0.0002, \ldots, 1.0000$. The results are shown in Table 6.5 which shows that $q_1 = 0.6$ is close to the point where $\mathcal{R}_0 - 1$ changes sign.

Value of q_1	Stability
0.0001, ... , 0.3066	$\mathcal{R}_0 < 1$, $\mathcal{R}_1$ not real $\mathbf{e}_1$ stable $\mathbf{e}_2$ infeasible
0.3067, ... , 0.5994	$\mathcal{R}_0 < 1$, $\mathcal{R}_1 > 1$ $\mathbf{e}_1$ stable $\mathbf{e}_2$ feasible, stable
0.5995, ... , 1	$\mathcal{R}_0 > 1$, $\mathcal{R}_1 > 1$ $\mathbf{e}_1$ unstable $\mathbf{e}_2$ feasible, stable

Table 6.5: Feasibility and stability of the deterministic equilibria in a subspace of the parameter space: grid for q_1

(β) Close to the point where $\mathcal{R}_1 - 1$ changes sign or changes between being positive and pure complex (and as long as $\mathcal{R}_0 < 1$) the deterministic system tends to $\mathbf{e}_1$ with both initial conditions $\mathbf{x}_0^1$, $\mathbf{x}_0^2$. For instance that happens with $\gamma_0 = 0.3244$ ($\mathcal{R}_1 > 1$) and $\gamma_0 = 0.3245$ ($\mathcal{R}_1$ non-real) and also for $p = 0.07$. Table 6.6 shows that the value $p = 0.07$ is close to the point where $\mathcal{R}_1$ changes from being non-real to real and greater than one.

It seems therefore that close to the point where $\mathcal{R}_1 - 1$ changes behaviour (and as long as $\mathcal{R}_0 < 1$) the system is attracted to $\mathbf{e}_1$; in most of the numerical solutions this happened at a slower rate than when the corresponding parameter value (that was varied) was far away from that point. On the other hand, close to the point where $\mathcal{R}_0 - 1$ changes sign (and as long as $\mathcal{R}_1$ remains greater than 1) the system is attracted to $\mathbf{e}_2$. In some cases it is more accurate to say that the system is ultimately attracted to $\mathbf{e}_2$; Figure 6.3 shows the value of z over a period of 6000 years ((a) shows the first 1000 years and (b) the whole period) for a system that begins from $\mathbf{x}_0^1 = (n-1, 0, 1, 0, 0)$ with the

Value of p	**Stability**
0.0001,... ,0.0672	$\mathcal{R}_0 < 1$, $\mathcal{R}_1$ not real $\mathbf{e}_1$ stable $\mathbf{e}_2$ infeasible
0.0673,... ,0.1057	$\mathcal{R}_0 < 1$, $\mathcal{R}_1 > 1$ $\mathbf{e}_1$ stable $\mathbf{e}_2$ feasible, stable
0.1058,... ,1	$\mathcal{R}_0 > 1$, $\mathcal{R}_1 > 1$ $\mathbf{e}_1$ unstable $\mathbf{e}_2$ feasible, stable

Table 6.6: Feasibility and stability of the deterministic equilibria in a subspace of the parameter space: grid for p

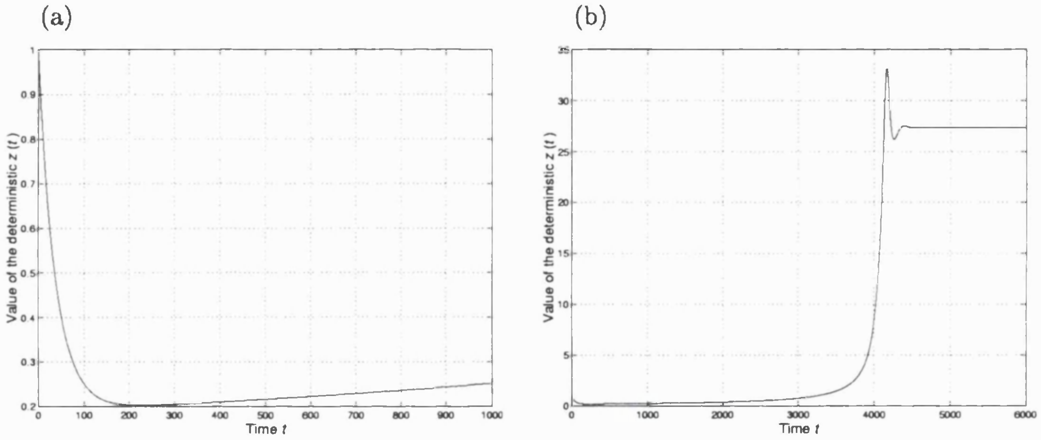


Figure 6.3: The value of $z(t)$ as obtained from numerical solution of the system (6.1). (a) shows the first 1000 years and (b) the first 6000 years. The parameter values are as shown in (6.16) and $\gamma_0 = 0.1588$. The initial conditions are $x_0 = n - 1$ and $z_0 = 1$.

parameter values shown in (6.16) and $\gamma_0 = 0.1588$. In this case $\mathcal{R}_0 < 1$ and $\mathcal{R}_1 > 1$ and both $\mathbf{e}_1$ and $\mathbf{e}_2$ are stable. Initially the value of z decreases and reaches its minimum at $z(226) = 0.2023$. Then it starts increasing slowly and after 3500 years it jumps to a peak of $z = 33.2$ and then drops to the endemic value $z_2^e = 27.38$. A similar behaviour was observed for y, w, and u. It has to be stressed that from these results it appears that the final equilibrium may not be so significant for any practical purposes, since it is reached after a very long time and for such a long timescale it is unreasonable to assume that the values of the parameters remain the same.

Summarising the results for the equilibrium of the deterministic system, we have found the following:

- The deterministic system has three equilibrium points: the disease-free equilibrium $\mathbf{e}_1$, and two endemic equilibria $\mathbf{e}_2$ and $\mathbf{e}_3$.
- $\mathbf{e}_1$ is locally stable when $\mathcal{R}_0 < 1$ and unstable when $\mathcal{R}_0 > 1$.
- $\mathbf{e}_2$ is feasible when $\mathcal{R}_1 \geq 1$ and infeasible otherwise. The numerical results presented in this section prove that $\mathbf{e}_2$ is stable in some subsets of the parameter space where $\mathcal{R}_1 > 1$ (and with either $\mathcal{R}_0 > 1$ or $\mathcal{R}_0 < 1$) and suggest that maybe this is the case throughout the region $\mathcal{R}_1 > 1$ (even when $\mathcal{R}_0 < 1$).
- $\mathbf{e}_3$ is unstable in the space where it is feasible.

Finally it has to be noted that the fact that the endemic equilibrium can be stable even when $\mathcal{R}_0 < 1$ has serious implications for the control of the disease. If public health policies aim at reducing $\mathcal{R}_0$ in order to control the disease, then for TB this is not enough: reducing $\mathcal{R}_0$ to a value less than one makes the disease-free equilibrium stable, but still the disease may not tend to extinction (depending on the initial conditions) if $\mathcal{R}_1$ is still greater than one. Therefore $\mathcal{R}_1$ has to be reduced (to a value less than 1) as well, in order to "guarantee" the extinction of the disease. In any case, though, the time until extinction (if extinction is achieved) can be very long, as the results in the following sections will show.

6.3 The stochastic model

6.3.1 The transient phase

Let $p_{\mathbf{x}}(t) = p(x, y, z, w, u; t)$ be the probability that there are x uninfected individuals, y infectious cases, z latents, w non-infectious cases, and u recovered cases in the population at time $t \geq 0$:

$$p_{\mathbf{x}}(t) = p(x, y, z, w, u; t) = \mathrm{P}[X(t) = x, Y(t) = y, Z(t) = z, W(t) = w, U(t) = u], \quad (6.17)$$

for $t \geq 0$, $\mathbf{x} \in \mathcal{S} \equiv \mathbb{Z}_+^5$, and $p_{\mathbf{x}}(t) = 0$ otherwise. The initial conditions are $p_{\mathbf{x}_0}(0) = 1$ and $p_{\mathbf{x}}(0) = 0$ for any $\mathbf{x} \neq \mathbf{x}_0$ where $\mathbf{x}_0 = (x_0, y_0, z_0, w_0, u_0) \in \mathcal{S}_0$ as defined in (6.3). The corresponding Kolmogorov forward equations for $p_{\mathbf{x}}(t)$ are given in the Appendix, equation (A.7). The joint probability generating function $\mathcal{P}(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5; t) =$

$\mathrm{E}[\theta_1^{X(t)}\theta_2^{Y(t)}\theta_3^{Z(t)}\theta_4^{W(t)}\theta_5^{U(t)}]$ satisfies the equation

$$\begin{aligned}
\frac{\partial \mathcal{P}}{\partial t} &= \lambda(\theta_1 - 1)\mathcal{P} + \mu(1-\theta_1)\frac{\partial \mathcal{P}}{\partial \theta_1} \\
&+ [(\mu+\mu_1)(1-\theta_2) + \gamma_0(\theta_5 - \theta_2)]\frac{\partial \mathcal{P}}{\partial \theta_2} \\
&+ [\mu(1-\theta_3) + q_2\beta(\theta_2-\theta_3) + (1-q_2)\beta(\theta_4-\theta_3)]\frac{\partial \mathcal{P}}{\partial \theta_3} \\
&+ [(\mu+\mu_2)(1-\theta_4) + \delta(\theta_2-\theta_4) + \delta_0(\theta_5-\theta_4)]\frac{\partial \mathcal{P}}{\partial \theta_4} \qquad (6.18) \\
&+ [\mu(1-\theta_5) + \epsilon_1(\theta_2-\theta_5) + \epsilon_2(\theta_4-\theta_5)]\frac{\partial \mathcal{P}}{\partial \theta_5} \\
&+ \frac{\alpha}{n}\theta_2[-\theta_1 + pq_1\theta_2 + (1-p)\theta_3 + p(1-q_1)\theta_4]\frac{\partial^2 \mathcal{P}}{\partial \theta_1 \partial \theta_2} \\
&+ \frac{\alpha_2}{n}\theta_2[q_3\theta_2 + (1-q_3)\theta_4 - \theta_3]\frac{\partial^2 \mathcal{P}}{\partial \theta_2 \partial \theta_3},
\end{aligned}$$

with the initial condition $\mathcal{P}(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5;0) = \theta_1^{x_0}\theta_2^{y_0}\theta_3^{z_0}\theta_4^{w_0}\theta_5^{u_0}$.

From equation (6.18) a system of differential equation for the first and second moments of X, Y, Z, W, and U is deduced; the equations for the means are the following

$$\begin{aligned}
\frac{d\mathrm{E}[X]}{dt} &= -\frac{\alpha}{n}\mathrm{E}[XY] - \mu\mathrm{E}[X] + \lambda \\
\frac{d\mathrm{E}[Y]}{dt} &= pq_1\frac{\alpha}{n}\mathrm{E}[XY] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZ] - \Gamma_s\mathrm{E}[Y] + q_2\beta\mathrm{E}[Z] + \delta\mathrm{E}[W] + \epsilon_1\mathrm{E}[U] \\
\frac{d\mathrm{E}[Z]}{dt} &= (1-p)\frac{\alpha}{n}\mathrm{E}[XY] - \frac{\alpha_2}{n}\mathrm{E}[YZ] - (\beta+\mu)\mathrm{E}[Z] \qquad (6.19) \\
\frac{d\mathrm{E}[W]}{dt} &= p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XY] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ] + (1-q_2)\beta\mathrm{E}[Z] - \Delta_s\mathrm{E}[W] + \epsilon_2\mathrm{E}[U] \\
\frac{d\mathrm{E}[U]}{dt} &= \gamma_0\mathrm{E}[Y] + \delta_0\mathrm{E}[W] - (\epsilon_1+\epsilon_2+\mu)\mathrm{E}[U],
\end{aligned}$$

where the Γ_s and Δ_s are as defined in (6.2) and the terms $\mathrm{E}[XY]$ and $\mathrm{E}[YZ]$ can be expressed as

$$\begin{aligned}
\mathrm{E}[XY] &= \mathrm{Cov}[X,Y] + \mathrm{E}[X]\mathrm{E}[Y] \\
\mathrm{E}[YZ] &= \mathrm{Cov}[Y,Z] + \mathrm{E}[Y]\mathrm{E}[Z].
\end{aligned}$$

The equations for the variances and covariances are given in the Appendix (Section A.3.2). The system of equations for the first and second moments contains higher-order moments and hence it is open and cannot be solved directly.

From the system (6.19) it follows that the expected value of the total population size satisfies the equation

$$\frac{d\mathrm{E}[N(t)]}{dt} = \lambda - \mu\mathrm{E}[N(t)] - \mu_1\mathrm{E}[Y(t)] - \mu_2\mathrm{E}[W(t)],$$

which with integration gives

$$\mathrm{E}[N(t)] = \frac{\lambda}{\mu} + e^{-\mu t}\left(n - \frac{\lambda}{\mu}\right) - e^{-\mu t}\int_0^t e^{\mu s}\{\mu_1 \mathrm{E}[Y(s)] + \mu_2 \mathrm{E}[W(s)]\}ds. \tag{6.20}$$

Equation (6.20) shows that the mean population size is bounded above by λ/μ if $\lambda \geq \mu n$ and by n if $\lambda < \mu n$. Therefore, $\mathrm{E}[X(t)]$, $\mathrm{E}[Y(t)]$, $\mathrm{E}[Z(t)]$, $\mathrm{E}[W(t)]$, $\mathrm{E}[U(t)]$, and $\mathrm{E}[N(t)]$ are bounded above by $\max\{n, \lambda/\mu\}$, for all $t \geq 0$.

6.3.2 The equilibrium state of the process

The process described in this chapter is a Markov process in continuous time with countable state space $\mathcal{S} = \mathbb{Z}_+^5$. Let $\mathcal{A}$ denote the subset of $\mathcal{S}$ that contains all the states of the form $(x, 0, 0, 0, 0)$ and $\mathcal{D}$ the remaining set of states:

$$\mathcal{A} = \{(x, 0, 0, 0, 0) \in \mathbb{Z}_+^5\}$$
$$\mathcal{D} = \mathcal{S} - \mathcal{A} = \{(x, y, z, w, u) \in \mathbb{Z}_+^5 : (y, z, w, u) \neq (0, 0, 0, 0)\}.$$

The sets $\mathcal{A}$ and $\mathcal{D}$ form two irreducible classes. The former is closed and absorbing, while the latter is open and transient. The fact that the class $\mathcal{A}$ is absorbing means that once the chain reaches one of the states in $\mathcal{A}$ then it will remain within $\mathcal{A}$ (because there are no infected individuals in the population and hence there will be no more new infections and the population will remain free from the infection). Using Theorem 5.3 we will show that the chain will be absorbed in $\mathcal{A}$ with probability one.

Following the notation in Definition 5.2, we define the functions $a_j(\mathbf{x})$, $d_j(\mathbf{x})$, $e_{ij}(\mathbf{x})$, for $\mathbf{x} = (x, y, z, w, u) \in \mathcal{S}$, $i, j = 1, 2, \ldots, 5$ and $i \neq j$, as follows: $a_1(\mathbf{x}) = \lambda$ and $a_j(\mathbf{x}) = 0$, for $j = 2, 3, 4, 5$; $d_1(\mathbf{x}) = \mu x$, $d_2(\mathbf{x}) = (\mu + \mu_1)y$, $d_3(\mathbf{x}) = \mu z$, $d_4(\mathbf{x}) = (\mu + \mu_2)w$, and $d_5(\mathbf{x}) = \mu u$. The definitions of e_{ij} are shown in Table 6.7.

j	$e_{1j}(\mathbf{x})$	$e_{2j}(\mathbf{x})$	$e_{3j}(\mathbf{x})$	$e_{4j}(\mathbf{x})$	$e_{5j}(\mathbf{x})$
1	$-$	0	0	0	0
2	$pq_1\frac{\alpha}{n}xy$	$-$	$q_2\beta z + q_3\frac{\alpha_2}{n}yz$	δw	$\epsilon_1 u$
3	$(1-p)\frac{\alpha}{n}xy$	0	$-$	0	0
4	$p(1-q_1)\frac{\alpha}{n}xy$	0	$(1-q_2)\beta z + (1-q_3)\frac{\alpha_2}{n}yz$	$-$	$\epsilon_2 u$
5	0	$\gamma_0 y$	0	$\delta_0 w$	$-$

Table 6.7: The functions e_{ij} from Reuter's Theorem

The functions d_j and e_{ij} satisfy the conditions (5.17) for all $j = 1, \ldots, 5$ and $i \neq j$. Also, by "freezing" the states $(x, 0, 0, 0, 0)$, i.e. assuming that $a_1(x, 0, 0, 0, 0) =$

$d_1(x,0,0,0,0) = 0$ for all $x = 0,1,\ldots,$ the states $(x,0,0,0,0)$ become absorbing and Theorem 5.3 can be applied. Let $\mathcal{A}_k$ denote the set of all states $(x,y,z,w,u) \in \mathcal{D}$ such that $x+y+z+w+u = k$ for $k = 1,2,\ldots.$ Then it follows that

$$r_k = \max_{\mathbf{x}\in\mathcal{A}_k} \sum_{i=1}^{5} a_i(\mathbf{x}) = \lambda \tag{6.21}$$

$$s_k = \min_{\mathbf{x}\in\mathcal{A}_k} \sum_{i=1}^{5} d_i(\mathbf{x}) = \mu k, \tag{6.22}$$

for all $k = 1,2,\ldots.$ From the statements (a), (b), and (c) of Theorem 5.3 the following results can be deduced:

- $S_1 \equiv \dfrac{1}{r_2} + \displaystyle\sum_{k=3}^{\infty}\left(\frac{1}{r_k} + \frac{s_k}{r_k r_{k-1}} + \cdots + \frac{s_k \cdots s_3}{r_k \cdots r_2}\right)$
  $= \dfrac{1}{\lambda} + \displaystyle\sum_{k=3}^{\infty}\left(\frac{1}{\lambda} + \frac{\mu k}{\lambda^2} + \cdots + \frac{k(k-1)\cdots 3\mu^{k-2}}{\lambda^{k-1}}\right) = \infty.$
  Hence the process is regular.
- $S_2 \equiv \displaystyle\sum_{k=1}^{\infty} \frac{s_1 \cdots s_k}{r_1 \cdots r_k} = \sum_{k=1}^{\infty}\left(\frac{\mu}{\lambda}\right)^k k! = \infty.$
  Therefore $\pi(\mathbf{x},\mathbf{y}) = 0$ if $\mathbf{y} \in \mathcal{D}$ and $\alpha(\mathbf{x}) = \sum_{\mathbf{y}\in\mathcal{A}} \pi(\mathbf{x},\mathbf{y}) = 1$ for all $\mathbf{x} \in \mathcal{D}$, where $\pi(\mathbf{x},\mathbf{y})$ is the limit as t tends to infinity of the probability $\mathrm{P}[\mathbf{X}(t) = \mathbf{y}|\mathbf{X}(0) = \mathbf{x}]$ for $\mathbf{x},\mathbf{y} \in \mathcal{S}$.
- $S_3 \equiv \dfrac{1}{s_1} + \displaystyle\sum_{k=2}^{\infty} \frac{r_1 r_2 \cdots r_{k-1}}{s_1 s_2 \cdots s_k} = \frac{1}{\lambda}\sum_{k=1}^{\infty} \frac{(\lambda/\mu)^k}{k!} = \frac{1}{\lambda}\left(e^{\lambda/\mu} - 1\right) < \infty.$
  Hence the mean time to reach $\mathcal{A}$, starting from any state i in $\mathcal{D}$, is finite.

Since $p_{\mathbf{x}_0}(0) = \mathrm{P}[\mathbf{X}(0) = \mathbf{x}_0] = 1$, the above results imply that $\sum_{\mathbf{y}\in\mathcal{A}} \pi(\mathbf{y}) = 1$ where

$$\pi(\mathbf{y}) = \lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = \mathbf{y}], \quad \text{for} \quad \mathbf{y} \in \mathcal{S},$$

so that the population will ultimately be free from the infection with probability 1. After the extinction of the infection, the process of the uninfected individuals can be described by a birth and death process with birth and death rates $\lambda_k = \lambda$ and $\mu_k = \mu k$, respectively, for $k = 0,1,\ldots.$ The limiting distribution is a Poisson process with parameter λ/μ. Summarising, the results of this section are:

- The process will be absorbed in $\mathcal{A}$ with probability 1, so that extinction of the infection is certain: $\sum_{\mathbf{y}\in\mathcal{A}} \pi(\mathbf{y}) = 1$.

- The mean time until extinction is finite.
- After extinction of the infection, the limiting distribution of the uninfected individuals is Poisson with parameter λ/μ and

$$\lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = \mathbf{y}] = \begin{cases} 0 & \text{if } \mathbf{y} \in \mathcal{D} \\ \dfrac{e^{\lambda/\mu}(\lambda/\mu)^k}{k!} & \text{if } \mathbf{y} = (k, 0, 0, 0, 0) \in \mathcal{A}. \end{cases}$$

6.3.3 The quasi-stationary distribution

The results in the previous section show that the TB infection will die out in finite time with probability one. The fact that extinction occurs in finite time, though, does not set any limits to the length of the extinction time, which can be arbitrarily large. Moreover, as was explained in Section 5.3.4, before extinction the process may settle down around a quasi-stationary level and remain there for a long time before proceeding to extinction. Some of the most important results on the theory of quasi-stationary and limiting-conditional distributions were presented in Section 5.3.4. Here we will briefly discuss the implications of the existence of quasi-stationary distributions.

For $\mathbf{X}(t) = (X(t), Y(t), Z(t), W(t), U(t))$, the state space $\mathcal{S} = \mathbb{Z}_+^5$ of the process $\{\mathbf{X}(t), t \geq 0\}$ consists of an absorbing class $\mathcal{A}$ and a transient class $\mathcal{D} = \mathcal{S} - \mathcal{A}$, where

$$\mathcal{A} = \{(x, 0, 0, 0, 0) : x \in \mathbb{Z}_+\}.$$

If the limits

$$\lim_{t\to\infty} P[\mathbf{X}(t) = j | \mathbf{X}(0) = i, \mathbf{X}(t) \notin \mathcal{A}] = \phi_j, \quad i, j \text{ in } \mathcal{D}$$

exist and are independent of the initial conditions i then the process has a limiting-conditional distribution. This means that the process conditioned on non-absorption in $\mathcal{A}$ has a stationary distribution over $\mathcal{D}$. Given the fact that ultimate absorption in $\mathcal{A}$ is certain, this implies that the (unconditional) process can go through this "stationary" distribution, where it will settle for some time and then eventually it will be absorbed in $\mathcal{A}$. Thus if we look at an individual realisation of the process we may observe an apparent stationarity (at a level different from the one corresponding to the extinction of the infection) while the process is still in the transient class $\mathcal{D}$ (and hence in the transient phase) before it reaches the absorbing class $\mathcal{A}$.

The numerical results that we present in the following sections exhibit this behaviour. For instance the means of X, Y, Z, W, U stabilise at a level different from

$(\lambda/\mu, 0, 0, 0, 0)$ which corresponds to the extinction of TB. Also the marginal distributions (Section 6.3.6) split into two distributions, one centred around the disease-free equilibrium and the other around an endemic equilibrium (the equilibrium of the conditional means).

It appears therefore that our process has a limiting-conditional distribution and for that reason the conditional properties of the process have been studied along with the unconditional ones in the following sections, where we present some numerical results (from simulations of the stochastic process). Analytically the conditional properties can be deduced from the differential equation (A.7) for the probabilities $p_{\mathbf{x}}(t)$; define

$$\begin{aligned} p_{\mathcal{A}}(t) &= \mathrm{P}[\mathbf{X}(t) \in \mathcal{A}] = \mathrm{P}[Y(t) = Z(t) = W(t) = U(t) = 0] = \sum_{\mathbf{x} \in \mathcal{A}} p_{\mathbf{x}}(t) \\ q_{\mathbf{x}}(t) &= \mathrm{P}[\mathbf{X}(t) = \mathbf{x} | \mathbf{X}(t) \notin \mathcal{A}] = \frac{p_{\mathbf{x}}(t)}{1 - p_{\mathcal{A}}(t)}, \end{aligned} \tag{6.23}$$

for $t \geq 0$, $\mathbf{x} \in \mathcal{D}$, and $q_{\mathbf{x}}(t) = 0$ otherwise. Then the differential equation for the probabilities $q_{\mathbf{x}}(t)$ can be deduced by differentiating with respect to time in (6.23) and substituting in (A.7). The conditional probabilities $q_{\mathbf{x}}(t)$ can also be evaluated numerically with the algorithm developed by Pollett & Stewart (1994) (see also Section 5.3.4).

6.3.4 Epidemiological indices

In this section we will study some epidemiological indices which are helpful in assessing the severity of an epidemic. The indices to be studied are the following:

Definition 6.1

- **Risk of infection and reinfection:** *The risk of infection at year i is the number of primary infections that occurred during the i-th year, expressed as proportion per* 100,000 *general population. Primary infections are the infections of uninfected individuals (so that this index does not account for reinfections). The risk of reinfection at year i is the number of reinfections that occurred during the i-th year, expressed as proportion per* 100,000 *general population.*

- **Incidence:** *The total incidence at year i is the number of new cases that developed during the i-th year per* 100,000 *general population. On some occasions the incidence of infectious TB (number of new infectious cases) and the incidence of non-infectious TB (number of new non-infectious cases) will be given separately.*

- **Mortality:** *Total mortality at year i is the number of TB-deaths during year i, per* 100,000 *general population. This index counts only the excess deaths due to TB and not the total number of deaths of TB cases. The mortality of infectious TB and that of non-infectious TB will be given separately on some occasions.*

- **Prevalence:** *Prevalence of infectious and non-infectious TB at year i is the number of infectious and non-infectious cases, respectively, per* 100,000 *general population at the end of year i. Prevalence of TB infection at year i is the number of infected individuals per* 100,000 *general population at the end of year i.*

For the incidence of infectious TB, new infectious cases developing during a certain year are all the transitions from the class of uninfected (X) and the class of latents (Z) to that of infectious cases (Y) during that year. From an epidemiological point of view the number of recovered (U) or non-infectious cases (W) who become infectious should not be included in this index (see, e.g., Styblo 1991). Similarly, for the incidence of non-infectious TB, new non-infectious cases developing during a certain year are all the transitions from the class of uninfected (X) and the class of latents (Z) to that of non-infectious cases (W) during that year. Finally, for the prevalence of TB infection, the number of infected individuals is the sum $Y + Z + W + U$.

Throughout this chapter these rates are presented as proportions per 10^5 general population and therefore the term "per 100,000 general population" will be occasionally suppressed. The rates presented in this section were calculated from simulations of the stochastic model (details of the implementation of the simulations can be found in the Appendix, Section A.3.3). These rates were also calculated for each year i conditional on non-extinction of the epidemic by time i. For simplicity they will be referred as the conditional rates.

The parameter values used in these simulations are shown in Table 6.8 (which also cites the references which justify these choices). The simulations were carried out for three population sizes $n = 100$, $n = 1000$, and $n = 10000$. For each of these three situations, the following six sets of initial conditions were used:

$$\begin{array}{ll} y_0 = 1, x_0 = n - 1 & w_0 = 10, x_0 = n - 10 \\ y_0 = 10, x_0 = n - 10 & u_0 = 10, x_0 = n - 10 \\ z_0 = 10, x_0 = n - 10 & y_0 = z_0 = w_0 = u_0 = 10. \end{array} \tag{6.24}$$

Parameter	Value	References
λ	μn	
μ	0.02	
μ_1	0.13	Grzybowski & Enarson (1978), Dolin et al. (1994), Bloom & Murray (1992), Springett (1971), Enarson & Rouillon (1998)
μ_2	0.1	Grzybowski & Enarson (1978), Dolin et al. (1994), Bloom & Murray (1992), Springett (1971), Enarson & Rouillon (1998)
α	10	Murray et al. (1993), Styblo (1991)
p	0.05	Murray et al. (1993), Styblo (1991), Sutherland, Švandová & Radhakrishna (1982), Vynnycky & Fine (1997), Enarson & Rouillon (1998)
q_1	0.55	Murray et al. (1993), Styblo (1991)
β	0.002	Krishnamurthy & Chaudhuri (1990), Dolin et al. (1994), Enarson & Rouillon (1998), Krishnamurthy, Nair, Gothi & Chakraborty (1976)
q_2	0.55	Dye et al. (1998), Blower et al. (1995), Vynnycky & Fine (1997), Krishnamurthy et al. (1976)
p_r	0.45	Dye et al. (1998), Sutherland et al. (1982), Vynnycky & Fine (1997)
p_3	0.05	Murray et al. (1993), Styblo (1991), Sutherland et al. (1982), Vynnycky & Fine (1997), Enarson & Rouillon (1998)
α_2	$p_3 p_r \alpha$	
q_3	0.55	Dye et al. (1998), Vynnycky & Fine (1997)
δ	0.02	Grosset (1989), Murray et al. (1993)
γ_0	0.066	Murray et al. (1991), Springett (1971), Enarson & Rouillon (1998), Grzybowski & Enarson (1978)
δ_0	0.066	Murray et al. (1991), Springett (1971), Enarson & Rouillon (1998), Grzybowski & Enarson (1978)
ϵ_1	0.015	Styblo (1991), Springett (1971), Campbell (1974)
ϵ_2	0.015	Styblo (1991), Springett (1971), Campbell (1974)

Table 6.8: Parameter values for model Zeus

The simulations were carried out up to time $t = 300$ years. Throughout this section time is measured in years.

Prevalence of infectious TB

The conditional and unconditional prevalences of infectious TB, as calculated from the simulations, are shown in Figure 6.4. For $y_0 = 10$ and $y_0 = z_0 = w_0 = u_0 = 10$ the prevalences are quite similar and higher than in the other cases. After these two, the next most severe situation is with $u_0 = 10$, then with $w_0 = 10$, then with $y_0 = 1$, and the least severe with $z_0 = 10$. The same ordering in severity is observed in the conditional prevalences, but there the differences decrease as time increases and the system tends

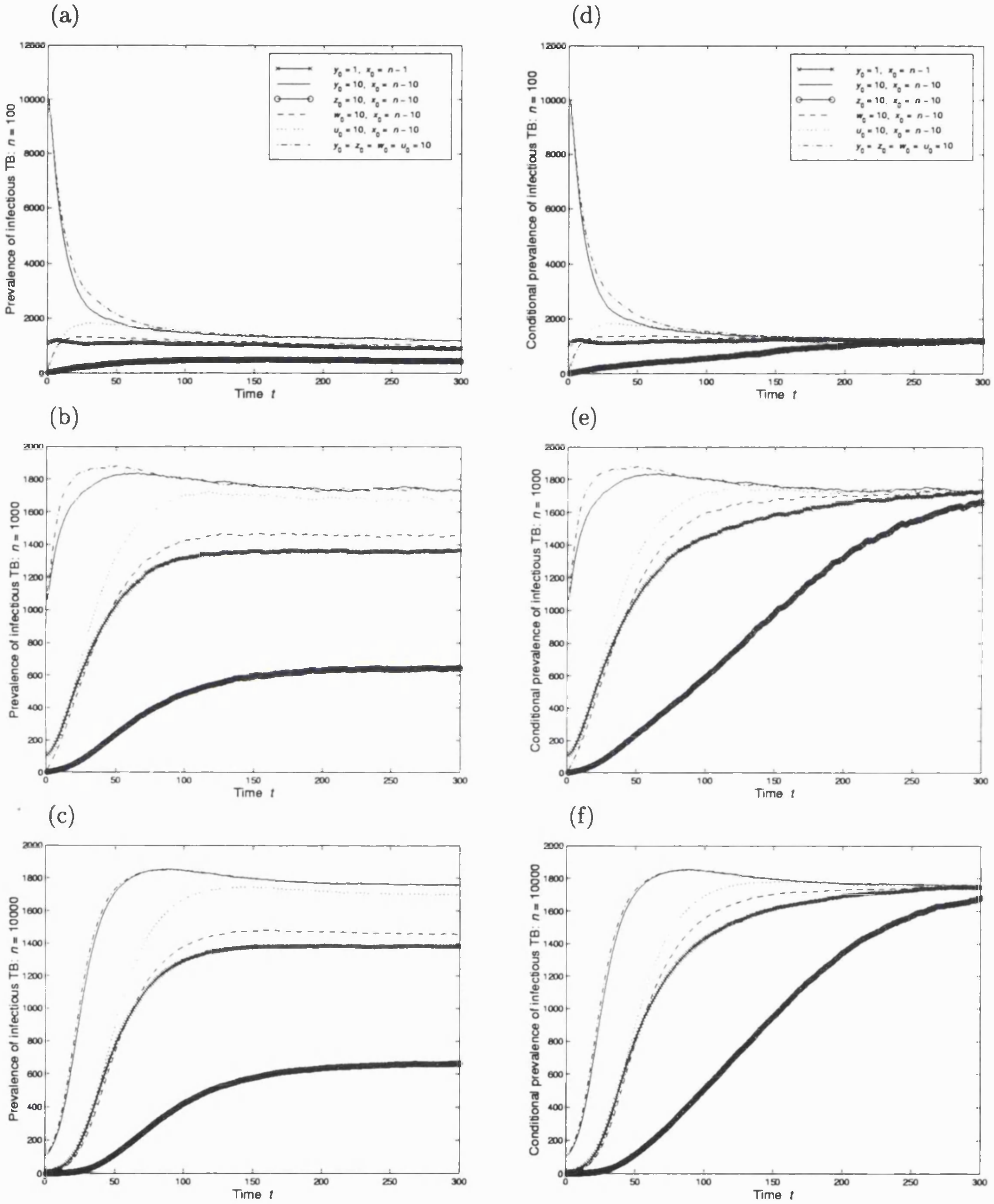


Figure 6.4: (a),(b),(c): Prevalence of infectious TB for $n = 100$, $n = 1000$, and $n = 10000$, respectively. (d),(e),(f): Conditional prevalence of infectious TB for $n = 100$, $n = 1000$, and $n = 10000$, respectively. All the rates were calculated as proportions per 10^5 general population. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24) (the curves are labelled as in Figures (a) and (d)). The parameter values used are shown in Table 6.8. Time is measured in years.

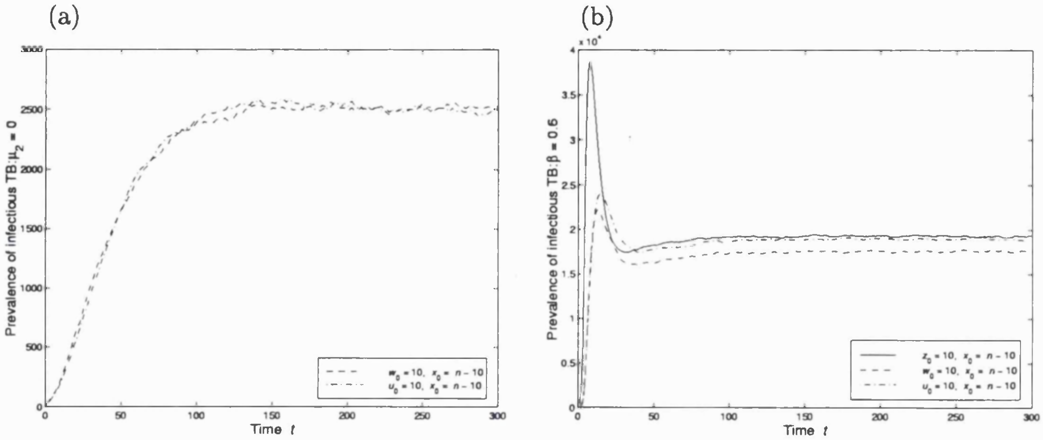


Figure 6.5: Prevalence of infectious TB for two special cases: the parameter values used are the ones shown in Table 6.8 except $\mu_2 = 0$ for (a) and $\beta = 0.6$ for (b). In both cases $n = 1000$. All the rates shown are proportions per 10^5 general population. Time is measured in years.

to quasi-stationarity. So that at time $t = 300$ the conditional prevalences are almost the same regardless of the initial conditions y_0, z_0, w_0, u_0. The differences observed for the various initial conditions can be explained by the following arguments:

- The prevalence is higher with larger values of y_0 because then there are more infections from the beginning of the epidemic which increase the reservoir of infected individuals thus boosting the severity of the epidemic.
- The prevalence is higher when $u_0 = 10$ than when $w_0 = 10$ because the total death rate $\mu + \mu_2$ for the class W decreases its size more than the normal death rate μ for the class U decreases the size of U. Therefore the number of infectious cases Y developing from U is greater than that from W and hence starting with $u_0 = 10$ increases the value of Y more (and faster) than starting with $w_0 = 10$. And having more infectious cases boosts the severity of the epidemic, as was explained above. This explanation can be supported from the results of another simulation shown in Figure 6.5(a). In this case we used $n = 1000$ and the parameter values shown in Table 6.8, except that μ_2 was set equal to zero. In this case the prevalence of infectious TB was almost identical throughout the time interval $\mathcal{I}_t = [1, 300]$ for the two cases ($u_0 = 10$ and $w_0 = 10$).
- The prevalence is the smallest when $z_0 = 10$ because the rate $q_2\beta$ at which the latents become infectious is very small and much smaller than the rates δ and ϵ_1 at which the non-infectious and the recovered, respectively, become infectious. Therefore the size of the infectious class Y increases faster (and more) when $u_0 = 10$ or $w_0 = 10$ than when $z_0 = 10$ leading to more severe epidemics. This explanation can be supported by the

results of the simulations shown in Figure 6.5(b). For this case we used a larger value for β ($\beta = 0.6$, the other parameters as shown in Table 6.8, and $n = 1000$); initially the prevalence of infectious TB is much higher with $z_0 = 10$ than with $w_0 = 10$ or $u_0 = 10$ and remains higher up to time $t = 300$.

These observations hold for all three cases $n = 100$, $n = 1000$, and $n = 10000$ and seem to suggest that the level of the prevalence of infectious TB depends (proportionately) on the rates at which individuals become infectious. Therefore the prevalence will be higher when the epidemic starts with 10 individuals in the class H, for $H = Z, W, U$, if the total contribution from H to Y is higher than the others (i.e. the more it increases Y, the more severe the epidemic will be).

Comparing the results for $n = 100$, $n = 1000$, and $n = 10000$ at the end of the interval $\mathcal{I}_t$ it can be observed that the prevalences for $n = 1000$ and $n = 10000$ are quite similar, while those for $n = 100$ are slightly lower than the others. This difference remains even between the conditional prevalences. This is apparently a result of the effect of the initial total population size n on the quasi-stationary distribution. As has been explained before, the system remains in a quasi-stationary state before it reaches the final stationary distribution (which corresponds to extinction of the infection), unless the values of $\mathcal{R}_0$ and/or n are too small to preserve an endemic infection. In this case $\mathcal{R}_0 = 4.59$ but $n = 100$ so that the system does not remain in quasi-stationarity, as it does when n is 1000 or 10000. Thus the prevalences steadily (although slightly) decrease when $n = 100$ while they seem to have stabilised when n is 1000 or 10000.

Prevalence of non-infectious TB and TB infection

The conditional prevalences of non-infectious TB are shown in Figure 6.6. The graphs for the unconditional prevalences are similar to the ones for infectious TB and are not shown here. The unconditional and conditional prevalences of TB infection are shown in Figure 6.7. The ordering in severity (highest prevalences rates) is the same as for the infectious TB: $y_0 = 10$ and $y_0 = z_0 = w_0 = u_0 = 10$ are the most severe, then $u_0 = 10$, $w_0 = 10$, $y_0 = 1$, and the least severe $z_0 = 10$. In the conditional prevalences the difference decreases as time increases and the system tends to quasi-stationarity.

At the end of the interval $\mathcal{I}_t$ the values for $n = 1000$ and $n = 10000$ are quite similar but remain greater than the ones for $n = 100$ even in the conditional prevalences.

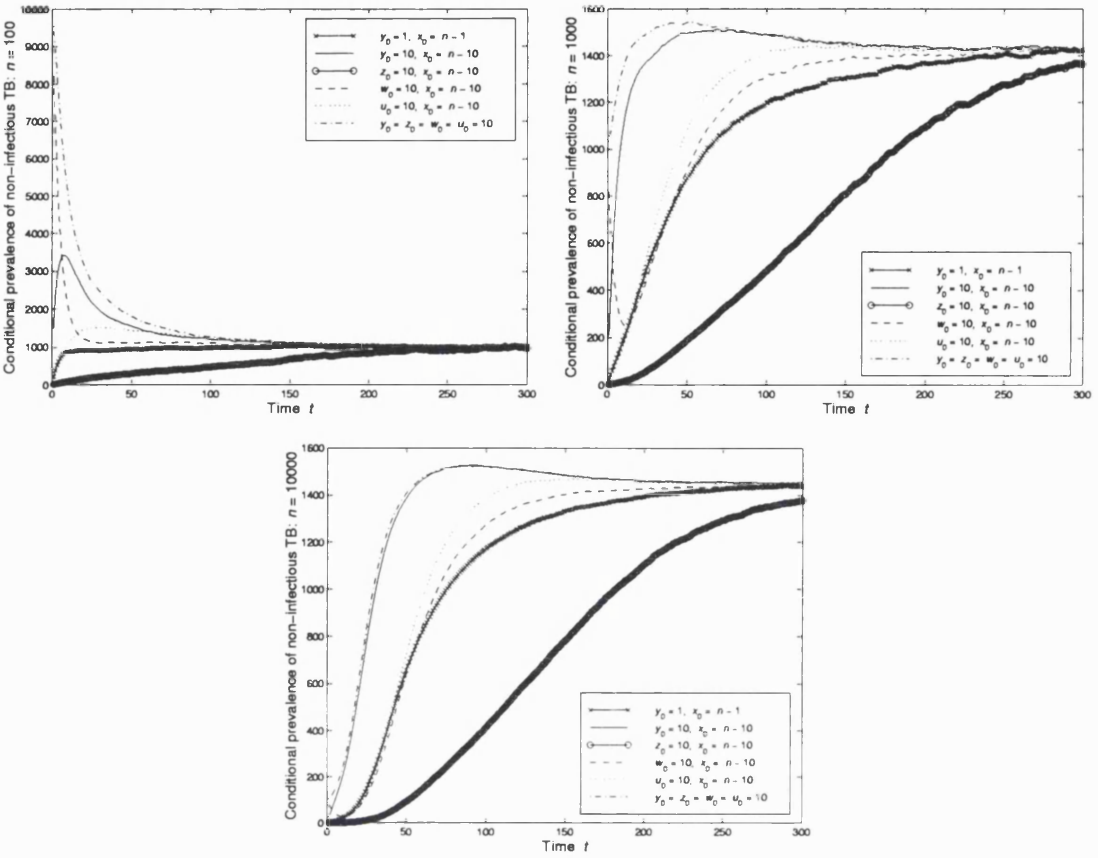


Figure 6.6: Conditional prevalence of non-infectious TB for $n = 100, 1000, 10000$. All the rates were calculated as proportions per 10^5 general population. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). The parameter values used are shown in Table 6.8. Time is measured in years.

The prevalences of non-infectious TB are of the same order as the prevalences of infectious TB (as was expected since the infectious cases are about half of the total number of cases; see, e.g., Murray et al. 1993).

Again for $n = 100$ the prevalence of TB infection declines slowly as time increases, so that we can observe the (slow) tendency to extinction, while the ones for $n = 1000$ and $n = 10000$ remain stable. It is interesting to note that the prevalence of TB infection reaches a maximum of about 95% when $n = 100$ and 85% when n is 1000 or 10000 for the cases with $y_0 = 10$ and $y_0 = z_0 = w_0 = u_0 = 10$ during the first 30 years and then declines. The proportion of the population that is infected with TB after 300 years conditioned on non-extinction by that point is about 60% when $n = 100$ and about 85% when n is 1000 or 10000.

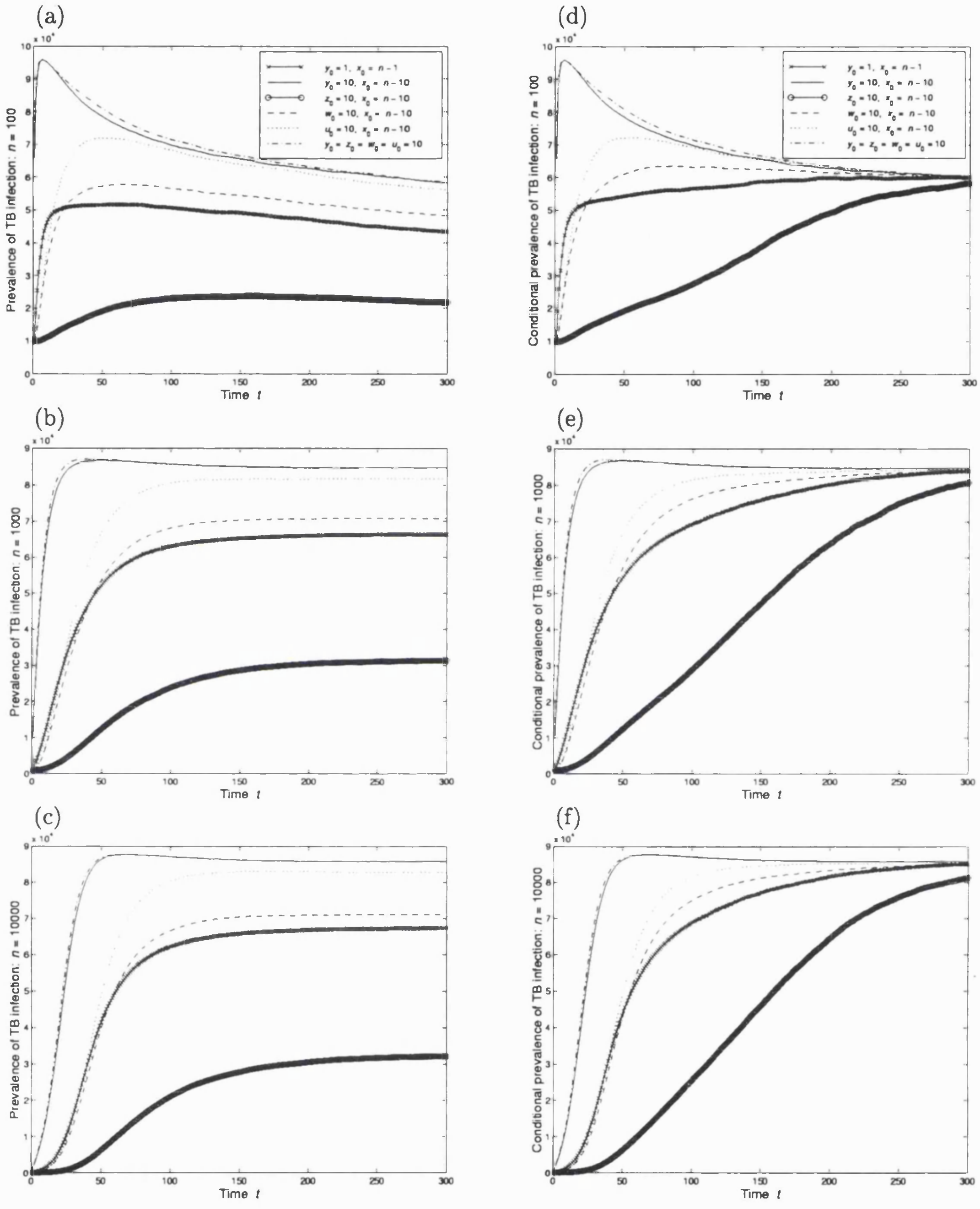


Figure 6.7: (a),(b),(c): Prevalence of TB infection for $n = 100$, $n = 1000$, and $n = 10000$, respectively. (d),(e),(f): Conditional prevalence of TB infection for $n = 100$, $n = 1000$, and $n = 10000$, respectively. All the rates were calculated as proportions per 10^5 general population. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24) (the curves are labelled as in Figures (a) and (d)). The parameter values used are shown in Table 6.8. Time is measured in years.

Incidence of infectious and non-infectious TB

The conditional incidences of infectious and non-infectious TB for $n = 100$ and $n = 1000$ are shown in Figure 6.8 (the results for $n = 10000$ are similar to the ones for $n = 1000$ and are omitted). The same qualitative behaviour is exhibited as for the prevalences. Conditional on non-extinction by time $t = 300$, the incidences at that time point are: 200 infectious and 170 non-infectious new cases per 10^5 general population when $n = 100$ and 270 infectious and 230 non-infectious new cases per 10^5 general population when $n = 1000$.

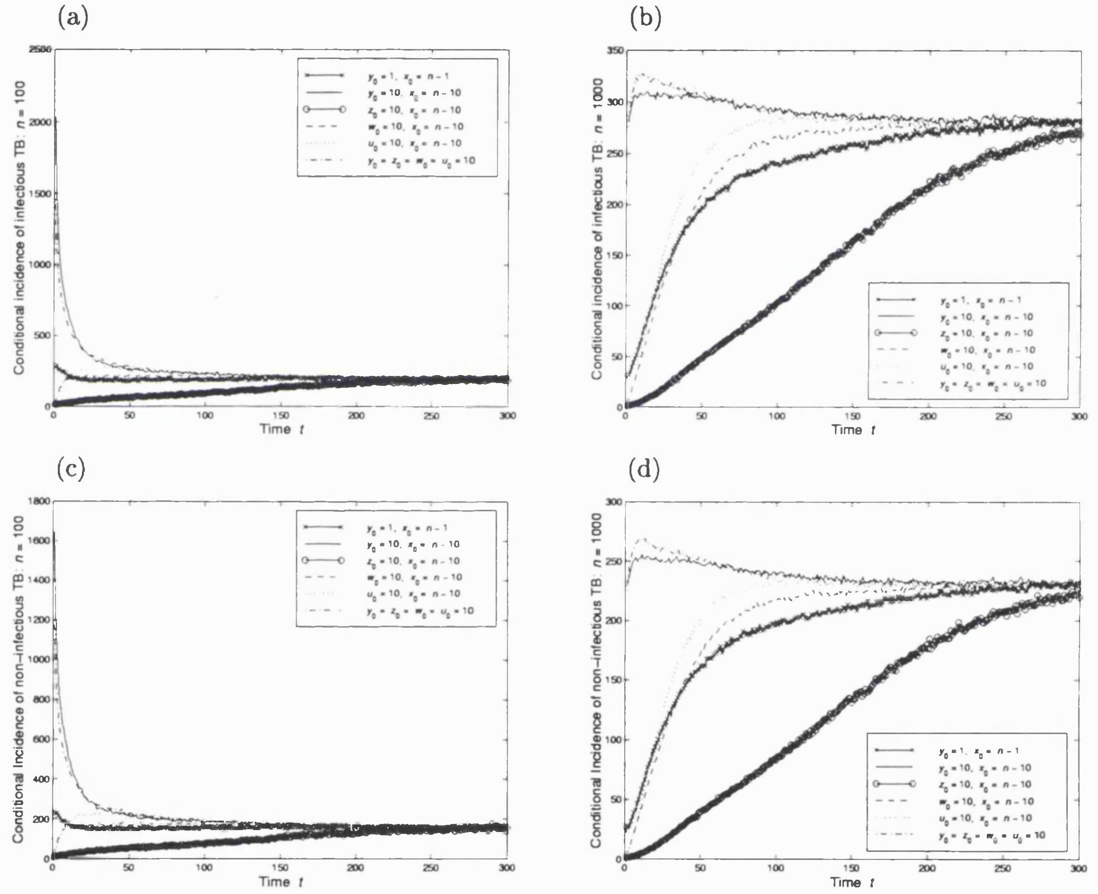


Figure 6.8: (a),(b): Conditional incidence of infectious TB for $n = 100$, $n = 1000$, respectively. (c),(d): Conditional incidence of non-infectious TB for $n = 100$, $n = 1000$, respectively. All the rates were calculated as proportions per 10^5 general population. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). The parameter values used are shown in Table 6.8. Time is measured in years.

