Estimating The Undiagnosed HIV-Positive Population
A Mathematical Modeling Study

by

Rebecca de Boer

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

In this thesis a mathematical model of HIV transmission and diagnosis is used to estimate the total size of the HIV-positive population and the HIV incidence from HIV case report data for the Province of Alberta.

Worldwide, estimates of the size of the HIV-positive population are used to allocate medical resources and target disease prevention efforts, while estimates of HIV incidence are used to evaluate the effectiveness of intervention programs and track changes in risk behaviours. Many HIV surveillance programs are based on reports of newly diagnosed cases. Estimating the total size of the HIV-positive population from this data is challenging as those who are HIV-positive but have not been diagnosed are not included. Furthermore, trends in HIV diagnosis do not reflect trends in incidence as the length of time newly diagnosed HIV patients have been infected is usually unknown.

Fitting the model used in this thesis is complicated by the presence of non-identifiable parameters. Non-identifiable parameters occur when all parameter values on a surface in the parameter space have identical model outcomes for the quantities represented in the data. Methods for systematic detection of this behaviour and resolution of non-identifiabilities are discussed in a general modelling framework and applied to the HIV model for the assessment of the Province of Alberta data.

Interval estimates for all parameters are obtained using an iterated Markov chain Monte Carlo (MCMC) method and the resulting fitted model is validated. The validated model is used to produce estimates of the total size of the HIV-positive population including those who have not been diagnosed for the years 2001 to 2020. Estimates of HIV incidence, time from infection to diagnosis, and the size of the undiagnosed population are also computed using the model. Uncertainty and sensitivity analysis

are used to determine how much uncertainty remains in these estimates and which parameters are most important to the model outcomes. Finally, the model is used to simulate several potential intervention strategies to reduce HIV incidence in the province.

The potential impact of antiretroviral drug resistant strains of HIV on a hypothetical "treatment as prevention" program in the context of a generalized HIV epidemic is studied using another model. This model includes the development and transmission of drug resistant viral strains. Sensitivity and uncertainty analyses are used to explore the potential outcomes.

Finally, the asymptotic behaviour of a simple disease model similar to the Alberta HIV model, but using more general forms of population dependent transmission, is analyzed mathematically. It is shown that for some types of population dependence this model can display complicated dynamical behaviours including backward bifurcations and Hopf bifurcations.

Preface

The data used in Chapters 2 to 4 is provided by the Surveillance and Assessment Branch of Alberta Health under a memorandum of understanding with the Information Research Laboratory at the University of Alberta.

The model described in Section 5.1 of this thesis was developed as part of an interdisciplinary collaboration at the University of Alberta lead by Dr. S. Houston of the Faculty of Medicine and Department of Public Health and Dr. M. Li of the Department of Mathematical and Statistical Sciences. I was responsible for model analysis including sensitivity analysis, uncertainty analysis and model simulations. I also contributed to model design and parameter estimation.

Section 5.2 of this thesis has been published as R. de Boer and M. Li, *Density dependence in disease incidence and its impacts on transmission dynamics*, Canadian Applied Mathematics Quarterly, 19 (2011). I was responsible for the mathematical analysis and the majority of the manuscript composition. M. Li was involved with the formulation of the problem and assisted with the manuscript composition.

Acknowledgements

I would like to thank the Alberta Health Surveillance and Assessment Branch for their generosity in providing the HIV data used in this thesis. In particular, thanks to Michael Sanderson and Larry Svenson for their assistance in acquiring and formatting the data along with their comments at the beginning of this project.

I would also like to thank my supervisor Dr Michael Li for his support and his guidance throughout this project.

Finally, I would like to thank my husband, Johnwill, for always being encouraging, and Daniel for just being himself.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Information about the size of human immunodeficiency virus (HIV) epidemics is vital for public health workers worldwide as this information can be used to allocate medical resources, target disease prevention efforts, and evaluate the effectiveness of interventions. Quantities of particular interest include the size of HIV-positive populations and the yearly number of new infections. It is rarely possible to measure these quantities directly. Instead they are estimated from HIV surveillance data.

In the Province of Alberta, as in many other industrialized contexts, HIV surveillance is based on reports of diagnosed cases. This type of data is relatively easy to collect, but is challenging to use to estimate the overall size of the HIV-positive population. This challenge is caused by the fact that case report data only includes the part of the HIV-positive population that has already been diagnosed. Those who are HIV-positive but undiagnosed are an important component of the HIV-positive population as they are the source of many new infections. Information about the size of the undiagnosed population can be used in healthcare planning and to target and evaluate diagnosis programs. Estimating the yearly number of new HIV cases is also challenging using case report data. HIV has a relatively long period with few symptoms and diagnosis may occur several years after infection, therefore case report data does not directly reflect trends in new infections.