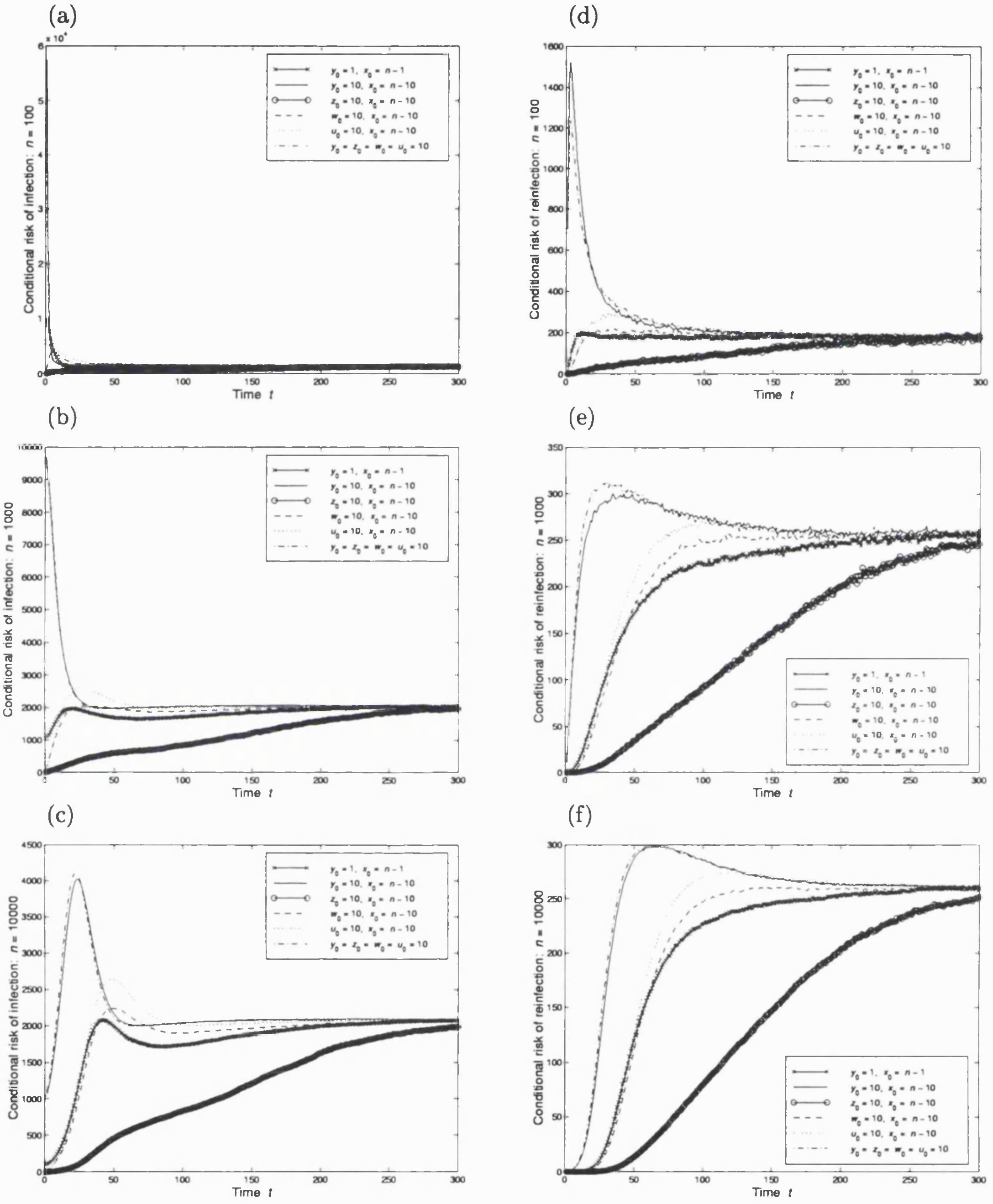


Figure 6.9: (a),(b),(c): Conditional risk of infection for $n = 100$, $n = 1000$, and $n = 10000$ respectively. (d),(e),(f): Conditional risk of reinfection for $n = 100$, $n = 1000$, and $n = 10000$, respectively. All the rates were calculated as proportions per 10^5 general population. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). The parameter values used are shown in Table 6.8. Time is measured in years.

Risk of infection and reinfection

The conditional risks of infection and reinfection are shown in Figure 6.9. Observations similar to the ones for prevalences can be made from these graphs. The highest risks correspond to the cases $y_0 = 10$ and $y_0 = z_0 = w_0 = u_0 = 10$, while the lowest for $z_0 = 10$. With $n = 1000$ and $n = 10000$, the risks are quite similar (about 2000 infections and 260 reinfections per 10^5 general population at $t = 300$) and higher than when $n = 100$ (about 1500 infections and 180 reinfections per 10^5 general population at $t = 300$).

Mortality rate

The conditional mortality rates of infectious TB are shown in Figure 6.10 for $n = 100$ and $n = 1000$. The results are qualitatively the same as for the prevalences. Conditional on non-extinction by time $t = 300$, the mortality rate at $t = 300$ is about 170 and 220 deaths per 10^5 general population when $n = 100$ and $n = 1000$, respectively. The results for $n = 10000$ (not shown here) are similar to the ones for $n = 1000$.

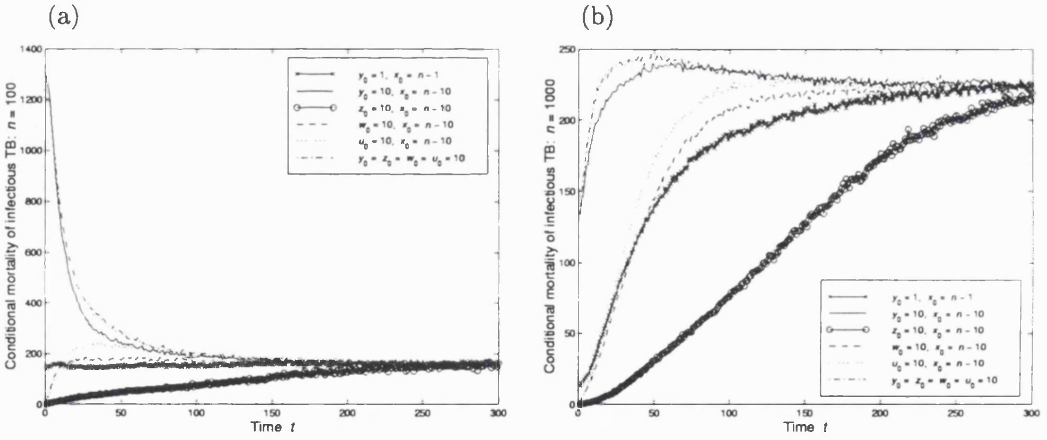


Figure 6.10: Conditional mortality rate of infectious TB (per 10^5 general population) for (a) $n = 100$ and (b) $n = 1000$. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). The parameter values used are shown in Table 6.8. Time is measured in years.

6.3.5 Moments

In this section the mean and standard deviation of X, Y, Z, W, and U are studied. The stochastic model was simulated with the parameter values shown in Table 6.8 and n equal to 100, 1000, and 10000. For each value of n the six sets of initial conditions shown

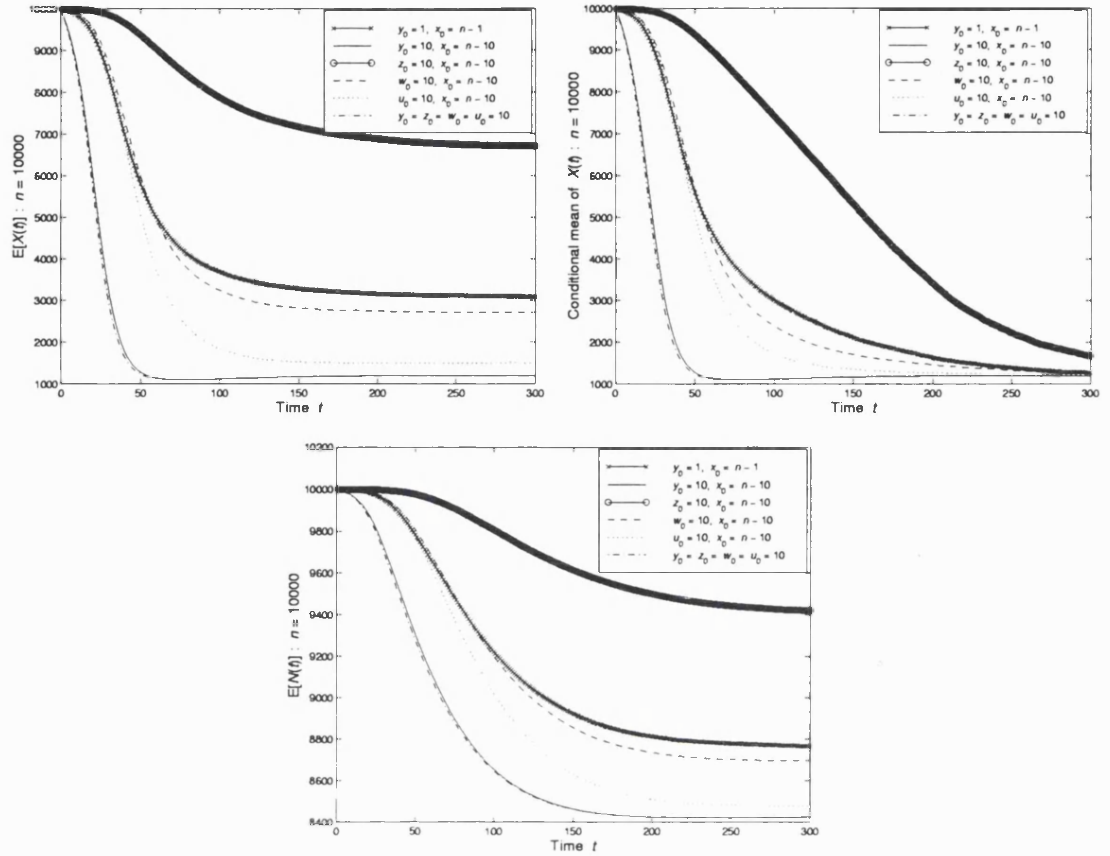


Figure 6.11: The unconditional and conditional mean of X and the mean of N. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). In all cases $n = 10000$ and the parameter values are as shown in Table 6.8. Time is measured in years.

in (6.24) were used. Details of the implementation of the simulations can be found in the Appendix (Section A.3.3).

Means

Figures 6.11, 6.12, and 6.13 show the conditional and unconditional means of X, Y, Z, W, and U and the unconditional mean of the total population size N for $n = 10000$. The results for $n = 100$ and $n = 1000$ do not differ qualitatively and are not shown.

The behaviour of the means of Y, Z, W, and U is similar to that for the prevalences (see Section 6.3.4) and the same ordering in the severity of the epidemic can be observed:

(a) The means of Y, Z, W, and U are higher if the epidemic starts with many infectious cases, for instance with $y_0 = 10$ and $y_0 = z_0 = w_0 = u_0 = 10$, compared to the cases $y_0 = 1$, $x_0 = n - 1$ and $y_0 = 0$, $z_0 + w_0 + u_0 = 10$. Also, the larger the value of y_0 the

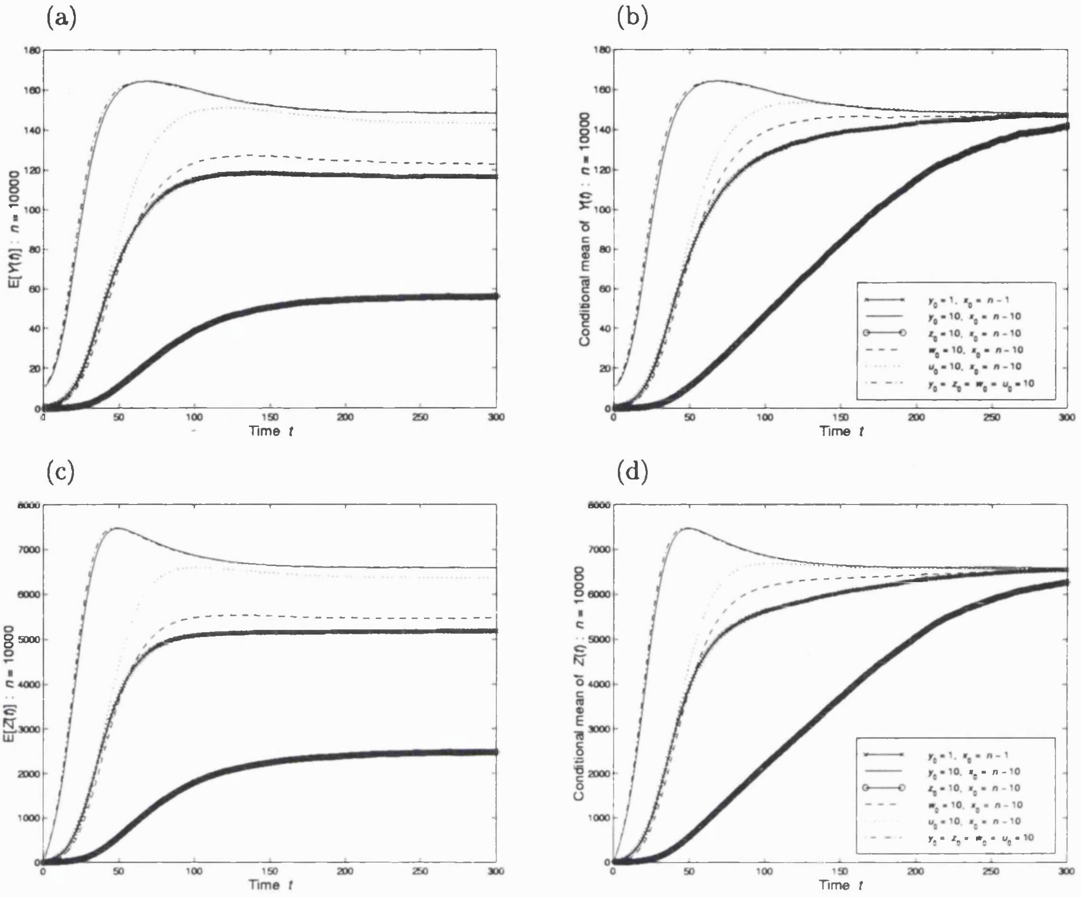


Figure 6.12: (a), (b) Unconditional and conditional means of Y. (c), (d) Unconditional and conditional means of Z. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24) (the curves are labelled as in Figures (b) and (d)). In all cases $n = 10000$ and the parameter values are as shown in Table 6.8. Time is measured in years.

larger the decrease in the size of the susceptible population and of the total population. Therefore the most severe cases are the ones that begin with large numbers of infectious individuals.

(b) Among the epidemics that begin with 10 infected but non-infectious individuals ($z_0 = 10$ or $w_0 = 10$ or $u_0 = 10$) the most severe case is when $u_0 = 10$ (largest means of Y, Z, W, U and smallest means of X, N), while the least severe is when $z_0 = 10$.

(c) For the conditional means these differences decrease in time and the conditional means converge to the same value, regardless of the initial conditions.

This ordering for the means depending on the initial value of $\mathbf{X}_0$ can be explained with the same argument as for the prevalences. The larger the initial number of infectious cases the more infections will take place in the beginning, thus boosting the epidemic to take off quickly and increasing the sizes of all the classes of infected individuals. If

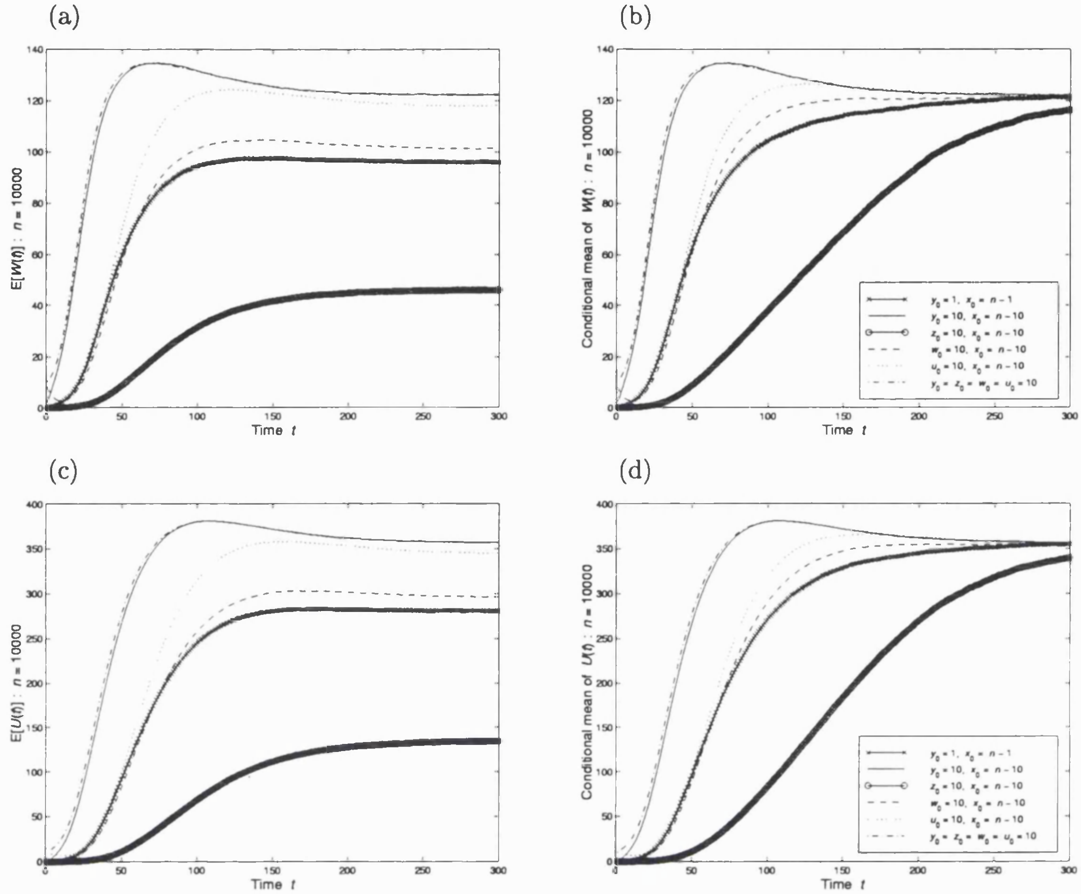


Figure 6.13: (a), (b) Unconditional and conditional means of W. (c), (d) Unconditional and conditional means of U. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24) (the curves are labelled as in Figures (b) and (d)). In all cases $n = 10000$ and the parameter values are as shown in Table 6.8. Time is measured in years.

the epidemic starts with a number of infected but non-infectious individuals ($y_0 = 0$, $z_0 + w_0 + u_0 > 0$) then it all depends on how quickly and how many of these infected become infectious. It has to be noted that this depends not only on the rates at which individuals from the Z, W, U classes move to the Y class, but also on the rates at which they move out of these classes, since these decrease the sizes of Z, W, U.

The deterministic values and the stochastic means

The conditional and unconditional means of X, Y, Z, W, U (as obtained from the simulations) and the corresponding deterministic values (obtained from numerical solutions of the system (6.1)) are compared for two cases: with $y_0 = 1$ and $y_0 = 10$ (Figure 6.14). In both cases $n = 10000$, $x_0 = n - y_0$, and the parameter values are as shown in Table 6.8. The cumulative distribution function of the extinction time for these initial conditions

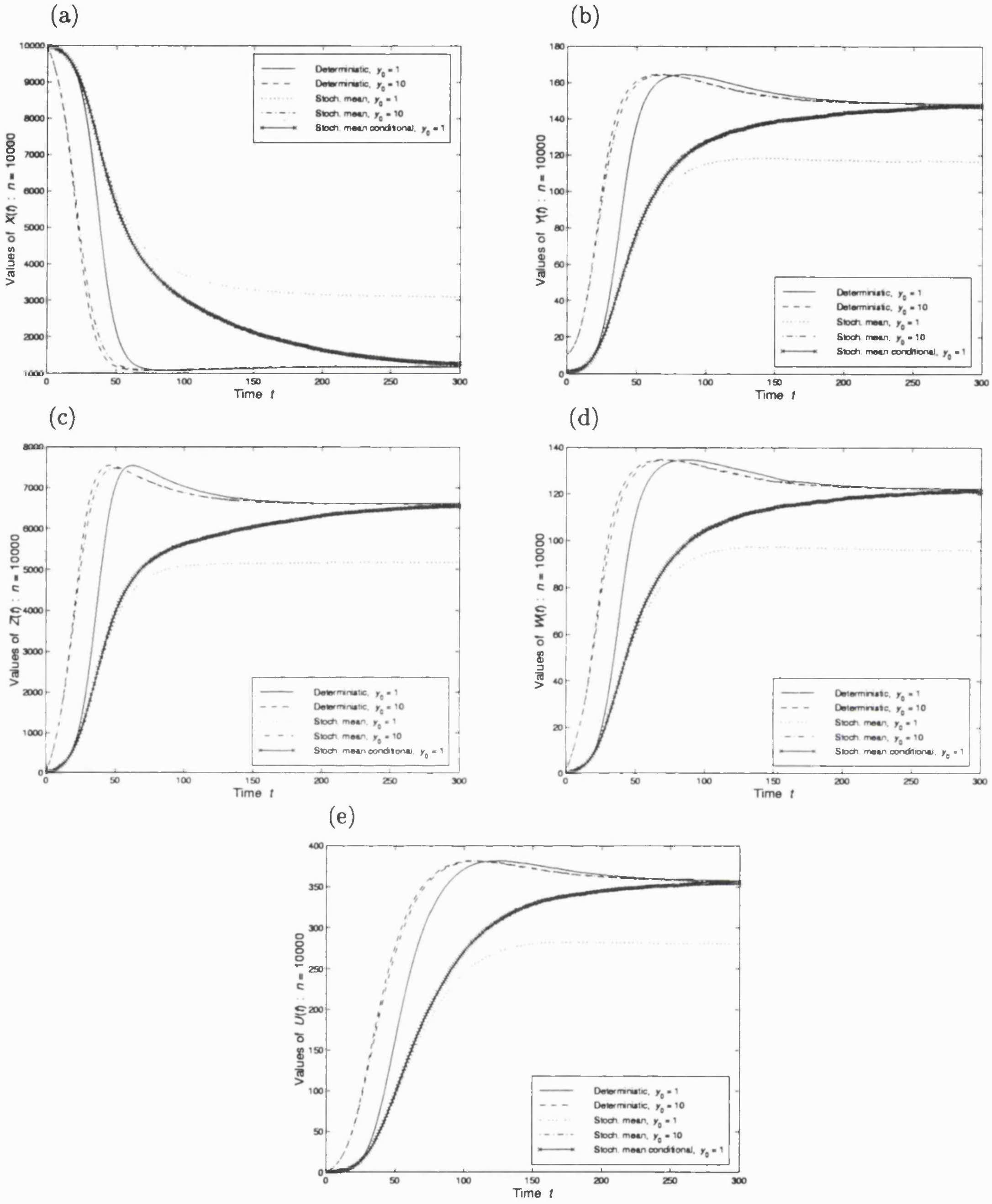


Figure 6.14: The deterministic values and the stochastic means of (a) X, (b) Y, (c) Z, (d) W, (e) U. In each graph there are five curves, two for the deterministic values with $y_0 = 1$ and $y_0 = 10$, two for the unconditional stochastic means with $y_0 = 1$ and $y_0 = 10$, and one for the conditional stochastic means with $y_0 = 1$ (when $y_0 = 10$, the conditional means are the same as the unconditional because none of the simulations ended with extinction). In all cases $n = 10000$, $x_0 = n - y_0$, and the parameter values are as shown in Table 6.8. Time is measured in years.

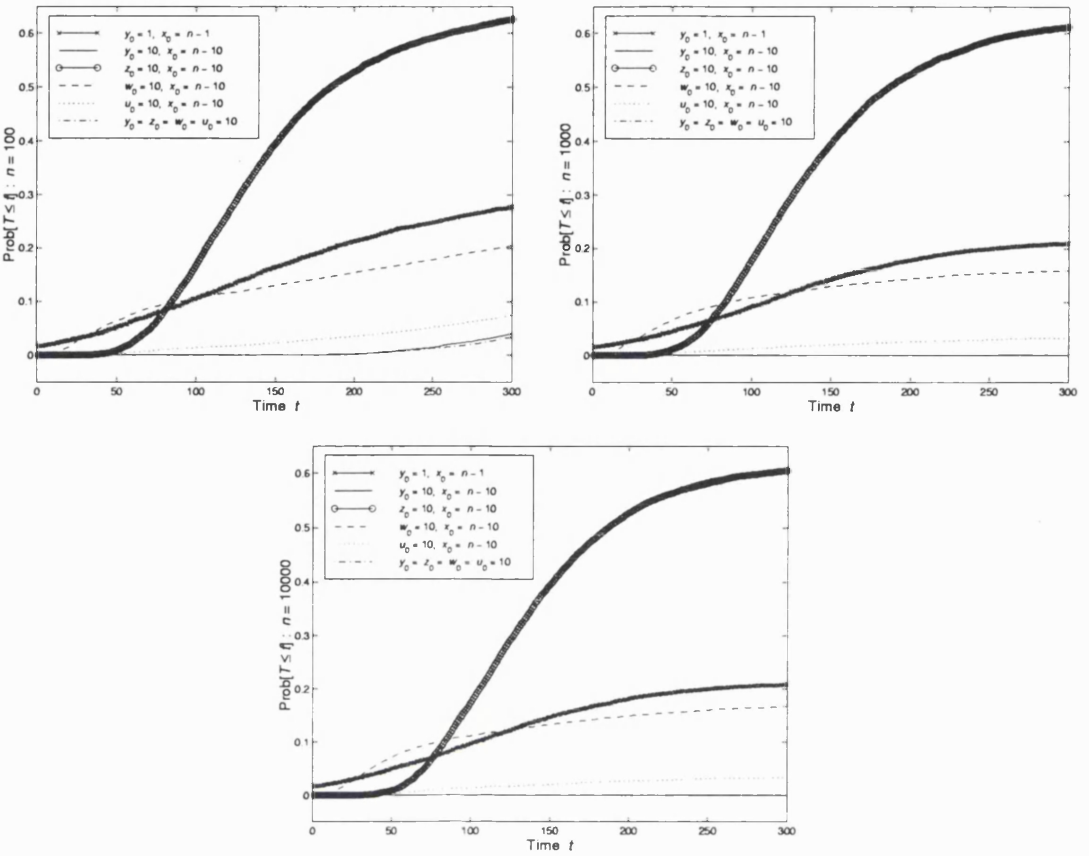


Figure 6.15: Cumulative distribution of extinction time, P[$\mathcal{T} \le t$], for $n = 100, 1000, 10000$. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). The parameter values are as shown in Table 6.8. Time is measured in years.

and parameter values is shown in Figure 6.15 (the distribution of the extinction time will be studied in more detail in Section 6.3.7). In the case with $y_0 = 10$ none of the individual simulation runs ended with extinction of the infection so that the conditional and unconditional means are the same (represented with one curve in Figure 6.14). These means are very similar to the deterministic values and basically they only differ slightly during the first 70 years.

In the case with $y_0 = 1$ though there are considerable differences between the three curves. The epidemic takes off more quickly with the deterministic formulation, so that the decrease in X and the increase in Y, Z, W, U is steeper in the deterministic values than in the stochastic means. The conditional and unconditional means are quite close in the beginning since the cumulative probability of extinction is very small (see Figure 6.15). As the probability of extinction increases though, the curve for the conditional means drifts away from the one for the unconditional means and moves closer to the deterministic curve.

In both cases the stochastic means seem to have reached a steady state by time $t = 300$, which is not the equilibrium of the process ($\mathrm{E}[X_e] = n$, $\mathrm{E}[Y_e] = \mathrm{E}[Z_e] = \mathrm{E}[W_e] = \mathrm{E}[U_e] = 0$). As was explained for the model in the previous chapter and in Section 6.3.3, this apparent stationarity can be explained by the existence of a quasi-stationary distribution or limiting-conditional distribution (see Sections 5.3.4 and 6.3.3).

Standard deviations

Results for the standard deviations of Y and Z for $n = 100$, $n = 1000$, and $n = 10000$ are shown in Figure 6.16, while Figure 6.17 shows the standard deviation of X, W, and U for $n = 10000$ (with $n = 100$ and $n = 1000$ the results are qualitatively similar, as it will be explained below, and are not shown here). The parameter values used are shown in Table 6.8 and the initial conditions in (6.24).

The results for the means presented earlier in this section suggest that by time $t = 300$ the process has settled down and hence we would expect that at that point the process either has reached the final equilibrium (extinction of TB) or it is still around the quasi-stationary level (this can be verified from the marginal distributions of X, Y, Z, W, U; see Section 6.3.6). Therefore we would expect to have small standard deviations for the cases in which the majority of the simulation runs ended at one of the two extremes (i.e. most runs had died out or most runs were at the quasi-stationary level).

For $n = 1000$ and $n = 10000$ the results in Figures 6.16 and 6.17 suggest that this must be the case. In both cases, at the end of the interval $\mathcal{I}_t$, the ordering in the value of the standard deviation, starting with the smallest, is the following:

$$\begin{gathered} y_0 = 10, x_0 = n - 10 \ \text{ and } \ y_0 = z_0 = w_0 = u_0 = 10 \\ u_0 = 10, x_0 = n - 10 \\ w_0 = 10, x_0 = n - 10 \\ y_0 = 1, x_0 = n - 1 \\ z_0 = 10, x_0 = n - 10. \end{gathered}$$

It should be noticed that this is exactly the same ordering as for the cumulative distribution function of the extinction time (Figure 6.15). For $n = 1000$ and $n = 10000$ with $y_0 = 10$ (and $x_0 = n - 10$ or $z_0 = w_0 = u_0 = 10$) none of the simulation runs

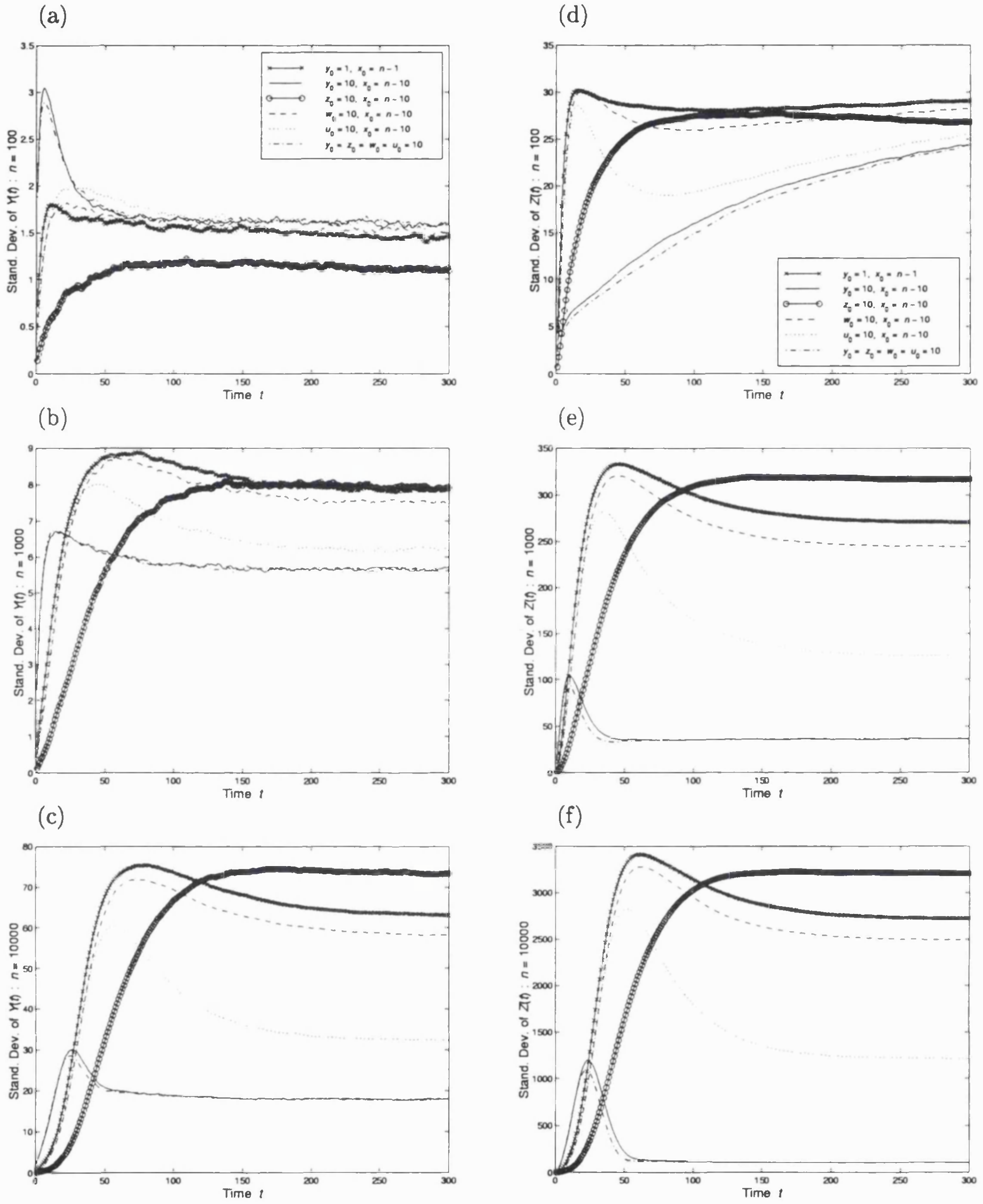


Figure 6.16: (a), (b), (c) Standard deviations of Y for $n = 100$, $n = 1000$, $n = 10000$, respectively; (d), (e), (f) standard deviations of Z for $n = 100$, $n = 1000$, $n = 10000$, respectively. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24) (the curves are labelled as in Figures (a) and (d)). The parameter values are as shown in Table 6.8. Time is measured in years.

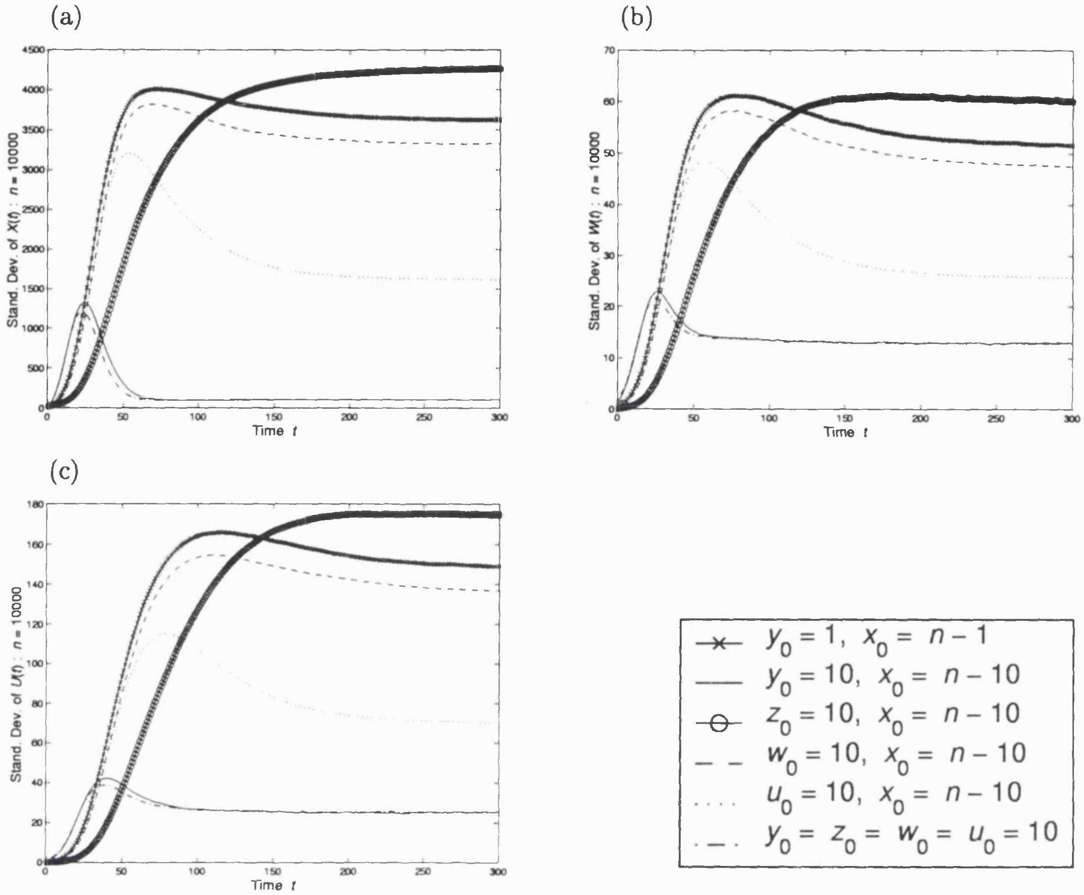


Figure 6.17: Standard deviations of (a) X, (b) W, and (c) U for $n = 10000$. In each graph there are six curves, one for each of the set of initial conditions shown in (6.24). The parameter values are as shown in Table 6.8. Time is measured in years.

ended with extinction and the process is most likely to be at the quasi-stationary level (i.e. most runs are at the quasi-stationary phase) so that the standard deviation is very small in this case. With $u_0 = 10, x_0 = n - 10$ only a few runs had died out so that the standard deviation is slightly larger. With $z_0 = 10, x_0 = n - 10$ about 60% of the runs ended with extinction so that the standard deviation is much larger in this case.

For $n = 100$ though things are slightly different. The standard deviation of X and Z is almost as for the larger values of n, but for Y, W, and U the ordering is exactly reversed: the cases with $y_0 = 10$ (and $x_0 = n - 10$ or $z_0 = w_0 = u_0 = 10$) have the largest standard deviation and the one with $z_0 = 10, x_0 = n - 10$ the smallest. This is probably a result of the fact that when $n = 100$ the population is quite small to preserve the infection, so that the process remains at the quasi-stationary level only for a short time and hence at time $t = 300$ some of the simulation runs are in the phase of falling from the quasi-stationary level to extinction. Therefore, although with large n each run

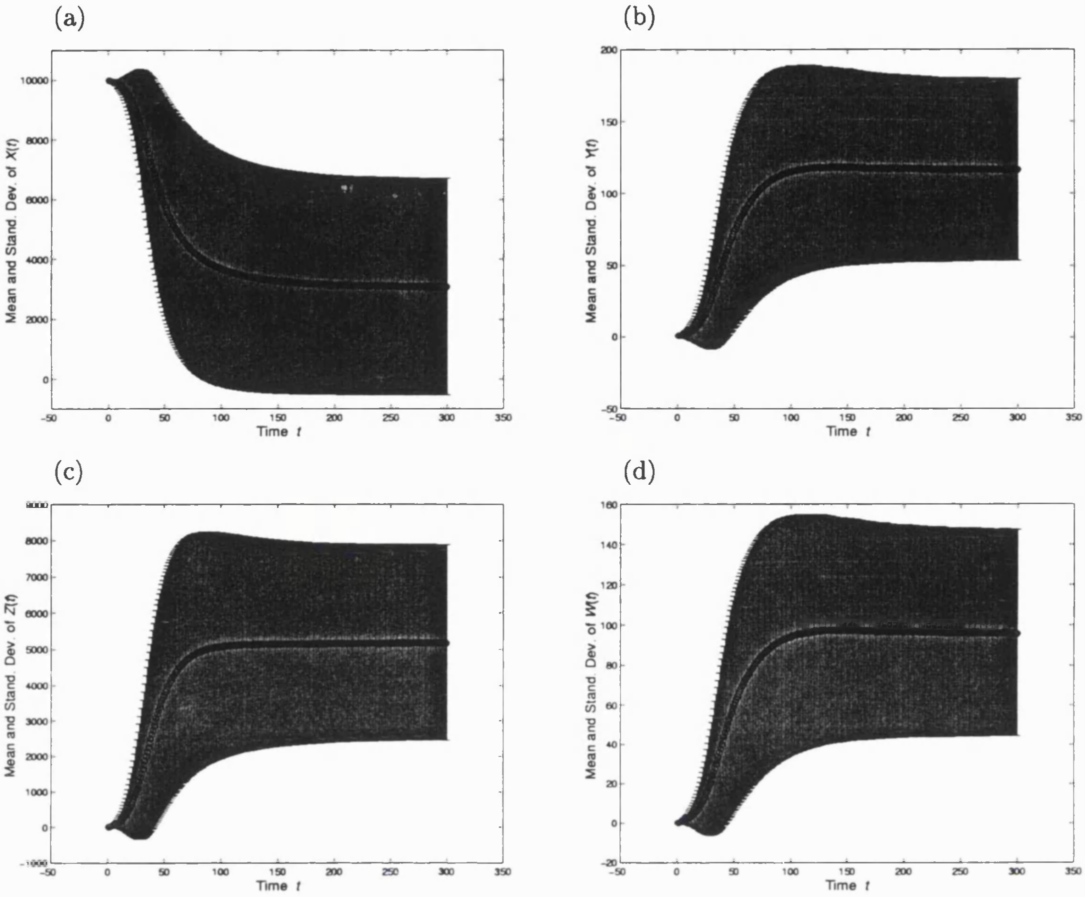


Figure 6.18: (a), (b), (c), (d) Mean and standard deviation of X, Y, Z, W, respectively, with $y_0 = 1$, $x_0 = n - y_0$, and $n = 10000$. The parameter values are as shown in Table 6.8. The centre of each circle represents the value of the mean at the respective time point and the length of the line segment above (and that below) the circle is equal to the standard deviation at the respective point. Time is measured in years.

was either around the quasi-stationary level or at extinction, with small n there are some runs which are at a level between extinction and quasi-stationarity. Therefore the ordering in the value of the standard deviation when $n = 100$ can be different than the one for $n = 1000$ and $n = 10000$ and also the ordering for the standard deviation of Y and Z may be different depending on how fast each of the variables X, Y, Z, W, U move towards extinction. The results from the marginal distributions presented in Section 6.3.6 support these arguments.

Finally in Figures 6.18 and 6.19 we present the results for the means and standard deviations combined, for the two cases $y_0 = 1$, $x_0 = n - 1$ and $y_0 = 10$, $x_0 = n - 10$. The value of n is 10000 and the parameters as shown in Table 6.8. The centre of each circle in Figures 6.18 and 6.19 represents the value of the mean at the respective time point and the length of the line segment above (and the one below) the circle is equal to the

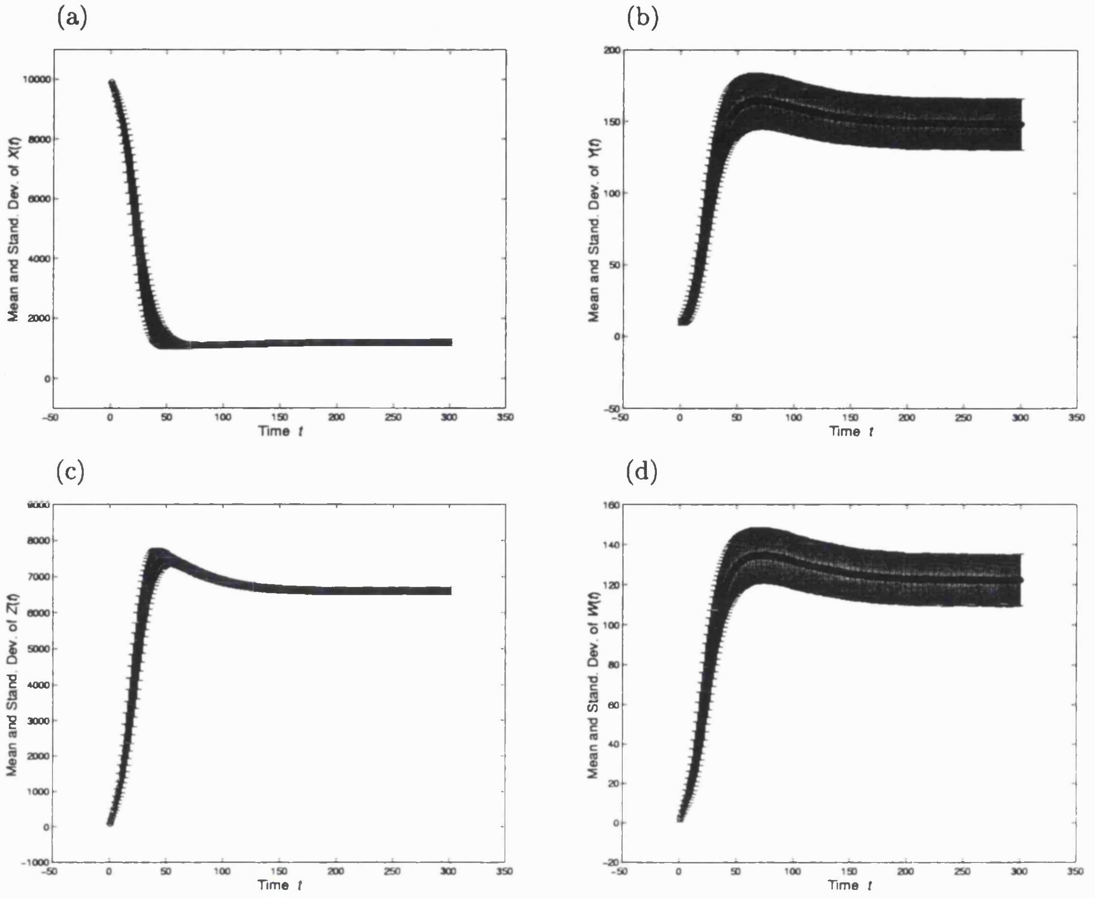


Figure 6.19: (a), (b), (c), (d) Mean and standard deviation of X, Y, Z, W, respectively, with $y_0 = 10$, $x_0 = n - y_0$, and $n = 10000$. The parameter values are as shown in Table 6.8. The centre of each circle represents the value of the mean at the respective time point and the length of the line segment above (and that below) the circle is equal to the standard deviation at the respective point. Time is measured in years.

standard deviation at the respective time point. Therefore the whole segment with the circle in the middle represents the interval $(\hat{\mu} - \hat{\sigma}, \hat{\mu} + \hat{\sigma})$ where $\hat{\mu}$ and $\hat{\sigma}$ are the estimates for the mean and standard deviation, respectively, at the particular time point.

These graphs show the enormous variation in the cases that begin with a small number of infectives and how much this variation is reduced if the epidemic begins with a larger number of infectives. As was explained above, this variation is a result of the initial phase until the epidemic takes off and the fact that the smaller the initial number of infectives (y_0) the larger the probability that the epidemic will die out. This is the reason why looking only at the means of X, Y, Z, W, U can be misleading, while graphs like those in Figures 6.18 and 6.19 can give a better picture of the evolution of the epidemic. Also they are helpful for comparisons with data from actual epidemics.

6.3.6 Marginal distributions

The marginal distributions of X, Y, Z, W, and U were calculated from simulations of the stochastic model for the following initial conditions:

$$\begin{array}{lll} n = 50, \quad y_0 = 1 & \qquad n = 100, \quad y_0 = 1 & \qquad n = 1000, \quad y_0 = 1 \\ n = 50, \quad y_0 = 5 & \qquad n = 100, \quad y_0 = 10 & \qquad n = 1000, \quad y_0 = 10, \end{array}$$

and $x_0 = n - y_0$. The parameter values used are as shown in Table 6.8. The cases with $n = 50$, $y_0 = 1$ and $n = 100$, $y_0 = 1$ were also examined with $\alpha = 4$ and $\alpha = 20$, respectively (and the other parameters as in Table 6.8).

The observations made from these cases are the same as the ones for the marginal distributions in the previous chapter (Section 5.3.5), so we present only the results for some of the cases here and briefly discuss about the others. Figure 6.20 shows the marginal distribution of Z for the cases with $n = 100$, $y_0 = 1$ and α equal to 10 and 20. The marginal distribution of Y for the cases with $n = 1000$ and y_0 equal to 1 and 10 is shown in Figure 6.21. As was explained in Section 5.3.5, the marginal distributions are bimodal, with one mode around the disease-free equilibrium and the other around the equilibrium of the conditional means (see Section 6.3.5 for the conditional means).

Figure 6.20 shows the effect of increasing $\mathcal{R}_0$: as the value of $\mathcal{R}_0$ increases more mass appears around the conditional means and less around the stochastic equilibrium. The same effect is observed by increasing y_0 (Figure 6.21) and n (results not shown here). This behaviour can be explained by the fact that for the more severe epidemics (with larger $\mathcal{R}_0$ or y_0) it takes more time until the infection dies out and it is more likely that the process will remain around the limiting-conditional distribution for a long time. The same holds for the cases with large initial population size n; the process fluctuates around the endemic conditional means for a long time unless the population size is too small to preserve the infection, which then dies out relatively soon (see, e.g., Bartlett 1957, 1960*a*). These results agree with the results for the distribution of the extinction time (see Section 6.3.7).