In this thesis a method is developed for estimating the total size of the HIV-positive population, the size of the undiagnosed population, and the number of yearly new infections from case report data.

1.1 The Size of an Epidemic

Two different quantities are often used to describe the size of an epidemic such as HIV: *prevalence* and *incidence*.

Prevalence: The proportion of the population infected with a disease is the *prevalence*. This quantity may be reported as a fraction or a percentage. In the case of HIV it

is most commonly reported as a number per 100 000 population. However it is reported, the prevalence describes the total number of people infected as a fraction of the total population.

Incidence: The number of new infections per unit time is the *incidence*. For HIV, the time unit used is usually one year. The incidence may also be reported as the proportion of the susceptible population who become infected in a fixed amount of time. In either case, the incidence describes how quickly new cases are occurring.

In the case of a disease like HIV which causes permanent infection which patients may survive for many years, the prevalence does not always give a good sense of the current state of disease transmission. Many prevalent cases have been infected for years and a change in incidence may not impact the prevalence significantly for several years. However, prevalence does indicate how much of the population is affected by HIV.

1.1.1 A Method for Estimating HIV Prevalence and Incidence

In the Province of Alberta and many other places worldwide, case reports for HIV and AIDS are routinely recorded for HIV surveillance. As has already been discussed, this type of data poses a problem in estimation of overall HIV incidence and prevalence because HIV infections may remain undiagnosed for many years and the size of the undiagnosed population is unknown. In this thesis, a novel method of solving this problem and producing estimates of HIV prevalence and incidence from case report data will be developed.

The method uses a deterministic, population based model of HIV transmission and diagnosis to describe an HIV epidemic. The model is calibrated using case report data. Once it has been calibrated the model is used to produce estimates of the total size of the HIV-positive population and HIV incidence, including past trends and future projections. The model may also be used to estimate other features of disease transmission and diagnosis such as the size of the undiagnosed HIV-positive population and the fraction of the undiagnosed population that has been undiagnosed for many years. The model can be used to simulate the potential results of intervention programs that target HIV transmission or diagnosis.

An introduction to the most common methods used worldwide in monitoring HIV prevalence and incidence is found in Section 1.2. The majority of these methods require data from HIV prevalence surveys. This data may be the result of cross-sectional surveys sampling from the entire population, or may be collected at sentinel surveillance sites such as health clinics. Collecting and interpreting this data is a challenge in itself. Cross-sectional surveys are large undertakings which cannot be repeated frequently while sentinel surveillance covers only specific sub-populations.

Prevalence surveys do not directly provide information about HIV incidence which must be estimated separately. The Spectrum package is a tool for HIV incidence esti-

mation developed by the Joint United Nations Programme on HIV/AIDS (UNAIDS). Spectrum uses data from HIV prevalence surveys to fit a mathematical model and produce estimates of HIV incidence along with past trends and future projections. Unlike the method developed in this thesis, the Spectrum model cannot be calibrated using case report data and cannot directly make estimates regarding the undiagnosed population.

The only commonly used method for HIV incidence estimation using case report data is back-projection. The method uses a statistical model for the time from infection to diagnosis and attempts to correct case report data for diagnosis delay. This produces estimates of incidence trends for the past. As many of those who have been recently infected will not yet be diagnosed, back projection is best used retrospectively. Unlike the model used in this thesis it does not allow for projection into the future.

The model that will be used in this thesis has some similarity to the models used by Wilson [142] and Bezemer [14]. However both of these authors use models with multiple disease stages and must determine values for many more parameters. The model used by Wilson is not calibrated to surveillance data and is not intended to provide prevalence or incidence estimates. The model used by Bezemer is calibrated using HIV and AIDS reports, but does not include HIV transmission dynamics.

1.2 Literature Review

A number of papers have recently been published discussing and comparing multiple methods for estimating the size of HIV/AIDS epidemics. The Working Group on Estimation of HIV Prevalence in Europe focuses on methods that can be used to estimate the number of HIV-positive people who are undiagnosed including a comparison of several back-projection based methods [144]. Brookmeyer discusses estimation and measurement methods more generally including a discussion of biomarker based methods for estimating HIV incidence [22]. Other papers focus on particular national contexts. For Kenya and Uganda, several methods to estimate HIV incidence are compared including Spectrum, sequential prevalence surveys, the BED immunoassay, and a cohort study [81]. For the Netherlands, several methods to estimate HIV prevalence are compared including the UNAIDS Workbook method, multiparameter evidence synthesis, and Spectrum [133]. These methods and others will be discussed in further detail in this section.