Because of this long time scale of the epidemic, the simulations have to be carried out for a very long time in order to demonstrate the ultimate extinction, so that our results here show only how the probability mass around the endemic level decreases in size and even then only for small values of n and y_0. For instance, in Figure 6.20 it can be observed how the probability $\mathrm{P}[Z(t) = 0]$ increases in time, although after 100 years

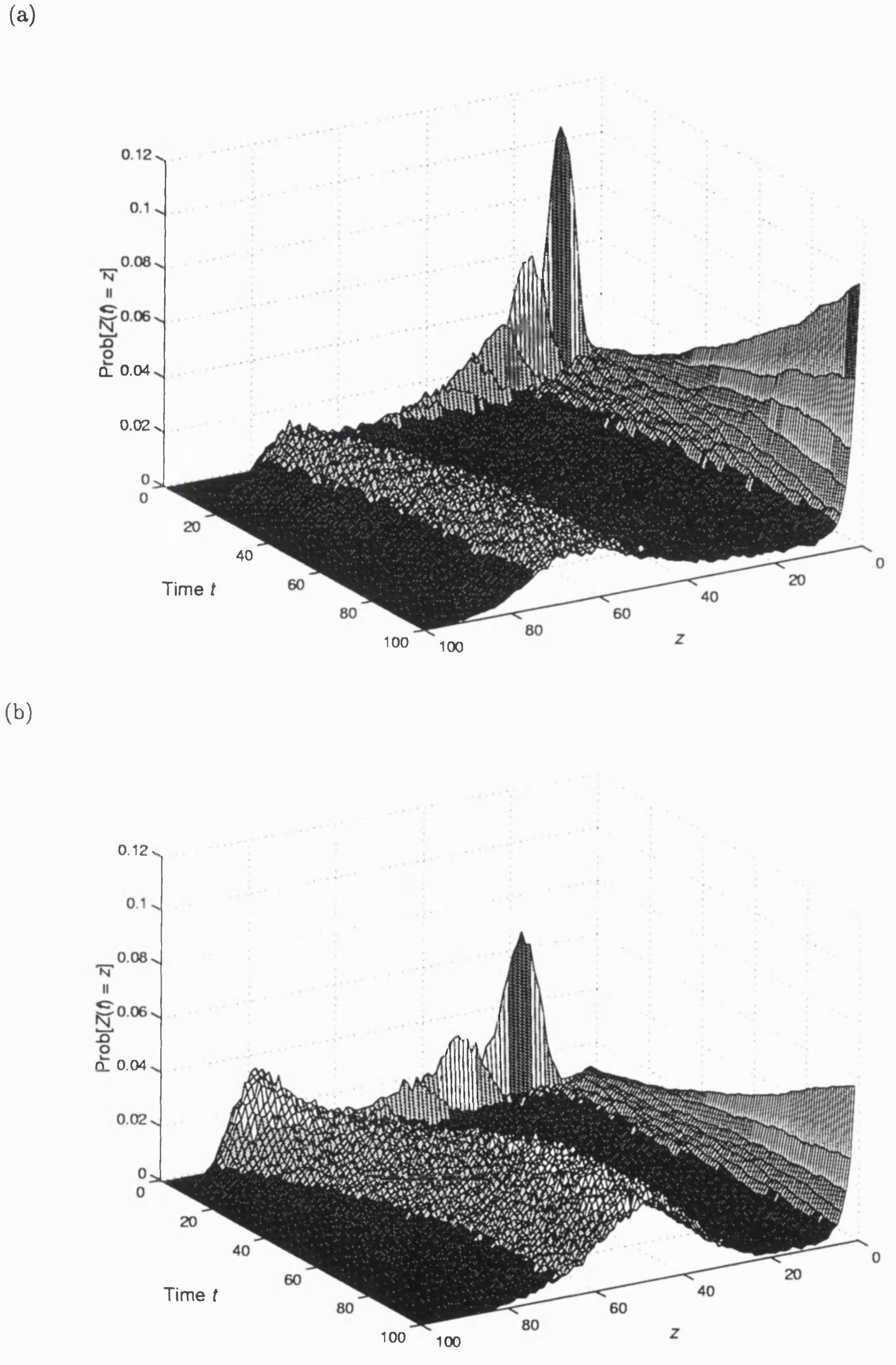


Figure 6.20: The marginal distribution of Z with $n = 100$, $y_0 = 1$, and $x_0 = n - 1$. The parameter values are as shown in Table 6.8 and (a) $\alpha = 10$ and (b) $\alpha = 20$. Time is measured in years.

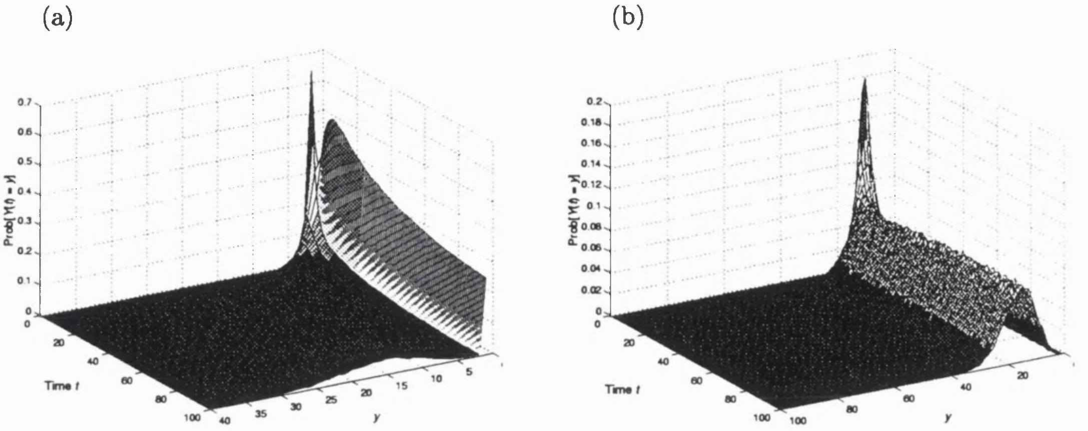


Figure 6.21: The marginal distribution of Y with (a) $y_0 = 1$, (b) $y_0 = 10$ and $n = 1000$, $x_0 = n - y_0$. The parameter values are as shown in Table 6.8 and time is measured in years.

it is still just around 0.1. One simulation was carried out for a long time and with small values of n, $\mathcal{R}_0$, and y_0 ($n = 50$, $\mathcal{R}_0 = 1.84$ and $y_0 = 1$, $x_0 = n - y_0$) and it took about 2500 years until the epidemic had died out in all 10^4 simulation runs (at that point the probability $\mathrm{P}[Y(t) = Z(t) = W(t) = U(t) = 0]$ was equal to 1).

6.3.7 The time until extinction

In this section we will study the distribution of the extinction time, $\mathcal{T}$. The results presented here were obtained from simulations of the stochastic model. Details for the implementation of the simulations can be found in the Appendix (Section A.3.3). The observations made for the distribution of $\mathcal{T}$ are the same as the ones for the distribution of the extinction time in the previous chapter (Section 5.3.6), so we present the results only for some cases here and briefly discuss about the others.

Processes starting with the introduction of one infectious case

First we consider only processes that begin with the introduction of one infectious TB case into an uninfected population, so that $y_0 = 1$ and $x_0 = n - 1$. The process was simulated for four different values of n (50, 100, 200, 400) and three values of α (2, 4, 6). The other parameter values are as shown in Table 6.8.

The statistics of the distribution of $\mathcal{T}$ and the value of $\mathcal{R}_0$ are shown in Table 6.9. Figure 6.22 shows the distribution of $\mathcal{T}$. As n and/or $\mathcal{R}_0$ increase, the mean and standard deviation of $\mathcal{T}$ increase and the probability $\mathrm{P}[\mathcal{T} \leq t]$ decreases for each $t \geq 0$, more so if both n and $\mathcal{R}_0$ increase. For instance when $\alpha = 2$ or $n = 50$, in more than 80% of the

		$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\alpha = 2$	Min:	0.002	0.001	0.000	0.000
	Q_1:	65.916	71.397	66.412	66.504
	Mean:	152.594	170.349	176.195	185.212
	Q_2:	131.974	140.352	139.136	141.613
	Q_3:	213.459	235.288	241.493	248.198
	Max:	1284.480	1244.439	1386.200	1762.013
	SD:	120.461	140.701	156.594	176.889
$\alpha = 4$	Min:	0.001	0.002	0.001	0.000
	Q_1:	108.765	112.541	116.114	116.062
	Mean:	233.716	319.459	549.948	1917.187
	Q_2:	194.455	225.987	259.410	267.739
	Q_3:	313.055	443.279	741.399	2449.284
	Max:	1885.835	2928.283	7237.947	34311.616
	SD:	181.468	299.629	697.133	3278.436
$\alpha = 6$	Min:	0.000	0.000	0.000	0.001
	Q_1:	143.062	151.371	161.539	161.294
	Mean:	303.350	549.674	2224.288	78132.950
	Q_2:	249.847	357.697	951.172	28572.778
	Q_3:	402.266	756.234	3263.992	114533.306
	Max:	2297.351	4657.944	26429.326	1384176.682
	SD:	232.413	572.194	2995.758	116489.417

Table 6.9: Statistics of the extinction time for processes that begin with one infectious case ($y_0 = 1$, $x_0 = n - 1$) and for various values of α and n. Q_i is the i-th quartile (for $i = 1, 2, 3$) and SD is the standard deviation. In some cases the minimum value is less than 0.0005 and hence it is rounded here to 0.000. The value of $\mathcal{R}_0$ is $\mathcal{R}_0 = 0.918, 1.836, 2.754$ for $\alpha = 2, 4, 6$, respectively.

10000 runs the epidemic had died out by time $t = 500$ while this proportion is about 40% when $\alpha = 6$ and n is 200 or 400.

The distribution of $\mathcal{T}$ is highly skewed to the right with a very high peak at the first 1–2 years. As n and/or α increases the distribution becomes less skewed and shifts to the right: the peak in the beginning has less mass and the tail of the distribution has more mass and becomes longer. The same qualitative behaviour was observed as $\mathcal{R}_0$ increases by changing the other parameters (results not shown here). We can conclude therefore that the effect of increasing $\mathcal{R}_0$ on the distribution of the extinction time is the same as the effect of increasing α, as described in this section.

These results suggest that as n and/or $\mathcal{R}_0$ increase it is more likely that the epidemic will last longer, either because the dynamics of the infection are strong (large $\mathcal{R}_0$) or because the population size is large enough to preserve the infection (see discussion at the end of the previous chapter).

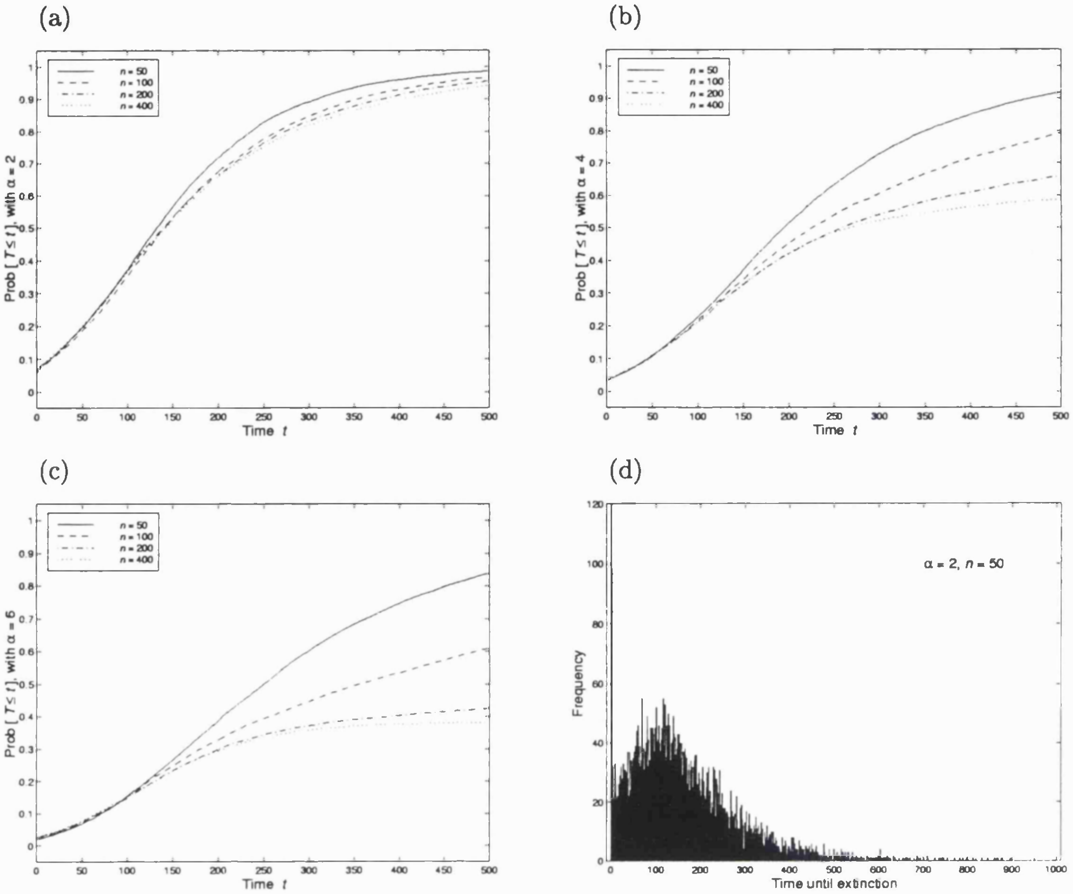


Figure 6.22: The distribution of the extinction time, $\mathcal{T}$, for processes that start with one infectious case ($y_0 = 1$, $x_0 = n - 1$). The parameter values are as shown in Table 6.8 except for α, which has the value indicated in each graph. Time is measured in years. (a), (b), (c) The cumulative distribution, $\mathrm{P}[\mathcal{T} \leq t]$, with $\alpha = 2, 4, 6$, respectively. (d) Histogram of the extinction time for the case with $\alpha = 2$ and $n = 50$.

Processes starting with more than one infected individual

In this part we consider only processes with $y_0 + z_0 + w_0 + u_0 > 1$. The process was simulated with three combinations of α and n: $\alpha = 2$, $n = 50$; $\alpha = 2$, $n = 100$; and $\alpha = 6$, $n = 50$; the other parameter values are as shown in Table 6.8. For each combination of α and n, the following initial conditions were used:

$$\begin{array}{ll} y_0 = 10\% \text{ of } n,\ x_0 = n - y_0 & w_0 = 10\% \text{ of } n,\ x_0 = n - w_0 \\ y_0 = 20\% \text{ of } n,\ x_0 = n - y_0 & w_0 = 20\% \text{ of } n,\ x_0 = n - w_0 \\ z_0 = 10\% \text{ of } n,\ x_0 = n - z_0 & u_0 = 10\% \text{ of } n,\ x_0 = n - u_0 \\ z_0 = 20\% \text{ of } n,\ x_0 = n - z_0 & u_0 = 20\% \text{ of } n,\ x_0 = n - u_0 \\ \multicolumn{2}{c}{y_0 = z_0 = w_0 = u_0 = 10\% \text{ of } n.} \end{array} \tag{6.25}$$

For each of these cases the statistics of the distribution of $\mathcal{T}$ are shown in Table 6.10.

	$\alpha = 2$ and $n = 50$								
	$y_0 = 5$	$y_0 = 10$	$z_0 = 5$	$z_0 = 10$	$w_0 = 5$	$w_0 = 10$	$u_0 = 5$	$u_0 = 10$	$y_0 = z_0 = 5$
									$w_0 = u_0 = 5$
Min	31.57	78.31	7.82	26.24	1.99	5.25	8.28	24.77	76.59
Q_1	180.00	196.96	72.88	109.64	38.55	110.09	107.01	176.95	199.41
Mean	257.36	276.01	135.08	179.79	146.80	200.23	203.52	258.87	278.03
Q_2	235.17	252.17	109.25	153.37	119.28	185.47	186.45	238.69	255.10
Q_3	307.42	327.98	166.25	222.40	217.99	267.45	273.73	317.12	330.96
Max	1286.40	1195.61	1092.33	949.07	1095.78	1263.28	1353.09	1140.39	1024.93
SD	112.61	112.52	95.35	101.49	128.63	125.83	127.38	121.41	112.05

	$\alpha = 2$ and $n = 100$								
	$y_0 = 10$	$y_0 = 20$	$z_0 = 10$	$z_0 = 20$	$w_0 = 10$	$w_0 = 20$	$u_0 = 10$	$u_0 = 20$	$y_0 = z_0 = 10$
									$w_0 = u_0 = 10$
Min	91.70	105.82	24.57	51.47	4.84	16.23	28.79	49.23	102.29
Q_1	246.56	260.60	109.96	153.23	122.53	196.76	195.30	249.62	265.29
Mean	341.58	357.09	185.85	240.50	231.17	294.33	295.05	344.80	360.49
Q_2	311.95	324.52	154.71	207.59	208.63	269.33	270.08	319.33	329.92
Q_3	403.45	418.10	226.14	291.79	310.58	365.20	368.91	408.62	420.37
Max	1421.15	1417.44	1147.84	1204.89	1332.44	1435.09	1610.13	1479.18	1349.60
SD	136.99	140.69	115.21	125.90	149.35	144.80	147.14	139.07	137.24

	$\alpha = 6$ and $n = 50$								
	$y_0 = 5$	$y_0 = 10$	$z_0 = 5$	$z_0 = 10$	$w_0 = 5$	$w_0 = 10$	$u_0 = 5$	$u_0 = 10$	$y_0 = z_0 = 5$
									$w_0 = u_0 = 5$
Min	36.72	85.72	10.05	23.16	1.10	3.19	6.54	25.44	91.58
Q_1	248.69	249.35	72.59	112.10	43.63	184.06	180.49	249.41	258.69
Mean	403.06	403.10	169.04	240.54	250.41	337.23	339.41	402.31	414.95
Q_2	342.86	344.54	112.25	170.52	198.49	288.36	296.11	346.07	353.95
Q_3	490.05	491.62	189.55	293.25	367.76	437.32	452.25	495.81	505.26
Max	2207.87	2548.87	1726.70	2094.49	2016.35	2185.46	2158.97	2331.14	2094.54
SD	225.96	218.38	169.30	201.99	243.56	236.84	235.88	228.04	224.50

Table 6.10: Statistics of the extinction time for processes that start with more than one infected individual: $y_0 + z_0 + w_0 + u_0 > 1$. Q_i is the i-th quartile (for $i = 1, 2, 3$) and SD is the standard deviation.

For the case with $a = 2$ and $n = 50$, Figure 6.23 shows the estimates of the cumulative distribution of $\mathcal{T}$ and Figure 6.24 shows histograms of $\mathcal{T}$ for some of the initial conditions examined. The other cases are qualitatively the same and are not shown here.

Comparing Figures 6.24 and 6.22(d), it can be observed that the distribution of $\mathcal{T}$ does not have the peak at time $t = 1$ that was observed in the cases with $y_0 = 1$. This happens because starting with 5, 10, or 20 infected individuals it is quite unlikely that all of them will die within a year. So that the minimum value of $\mathcal{T}$ is larger and the probability $\mathrm{P}[\mathcal{T} \leq t]$ is smaller (than when $y_0 = 1$).

With larger values of y_0 and/or u_0 it becomes less likely that the infection will die out soon (compared to cases with $y_0 = 1$), but this difference is not that noticeable for the cases with $z_0 > 0$ and $w_0 > 0$. For instance, the second peak in the distribution of $\mathcal{T}$ when $y_0 = 1$ shifts to the right as y_0 and/or u_0 increase, but with $z_0 > 0$ and $w_0 > 0$

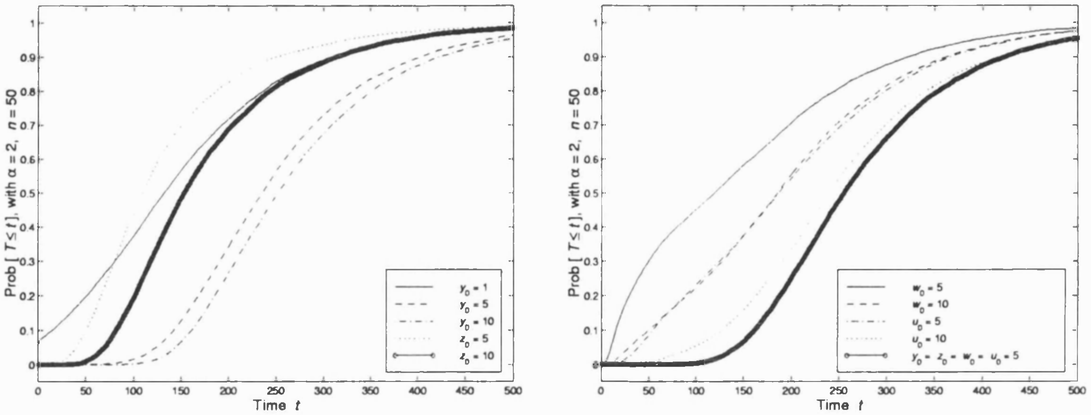


Figure 6.23: The cumulative distribution, $P[\mathcal{T} \le t]$, of the extinction time for processes that begin with more than one infected individual with $\alpha = 2$ and $n = 50$. The other parameter values are as in Table 6.8. In each graph the different curves are for each set of initial conditions shown in (6.25). Time is measured in years.

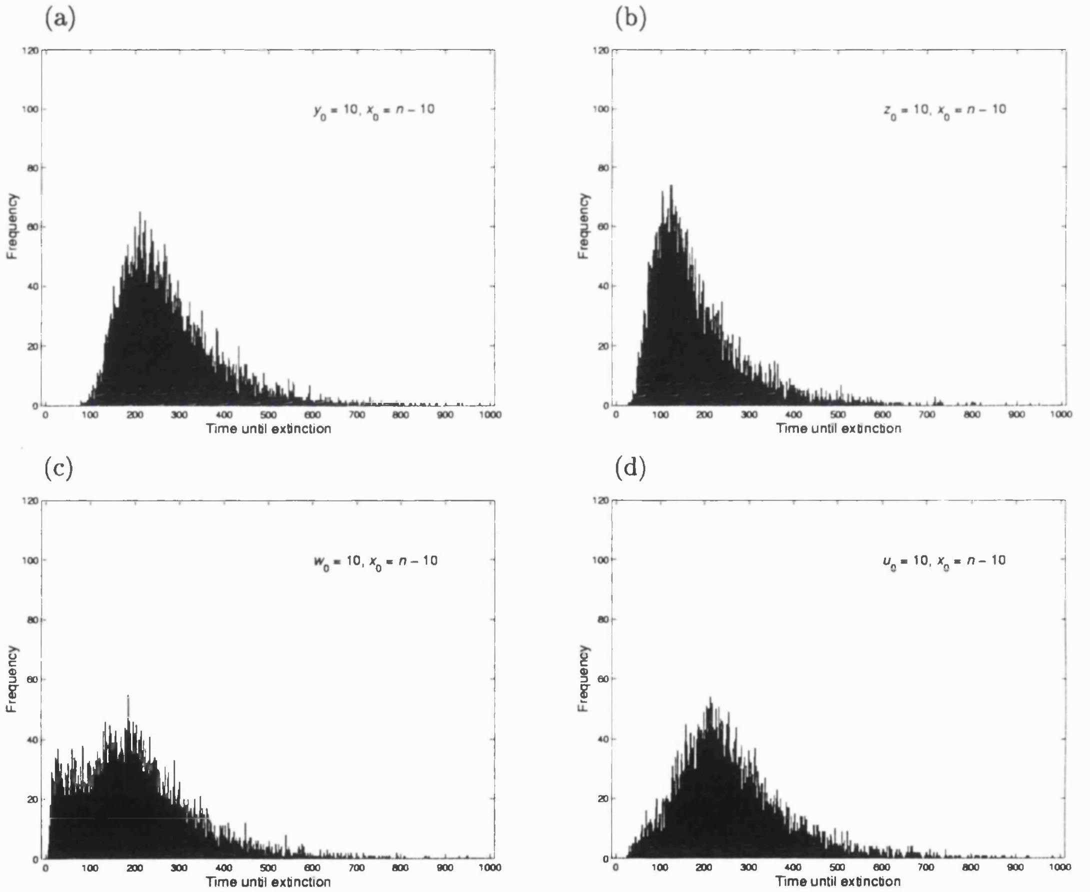


Figure 6.24: Histograms of the extinction time, $\mathcal{T}$, for processes that start with more than one infected individual: (a) $y_0 = 10$ and $x_0 = n - y_0$; (b) $z_0 = 10$ and $x_0 = n - z_0$; (c) $w_0 = 10$ and $x_0 = n - w_0$; (d) $u_0 = 10$ and $x_0 = n - u_0$. In all cases $n = 50$, $\alpha = 2$ and the other parameter values are as shown in Table 6.8. Time is measured in years.

there is still a lot of mass of probability during the first 50 years. Also the probability $\mathrm{P}[\mathcal{T} \leq t]$ decreases and the mean of $\mathcal{T}$ increases considerably as y_0 and u_0 increase.

This result agrees with the results in the Section 6.3.4, where it was shown that the prevalence, incidence, risk of infection, and mortality were higher for the cases with $y_0 > 1$ and $u_0 > 1$ than those with $z_0 > 1$ and $w_0 > 1$ (and the explanation is the same as in Section 6.3.4). Clearly, though, as each of the y_0, z_0, w_0, u_0 increases then it is more likely that the epidemic will last for a long time (for instance the mean of $\mathcal{T}$ increases and the $\mathrm{P}[\mathcal{T} \leq t]$ decreases) and more so for larger values of n and α.

6.3.8 Linear approximation

In this section we will study a model that approximates the stochastic model presented in this chapter. The model is basically the same as the stochastic model, the only difference is the assumption that X and Z (the numbers of uninfected and latents, respectively) evolve deterministically in time so that only $Y(t)$, $W(t)$, and $U(t)$ are stochastic variables. With this modification the stochastic model becomes linear, which is the reason why it is called a linear approximation. In the rest of this section this model will be referred to as the linear model or the linear approximation to model Zeus.

The assumption that X and Z evolve deterministically is equivalent to assuming that the rates of all the transitions that are functions of the numbers of uninfected and latents (X and Z, respectively) are functions of their mean values, $\mathrm{E}[X]$ and $\mathrm{E}[Z]$, instead. Hence, the means from the linear model are equal to the deterministic values. The advantage of the linear approximation is that it "adds" some stochasticity to the corresponding deterministic model and hence can be viewed as an approximation to the stochastic model. It is possible that the resulting linear model will not give accurate information about the original stochastic model, since some of the randomness of the original model is lost. Nevertheless, it has been observed that in some situations the linear model provides good estimates of the second moments of the remaining stochastic variables (see, e.g., Herbert 1998, Isham 1991), so in this section we investigate some of the properties of the linear model and compare the numerical results obtained with results from simulations of the stochastic model. Further comments about the applicability of this approximation can be found at the end of this section.

Let x, z denote the deterministic variables and Y_ℓ, W_ℓ, U_ℓ the stochastic variables for this model. The initial conditions are $x(0) = x_0$, $Y_\ell(0) = y_0$, $z(0) = z_0$, $W_\ell(0) = w_0$, $U_\ell(0) = u_0$, where $(x_0, y_0, z_0, w_0, u_0) \in S_0$ and S_0 is as defined in (6.3).

The state probabilities are defined as

$$p^\ell_{ywu}(t) = \mathrm{P}[Y_\ell(t) = y, W_\ell(t) = w, U_\ell(t) = u], \quad t \geq 0,$$

for $(y, w, u) \in S_\ell$ and zero otherwise, where

$$S_\ell = \{(y, w, u) \in \mathbb{Z}^3 : y, w, u \geq 0\}.$$

At time zero $p^\ell_{y_0w_0u_0}(0) = 1$ and $p^\ell_{ywu}(0) = 0$ for any $(y, w, u) \neq (y_0, w_0, u_0)$. The corresponding Kolmogorov equations for $p^\ell_{ywu}(t)$ are given in the Appendix (Section A.3.4).

The joint probability generating function for Y_ℓ, W_ℓ, U_ℓ,

$$\mathcal{P}_\ell(\theta_2, \theta_4, \theta_5; t) = \mathrm{E}[\theta_2^{Y_\ell(t)} \theta_4^{W_\ell(t)} \theta_5^{U_\ell(t)}]$$

satisfies the equation

$$\begin{aligned}
\frac{\partial \mathcal{P}_\ell}{\partial t} = {} & \beta z(t)[q_2(\theta_2 - 1) + (1 - q_2)(\theta_4 - 1)]\mathcal{P}_\ell \\
& + [L_1(t)\theta_2(\theta_2 - 1) + L_2(t)\theta_2(\theta_4 - 1) + (\mu + \mu_1)(1 - \theta_2) + \gamma_0(\theta_5 - \theta_2)]\frac{\partial \mathcal{P}_\ell}{\partial \theta_2} \\
& + [\delta(\theta_2 - \theta_4) + \delta_0(\theta_5 - \theta_4) + (\mu + \mu_2)(1 - \theta_4)]\frac{\partial \mathcal{P}_\ell}{\partial \theta_4} \\
& + [\epsilon_1(\theta_2 - \theta_5) + \epsilon_2(\theta_4 - \theta_5) + \mu(1 - \theta_5)]\frac{\partial \mathcal{P}_\ell}{\partial \theta_5},
\end{aligned} \tag{6.26}$$

where

$$\begin{aligned}
L_1(t) &= pq_1\frac{\alpha}{n}x(t) + q_3\frac{\alpha_2}{n}z(t) \\
L_2(t) &= p(1 - q_1)\frac{\alpha}{n}x(t) + (1 - q_3)\frac{\alpha_2}{n}z(t),
\end{aligned} \tag{6.27}$$

with the initial condition $\mathcal{P}_\ell(\theta_2, \theta_4, \theta_5; 0) = \theta_2^{y_0}\theta_4^{w_0}\theta_5^{u_0}$. The $x(t)$, $z(t)$ that appear in equation (6.26), as well as in (A.9), are the deterministic values of x and z at time t which are deduced from the system (6.1).

From equation (6.26) a system of differential equations for the first and second moments of Y_ℓ, W_ℓ, U_ℓ is deduced; the equations for the means are

$$\begin{aligned}
\frac{d\mathrm{E}[Y_\ell(t)]}{dt} &= [L_1(t) - \Gamma_s]\mathrm{E}[Y_\ell] + \delta\mathrm{E}[W_\ell] + \epsilon_1\mathrm{E}[U_\ell] + q_2\beta z(t) \\
\frac{d\mathrm{E}[W_\ell(t)]}{dt} &= L_2(t)\mathrm{E}[Y_\ell] - \Delta_s\mathrm{E}[W_\ell] + \epsilon_2\mathrm{E}[U_\ell] + (1 - q_2)\beta z(t) \\
\frac{d\mathrm{E}[U_\ell(t)]}{dt} &= \gamma_0\mathrm{E}[Y_\ell] + \delta_0\mathrm{E}[W_\ell] - \mathrm{E}_s\mathrm{E}[U_\ell],
\end{aligned} \tag{6.28}$$

where the Γ_s, Δ_s, and E_s are as defined in (6.2). The equations for the variances and covariances are given in the Appendix (Section A.3.4).

The system (6.28) for the first moments of Y_ℓ, W_ℓ, U_ℓ depends on the values of $x(t)$, $z(t)$ but not on the second moments, so that it is a closed system. Similarly the system for the second moments (see Section A.3.4) does not contain moments of higher order. This is one of the advantages of the linear approximation, which basically aims at linearising the model making the equations for the moments linear. Therefore the systems for the moments of any order are closed and no further approximation is needed in order to deduce their values.

In general, the variables to be taken as deterministic will be chosen among the variables that appear in the non-linear terms in the original model (so that these terms will become linear), provided that their size is large enough. The reason for that is that in general the behaviour of the mean of a stochastic variable can be similar to that of the corresponding deterministic variable provided that its size is relatively large (see, e.g., Bailey 1975, Ch. 5). For this model for example, X, Y, Z are the variables involved in non-linear terms. From the results obtained thus far it appears that the sizes of both X and Z will be quite large, but not that of Y, therefore assuming that X and Z evolve deterministically may give a good approximation.

Also it has to be noted that the equations (6.28) for the means of Y_ℓ, W_ℓ, U_ℓ are the same as the second, fourth, and fifth equations of the system (6.1) for the deterministic y, w, u. Therefore, the values of the stochastic means $\mathrm{E}[Y_\ell(t)]$, $\mathrm{E}[W_\ell(t)]$, $\mathrm{E}[U_\ell(t)]$ will be the same as the values of the deterministic $y(t)$, $w(t)$, $u(t)$ for any $t \geq 0$.

The equilibrium of the linear model

Define the variables $Y_{\ell e}$, $W_{\ell e}$, $U_{\ell e}$ by

$$\begin{aligned} q^{\ell e}_{ywu} &= \mathrm{P}[Y_{\ell e} = y, W_{\ell e} = w, U_{\ell e} = u] \\ &\equiv \lim_{t\to\infty} \mathrm{P}[Y_\ell(t) = y, W_\ell(t) = w, U_\ell(t) = u], \end{aligned}$$

for $(y, w, u) \in S_\ell \equiv \mathbb{Z}^3_+$. Also, let x_e, z_e be the equilibrium values of the deterministic $x(t)$, $z(t)$, respectively:

$$x_e = \lim_{t\to\infty} x(t) \qquad\qquad z_e = \lim_{t\to\infty} z(t).$$

Taking the limits in equation (A.9) as t tends to infinity, we deduce the equivalent difference equation for the probabilities $q^{\ell e}_{ywu}$. It is easy to show that for $y = w = u = 0$ this equation gives

$$\beta z_e q^{\ell e}_{000} = (\mu + \mu_1) q^{\ell e}_{100} + (\mu + \mu_2) q^{\ell e}_{010} + \mu q^{\ell e}_{001},$$

from which it follows immediately that if $z_e \neq 0$ then $q^{\ell e}_{000} \neq 1$, while if $z_e = 0$ then $q^{\ell e}_{100} = q^{\ell e}_{010} = q^{\ell e}_{001} = 0$. Using this result and taking recursively the equation (A.9) at equilibrium for $y = 0, 1, 2, \dots ; w = 0, 1, 2, \dots ; u = 0, 1, 2, \dots$, it can be shown that if $z_e = 0$ then $q^{\ell e}_{ywu} = 0$ for any $(y, w, u) \in S_\ell - \{(0, 0, 0)\}$. Assuming that $\sum q^{\ell e}_{ywu} = 1$ (summing over all $(y, w, u) \in S_\ell$) then follows that $q^{\ell e}_{000} = 1$. Therefore with the linear formulation, extinction of the infection is not always certain as with the stochastic model (see Section 6.3.2); the infection ultimately dies out with probability one only when $z_e = 0$, but otherwise the probability of extinction $q^{\ell e}_{000}$ is less than one.

Since z_e is the equilibrium value of the deterministic $z(t)$ the results from Section 6.2 can be used in order to determine when $z_e = 0$. In Section 6.2 it was shown that the deterministic model has three possible equilibria $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ (see equations (6.6) and (6.7) for definitions), where $\mathbf{e}_1$ is the point that corresponds to extinction (and hence with $z_e = 0$). The point $\mathbf{e}_3$ is unstable whenever it is feasible. If $\mathcal{R}_0 < 1$, $\mathcal{R}_1 < 1$ (see (6.9) and (6.13) for definitions) then $\mathbf{e}_1$ is the only feasible equilibrium and it is stable; if $\mathcal{R}_0 < 1$ and $\mathcal{R}_1 > 1$ then $\mathbf{e}_1$ is stable, but also $\mathbf{e}_2$ is feasible and possibly stable; if $\mathcal{R}_0 > 1$ and $\mathcal{R}_1 > 1$ then $\mathbf{e}_1$ is unstable and $\mathbf{e}_2$ is feasible and possibly stable.

It follows from these results that if both $\mathcal{R}_0$ and $\mathcal{R}_1$ are less than one then $z_e = 0$ and $q^{\ell e}_{000} = 1$; if both $\mathcal{R}_0$ and $\mathcal{R}_1$ are greater than one then $z_e \neq 0$ and hence $q^{\ell e}_{000} \neq 1$ and hence extinction is not certain; when $\mathcal{R}_0 < 1$ and $\mathcal{R}_1 > 1$ then z_e can be equal to zero, depending on the initial conditions.

Numerical results

In Section 5.4 the linear approximation was used for the model presented in Chapter 5. As was explained in Section 5.4.2, one of the difficulties encountered in most epidemic models is the fact that the system of differential equations for the first and second moments involves higher order moments, so that it is open and cannot be solved. One way to overcome this problem is to approximate the higher order moments by expressions that involve only first and second order moments. For instance, if $(X_1, X_2, \dots , X_k)'$ has a multivariate normal distribution then

$$\mathrm{E}[X_1 X_2 X_3] = \mu_1\mu_2\mu_3 + \mu_1\sigma_{23} + \mu_2\sigma_{13} + \mu_3\sigma_{12} \tag{6.29}$$

where $\mu_i = \mathrm{E}[X_i]$ and $\sigma_{ij} = \mathrm{Cov}[X_i, X_j]$, and with similar expressions for $\mathrm{E}[X_1^2 X_2]$, $\mathrm{E}[X_1 X_2^2]$ and so on. Substituting for the third order moments from these expressions, makes the system for the first and second moments closed and hence it can be solved at least numerically. The normal approximation can be justified by results showing that certain Markov processes can be approximated by a normal distribution as the initial total population size tends to infinity (see, e.g., Whittle 1957, Kurtz 1970, 1971, 1981).

The linear approximation is another way to overcome this problem, since the model is linear and hence the system for the first and second moments is closed. In this section we present numerical results from both the normal and the linear approximations and compare them with results from the simulations. For the linear approximation the system of differential equations for the means, variances, and covariances was solved numerically using routines from the NAG library for Fortran. For the normal approximation the third order moments appearing in the equations for the first and second moments (see equations (6.19) for the means and Section A.3.2 for the variances and covariances) were expressed in terms of the first and second moments (using the formulae like (6.29)) and the resulting system was solved numerically using routines from the NAG library. Details for the implementation of the simulations can be found in the Appendix (Section A.3.3).

For the results discussed in this section we used the parameter values shown in Table 6.8 and the following sets of initial conditions:

$$n = 1000 \qquad y_0 = 10, x_0 = n - y_0 \tag{6.30}$$

$$n = 1000 \qquad y_0 = z_0 = w_0 = u_0 = 10 \tag{6.31}$$

$$n = 10000 \qquad y_0 = 10, x_0 = n - y_0 \tag{6.32}$$

$$n = 10000 \qquad y_0 = 100, x_0 = n - y_0. \tag{6.33}$$

Also the case with $n = 1000$, $y_0 = 10$, $x_0 = n - y_0$ was examined with two slightly different sets of parameter values: (a) $\alpha = 9$ and the rest of the parameters as in Table 6.8, (b) $\beta = 0.001$ and the rest of the parameters as in Table 6.8. For all cases the three methods were carried out up to time $t = 300$ so that the results discussed here are for the time interval $\mathcal{I}_t = [1, 300]$.

Figure 6.25 shows the results for the means and Figure 6.26 the results for standard deviations and covariances for the case (6.30). In each graph there are three curves, one from the linear model, one from the normal approximation, and one from the simulations. For the means of X, Z the values from the linear model correspond to the values of the deterministic x, z. Also the means of Y, W, U from the linear model are equal to the deterministic y, w, u, respectively. The results for the cases (6.31), (6.32), and (6.33) and the ones with $\alpha = 9$ and $\beta = 0.001$ are qualitatively the same and are not presented here.

In all the cases examined the results for the means, standard deviations, and covariances from the normal approximation and the simulations were very close (in some cases the two curves seem to coincide). The results from the linear approximation are very close to the cthers in the beginning (the first 10–15 years) but then deviate, in some cases substantially (for instance for the covariances in Fig. 6.26). For the means in case (6.32) the difference between the three curves became smaller in time (see Fig. 6.27 for the means of Y and Z) and after the first 100 years the three curves can be hardly distinguished. That was observed only for the means in case (6.32) though; in all other situations the difference between the linear and the other two curves remained up to the end, $t = 300$.

It is also interesting to note that the standard deviations and covariances calculated from the linear approximation were larger than those from the normal approximation and the simulations: during the first 10–15 years all three values (linear, normal, simulation) are very close, but the linear is slightly bigger. The standard deviations and covariances from the linear approximation increase more rapidly than those from the normal approximation and the simulation and the curves for the linear model remain higher than the other two until the end, $t = 300$. This was observed in all cases except (6.32): in this case (Figure 6.28) all standard deviations and covariances from the linear model are smaller than the other two during the first 30–40 years and then the linear curve crosses over the other two and remains higher for the rest of the interval $\mathcal{I}_t$.

The discrepancies from the linear model may be explained from the sizes of X and Z (the variables that are taken as deterministic). As can be observed from Figure 6.25, in the beginning the value of X is very large, although Z is rather small, and the linear approximation agrees with the other two methods. As the value of X falls, the

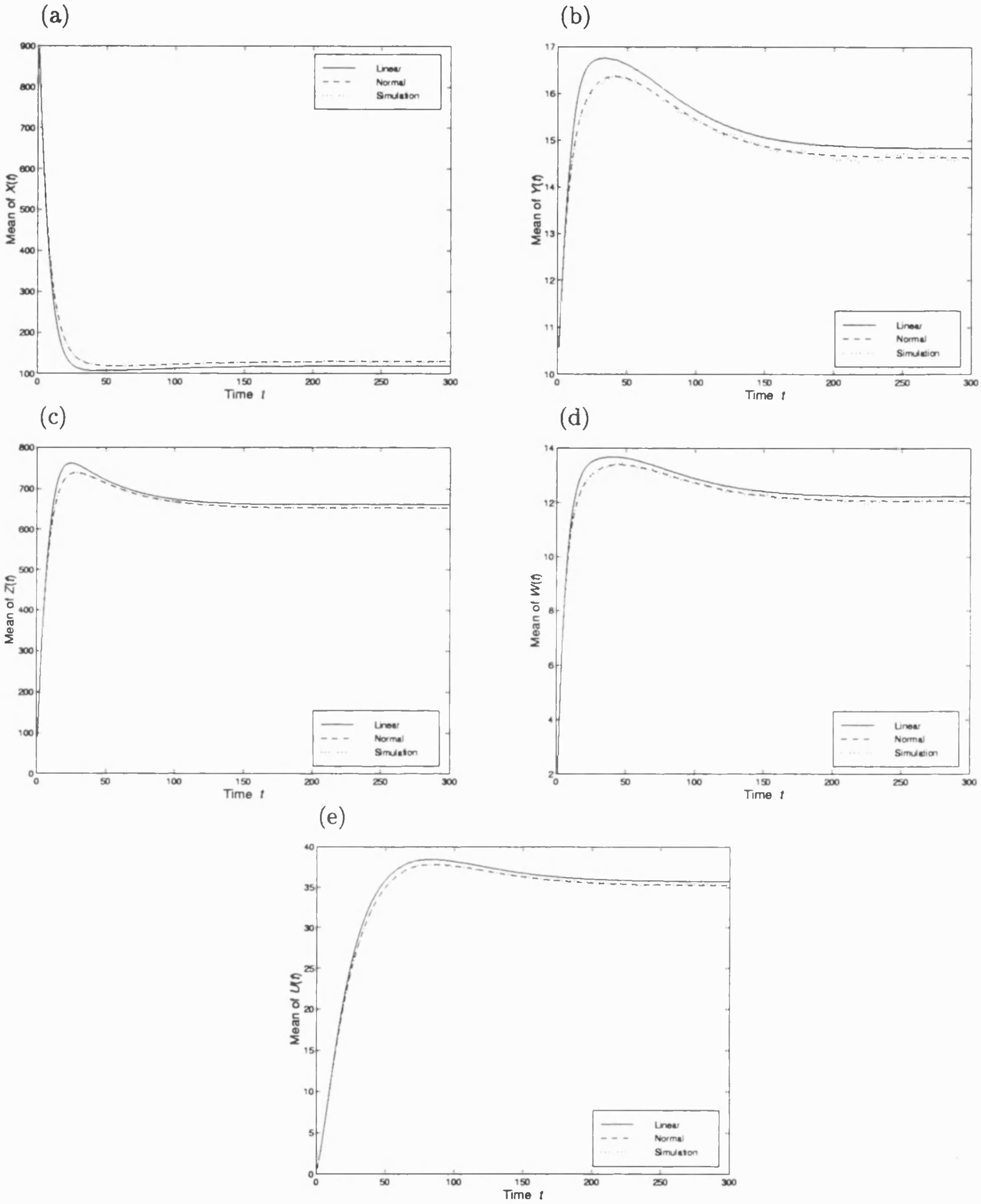


Figure 6.25: The means of (a) X, (b) Y, (c) Z, (d) W, and (e) U from the linear and normal approximations and simulations of the stochastic model. Each graph has three curves one for the linear model, one for the normal, and one from the simulations. The values for X and Z from the linear model are the deterministic x and z, respectively. The parameter values are as shown in Table 6.8 and $n = 1000$, $y_0 = 10$, $x_0 = n - y_0$. Time is measured in years.

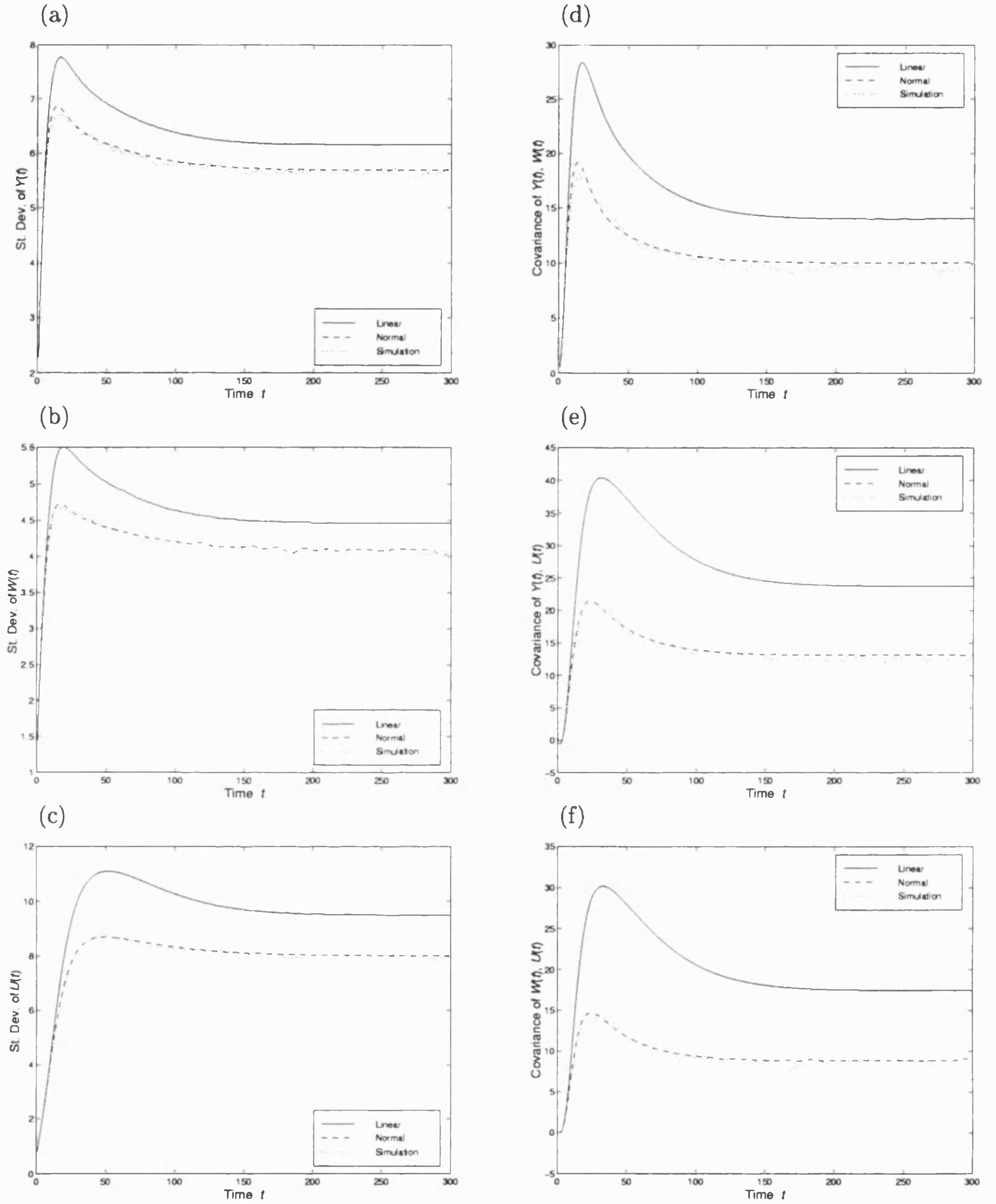


Figure 6.26: (a), (b), (c) Standard deviation of Y, W, U, respectively. (d), (e), (f) Covariance of Y, W; Y, U; and W, U, respectively. Each graph has three curves, one for the values from the linear model, one from the normal approximation, and one from the simulations. The parameter values are as shown in Table 6.8 and $n = 1000$, $y_0 = 10$, $x_0 = n - y_0$. Time is measured in years.

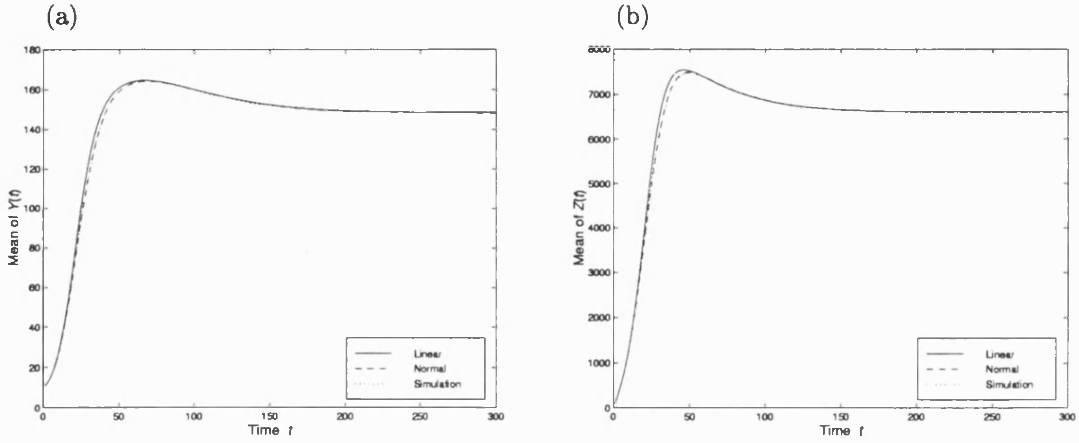


Figure 6.27: The means of (a) Y and (b) Z with $n = 10000$, $y_0 = 10$, $x_0 = n - y_0$. Each graph has three curves one for the values from the linear model, one from the normal approximation, and one from the simulations. The values for Z from the linear model are the deterministic values. The parameter values are as shown in Table 6.8. Time is measured in years.

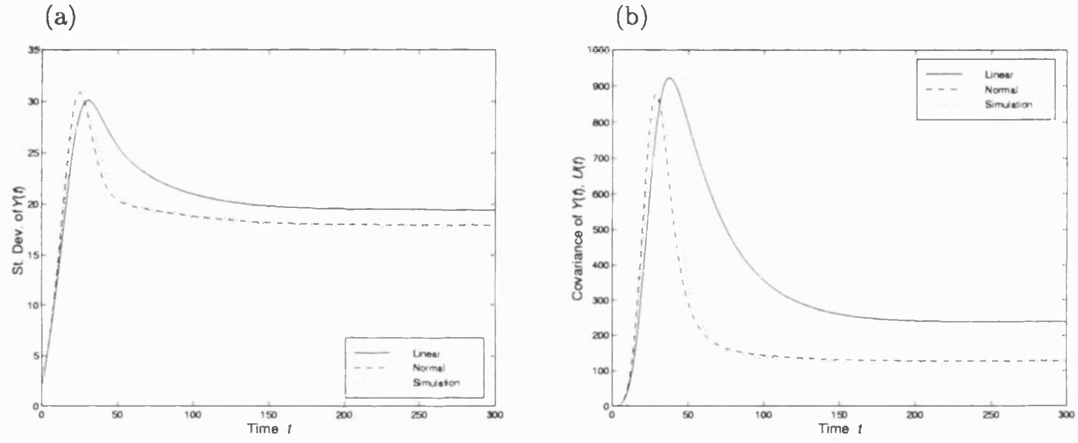


Figure 6.28: (a) Standard deviation of Y and (b) covariance of Y, U as obtained from the linear and normal approximations and simulations of the stochastic model. The parameter values are as shown in Table 6.8 and $n = 10000$, $y_0 = 10$, $x_0 = n - y_0$. Time is measured in years.

discrepancies in the second moments from the linear approximation become larger. This may be simply due to the fact that $X(t)$ is not large enough any more for the linear model to give good approximation. Although the size of Z is small in the beginning and large later (when the discrepancies in the second moments appear), this may not affect the goodness of the linear approximation, because the effective contact rate (α_2) for the latents is much smaller than that (α) for the uninfected (and hence it is mainly the size of X that will determine the quality of the linear approximation).

Overall, from the results examined thus far it appears that the results from the normal approximations and the simulations agree in all cases, and throughout the time interval until the epidemic reaches its endemic level. In contrast, the results from the linear approximation agree very closely in the beginning of the epidemic only, but then they deviate, slightly for the means but more significantly for the standard deviations and covariances. In particular the linear approximation tends to overestimate the second moments, a phenomenon that has been observed in other situations as well (see, e.g., Isham 1991).

Discussion

The device of linearising a stochastic model is clearly advantageous from the point of view that it makes the model simpler and more attractive and manageable mathematically. Unfortunately, for the model presented in this chapter, taking two of the variables as deterministic means that there are still three stochastic variables, so that the model is not that simple. Also having x and z evolving deterministically does not mean that these two variables are completely eliminated from the model, since the values of the deterministic $x(t)$ and $z(t)$ still appear in the equations for the state probabilities $p^{\ell}_{ywu}(t)$ and the probability generating function. Overall therefore the linear model is simpler than the original stochastic one, but it is still complicated enough for analytical results to be difficult to deduce from it. Clearly for a model like the one presented in the previous chapter the linear approximation yields a more manageable linear model, but for model Zeus more simplifying assumptions have to be made in order to deduce a really simple model.

On the other hand, from the point of deducing numerical results, the linear approximation seems more promising. It can be considered as a method for moment closure, giving a closed system for the first and second moments that approximate those of the original model. Also, compared to simulations it is a much quicker method, since the results are easily obtained by numerical solutions of systems of differential equations. Finally, compared to the deterministic model, the linear model has clearly the advantage of giving information about the variation of some of the variables: the means from the linear model are the same as the deterministic values, but from the linear model the variances and covariances (and hence confiᵈence intervals) of the non-deterministic variables can also be deduced. This information is completely lost with the deterministic formulation.

Nevertheless, the linear approximation still has some drawbacks. First of all there is the loss of information about the variances and covariances of the variables that are taken as deterministic. For situations where it is not enough to know only the means of these variables, this can be a significant disadvantage. Moreover, the fact that some of the variables are taken as deterministic means that the linear model accounts for less variation within the whole model and that can affect even the means and covariances of the remaining stochastic variables. Also, from the results presented in this section it appears that although the means can be quite well approximated by the linear model (i.e. the deterministic values), this does not always hold for the variances and covariances, which are significantly overestimated by the linear model in some cases. From that respect, the normal approximation seems to be better than the linear, since the moments from the simulations and the normal approximation closely agree, but they do not agree that well with those from the linear approximation. Finally it has to be noted that the accuracies of both the normal and the linear approximation depend on the sizes of the total population and/or the individual classes.

Chapter 7

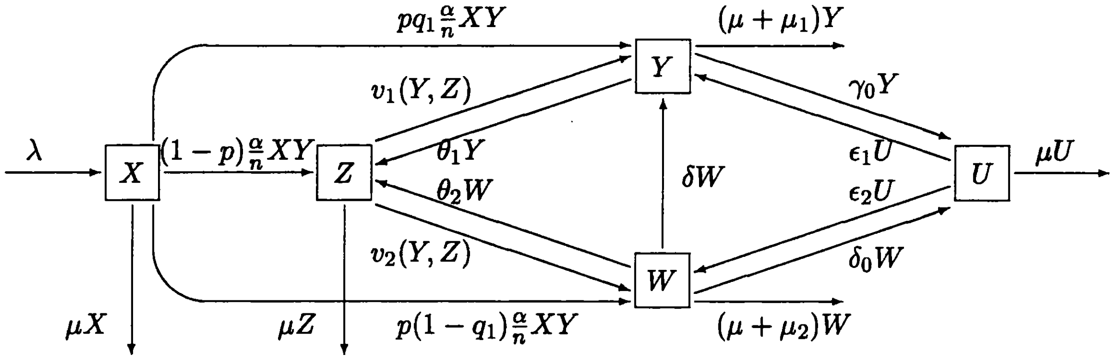
Model Clio: a model for chemotherapy

7.1 Introduction

In this chapter we will study the effects of chemotherapy. We assume that we have a population in which TB is introduced at some point and the natural evolution of the infection (in the absence of any control measures) is described by the model Zeus (presented in the previous chapter). Therefore, we have a population divided into five classes, the uninfected (X), the latents (Z), the infectious TB cases (Y), the non-infectious TB cases (W), and those naturally recovered (U). The possible transitions and the assumptions made for model Zeus are described in detail in Section 6.1. The main characteristics of the model are briefly described below.

If at time t there are $X(t)$ uninfected and $Y(t)$ infectious cases in the population then the probability of one new infection occurring in the interval $[t, t + dt]$ is $\alpha X(t)Y(t)dt/n + o(dt)$ where α is the effective contact rate. Among those who get infected a proportion p develop clinical TB within a year after infection (primary TB) and the remaining proportion, $1 - p$, become latents; those who develop TB are infectious or non-infectious with probabilities q_1 and $1 - q_1$, respectively.

Latents may develop clinical disease at some point as a result of exogenous reinfection (acquiring a new infection) or endogenous reactivation of an old infection. The reactivation rate is denoted by β. After reactivation the individual has infectious or non-infectious TB with probabilities q_2 and $1 - q_2$, respectively. The effective contact rate between latents and infectious cases is $p_r\alpha$. After reinfection an individual develops clinical disease within a year (primary TB) or remains latent with probabilities p_3 and $1-p_3$, respectively. For simplicity we denote $\alpha_2 = p_3 p_r \alpha$.



$$v_1(Y,Z) = q_2\beta Z + q_3\tfrac{\alpha_2}{n}YZ \qquad v_2(Y,Z) = (1-q_2)\beta Z + (1-q_3)\tfrac{\alpha_2}{n}YZ$$

Figure 7.1: A model for chemotherapy: Model Clio

Non-infectious TB cases become infectious at a rate δW. Infectious and non-infectious cases recover spontaneously at rates γ_0 and δ_0 per capita, respectively, and those who have recovered may relapse later and become infectious or non-infectious cases at rates $\epsilon_1 U$ and $\epsilon_2 U$, respectively.

Finally, there is immigration of susceptibles at a constant rate λ, normal death at a rate μ per capita, and excess death due to TB at rates μ_1 and μ_2 (per capita) for the infectious and non-infectious cases, respectively. At some points the special case $\lambda = \mu n$ is also investigated.

As was explained in the previous chapter, after the introduction of the infection into the population, the infection spreads and finally settles down at an endemic level, described by the mean of the quasi-stationary distribution (see Sections 6.3.3 and 6.3.6). In this chapter we will assume that chemotherapy is introduced in the population after the infection has reached this steady endemic level and study how that will affect the further progress of the epidemic.

Suppose that a proportion θ_{1da} of the infectious TB cases, Y, are diagnosed and a proportion θ_{1db} of those diagnosed receive treatment. Let $\theta_{1d} = \theta_{1da}\theta_{1db}$. In most countries nowadays $\theta_{1db} = 1$, which means that any individual diagnosed with TB receives treatment, so the rate θ_{1d} is called the detection rate (for those with infectious

TB). Among those treated, a proportion θ_{1c} is cured and become non-infectious. The remaining proportion, $1 - \theta_{1c}$, although receiving treatment, fail to convert to sputum-negative or become non-infectious for a very short time (say less than a year) and relapse later. This proportion covers those individuals who completed their treatment but the treatment itself is not effective for the particular individual and also those patients who did not complete the therapy or did not follow the therapy correctly, for various reasons (it is a fact that a proportion of those treated remain infectious, see, e.g., Dye et al. 1998, Grosset 1989, Styblo 1989, Murray et al. 1993).

Patients who are treated but do not convert to sputum-negative may be less infectious than those in the Y class (since they have received at least some treatment, which possibly has had some effect, i.e. killed a number of tubercle bacilli) but either they are slightly less infectious than those in the Y class or they are less infectious than the Y only for a short time and then they are equally infectious, so that we can assume that they are approximately as infectious as those in the Y class (see, e.g., Grosset 1989, Murray et al. 1993). Therefore the assumption made here is that a proportion $\theta_1 = \theta_{1d}\theta_{1c}$ of the infectious TB cases is cured and removed out of the class Y.

Similar assumptions are made for the non-infectious cases, W. A proportion θ_{2da} of the non-infectious cases is diagnosed and a proportion θ_{2db} of those diagnosed receive treatment. Let $\theta_{2d} = \theta_{2da}\theta_{2db}$. Among those treated a proportion θ_{2c} is cured. The remaining proportion, $1 - \theta_{2c}$, although receiving treatment, remain ill, for reasons similar to those mentioned above for the proportion $1-\theta_{1c}$ of the infectious cases who are not cured (see, e.g., Dye et al. 1998, Grosset 1989). Therefore a proportion $\theta_2 = \theta_{2d}\theta_{2c}$ of the non-infectious cases is cured and removed from the W class.

The parameter θ_{2c} can be taken as equal to θ_{1c}, since there is no evidence that treatment is more or less effective depending on whether the patient is infectious or not (see, e.g., Dye et al. 1998). In this chapter we assume that $\theta_{2c} = \theta_{1c}$ and hence $\theta_2 = \theta_{2d}\theta_{1c}$. From a public health point of view, it is helpful to express the detection rate θ_{2d} for non-infectious cases as a proportion of the detection rate θ_{1d} for infectious cases (since the control programs are aiming firstly on the infectious cases and secondly on the non-infectious, so that it is useful to know the relative ratio θ_{2d}/θ_{1d}). Therefore in the rest of this chapter θ_{2d} will be expressed as $\theta_{2d} = \theta_{2r}\theta_{1d}$ in some cases, and hence $\theta_2 = \theta_{2r}\theta_1$, where $0 \leq \theta_{2r} \leq 1$ gives the relative detection rate for non-infectious cases.

Among those who are treated with chemotherapy and cured, a proportion will relapse later and become TB cases (see, e.g., Styblo 1991, Chan & Yew 1998). When they relapse, there are two possible scenarios:
(a) they become infectious if they were infectious before treatment, and non-infectious if they were non-infectious before treatment.
(b) they become infectious or non-infectious with probabilities q_2' and $1-q_2'$, respectively, independently of their status (infectious or non-infectious) before treatment.

There is no study in the medical literature that proves either of the two scenarios and possibly the truth lies between the two. Most modellers assume that the status (infectious or non-infectious) after treatment is independent of the status before treatment (see, e.g., Dye et al. 1998, Joesoef et al. 1989). It is not unreasonable to assume that they are independent since, for instance, treatment may decrease the number of tubercle bacilli, so that a previously infectious case that relapses may become non-infectious. Therefore for the model presented in this chapter we assume that they are independent, and hence the infectiousness of a patient who relapses after effective treatment does not depend on his/her infectiousness before treatment.

There is also evidence that reinfection after successful treatment is possible (see, e.g., Small et al. 1993). Unfortunately there have been very few studies on reinfection of cured individuals, so that it is not clear how likely this is. Reinfection of those cured may be a very rare event (except in some high risk groups, like individuals with immunosuppression) so that the probability can be considered as negligible. On the other hand, maybe it is negligible only in areas with low risk of infection, while in areas with high risk it is very likely to happen (as with reinfection of those with a latent infection). The model presented here (as the model Zeus) is intended mainly for studying the effects of chemotherapy in developing countries, where the risk of infection and prevalence are high, so that reinfection even of those cured after successful treatment may not be negligible, and therefore it is included as a possibility in this model.

For those who get reinfected after successful treatment there is again the question of whether their status (infectious or non-infectious) after reinfection is independent of their status before treatment or not. Unfortunately there is not enough evidence in the medical literature to support either of the two theories and since it is not unreasonable to assume that they are independent we will adopt this assumption.

We assume that the effective contact rate between infectious cases and those cured after treatment is $p'_r\alpha$. If a patient who has been cured by treatment is reinfected, he/she develops primary TB (within a year) with probability p'_3, and then becomes infectious or non-infectious with probability q'_3, $1 - q'_3$, respectively. Let $\alpha'_2 = p'_3p'_r\alpha$. Then one way to incorporate chemotherapy in model Zeus is to add a class U' for those cured after treatment and the following transitions from state (X, Y, Z, W, U, U') in the interval $[t, t + dt]$:

to the state	at rate
$X, Y - 1, Z, W, U, U' + 1$	$\theta_1 Y$
$X, Y + 1, Z, W, U, U' - 1$	$q'_2\beta'$
$X, Y + 1, Z, W, U, U' - 1$	$q'_3\alpha'_2 YU'/n$
$X, Y, Z, W - 1, U, U' + 1$	$\theta_2 W$
$X, Y, Z, W + 1, U, U' - 1$	$(1 - q'_2)\beta'$
$X, Y, Z, W + 1, U, U' - 1$	$(1 - q'_3)\alpha'_2 YU'/n.$

It has to be noted that the class U' has to be separate from the class U because the relapse rates (from U, U' to Y, W) for those naturally recovered are much higher than those for the recovered after treatment (see, e.g., Styblo 1991, Springett 1971, Grosset 1989). This is also the reason why reinfection of the naturally recovered has not been included in the models (neither in Zeus nor in Clio), since the relapse rate for this class is so high that the possibility of reinfection can be ignored (see, e.g., Styblo 1991, Springett 1971).

On the other hand, if

$$\begin{matrix} q_2 = q'_2 & \quad \text{and} \quad & \beta = \beta' \\ q_3 = q'_3 & & \alpha_2 = \alpha'_2, \end{matrix} \tag{7.1}$$

then the classes Z and U' do not have to be separate. Again, there is not enough evidence in the medical literature to prove whether the equalities (7.1) hold or not. If we assume that they are not equal then the classes Z and U' have to be separate, so that we increase the number of variables by one and the number of parameters by four, making the model more complicated and (mathematically) unattractive. In addition, by assuming that the equalities (7.1) do not hold, the uncertainty about the actual values of the eight parameters involved in (7.1) may result in errors in the numerical results deduced from this model which balance the errors incurred if we assume that (7.1) hold (if in reality

$X(t)$	Number of uninfected individuals at time t
$Y(t)$	Number of infectious TB cases at time t
$Z(t)$	Number of inactive cases at time t
$W(t)$	Number of non-infectious TB cases at time t
$U(t)$	Number of naturally recovered patients at time t
λ	Immigration of uninfected individuals
μ	Normal death rate (per capita)
μ_1	Excess death rate due to TB for infectious cases (per capita)
μ_2	Excess death rate due to TB for non-infectious cases (per capita)
α	The effective contact rate between uninfected and infectious cases
p	Probability of developing primary TB (after first infection)
q_1	Probability of developing infectious TB for those with primary TB (after the first infection)
β	Reactivation/relapse rate for inactive cases
q_2	Probability that reactivation/relapse of inactive cases leads to infectious TB
$p_r\alpha$	The effective contact rate between inactive and infectious cases
p_3	Probability of developing primary TB (after reinfection)
α_2	$\alpha_2 = p_3 p_r \alpha$
q_3	Probability of developing infectious TB for those with primary TB (after reinfection)
δ	Rate at which non-infectious cases become infectious
γ_0	Natural recovery rate for infectious cases
δ_0	Natural recovery rate for non-infectious cases
ϵ_1	Relapse rate to the infectious class (for those naturally recovered)
ϵ_2	Relapse rate to the non-infectious class (for those naturally recovered)
θ_1	Recovery rate after successful treatment with chemotherapy for the infectious cases
θ_2	Recovery rate after successful treatment with chemotherapy for the non-infectious cases
θ_{2r}	Relative detection rate for non-infectious cases (assuming $\theta_{2c} = \theta_{1c}$)
n	Initial total population size

Table 7.1: Variables and parameters used in model Clio

they do not). Finally, several results in the literature indicate that the relapse rates and the effective contact rates for the latents and those cured with chemotherapy are very small (see, e.g., Sutherland et al. 1982, Chan & Yew 1998, Enarson & Rouillon 1998, Dolin et al. 1994), and it seems unlikely that they will differ substantially. Taking all these things into consideration, we will assume from this point on that the equalities (7.1) do hold and therefore the Z and U' classes can be combined in one class which will be referred to as the inactive class.

The definitions of the variables and parameters used for this model are summarised in Table 7.1. The possible transitions and their rates are illustrated in Figure 7.1. If $\theta_1 = \theta_2 = 0$ then this model is the same as the model presented in the previous chapter.

7.2 The deterministic model

For the corresponding deterministic model, let $x(t)$, $y(t)$, $z(t)$, $w(t)$, and $u(t)$ denote the number of uninfected, infectious cases, inactive cases, non-infectious cases, and recovered, respectively, at time t. The differential equations for x, y, z, w, and u are:

$$\begin{aligned}
\frac{dx}{dt} &= -\frac{\alpha}{n}xy - \mu x + \lambda \\
\frac{dy}{dt} &= pq_1\frac{\alpha}{n}xy + q_3\frac{\alpha_2}{n}yz - (\gamma_0 + \theta_1 + \mu + \mu_1)y + q_2\beta z + \delta w + \epsilon_1 u \\
\frac{dz}{dt} &= (1-p)\frac{\alpha}{n}xy - \frac{\alpha_2}{n}yz + \theta_1 y - (\beta + \mu)z + \theta_2 w \\
\frac{dw}{dt} &= p(1-q_1)\frac{\alpha}{n}xy + (1-q_3)\frac{\alpha_2}{n}yz + (1-q_2)\beta z - (\delta + \delta_0 + \theta_2 + \mu + \mu_2)w + \epsilon_2 u \\
\frac{du}{dt} &= \gamma_0 y + \delta_0 w - (\epsilon_1 + \epsilon_2 + \mu)u.
\end{aligned} \tag{7.2}$$

Here x, y, z, w, and u are non-negative continuous functions. The initial conditions are $(x(0), y(0), z(0), w(0), u(0)) = (x_0, y_0, z_0, w_0, u_0) \in \mathcal{S}_0$, where

$$\mathcal{S}_0 = \{\mathbf{x} = (x, y, z, w, u) \in \mathbb{Z}_+^5 : 1 \le x \le n-1, x + y + z + w + u = n\}, \tag{7.3}$$

and $n \ge 2$ is the initial total population size: $n = x_0 + y_0 + z_0 + w_0 + u_0$. From the system (7.2) a differential equation is deduced for the total population size, $N(t) = x(t) + y(t) + z(t) + w(t) + u(t)$, which with integration gives

$$N(t) = \frac{\lambda}{\mu} + e^{-\mu t}\left(n - \frac{\lambda}{\mu}\right) - e^{-\mu t}\int_0^t e^{\mu s}[\mu_1 y(s) + \mu_2 w(s)]ds, \tag{7.4}$$

for $t \ge 0$. This equation is the same as equation (6.5) for the total population size in the previous chapter. From (7.4) it follows that $N(t)$ (and hence $x(t)$, $y(t)$, $z(t)$, $w(t)$, and $u(t)$ as well) is always bounded above by n if $\lambda \le \mu n$ and by λ/μ if $\lambda > \mu n$.

Solving the system (7.2) with the derivatives on the left-hand side equal to zero follows that the system (7.2) admits three possible equilibria, $\mathbf{e}_i = (x_i^e, y_i^e, z_i^e, w_i^e, u_i^e)$, for $i = 1, 2, 3$. The first of these points, $\mathbf{e}_1$, corresponds to the extinction of the infection: $\mathbf{e}_1 = (\lambda/\mu, 0, 0, 0, 0)$. The coordinates of the other two points, $\mathbf{e}_2$ and $\mathbf{e}_3$, are given in the Appendix (Section A.4.1).

Depending on the parameter values, the coordinates of $\mathbf{e}_2$ and $\mathbf{e}_3$ may be positive, negative, or complex. So we will call an equilibrium point $\mathbf{e}$ feasible if all its coordinates are non-negative. $\mathbf{e}_1$ is always feasible. It can be shown that $\mathbf{e}_2$ and $\mathbf{e}_3$ are feasible if and only if $0 \le x_2^e \le \lambda/\mu$ and $0 \le x_3^e \le \lambda/\mu$, respectively. Using the Routh-Hurwitz criterion (Theorem 5.1), it can be shown that $\mathbf{e}_1$ is unstable if $\mathcal{R}_0 > 1$, where $\mathcal{R}_0$ is the basic reproduction ratio

$$\mathcal{R}_0 = \frac{\alpha}{n}\frac{\lambda}{\mu}\frac{H_1}{H_2} \tag{7.5}$$

(see Section A.4.1 for the calculation of $\mathcal{R}_0$), with H_1 and H_2 defined as

$$\begin{aligned} H_1 &= pq_1[\Phi_s(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) - \theta_2 \mathrm{E}_s(1-q_2)\beta] \\ &\quad + p(1-q_1)[\Phi_s(\delta \mathrm{E}_s + \delta_0\epsilon_1) + \theta_2 \mathrm{E}_s q_2 \beta] \\ &\quad + (1-p)[q_2\beta(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_2)\beta(\delta \mathrm{E}_s + \delta_0\epsilon_1)] \\ H_2 &= \Phi_s[\Gamma_s(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) - \gamma_0(\delta\epsilon_2 + \epsilon_1\Delta_s)] \\ &\quad - \theta_1[q_2\beta(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_2)\beta(\delta \mathrm{E}_s + \delta_0\epsilon_1)] \\ &\quad - \theta_2[q_2\beta\gamma_0\epsilon_2 + (1-q_2)\beta(\Gamma_s \mathrm{E}_s - \gamma_0\epsilon_1)], \end{aligned}$$

where, for simplicity, the parameter values have been grouped as follows:

$$\begin{aligned} \Gamma_s &= \gamma_0 + \theta_1 + \mu + \mu_1 \qquad & \mathrm{E}_s &= \epsilon_1 + \epsilon_2 + \mu \\ \Delta_s &= \delta + \delta_0 + \theta_2 + \mu + \mu_2 \qquad & \Phi_s &= \beta + \mu. \end{aligned} \tag{7.6}$$

Unfortunately the feasibility and stability of the equilibrium points cannot be further investigated analytically (due to the complexity of the algebra involved), but for specific parameter values of interest they can be studied numerically. Figure 7.2 shows the values of x_1^e, x_2^e, x_3^e for a particular set of parameter values (see legend of the graph). These values are not all representative of TB, but they are used here in order to show that it is possible to have multiple stable equilibria when $\mathcal{R}_0 < 1$. The value of θ_1 varied over the values $\theta_1 = 0.0001, 0.0002, \ldots, 0.9999, 1.0000$ thus giving a grid of the parameter space with respect to the value of θ_1.

From Figure 7.2 it can be observed that for $0.0001 \leq \theta_1 \leq 0.3140$ all three points $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are feasible (since $0 \leq x_i^e \leq \lambda/\mu = 100$, for all $i = 1, 2, 3$). Using the Routh-Hurwitz criterion (Theorem 5.1) it can be shown that $\mathbf{e}_1$ and $\mathbf{e}_2$ are stable, but $\mathbf{e}_3$ unstable. The value of $\mathcal{R}_0$ is less than one in this subspace of the parameter space. On the other hand, for $\theta_1 = 0.3141, \ldots, 1.0$ ($\mathcal{R}_0 < 1$ also here) only $\mathbf{e}_1$ is feasible and stable, while $\mathbf{e}_2$ and $\mathbf{e}_3$ are infeasible (x_2^e and x_3^e are complex).

These results show that, as for the model presented in the previous chapter, the behaviour of the deterministic model does not conform to the usual behaviour observed in most epidemic models (where for $\mathcal{R}_0 < 1$ the disease-free equilibrium is stable and

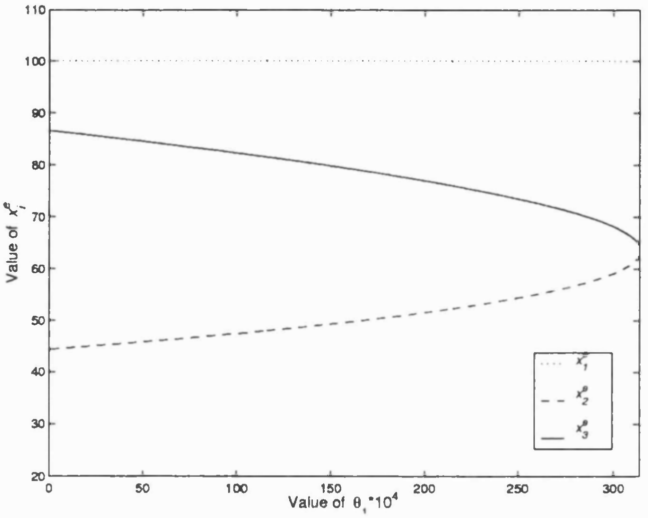


Figure 7.2: The values of x_1^e, x_2^e, x_3^e for a particular set of parameter values: $n = 100$, $\mu = 0.0222$, $\mu_1 = 0.3$, $\mu_2 = 0.21$, $\lambda = \mu n$, $q_1 = 0.55$, $q_2 = 0.55$, $q_3 = 0.055$, $p = 0.1$, $p_r = 0.6$, $\alpha = 6$, $\alpha_2 = p_r\alpha$, $\beta = 0.0001$, $\delta = 0.015$, $\gamma_0 = 0.2$, $\delta_0 = 0.2$, $\epsilon_1 = 0.03$, $\epsilon_2 = 0.03$, $\theta_2 = 0.1$, and θ_1 varies over $\theta_1 = 0.0001, 0.0002, \ldots, 0.3139, 0.3140$. For $\theta_1 \geq 0.3141$ the values of x_2^e and x_3^e are not real.

the endemic equilibrium unstable or even infeasible, while for $\mathcal{R}_0 > 1$ the disease-free equilibrium is unstable and the endemic stable). Here, it is possible to have $\mathcal{R}_0 < 1$ and still two equilibria stable, one corresponding to extinction and one endemic.

Finally, for the special case with $q_1 = q_2 = q_3 = q$ (most of the numerical results presented in this chapter are for this case), it can be shown that the third equilibrium point, $\mathbf{e}_3$, is not feasible throughout the parameter space, while $\mathbf{e}_2$ is feasible if and only if $\mathcal{R}_0 > 1$. Therefore, in this case, we have only two possible equilibria, $\mathbf{e}_1$ and $\mathbf{e}_2$. If $\mathcal{R}_0 < 1$ only $\mathbf{e}_1$ is feasible, but when $\mathcal{R}_0 > 1$, $\mathbf{e}_1$ is unstable and $\mathbf{e}_2$ feasible.

7.3 The stochastic model

7.3.1 The transient phase

Let $p_{\mathbf{x}}(t) = p(x, y, z, w, u; t)$ be the probability that there are x uninfected individuals, y infectious cases, z inactive cases, w non-infectious cases, and u recovered in the population at time $t \geq 0$, for $\mathbf{x} = (x, y, z, w, u) \in \mathcal{S} = \mathbb{Z}_+^5$, $t \geq 0$ and $p_{\mathbf{x}}(t) = 0$ whenever $\mathbf{x} \notin \mathcal{S}$. The initial conditions are $p_{\mathbf{x}_0}(0) = 1$ and $p_{\mathbf{x}}(0) = 0$ for any $\mathbf{x} \neq \mathbf{x}_0 = (x_0, y_0, z_0, w_0, u_0)$, where $\mathbf{x}_0 \in \mathcal{S}_0$ and $\mathcal{S}_0$ is as defined in (7.3). The corresponding Kolmogorov forward equations for $p_{\mathbf{x}}(t)$ are given in the Appendix (Section A.4.2). The probability generating function, $\mathcal{P}(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5; t) = \mathrm{E}[\phi_1^{X(t)}\phi_2^{Y(t)}\phi_3^{Z(t)}\phi_4^{W(t)}\phi_5^{U(t)}]$, satisfies the equation

$$\begin{aligned}
\frac{\partial \mathcal{P}}{\partial t} &= \lambda(\phi_1 - 1)\mathcal{P} + \mu(1-\phi_1)\frac{\partial \mathcal{P}}{\partial \phi_1} \\
&\quad + [(\mu+\mu_1)(1-\phi_2) + \gamma_0(\phi_5 - \phi_2) + \theta_1(\phi_3 - \phi_2)]\frac{\partial \mathcal{P}}{\partial \phi_2} \\
&\quad + [\mu(1-\phi_3) + q_2\beta(\phi_2 - \phi_3) + (1-q_2)\beta(\phi_4 - \phi_3)]\frac{\partial \mathcal{P}}{\partial \phi_3} \\
&\quad + [(\mu+\mu_2)(1-\phi_4) + \delta(\phi_2 - \phi_4) + \delta_0(\phi_5 - \phi_4) + \theta_2(\phi_3 - \phi_4)]\frac{\partial \mathcal{P}}{\partial \phi_4} \qquad (7.7) \\
&\quad + [\mu(1-\phi_5) + \epsilon_1(\phi_2 - \phi_5) + \epsilon_2(\phi_4 - \phi_5)]\frac{\partial \mathcal{P}}{\partial \phi_5} \\
&\quad + \frac{\alpha}{n}\phi_2[-\phi_1 + pq_1\phi_2 + (1-p)\phi_3 + p(1-q_1)\phi_4]\frac{\partial^2 \mathcal{P}}{\partial \phi_1 \partial \phi_2} \\
&\quad + \frac{\alpha_2}{n}\phi_2[q_3\phi_2 + (1-q_3)\phi_4 - \phi_3]\frac{\partial^2 \mathcal{P}}{\partial \phi_2 \partial \phi_3},
\end{aligned}$$

with the initial condition $\mathcal{P}(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5; 0) = \phi_1^{x_0}\phi_2^{y_0}\phi_3^{z_0}\phi_4^{w_0}\phi_5^{u_0}$.

From equation (7.7) a system of differential equations for the first and second moments of X, Y, Z, W, and U is deduced; the equations for the means are

$$\begin{aligned}
\frac{d\mathrm{E}[X]}{dt} &= -\frac{\alpha}{n}\mathrm{E}[XY] - \mu\mathrm{E}[X] + \lambda \\
\frac{d\mathrm{E}[Y]}{dt} &= pq_1\frac{\alpha}{n}\mathrm{E}[XY] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZ] - \Gamma_s\mathrm{E}[Y] + q_2\beta\mathrm{E}[Z] + \delta\mathrm{E}[W] + \epsilon_1\mathrm{E}[U] \\
\frac{d\mathrm{E}[Z]}{dt} &= (1-p)\frac{\alpha}{n}\mathrm{E}[XY] - \frac{\alpha_2}{n}\mathrm{E}[YZ] + \theta_1\mathrm{E}[Y] - (\beta+\mu)\mathrm{E}[Z] + \theta_2\mathrm{E}[W] \\
\frac{d\mathrm{E}[W]}{dt} &= p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XY] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ] + (1-q_2)\beta\mathrm{E}[Z] \qquad (7.8) \\
&\quad - \Delta_s\mathrm{E}[W] + \epsilon_2\mathrm{E}[U] \\
\frac{d\mathrm{E}[U]}{dt} &= \gamma_0\mathrm{E}[Y] + \delta_0\mathrm{E}[W] - (\epsilon_1 + \epsilon_2 + \mu)\mathrm{E}[U],
\end{aligned}$$

where the Γ_s and Δ_s are as defined in (7.6) and the terms $\mathrm{E}[XY]$ and $\mathrm{E}[YZ]$ can be expressed as $\mathrm{E}[XY] = \mathrm{Cov}[X,Y] + \mathrm{E}[X]\mathrm{E}[Y]$ and $\mathrm{E}[YZ] = \mathrm{Cov}[Y,Z] + \mathrm{E}[Y]\mathrm{E}[Z]$. The equations for the variances and covariances are given in the Appendix (Section A.4.2).

From the system (7.8) it follows that the expected value of the total population size, $N(t) = X(t) + Y(t) + Z(t) + W(t) + U(t)$, satisfies the equation

$$\frac{d\mathrm{E}[N(t)]}{dt} = \lambda - \mu\mathrm{E}[N(t)] - \mu_1\mathrm{E}[Y(t)] - \mu_2\mathrm{E}[W(t)],$$

which with integration gives

$$\mathrm{E}[N(t)] = \frac{\lambda}{\mu} + e^{-\mu t}\left(n - \frac{\lambda}{\mu}\right) - e^{-\mu t}\int_0^t e^{\mu s}\{\mu_1\mathrm{E}[Y(s)] + \mu_2\mathrm{E}[W(s)]\}ds. \qquad (7.9)$$

This equation is the same as equation (6.20) for the mean total population size in the previous chapter.

7.3.2 The equilibrium state of the process

The process described in this chapter is a Markov process in continuous time with countable state space $\mathcal{S} = \mathbb{Z}_+^5$. Let $\mathcal{A}$ denote the subset of $\mathcal{S}$ that contains all the states of the form $(x, 0, 0, 0, 0)$ and $\mathcal{D}$ the remaining set of states:

$$\begin{aligned} \mathcal{A} &= \{(x,0,0,0,0) \in \mathbb{Z}_+^5\} \\ \mathcal{D} &= \mathcal{S} - \mathcal{A} = \{(x,y,z,w,u) \in \mathbb{Z}_+^5 : (y,z,w,u) \neq (0,0,0,0).\} \end{aligned}$$

The sets $\mathcal{A}$ and $\mathcal{D}$ form two irreducible classes. The former is closed and absorbing, while the latter is open and transient. Using Theorem 5.3 we will show that the chain will be absorbed in $\mathcal{A}$ with probability one.

Following the notation in Definition 5.2, the functions $a_j(\mathbf{x})$, $d_j(\mathbf{x})$, $e_{ij}(\mathbf{x})$, for $\mathbf{x} \in \mathcal{S}$, $i,j = 1, \dots, 5$ and $i \neq j$, are defined as follows: $a_1(\mathbf{x}) = \lambda$ and $a_j(\mathbf{x}) = 0$, for $j = 2,3,4,5$; $d_1(\mathbf{x}) = \mu x$, $d_2(\mathbf{x}) = (\mu + \mu_1)y$, $d_3(\mathbf{x}) = \mu z$, $d_4(\mathbf{x}) = (\mu + \mu_2)w$, and $d_5(\mathbf{x}) = \mu u$. The definitions of e_{ij} are shown in Table 7.2.

j	$e_{1j}(\mathbf{x})$	$e_{2j}(\mathbf{x})$	$e_{3j}(\mathbf{x})$	$e_{4j}(\mathbf{x})$	$e_{5j}(\mathbf{x})$
1	$-$	0	0	0	0
2	$pq_1\frac{\alpha}{n}xy$	$-$	$q_2\beta z + q_3\frac{\alpha_2}{n}yz$	δw	$\epsilon_1 u$
3	$(1-p)\frac{\alpha}{n}xy$	$\theta_1 y$	$-$	$\theta_2 w$	0
4	$p(1-q_1)\frac{\alpha}{n}xy$	0	$(1-q_2)\beta z + (1-q_3)\frac{\alpha_2}{n}yz$	$-$	$\epsilon_2 u$
5	0	$\gamma_0 y$	0	$\delta_0 w$	$-$

Table 7.2: The functions e_{ij} from Reuter's Theorem

The functions d_j and e_{ij} satisfy the conditions (5.17) for all $i,j = 1, \dots, 5$ and $i \neq j$. Also, by "freezing" the states $\mathbf{x} \in \mathcal{A}$, i.e. assuming that $a_1(\mathbf{x}) = d_1(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{A}$, all the states in $\mathcal{A}$ become absorbing and Theorem 5.3 can be applied. Define $\mathcal{A}_k = \{(x,y,z,w,u) \in \mathcal{D} : x+y+z+w+u = k\}$, for $k = 1, 2, \dots$. Then it follows that

$$\begin{aligned} r_k &= \max_{\mathbf{x} \in \mathcal{A}_k} \sum_{i=1}^{5} a_i(\mathbf{x}) = \lambda \\ s_k &= \min_{\mathbf{x} \in \mathcal{A}_k} \sum_{i=1}^{5} d_i(\mathbf{x}) = \mu k, \end{aligned}$$

for all $k = 1, 2, \dots$. The functions r_k and s_k are the same as the r_k and s_k for the model in the previous chapter (see equations (6.21) and (6.22)) and hence Theorem 5.3 gives the same results, which can be summarised as follows (see also Section 6.3.2):

- The process will be absorbed in $\mathcal{A}$ with probability 1, so that extinction of the infection is certain: $\sum_{\mathbf{y}\in\mathcal{A}} \pi(\mathbf{y}) = 1$.
- The mean time until extinction is finite.
- After extinction of the infection, the limiting distribution of the uninfected individuals is Poisson with parameter λ/μ and hence

$$\lim_{t\to\infty} \mathrm{P}[\mathbf{X}(t) = \mathbf{y}] = \begin{cases} 0 & \text{if } \mathbf{y} \in \mathcal{D} \\ \dfrac{e^{\lambda/\mu}(\lambda/\mu)^k}{k!} & \text{if } \mathbf{y} = (k, 0, 0, 0, 0) \in \mathcal{A}. \end{cases}$$

7.4 Epidemiology

7.4.1 Definitions

In this section we will study some epidemiological indices which are helpful in assessing the severity of an epidemic and hence the effect of chemotherapy in the spread (and/or the endemicity) of TB in a population. The epidemiological indices to be studied are the risk of infection and reinfection, incidence, prevalence, and mortality, as defined in Definition 6.1.

For this model, new infectious cases developing during a certain year are all the transitions from the class of uninfected (X) and the class of inactive cases (Z) to that of infectious cases (Y) during that year. From an epidemiological point of view the number of recovered (U) or non-infectious cases (W) who become infectious should not be included in this index (see, e.g., Styblo 1991), but the number of cured (after treatment) who become infectious are included, since they are in the same class (Z) as the latents. Similarly, new non-infectious cases developing during a certain year are all the transitions from the class of uninfected (X) and the class of inactive cases (Z) to that of non-infectious cases (W) during that year.

Throughout this chapter the epidemiological indices are presented as proportions per $100,000$ general population. The indices presented in this section were calculated from simulations of the stochastic model shown in Figure 7.1. Details of the implementation of the simulations can be found in the Appendix (Section A.4.3).

The parameters were chosen to have values that are representative for TB and are shown in Table 6.8; they are the same as the values used in the respective Section 6.3.4 for the epidemiological indices in the previous chapter. The values of θ_1 and θ_2 depend on the particular control program implemented. In this section we present results for various values of θ_1 and θ_2, in order to study the relationship of the effect of chemotherapy with the detection and cure rates. For all the results presented in this section we give only the total value of θ_1 and not separately the values of θ_{1d} and θ_{1c}, since in the simulations these two parameters appear only in the product $\theta_{1d}\theta_{1c} = \theta_1$; hence the results for a specific value of θ_1 can be translated as results for all the combinations of values of θ_{1d} and θ_{1d} whose product is equal to that value of θ_1. Also, we assume that $\theta_{2c} = \theta_{1c}$ and $\theta_{2d} = \theta_{2r}\theta_{1d}$, so that $\theta_2 = \theta_{2r}\theta_1$.

The initial conditions were taken as the equilibrium values from model Zeus. Model Zeus describes the natural evolution of TB and is the same as model Clio with $\theta_1 = \theta_2 = 0$. As was explained in Section 7.1, the epidemic begins with the introduction of the infection into a population of size n (for the results presented in this section n was taken equal to 1000). The infection spreads and finally settles down at an endemic stable equilibrium (the quasi-stationary distribution of model Zeus; see Sections 6.3.3 and 6.3.6). We assume that chemotherapy is introduced after the process has reached this steady endemic level. Therefore the most natural way to simulate this process is to simulate model Zeus (or equivalently model Clio with $\theta_1 = \theta_2 = 0$) until the process reaches the endemic stable state and then simulate model Clio (with the particular values of θ_1 and θ_2 desired) taking the initial value of the vector $(X,\ Y,\ Z,\ W,\ U)$ to be equal to the equilibrium value of $(X,\ Y,\ Z,\ W,\ U)$ deduced from the simulations of model Zeus. Details of how the simulations were implemented can be found in the Appendix (Section A.4.3).

In the following two sections we present results for the percentage decline in the epidemiological indices. The value of each index, say $\mathcal{F}(t)$, t years after the introduction of chemotherapy, was calculated from the simulations of model Clio (with initial conditions as explained above). Then the decline, t years after the introduction of chemotherapy, was calculated as $\mathcal{F}_e - \mathcal{F}(t)$, where $\mathcal{F}_e$ is the value of the respective index at the equilibrium level for model Zeus.

7.4.2 The first 30 years after the introduction of chemotherapy

First we will study the effects of chemotherapy during the first 30 years after its introduction. The simulations were carried out up to time $t = 30$ for $\theta_1 = 15, 20, \ldots, 55, 60\%$, $\theta_{2r} = 0, 0.5, 0.7$ and $\theta_2 = \theta_{2r}\theta_1$. It has to be stressed that the results with $\theta_{2r} > 0$

show the effect of treating the non-infectious TB cases, as $\theta_{2r} = 0$ means that the non-infectious are given no treatment at all. The percentage decrease was calculated for all the epidemiological indices listed in Definition 6.1. Figure 7.3 shows the results for the total incidence and the prevalence of TB infection at 10, 20, and 30 years and Figure 7.4 for the prevalence of infectious and non-infectious TB, mortality, and risk of infection at 10 years. The other results were qualitatively similar and are not presented here.

The first point that should be noticed from Figures 7.3 and 7.4 is that for most of the epidemiological indices under study, the value of θ_{2r} does not substantially affect their decline. Especially during the first 10 years the results from the three different values of θ_{2r} are almost the same. As time increases the difference between the three curves (for the three different values of θ_{2r}) increases, but still it remains very small even after 30 years. The only cases were the value of θ_{2r} made a significant difference were for the prevalence and the mortality of non-infectious TB and the total mortality.

This result is quite reasonable since $\theta_2 = \theta_{2r}\theta_1$ and θ_2 is the rate at which the non-infectious cases are removed from the W class (to the inactive class, Z, after successful treatment). Therefore even a small increase in θ_2 will decrease the number of non-infectious cases, W, and hence the prevalence and mortality of W (and the total mortality, which is the sum of the mortality of Y and W). On the other hand, all the other epidemiological indices depend on the value of W only indirectly, because they depend mainly on the value of infectious cases, Y (which of course is the key element that drives the progress of the epidemic). Hence, the chain of reactions that we would expect is that an increase in θ_2 will decrease W and that will decrease Y (because of the transitions from W to Y), which in turn will decrease the other epidemiological indices. Since the value of δ (the rate of transitions from W to Y) is quite small, the epidemiological indices depend on W only slightly and indirectly and therefore they will not be affected that much by a small increase in θ_2.

In contrast, the value of θ_1 has more effect on the epidemiological indices. Even after 10 years, increasing θ_1 from 0.15 to 0.60 increases the decline in most epidemiological indices by 10–15%. As was explained above this is a result of the fact that the main force driving the epidemic is the number of infectious cases, Y. Therefore with an increase in the value of θ_1, the value of Y decreases, thus decreasing the prevalence of Y and hence the number of infections (counted by the risk of infection and reinfection), the

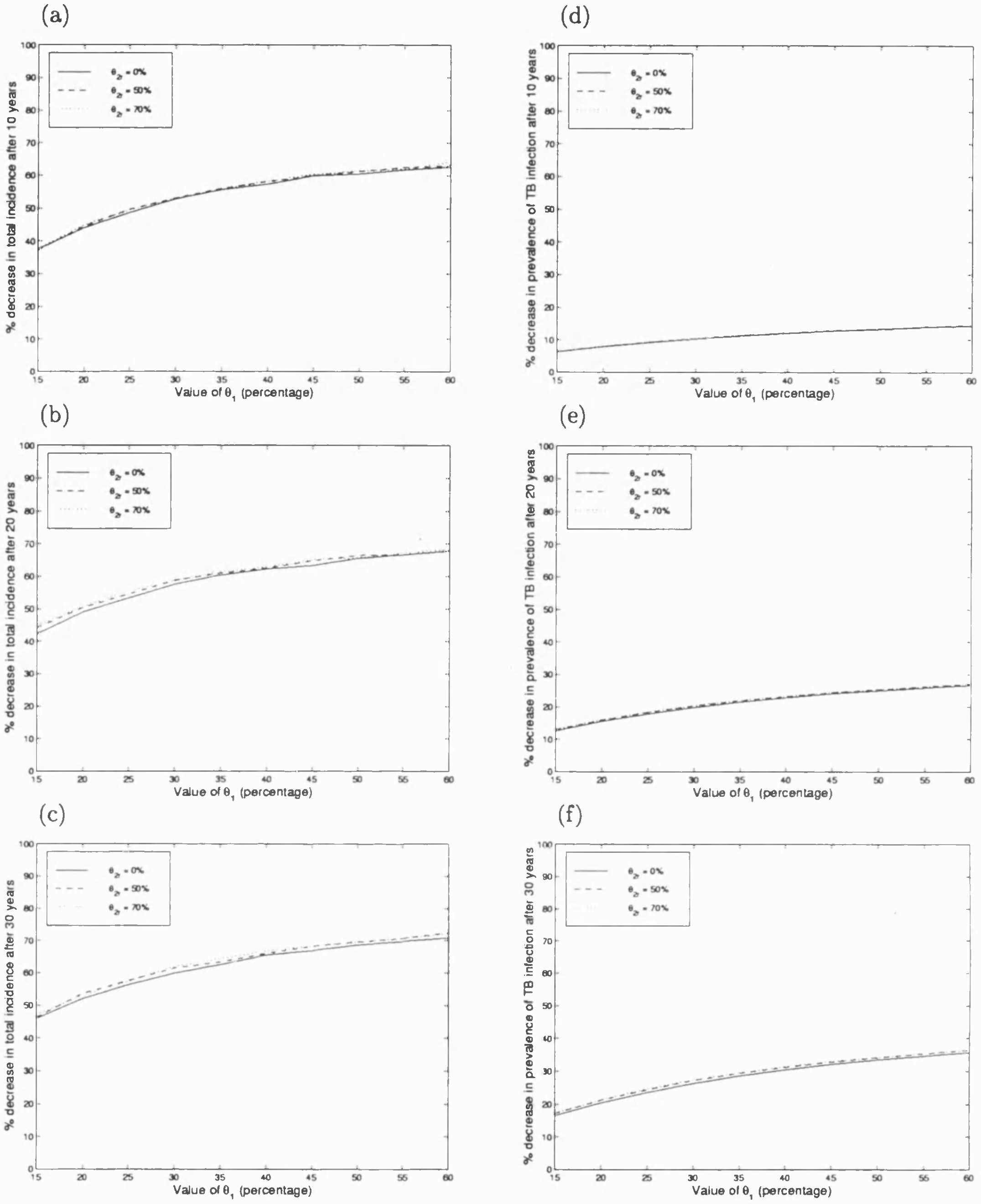


Figure 7.3: (a), (b), (c): Percentage decline in the total incidence of TB as a function of θ_1. The three graphs are for the decline 10, 20, and 30 years, respectively, after the introduction of chemotherapy. (d), (e), (f): Percentage decline in the prevalence of TB infection as a function of θ_1. The three graphs are for the decline 10, 20, and 30 years, respectively, after the introduction of chemotherapy. All the rates were calculated as proportions per 10^5 general population. In each graph there are three curves, one for each of the following values of θ_{2r}: 0, 0.5, and 0.7. The other parameter values used are shown in Table 6.8. The initial conditions are taken from the endemic steady level of the natural evolution of TB (see Section 7.4.1 for details). Time is measured in years after the introduction of chemotherapy. The data used for these graphs were available only for the discrete points $\theta_1 = 15, 20, \ldots, 55, 60\%$ and not for all the values of θ_1 between 15% and 60%.