The World Health Organization defines three different epidemic types.

- In *low level epidemics*, HIV has not spread significantly in any subpopulation including those at high risk such as intravenous drug users, men who have sex with men, or sex workers and their clients.
- In *concentrated epidemics*, HIV is well established in subpopulations at high risk, but is not commonly transmitted to the general population.

- In *generalized epidemics*, HIV is established in the general population and many cases are not related to individuals at increased risk.

The type of epidemic occurring in a region determines which methods of estimating prevalence and incidence are most appropriate[145].

1.2.1 Prevalence

Several well established methods can be used to estimate HIV prevalence. These methods do not, by themselves, give any information about HIV incidence but may provide information required by methods to estimate incidence.

Sentinel Surveillance and The Direct Method: In much of the world, the most readily available source of data on HIV prevalence comes from the testing of blood samples collected at antenatal clinics. In generalized epidemics, sexually active women of childbearing age are at moderate to high risk for HIV and the HIV prevalence in this population is assumed to be correlated with HIV prevalence in the general population. Thus many national estimates of HIV prevalence depend heavily on information from sentinel surveillance based in antenatal clinics.

In low level and concentrated epidemics, other surveillance programs can be targeted at populations at high risk to provide more information about the overall epidemic. For example, surveillance programs using sexually transmitted infections clinics can provide information about all those who are sexually active including both heterosexual men and men who have sex with men (MSM). Other surveillance programs may target drug users or sex workers. This type of data can be used directly to examine prevalence within the particular subpopulations being surveyed. For example, surveillance programs often target injection drug users [103, 127], men who have sex with men [85, 84], and sex workers [111, 3].

Prevalence data for groups at higher risk can be combined with estimates of the size of the relevant subpopulations to determine an overall prevalence estimate. This method is sometimes called the *direct method* and is the basis for the UNAIDS Workbook, a spreadsheet based guide to using the direct method to construct an estimate of HIV prevalence for a region. The UNAIDS Workbook has been recommended for use in low level and concentrated epidemics [91]. China and Ukraine, among others, have used the UNAIDS Workbook to produce prevalence estimates [140, 83] while the United Kingdom has used a customized version of the direct method [97].

Multiparameter Evidence Synthesis: An alternative to the direct method for combining surveillance data for high risk groups into overall estimates of HIV prevalence is *multiparameter evidence synthesis* [1]. This method uses similar data to the direct method, but defines a Bayesian framework and fits the relevant prevalence parameters

using a Markov chain Monte Carlo (MCMC) method. This method is intended to provide a more careful quantification of uncertainty than the direct method allows and it can also be used in cases where some of the data has known or suspected biases. Multiparameter evidence synthesis has been applied in the United Kingdom [108] and the Netherlands [133].

Cross-Sectional Surveys: Another approach has been to include HIV testing in large scale surveys at the national level. These projects require selecting a representative sample of the population and usually involve both an interview and collection of specimens for HIV testing. National demographic and health surveys including HIV testing have been done in a number of different countries since the year 2000 [18, 98, 46]. A similar survey was performed in New York City in 2004 [102].

Cross-sectional surveys such as these avoid some of the concerns with basing HIV prevalence estimates on antenatal clinic surveillance, but they may still be subject to bias. In particular, response bias may be a factor in cross-sectional surveys as those at highest risk may not be available to complete the survey [8]. Furthermore, those who are already aware that they are HIV-positive may refuse to participate. Nonetheless, cross-sectional surveys are often very successful, with only a small number of non-responders [98].

Cross-sectional surveys are expensive and can be difficult to conduct, requiring significant human and laboratory resources. As a result, cross-sectional surveys cannot be repeated frequently and, while they provide a good estimate of HIV prevalence, they cannot be used to track short term changes in the epidemic. They are particularly difficult in low-level epidemics where infection is relatively rare and a large sample will be required to accurately estimate prevalence [47]. In concentrated epidemics some of the populations most at risk for HIV are difficult to include in large scale surveys which often use household based sampling procedures.

1.2.2 Incidence

Measuring HIV incidence is more difficult than measuring prevalence. Measuring incidence requires information about new cases which are often not easy to distinguish from long-standing cases as regular testing is not common and many HIV cases are not diagnosed until long after infection.

Repeated Prevalence Estimates: One method for constructing incidence estimates requires age specific prevalence estimates to be repeated at multiple points in time. Assumptions about population aging and deaths are applied to estimate how many new cases have occurred during the time between the prevalence estimates. This process has been applied to estimate HIV incidence in South Africa [114] and a number of other countries [62].