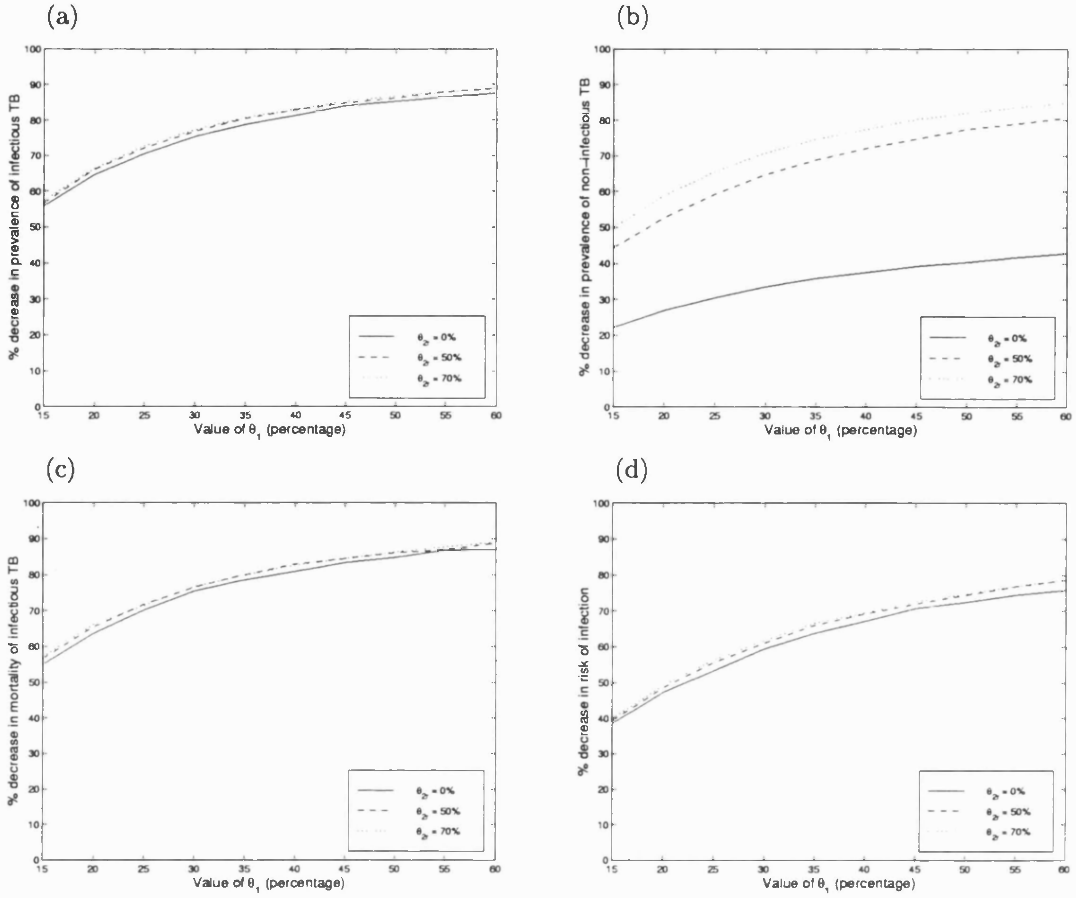


Figure 7.4: Percentage decline in (a) prevalence of infectious TB, (b) prevalence of non-infectious TB, (c) mortality of infectious TB, and (d) risk of infection, as a function of θ_1 (10 years after the introduction of chemotherapy). All the rates were calculated as proportions per 10^5 general population. The parameter values used are shown in Table 6.8. The initial conditions are taken from the endemic steady level of the natural evolution of TB (see Section 7.4.1 for details). The data used for these graphs were available only for the discrete points $\theta_1 = 15, 20, \ldots, 55, 60\%$ and not for all the values of θ_1 between 15% and 60%.

number of new cases developing (incidence) and so on. The difference is bigger for small values of θ_1, while for θ_1 greater than 0.45 the slope of the curves shown in Figures 7.3 and 7.4 decreases.

Also it should be noticed that even for the smallest values of θ_1 and θ_{2r} examined here and even after only 10 years, there was a significant decrease (about 35% to 55% in most cases) in all the epidemiological indices, except for the prevalence of TB infection and the prevalence of the Z class (not shown here). For these two indices the decline was rather small and only for the largest values of θ_1 and θ_{2r} did the decline reach a level of about 30% after 30 years.

This is probably a result of the fact that in the beginning there is still a large number of infectious, Y, and non-infectious, W, cases (from the endemic steady state of the natural evolution, before the introduction of chemotherapy). After successful treatment these move to the inactive class, Z. Therefore there is an influx into the Z class, which means that the number of inactive cases does not decrease significantly. On the other hand, the size of the Z class does not substantially increase, because the influx from the infectious and non-infectious classes into the inactive class increases the outflow from the inactive class due to endogenous reactivation and exogenous reinfection (at rates βZ and $\alpha_2 YZ/n$, respectively). Therefore the prevalence of Z decreases only slightly. On the other hand the prevalences of infectious and non-infectious TB decline significantly because of the outflow due to successful treatment and hence the other indices decrease as well, as a result of the decreased number of infectious cases. Overall, the prevalence of TB infection (which is the sum of prevalences of the Y, Z, W, and U classes) does not decrease significantly because it is the sizes of each individual class that are affected in the beginning (for instance by "moving" individuals from the Y to the Z class), but the overall size of $Y + Z + W + U$ does not change that much. It takes more time to see the effect of chemotherapy in the size of $Y + Z + W + U$ and hence the prevalence of TB infection as well, as the results in the next section will show.

7.4.3 The long run behaviour

Now we will study the long run behaviour of the process. The simulations were carried out up to time $t = 300$. From a practical point of view the results for such a long period are of no interest, since it is unreasonable to assume that the values of the parameters will remain the same for 300 years. Nevertheless, the information about how the epidemic could evolve (if things remained the same with respect to possible transitions and the values of the parameters) can be helpful in planning long-term policies or as an indication of the possible long-run effects of chemotherapy.

The simulations were carried out with $\theta_1 = 20, 40, 60\%$, $\theta_{2r} = 0, 0.5, 0.7$ and $\theta_2 = \theta_{2r}\theta_1$. The percentage decline was calculated for all the epidemiological indices listed in Definition 6.1. Figure 7.5 shows the results for the total incidence and the prevalence of non-infectious TB for $\theta_{2r} = 0, 0.5, 0.7$ and Figure 7.6 for the prevalence of infectious TB and TB infection and the risks of infection and reinfection for $\theta_{2r} = 0.5$. The other results were qualitatively similar and are not presented here.

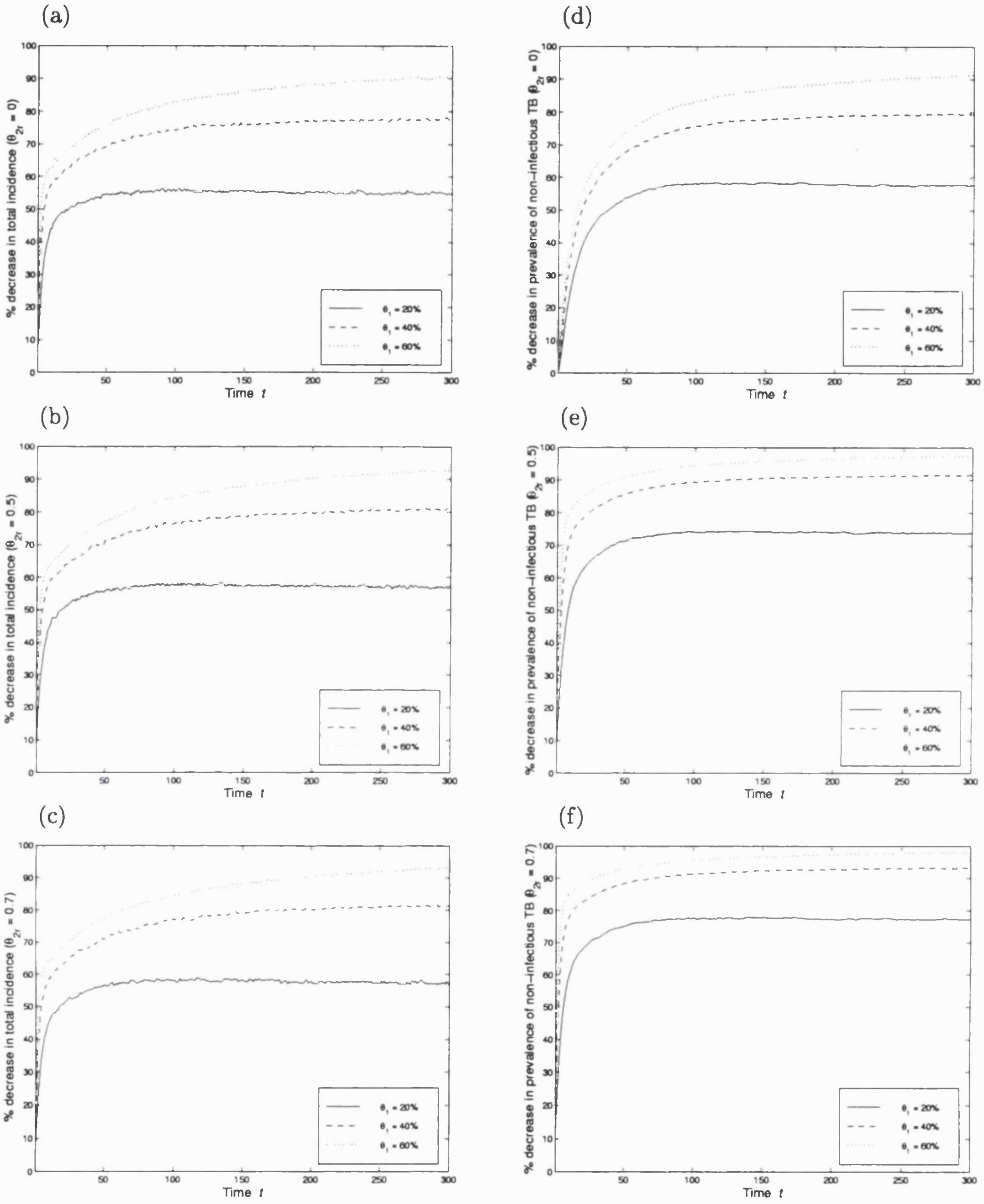


Figure 7.5: (a), (b), (c): Percentage decline in the total incidence of TB as a function of time. The three graphs are for the following three values of θ_{2r}: 0, 0.5, and 0.7. (d), (e), (f): Percentage decline in the prevalence of non-infectious TB as a function time. The three graphs are for the following three values of θ_{2r}: 0, 0.5, and 0.7. All the rates were calculated as proportions per 10^5 general population. In each graph there are three curves, one for each of the following values of θ_1: 20%, 40%, and 60%. The other parameter values used are shown in Table 6.8. The initial conditions are taken from the endemic steady level of the natural evolution of TB (see Section 7.4.1 for details). Time t is measured in years after the introduction of chemotherapy.

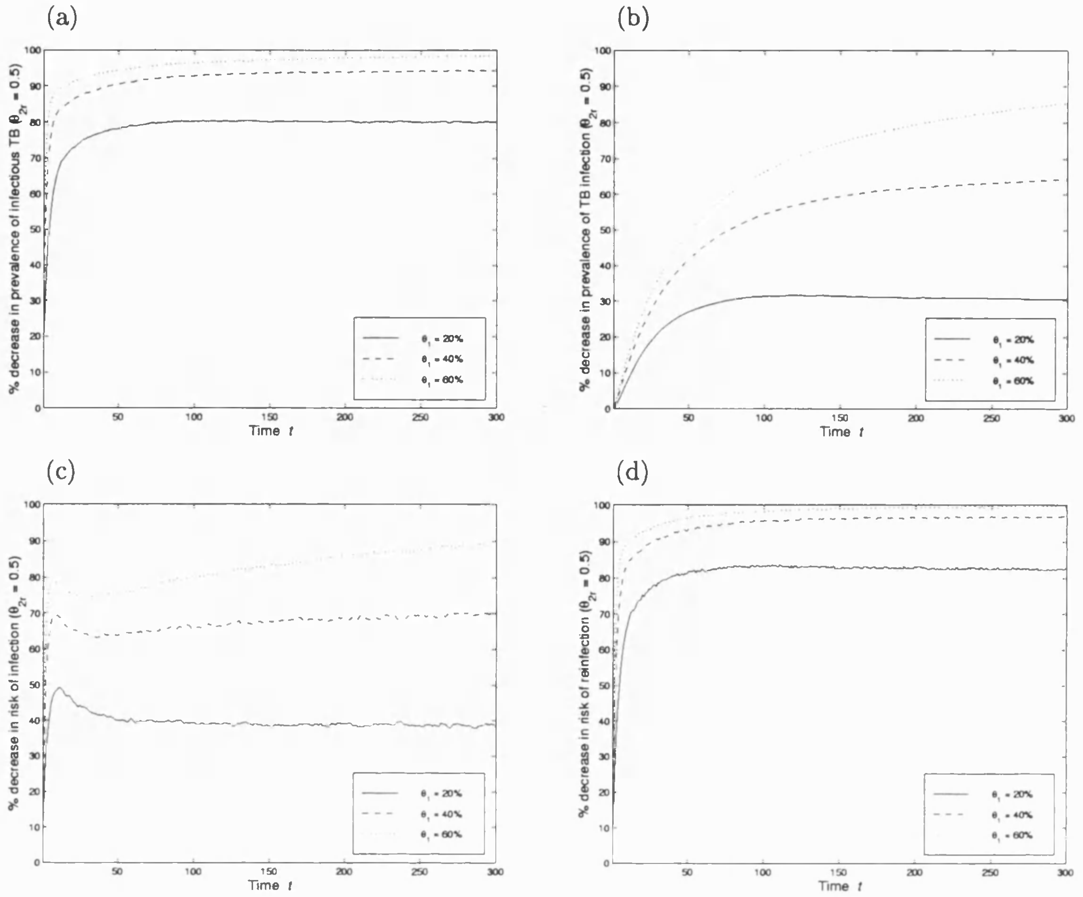


Figure 7.6: Percentage decline in (a) prevalence of infectious TB, (b) prevalence of TB infection, (c) risk of infection, and (d) risk of reinfection as a function of time, with $\theta_{2r} = 0.5$. All the rates were calculated as proportions per 10^5 general population. The parameter values used are shown in Table 6.8. The initial conditions are taken from the endemic steady level of the natural evolution of TB (see Section 7.4.1 for details). Time t is measured in years after the introduction of chemotherapy.

As for the results in the previous section, the decline in the epidemiological indices does not vary significantly depending on the value of θ_{2r}. The difference between the curves for θ_{2r} equal to 0, 0.5, and 0.7 is quite small, although it increases slightly after the first 40–50 years (at time $t = 300$ the difference is 5–10% for the various indices). The only exception is for the prevalence and the mortality of non-infectious cases, where the difference reaches its maximum during the first 30–40 years and then declines. The effect of changing the value of θ_{2r} during the initial phase was explained in the previous section: an increase in θ_{2r} will decrease the size of W; hence the prevalence and mortality of non-infectious cases will decline significantly (since these indices directly depend on the size of W), but the other indices will change only slightly (since they depend on W only indirectly). In the long run though, this effect of changing θ_{2r} wanes, as it affects mainly the size of W. The main force that drives the epidemic is the size of Y; since this is not altered significantly by changing the value of θ_{2r}, eventually the size of W and hence the prevalence and mortality of non-infectious TB will not be affected considerably.

Some other interesting elements of the behaviour of the decline in the epidemiological indices can be grouped in the following three classes:

Mortality and prevalence of infectious cases, risk of reinfection: The slope of the curves is very high in the beginning and the percentage decline rapidly increases during the first 20–30 years, reaching a level of more than 70%. After that it increases more slowly and at time $t = 300$ the percentage decline is more than 95% for $\theta_1 = 60\%$ and around 75–80% for $\theta_1 = 20\%$. The difference between the curves depending on the value of θ_1 is much smaller than for the other indices and increases slightly in time.

Total mortality and incidence, incidence of infectious TB, prevalence of non-infectious TB: The slope of the curves is still quite high in the beginning although lower than for the indices mentioned above. The percentage decline increases quite rapidly during the first 20–30 years, reaching a level around 50–80%. After that it increases more slowly and at time $t = 300$ it is about 90–95% for $\theta_1 = 60\%$ and around 55–70% for $\theta_1 = 20\%$. The difference between the curves depending on the value of θ_1 is more considerable than for the indices mentioned above, but increases slightly in time.

Prevalence of TB infection and of the class Z : The slope of the curves in the beginning is much lower than for the other indices; the percentage decline increases much more slowly and smoothly during the first 50 years, and for $\theta_1 = 60\%$ it continues to increase significantly up to $t = 100$. The difference between the curves for the different values of θ_1 is very considerable here and increases a lot in time: the percentage decline in these indices at $t = 50$ is about 40–45% for $\theta_1 = 60\%$ and about 20–25% for $\theta_1 = 20\%$, while at $t = 300$ it is around 80% for $\theta_1 = 60\%$ and around 25–30% for $\theta_1 = 20\%$.

Finally, for the risk of infection the behaviour is similar to the one for the second class described above, the only difference being that in the beginning the curves have a peak (around $t = 20$). The percentage decline then decreases up to $t = 50$ and then remains almost steady for $\theta_1 = 20, 40\%$ and slightly increases for $\theta_1 = 60\%$.

The behaviour described above can be explained by the effect of the class of infectious cases. The epidemiological indices that depend more heavily and/or directly on the size of the Y class are the ones which are more affected by the different values of θ_1 and have the highest and more rapid decrease. Therefore the mortality and prevalence of infectious cases decreases more, and more rapidly, than the prevalence of the other classes and the mortality of non-infectious cases. Also, they decrease more than the incidence since this latter index depends also on the sizes of X and Z, and only indirectly on the size of Y (the infections and reinfections that develop into cases). The prevalence of TB infection and that of the class Z present the smallest and slowest decrease for the reasons explained in the previous section for the first years after the introduction of chemotherapy: the influx from the classes of infectious and non-infectious cases (due to successful treatment) makes the decrease in these indices slower.

The peak observed in the decline of the risk of infection can be explained by the interaction between the classes X and Y. After the introduction of chemotherapy the size of Y rapidly decreases (as the prevalence of infectious TB shows) thus decreasing the number of new infections quite significantly. That results in an increase in the size of the X class (results not shown here). Therefore, although there is a smaller number of infectious cases to transmit the disease, there is a larger pool of uninfected individuals to get infected, and hence their product (XY) can increase.

One final observation that should be made with respect to the effect of the value of θ_1 is that for $\theta_1 = 20\%$ the epidemiological indices seem to stabilise after 100 years, while for $\theta_1 = 40\%$ this point is reached somewhat later. For $\theta_1 = 60\%$ though they continue to decrease slightly even up to time $t = 300$, indicating that the process has not reached the steady endemic level yet.

7.4.4 Variation in the epidemiological indices

Finally we conclude the study of the decline in the epidemiological indices with a few comments about their variation. Figure 7.7 shows the standard deviation of the prevalence of infectious and non-infectious TB 10 years after the introduction of chemotherapy and during the first 300 years after its introduction (details of how the standard deviations were calculated can be found in the Appendix, Section A.4.3). For the other indices the results were qualitatively similar and are not shown here.

The value of θ_1 affects the level of variation substantially. As θ_1 increases the standard deviation decreases, so for the largest values of θ_1 the results presented in the

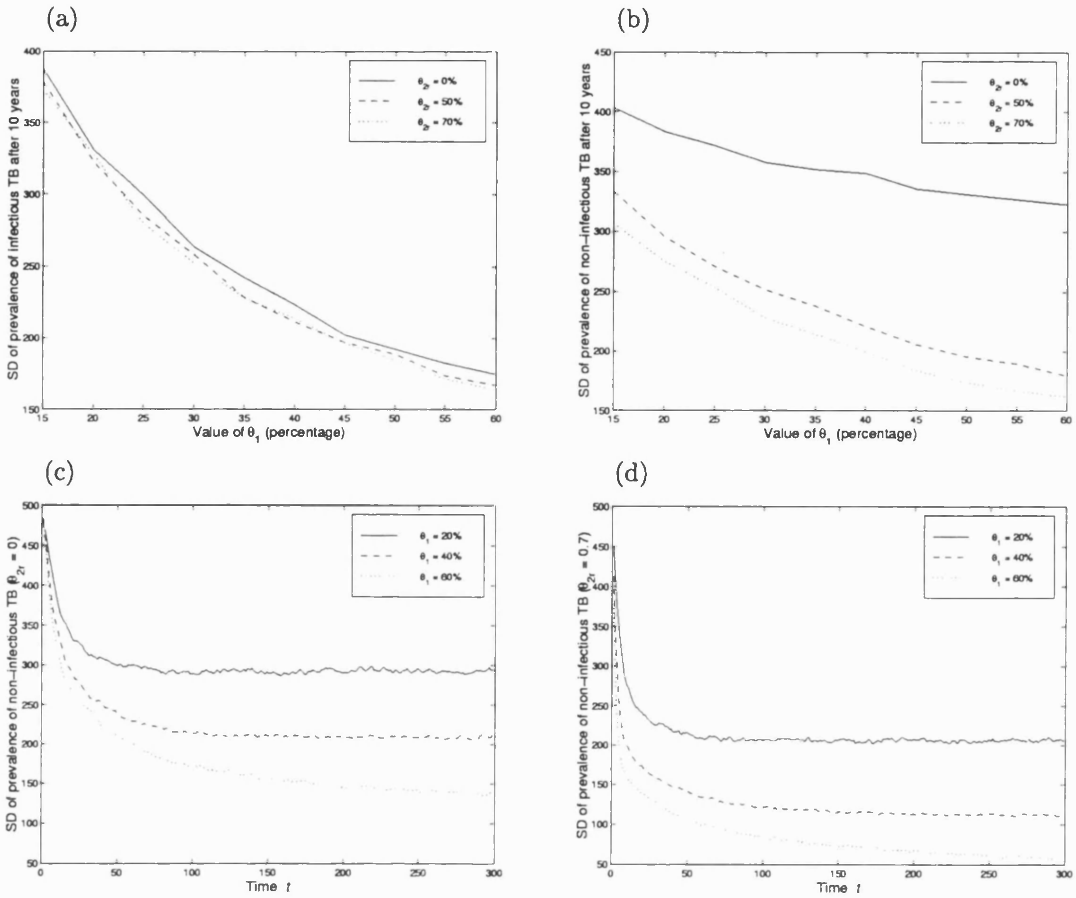


Figure 7.7: (a), (b) Standard deviation of the prevalence of infectious and non-infectious TB, respectively, 10 years after the introduction of chemotherapy. The data used for these graphs were available only for the discrete points $\theta_1 = 15, 20, \ldots, 55, 60\%$ and not for all the values of θ_1 between 15% and 60%. (c), (d) Standard deviation of the prevalence of non-infectious TB as a function of time with $\theta_{2r} = 0, 0.7$, respectively. In all cases the prevalences were calculated as proportions per 10^5 general population. The parameter values used are shown in Table 6.8. The initial conditions are taken from the endemic steady level of the natural evolution of TB (see Section 7.4.1 for details). Time is measured in years after the introduction of chemotherapy.

previous sections are more accurate (in the sense that there is less variability in the results obtained for the estimates of the epidemiological indices). In contrast, the value of θ_{2r} does not affect the variability that much, and the standard deviations decrease only slightly as θ_{2r} increases.

A clearer picture about the decline in the epidemiological indices and the comparisons for the relative effect of the various levels of θ_1 and θ_{2r} can be obtained by Figure 7.8. In this figure we present results for 95% confidence intervals for the percentage decline in the prevalence and the incidence of infectious TB 10 years after the introduction of chemotherapy and for θ_{2r} equal to 0 and 0.7 and $\theta_1 = 15, 20, \ldots, 55, 60$. It is obvious from these graphs that the value of θ_1 significantly affects the decline, espe-

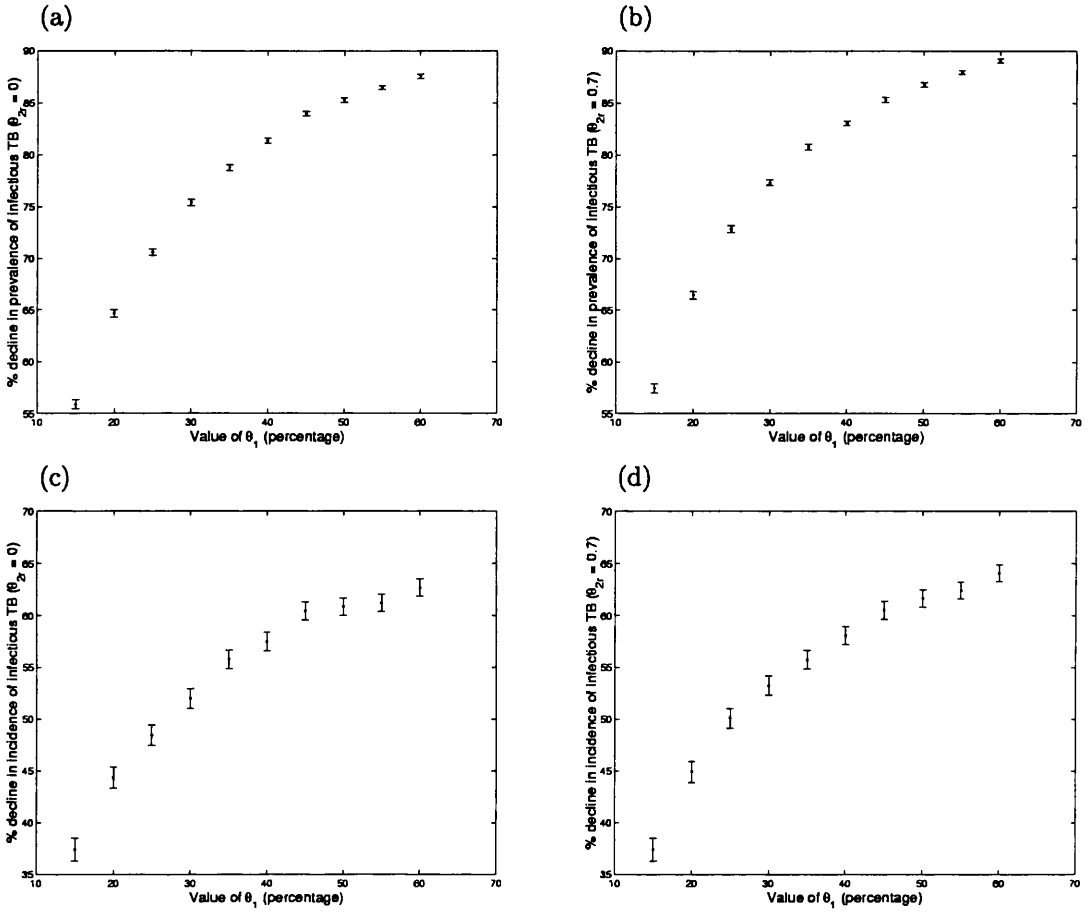


Figure 7.8: 95% confidence intervals for the percentage decline in (a), (b) the prevalence of infectious TB, with $\theta_{2r} = 0, 0.7$, respectively, and (c), (d) the incidence of infectious TB, with $\theta_{2r} = 0, 0.7$, respectively. The indices were calculated as proportions per 10^5 general population, 10 years after the introduction of chemotherapy. The parameter values used are shown in Table 6.8 and $\theta_1 = 15, 20, \ldots, 55, 60\%$. For each value of θ_1 the dot in the figures above presents the percentage decline in the prevalence and incidence, and the length of the line segment above and below the dot is equal to the length of the 95% confidence interval. The initial conditions are taken from the endemic steady level of the natural evolution of TB (see Section 7.4.1 for details).

cially for small values of θ_1 (up to 35–40%), while varying θ_{2r} even from 0 to 0.7 results in only minor differences in the decline.

7.4.5 Conclusions

The results presented in the previous sections suggest that the level of the recovery rate (after successful treatment) for the non-infectious cases, θ_2, does not substantially affect the epidemiological indices. Surprisingly, even with $\theta_2 = 0$ the decline in the epidemiological indices is not substantially smaller than the decline with positive values of θ_2, suggesting that treatment of non-infectious cases may not significantly contribute in the reduction of the TB problem.

On the other hand, the level of the recovery rate for the infectious cases, θ_1, has a significant impact on the decline of the epidemic. For instance by varying θ_1 from 0.4 to 0.6 the difference in the decline of the epidemiological indices can be 5–10% in the short term and 10–30% in the long run. The level of decline depends on the epidemiological index studied, but even with a moderate value of $\theta_1 = 0.4$ the decrease in incidence, mortality, and prevalence of infectious TB is more than 50% 10 years after the introduction of chemotherapy. For the prevalence of the infection, however, the decrease is only about 10%.

In interpreting these results and determining the best policy to be implemented there are several things that have to be taken into account. First of all there is the fact that TB control policies have long-run effects, in the sense that it may take some time until substantial effects can be observed, but also that the reduction achieved can be sustained for a long time, even after terminating the control policy (Enarson & Rouillon 1998; see also Azuma 1975 for modelling the decline after termination of the controls).

On the other hand, low cure rates may also have the exactly opposite effect. In the short term they decrease the number of infectious cases and hence the number of new infections, number of deaths and so on. In particular the mortality rates may be reduced even in the long run, since treatment always has some effect on the patient, reducing the number of tubercle bacilli, for instance, thus reducing the mortality for these individuals. But low cure rates result in high treatment failures, persons who relapse and become infectious again later, and therefore increase the rate of transmission of the disease (see, e.g., Dye et al. 1998). In other words, with low cure rates most of the infectious cases treated will be cured only temporarily or partially and hence they will remain alive and infectious for longer time thus prolonging their infectious periods. Therefore, although the mortality rate may be decreased, the risk of infection may not be substantially reduced in the long run, which is also one of the reasons why the mortality rates are no longer useful epidemiological indices of the severity of the epidemic (see, e.g., Styblo 1991). The results presented in the previous sections agree with this point since the decline in the risk of infection is smaller than the decline in mortality and also the difference in the decline in the various indices, depending on the value of θ_1, increased in time.

Another point that has to be taken into account with respect to the treatment of non-infectious cases is the fact that treating these patients helps in reducing the transmission of TB before diagnosis (for those non-infectious cases who finally convert to infectious TB), which accounts for a significant proportion of the total transmission (Murray et al. 1991). Therefore treating non-infectious cases has a direct effect on the reduction of the rate of transmission, which cannot be achieved by treating only infectious cases. The only way of compensating for this effect is by increasing the detection rate θ_{1d} so that the non-infectious who become infectious will be detected early enough and hence pre-diagnosis transmission will not be significant.

Finally it has to be mentioned that the effectiveness of any control program also depends on cultural and behavioural indices in the country/area where it is implemented. Nevertheless, it has been shown that despite these differences, high cure rates can be achieved, for instance 80–90% in Malawi, Mozambique, and Tanzania (Murray et al. 1991) and 85–90% in China (China Tuberculosis Control Collaboration 1996). If these high cure rates are combined with high detection rates, then the overall successful treatment rate will be high, thus contributing in the reduction of the TB problem. The importance of an effective case-finding system has been extensively emphasised in the literature, especially after the publication of results that document high successful treatment rates with intensive control programmes in countries where previous (not intensive) programmes had been consistently unsuccessful (see, e.g., Styblo 1983, China Tuberculosis Control Collaboration 1996, Murray et al. 1993). This is the reason why case-finding and treatment are no longer seen as separate or inequivalent elements of a TB control program, but have to be targeted as one entity in order to achieve the best possible results. The WHO targets for case-finding and cure are 70% and 80–85%, respectively, giving a value of θ_1 around 0.6 (which is the maximum value of θ_1 studied in the previous sections).

Chapter 8

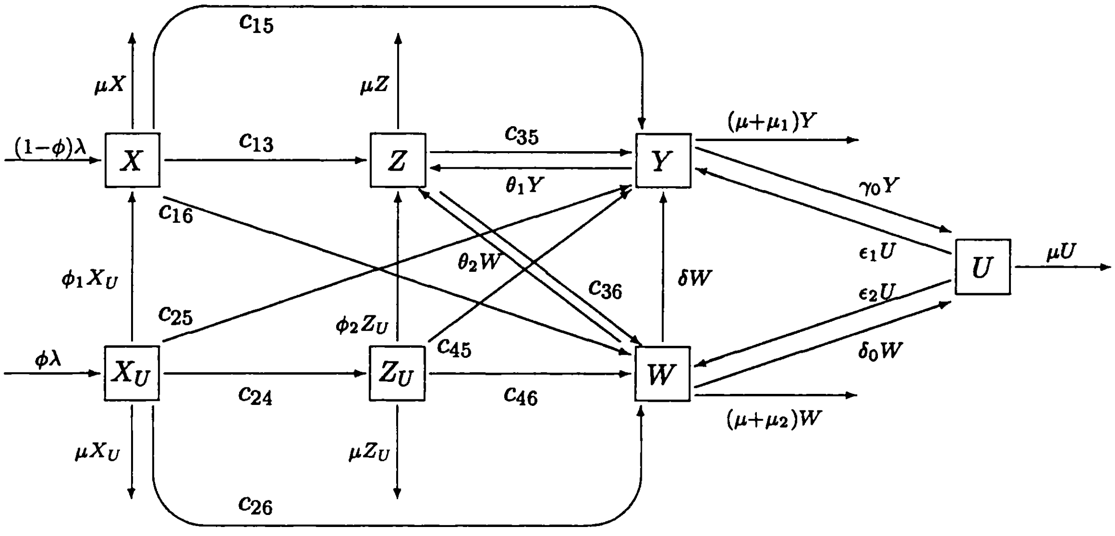
Model Erato: a model for the BCG vaccine

8.1 Introduction

The Bacille Calmette-Guérin (BCG) vaccine for tuberculosis is one of the most widely used, but also one of the most controversial vaccines (Fine 1995). Its effectiveness varies from zero to 80% and for this reason it has never been routinely recommended in some countries, except for some specific classes of the population (Murray et al. 1993, LaScolea & Rangoonwala 1996). BCG offers protection against development of disease but not against infection (or reinfection) and even that is for a limited length of time (about 10–15 years) (Huebner 1996, Murray et al. 1993). It is recommended that the vaccine should be given as early in life as possible, but the effect of revaccination or vaccination at older ages is unknown (Smith & Fine 1998, Murray et al. 1993). Since its protection lasts only for 10–15 years, the vaccinated individual is protected mainly during childhood and adolescence. During these periods of life, though, the infectious forms of TB are rare, and this limits the effect of BCG as a control method.

Because of these limitations in the effectiveness of BCG, even total BCG coverage of the population will have little effect on the annual risk of infection and hence in controlling tuberculosis (Murray et al. 1993). Consequently no control program can depend solely on BCG, but when a mass-vaccination program is implemented along with an intense chemotherapy program, it can help in controlling the TB problem. Therefore in this chapter the effect of BCG will be examined by comparing the effect of a control program implementing both BCG and chemotherapy with a program implementing only chemotherapy. To this end, the model for chemotherapy (as presented in the previous chapter) is extended to account for the fact that some members of the population will be vaccinated.

The definitions of the variables and parameters used for the model for chemotherapy (model Clio) are summarised in Table 7.1. The additional variables and parameters used for the model for BCG are shown in Table 8.1. The possible transitions and their rates are illustrated in Figure 8.1.



$$c_{13}(X,Y) = (1-p)\tfrac{\alpha}{n}XY \qquad c_{24}(X_v,Y) = (1-p')\tfrac{\alpha}{n}X_vY$$
$$c_{15}(X,Y) = pq_1\tfrac{\alpha}{n}XY \qquad c_{25}(X_v,Y) = p'q_1\tfrac{\alpha}{n}X_vY$$
$$c_{16}(X,Y) = p(1-q_1)\tfrac{\alpha}{n}XY \qquad c_{26}(X_v,Y) = p'(1-q_1)\tfrac{\alpha}{n}X_vY$$
$$c_{35}(Y,Z) = q_2\beta Z + q_3\tfrac{\alpha_2}{n}YZ \qquad c_{45}(Y,Z_v) = q_2\beta' Z_v + q_3\tfrac{\alpha_2'}{n}YZ_v$$
$$c_{36}(Y,Z) = (1-q_2)\beta Z + (1-q_3)\tfrac{\alpha_2}{n}YZ \qquad c_{46}(Y,Z_v) = (1-q_2)\beta' Z_v + (1-q_3)\tfrac{\alpha_2'}{n}YZ_v$$

Figure 8.1: Model Erato: a model for chemotherapy and BCG vaccine

The population is divided into 7 classes, X, X_v, Z, Z_v, Y, W, U. The classes X, Z, Y, W, and U and the possible transitions between them are defined exactly as in model Clio (see Section 7.1). X_v are vaccinated individuals who have never been infected and Z_v are vaccinated individuals who have been infected but are inactive (they will be referred to as uninfected vaccinated and inactive vaccinated classes, respectively). We assume that all vaccinations are given soon after birth. If ϕ is the proportion of newborns

$X_v(t)$	Number of uninfected vaccinated individuals at time t
$Z_v(t)$	Number of vaccinated inactive cases at time t
ϕ	Proportion of newborns who are vaccinated
ϕ_1^{-1}	Average length of protection from BCG, for uninfected individuals
ϕ_2^{-1}	Average length of protection from BCG, for infected individuals
v_1	Reduction (due to BCG) in the probability p of developing primary TB (within a year after infection)
v_2	Reduction (due to BCG) in the probability β of developing secondary TB (endogenous reactivation)
p'	$v_1 p$ (probability of developing primary TB for vaccinated individuals)
β'	$v_2 \beta$ (reactivation rate for vaccinated individuals)
α_2'	$v_1 \alpha_2$

Table 8.1: Additional variables and parameters, which are used in model Erato over and above those defined in Table 7.1 for model Clio

who are vaccinated, then the immigration rate into the uninfected vaccinated class X_v is $\phi\lambda$ and the immigration rate into the uninfected class X is $(1 - \phi)\lambda$.

The protection from BCG lasts only for a few years, so we assume that vaccinated individuals who have not been infected (X_v) move to the class X at a rate $\phi_1 X_v$ and hence they are not protected anymore. Similarly, inactive cases who have been vaccinated but not developed TB yet (Z_v) move to the class Z at a rate $\phi_2 Z_v$ (with $\phi_2 \geq \phi_1$) and they are not protected anymore from the vaccine.

Since BCG does not protect against infection, new infections of vaccinated individuals (X_v) occur at the same rate as for the non-vaccinated ($\alpha Y/n$ per capita). The vaccine does, however, protect against the development of disease and hence the probability p' that a newly infected individual will develop TB soon after infection is decreased: $p' = v_1 p$, with $0 < v_1 < 1$. There is no evidence that BCG offers more protection against infectious forms of TB than against the non-infectious ones (Rouillon & Waaler 1976), so we assume that the probability q_1 that a newly infected individual will develop infectious TB is the same regardless of whether the individual is vaccinated or not.

Similarly for the reinfections, the effective contact rate between vaccinated inactive cases and infectious cases is the same ($p_r\alpha$) as for the non-vaccinated inactive cases. After reinfection, an individual develops clinical disease within a year (primary TB) or remains inactive with probabilities p_3' and $1-p_3'$, respectively, where $p_3' = v_1 p_3$ (and p_3 is the corresponding probability for the non-vaccinated inactive cases). For simplicity we will denote $\alpha_2' = p_3' p_r \alpha = v_1 \alpha_2$. Those who develop TB are infectious or non-infectious with probabilities q_3 and $1 - q_3$, respectively (equal to those for the non-vaccinated).

Finally, vaccinated inactive cases may develop TB due to endogenous reactivation at a rate $\beta' = v_2 \beta$, with $0 < v_2 < 1$. The reduction v_2 in the reactivation rate β is either the same or greater than v_1 (the reduction in the rate of development of primary TB). There is not enough evidence to support either alternative, but since the efficacy of BCG wanes in time (Ferebee 1970, Murray et al. 1993), it is possible that v_2 is greater than v_1. For the moment we will assume that v_1 and v_2 are different, and later in Section 8.3, where we study the effect of BCG numerically, both cases (with $v_1 = v_2$ and $v_1 < v_2$) are examined. Since there is not enough evidence that the level of protection from the vaccine is different for different kinds (or severity) of TB (Rouillon & Waaler 1976), we assume that the probability q_2, that after reactivation the individual has infectious TB, is the same for vaccinated and unvaccinated inactive cases.

8.2 Model equations

The corresponding deterministic model is described by the following equations:

$$\begin{aligned}
\frac{dx}{dt} &= -\frac{\alpha}{n}xy - \mu x + (1-\phi)\lambda + \phi_1 x_v \\
\frac{dx_v}{dt} &= -\frac{\alpha}{n}x_v y - (\phi_1 + \mu)x_v + \phi\lambda \\
\frac{dz}{dt} &= (1-p)\frac{\alpha}{n}xy - \frac{\alpha_2}{n}yz - (\beta+\mu)z + \phi_2 z_v + \theta_1 y + \theta_2 w \\
\frac{dz_v}{dt} &= (1-p')\frac{\alpha}{n}x_v y - \frac{\alpha_2'}{n}yz_v - (\phi_2 + \beta' + \mu)z_v \\
\frac{dy}{dt} &= pq_1\frac{\alpha}{n}xy + p'q_1\frac{\alpha}{n}x_v y + q_3\frac{\alpha_2}{n}yz + q_3\frac{\alpha_2'}{n}yz_v - (\gamma_0 + \theta_1 + \mu + \mu_1)y \\
&\quad + q_2\beta z + q_2\beta' z_v + \delta w + \epsilon_1 u \\
\frac{dw}{dt} &= p(1-q_1)\frac{\alpha}{n}xy + p'(1-q_1)\frac{\alpha}{n}x_v y + (1-q_3)\frac{\alpha_2}{n}yz + (1-q_3)\frac{\alpha_2'}{n}yz_v \\
&\quad + (1-q_2)\beta z + (1-q_2)\beta' z_v - (\delta + \delta_0 + \theta_2 + \mu + \mu_2)w + \epsilon_2 u \\
\frac{du}{dt} &= \gamma_0 y + \delta_0 w - (\epsilon_1 + \epsilon_2 + \mu)u,
\end{aligned}$$

where x, x_v, z, z_v, y, w, and u are non-negative continuous functions. The initial conditions are $\mathbf{x}_0 = (x(0), x_v(0), z(0), z_v(0), y(0), w(0), u(0)) \in \mathcal{S}_0$, where

$$\mathcal{S}_0 = \{(h_1, h_2, \ldots, h_7) \in \mathbb{Z}_+^7 : 1 \le h_1 + h_2 \le n-1, h_1 + \cdots + h_7 = n\}, \tag{8.1}$$

and $n \geq 2$ is the initial total population size.

For the stochastic model, the probabilities $p(\mathbf{x};t) = p(x, x_v, z, z_v, y, w, u; t)$ are defined as

$$p(\mathbf{x};t) = \mathrm{P}[X(t) = x, X_v(t) = x_v, Z(t) = z, Z_v(t) = z_v, Y(t) = y, W(t) = w, U(t) = u],$$

for $t \geq 0$ and $\mathbf{x} = (x, x_v, z, z_v, y, w, u) \in \mathcal{S} = \mathbb{Z}^7_+$ and $p(\mathbf{x};t) = 0$ whenever $\mathbf{x} \notin \mathcal{S}$. The initial conditions are $p(\mathbf{x}_0; 0) = 1$ and $p(\mathbf{x}; 0) = 0$ for any $\mathbf{x} \neq \mathbf{x}_0$, where $\mathbf{x}_0 \in \mathcal{S}_0$ as defined in (8.1). The corresponding Kolmogorov forward equations for $p(\mathbf{x};t)$ are given in the Appendix (Section A.5.1). The joint probability generating function

$$\mathcal{P}(h_1, \ldots, h_7; t) = \mathrm{E}[h_1^{X(t)} h_2^{X_v(t)} h_3^{Z(t)} h_4^{Z_v(t)} h_5^{Y(t)} h_6^{W(t)} h_7^{U(t)}]$$

satisfies the equation

$$\begin{aligned}
\frac{\partial \mathcal{P}}{\partial t} &= \lambda[(1-\phi)h_1 + \phi h_2 - 1)\mathcal{P} + \mu(1 - h_1)\frac{\partial \mathcal{P}}{\partial h_1} \\
&+ [\mu(1-h_2) + \phi_1(h_1 - h_2)]\frac{\partial \mathcal{P}}{\partial h_2} \\
&+ [\mu(1-h_3) + q_2\beta(h_5 - h_3) + (1-q_2)\beta(h_6 - h_3)]\frac{\partial \mathcal{P}}{\partial h_3} \\
&+ [\mu(1-h_4) + q_2\beta'(h_5 - h_4) + (1-q_2)\beta'(h_6 - h_4) + \phi_2(h_3 - h_4)]\frac{\partial \mathcal{P}}{\partial h_4} \\
&+ [(\mu + \mu_1)(1-h_5) + \gamma_0(h_7 - h_5) + \theta_1(h_3 - h_5)]\frac{\partial \mathcal{P}}{\partial h_5} \\
&+ [(\mu + \mu_2)(1-h_6) + \delta(h_5 - h_6) + \delta_0(h_7 - h_6) + \theta_2(h_3 - h_6)]\frac{\partial \mathcal{P}}{\partial h_6} \\
&+ [\mu(1-h_7) + \epsilon_1(h_5 - h_7) + \epsilon_2(h_6 - h_7)]\frac{\partial \mathcal{P}}{\partial h_7} \\
&+ \frac{\alpha}{n}h_5[-h_1 + pq_1h_5 + (1-p)h_3 + p(1-q_1)h_6]\frac{\partial^2 \mathcal{P}}{\partial h_1 \partial h_5} \\
&+ \frac{\alpha}{n}h_5[-h_2 + p'q_1h_5 + (1-p')h_4 + p'(1-q_1)h_6]\frac{\partial^2 \mathcal{P}}{\partial h_2 \partial h_5} \\
&+ \frac{\alpha_2}{n}h_5[q_3h_5 + (1-q_3)h_6 - h_3]\frac{\partial^2 \mathcal{P}}{\partial h_3 \partial h_5} \\
&+ \frac{\alpha_2'}{n}h_5[q_3h_5 + (1-q_3)h_6 - h_4]\frac{\partial^2 \mathcal{P}}{\partial h_4 \partial h_5},
\end{aligned}$$

with the initial condition $\mathcal{P}(h_1, \ldots, h_7; 0) = h_1^{X(0)} h_2^{X_v(0)} h_3^{Z(0)} h_4^{Z_v(0)} h_5^{Y(0)} h_6^{W(0)} h_7^{U(0)}$.

From the equation for the probability generating function a system of differential equations for the first and second moments of X, X_v, Z, Z_v, Y, W, and U is deduced;

the equations for the means are

$$
\begin{aligned}
\frac{d\mathrm{E}[X]}{dt} &= -\frac{\alpha}{n}\mathrm{E}[XY] - \mu\mathrm{E}[X] + (1-\phi)\lambda + \phi_1\mathrm{E}[X_v] \\
\frac{d\mathrm{E}[X_v]}{dt} &= -\frac{\alpha}{n}\mathrm{E}[X_vY] - (\phi_1+\mu)\mathrm{E}[X_v] + \phi\lambda \\
\frac{d\mathrm{E}[Z]}{dt} &= (1-p)\frac{\alpha}{n}\mathrm{E}[XY] - \frac{\alpha_2}{n}\mathrm{E}[YZ] - (\beta+\mu)\mathrm{E}[Z] + \phi_2\mathrm{E}[Z_v] + \theta_1\mathrm{E}[Y] + \theta_2\mathrm{E}[W] \\
\frac{d\mathrm{E}[Z_v]}{dt} &= (1-p')\frac{\alpha}{n}\mathrm{E}[X_vY] - \frac{\alpha_2'}{n}\mathrm{E}[YZ_v] - (\phi_2+\beta'+\mu)\mathrm{E}[Z_v] \\
\frac{d\mathrm{E}[Y]}{dt} &= pq_1\frac{\alpha}{n}\mathrm{E}[XY] + p'q_1\frac{\alpha}{n}\mathrm{E}[X_vY] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZ] + q_3\frac{\alpha_2'}{n}\mathrm{E}[YZ_v] + q_2\beta\mathrm{E}[Z] \\
&\quad - (\gamma_0+\theta_1+\mu+\mu_1)\mathrm{E}[Y] + q_2\beta'\mathrm{E}[Z_v] + \delta\mathrm{E}[W] + \epsilon_1\mathrm{E}[U] \\
\frac{d\mathrm{E}[W]}{dt} &= (1-q_1)\frac{\alpha}{n}(p\mathrm{E}[XY] + p'\mathrm{E}[X_vY]) + (1-q_3)(\frac{\alpha_2}{n}\mathrm{E}[YZ] + \frac{\alpha_2'}{n}\mathrm{E}[YZ_v]) \\
&\quad + (1-q_2)(\beta\mathrm{E}[Z] + \beta'\mathrm{E}[Z_v]) - (\delta+\delta_0+\theta_2+\mu+\mu_2)\mathrm{E}[W] + \epsilon_2\mathrm{E}[U] \\
\frac{d\mathrm{E}[U]}{dt} &= \gamma_0\mathrm{E}[Y] + \delta_0\mathrm{E}[W] - (\epsilon_1+\epsilon_2+\mu)\mathrm{E}[U],
\end{aligned}
$$

where the term $\mathrm{E}[XY]$ can be expressed as $\mathrm{E}[XY] = \mathrm{Cov}[X,Y] + \mathrm{E}[X]\mathrm{E}[Y]$, and similarly for the terms $\mathrm{E}[X_vY]$, $\mathrm{E}[YZ]$, and $\mathrm{E}[YZ_v]$. The equations for the variances and covariances are given in the Appendix (Section A.5.1).

8.3 Epidemiology

In this section we present some numerical results for the epidemiological indices defined in Definition 6.1. For this model the risk of infection is the number of primary infections of both vaccinated and unvaccinated individuals (X and X_v). Similarly the risk of reinfection is the number of reinfections of vaccinated and unvaccinated inactive cases (Z and Z_v). The incidence of infectious TB is the number of new infectious cases that developed during a year, which counts all the transitions from the uninfected and inactive classes (X, X_v, Z, Z_v) to the class of infectious cases (Y). The incidence of non-infectious TB counts all the transitions from the uninfected and inactive classes (X, X_v, Z, Z_v) to the class of non-infectious cases (W). Finally, the prevalence of TB infection is the proportion of infected individuals ($Z + Z_v + Y + W + U$) in the population. All these rates are defined (and were calculated from the numerical simulations) as proportions per 10^5 general population. Details of the implementation of the simulations can be found in the Appendix (Section A.5.2).

The parameter values are the same as in the previous chapter (see Table 6.8), with $\theta_1 = 0.4$, $\theta_{2r} = 0.5$, and $\theta_2 = \theta_{2r}\theta_1$. For v_1 and v_2, the reduction in the probability of developing TB (primary and secondary, respectively), we used the following values: $v_1 = 20, 40, 60, 80\%$ and $v_2 = v_1$ or $v_2 = v_1 + 10\%$.

For the proportion ϕ of newborns vaccinated, we used two different values, 50% and 95%. The value $\phi = 95\%$ is rather extreme (since it assumes that almost all newborns are vaccinated), but still it is a value that has been achieved (see Rouillon & Waaler 1976, Fine 1995) and also it gives a limit for the maximum effect that BCG can have.