The assumptions about deaths in the HIV-positive population are particularly crucial to this method and changing death rates due to the increasing availability of treatment can bias the results [61]. For example, an increase in disease prevalence could be caused by more new infections or it could be caused by infected persons surviving longer due to treatment. If the death rate is assumed to be constant when it is actually decreasing, this method will overestimate HIV incidence. Reliable information on the death rate in the HIV-positive population is not always available, limiting the usefulness of this method.

Disease Progression Biomarkers: Another method of estimating HIV incidence involves laboratory tests for infection recency. Such tests indicate not only whether or not a specimen is HIV-positive but also whether or not the HIV infection is recent.

One early technique used a less sensitive *detuned* HIV antibody assay to discriminate recently acquired infections from longer term infections. This technique uses the fact that those who have been recently infected generally have a lower concentration of HIV antibodies than those who have been infected for a longer period. Unfortunately, this technique may also classify many people with advanced AIDS as having a recently acquired infection as their HIV antibodies are waning due to immune system damage. Furthermore, those who have long standing infections that are successfully suppressed using antiretroviral therapy (ART) may also appear to be recently infected using this test [27].

Other tests for recent infection use different aspects of the immune response to distinguish between recently acquired and longstanding infections. The most commonly used of these is the BED IgG-Capture Enzyme Immunoassay [9]. While the BED Immunoassay outperforms older detuning techniques, it is still known to misclassify some of those who have long term infections. When applying biomarker methods to determine HIV incidence, the possibility for false recent results must be accounted for. This may occur prior to the testing phase, so that biomarker tests are not used to determine infection recency for those who are known to have long standing infections, or statistical methods may be used to remove this type of error in the analysis phase [141].

Incorporating such tests with sentinel surveillance or cross-sectional survey methods for estimating HIV prevalence allows these survey methods to also provide an estimate of recently acquired HIV. Interpreting such estimates in terms of HIV incidence requires knowledge about the average time during which a person will test positive for recent infection. The value of this parameter can be calculated during cohort studies [105] but, as these are rare and typically small, a great deal of uncertainty may remain.

The BED Immunoassay was included in the 2005 national household survey in South Africa [113] in order to estimate HIV incidence. The BED Immunoassay is also used in the United States where data collected using this method is combined with back-projection methods for estimating HIV prevalence and incidence trends [79, 107].

Cohort Studies: The most direct method of measuring HIV incidence is to perform a cohort study in which a defined population is followed up regularly for repeated HIV testing over some time period. While this provides a direct measurement of HIV incidence over the study period, cohort studies are costly and it is difficult to repeatedly follow up with a large population. Therefore, this type of study is usually done on a smaller scale, targeting some specific subpopulation, for example agricultural workers in Kericho Kenya [115]. Cohort studies are sometimes intended to investigate the impact of particular interventions [78] or to determine the influence of specific risk factors [124]. The resulting measurements of HIV incidence may not be generalizable to a larger population. Additionally, participation in a cohort study and the associated repeated HIV testing is thought to influence behaviours and bias the resulting estimates [22].

1.2.3 Trends and Projections

In addition to estimates of epidemic size at a single point in time, information on how HIV prevalence and incidence have changed in the past and how they may continue to change in the future is extremely valuable. A number of specialized methods have been developed for determining trends in HIV incidence and prevalence.

1.2.3.1 Back-projection

Back-projection, also known as back-calculation, is one of the oldest methods for reconstructing historical HIV incidence curves. It is a statistical method using HIV and/or AIDS diagnosis data. As already mentioned, this information is routinely tracked by public health surveillance programs in the industrialized world as this data is relatively straight-forward to collect. However, the delay between infection and diagnosis requires care in interpreting this information in terms of prevalence and incidence. Newly diagnosed HIV patients may have been HIV-positive for some time depending on a number of factors including the individual's perceived level of risk and exposure to HIV testing. Additionally, newly diagnosed AIDS patients may have been HIV-positive for many years depending on their adherence to treatment.

The back projection method was originally designed for use with AIDS case report data but has been updated more recently to utilize other data sources, such as HIV case reports and infection recency information [11, 123]. The updated back-projection method involves choosing an appropriate statistical distribution for the time from infection to diagnosis. This distribution is used together with HIV diagnosis data to estimate the number of new infections in previous years. Usually the distribution of time to diagnosis is chosen as a combination of several distributions accounting for various motivations for testing such as development of symptoms or a suspected exposure [101]. These distributions must additionally be time dependent if they are to be applicable to both the beginning of the epidemic and more recent data.

It may be possible to estimate some of the parameters required to specify these distributions from data, however the available data is usually insufficient to compute all the required parameters. This may be addressed either by making additional assumptions to reduce the number of parameters required, or by including additional data. In particular, the method described in [148] uses partial infection recency data in specifying distribution parameters.