The values of ϕ_1 and ϕ_2 depend on the length of the protection conferred by BCG. It is believed that the protection can last for 10–15 years, but there is no evidence that it can last for more than 15 years (see, e.g., Murray et al. 1993, Styblo 1991). For the numerical results presented in this section we used an average length of protection of 12 years, so that $\phi_1 = 0.0833$. The value of ϕ_2 is greater than or equal to ϕ_1. If the vaccinated individual gets infected and becomes inactive soon after vaccination then $\phi_2 = \phi_1$; in all other situations $\phi_2 > \phi_1$. The difference $\phi_2 - \phi_1$ depends on the length of the time between vaccination and infection. The sooner the vaccinated individual gets infected, the smaller the difference $\phi_2 - \phi_1$ will be. Therefore the value of ϕ_2 depends on the particular individual and the only knowledge we have for its value is that it will be at least equal to ϕ_1. For the numerical results presented in this section we assumed that $\phi_2 = \phi_1$ so that will give a view of the maximum effect that BCG can have.

The initial conditions were taken as in model Clio (see Section 7.4.1). Model Zeus (Chapter 6) describes the natural evolution of TB and is the same as model Clio with $\theta_1 = \theta_2 = 0$ and the same as model Erato with θ_1, θ_2, ϕ, ϕ_1, ϕ_2, v_1, and v_2 all equal to zero. The epidemic begins with the introduction of the infection into a population of size n (for the results presented in this section n was taken equal to 1000). In the absence of any control measures, the infection spreads and finally settles down at an endemic stable equilibrium (the quasi-stationary distribution of model Zeus; see Sections 6.3.3 and 6.3.6). For the numerical results presented in this chapter we considered only the situations where the control program is introduced after the process has reached this steady endemic level, since in most countries the controls were introduced when the infection was endemic. Other initial conditions have not been considered here, although they may give different results for the effectiveness of the vaccine (for instance, if the prevalences are at a lower level when the controls are introduced, then the progress of the epidemic can be seen from the figures presented in this section from the time point that the prevalences have reached this level). The following two control programs are considered in this chapter:

(a) a program implementing only chemotherapy (the progress of the epidemic then is described by model Clio).

(b) a program implementing both chemotherapy and BCG vaccination (the progress of the epidemic then is described by model Erato).

The initial conditions for both Clio and Erato are taken from the endemic stable state of Zeus: model Zeus is simulated (with the same parameter values) until the process reaches the stable state; then both Clio and Erato are simulated by taking the initial value of the vector (X, Y, Z, W, U) to be equal to the equilibrium value of (X, Y, Z, W, U) deduced from the simulations of model Zeus (and $X_v(0) = Z_v(0) = 0$ for Erato, since there are no vaccinated individuals at time $t = 0$). Details of how the simulations were implemented can be found in the Appendix (Section A.5.2).

Both models were simulated with the parameters shown in Table 6.8 and $\theta_1 = 0.4$, $\theta_{2r} = 0.5$, $\theta_2 = \theta_{2r}\theta_1$, and for Erato the values of ϕ, ϕ_1, ϕ_2, v_1, v_2 mentioned above (the results for Clio are also shown in Section 7.4). Estimates of each epidemiological index were calculated from the models Clio and Erato, say $F_c(t)$ and $F_\varepsilon(t)$, respectively (t years after the introduction of the respective control program). Then the percentage decline due to the vaccine was calculated from the formula

$$D(t) = \frac{F_c(t) - F_\varepsilon(t)}{F_c(t)}100. \tag{8.2}$$

In the remaining part of this section we present results (obtained from numerical simulations of the models Clio and Erato) for the percentage decline in the epidemiological indices and the variation of these estimates. Figure 8.2 shows the decline in the prevalence of TB infection for $\phi = 50, 95\%$ and v_2 equal to v_1 and $v_1 + 10\%$. Figure 8.3 shows the decline in the risk of infection and the incidence of infectious TB with $\phi = 95\%$ and $v_2 = v_1$. For the other cases the results were qualitatively similar and are not shown here. Also, the 95% confidence intervals for the estimates of the incidence and prevalence of infectious TB and the risk of infection and reinfection are shown in Table 8.2.

As the values of v_1 and v_2 decrease (which means that the protection from BCG

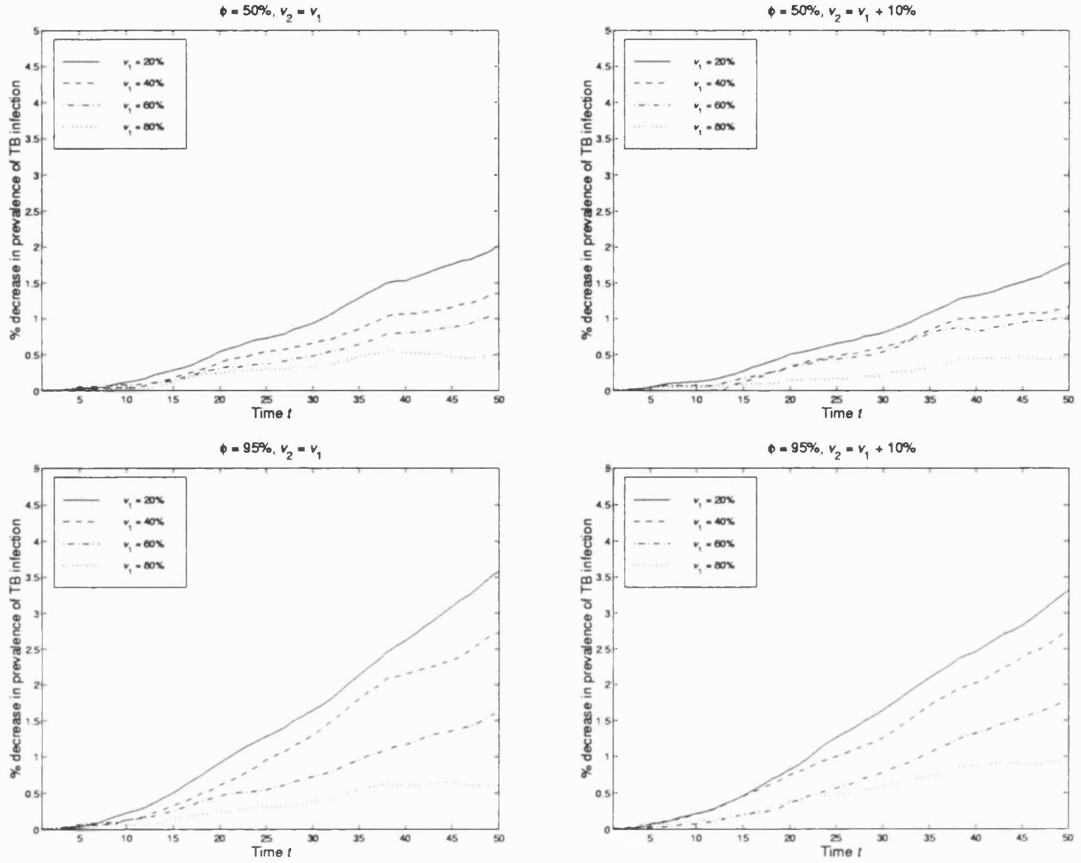


Figure 8.2: Percentage decline in the prevalence of TB infection due to BCG vaccination (calculated from the formula (8.2)). The prevalence was calculated as proportion per 10^5 general population. The parameter values are as shown in Table 6.8 and $\theta_1 = 0.4$, $\theta_{2r} = 0.5$, $\theta_2 = \theta_{2r}\theta_1$, $\phi_1 = \phi_2 = 0.0833$. The initial conditions are taken from the endemic steady level of the natural evolution of TB. Time t is measured in years after the introduction of the control policy.

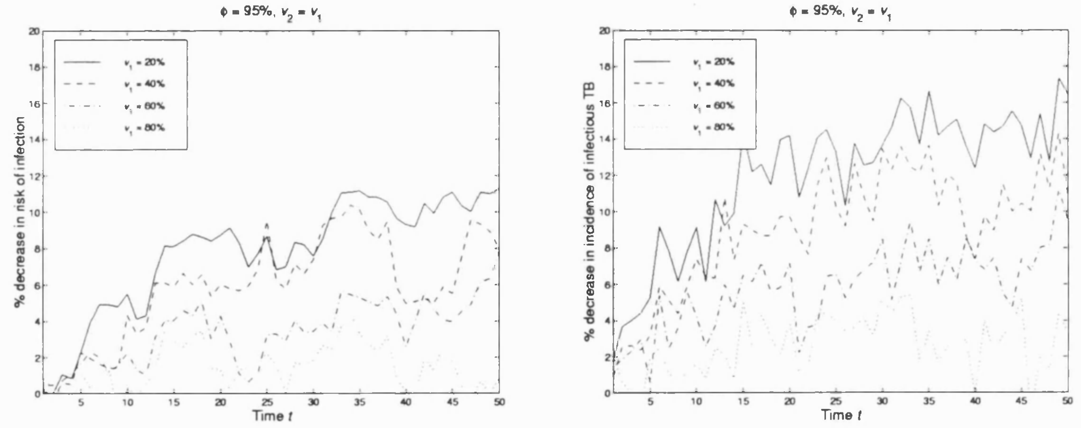


Figure 8.3: Percentage decline in the risk of infection and the incidence of infectious TB due to BCG vaccination (calculated from the formula (8.2)). The values of the rates are proportions per 10^5 general population. The parameter values are as shown in Table 6.8 and $\theta_1 = 0.4$, $\theta_{2r} = 0.5$, $\theta_2 = \theta_{2r}\theta_1$, $\phi_1 = \phi_2 = 0.0833$. The initial conditions are taken from the endemic steady level of the natural evolution of TB. Time t is measured in years after the introduction of the control policy.

Year	Incidence of	Prevalence of	Risk of	Risk of
	infectious TB	infectious TB	infection	reinfection
1	(242.9, 250.3)	(1173.8, 1193.6)	(1713.8, 1738.7)	(210.0, 217.5)
2	(193.3, 199.7)	(837.8, 853.6)	(1246.1, 1267.9)	(145.8, 151.7)
3	(161.9, 167.8)	(630.4, 643.3)	(967.3, 987.3)	(106.0, 111.1)
4	(142.5, 148.1)	(502.9, 514.1)	(804.8, 823.6)	(82.5, 87.0)
5	(130.0, 135.3)	(416.1, 426.1)	(704.3, 722.4)	(66.0, 70.0)
6	(119.8, 124.8)	(362.0, 371.3)	(640.5, 658.0)	(54.3, 57.9)
7	(115.5, 120.3)	(325.4, 334.0)	(605.6, 623.0)	(46.5, 49.7)
8	(113.2, 118.1)	(301.7, 310.0)	(588.5, 605.9)	(43.4, 46.5)
9	(109.5, 114.3)	(287.9, 296.0)	(585.5, 603.1)	(39.5, 42.5)
10	(105.8, 110.4)	(276.1, 284.0)	(587.3, 605.3)	(38.7, 41.7)
11	(104.8, 109.4)	(264.8, 272.6)	(594.9, 613.4)	(36.3, 39.1)
12	(100.7, 105.3)	(256.1, 263.7)	(601.9, 620.7)	(33.6, 36.3)
13	(101.8, 106.4)	(248.7, 256.2)	(601.5, 620.6)	(32.0, 34.8)
14	(98.2, 102.7)	(240.2, 247.5)	(608.3, 627.8)	(29.7, 32.2)
15	(94.6, 99.0)	(234.4, 241.6)	(614.1, 633.9)	(29.3, 31.8)
16	(94.7, 99.0)	(229.2, 236.4)	(621.7, 641.9)	(28.4, 31.0)
17	(93.0, 97.4)	(225.1, 232.2)	(626.7, 647.3)	(28.2, 30.7)
18	(93.0, 97.4)	(219.7, 226.7)	(633.5, 654.3)	(26.2, 28.6)
19	(89.4, 93.7)	(214.6, 221.5)	(634.7, 655.7)	(24.9, 27.2)
20	(88.9, 93.2)	(209.4, 216.3)	(639.9, 661.4)	(24.5, 26.8)

Table 8.2: 95% confidence intervals for the estimates of the incidence and prevalence of infectious TB and the risk of infection and reinfection. The values of the rates are proportions per 10^5 general population. The parameter values are as shown in Table 6.8 and $\theta_1 = 0.4$, $\theta_{2r} = 0.5$, $\theta_2 = \theta_{2r}\theta_1$, $\phi_1 = \phi_2 = 0.0833$ $\phi = 95\%$, $v_1 = v_2 = 20\%$. The initial conditions are taken from the endemic steady level of the natural evolution of TB. Time is measured in years after the introduction of the control policy.

increases) the epidemic becomes less severe and hence the epidemiological indices decrease. The decline becomes more significant as time increases, since the number of infectious cases slowly decreases and that reduces the TB problem in the long run. For some indices though the reduction is very small. For instance the decline in the prevalence of TB infection is just 3.5% after 50 years with $\phi = 95\%$ and $v_1 = v_2 = 20\%$. For the same parameters the decline in the risk of infection is more considerable (about 4% at $t = 5$ and around 10% after 30 years) and even more for the incidence of infectious TB (around 4% at $t = 5$ and between 10% and 16% after 15 years). This was observed with the other values of ϕ and v_1, v_2 as well.

This difference can be explained by the fact that BCG offers protection against development of disease, but not against infection. The level of protection is determined by the parameters v_1 and v_2, which reduce the rates of development of disease after infection (primary and secondary, respectively). Reducing these rates means that the numbers of new cases that develop each year are reduced. The incidence rate represents exactly this number of new cases and hence the most significant effect of BCG will be on the incidence rate.

Since the vaccine does not protect against infection, the number of new infections will be the same initially (after the introduction of the BCG campaign). When the number of new cases developing each year starts to decrease, the number of infectious cases (Y) will decrease and hence that of new infections as well (since new infections occur at a rate $\alpha XY/n$). The risk of infection represents precisely this number of new infections, but since it is more indirectly related to the reduction in the number of new cases (than the incidence rate), the decline in the risk of infection will be smaller than the decline in the incidence.

On the other hand, the prevalence of TB infection is the index that we would expect to be the least affected by BCG. The prevalence of infection is the proportion of infected individuals in the population and that will be significantly reduced when the sizes of all Z, Z_v, Y, W, U classes are reduced. Since BCG does not protect against infection, the sizes of these classes will be more or less the same in the beginning. After the number of new cases developing each year starts decreasing, the sizes of Y, W, and U will decrease, but it will take more time until the number of inactive cases is considerably reduced (since for those in the Z_v class the rate of development to disease is reduced, and hence most of them will remain in this class). For instance, with $\phi = 95\%$ and $v_1 = v_2 = 20\%$, the percentage decline in the number of inactive cases (i.e. the difference between Z from model Clio and $Z + Z_v$ from Erato) is less than 0.5% during the first 15 years and around 3.5% at $t = 50$.

8.4 Discussion and conclusions

The BCG vaccine is one of the most controversial vaccines because of the uncertainty about its effectiveness (in protecting those vaccinated) and its impact on the general TB problem. The results presented in the previous section can provide some insight into the effect of a mass-vaccination campaign for newborns, but there are a few things that have to be taken into account when interpreting these results.

First of all, the model presented in this chapter does not take into account the age distribution of the population. This can affect the results for the effectiveness of the vaccine, since the protection from BCG lasts only for one or two decades and the age distribution of the infectious forms of TB is not uniform. For instance, the proportion of infectious cases following primary infection is about 2% for those aged 0–14 and about 24% for those aged 15–29 (Styblo 1991). An age-dependent model might be more suitable for the study of the effectiveness of BCG.

Another drawback is the uncertainty about the values of some of the parameters used in this model, for instance the values of ϕ_2 (which depends on the typical length of the time between vaccination and infection), q_1', q_2', q_3' (whether they are equal to q_1, q_2, q_3, respectively, or not), and the difference between v_1 and v_2. Although small differences in the assumptions we have made in this chapter may not affect the results substantially, this is a matter that should also be taken into consideration.

Apart from these model-related considerations, there are other matters that make the use of BCG controversial. For instance, it is believed that vaccination at school-leaving age may also have a significant effect in controlling TB, both directly and indirectly. The direct effect is by decreasing the probability of developing TB for those vaccinated. The indirect effect is that by preventing the development of disease (for those vaccinated and infected), the number of infectious cases decreases and hence the number of new infections as well, thus decreasing the probability of developing TB even for the unvaccinated individuals. Since the infectious forms of TB are more common among teenagers and young adults than among infants, vaccination at school-leaving age may prevent more infectious cases and hence have a more significant effect than vaccination of newborns (Styblo 1991). On the other hand, if vaccination is not given at birth but at an older age, then a proportion of those vaccinated will have already been infected and there is no evidence that BCG can offer any protection if given after infection (Styblo 1991).

The main advantages of BCG are that its financial costs are low, it is simple to administer, and causes few complications (Smith & Fine 1998, Comstock & Geiter 1994). Therefore in areas with a high risk of infection, BCG can have a significant impact over time in fighting TB. The major disadvantage is that it invalidates the tuberculin skin test, thus interfering with the diagnosis of TB (Huebner 1996, Murray et al. 1993). Combined with the uncertainty about its efficacy, this makes most health-policy makers believe that in areas with low risk of infection there is little need for a mass-vaccination campaign. In these areas most TB cases occur among people who have been infected in the past; it is crucial to identify these individuals as early as possible and the tuberculin test is the easiest and most immediate diagnostic tool. BCG vaccination may invalidate the skin test, and possibly without offering any protection.

Finally it has to be mentioned that the impact of HIV on tuberculosis and the emergence of multi-drug resistant TB has lead to reassessment of the vaccination policies in areas with high risk of infection (Smith & Fine 1998).

The results presented in the previous section show that a mass-vaccination campaign among newborns can contribute in the control of TB. Comparing a policy that implements only chemotherapy with a policy implementing both chemotherapy and BCG vaccination of newborns, the decline in the epidemiological indices is not very high in most situations, but when this decline is viewed in terms of human lives saved it cannot be ignored, no matter how small it is. BCG has been characterised as the vaccine that is "given to the young to protect against a disease that is most common in adulthood" (LaScolea & Rangoonwala 1996) and it is believed that the absolute number of cases prevented by BCG is very small. Since TB is highly fatal, though, this small number of cases among children can be translated into a significant number of man-years saved.

The question that health-policy makers are facing is whether this effect is worth the time and money invested in a mass-BCG campaign. For instance it may be the case that if this investment were made in increasing the rate of diagnosing and treating infectious cases instead, then there would be a more significant effect in the decline of tuberculosis. Unfortunately it is very hard to estimate the cost-effectiveness of BCG (Murray et al. 1993). Rouillon & Waaler (1976) developed a decision model that takes into account epidemiological, economical, and psychological aspects of a mass-BCG campaign in order to assess when (and if) it should be terminated. Their final conclusion was that "... the decision itself — to start, to pursue or to stop BCG on a mass scale — is of political nature (public health policy): it always implies a value judgement — from the part of the providers ... and/or, preferably, from the part of the users, i.e. a preference which is a result of a 'weighing' between advantages and inevitable disadvantages ...".

Chapter 9

Discussion

9.1 Conclusions

The main use of mathematical modelling in epidemiology is to improve our understanding of the dynamics of the infection under study and, in particular, how the progress of the infection at the individual level affects its progress at the population level. In this thesis we have presented models for the evolution of TB in the absence and in the presence of control measures. The study of models for the natural evolution of TB can improve our understanding of the epidemiology of the disease (since some aspects of its dynamics will be obscured by the effects of medical treatment) and show how predictions for the future patterns of TB can be made. The models that include control measures can provide insights about the effectiveness of the controls and allow comparisons to be made between different control policies.

In Chapters 4 and 5 we presented a simple model for a population with constant and variable sizes, respectively. The dynamics of *M. tuberculosis* are quite complicated and these models may seem oversimplified. Nevertheless, they do account for the most important determinants of the spread of TB and provide insights for the development of a TB epidemic, which were verified by the more detailed model presented in Chapter 6.

The progress of an epidemic will be determined mainly by the initial size of the population and the value of the basic reproduction ratio, $\mathcal{R}_0$. In an open population, the infection will ultimately die out with probability one. Before extinction, though, the epidemic may settle down at an endemic level, unless the value of $\mathcal{R}_0$ and/or the initial population size are too small to preserve an endemic infection. The results from

Chapter 6 also show that, in contrast with other epidemic models, the condition $\mathcal{R}_0 < 1$ may not be enough to ensure that there will not be a major epidemic (since for the corresponding deterministic model both the endemic and the disease-free equilibrium can be stable when $\mathcal{R}_0 < 1$). It has to be stressed though that even in the situations where there is no major epidemic, the time until the final extinction of the infection (i.e. until there is no infected individual in the population) can be very long.

These results agree with the history of TB epidemics in the pre-chemotherapy era. The existence of TB has been traced back to Egyptian mummies (4000–2000 BC), but major epidemics did not arise until the 1600's in Europe and somewhat later in North America (Bloom & Murray 1992). From the beginning of the 20th century these epidemics were in decline, although effective therapy had not been introduced (Murray et al. 1993).

This pattern can be explained by the fact that until the 1600's there were only a few infectious cases which, given the small size of the populations of that era, could not give rise to major epidemics. On the other hand, the number of infectives was large enough to keep the infection within the population (since these few infectives cause new infections and thus increase the pool of infected individuals).

The population growth, urbanisation, and industrialisation in Europe in the 17th century created the conditions needed for a major epidemic: (a) increasing population sizes and (b) crowding of people (increased contact rates), increased poverty, unsanitary living conditions, malnutrition, resulting in poor health and hence weak immune system (higher reactivation and reinfection rates), overall increased $\mathcal{R}_0$. It is possible therefore that an increase in the population sizes and the value of $\mathcal{R}_0$ caused the outbreaks of the 17th century, which peaked and then fell to their endemic stable level.

Further results on the distribution of the numbers of infectious cases and infected individuals, the size of the epidemic, and the epidemiological indices (incidence, prevalence, mortality, and risk of infection) have been obtained by numerical simulations and approximations of the stochastic models presented in Chapters 4–6. In particular, the normal and linear approximations proved to be quite efficient in estimating the moments of the sizes of the various classes into which the population is divided. The advantage of these approximations is that numerical results are obtained more quickly (since the results are deduced by solving systems of differential equations) and still, compared to the deterministic models, they can provide information about the variability (variance, covariance, confidence intervals) of the variables of interest.

We continued our study with two models that allow the implementation of a control policy: in Chapter 7 we considered a policy implementing only chemotherapy and in Chapter 8 a policy implementing chemotherapy and BCG vaccination of newborns. Numerical results for the percentage decline in the epidemiological indices were obtained for various levels of detection and cure rates for chemotherapy and various coverage and protective levels for BCG. Our results show that treatment of non-infectious cases and vaccination cannot significantly contribute to the reduction in the TB problem. Treatment of infectious cases is the most important tool in controlling TB, and since the currently recommended regimens have proved to be effective (Murray et al. 1991, China Tuberculosis Control Collaboration 1996), the main aim of any control program should be the immediate detection of those who are infectious and their successful treatment.

9.2 Recommendations for further research

One of the areas in which the work of this thesis can be extended is with regard to the quasi-stationary distributions. Several issues still remain open, for instance the formal proof of the existence of quasi-stationary distributions, the distribution of the time until the process reaches quasi-stationarity and how long it remains there. The importance of the quasi-stationary distributions arises from the fact that for any practical purposes for TB, it is not the actual stochastic equilibrium (extinction) that will be observed, but the endemic quasi-stationary distribution.

For the same reason, it would be interesting to investigate further the stability of the deterministic equilibria for the models presented in Chapters 6–8, since the quasi-stationary distributions are centred around a level close to the deterministic endemic equilibrium.

Another area for further research is the analysis of data for parameter estimation, model validation, and curve-fitting (see, e.g., Vynnycky (1996), Vynnycky & Fine (1997), Becker (1989) for statistical models for TB and other infectious diseases and Styblo (1991), Murray et al. (1993) for discussions about available data for TB).

It is also possible to modify the models presented in this thesis in order to account for other features of TB that have not been included here. For instance, the population can be divided into specific age-groups (as well as clinical states) in order to account for the fact that the probability of developing infectious (vs. non-infectious) TB and the reactivation rate vary depending on the age of the infected individual.

Also, some parameters can be taken as functions of time, for instance the contact rate or the reactivation rate, in order to account for behavioural changes, public awareness of the TB problem, better nutrition and so on.

In addition, it might be helpful to model the dynamics of TB using a cluster model (taking as clusters the circle of close contacts, for instance home, school, work place), since the contact rate, and hence the probability of being infected, is much higher for the individuals that have daily contact with an infectious case than for those who have only casual contact.

The models presented in this thesis can also be extended to account for the effects of preventive therapy (chemoprophylaxis), multidrug-resistant TB, and HIV infection. In particular for the effect of preventive therapy, as well as the effect of BCG vaccination, it would be interesting to take into account the cost-benefits of these controls, since no control program can depend solely on BCG or chemoprophylaxis (Murray et al. 1993, Styblo 1991) and the decision on whether to implement one of these controls along with chemotherapy will be based on whether their contribution in reducing the TB problem is worth the means invested in them.

Appendix A

A.1 The first model

A.1.1 The variances and covariance of X and Y

The variances of X and Y (σ_{XX}, σ_{YY}) and the covariance of X, Y (σ_{XY}) satisfy the following equations

$$\begin{aligned}
\frac{d\sigma_{XX}}{dt} &= \frac{\alpha}{n}(\sigma_{XY} + \mu_X\mu_Y)(1 + 2\mu_X) - 2\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
\frac{d\sigma_{YY}}{dt} &= 2\frac{\alpha(1-\rho)}{n}\mathrm{E}[XY^2] + \beta(n - \mu_X) + (\gamma - \beta)\mu_Y - 2(\beta + \gamma)\sigma_{YY} \\
&\quad - 2\beta\sigma_{XY} + \frac{\alpha(1-\rho)}{n}(\sigma_{XY} + \mu_X\mu_Y)(1 - 2\mu_Y) \qquad\qquad (\mathrm{A.1}) \\
\frac{d\sigma_{XY}}{dt} &= \frac{\alpha(1-\rho)}{n}\mathrm{E}[X^2Y] - \frac{\alpha}{n}\mathrm{E}[XY^2] - \beta\sigma_{XX} - (\beta + \gamma)\sigma_{XY} \\
&\quad - \frac{\alpha}{n}(\sigma_{XY} + \mu_X\mu_Y)[(1-\rho)(1 + \mu_X) - \mu_Y].
\end{aligned}$$

A.1.2 Numerical results from simulations

For the simulation of this model there are 4 different kinds of events that can occur. The events and their rates are defined in Table A.1. The simulation clock, $\mathcal{T}$, gives the current time and the vector $\mathbf{X}_c = (X_c, Y_c)$ gives the current value of the state of the process, $\mathbf{X} = (X, Y)$. At time $t = 0$ the simulation clock is set equal to 0 and $\mathbf{X}_c$ is set equal to the initial value $\mathbf{X}_0 = (x_0, y_0)$.

The rates of the events shown in Table A.1 are calculated using the current value of $\mathbf{X}_c$. The sum of these rates gives the rate of the exponentially distributed time until the next event occurs. A variate from the exponential distribution (with this rate) is generated to determine the time of the next event and then a Uniform$(0, 1)$ variate is generated to determine what kind of event it is. The simulation clock, $\mathcal{T}$, is advanced to

Event	Rate
$(X,Y,Z) \to (X-1,Y+1,Z)$	$(1-\rho)\frac{\alpha}{n}XY$
$(X,Y,Z) \to (X-1,Y,Z+1)$	$\rho\frac{\alpha}{n}XY$
$(X,Y,Z) \to (X,Y-1,Z+1)$	γY
$(X,Y,Z) \to (X,Y+1,Z-1)$	βZ

Table A.1: Events for the simulations of the first model

the time of the next event, the value of $\mathbf{X}_c$ is updated according to what kind of event it is, and also the statistics of interest are updated.

The stochastic model was simulated $R = 10^4$ times. The estimates for the means of X at time t were calculated using the formula $\hat{\mu} = \sum_{i=1}^{R} x_i/R$, where x_i is the value of X at time t in the i-th individual simulation run. The mean of Y was calculated similarly. The mean of Z was deduced from those of X and Y, as $\mathrm{E}[Z] = n - \mathrm{E}[X] - \mathrm{E}[Y]$.

The simulations were terminated after 100 years and the number of susceptibles was zero at that point in all 10^4 individual runs. Consequently the $R = 10^4$ runs yielded a sample of R independent values $\tau_1, \tau_2, \dots, \tau_R$, where τ_i is the time that the last susceptible got infected in the i-th run. The sample $\tau = \{\tau_1, \tau_2, \dots, \tau_R\}$ was then used to calculate the distribution of T shown in Figure 4.6.

A.2 The second model

A.2.1 The variances and covariances of X, Y, and Z

The variances and covariances of X, Y, and Z satisfy the following equations:

$$
\begin{aligned}
\frac{d\mathrm{Var}[X(t)]}{dt} &= b + \mu\mathrm{E}[X] - 2\mu\mathrm{Var}[X] + \frac{\alpha}{n}(1 + 2\mathrm{E}[X])\mathrm{E}[XY] - 2\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
\frac{d\mathrm{Var}[Y(t)]}{dt} &= (\gamma + \mu + \delta)\mathrm{E}[Y] + \beta\mathrm{E}[Z] - 2(\gamma + \mu + \delta)\mathrm{Var}[Y] + 2\beta\mathrm{Cov}[Y, Z] \\
&\quad + (1-\rho)\frac{\alpha}{n}(1 - 2\mathrm{E}[Y])\mathrm{E}[XY] + 2(1-\rho)\frac{\alpha}{n}\mathrm{E}[XY^2] \\
\frac{d\mathrm{Var}[Z(t)]}{dt} &= \gamma\mathrm{E}[Y] + (\beta + \mu)\mathrm{E}[Z] + 2\gamma\mathrm{Cov}[Y, Z] - 2(\beta + \mu)\mathrm{Var}[Z] \\
&\quad + \rho\frac{\alpha}{n}(1 - 2\mathrm{E}[Z])\mathrm{E}[XY] + 2\rho\frac{\alpha}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[X, Y]}{dt} &= \beta\mathrm{Cov}[X, Z] - (\gamma + \delta + 2\mu)\mathrm{Cov}[X, Y] + (1-\rho)\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
&\quad - \frac{\alpha}{n}\mathrm{E}[XY^2] - \frac{\alpha}{n}((1-\rho) + (1-\rho)\mathrm{E}[X] - \mathrm{E}[Y])\mathrm{E}[XY]
\end{aligned}
$$

$$
\begin{aligned}
\frac{d\mathrm{Cov}[X,Z]}{dt} &= \gamma\mathrm{Cov}[X,Y] - (\beta+2\mu)\mathrm{Cov}[X,Z] - \frac{\alpha}{n}(\rho+\rho\mathrm{E}[X]-\mathrm{E}[Z])\mathrm{E}[XY] \\
&\quad + \rho\frac{\alpha}{n}\mathrm{E}[X^2Y] - \frac{\alpha}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[Y,Z]}{dt} &= -\gamma\mathrm{E}[Y] + \gamma\mathrm{Var}[Y] - \beta\mathrm{E}[Z] + \beta\mathrm{Var}[Z] - (\beta+\gamma+\delta+2\mu)\mathrm{Cov}[Y,Z] \\
&\quad + \rho\frac{\alpha}{n}\mathrm{E}[XY^2] + (1-\rho)\frac{\alpha}{n}\mathrm{E}[XYZ] - \frac{\alpha}{n}(\rho\mathrm{E}[Y] + (1-\rho)\mathrm{E}[Z])\mathrm{E}[XY].
\end{aligned}
$$

A.2.2 Numerical results from simulations

The simulations were implemented as explained in Section A.1.2 for the first model. For this model there are 8 different kinds of events (as shown in Table A.2) and the vector $\mathbf{X}_c = (X_c, Y_c, Z_c)$ gives the current value of the state of the process $\mathbf{X} = (X, Y, Z)$. The initial conditions are $\mathbf{X}_0 = (x_0, y_0, z_0)$. For each set on initial conditions and parameter values the simulation was repeated $R = 10^4$ times.

Event	Rate
$(X,Y,Z) \to (X+1,Y,Z)$	b
$(X,Y,Z) \to (X-1,Y,Z)$	μX
$(X,Y,Z) \to (X,Y-1,Z)$	$(\mu+\delta)Y$
$(X,Y,Z) \to (X,Y,Z-1)$	μZ
$(X,Y,Z) \to (X-1,Y+1,Z)$	$(1-\rho)\frac{\alpha}{n}XY$
$(X,Y,Z) \to (X-1,Y,Z+1)$	$\rho\frac{\alpha}{n}XY$
$(X,Y,Z) \to (X,Y-1,Z+1)$	γY
$(X,Y,Z) \to (X,Y+1,Z-1)$	βZ

Table A.2: Events for the simulations of the second model

The moments of X, Y, and Z

The estimates for the mean and standard deviation of H and the covariance of H, H' at time t, for $H, H' = X, Y, Z$, were calculated from the formulae

$$\hat{\mu}_H = \frac{1}{R}\sum_{i=1}^{R} H_i \tag{A.2}$$

$$\hat{\sigma}_H = \left[\frac{1}{R-1}\sum_{i=1}^{R}(H_i - \hat{\mu}_H)^2\right]^{1/2} \tag{A.3}$$

$$\hat{\sigma}_{HH'} = \frac{1}{R-1}\left[\sum_{i=1}^{R} H_i H_i' - \frac{1}{R}\left(\sum_{i=1}^{R} H_i\right)\left(\sum_{i=1}^{R} H_i'\right)\right], \tag{A.4}$$

where H_i and H_i' are the values of H and H', respectively, at time t in the i-th individual simulation run. The conditional mean of H was calculated with the same formula as above for the mean, but with R equal to the number of runs in which the epidemic had not died out by time t.

The marginal distributions of X, Y, and Z

The marginal distributions were calculated from the following formula

$$\mathrm{P}[H(t)=k]=\frac{\mathcal{M}_k(t)}{R}$$

for $H=X,Y,Z$, where $\mathcal{M}_k(t)$ is the number of runs in which $H(t)=k$, for $t=1,2,\dots$ and $k=0,1,\dots$.

Time until extinction and size of the epidemic

For each set of parameter values and initial conditions the stochastic model was simulated 10^4 times. The simulations were terminated at a time point large enough such that all $R=10^4$ runs ended with extinction of the infection by that point. From each individual run i the extinction time τ_i and the size of the epidemic s_i were obtained (for $i=1,2,\dots,R$), thus yielding a sample $\tau=\{\tau_1,\dots,\tau_R\}$ from the distribution of the extinction time and a sample $s=\{s_1,\dots,s_R\}$ from the distribution of the size of the epidemic. These two samples were used to calculate the statistics of the extinction time and those of the size, presented in Section 5.3.6.

A.3 Model Zeus

A.3.1 The deterministic model

In this model the infected individuals can be in four different states: latent, infectious cases, non-infectious cases, and recovered. State transitions occur according to the rate matrix S, where

$$\mathsf{S}=\begin{bmatrix} -\gamma_0 & q_2\beta & \delta & \epsilon_1 \\ 0 & -\beta & 0 & 0 \\ 0 & (1-q_2)\beta & -(\delta+\delta_0) & \epsilon_2 \\ \gamma_0 & 0 & \delta_0 & -(\epsilon_1+\epsilon_2) \end{bmatrix},$$

and death occurs according to the diagonal matrix $\mathsf{D}=\operatorname{diag}\{\mu+\mu_1,\mu,\mu+\mu_2,\mu\}$. Let T be the matrix whose (i,j) element is the rate at which an infected individual with state

j produces secondary cases with state i, such that

$$\mathsf{T} = \begin{bmatrix} pq_1\frac{\alpha}{n}\frac{\lambda}{\mu} & 0 & 0 & 0 \\ (1-p)\frac{\alpha}{n}\frac{\lambda}{\mu} & 0 & 0 & 0 \\ p(1-q_1)\frac{\alpha}{n}\frac{\lambda}{\mu} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{A.5}$$

Then $\mathcal{R}_0$ is the dominant eigenvalue of the matrix $\mathsf{K} = -\mathsf{T}(\mathsf{S}-\mathsf{D})^{-1}$ (see Definition 3.1) and so the formula (6.9) for $\mathcal{R}_0$ is deduced.

Equilibrium points

The coefficients M_{1j}, $j = 1, 2, 3, 4$, for the characteristic polynomial $P_1(\tau)$ of $DF(\mathrm{e}_1)$ (see equation 6.11) are

$$\begin{aligned} M_{11} &= \Phi_s + \Delta_s + \mathrm{E}_s + m_4 \\ M_{12} &= -(1-p)\frac{\alpha}{n}\frac{\lambda}{\mu}q_2\beta + (\Delta_s + \mathrm{E})(\Phi_s + m_4) + \Phi_s m_4 - \delta p(1-q_1)\frac{\alpha}{n}\frac{\lambda}{\mu} - \epsilon_1\gamma_0 + m_1 \\ M_{13} &= -(1-p)\frac{\alpha}{n}\frac{\lambda}{\mu}\beta[q_2(\Delta_s + \mathrm{E}) + (1-q_2)\delta] - \Phi_s\delta p(1-q_1)\frac{\alpha}{n}\frac{\lambda}{\mu} - \Phi_s\epsilon_1\gamma_0 \\ &\quad + \Phi_s(\Delta_s + \mathrm{E})m_4 + (\Phi_s + m_4)m_1 - p(1-q_1)\frac{\alpha}{n}\frac{\lambda}{\mu}m_2 - \gamma_0 m_3 \\ M_{14} &= -(1-p)\frac{\alpha}{n}\frac{\lambda}{\mu}\beta[q_2 m_1 + (1-q_2)m_2] + \Phi_s m_4 m_1 - \Phi_s p(1-q_1)\frac{\alpha}{n}\frac{\lambda}{\mu}m_2 - \Phi_s\gamma_0 m_3, \end{aligned}$$

where $m_1 = \Delta_s\mathrm{E}_s - \delta_0\epsilon_2$, $m_2 = \delta\mathrm{E}_s + \delta_0\epsilon_1$, $m_3 = \delta\epsilon_2 + \epsilon_1\Delta_s$, and $m_4 = \Gamma_s - pq_1(\alpha\lambda)/(\mu n)$. Now for $i = 2, 3$ define the quantities

$$\begin{aligned} c_1(i) &= \mu + \frac{\alpha}{n}y_i^e \\ c_2(i) &= \Phi_s + \frac{\alpha_2}{n}y_i^e \\ c_3(i) &= q_3\frac{\alpha_2}{n}y_i^e + q_2\beta \\ c_4(i) &= (1-q_3)\frac{\alpha_2}{n}y_i^e + (1-q_2)\beta \\ c_5(i) &= \Gamma_s - pq_1\frac{\alpha}{n}x_i^e - q_3\frac{\alpha_2}{n}z_i^e \\ c_6(i) &= \gamma + p(1-q_1)\frac{\alpha}{n}x_i^e + (1-q_3)\frac{\alpha_2}{n}z_i^e \\ c_7(i) &= (1-p)\frac{\alpha}{n}x_i^e - \frac{\alpha_2}{n}z_i^e, \end{aligned}$$

where $\mathbf{e}_i = (x_i^e, y_i^e, z_i^e, w_i^e, u_i^e)$ for $i = 2, 3$ are the equilibrium points as defined in (6.6) and (6.7). Using these formulae, the coefficients M_{ij} for $i = 2, 3$ and $j = 1, \ldots, 5$, of the characteristic polynomial $P_i(\tau)$ of $DF(\mathbf{e}_i)$ (see equation 6.12) are the following:

$$\begin{aligned}
M_{i1} &= c_1(i) + \Delta_s + \mathrm{E}_s + c_2(i) + c_5(i)\\
M_{i2} &= c_1(i)[\Delta_s + \mathrm{E}_s + c_2(i) + c_5(i)] - c_7(i)c_3(i) + m_1 + (\Delta_s + \mathrm{E}_s)[c_2(i) + c_5(i)]\\
&\quad + c_2(i)c_5(i) - \epsilon_1\gamma_0 - \delta c_6(i) + \left(\frac{\alpha}{n}\right)^2 pq_1 x_i^e y_i^e\\
M_{i3} &= c_1(i)\{-c_7(i)c_3(i) + m_1 + (\Delta_s + \mathrm{E}_s)[c_2(i) + c_5(i)] + c_2(i)c_5(i) - \delta c_6(i) - \epsilon_1\gamma_0\}\\
&\quad - (\Delta_s + \mathrm{E}_s)c_7(i)c_3(i) - \delta c_7(i)c_4(i) + (\Delta_s + \mathrm{E}_s)c_2(i)c_5(i) + m_1[c_2(i) + c_5(i)]\\
&\quad - c_6(i)[m_2 + \delta c_2(i)] - \gamma_0[m_3 + \epsilon_1 c_2(i)]\\
&\quad + \left(\frac{\alpha}{n}\right)^2 x_i^e y_i^e[(1-p)c_3(i) + pq_1c_2(i) + pq_1(\Delta_s + \mathrm{E}_s) + p(1-q_1)\delta]\\
M_{i4} &= c_1(i)\{(\Delta_s + \mathrm{E}_s)[c_2(i)c_5(i) - c_3(i)c_7(i)] - \delta c_7(i)c_4(i) + m_1[c_2(i) + c_5(i)]\\
&\quad - c_6(i)[m_2 + \delta c_2(i)] - \gamma_0[m_3 + \epsilon_1 c_2(i)]\} - m_1c_7(i)c_3(i) - m_2c_7(i)c_4(i)\\
&\quad + m_1c_2(i)c_5(i) - \gamma_0c_2(i)m_3 - m_2c_2(i)c_6(i) + \left(\frac{\alpha}{n}\right)^2 x_i^e y_i^e\{pq_1m_1\\
&\quad + (\Delta_s + \mathrm{E}_s)[(1-p)c_3(i) + pq_1c_2(i)] + (1-p)\delta c_4(i) + p(1-q_1)[\delta c_2(i) + m_2]\}\\
M_{i5} &= c_1(i)\{-m_1c_7(i)c_3(i) - m_2c_7(i)c_4(i) + m_1c_2(i)c_5(i) - m_3\gamma_0c_2(i) - m_2c_2(i)c_6(i)\}\\
&\quad + \left(\frac{\alpha}{n}\right)^2 x_i^e y_i^e\{(1-p)m_1c_3(i) + pq_1m_1c_2(i) + (1-p)m_2c_4(i) + p(1-q_1)m_2c_2(i)\}.
\end{aligned}$$

Numerical minimisation

The programs for the numerical minimisation were written in Fortran using the routines E04UCF and E04UEF of the NAG library (Mark 18A). E04UCF is essentially identical to the subroutine NPSOL described in Gill, Murray, Saunders & Wright (1986) and is used to solve the following minimisation problem

$$\text{Minimise } F(x) \text{ subject to } \left\{ \begin{array}{c} \mathbf{l}_1 \leq \mathbf{x} \leq \mathbf{u}_1 \\ \mathbf{l}_2 \leq A\mathbf{x} \leq \mathbf{u}_2 \\ \mathbf{l}_3 \leq c(\mathbf{x}) \leq \mathbf{u}_3 \end{array} \right\}, \qquad \text{(A.6)}$$

where the conditions are bound, linear, and nonlinear conditions, respectively. The basic structure of E04UCF involves major and minor iterations. The major iterations generate a sequence of iterates x_k that converge to x^*, a first-order Kuhn-Tucker point of the problem (A.6). At a typical major iteration, the new iterate x' is defined by $x' = x + ap$, where x is the current iterate, a is the step length, and p is the search direction. The search direction p is the solution to a quadratic subproblem (the minor

	D_{22}	D_{23}	D_{24}
λ	$2.783776\cdot10^{-10}$	$6.347017\cdot10^{-10}$	2.000000
μ	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	$2.000002\cdot10^{-02}$
μ_1	1.000000	1.000000	0.130000
μ_2	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	0.100000
p	0.267919	1.000000	$4.999996\cdot10^{-02}$
q_1	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	0.550000
α/n	1.000000	0.669613	0.100000
α_2/n	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$
β	$1.000000\cdot10^{-10}$	1.000000	$1.000000\cdot10^{-10}$
δ	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	$1.999998\cdot10^{-02}$
ϵ_1	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	$1.499995\cdot10^{-02}$
γ_0	$1.000000\cdot10^{-10}$	1.000000	$6.600000\cdot10^{-02}$
ϵ_2	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	$1.499999\cdot10^{-02}$
δ_0	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	$6.599999\cdot10^{-02}$
q_2	0.552488	$1.000000\cdot10^{-10}$	0.550000
q_3	$1.000000\cdot10^{-10}$	$1.000000\cdot10^{-10}$	0.550000
Obj. Val.	$0.812823\cdot10^{-09}$	$0.645001\cdot10^{-08}$	$0.596063\cdot10^{-09}$

Table A.3: Results of the numerical minimisation of the principal minors D_{22}, D_{23}, and D_{24} over the whole parameter space with $\mathcal{R}_0 > 1$. The last row of the table gives the optimal values of the objective functions D_{22}, D_{23}, D_{24} and the other rows the values of the parameters at the point where these optimal values are achieved.

iterations of E04UCF are for the solution of this subproblem). The major iteration proceeds by determining a step length a that produces a "sufficient decrease" in an augmented Lagrangian merit function. The method used by E04UCF is described in detail in the manual of the NAG library (see also Gill et al. 1986).

For our problem, there are three functions to be minimised: $f_1(\mathbf{h}) = D_{22}$, $f_2(\mathbf{h}) = D_{23}$, $f_3(\mathbf{h}) = D_{24}$, as functions of the parameters (i.e. $\mathbf{h}$ is a vector with elements the parameters, μ, μ_1, μ_2, α, and so on). The programs were run separately minimising first under the condition $\mathcal{R}_0 > 1$ and then under the conditions $\{\mathcal{R}_0 < 1$ and $\mathcal{R}_1 > 1\}$. Only the minimum of D_{24} when $\mathcal{R}_0 > 1$ was found without floating underflow and that was positive. For all the other cases, floating underflow occurred during the process of finding the minimum value, but the optimal solution found was positive. Table A.3 shows the optimal solutions (the values of the objective functions D_{22}, D_{23}, D_{24}) and the values of the parameters at the point where these values are achieved for $\mathcal{R}_0 > 1$.

A.3.2 The stochastic model

The probabilities $p_{\mathbf{x}}(t) = p(x, y, z, w, u; t)$, as defined in (6.17), satisfy the following equations:

$$\begin{aligned}
&\frac{dp(x,y,z,w,u)}{dt} \\
&= \lambda p(x-1,y,z,w,u) + \mu(x+1)p(x+1,y,z,w,u) \\
&\quad + (\mu+\mu_1)(y+1)p(x,y+1,z,w,u) + \mu(z+1)p(x,y,z+1,w,u) \\
&\quad + (\mu+\mu_2)(w+1)p(x,y,z,w+1,u) + \mu(u+1)p(x,y,z,w,u+1) \\
&\quad + \frac{\alpha}{n}(x+1)[pq_1(y-1)p(x+1,y-1,z,w,u) \qquad \text{(A.7)} \\
&\quad + (1-p)yp(x+1,y,z-1,w,u) + p(1-q_1)yp(x+1,y,z,w-1,u)] \\
&\quad + q_2\beta(z+1)p(x,y-1,z+1,w,u) + (1-q_2)\beta(z+1)p(x,y,z+1,w-1,u) \\
&\quad + \frac{\alpha_2}{n}(z+1)[q_3(y-1)p(x,y-1,z+1,w,u) + (1-q_3)yp(x,y,z+1,w-1,u)] \\
&\quad + \delta(w+1)p(x,y-1,z,w+1,u) \\
&\quad + \gamma_0(y+1)p(x,y+1,z,w,u-1) + \delta_0(w+1)p(x,y,z,w+1,u-1) \\
&\quad + \epsilon_1(u+1)p(x,y-1,z,w,u+1) + \epsilon_2(u+1)p(x,y,z,w-1,u+1) \\
&\quad - (\lambda + \mu x + \frac{\alpha}{n}xy + \frac{\alpha_2}{n}yz + \Gamma_s y + \Phi_s z + \Delta_s w + \mathrm{E}_s u)p(x,y,z,w,u),
\end{aligned}$$

where for simplicity the dependence on t was suspended in the right-hand side of (A.7). The terms Γ_s, Δ_s, Φ_s, and E_s are defined in (6.2). Equation (A.7) holds for all $(x,y,z,w,u) \in \mathcal{S}$ and $p_{\mathbf{x}}(t) = 0$ for any $\mathbf{x} \notin \mathcal{S}$. The initial conditions are $p_{\mathbf{x}_0}(0) = 1$ and $p_{\mathbf{x}}(0) = 0$ for any $\mathbf{x} \neq \mathbf{x}_0$, where $\mathcal{S} = \mathbb{Z}_+^5$ and $\mathbf{x}_0 \in \mathcal{S}_0$ as defined in (6.3).

The variances of X, Y, Z, W, and U satisfy the following differential equations:

$$\begin{aligned}
\frac{d\mathrm{Var}[X(t)]}{dt} &= \lambda + \mu\mathrm{E}[X] - 2\mu\mathrm{Var}[X] + \frac{\alpha}{n}(1+2\mathrm{E}[X])\mathrm{E}[XY] - 2\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
\frac{d\mathrm{Var}[Y(t)]}{dt} &= \Gamma_s\mathrm{E}[Y] + q_2\beta\mathrm{E}[Z] + \delta\mathrm{E}[W] + \epsilon_1\mathrm{E}[U] - 2\Gamma_s\mathrm{Var}[Y] + 2\delta\mathrm{Cov}[Y,W] \\
&\quad + 2q_2\beta\mathrm{Cov}[Y,Z] + 2\epsilon_1\mathrm{Cov}[Y,U] + pq_1\frac{\alpha}{n}(1-2\mathrm{E}[Y])\mathrm{E}[XY] \\
&\quad + q_3\frac{\alpha_2}{n}(1-2\mathrm{E}[Y])\mathrm{E}[YZ] + 2pq_1\frac{\alpha}{n}\mathrm{E}[XY^2] + 2q_3\frac{\alpha_2}{n}\mathrm{E}[Y^2Z] \\
\frac{d\mathrm{Var}[Z(t)]}{dt} &= \Phi_s\mathrm{E}[Z] - 2\Phi_s\mathrm{Var}[Z] + (1-p)\frac{\alpha}{n}(1-2\mathrm{E}[Z])\mathrm{E}[XY] \\
&\quad + \frac{\alpha_2}{n}(1+2\mathrm{E}[Z])\mathrm{E}[YZ] + 2(1-p)\frac{\alpha}{n}\mathrm{E}[XYZ] - 2\frac{\alpha_2}{n}\mathrm{E}[YZ^2]
\end{aligned}$$

$$\begin{aligned}
\frac{d\mathrm{Var}[W(t)]}{dt} &= (1-q_2)\beta \mathrm{E}[Z] + \Delta_s \mathrm{E}[W] + \epsilon_2 \mathrm{E}[U] - 2\Delta_s \mathrm{Var}[W] \\
&\quad + 2\epsilon_2 \mathrm{Cov}[W,U] + 2(1-q_2)\beta \mathrm{Cov}[Z,W] \\
&\quad + p(1-q_1)\frac{\alpha}{n}(1-2\mathrm{E}[W])\mathrm{E}[XY] + (1-q_3)\frac{\alpha_2}{n}(1-2\mathrm{E}[W])\mathrm{E}[YZ] \\
&\quad + 2p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XYW] + 2(1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZW] \\
\frac{d\mathrm{Var}[U(t)]}{dt} &= \gamma_0(\mathrm{E}[Y] + 2\mathrm{Cov}[Y,U]) + \delta_0(\mathrm{E}[W] + 2\mathrm{Cov}[W,U]) \\
&\quad + \mathrm{E}_s(\mathrm{E}[U] - 2\mathrm{Var}[U]),
\end{aligned}$$

where the terms $\mathrm{E}[XY]$ and $\mathrm{E}[YZ]$ can be expressed as $\mathrm{E}[XY] = \mathrm{Cov}[X,Y] + \mathrm{E}[X]\mathrm{E}[Y]$ and $\mathrm{E}[YZ] = \mathrm{Cov}[Y,Z] + \mathrm{E}[Y]\mathrm{E}[Z]$. The covariances of X, Y, Z, W, and U satisfy the following differential equations:

$$\begin{aligned}
\frac{d\mathrm{Cov}[X,Y]}{dt} &= q_2\beta \mathrm{Cov}[X,Z] + \delta \mathrm{Cov}[X,W] + \epsilon_1 \mathrm{Cov}[X,U] - (\Gamma_s + \mu)\mathrm{Cov}[X,Y] \\
&\quad - \frac{\alpha}{n}(pq_1 + pq_1\mathrm{E}[X] - \mathrm{E}[Y])\mathrm{E}[XY] - q_3\frac{\alpha_2}{n}\mathrm{E}[X]\mathrm{E}[YZ] + pq_1\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
&\quad - \frac{\alpha}{n}\mathrm{E}[XY^2] + q_3\frac{\alpha_2}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[X,Z]}{dt} &= -(\Phi_s + \mu)\mathrm{Cov}[X,Z] - \frac{\alpha}{n}(1 - p + (1-p)\mathrm{E}[X] - \mathrm{E}[Z])\mathrm{E}[XY] \\
&\quad + \frac{\alpha_2}{n}\mathrm{E}[X]\mathrm{E}[YZ] + (1-p)\frac{\alpha}{n}\mathrm{E}[X^2Y] - \left(\frac{\alpha}{n} + \frac{\alpha_2}{n}\right)\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[X,W]}{dt} &= (1-q_2)\beta \mathrm{Cov}[X,Z] + \epsilon_2 \mathrm{Cov}[X,U] - \frac{\alpha}{n}\mathrm{E}[XYW] \\
&\quad - (\Delta_s + \mu)\mathrm{Cov}[X,W] - \frac{\alpha}{n}(p(1-q_1) + p(1-q_1)\mathrm{E}[X] - \mathrm{E}[W])\mathrm{E}[XY] \\
&\quad - (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[X]\mathrm{E}[YZ] + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[X^2Y] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[X,U]}{dt} &= \gamma_0 \mathrm{Cov}[X,Y] + \delta_0 \mathrm{Cov}[X,W] - (\mathrm{E}_s + \mu)\mathrm{Cov}[X,U] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[U]\mathrm{E}[XY] - \frac{\alpha}{n}\mathrm{E}[XYU] \\
\frac{d\mathrm{Cov}[Y,Z]}{dt} &= -q_2\beta \mathrm{E}[Z] + q_2\beta \mathrm{Var}[Z] + \delta \mathrm{Cov}[Z,W] + \epsilon_1 \mathrm{Cov}[Z,U] - (\Gamma_s + \Phi_s)\mathrm{Cov}[Y,Z] \\
&\quad - \frac{\alpha}{n}\{(1-p)\mathrm{E}[Y] + pq_1\mathrm{E}[Z]\}\mathrm{E}[XY] - \frac{\alpha_2}{n}(q_3 + q_3\mathrm{E}[Z] - \mathrm{E}[Y])\mathrm{E}[YZ] \\
&\quad + pq_1\frac{\alpha}{n}\mathrm{E}[XYZ] + (1-p)\frac{\alpha}{n}\mathrm{E}[XY^2] - \frac{\alpha_2}{n}\mathrm{E}[Y^2Z] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZ^2]
\end{aligned}$$

$$
\begin{aligned}
\frac{d\mathrm{Cov}[Y,W]}{dt} &= -\delta\mathrm{E}[W] + \delta\mathrm{Var}[W] + (1-q_2)\beta\mathrm{Cov}[Y,Z] \\
&\quad + \epsilon_2\mathrm{Cov}[Y,U] - (\Gamma_s + \Delta_s)\mathrm{Cov}[Y,W] + \epsilon_1\mathrm{Cov}[W,U] + q_2\beta\mathrm{Cov}[Z,W] \\
&\quad - \frac{\alpha}{n}\{p(1-q_1)\mathrm{E}[Y] + pq_1\mathrm{E}[W]\}\mathrm{E}[XY] + pq_1\frac{\alpha}{n}\mathrm{E}[XYW] \\
&\quad - \frac{\alpha_2}{n}\{(1-q_3)\mathrm{E}[Y] + q_3\mathrm{E}[W]\}\mathrm{E}[YZ] + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XY^2] \\
&\quad + q_3\frac{\alpha_2}{n}\mathrm{E}[YZW] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[Y^2Z] \\
\frac{d\mathrm{Cov}[Y,U]}{dt} &= -\epsilon_1\mathrm{E}[U] - \gamma_0\mathrm{E}[Y] + \epsilon_1\mathrm{Var}[U] + \gamma_0\mathrm{Var}[Y] + q_2\beta\mathrm{Cov}[Z,U] \\
&\quad + \delta\mathrm{Cov}[W,U] - (\Gamma_s + \mathrm{E}_s)\mathrm{Cov}[Y,U] + \delta_0\mathrm{Cov}[Y,W] \\
&\quad - pq_1\frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[U] - q_3\frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[U] + pq_1\frac{\alpha}{n}\mathrm{E}[XYU] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZU] \\
\frac{d\mathrm{Cov}[Z,W]}{dt} &= -(1-q_2)\beta(\mathrm{E}[Z] - \mathrm{Var}[Z]) - (\Delta_s + \Phi_s)\mathrm{Cov}[Z,W] \\
&\quad + \epsilon_2\mathrm{Cov}[Z,U] - \frac{\alpha}{n}\{p(1-q_1)\mathrm{E}[Z] + (1-p)\mathrm{E}[W]\}\mathrm{E}[XY] \\
&\quad + (1-p)\frac{\alpha}{n}\mathrm{E}[XYW] + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XYZ] - \frac{\alpha_2}{n}\mathrm{E}[YZW] \\
&\quad + \frac{\alpha_2}{n}\{\mathrm{E}[W] - (1-q_3)\mathrm{E}[Z] - (1-q_3)\}\mathrm{E}[YZ] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ^2] \\
\frac{d\mathrm{Cov}[Z,U]}{dt} &= \gamma_0\mathrm{Cov}[Y,Z] + \delta_0\mathrm{Cov}[Z,W] - (\mathrm{E}_s + \Phi_s)\mathrm{Cov}[Z,U] - (1-p)\frac{\alpha}{n}\mathrm{E}[U]\mathrm{E}[XY] \\
&\quad + \frac{\alpha_2}{n}\mathrm{E}[U]\mathrm{E}[YZ] + (1-p)\frac{\alpha}{n}\mathrm{E}[XYU] - \frac{\alpha_2}{n}\mathrm{E}[YZU] \\
\frac{d\mathrm{Cov}[W,U]}{dt} &= -\delta_0(\mathrm{E}[W] - \mathrm{Var}[W]) - \epsilon_2(\mathrm{E}[U] - \mathrm{Var}[U]) + \gamma_0\mathrm{Cov}[Y,W] \\
&\quad - (\Delta_s + \mathrm{E}_s)\mathrm{Cov}[W,U] + (1-q_2)\beta\mathrm{Cov}[Z,U] \\
&\quad - p(1-q_1)\frac{\alpha}{n}\mathrm{E}[U]\mathrm{E}[XY] - (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[U]\mathrm{E}[YZ] \\
&\quad + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XYU] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZU].
\end{aligned}
$$

A.3.3 Numerical results from simulations

For the simulation of this model there are 20 different kinds of events that can occur. The events and their rates are defined in Table A.4. The simulations were implemented as described in Section A.1.2, where the vector $\mathbf{X}_c = (X_c, Y_c, Z_c, W_c, U_c)$ gives the current value of the state of the process $\mathbf{X} = (X, Y, Z, W, U)$ and the initial value at time $t = 0$ is $\mathbf{X}_0 = (x_0, y_0, z_0, w_0, u_0)$. For each set of initial conditions and parameter values the simulation was repeated $R = 10^4$ times.

Event	Rate	Event	Rate
birth (of X)	λ	$X \to Y$	$pq_1\frac{\alpha}{n}XY$
death of X	μX	$X \to Z$	$(1-p)\frac{\alpha}{n}XY$
normal death of Y	μY	$X \to W$	$p(1-q_1)\frac{\alpha}{n}XY$
TB death of Y	$\mu_1 Y$	$Z \to Y$ (reactivation)	$q_2\beta Z$
death of Z	μZ	$Z \to Y$ (reinfection)	$q_3\frac{\alpha_2}{n}YZ$
normal death of W	μW	$Z \to W$ (reactivation)	$(1-q_2)\beta Z$
TB death of W	$\mu_2 W$	$Z \to W$ (reinfection)	$(1-q_3)\frac{\alpha_2}{n}YZ$
death of U	μU	$W \to Y$	δW
$Y \to U$	$\gamma_0 Y$	$U \to Y$	$\epsilon_1 U$
$W \to U$	$\delta_0 W$	$U \to W$	$\epsilon_2 U$

Table A.4: Events for the simulations of model Zeus

Epidemiology

For the calculation of the estimates of the epidemiological indices of interest and the corresponding variance at a particular time point t we used the following formulae:

$$\begin{aligned}\hat{\mu}_e &= \frac{1}{R}\sum_{i=1}^{R}\frac{d_i}{N_i}10^5 \\ \hat{\sigma}_e^2 &= \frac{1}{R-1}\left[\sum_{i=1}^{R}\left(\frac{d_i}{N_i}10^5\right)^2 - R\hat{\mu}_e^2\right],\end{aligned} \tag{A.8}$$

where R is the number of individual simulation runs ($R = 10^4$). Let NB_i and NE_i denote the total population sizes in the beginning and the end of year t, respectively, in the i-th run. Then the d_i and N_i are defined as follows:

- *for the mortality:* d_i is the number of excess deaths due to TB during year t in the i-th run and $N_i = NB_i$. For the separate rates, mortality of infectious and mortality of non-infectious, the d_i counts either the excess deaths from the Y or the W class, respectively.
- *for the risk of infection:* d_i is the number of new infections (i.e. transitions from X to Y, from X to Z, and from X to W) that occurred during year t in the i-th run and $N_i = (NB_i + NE_i)/2$.
- *for the risk of reinfection:* d_i is the number of reinfections during year t in the i-th run and $N_i = (NB_i + NE_i)/2$.
- *for the incidence of infectious TB:* d_i is the number of transitions from the X or the Z class to Y, during year t in the i-th run and $N_i = (NB_i + NE_i)/2$.
- *for the incidence of non-infectious TB:* d_i is the number of transitions from the X or the Z class to W, during year t in the i-th run and $N_i = (NB_i + NE_i)/2$.

• *for the prevalence of infectious and non-infectious TB:* d_i is the value of Y and W, respectively, at the end of year t in the i-th run and $N_i = NE_i$.

• *for the prevalence of TB infection:* d_i is the value of $Y + Z + W + U$ at the end of year t in the i-th run and $N_i = NE_i$.

• *for the respective conditional factors:* The d_i and N_i are as defined for each index above, the only difference being that R is equal to the number of runs for which the epidemic had not died out up to time t.

Moments, marginal distributions, and statistics for the extinction time

The estimates for the means and standard deviations of H at time t and the covariance of H and H' at time t, for $H, H' = X, Y, Z, W, U$, were calculated from the equations A.2, A.3, and A.4, where H_i, H_i' are the values of H, H', respectively, at time t in the i-th run. The conditional mean of H was calculated with the same formula as for the mean, but with R equal to the number of runs in which the epidemic had not died out by time t.

The marginal distribution of $H = X, Y, Z, W, U$ was calculated from the formula $\mathrm{P}[H(t) = k] = \mathcal{M}_k(t)/R$, where $\mathcal{M}_k(t)$ is the number of runs in which $H(t) = k$, with $t = 1, 2, \ldots$ and $k = 0, 1, \ldots$.

For the extinction time the simulations were terminated at a time point large enough such that all 10^4 runs ended with extinction of the infection by that point. Consequently the $R = 10^4$ runs yielded a sample of R independent values $\tau_1, \tau_2, \ldots, \tau_R$, where τ_i is the extinction time in the i-th run. The sample $\tau = \{\tau_1, \tau_2, \ldots, \tau_R\}$ was then used to calculate the statistics presented in Section 6.3.7.

A.3.4 Linear approximation

The probabilities $p^{\ell}_{ywu}(t)$ satisfy the following equations

$$
\begin{aligned}
\frac{dp^{\ell}_{ywu}(t)}{dt} &= [L_1(t)(y-1) + q_2\beta z(t)]p^{\ell}_{y-1,w,u} + [L_2(t)y + (1-q_2)\beta z(t)]p^{\ell}_{y,w-1,u} \\
&+ (\mu+\mu_1)(y+1)p^{\ell}_{y+1,w,u} + \delta(w+1)p^{\ell}_{y-1,w+1,u} \\
&+ (\mu+\mu_2)(w+1)p^{\ell}_{y,w+1,u} + \gamma_0(y+1)p^{\ell}_{y+1,w,u-1} \\
&+ \epsilon_1(u+1)p^{\ell}_{y-1,w,u+1} + \epsilon_2(u+1)p^{\ell}_{y,w-1,u+1} \\
&+ \delta_0(w+1)p^{\ell}_{y,w+1,u-1} + \mu(u+1)p^{\ell}_{y,w,u+1} \\
&- \left\{\beta z(t) + \Delta_s w + \mathrm{E}_s u + \left[p\frac{\alpha}{n}x(t) + \frac{\alpha_2}{n}z(t) + \Gamma_s\right] y\right\} p^{\ell}_{y,w,u},
\end{aligned}
\tag{A.9}
$$

where the $L_1(t)$, $L_2(t)$ are as defined in (6.27).

From equation (6.26) for the probability generating function of Y_ℓ, W_ℓ, U_ℓ we obtain the following equations for the variances and covariances:

$$
\begin{aligned}
\frac{d\mathrm{Var}[Y_\ell]}{dt} &= [\Gamma_s + L_1(t)]\mathrm{E}[Y_\ell] + 2[L_1(t) - \Gamma_s]\mathrm{Var}[Y_\ell] + 2\delta\mathrm{Cov}[Y_\ell, W_\ell] \\
&+ 2\epsilon_1\mathrm{Cov}[Y_\ell, U_\ell] + \delta\mathrm{E}[W_\ell] + \epsilon_1\mathrm{E}[U_\ell] + q_2\beta z(t) \\
\frac{d\mathrm{Var}[W_\ell]}{dt} &= -2\Delta_s\mathrm{Var}[W_\ell] + 2L_2(t)\mathrm{Cov}[Y_\ell, W_\ell] + 2\epsilon_2\mathrm{Cov}[W_\ell, U_\ell] + \Delta_s\mathrm{E}[W_\ell] \\
&+ L_2(t)\mathrm{E}[Y_\ell] + (1-q_2)\beta z(t) + \epsilon_2\mathrm{E}[U_\ell] \\
\frac{d\mathrm{Var}[U_\ell]}{dt} &= \mathrm{E}_s\mathrm{E}[U_\ell] - 2\mathrm{E}_s\mathrm{Var}[U_\ell] + \gamma_0\mathrm{E}[Y_\ell] + \delta_0\mathrm{E}[W_\ell] + 2\gamma_0\mathrm{Cov}[Y_\ell, U_\ell] \\
&+ 2\delta_0\mathrm{Cov}[W_\ell, U_\ell] \\
\frac{d\mathrm{Cov}[Y_\ell, W_\ell]}{dt} &= [L_1(t) - \Gamma_s - \Delta_s]\mathrm{Cov}[Y_\ell, W_\ell] - \delta\mathrm{E}[W_\ell] + \epsilon_1\mathrm{Cov}[W_\ell, U_\ell] \\
&+ L_2(t)\mathrm{Var}[Y_\ell] + \epsilon_2\mathrm{Cov}[Y_\ell, U_\ell] + \delta\mathrm{Var}[W_\ell] \\
\frac{d\mathrm{Cov}[Y_\ell, U_\ell]}{dt} &= [L_1(t) - \Gamma_s - \mathrm{E}_s]\mathrm{Cov}[Y_\ell, U_\ell] - \epsilon_1\mathrm{E}[U_\ell] + \delta_0\mathrm{Cov}[Y_\ell, W_\ell] + \epsilon_1\mathrm{Var}[U_\ell] \\
&+ \gamma_0\mathrm{Var}[Y_\ell] - \gamma_0\mathrm{E}[Y_\ell] + \delta\mathrm{Cov}[W_\ell, U_\ell] \\
\frac{d\mathrm{Cov}[W_\ell, U_\ell]}{dt} &= L_2(t)\mathrm{Cov}[Y_\ell, U_\ell] - \epsilon_2\mathrm{E}[U_\ell] + \gamma_0\mathrm{Cov}[Y_\ell, W_\ell] - (\Delta_s + \mathrm{E}_s)\mathrm{Cov}[W_\ell, U_\ell] \\
&+ \epsilon_2\mathrm{Var}[U_\ell] + \delta_0\mathrm{Var}[W_\ell] - \delta_0\mathrm{E}[W_\ell].
\end{aligned}
$$

A.4 Model Clio

A.4.1 The deterministic model

Define $\pi_1, \pi_2, \ldots, \pi_8$ as follows:

$$\begin{aligned}
\pi_1 &= (1-p)(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + p(1-q_1)\theta_2 \mathrm{E}_s \\
\pi_2 &= \ -(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_3)\theta_2 \mathrm{E}_s \\
\pi_3 &= \theta_1(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + \theta_2\gamma_0\epsilon_2 \\
\pi_4 &= \ -\Phi_s(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_2)\beta\theta_2 \mathrm{E}_s \\
\pi_5 &= p[q_1(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_1)(\delta \mathrm{E}_s + \delta_0\epsilon_1)] \\
\pi_6 &= q_3(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_3)(\delta \mathrm{E}_s + \delta_0\epsilon_1) \\
\pi_7 &= \frac{1}{\mathrm{E}_s}[-(\Gamma_s \mathrm{E}_s - \gamma_0\epsilon_1)(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + \gamma_0\epsilon_2(\delta \mathrm{E}_s + \delta_0\epsilon_1)] \\
\pi_8 &= \beta[q_2(\Delta_s \mathrm{E}_s - \delta_0\epsilon_2) + (1-q_2)(\delta \mathrm{E}_s + \delta_0\epsilon_1)],
\end{aligned}$$

where the Γ_s, Δ_s, E_s, and Φ_s are as defined in (7.6). Now, define φ_1, φ_2, and φ_3 as

$$\begin{aligned}
\varphi_1 &= \frac{\frac{\alpha}{n}\pi_2}{\pi_4\pi_6 - \pi_2\pi_8}\left[\frac{\alpha}{n}(\pi_4\pi_5 - \pi_1\pi_8) - \mu\frac{\alpha_2}{n}(\pi_2\pi_5 - \pi_1\pi_6)\right] \\
\varphi_2 &= \frac{\pi_2}{\pi_4\pi_6 - \pi_2\pi_8}\left[\frac{\alpha}{n}(\pi_4\pi_7 - \pi_3\pi_8) - \mu\frac{\alpha_2}{n}(\pi_2\pi_7 - \pi_3\pi_6) + \lambda\frac{\alpha}{n}\frac{\alpha_2}{n}(\pi_2\pi_5 - \pi_1\pi_6)\right] \\
\varphi_3 &= \lambda\frac{\alpha_2}{n}\pi_2\frac{\pi_2\pi_7 - \pi_3\pi_6}{\pi_4\pi_6 - \pi_2\pi_8}.
\end{aligned}$$

With the notation above the equilibrium points $\mathbf{e}_i = (x_i^e, y_i^e, z_i^e, w_i^e, u_i^e)$ for $i = 2, 3$ are defined from the following relationships:

$$x_2^e = \frac{-\varphi_2 - \sqrt{\mathcal{D}}}{2\varphi_1} \quad \text{and} \quad x_3^e = \frac{-\varphi_2 + \sqrt{\mathcal{D}}}{2\varphi_1},$$

where $\mathcal{D} = \varphi_2^2 - 4\varphi_1\varphi_3$, and for $i = 2, 3$

$$\begin{aligned}
y_i^e &= \frac{\lambda - \mu x_i^e}{\alpha x_i^e / n} \\
z_i^e &= -\frac{\pi_1 \frac{\alpha}{n} x_i^e y_i^e + \pi_3 y_i^e}{\pi_2 \frac{\alpha_2}{n} y_i^e + \pi_4} \\
w_i^e &= \frac{\mathrm{E}_s}{\Delta_s \mathrm{E}_s - \delta_0\epsilon_2}[p(1-q_1)\frac{\alpha}{n}x_i^e y_i^e + (1-q_3)\frac{\alpha_2}{n}y_i^e z_i^e + \frac{\gamma_0\epsilon_2}{\mathrm{E}_s}y_i^e + (1-q_2)\beta z_i^e] \\
u_i^e &= \frac{1}{\mathrm{E}_s}(\gamma_0 y_i^e + \delta_0 w_i^e).
\end{aligned}$$

The value of $\mathcal{R}_0$ is calculated from Definition 3.1. In this model the infected individuals can be in four different states: inactive, infectious, and non-infectious cases, and recovered. Then $\mathcal{R}_0$ is the dominant eigenvalue of the matrix $\mathsf{K} = -\mathsf{T}(\mathsf{S}-\mathsf{D})^{-1}$, where

$$\mathsf{S} = \begin{bmatrix} -(\gamma_0+\theta_1) & q_2\beta & \delta & \epsilon_1 \\ \theta_1 & -\beta & \theta_2 & 0 \\ 0 & (1-q_2)\beta & -(\delta+\delta_0+\theta_2) & \epsilon_2 \\ \gamma_0 & 0 & \delta_0 & -(\epsilon_1+\epsilon_2) \end{bmatrix},$$

D is the diagonal matrix $\mathsf{D} = \text{diag}\{\mu+\mu_1, \mu, \mu+\mu_2, \mu\}$, and the matrix T is as defined in (A.5).

A.4.2 The stochastic model

The Kolmogorov forward equations for the probabilities $p_{\mathbf{x}}(t) = p(x,y,z,w,u;t)$ are

$$\begin{aligned}
\frac{dp(x,y,z,w,u)}{dt} &= \lambda p(x-1,y,z,w,u) + \mu(x+1)p(x+1,y,z,w,u) \\
&+ (\mu+\mu_1)(y+1)p(x,y+1,z,w,u) + \mu(z+1)p(x,y,z+1,w,u) \\
&+ (\mu+\mu_2)(w+1)p(x,y,z,w+1,u) + \mu(u+1)p(x,y,z,w,u+1) \\
&+ \frac{\alpha}{n}(x+1)[pq_1(y-1)p(x+1,y-1,z,w,u) \\
&+ (1-p)yp(x+1,y,z-1,w,u) + p(1-q_1)yp(x+1,y,z,w-1,u)] \\
&+ q_2\beta(z+1)p(x,y-1,z+1,w,u) + (1-q_2)\beta(z+1)p(x,y,z+1,w-1,u) \\
&+ \frac{\alpha_2}{n}(z+1)[q_3(y-1)p(x,y-1,z+1,w,u) + (1-q_3)yp(x,y,z+1,w-1,u)] \\
&+ \theta_1(y+1)p(x,y+1,z-1,w,u) + \theta_2(w+1)p(x,y,z-1,w+1,u) \\
&+ \delta(w+1)p(x,y-1,z,w+1,u) \\
&+ \gamma_0(y+1)p(x,y+1,z,w,u-1) + \delta_0(w+1)p(x,y,z,w+1,u-1) \\
&+ \epsilon_1(u+1)p(x,y-1,z,w,u+1) + \epsilon_2(u+1)p(x,y,z,w-1,u+1) \\
&- (\lambda + \mu x + \frac{\alpha}{n}xy + \frac{\alpha_2}{n}yz + \Gamma_s y + \Phi_s z + \Delta_s w + \mathrm{E}_s u)p(x,y,z,w,u),
\end{aligned}$$

where for simplicity, the dependence on t in the terms $p(x,y,z,w,u;t)$ has been suppressed, and the Γ_s, Δ_s, Φ_s, and E_s are defined in (7.6).

The variances of X, Y, Z, W, and U satisfy the following differential equations:

$$\frac{d\text{Var}[X(t)]}{dt} = \lambda + \mu\text{E}[X] + 2\mu\text{Var}[X] + \frac{\alpha}{n}\text{E}[XY](1+2\text{E}[X]) - 2\frac{\alpha}{n}\text{E}[X^2Y]$$

$$\begin{aligned}
\frac{d\mathrm{Var}[Y(t)]}{dt} &= \Gamma_s\mathrm{E}[Y] + q_2\beta\mathrm{E}[Z] + \delta\mathrm{E}[W] + \epsilon_1\mathrm{E}[U] - 2\Gamma_s\mathrm{Var}[Y] + 2\delta\mathrm{Cov}[Y,W] \\
&\quad + 2\epsilon_1\mathrm{Cov}[Y,U] + 2q_2\beta\mathrm{Cov}[Y,Z] + pq_1\frac{\alpha}{n}\mathrm{E}[XY](1-2\mathrm{E}[Y]) \\
&\quad + q_3\frac{\alpha_2}{n}\mathrm{E}[YZ](1-2\mathrm{E}[Y]) + 2q_3\frac{\alpha_2}{n}\mathrm{E}[Y^2Z] + 2pq_1\frac{\alpha}{n}\mathrm{E}[XY^2] \\
\frac{d\mathrm{Var}[Z(t)]}{dt} &= \theta_1\mathrm{E}[Y] + \Phi_s\mathrm{E}[Z] + \theta_2\mathrm{E}[W] - 2\Phi_s\mathrm{Var}[Z] + 2\theta_1\mathrm{Cov}[Y,Z] + 2\theta_2\mathrm{Cov}[Z,W] \\
&\quad + (1-p)\frac{\alpha}{n}\mathrm{E}[XY](1-2\mathrm{E}[Z]) + \frac{\alpha_2}{n}\mathrm{E}[YZ](1+2\mathrm{E}[Z]) - 2\frac{\alpha_2}{n}\mathrm{E}[YZ^2] \\
&\quad + 2(1-p)\frac{\alpha}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Var}[W(t)]}{dt} &= (1-q_2)\beta\mathrm{E}[Z] + \Delta_s\mathrm{E}[W] + \epsilon_2\mathrm{E}[U] - 2\Delta_s\mathrm{Var}[W] \\
&\quad + 2(1-q_2)\beta\mathrm{Cov}[Z,W] + 2\epsilon_2\mathrm{Cov}[W,U] \\
&\quad + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XY](1-2\mathrm{E}[W]) + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ](1-2\mathrm{E}[W]) \\
&\quad + 2p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XYW] + 2(1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZW] \\
\frac{d\mathrm{Var}[U(t)]}{dt} &= \gamma_0\mathrm{E}[Y] + \delta_0\mathrm{E}[W] + \mathrm{E}_s\mathrm{E}[U] - 2\mathrm{E}_s\mathrm{Var}[U] + 2\gamma_0\mathrm{Cov}[Y,U] + 2\delta_0\mathrm{Cov}[W,U].
\end{aligned}$$

The covariances of X, Y, Z, W, and U satisfy the following differential equations:

$$\begin{aligned}
\frac{d\mathrm{Cov}[X,Y]}{dt} &= -(\Gamma_s+\mu)\mathrm{Cov}[X,Y] + q_2\beta\mathrm{Cov}[X,Z] + \epsilon_1\mathrm{Cov}[X,U] + \delta\mathrm{Cov}[X,W] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY](\mathrm{E}[Y] - pq_1\mathrm{E}[X] - pq_1) - q_3\frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[X] - \frac{\alpha}{n}\mathrm{E}[XY^2] \\
&\quad + pq_1\frac{\alpha}{n}\mathrm{E}[X^2Y] + q_3\frac{\alpha_2}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[X,Z]}{dt} &= -(\Phi_s+\mu)\mathrm{Cov}[X,Z] + \theta_1\mathrm{Cov}[X,Y] + \theta_2\mathrm{Cov}[X,W] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY](\mathrm{E}[Z] - (1-p)\mathrm{E}[X] - (1-p)) + \frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[X] \\
&\quad - (\frac{\alpha}{n} + \frac{\alpha_2}{n})\mathrm{E}[XYZ] + (1-p)\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
\frac{d\mathrm{Cov}[X,W]}{dt} &= -(\Delta_s+\mu)\mathrm{Cov}[X,W] + (1-q_2)\beta\mathrm{Cov}[X,Z] + \epsilon_2\mathrm{Cov}[X,U] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY](\mathrm{E}[W] - p(1-q_1)\mathrm{E}[X] - p(1-q_1)) \\
&\quad - (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[X] - \frac{\alpha}{n}\mathrm{E}[XYW] + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
&\quad + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[XYZ] \\
\frac{d\mathrm{Cov}[X,U]}{dt} &= -(\mathrm{E}_s+\mu)\mathrm{Cov}[X,U] + \gamma_0\mathrm{Cov}[X,Y] + \delta_0\mathrm{Cov}[X,W] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[U] - \frac{\alpha}{n}\mathrm{E}[XYU]
\end{aligned}$$

$$
\begin{aligned}
\frac{d\mathrm{Cov}[Y,Z]}{dt} = {} & -\theta_1\mathrm{E}[Y] - q_2\beta\mathrm{E}[Z] + q_2\beta\mathrm{Var}[Z] + \theta_1\mathrm{Var}[Y] + \delta\mathrm{Cov}[Z,W] \\
& + \epsilon_1\mathrm{Cov}[Z,U] - (\Gamma_s + \Phi_s)\mathrm{Cov}[Y,Z] + \theta_2\mathrm{Cov}[Y,W] \\
& - \frac{\alpha}{n}\mathrm{E}[XY]((1-p)\mathrm{E}[Y] + pq_1\mathrm{E}[Z]) - \frac{\alpha_2}{n}\mathrm{E}[YZ](\mathrm{E}[Y] - q_3\mathrm{E}[Z] - q_3) \\
& + (1-p)\frac{\alpha}{n}\mathrm{E}[XY^2] + pq_1\frac{\alpha}{n}\mathrm{E}[XYZ] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZ^2] - \frac{\alpha_2}{n}\mathrm{E}[Y^2Z] \\
\frac{d\mathrm{Cov}[Y,W]}{dt} = {} & -\delta\mathrm{E}[W] + \delta\mathrm{Var}[W] + q_2\beta\mathrm{Cov}[Z,W] \\
& + \epsilon_1\mathrm{Cov}[W,U] - (\Gamma_s + \Delta_s)\mathrm{Cov}[Y,W] + (1-q_2)\beta\mathrm{Cov}[Y,Z] \\
& + \epsilon_2\mathrm{Cov}[Y,U] - \frac{\alpha}{n}\mathrm{E}[XY](p(1-q_1)\mathrm{E}[Y] + pq_1\mathrm{E}[W]) \\
& - \frac{\alpha_2}{n}\mathrm{E}[YZ]((1-q_3)\mathrm{E}[Y] + q_3\mathrm{E}[W]) + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XY^2] \\
& + pq_1\frac{\alpha}{n}\mathrm{E}[XYW] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZW] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[Y^2Z] \\
\frac{d\mathrm{Cov}[Y,U]}{dt} = {} & -\epsilon_1\mathrm{E}[U] - \gamma_0\mathrm{E}[Y] + \epsilon_1\mathrm{Var}[U] + \gamma_0\mathrm{Var}[Y] + q_2\beta\mathrm{Cov}[Z,U] \\
& + \delta\mathrm{Cov}[W,U] - (\Gamma_s + \mathrm{E}_s)\mathrm{Cov}[Y,U] + \delta_0\mathrm{Cov}[Y,W] - pq_1\frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[U] \\
& - q_3\frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[U] + pq_1\frac{\alpha}{n}\mathrm{E}[XYU] + q_3\frac{\alpha_2}{n}\mathrm{E}[YZU] \\
\frac{d\mathrm{Cov}[Z,W]}{dt} = {} & -\theta_2\mathrm{E}[W] - (1-q2)\beta\mathrm{E}[Z] + \theta_2\mathrm{Var}[W] + (1-q_2)\beta\mathrm{Var}[Z] \\
& + \theta_1\mathrm{Cov}[Y,W] - (\Delta_s + \Phi_s)\mathrm{Cov}[Z,W] + \epsilon_2\mathrm{Cov}[Z,U] \\
& - \frac{\alpha}{n}\mathrm{E}[XY](p(1-q_1)\mathrm{E}[Z] + (1-p)\mathrm{E}[W]) \\
& + \frac{\alpha_2}{n}\mathrm{E}[YZ](\mathrm{E}[W] - (1-q_3)\mathrm{E}[Z] - (1-q_3)) + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XYZ] \\
& + (1-p)\frac{\alpha}{n}\mathrm{E}[XYW] - \frac{\alpha_2}{n}\mathrm{E}[YZW] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ^2] \\
\frac{d\mathrm{Cov}[Z,U]}{dt} = {} & \theta_1\mathrm{Cov}[Y,U] + \theta_2\mathrm{Cov}[W,U] - (\Phi_s + \mathrm{E}_s)\mathrm{Cov}[Z,U] + \gamma_0\mathrm{Cov}[Y,Z] \\
& + \delta_0\mathrm{Cov}[Z,W] - (1-p)\frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[U] + \frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[U] \\
& + (1-p)\frac{\alpha}{n}\mathrm{E}[XYU] - \frac{\alpha_2}{n}\mathrm{E}[YZU] \\
\frac{d\mathrm{Cov}[W,U]}{dt} = {} & -\epsilon_2\mathrm{E}[U] - \delta_0\mathrm{E}[W] + \epsilon_2\mathrm{Var}[U] + \delta_0\mathrm{Var}[W] + (1-q_2)\beta\mathrm{Cov}[Z,U] \\
& - (\Delta_s + \mathrm{E}_s)\mathrm{Cov}[W,U] + \gamma_0\mathrm{Cov}[Y,W] \\
& - p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[U] - (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[U] \\
& + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[XYU] + (1-q_3)\frac{\alpha_2}{n}\mathrm{E}[YZU].
\end{aligned}
$$

A.4.3 Numerical results from simulations

For the simulation of model Clio there are 22 different kinds of events that can occur. The events and their rates are defined in Table A.5. For each set of parameter values the simulation was repeated $R = 10^4$ times.

Event	Rate	Event	Rate
birth (of X)	λ	$X \to Y$	$pq_1 \frac{\alpha}{n} XY$
death of X	μX	$X \to Z$	$(1-p)\frac{\alpha}{n} XY$
normal death of Y	μY	$X \to W$	$p(1-q_1)\frac{\alpha}{n} XY$
TB death of Y	$\mu_1 Y$	$Z \to Y$ (reactivation)	$q_2 \beta Z$
death of Z	μZ	$Z \to Y$ (reinfection)	$q_3 \frac{\alpha_2}{n} YZ$
normal death of W	μW	$Z \to W$ (reactivation)	$(1-q_2)\beta Z$
TB death of W	$\mu_2 W$	$Z \to W$ (reinfection)	$(1-q_3)\frac{\alpha_2}{n} YZ$
death of U	μU	$Y \to U$	$\gamma_0 Y$
$Y \to Z$	$\theta_1 Y$	$U \to Y$	$\epsilon_1 U$
$W \to Z$	$\theta_2 W$	$W \to U$	$\delta_0 W$
$W \to Y$	δW	$U \to W$	$\epsilon_2 U$

Table A.5: Events for the simulations of model Clio

For this model we assumed that at time $t = 0$, when chemotherapy is introduced, the epidemic has already reached the steady endemic level described by the quasi-stationary distribution of model Zeus (for the natural evolution of TB). Therefore, the vector of initial conditions, $\mathbf{X}_0 = (x_0, y_0, z_0, w_0, u_0)$, for model Clio must be a variate from the quasi-stationary distribution of model Zeus.

In order to implement that, 10^4 simulation runs of model Zeus were carried out with the parameter values of interest. Let $\mathbf{X}_i^{\mathcal{Z}}(t)$ denote the value of the vector $(X(t), Y(t), Z(t), W(t), U(t))$ at time t from the i-th simulation run of model Zeus, for $i = 1, 2, \ldots, 10000$, and similarly $\mathbf{X}_i^{\mathcal{C}}(t)$ for Clio.

The results from the simulations of model Zeus (see Sections 6.3.5 and 6.3.4) show that by time $t = 300$ the process has either died out or reached the steady endemic level. In particular the simulations that were carried out with a large value of y_0 (the initial number of infectious cases) all ended at the steady endemic level. So, the 10^4 runs of Zeus were carried out with the parameter values of interest and $\mathbf{X}_i^{\mathcal{Z}}(0) = (n - 10, 10, 0, 0, 0)$, which ensured that all 10^4 runs ended at the steady endemic level and hence each of the 10^4 vectors $\mathbf{X}_i^{\mathcal{Z}}(300)$ is a variate from the quasi-stationary distribution of Zeus. Then the initial conditions for the simulations of Clio where taken as $\mathbf{X}_i^{\mathcal{C}}(0) = \mathbf{X}_i^{\mathcal{Z}}(300)$, for $i = 1, 2, \ldots, 10000$, ensuring that the initial conditions of Clio are variates from the steady endemic level of the natural evolution of TB.

Similarly for the various statistics of interest, the final values calculated from the simulations of Zeus where used as the initial values for Clio. For instance, if $\mathcal{V}^{\mathcal{Z}}(t)$ and $\mathcal{V}^{\mathcal{C}}(t)$ denote the prevalence of infectious cases at time t for model Zeus and Clio, respectively, then we take $\mathcal{V}^{\mathcal{C}}(0) = \mathcal{V}^{\mathcal{Z}}(300)$.

Epidemiology

The estimates of the epidemiological indices and the corresponding variances at a particular time point t were calculated using formulae (A.8), where R is the number of individual simulation runs ($R = 10^4$) and the d_i and N_i are defined as for the equations (A.8). The percentage decline at time t for each index was calculated from the formula

$$\mathcal{PDF}(t) = \frac{\mathcal{F}(0) - \mathcal{F}(t)}{\mathcal{F}(0)} 100, \tag{A.10}$$

and the 95% confidence interval for the percentage decline from the formula

$$\frac{\mathcal{F}(0) - \mathcal{F}(t)}{\mathcal{F}(0)} 100 \pm 1.96 \frac{\mathcal{S}_{\mathcal{F}}(t)}{\sqrt{R}} \frac{100}{\mathcal{F}(0)}, \tag{A.11}$$

where $\mathcal{F}(t)$ is the value of the particular index at time t, $\mathcal{F}(0)$ is the respective value at $t = 0$, $\mathcal{S}_{\mathcal{F}}(t)$ is the standard deviation of the index at time t, and $\mathcal{PDF}(t)$ is the percentage decrease in this factor at time t.

A.5 Model Erato

A.5.1 Model equations

For simplicity, we introduce the notation

$$\begin{aligned} \Gamma_s &= \gamma_0 + \theta_1 + \mu + \mu_1 & \Phi_s &= \phi_1 + \mu \\ \Delta_s &= \delta + \delta_0 + \theta_2 + \mu + \mu_2 & \mathrm{B}'_s &= \beta' + \phi_2 + \mu \\ \mathrm{E}_s &= \epsilon_1 + \epsilon_2 + \mu & \mathrm{B}_s &= \beta + \mu. \end{aligned}$$

Let $\mathbf{e}_i$ denote the vector of $\mathbb{Z}_+^7$ whose i-th component is 1 and all other components are zero. For $\mathbf{x} = (x, x_v, z, z_v, y, w, u) \in \mathbb{Z}_+^7$ the probabilities $p(\mathbf{x}; t)$ satisfy the following equations:

$$
\begin{aligned}
\frac{dp(\mathbf{x})}{dt} &= (1-\phi)\lambda p(\mathbf{x}-\mathbf{e}_1) + \phi\lambda p(\mathbf{x}-\mathbf{e}_2) + \mu(x+1)p(\mathbf{x}+\mathbf{e}_1) \\
&\quad + \mu(x_v+1)p(\mathbf{x}+\mathbf{e}_2) + \mu(z+1)p(\mathbf{x}+\mathbf{e}_3) + \mu(z_v+1)p(\mathbf{x}+\mathbf{e}_4) \\
&\quad + (\mu+\mu_1)(y+1)p(\mathbf{x}+\mathbf{e}_5) + (\mu+\mu_2)(w+1)p(\mathbf{x}+\mathbf{e}_6) + \mu(u+1)p(\mathbf{x}+\mathbf{e}_7) \\
&\quad + \phi_1(x_v+1)p(\mathbf{x}-\mathbf{e}_1+\mathbf{e}_2) + \phi_2(z_v+1)p(\mathbf{x}-\mathbf{e}_3+\mathbf{e}_4) + \delta(w+1)p(\mathbf{x}-\mathbf{e}_5+\mathbf{e}_6) \\
&\quad + \frac{\alpha}{n}(x+1)[pq_1(y-1)p(\mathbf{x}-\mathbf{e}_5+\mathbf{e}_1) + (1-p)yp(\mathbf{x}-\mathbf{e}_3+\mathbf{e}_1) \\
&\quad + p(1-q_1)yp(\mathbf{x}-\mathbf{e}_6+\mathbf{e}_1)] + \frac{\alpha}{n}(x_v+1)[p'q_1(y-1)p(\mathbf{x}-\mathbf{e}_5+\mathbf{e}_2) \\
&\quad + (1-p')yp(\mathbf{x}-\mathbf{e}_4+\mathbf{e}_2) + p'(1-q_1)yp(\mathbf{x}-\mathbf{e}_6+\mathbf{e}_2)] + \theta_1(y+1)p(\mathbf{x}-\mathbf{e}_3+\mathbf{e}_5) \\
&\quad + \theta_2(w+1)p(\mathbf{x}-\mathbf{e}_3+\mathbf{e}_6) + [q_2\beta + q_3\frac{\alpha_2}{n}(y-1)](z+1)p(\mathbf{x}-\mathbf{e}_5+\mathbf{e}_3) \\
&\quad + [(1-q_2)\beta + (1-q_3)\frac{\alpha_2}{n}y](z+1)p(\mathbf{x}-\mathbf{e}_6+\mathbf{e}_3) + \gamma_0(y+1)p(\mathbf{x}-\mathbf{e}_7+\mathbf{e}_5) \\
&\quad + [(1-q_2)\beta' + (1-q_3)\frac{\alpha_2'}{n}y](z_v+1)p(\mathbf{x}-\mathbf{e}_6+\mathbf{e}_4) + \delta_0(w+1)p(\mathbf{x}-\mathbf{e}_7+\mathbf{e}_6) \\
&\quad + [q_2\beta' + q_3\frac{\alpha_2'}{n}(y-1)](z_v+1)p(\mathbf{x}-\mathbf{e}_5+\mathbf{e}_4) + \epsilon_1(u+1)p(\mathbf{x}-\mathbf{e}_5+\mathbf{e}_7) \\
&\quad + \epsilon_2(u+1)p(\mathbf{x}-\mathbf{e}_6+\mathbf{e}_7) - [\lambda + \mu x + \Phi_s x_v + \mathrm{B}_s z + \mathrm{B}_s' z_v + \Gamma_s y + \Delta_s w \\
&\quad + \mathrm{E}_s u + \frac{\alpha}{n}(x+x_v)y + \frac{\alpha_2}{n}yz + \frac{\alpha_2'}{n}yz_v]p(\mathbf{x}),
\end{aligned}
$$

where for simplicity, the dependence on t has been suppressed in all the terms $p(\mathbf{x};t)$. The variances of X, X_v, Z, Z_v, Y, W, and U satisfy the following equations:

$$
\begin{aligned}
\frac{d\mathrm{Var}[X(t)]}{dt} &= (1-\phi)\lambda + \mu\mathrm{E}[X] - 2\mu\mathrm{Var}[X] + \frac{\alpha}{n}(1+2\mathrm{E}[X])\mathrm{E}[XY] - 2\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
&\quad + \phi_1\mathrm{E}[X_v] + 2\phi_1\mathrm{Cov}[X,X_v] \\
\frac{d\mathrm{Var}[X_v(t)]}{dt} &= \phi\lambda + \Phi_s\mathrm{E}[X_v] - 2\Phi_s\mathrm{Var}[X_v] + \frac{\alpha}{n}(1+2\mathrm{E}[X_v])\mathrm{E}[X_vY] - 2\frac{\alpha}{n}\mathrm{E}[X_v^2Y] \\
\frac{d\mathrm{Var}[Z(t)]}{dt} &= \mathrm{B}_s\mathrm{E}[Z] + \phi_2\mathrm{E}[Z_v] + \theta_2\mathrm{E}[W] + \theta_1\mathrm{E}[Y] + 2\phi_2\mathrm{Cov}[Z,Z_v] + 2\theta_2\mathrm{Cov}[Z,W] \\
&\quad + 2\theta_1\mathrm{Cov}[Y,Z] - 2\mathrm{B}_s\mathrm{Var}[Z] - 2\frac{\alpha_2}{n}\mathrm{E}[YZ^2] + 2(1-p)\frac{\alpha}{n}\mathrm{E}[XYZ] \\
&\quad + (1-p)\frac{\alpha}{n}(1-2\mathrm{E}[Z])\mathrm{E}[XY] + \frac{\alpha_2}{n}(1+2\mathrm{E}[Z])\mathrm{E}[YZ] \\
\frac{d\mathrm{Var}[Z_v(t)]}{dt} &= \mathrm{B}_s'(\mathrm{E}[Z_v] - 2\mathrm{Var}[Z]) - 2\frac{\alpha_2'}{n}\mathrm{E}[YZ_v^2] + 2(1-p')\frac{\alpha}{n}\mathrm{E}[X_vYZ_v] \\
&\quad + (1-p')\frac{\alpha}{n}(1-2\mathrm{E}[Z_v])\mathrm{E}[X_vY] + \frac{\alpha_2'}{n}(1+2\mathrm{E}[Z_v])\mathrm{E}[YZ_v]
\end{aligned}
$$

$$
\begin{aligned}
\frac{d\mathrm{Var}[Y(t)]}{dt} &= \Gamma_s \mathrm{E}[Y] + q_2\beta\mathrm{E}[Z] + \delta\mathrm{E}[W] + q_2\beta'\mathrm{E}[Z_v] + \epsilon_1\mathrm{E}[U] - 2\Gamma_s\mathrm{Var}[Y] \\
&\quad + 2pq_1\frac{\alpha}{n}\mathrm{E}[XY^2] + 2q_3\frac{\alpha_2}{n}\mathrm{E}[Y^2Z] + 2p'q_1\frac{\alpha}{n}\mathrm{E}[X_vY^2] + 2q_3\frac{\alpha_2'}{n}\mathrm{E}[Y^2Z_v] \\
&\quad + 2\delta\mathrm{Cov}[Y,W] + 2\epsilon_1\mathrm{Cov}[Y,U] + 2q_2\beta\mathrm{Cov}[Y,Z] + 2q_2\beta'\mathrm{Cov}[Y,Z_v] \\
&\quad + q_1\frac{\alpha}{n}(1-2\mathrm{E}[Y])(p\mathrm{E}[XY] + p'\mathrm{E}[X_vY]) \\
&\quad + q_3(1-2\mathrm{E}[Y])\left(\frac{\alpha_2}{n}\mathrm{E}[YZ] + \frac{\alpha_2'}{n}\mathrm{E}[YZ_v]\right) \\
\frac{d\mathrm{Var}[W(t)]}{dt} &= \Delta_s\mathrm{E}[W] + (1-q_2)(\beta\mathrm{E}[Z] + \beta'\mathrm{E}[Z_v]) + \epsilon_2\mathrm{E}[U] - 2\Delta_s\mathrm{Var}[W] \\
&\quad + 2(1-q_1)\frac{\alpha}{n}(p\mathrm{E}[XYW] + p'\mathrm{E}[X_vYW]) + 2\epsilon_2\mathrm{Cov}[W,U] \\
&\quad + 2(1-q_3)\left(\frac{\alpha_2}{n}\mathrm{E}[YZW] + \frac{\alpha_2'}{n}\mathrm{E}[YZ_vW]\right) + 2(1-q_2)(\beta\mathrm{Cov}[Z,W] \\
&\quad + \beta'\mathrm{Cov}[Z_v,W]) + (1-q_1)\frac{\alpha}{n}(1-2\mathrm{E}[W])(p\mathrm{E}[XY] + p'\mathrm{E}[X_vY]) \\
&\quad + (1-q_3)(1-2\mathrm{E}[W])\left(\frac{\alpha_2}{n}\mathrm{E}[YZ] + \frac{\alpha_2'}{n}\mathrm{E}[YZ_v]\right) \\
\frac{d\mathrm{Var}[U(t)]}{dt} &= \gamma_0(\mathrm{E}[Y] + 2\mathrm{Cov}[Y,U]) + \delta_0(\mathrm{E}[W] + 2\mathrm{Cov}[W,U]) + \mathrm{E}_s(\mathrm{E}[U] - 2\mathrm{Var}[U]).
\end{aligned}
$$