There are several versions of the back-projection method in use in different parts of the world depending on the data available locally. One method, applied in Canada and Australia, uses both HIV and AIDS reports but does not require that these two datasets be linked [149, 139]. Another method, applied in the United States, uses similar HIV and AIDS report data, but requires that the data be linked so that time between HIV and AIDS diagnosis is available [59].

The back-projection method can be used to estimate historical trends in incidence. A drawback of the back-projection method is that it is most effective at computing incidence several years in the past. Estimates for more recent years are prone to bias as some of the HIV-positive population infected recently may not yet be diagnosed. This also means that back calculation methods are not well suited to projection into the future [92].

1.2.3.2 Spectrum

The Spectrum package is a collection of tools for estimation of HIV prevalence and incidence trends developed and maintained by UNAIDS. It was originally developed as two separate tools: Spectrum and the Estimation and Projection Package (EPP) [129]. These tools are updated regularly to include new features, modelling assumptions, and estimation procedures [6, 122, 24, 25, 23, 49]. Most recently, these tools have been merged together to form a single software package known as Spectrum [7, 121]. These tools are primarily intended for use in estimating generalized epidemics but can also be used along with the UNAIDS Workbook methods discussed earlier in the case of low-level and concentrated epidemics.

From a mathematical perspective, the core of Spectrum is a deterministic disease transmission model. Some of the model parameters, such as the parameters for the AIDS survival time distribution, are fixed to appropriate values, while others, such as the time dependent birth rate, are set by the user to reflect the local situation. Four parameters remain to be fitted using data: the transmission coefficient, the time of the start of the epidemic, the fraction of the population entering the at risk group, and a behavioural response parameter. The parameters are fit using antenatal clinic sentinel surveillance data and/or prevalence data from cross-sectional surveys.

Changes to EPP/Spectrum have been implemented in response to a variety of concerns. For example, as antenatal clinic and other surveillance programs have expanded, the data on the epidemic has become more complete including data from lower risk areas.

Cross-sectional surveys are another relatively new source of HIV prevalence data. The 2005 version of EPP/Spectrum included tools to use data from an expanding surveillance system [23].

The increasing availability of antiretroviral therapy (ART) has also required changes in EPP/Spectrum. In the 2009 version, changes to the underlying model have been made to account for the effects of ART on disease transmission [24, 122]. This change added more parameter requirements to the model. The user must now provide a number of details about ART availability and effectiveness in the local context.

Despite these and other changes, EPP/Spectrum has not always been able to estimate the size of all epidemics sufficiently well. In response to this, the 2011 version of EPP/Spectrum changes the model substantially. The model is streamlined somewhat and fewer parameters are fitted, but the transmission coefficient is now stochastic, varying with some randomness over the course of the epidemic [6].

Whichever version of Spectrum is used, the calibrated model produces historical prevalence and incidence curves from beginning of the epidemic and projections into the future.

1.2.3.3 Other Models

A variety of other models of HIV epidemics have been developed. These models take a variety of forms including deterministic population based models [14], individual based microsimulation models [13], and statistical models based on back calculation [60].

The models are often tailored to specific purposes and many of these models are not intended primarily for estimating the size of the epidemic. Instead they are often created to examine the potential impact of one or more intervention programs. For example, interventions that have recently been explored using models include changing treatment guidelines [69] and effectiveness [17], antiretroviral treatment for discordant couples [44] and sex workers [134], promoting testing for men who have sex with men [142], and changes in condom usage [68]. A variety of different models have been developed to address the question of the effectiveness of treatment programs in preventing new HIV infections [40]. Other models are intended to estimate the costs and savings associated with these potential interventions [122, 135]. While only a few of these models are designed specifically to estimate HIV prevalence and incidence, comparisons of prevalence and incidence in multiple scenarios are often used to evaluate the impact of intervention programs [47].

As the model structures vary widely, so too does the information used in calibration. Often some of the parameters are fitted using data. Most often this is prevalence data such as the results of cross-sectional surveys or antenatal clinic surveillance [40]. Occasionally, HIV or AIDS case report data may be used [14, 60]. In most cases, at least some of the parameters are chosen either by assumption or with reference to other literature such as cohort studies and ART drug trials.