The covariances of X, X_v, Z, Z_v, Y, W, and U satisfy the following equations:

$$
\begin{aligned}
\frac{d\mathrm{Cov}[X,X_v]}{dt} &= -\phi_1\mathrm{E}[X_v] + \phi_1\mathrm{Var}[X_v] - 2\frac{\alpha}{n}\mathrm{E}[XX_vY] - (\Phi_s + \mu)\mathrm{Cov}[X,X_v] \\
&\quad + \frac{\alpha}{n}(\mathrm{E}[X]\mathrm{E}[X_vY] + \mathrm{E}[X_v]\mathrm{E}[XY]) \\
\frac{d\mathrm{Cov}[X,Z]}{dt} &= \theta_1\mathrm{Cov}[X,Y] + \theta_2\mathrm{Cov}[X,W] - (\mathrm{B}_s + \mu)\mathrm{Cov}[X,Z] + \phi_1\mathrm{Cov}[X_v,Z] \\
&\quad + \phi_2\mathrm{Cov}[X,Z_v] - \mathrm{E}[XYZ](\alpha + \alpha_2)/n + (1-p)\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY](\mathrm{E}[Z] - (1-p)\mathrm{E}[X] - (1-p)) + \frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[X] \\
\frac{d\mathrm{Cov}[X,Z_v]}{dt} &= \phi_1\mathrm{Cov}[X_v,Z_v] - \mathrm{E}[XYZ_v](\alpha + \alpha_2')/n + (1-p')\frac{\alpha}{n}\mathrm{E}[XX_vY] \\
&\quad - (\mathrm{B}_s' + \mu)\mathrm{Cov}[X,Z_v] + \frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[Z_v] - (1-p')\frac{\alpha}{n}\mathrm{E}[X]\mathrm{E}[X_vY] \\
&\quad + \frac{\alpha_2'}{n}\mathrm{E}[X]\mathrm{E}[YZ_v] \\
\frac{d\mathrm{Cov}[X,Y]}{dt} &= -(\Gamma_s + \mu)\mathrm{Cov}[X,Y] + \delta\mathrm{Cov}[X,W] + \epsilon_1\mathrm{Cov}[X,U] + q_2\beta\mathrm{Cov}[X,Z] \\
&\quad + q_2\beta'\mathrm{Cov}[X,Z_v] + \phi_1\mathrm{Cov}[X_v,Y] - \frac{\alpha}{n}\mathrm{E}[XY^2] + q_3\frac{\alpha_2}{n}\mathrm{E}[XYZ] \\
&\quad + p'q_1\frac{\alpha}{n}\mathrm{E}[XX_vY] + q_3\frac{\alpha_2'}{n}\mathrm{E}[XYZ_v] + pq_1\frac{\alpha}{n}\mathrm{E}[X^2Y] + p'q_1\frac{\alpha}{n}\mathrm{E}[X]\mathrm{E}[X_vY] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY](\mathrm{E}[Y] - pq_1\mathrm{E}[X] - pq_1) - q_3\mathrm{E}[X](\alpha_2\mathrm{E}[YZ] + \alpha_2'\mathrm{E}[YZ_v])/n
\end{aligned}
$$

$$
\begin{aligned}
\frac{d\mathrm{Cov}[X,W]}{dt} &= \phi_1\mathrm{Cov}[X_v,W] - (\Delta_s+\mu)\mathrm{Cov}[X,W] + \epsilon_2\mathrm{Cov}[X,U] - \frac{\alpha}{n}\mathrm{E}[XYW] \\
&\quad + (1-q_2)(\beta\mathrm{Cov}[X,Z] + \beta'\mathrm{Cov}[X,Z_v]) + p'(1-q_1)\frac{\alpha}{n}\mathrm{E}[XX_vY] \\
&\quad + (1-q_3)(\alpha_2\mathrm{E}[XYZ] + \alpha_2'\mathrm{E}[XYZ_v])/n + p(1-q_1)\frac{\alpha}{n}\mathrm{E}[X^2Y] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY](\mathrm{E}[W] - p(1-q_1)\mathrm{E}[X] - p(1-q_1)) \\
&\quad - (1-q_3)\mathrm{E}[X](\alpha_2\mathrm{E}[YZ] + \alpha_2'\mathrm{E}[YZ_v])/n - p'(1-q_1)\frac{\alpha}{n}\mathrm{E}[X]\mathrm{E}[X_vY] \\
\frac{d\mathrm{Cov}[X,U]}{dt} &= \phi_1\mathrm{Cov}[X_v,U] - (\mathrm{E}_s+\mu)\mathrm{Cov}[X,U] + \gamma_0\mathrm{Cov}[X,Y] + \delta_0\mathrm{Cov}[X,W] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[U] - \frac{\alpha}{n}\mathrm{E}[XYU] \\
\frac{d\mathrm{Cov}[X_v,Z]}{dt} &= -(\Phi_s+\mathrm{B}_s)\mathrm{Cov}[X_v,Z] + \phi_2\mathrm{Cov}[X_v,Z_v] + \theta_1\mathrm{Cov}[X_v,Y] \\
&\quad + \theta_2\mathrm{Cov}[X_v,W] - \mathrm{E}[X_vYZ](\alpha+\alpha_2)/n + (1-p)\frac{\alpha}{n}\mathrm{E}[XX_vY] \\
&\quad - (1-p)\frac{\alpha}{n}\mathrm{E}[XY]\mathrm{E}[X_v] + \frac{\alpha_2}{n}\mathrm{E}[YZ]\mathrm{E}[X_v] + \frac{\alpha}{n}\mathrm{E}[X_vY]\mathrm{E}[Z] \\
\frac{d\mathrm{Cov}[X_v,Z_v]}{dt} &= -(\mathrm{B}_s'+\Phi_s)\mathrm{Cov}[X_v,Z_v] - \mathrm{E}[X_vYZ_v](\alpha+\alpha_2')/n + (1-p')\frac{\alpha}{n}\mathrm{E}[X_v^2Y] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[X_vY](\mathrm{E}[Z_v] - (1-p')\mathrm{E}[X_v] - (1-p')) + \frac{\alpha_2'}{n}\mathrm{E}[X_v]\mathrm{E}[YZ_v] \\
\frac{d\mathrm{Cov}[X_v,Y]}{dt} &= -(\Gamma_s+\Phi_s)\mathrm{Cov}[X_v,Y] + q_2\beta\mathrm{Cov}[X_v,Z] + q_2\beta'\mathrm{Cov}[X_v,Z_v] \\
&\quad + \delta\mathrm{Cov}[X_v,W] + \epsilon_1\mathrm{Cov}[X_v,U] - \frac{\alpha}{n}\mathrm{E}[X_vY^2] + pq_1\frac{\alpha}{n}\mathrm{E}[XX_vY] \\
&\quad + q_3\frac{\alpha_2}{n}\mathrm{E}[X_vYZ] + q_3\frac{\alpha_2'}{n}\mathrm{E}[X_vYZ_v] + p'q_1\frac{\alpha}{n}\mathrm{E}[X_v^2Y] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[X_vY](\mathrm{E}[Y] - p'q_1\mathrm{E}[X_v] - p'q_1) - pq_1\frac{\alpha}{n}\mathrm{E}[X_v]\mathrm{E}[XY] \\
&\quad - q_3\mathrm{E}[X_v](\alpha_2\mathrm{E}[YZ] + \alpha_2'\mathrm{E}[YZ_v])/n \\
\frac{d\mathrm{Cov}[X_v,W]}{dt} &= -(\Delta_s+\Phi_s)\mathrm{Cov}[X_v,W] + \epsilon_2\mathrm{Cov}[X_v,U] - \frac{\alpha}{n}\mathrm{E}[X_vYW] \\
&\quad + (1-q_2)(\beta\mathrm{Cov}[X_v,Z] + \beta'\mathrm{Cov}[X_v,Z_v]) - p(1-q_1)\frac{\alpha}{n}\mathrm{E}[X_v]\mathrm{E}[XY] \\
&\quad + (1-q_3)(\alpha_2\mathrm{E}[X_vYZ] + \alpha_2'\mathrm{E}[X_vYZ_v])/n + (1-q_1)\frac{\alpha}{n}(p\mathrm{E}[XX_vY] \\
&\quad + p'\mathrm{E}[X_v^2Y]) + \frac{\alpha}{n}\mathrm{E}[X_vY](\mathrm{E}[W] - p'(1-q_1)\mathrm{E}[X_v] - p'(1-q_1)) \\
&\quad - (1-q_3)\mathrm{E}[X_v](\alpha_2\mathrm{E}[YZ] + \alpha_2'\mathrm{E}[YZ_v])/n \\
\frac{d\mathrm{Cov}[X_v,U]}{dt} &= -(\mathrm{E}_s+\Phi_s)\mathrm{Cov}[X_v,U] + \gamma_0\mathrm{Cov}[X_v,Y] + \delta_0\mathrm{Cov}[X_v,W] \\
&\quad + \frac{\alpha}{n}\mathrm{E}[X_vY]\mathrm{E}[U] - \frac{\alpha}{n}\mathrm{E}[X_vYU]
\end{aligned}
$$

$$
\begin{aligned}
\frac{d\mathrm{Cov}[Z,Z_v]}{dt} = & -(\mathrm{B}_s+\mathrm{B}'_s)\mathrm{Cov}[Z,Z_v]+\theta_1\mathrm{Cov}[Y,Z_v]+\theta_2\mathrm{Cov}[W,Z_v]+\phi_2\mathrm{Var}[Z_v]\\
& -\phi_2\mathrm{E}[Z_v]+(1-p)\frac{\alpha}{n}\mathrm{E}[XYZ_v]-\mathrm{E}[ZZ_vY](\alpha_2+\alpha'_2)/n\\
& +(1-p')\frac{\alpha}{n}(\mathrm{E}[X_vZY]-\mathrm{E}[Z]\mathrm{E}[X_vY])+\frac{\alpha'_2}{n}\mathrm{E}[Z]\mathrm{E}[Z_vY]\\
& -(1-p)\frac{\alpha}{n}\mathrm{E}[Z_v]\mathrm{E}[XY]+\frac{\alpha_2}{n}\mathrm{E}[Z_v]\mathrm{E}[ZY]\\
\frac{d\mathrm{Cov}[Z,Y]}{dt} = & \phi_2\mathrm{Cov}[Z_v,Y]+\theta_2\mathrm{Cov}[Y,W]+\theta_1\mathrm{Var}[Y]-\theta_1\mathrm{E}[Y]+\epsilon_1\mathrm{Cov}[Z,U]\\
& -(\Gamma_s+\mathrm{B}_s)\mathrm{Cov}[Y,Z]+\delta\mathrm{Cov}[Z,W]+q_2\beta(\mathrm{Var}[Z]-\mathrm{E}[Z])\\
& +(1-p)\frac{\alpha}{n}\mathrm{E}[XY^2]+\frac{\alpha_2}{n}\mathrm{E}[YZ](\mathrm{E}[Y]-q_3\mathrm{E}[Z]-q_3)-\frac{\alpha_2}{n}\mathrm{E}[Y^2Z]\\
& +pq_1\frac{\alpha}{n}\mathrm{E}[XYZ]+p'q_1\frac{\alpha}{n}\mathrm{E}[X_vYZ]+q_2\beta'\mathrm{Cov}[ZZ_v]+q_3\frac{\alpha'_2}{n}\mathrm{E}[ZZ_vY]\\
& +q_3\frac{\alpha_2}{n}\mathrm{E}[YZ^2]-\frac{\alpha}{n}\mathrm{E}[XY](pq_1\mathrm{E}[Z]+(1-p)\mathrm{E}[Y])\\
& -p'q_1\frac{\alpha}{n}\mathrm{E}[Z]\mathrm{E}[X_vY]-q_3\frac{\alpha'_2}{n}\mathrm{E}[Z]\mathrm{E}[Z_vY]\\
\frac{d\mathrm{Cov}[Z,W]}{dt} = & \theta_1\mathrm{Cov}[Y,W]+\epsilon_2\mathrm{Cov}[Z,U]+\theta_2(\mathrm{Var}[W]-\mathrm{E}[W])+\phi_2\mathrm{Cov}[Z_v,W]\\
& +(1-q_2)\beta(\mathrm{Var}[Z]-\mathrm{E}[Z])-(\Delta_s+\mathrm{B}_s)\mathrm{Cov}[Z,W]+\frac{\alpha_2}{n}\mathrm{E}[W]\mathrm{E}[YZ]\\
& +(1-q_2)\beta'\mathrm{Cov}[Z,Z_v]+(1-p)\frac{\alpha}{n}(\mathrm{E}[XYW]-\mathrm{E}[W]\mathrm{E}[XY])\\
& +p'(1-q_1)\frac{\alpha}{n}(\mathrm{E}[X_vZY]-\mathrm{E}[Z]\mathrm{E}[X_vY])+p(1-q_1)\frac{\alpha}{n}(\mathrm{E}[XZY]\\
& -\mathrm{E}[Z]\mathrm{E}[XY])+(1-q_3)\frac{\alpha_2}{n}(\mathrm{E}[Z^2Y]-\mathrm{E}[YZ](1-\mathrm{E}[Z]))\\
& -\frac{\alpha_2}{n}\mathrm{E}[YZW]+(1-q_3)\frac{\alpha'_2}{n}(\mathrm{E}[ZZ_vY]-\mathrm{E}[Z]\mathrm{E}[Z_vY])\\
\frac{d\mathrm{Cov}[Z,U]}{dt} = & \phi_2\mathrm{Cov}[Z_v,U]+\theta_1\mathrm{Cov}[Y,U]+\theta_2\mathrm{Cov}[W,U]+\gamma_0\mathrm{Cov}[Y,Z]\\
& +\delta_0\mathrm{Cov}[Z,W]-(\mathrm{B}_s+\mathrm{E}_s)\mathrm{Cov}[Z,U]\\
& +(1-p)\frac{\alpha}{n}(\mathrm{E}[XYU]-\mathrm{E}[XY]\mathrm{E}[U])+\frac{\alpha_2}{n}(\mathrm{E}[YZ]\mathrm{E}[U]-\mathrm{E}[YZU])\\
\frac{d\mathrm{Cov}[Z_v,Y]}{dt} = & -(\Gamma_s+\mathrm{B}'_s)\mathrm{Cov}[Y,Z_v]+q_2\beta\mathrm{Cov}[Z,Z_v]+\delta\mathrm{Cov}[Z_v,W]+\epsilon_1\mathrm{Cov}[Z_v,U]\\
& +q_2\beta'(\mathrm{Var}[Z_v]-\mathrm{E}[Z_v])+(1-p')\frac{\alpha}{n}(\mathrm{E}[X_vY^2]-\mathrm{E}[Y]\mathrm{E}[X_vY])\\
& +p'q_1\frac{\alpha}{n}(\mathrm{E}[X_vZ_vY]-\mathrm{E}[Z_v]\mathrm{E}[X_vY])+q_3\frac{\alpha_2}{n}(\mathrm{E}[ZZ_vY]-\mathrm{E}[Z_v]\mathrm{E}[ZY])\\
& +pq_1\frac{\alpha}{n}(\mathrm{E}[XZ_vY]-\mathrm{E}[Z_v]\mathrm{E}[XY])-\frac{\alpha'_2}{n}(\mathrm{E}[Z_vY^2]-\mathrm{E}[Y]\mathrm{E}[Z_vY])\\
& +q_3\frac{\alpha'_2}{n}(\mathrm{E}[Z_v^2Y]-\mathrm{E}[YZ_v](1+\mathrm{E}[Z_v]))
\end{aligned}
$$

$$\begin{aligned}
\frac{d\mathrm{Cov}[Z_v, W]}{dt} = & -(\Delta_s + \mathrm{B}'_s)\mathrm{Cov}[Z_v, W] + \epsilon_2 \mathrm{Cov}[Z_v, U] + (1 - q_2)\beta'(\mathrm{Var}[Z_v] - \mathrm{E}[Z_v]) \\
& + (1 - q_2)\beta \mathrm{Cov}[Z, Z_v] + (1 - p')\frac{\alpha}{n}(\mathrm{E}[X_v Y W] - \mathrm{E}[W]\mathrm{E}[X_v Y]) \\
& + p'(1 - q_1)\frac{\alpha}{n}(\mathrm{E}[X_v Z_v Y] - \mathrm{E}[Z_v]\mathrm{E}[X_v Y]) + p(1 - q_1)\frac{\alpha}{n}(\mathrm{E}[X Z_v Y] \\
& - \mathrm{E}[Z_v]\mathrm{E}[XY]) + (1 - q_3)\frac{\alpha_2}{n}(\mathrm{E}[Z Z_v Y] - \mathrm{E}[Z_v]\mathrm{E}[ZY]) - \frac{\alpha'_2}{n}\mathrm{E}[Z_v Y W] \\
& + (1 - q_3)\frac{\alpha'_2}{n}(\mathrm{E}[Z_v^2 Y] - \mathrm{E}[Y Z_v](1 + \mathrm{E}[Z_v])) + \frac{\alpha'_2}{n}\mathrm{E}[W]\mathrm{E}[Z_v Y] \\
\frac{d\mathrm{Cov}[Z_v, U]}{dt} = & -(\mathrm{B}'_s + \mathrm{E}_s)\mathrm{Cov}[Z_v, U] + \gamma_0 \mathrm{Cov}[Y, Z_v] + \delta_0 \mathrm{Cov}[Z, W_v] \\
& + \frac{\alpha'_2}{n}(\mathrm{E}[Y Z_v]\mathrm{E}[U] - \mathrm{E}[Y Z_v U]) + (1 - p')\frac{\alpha}{n}(\mathrm{E}[X_v Y U] - \mathrm{E}[X_v Y]\mathrm{E}[U]) \\
\frac{d\mathrm{Cov}[Y, W]}{dt} = & -(\Gamma_s + \Delta_s)\mathrm{Cov}[Y, W] + \epsilon_1 \mathrm{Cov}[W, U] + \epsilon_2 \mathrm{Cov}[Y, U] + \delta \mathrm{Var}[W] \\
& - \delta \mathrm{E}[W] + q_2 \beta \mathrm{Cov}[Z, W] + q_2 \beta' \mathrm{Cov}[Z_v, W] + (1 - q_2)(\beta \mathrm{Cov}[Y, Z] \\
& + \beta' \mathrm{Cov}[Y, Z_v]) + (1 - q_3)\frac{\alpha'_2}{n}(\mathrm{E}[Z_v Y^2] - \mathrm{E}[Y]\mathrm{E}[Y Z_v]) \\
& + p q_1 \frac{\alpha}{n}(\mathrm{E}[XYW] - \mathrm{E}[W]\mathrm{E}[XY]) + (1 - q_3)\frac{\alpha_2}{n}\mathrm{E}[Z Y^2] \\
& + p' q_1 \frac{\alpha}{n}(\mathrm{E}[X_v Y W] - \mathrm{E}[W]\mathrm{E}[X_v Y]) + q_3 \frac{\alpha_2}{n}(\mathrm{E}[YZW] - \mathrm{E}[W]\mathrm{E}[YZ]) \\
& + (1 - q_1)\frac{\alpha}{n}\{p'\mathrm{E}[X_v Y^2] + p\mathrm{E}[XY^2] - \mathrm{E}[Y](p\mathrm{E}[XY] + p'\mathrm{E}[X_v Y])\} \\
& + q_3 \frac{\alpha'_2}{n}(\mathrm{E}[Z_v Y W] - \mathrm{E}[W]\mathrm{E}[Y Z_v]) - (1 - q_3)\frac{\alpha_2}{n}\mathrm{E}[Y]\mathrm{E}[YZ] \\
\frac{d\mathrm{Cov}[Y, U]}{dt} = & -(\Gamma_s + \mathrm{E}_s)\mathrm{Cov}[Y, U] + \delta \mathrm{Cov}[W, U] + \delta_0 \mathrm{Cov}[Y, W] + \gamma_0 \mathrm{Var}[Y] \\
& - \gamma_0 \mathrm{E}[Y] + \epsilon_1 \mathrm{Var}[U] - \epsilon_1 \mathrm{E}[U] + q_2 \beta \mathrm{Cov}[Z, U] + q_2 \beta' \mathrm{Cov}[Z_v, U] \\
& - q_1 \frac{\alpha}{n}\mathrm{E}[U](p\mathrm{E}[XY] + p'\mathrm{E}[X_v Y]) + q_1 \frac{\alpha}{n}(p\mathrm{E}[XYU] + p'\mathrm{E}[X_v Y U]) \\
& - q_3 \mathrm{E}[U](\alpha_2 \mathrm{E}[YZ] + \alpha'_2 \mathrm{E}[Y Z_v])/n + q_3(\alpha_2 \mathrm{E}[YZU] + \alpha'_2 \mathrm{E}[Y Z_v U])/n \\
\frac{d\mathrm{Cov}[W, U]}{dt} = & -(\Delta_s + \mathrm{E}_s)\mathrm{Cov}[W, U] - \epsilon_2 \mathrm{E}[U] - \delta_0 \mathrm{E}[W] + \epsilon_2 \mathrm{Var}[U] + \delta_0 \mathrm{Var}[W] \\
& + \gamma_0 \mathrm{Cov}[Y, W] + (1 - q_2)(\beta \mathrm{Cov}[Z, U] + \beta' \mathrm{Cov}[Z_v, U]) \\
& + (1 - q_1)\frac{\alpha}{n}\{p\mathrm{E}[XYU] + p'\mathrm{E}[X_v Y U] - \mathrm{E}[U](p\mathrm{E}[XY] + p'\mathrm{E}[X_v Y])\} \\
& + (1 - q_3)\frac{1}{n}\{\alpha_2 \mathrm{E}[YZU] + \alpha'_2 \mathrm{E}[Z_v Y U] - \mathrm{E}[U](\alpha_2 \mathrm{E}[YZ] + \alpha'_2 \mathrm{E}[Z_v Y])\}.
\end{aligned}$$

A.5.2 Numerical results from simulations

For the simulation of model Erato there are 34 different kinds of events that can occur. The events and their rates are defined in Table A.6. The vector $\mathbf{X}_c$ gives the current value of $\mathbf{X} = (X, X_v, Z, Z_v, Y, W, U)$, which at time $t = 0$ has the value $\mathbf{X}_0 = (X(0), X_v(0), Z(0), Z_v(0), Y(0), W(0), U(0))$.

Event	Rate	Event	Rate
birth of X	$(1-\phi)\lambda$	$Z \rightarrow Y$ (reactiv)	$q_2\beta Z$
birth of X_v	$\phi\lambda$	$Z \rightarrow Y$ (reinf)	$q_3\frac{\alpha_2}{n}YZ$
death of X	μX	$Z \rightarrow W$ (reactiv)	$(1-q_2)\beta Z$
death of X_v	μX_v	$Z \rightarrow W$ (reinf)	$(1-q_3)\frac{\alpha_2}{n}YZ$
death of Z	μZ	$Z_v \rightarrow Y$ (reactiv)	$q_2\beta' Z_v$
death of Z_v	μZ_v	$Z_v \rightarrow Y$ (reinf)	$q_3\frac{\alpha_2'}{n}YZ_v$
natural death of Y	μY	$Z_v \rightarrow W$ (reactiv)	$(1-q_2)\beta' Z_v$
TB death of Y	$\mu_1 Y$	$Z_v \rightarrow W$ (reinf)	$(1-q_3)\frac{\alpha_2'}{n}YZ_v$
natural death of W	μW	$Y \rightarrow U$	$\gamma_0 Y$
TB death of W	$\mu_2 W$	$U \rightarrow Y$	$\epsilon_1 U$
death of U	μU	$W \rightarrow U$	$\delta_0 W$
$X \rightarrow Y$	$pq_1\frac{\alpha}{n}XY$	$U \rightarrow W$	$\epsilon_2 U$
$X \rightarrow Z$	$(1-p)\frac{\alpha}{n}XY$	$Y \rightarrow Z$	$\theta_1 Y$
$X \rightarrow W$	$p(1-q_1)\frac{\alpha}{n}XY$	$W \rightarrow Z$	$\theta_2 W$
$X_v \rightarrow Y$	$p'q_1\frac{\alpha}{n}X_vY$	$W \rightarrow Y$	δW
$X_v \rightarrow Z_v$	$(1-p')\frac{\alpha}{n}X_vY$	$X_v \rightarrow X$	$\phi_1 X_v$
$X_v \rightarrow W$	$p'(1-q_1)\frac{\alpha}{n}X_vY$	$Z_v \rightarrow Z$	$\phi_2 Z_v$

Table A.6: Events for the simulations of model Erato

We assume that at time $t = 0$, when chemotherapy and BCG vaccination are introduced, the epidemic has already reached the steady endemic level for the natural evolution of TB (described by the quasi-stationary distribution of model Zeus). Therefore, the vector of initial conditions, $\mathbf{X}_0$, for model Erato must be a variate from the quasi-stationary distribution of model Zeus.

This was implemented as for model Clio (see Section A.4.3). $R = 10^4$ simulation runs of model Zeus were carried out with the parameter values of interest. Let $\mathbf{X}_i^z(t)$ denote the value of the vector $(X(t), Y(t), Z(t), W(t), U(t))$ at time t from the i-th simulation run of model Zeus, for $i = 1, 2, \ldots, R$ and $\mathbf{X}_i^{\varepsilon 1}(t)$, $\mathbf{X}_i^{\varepsilon 2}(t)$ the values of the vectors $(X(t), Y(t), Z(t), W(t), U(t))$ and $(X_v(t), Z_v(t))$, respectively, at time t from the i-th simulation run of Erato.

The results from the simulations of model Zeus (see Sections 6.3.5 and 6.3.4) show that by time $t = 300$ the process has either died out or reached the steady endemic level. In particular the simulations that were carried out with a large value of y_0 (the initial number of infectious cases) all ended at the steady endemic level. So, the 10^4 runs of Zeus were carried out with the parameter values of interest and $\mathbf{X}_i^{z}(0) = (n - 10, 10, 0, 0, 0)$, which ensured that all 10^4 runs ended at the steady endemic level and hence each of the 10^4 vectors $\mathbf{X}_i^{z}(300)$ is a variate from the quasi-stationary distribution of Zeus. Then the initial conditions for the simulations of Erato where taken as $\mathbf{X}_i^{\varepsilon 1}(0) = \mathbf{X}_i^{z}(300)$ and $\mathbf{X}_i^{\varepsilon 2}(0) = (0, 0)$, for $i = 1, 2, \dots, R$, ensuring that the initial conditions of Erato are variates from the steady endemic level of the natural evolution of TB.

The estimates of the epidemiological indices and the corresponding variances were calculated from formulae (A.8), where R is the number of individual simulation runs ($R = 10^4$). The d_i and N_i are defined as for the equations (A.8), except

• *for the risk of infection:* d_i is the number of new infections (i.e. transitions from X to Y, Z, and W and transitions from X_v to Y, Z_v, and W) that occurred during year t in the i-th run.

• *for the incidence of infectious TB:* d_i is the number of all the transitions from the classes X, X_v, Z, Z_v to the class Y, during year t in the i-th run.

• *for the incidence of non-infectious TB:* d_i is the number of all the transitions from the classes X, X_v, Z, Z_v to the class W, during year t in the i-th run.

The percentage decline (due to BCG) in the particular epidemiological index at time t was calculated from the formula

$$D(t) = \frac{F_c(t) - F_\varepsilon(t)}{F_c(t)} 100,$$

where $F_\varepsilon(t)$ and $F_c(t)$ are the values of the particular index at time t, as calculated from models Erato and Clio, respectively. Also, if $S_\varepsilon(t)$ denotes the standard deviation of the index at time t, then the corresponding 95% confidence interval for the estimate of the index was calculated from the formula

$$F_\varepsilon(t) \pm 1.96 \frac{S_\varepsilon(t)}{\sqrt{R}}.$$

Bibliography

Abramov, V. (1994), 'On the asymptotic distribution of the maximum number of infectives in epidemic models with immigration', *J. Appl. Prob.* **31**, 606–613.

Allen, L. & Burgin, A. (2000), 'Comparison of deterministic and stochastic SIS and SIR models in discrete time', *Math. Biosciences* **163**, 1–33.

American Thoracic Society (1990), 'Diagnostic standards and classification of tuberculosis', *Amer. Rev. Resp. Dis.* **142**, 725–735.

Anderson, R. & May, R. (1991), *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford.

Aparicio, J., Capurro, A. & Castillo-Chavez, C. (2001), 'Social clustering and the transmission and dynamics of tuberculosis'. To appear.

Azuma, Y. (1975), 'A simple simulation model of tuberculosis epidemiology for use without large-scale computers', *Bull. World Health Org.* **52**, 313–322.

Bailey, N. (1964), *The Elements of Stochastic Processes*, John Wiley, New York.

Bailey, N. (1975), *The Mathematical Theory of Infectious Diseases and its Applications*, Griffin, London.

Ball, F. (1983), 'The threshold behaviour of epidemic models', *J. Appl. Prob.* **20**, 227–241.

Ball, F. (1995), Coupling methods in epidemic theory, *in* D. Mollison, ed., 'Epidemic models. Their structure and relation to data', Cambridge University Press, Cambridge, pp. 34–52.

Barbour, A. & Mollison, D. (1990), Epidemics and random graphs, *in* J.-P. Gabriel, C. Lefèvre & P. Picard, eds, 'Lecture Notes in Biomathematics, number 86', Springer, Berlin, pp. 86–89.

Barlow, N. (1993), 'A model for the spread of bovine Tb in New Zealand possum populations', *J. Appl. Ecol.* **30**, 156–164.

Bartlett, M. (1956), 'Deterministic and stochastic models for recurrent epidemics', *Proc. Third Berkeley Symp. on Math. Statist. and Prob.* **4**, 81–109.

Bartlett, M. (1957), 'Measles periodicity and community size', *J. R. Statist. Soc.* **A120**, 48–70.

Bartlett, M. (1960*a*), 'The critical community size for measles in the United States', *J. R. Statist. Soc.* **A123**, 37–44.

Bartlett, M. (1960*b*), *Stochastic Population Models in Ecology and Epidemiology*, Methuen, London.

Bartoszyński, R. (1967), 'Branching processes and the theory of epidemics', *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.* **4**, 259–269.

Becker, N. (1989), *Analysis of Infectious Disease Data*, Chapman and Hall, London.

Bentil, D. & Murray, J. (1993), 'Modelling bovine tuberculosis in badgers', *J. Anim. Ecol.* **62**, 239–250.

Bloom, B. & Murray, C. (1992), 'Tuberculosis: commentary on a reemergent killer', *Science* **257**, 1055–1064.

Blower, S., McLean, A., Porco, T., Small, P., Hopewell, P., Sanchez, M. & Moss, A. (1995), 'The intrinsic transmission dynamics of tuberculosis epidemics', *Nature Med.* **1**, 815–821.

Blower, S., Small, P. & Hopewell, P. (1996), 'Control strategies for tuberculosis epidemics: new models for old problems', *Science* **273**, 497–500.

Brewer, T., Heymann, S., Colditz, G., Wilson, M., Auerbach, K., Kane, D. & Fineberg, H. (1996), 'Evaluation of tuberculosis control policies using computer simulation', *J. Amer. Med. Assoc.* **276**, 1898–1903.

Campbell, A. (1974), 'Relapse in patients with tuberculosis', *Bull. Intern. Union Against Tuberc.* **1**, 219–222.

Castillo-Chavez, C. & Feng, Z. (1997), 'To treat or not to treat: the case of tuberculosis', *J. Math. Biol.* **35**, 629–656.

Chan, S. & Yew, W. (1998), Chemotherapy, *in* P. Davies, ed., 'Clinical Tuberculosis', Chapman and Hall, London, pp. 243–263. Second Edition.

China Tuberculosis Control Collaboration (1996), 'Results of directly observed short-course chemotherapy in 112842 Chinese patients with smear-positive tuberculosis', *Lancet* **347**, 358–362.

Chorba, R. & Sanders, J. (1971), 'Planning models for tuberculosis control programs', *Health Serv. Res.* **6**, 144–164.

Clancy, L. (1990), 'Infectiousness of tuberculosis', *Bull. Intern. Union Against Tub. Lung Dis.* **65**, 70.

Cohen, F. (1995), The epidemiology of tuberculosis, *in* F. Cohen & J. Durham, eds, 'Tuberculosis. A Sourcebook for Nursing Practice', Springer Publishing Company, New York, pp. 29–51.

Cohen, F. & Durham, J. (1995), Tuberculosis: an introduction, *in* F. Cohen & J. Durham, eds, 'Tuberculosis. A Sourcebook for Nursing Practice', Springer Publishing Company, New York, pp. 3–14.

Cohen, F., Harriman, C. & Madsen, L. (1995), Symptoms and diagnosis of tuberculosis, *in* F. Cohen & J. Durham, eds, 'Tuberculosis. A Sourcebook for Nursing Practice', Springer Publishing Company, New York, pp. 55–66.

Comstock, G. & Geiter, L. (1994), Prophylaxis, *in* D. Schlossberg, ed., 'Tuberculosis', Springer-Verlag, New York, pp. 89–94. Third Edition.

Cox, D. & Miller, H. (1965), *The Theory of Stochastic Processes*, Chapman and Hall, London.

Crofton, J., Horne, N. & Miller, F. (1992), *Clinical Tuberculosis*, MacMillan Education, London.

Daley, D. & Gani, J. (1999), *Epidemic Modelling*, Cambridge University Press, Cambridge.

Daniels, H. (1967), 'The distribution of the total size of an epidemic', *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.* **4**, 281–293.

Daniels, H. (1974), 'The maximum size of a closed epidemic', *Adv. Appl. Prob.* **6**, 607–621.

Darroch, J. & Seneta, E. (1965), 'On quasi-stationary distributions in absorbing discrete-time finite Markov chains', *J. Appl. Prob.* **2**, 88–100.

Darroch, J. & Seneta, E. (1967), 'On quasi-stationary distributions in absorbing continuous-time finite Markov chains', *J. Appl. Prob.* **4**, 192–196.

De Cock, K., Binkin, N., Zuber, P., Tappero, J. & Castro, K. (1996), 'Research issues involving HIV-associated tuberculosis in resource-poor countries', *J. Amer. Med. Assoc.* **276**, 1502–1507.

Diekmann, O. & Heesterbeek, J. (2000), *Mathematical Epidemiology of Infectious Diseases*, John Wiley, New York.

Diekmann, O., Heesterbeek, J. & Metz, J. (1990), 'On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations', *J. Math. Biol.* **28**, 365–382.

Dietz, K. (1967), 'Epidemics and rumours: a survey', *J. R. Statist. Soc.* **A130**, 505–528.

Dietz, K. & Schenzle, D. (1985), Mathematical models for infectious disease statistics, *in* A. Atkinson & S. Fienberg, eds, 'A Celebration of Statistics', Springer Verlag, New York, pp. 167–204.

Dolin, P., Raviglione, M. & Kochi, A. (1994), 'Global tuberculosis incidence and mortality during 1990–2000', *Bull. World Health Org.* **72**, 213–220.

Downton, F. (1967), 'Epidemics with carriers: a note on a paper of Dietz', *J. Appl. Prob.* **4**, 264–270.

Dye, C., Garnett, G., Sleeman, K. & Williams, B. (1998), 'Prospects for worldwide tuberculosis control under the WHO DOTS strategy', *Lancet* **352**, 1886–1891.

Dye, C., Scheele, S., Dolin, P., Pathania, V. & Raviglione, M. (1999), 'Global burden of tuberculosis', *J. Amer. Med. Assoc.* **282**, 677–686.

Enarson, D. & Rouillon, A. (1998), The epidemiological basis of tuberculosis, *in* P. Davies, ed., 'Clinical Tuberculosis', Chapman and Hall, London, pp. 35–52. Second Edition.

Ferebee, S. (1970), 'Controlled chemoprophylaxis trials in tuberculosis. A general review', *Adv. Tub. Res.* **17**, 28–106.

Fine, P. (1995), 'Variation in protection by BCG: implications of and for heterologous immunity', *Lancet* **346**, 1339–1345.

Flaspohler, D. (1974), 'Quasi-stationary distributions for absorbing continuous-time denumerable Markov chains', *Ann. Inst. Statist. Math.* **26**, 351–356.

Gani, J. (1965), 'On a partial differential equation of epidemic theory. I.', *Biometrika* **52**, 617–622.

Gill, P., Murray, W., Saunders, M. & Wright, M. (1986), User's guide for NPSOL (Version 4.0), Technical report, Department of Operations Research, Stanford University. Report SOL 86-2.

Goh, E. & Fam, K. (1981), 'A dynamic model of tuberculosis epidemiology for Singapore', *Ann. Acad. Med.* **10**, 40–49.

Good, P. (1968), 'The limiting behavior of transient birth and death processes conditioned on survival', *J. Austr. Math. Soc.* **8**, 712–722.

Greenwood, M. (1931), 'On the statistical measure of infectiousness', *J. Hyg., Camb.* **31**, 336–351.

Grosset, J. (1989), 'Present status of chemotherapy for tuberculosis', *Rev. Infect. Dis.* **11**, S347–S352.

Grzybowski, S. & Enarson, D. (1978), 'The fate of cases of pulmonary tuberculosis under various treatment programmes', *Bull. Intern. Union Against Tuberc.* **53**, 70–75.

Hamer, W. (1906), 'Epidemic disease in England', *Lancet* **1**, 733–739.

Heesterbeek, J. (1992), 'R_0', PhD thesis, Centrum voor Wiskunde en Informatica, Amsterdam.

Herbert, J. (1998), 'Stochastic processes for parasite dynamics', PhD thesis, University of London.

Herbert, J. & Isham, V. (2000), 'Stochastic host-parasite interaction models', *J. Math. Biol.* **40**, 343–371.

Hitchcock, S. (1986), 'Extinction probabilities in predator-prey models', *J. Appl. Prob.* **23**, 1–13.

Horne, N. (1990), *Modern Drug Treatment of Tuberculosis*, The Chest, Heart and Stroke Association, London. Seventh Edition.

Huebner, R. (1996), BCG vaccination in the control of tuberculosis, *in* T. Shinnick, ed., 'Tuberculosis', Springer, Berlin, pp. 263–282.

Isham, V. (1991), 'Assessing the variability of stochastic epidemics', *Math. Biosciences* **107**, 209–224.

Isham, V. (1993), 'Stochastic models for epidemics with special reference to AIDS', *Ann. Appl. Prob.* **3**, 1–27.

Jacquez, J. & O'Neill, P. (1991), 'Reproduction numbers and thresholds in stochastic epidemic models. I. Homogeneous populations', *Math. Biosciences* **107**, 161–186.

Jacquez, J. & Simon, C. (1993), 'The stochastic SI model with recruitment and deaths. I. Comparison with the closed SIS model', *Math. Biosciences* **117**, 77–125.

Joesoef, M., Remington, P. & Tjiptoherijanto, P. (1989), 'Epidemiological model and cost-effectiveness analysis of tuberculosis treatment programmes in Indonesia', *Intern. J. Epid.* **18**, 174–179.

Kanai, K. (1990), *Introduction to Tuberculosis and Mycobacteria*, South East Asian Information Center, International Medical Foundation of Japan.

Kendall, D. (1956), 'Deterministic and stochastic epidemics in closed populations', *Proc. Third Berkeley Symp. on Math. Statist. and Prob.* **4**, 149–165.

Kendall, W. & Saunders, I. (1983), 'Epidemics in competition II: the general epidemic', *J. R. Statist. Soc.* **B45**, 238–244.

Kermack, W. & McKendrick, A. (1927), 'A contribution to the mathematical theory of epidemics', *Proc. R. Soc. Lond.* **A115**, 700–721.

Kingman, J. (1963), 'The exponential decay of Markov transition probabilities', *Proc. Lond. Math. Soc.* **13**, 337–358.

Kochi, A., Vareldzis, B. & Styblo, K. (1993), 'Multidrug-resistant tuberculosis and its control', *Res. Microb.* **144**, 104–110.

Kribs-Zaleta, C. (2001), 'Center manifolds and normal forms in epidemic models', *J. Math. Biol.* . To appear.

Kribs-Zaleta, C. & Velasco-Hernández, J. (2000), 'A simple vaccination model with multiple endemic states', *Math. Biosciences* **164**, 183–201.

Krishnamurthy, V. & Chaudhuri, K. (1990), 'Risk of pulmonary tuberculosis associated with exogenous reinfection and endogenous reactivation in a South Indian rural population – A mathematical estimate', *Ind. J. Tub.* **37**, 63–67.

Krishnamurthy, V., Nair, S., Gothi, G. & Chakraborty, A. (1976), 'Incidence of tuberculosis among newly infected population and in relation to the duration of infected status', *Ind. J. Tub.* **23**, 3–7.

Kurtz, T. (1970), 'Solutions of ordinary differential equations as limits of pure jump Markov processes', *J. Appl. Prob.* **7**, 49–58.

Kurtz, T. (1971), 'Limit theorems for sequences of jump Markov processes approximating ordinary differential processes', *J. Appl. Prob.* **8**, 344–356.

Kurtz, T. (1981), *Approximation of Population Processes*, SIAM, Philadelphia.

LaScolea, L. & Rangoonwala, R. (1996), *Quinolones in Pulmonary Tuberculosis Management*, Hoechst, Frankfurt.

Lefèvre, C. (1990), Stochastic epidemic models for S-I-R infectious diseases: a brief survey of the recent general theory, *in* J.-P. Gabriel, C. Lefèvre & P. Picard, eds, 'Lecture Notes in Biomathematics, number 86', Springer, Berlin, pp. 1–12.

Matis, J. & Kiffe, T., eds (2000), *Stochastic Population Models*, number 145 *in* 'Lecture Notes in Statistics', Springer, New York.

May, R. (1995), 'The rise and fall and rise of tuberculosis', *Nature Med.* **1**, 752.

McKendrick, A. (1926), 'Applications of mathematics to medical problems', *Proc. Edin. Math. Soc.* **14**, 98–130.

Mollison, D. (1995), *Epidemic Models. Their Structure and Relation to Data*, Cambridge University Press, Cambridge.

Mollison, D., Isham, V. & Grenfell, B. (1994), 'Epidemics: models and data', *J. R. Statist. Soc.* **A157**, 115–149.

Murray, C., DeJonghe, E., Chum, H., Nyangulu, D., Salomao, A. & Styblo, K. (1991), 'Cost effectiveness of chemotherapy for pulmonary tuberculosis in three sub-Saharan African countries', *Lancet* **338**, 1305–1308.

Murray, C., Styblo, K. & Rouillon, A. (1993), Tuberculosis, *in* D. Jamison, W. Mosley, A. Measham & J. Bodadilla, eds, 'Disease Control Priorities in Developing Countries', Oxford University Press, Oxford, pp. 233–259.

Nair, M. & Pollett, P. (1993), 'On the relationship between μ-invariant measures and quasi-stationary distributions for continuous-time Markov chains', *Adv. Appl. Prob.* **25**, 82–102 and 717–719.

Oppenheim, I., Shuler, K. & Weiss, G. (1977), 'Stochastic theory of nonlinear rate processes with multiple stationary states', *Physica* **88A**, 191–214.

Pakes, A. (1973), 'Conditional limit theorems for a left-continuous random walk', *J. Appl. Prob.* **10**, 39–53.

Pan American Health Organization (1986), *Tuberculosis Control: A Mannual on Methods and Procedures for Integrated Programmes*, Scientific Publication, No. 498, Washington.

Pollett, P. (1988), 'Reversibility, invariance and μ-invariance', *Adv. Appl. Prob.* **20**, 600–621.

Pollett, P. & Stewart, D. (1994), 'An efficient procedure for computing quasi-stationary distributions of Markov chains with sparse transition structure', *Adv. Appl. Prob.* **26**, 68–79.

Pollett, P. & Vere-Jones, D. (1992), 'A note on evanescent processes', *Austr. J. Statist.* **34**, 531–536.

Raviglione, M., Rieder, H., Styblo, K., Khomenko, A., Esteves, K. & Kochi, A. (1994), 'Tuberculosis trends in Eastern Europe and the former USSR', *Tuber. Lung Dis.* **75**, 400–416.

Reinhard, H. (1986), *Differential Equations*, North Oxford Academic Publishers, London.

Renshaw, E. (1993), *Modelling Biological Populations in Space and Time*, Cambridge University Press, Cambridge.

Renshaw, E. (1998), 'Saddlepoint approximations for stochastic processes with truncated cumulant generating functions', *IMA J. Math. Appl. Med. Biol.* **15**, 41–52.

Renshaw, E. (2001), 'Applying the saddlepoint approximation to bivariate stochastic processes'. To appear.

Reuter, G. (1957), 'Denumerable Markov processes and the associated contraction semigroups on *l*', *Acta Mathematica* **97**, 1–46.

Reuter, G. (1961), 'Competition processes', *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.* **2**, 421–430.

ReVelle, C., Feldmann, F. & Lynn, W. (1969), 'An optimization model of tuberculosis epidemiology', *Management Science* **16**, B190–B211.

Ridler-Rowe, C. (1967), 'On a stochastic model of an epidemic', *J. Appl. Prob.* **4**, 19–33.

Ross, R. (1908), *Report on the Prevention of Malaria in Mauritius*, London.

Rouillon, A. & Waaler, H. (1976), 'BCG vaccination and epidemiological situation', *Adv. Tuberc. Res.* **19**, 64–126.

Rusu, G. (1973*a*), 'A Markovian model in tuberculosis epidemiology', *Ftiziologia* **22**, 585–592. (in Romanian).

Rusu, G. (1973*b*), 'An operational research model in tuberculosis prevention', *Ftiziologia* **22**, 593–598. (in Romanian).

Schulzer, M., Enarson, D., Grzybowski, S., Hong, Y., Kim, S. & Lin, T. (1987), 'An analysis of pulmonary tuberculosis data in Taiwan and Korea', *Intern. J. Epid.* **16**, 584–589.

Schulzer, M., Radhamani, M., Grzybowski, S., Mak, E. & Fitzgerald, J. (1994), 'A mathematical model for the prediction of the impact of HIV infection on tuberculosis', *Intern. J. Epid.* **23**, 400–407.

Sellke, T. (1983), 'On the asymptotic distribution of the size of a stochastic epidemic', *J. Appl. Prob.* **20**, 390–394.

Seneta, E. (1966), 'Quasi-stationary behaviour in the random walk with continuous time', *Austr. J. Statist.* **8**, 92–98.

Shekleton, M. (1995), The etiology, transmission, and pathogenesis of tuberculosis, *in* F. Cohen & J. Durham, eds, 'Tuberculosis. A Sourcebook for Nursing Practice', Springer Publishing Company, New York, pp. 15–27.

Siskind, V. (1965), 'Miscellanea. A solution of the general stochastic epidemic', *Biometrika* **52**, 613–616.

Small, P., Shafer, R., Hopewell, P., Singh, S., Murphy, M., Desmond, E., Sierra, M. & Schoolnik, G. (1993), 'Exogenous reinfection with multidrug-resistant *Mycobacterium tuberculosis* in patients with advanced HIV infection', *New Engl. J. Med.* **328**, 1137–1144.

Smith, P. & Fine, P. (1998), BCG vaccination, *in* P. Davies, ed., 'Clinical Tuberculosis', Chapman and Hall, London, pp. 417–431. Second Edition.

Soper, H. (1929), 'The interpretation of periodicity in disease prevalence', *J. R. Statist. Soc.* **A92**, 34–73.

Springett, V. (1971), 'Ten-year results during the introduction of chemotherapy for tuberculosis', *Tubercle* **52**, 73–87.

Stead, W. & Dutt, A. (1988), Changing faces of clinical tuberculosis, *in* M. Bendinelli & H. Friedman, eds, 'Mycobacterium Tuberculosis: Interaction with the Immune System', Plenum Press, New York, pp. 371–388.

Stirzaker, D. (1975), 'A perturbation method for the stochastic recurrent epidemic', *J. Inst. Maths Applics* **15**, 135–160.

Styblo, K. (1983), 'Tuberculosis and its control: lessons to be learned from past experience, and implications for leprosy control programmes', *Ethiop. Med. J.* **21**, 101–122.

Styblo, K. (1989), 'Overview and epidemiologic assessment of the current global tuberculosis situation with an emphasis on control in developing countries', *Rev. Infect. Dis.* **11**, S339–S346.

Styblo, K. (1991), 'Epidemiology of tuberculosis', *Royal Netherl. Assoc. Tub., Sel. Pap.* **24**, 9–136.

Sutherland, I., Švandová, E. & Radhakrishna, S. (1982), 'The development of clinical tuberculosis following infection with tubercle bacilli', *Tubercle* **63**, 255–268.

Tan, W. & Hsu, H. (1989), 'Some stochastic models of AIDS spread', *Statistics in Medicine* **8**, 121–136.

Trefny, J. & Hejdova, E. (1982), 'A model of the epidemiology of tuberculosis in the Czech Socialist Republic', *Bull. Intern. Union Against Tuberc.* **57**, 206–211.

van Doorn, E. (1991), 'Quasi-stationary distributions and convergence to quasi-stationarity of birth-death processes', *Adv. Appl. Prob.* **23**, 683–700.

Vere-Jones, D. (1969), 'Some limit theorems for evanescent processes', *Austr. J. Statist.* **11**, 67–78.

Vynnycky, E. (1996), 'An investigation of the transmission dynamics of *M. tuberculosis*', PhD thesis, University of London.

Vynnycky, E. & Fine, P. (1997), 'The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection', *Epidemiol. Infect.* **119**, 183–201.

Waaler, H., Geser, A. & Andersen, S. (1962), 'The use of mathematical models in the study of the epidemiology of tuberculosis', *Amer. J. Public Health* **52**, 1002–1013.

Waaler, H. & Piot, M. (1969), 'Use of an epidemiological model for estimating the effectiveness of tuberculosis control measures. Sensitivity of the effectiveness of tuberculosis control measures to the coverage of the population', *Bull. World Health Org.* **41**, 75–93.

Waaler, H. & Piot, M. (1970), 'Use of an epidemiological model for estimating the effectiveness of tuberculosis control measures. Sensitivity of the effectiveness of tuberculosis control measures to the social time preference', *Bull. World Health Org.* **43**, 1–16.

Whittle, P. (1955), 'The outcome of a stochastic epidemic – A note on Bailey's paper', *Biometrika* **42**, 116–122.

Whittle, P. (1957), 'On the use of the normal approximation in the treatment of stochastic processes', *J. R. Statist. Soc.* **B19**, 268–281.

Wolff, R. (1989), *Stochastic Modeling and the Theory of Queues*, Prentice Hall, New Jersey.

World Health Organization (1993), *Treatment of Tuberculosis: Guidelines for National Programmes*, WHO, Geneva.

Yaglom, A. (1947), 'Certain limit theorems of the theory of branching stochastic processes', *Dokl. Akad. Nauk SSSR (N.S.)* **56**, 795–798. (in Russian).