1.3 Thesis Outline

In the remainder of this thesis, a new method of estimating the size of an HIV epidemic will be developed and used to produce estimates for the province of Alberta. Chapters 2 to 4 detail the process of creating the model, validating it on additional data, and using it to estimate the size of the HIV epidemic in Alberta. In Chapter 2, the model is developed and calibrated to the available data using both a non-linear least squares method and a Bayesian approach. The concept of identifiability is discussed as we must determine which parameters can be successfully estimated using the data. In Chapter 3, the techniques that will be used to validate the model are explained and the model is validated using additional data from several different sources. The validation methods are based on Bayesian hypothesis testing using the results of the Bayesian fitting routine from Chapter 2. In Chapter 4, the validated model is used to estimate HIV prevalence and incidence including both historical trends and future projections. This chapter also contains estimates of a number of other quantities of interest including the size of the undiagnosed HIV-positive population, the time to diagnosis, and the potential impact of some intervention strategies. Uncertainty analysis is used to quantify the potential variation in outcomes while sensitivity analysis is used to identify parameters that may be of particular interest for interventions.

Chapter 5 includes some examples of the possibilities for disease modelling in the absence of data appropriate for model fitting techniques. Section 5.1 takes a computational approach to determining the potential effects of an intervention program for HIV on the prevalence of drug resistant viral strains. Parameters are specified from medical and epidemiological literature and tools from sensitivity and uncertainty analysis are used to explore the possible range of outcomes. Section 5.2 on the other hand, takes a theoretical approach. In this section the asymptotic behaviour of a simple disease model with a more general transmission term is investigated. A discussion of the equilibria and stability of the model is included and the existence of backwards and Hopf bifurcations is demonstrated.

Finally, Chapter 6 contains conclusions as well as some discussion of potential future directions for this work.

Chapter 2

Creating the Model

In this chapter, the model that will be used to estimate the total size of the HIV-positive population is developed. The chapter begins with an exploration of the available data followed by a description of the model and a discussion of model calibration. The model calibration step is complicated by the presence of nonidentifiable parameters. These must be detected and resolved before model calibration can proceed. Two methods of model calibration are used: nonlinear least squares which provides rapid results, and Bayesian parameter estimation which provides a natural way to validate the model and quantify parameter uncertainty. Validation and uncertainty analysis will be discussed in detail in Chapters 3 and 4. While the available data is for HIV in the Province of Alberta and estimating the size of the HIV-positive population in Alberta will be the focus for the next several chapters of this thesis, the methods described here may be applicable to other contexts.

2.1 Data

The major source of data for this project is HIV case reports and HIV deaths. Only the aggregated quantities are used, ie. number of reported cases and deaths each year. These two quantities allow the estimation of the total number of living people diagnosed with HIV, another useful quantity.

2.1.1 Province of Alberta

The data for the Province of Alberta is provided by Alberta Health and is gathered as part of the notifiable disease program in the province. The data includes cases that date back to the beginning of AIDS reporting in the province. However, HIV did not become a notifiable disease until 1998 so the data from earlier years is incomplete and consists only of AIDS patients. The last complete year included in this data is 2010. The data was acquired in May of 2011 and so only partial data is included for that year. Figures 2.1, 2.2, and 2.3 illustrate the data.

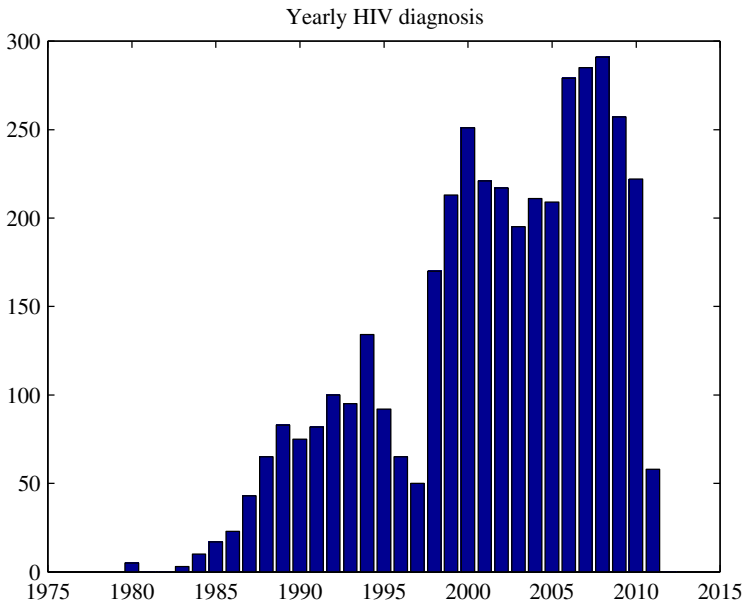


Figure 2.1: Yearly number of reported HIV and AIDS cases.

The jump in reported cases, in the late 1990s corresponds to the beginning of the requirement that HIV cases be reported. Prior to this, the reported cases were exclusively AIDS cases. The decline in deaths around the same time is probably due to increasingly effective treatments. Both of these contribute to the increasing trend in the cumulative diagnosed cases.

In order to successfully fit the model, information is also required about the size of the overall population. This is available from Statistics Canada. In particular data is available for population size [119] and annual deaths [120].

The Province of Alberta data also includes information on exposure categories. Using this additional information, more detailed models could be created to investigate questions about the sources of HIV transmission. This possibility is discussed in Section 6.2.

2.2 Proposed Model

The model that will be used for much of the remainder of this thesis is a compartmental disease transmission model described by the system of differential equations

$$\begin{aligned} \dot{S} &= \Lambda - (\beta_I I + \beta_D D)\frac{S}{N} - d_S S \\ \dot{I} &= (\beta_I I + \beta_D D)\frac{S}{N} - \alpha I - d_I I \\ \dot{D} &= \alpha I - d_D D. \end{aligned} \tag{2.1}$$

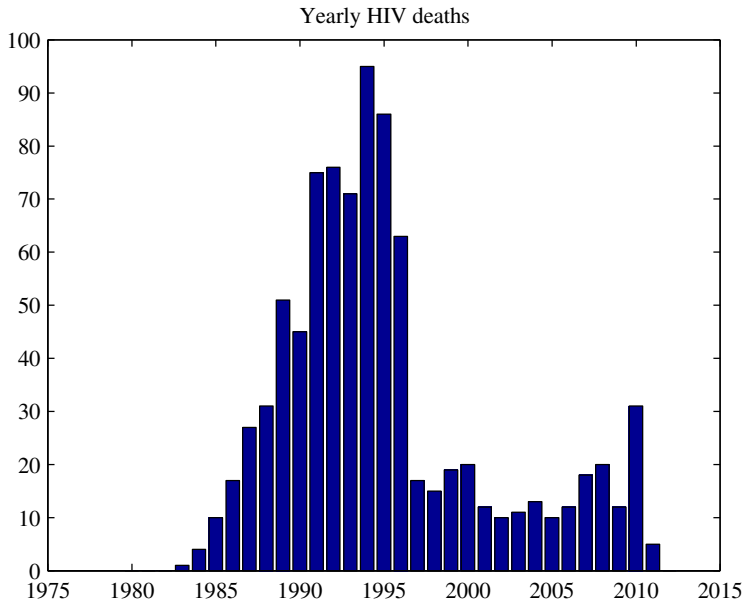


Figure 2.2: Yearly number of deaths for reported HIV and AIDS cases.

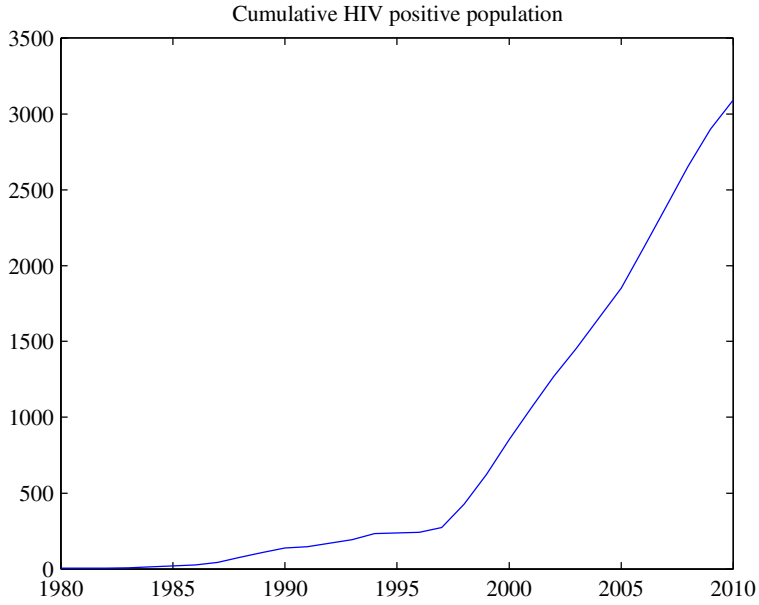


Figure 2.3: Cummulative number of diagnosed HIV and AIDS patients.

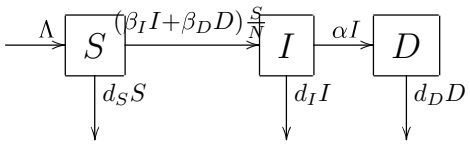


Figure 2.4: Model diagram for an HIV model including transmission, and diagnosis.

The model compartments, S, I, D, and N, are described in Table 2.1 while the parameters are summarized in Table 2.2. The model diagram for this model is illustrated in Figure 2.4. Each of the equations represents the rate of change of the size of one of the subpopulations that we are considering. The terms on the right hand side of the equations describe how the population moves between the compartments.

This model is related to the standard SI model, but has been modified to include the effects of diagnosis. In order to compare to the SI model framework the D compartment could be considered as an additional disease stage. Theoretical results exist dealing with the equilibria and stability of this type of model [55].

While these theoretical results are useful in determining the qualitative behaviour of the model on long time scales, this type of result may not be useful from a public health perspective. Discussions of equilibria and stability generally only give information about the long term behaviour of the system. This type of analysis requires that parameter values be fixed for all time and does not generally consider how long it could take for the system to approach equilibrium. On the other hand, it is not realistic to assume that parameter values will be fixed or continue to change predictably into the future. This in turn means that models are not expected to be valid over the long term. Instead we consider model behaviour in detail over shorter periods of time.

The structure of the model is chosen to allow parameter fitting using the available case report data. As this data captures diagnosis, the model must distinguish between those who are diagnosed as HIV-positive and those who are infected but have not been diagnosed. Since one of the major goals of this thesis is to estimate the size of the HIV-positive population that is not diagnosed, we will be particularly interested in the I compartment of the model. Additionally, the transmission coefficients, especially β_I, are particularly difficult to estimate directly and so these are of particular interest in the model fitting that will follow.

This model makes a number of assumptions. First, by not distinguishing between sexes, risk groups, or geographic locations we assume that it is reasonable to model the overall disease transmission scenario without these features. Although there may be important differences between groups, the simple model we use simplifies parameter fitting and may allow a clearer understanding of the results. When the results are interpreted we will have to keep in mind that the population described by the simple model is a

S	Susceptible population
I	Undiagnosed HIV-positive population
D	Diagnosed population
N	Total population $N = S + I + D$

Table 2.1: Description of model compartments for the HIV transmission model illustrated in Figure 2.4.

β_I, β_D	Transmissibility coefficients
α	Diagnosis rate
d_S, d_I, d_D,	Death rates
S_0, I_0, D_0,	Initial population sizes

Table 2.2: Description of parameters of the HIV transmission model illustrated in Figure 2.4.

combination of many different subpopulations which may have differing levels of risk. The parameter values we will end up selecting must be considered to be weighted averages of the parameter values that would be used in the case of a more complicated model structure containing these additional features. Models which divide the population into risk groups are discussed briefly in Section 6.2

Since the Alberta data does not include specific information on treatment, the model does not consider treatment separately from diagnosis. The result is that the D compartment contains all those who have been diagnosed HIV-positive. The majority are assumed to be well treated, but a few may have not yet started treatment, while others may have stopped treatment. The parameter β_D is a weighted average transmission coefficient for those who are effectively treated along with those who are simply diagnosed. Similarly, the death rate d_D is a weighted average for the diagnosed and treated populations.

2.3 Parameter Selection

In order to create a model that will describe the HIV population effectively, it is necessary to select appropriate parameter values for the model. Estimated values for some of the parameters may be available in medical and public health literature, however we will attempt to find as many of these parameter values as possible using parameter fitting techniques with the available data.

Although it may be simpler to select the values of some parameters in advance from medical and public health literature, we attempt to fit all of them in order to investigate how much information about the parameter values can be found solely from the case report data. As a side effect, this could allow us to compare estimated parameter

values to those found in medical and public health literature in order to check that the modelling results are reasonable.

Parameter selection will proceed in two steps. First, the population parameters Λ and d_S will be fit using overall population data including total population and yearly deaths for the Province of Alberta. These parameters are fit using only the S compartment of the model. The HIV-positive population is small compared to the total population of the province so it is reasonable to assume that the addition of the disease compartments will not have a significant impact the population parameters.

Some reparameterizations of the disease parameters will be used in order to enforce relationships between the parameters. Most of those who are diagnosed as HIV-positive are treated. Treatment is thought to greatly reduce the chance of transmission and so $\beta_D < \beta_I$. In order to enforce this, we use $\beta_D = a\beta_I$ with $a \ll 1$. Additionally, we look for the initial condition I_0 as a multiple of the total diagnosed population at the initial time $I_0 = rD_0$. The initial conditions D_0 and $N_0 = S_0 + I_0 + D_0$ can be computed directly from the data and are considered to be known before model fitting begins.

2.3.1 Identifiability

The parameter fitting step is complicated by the presence of nonidentifiable parameters. This is a common problem when parameter values must be estimated from data. When two or more parameters have similar effects on the model outcomes that are observed in the data, it may not be possible to determine appropriate values for these parameters. Concerns about model identifiability arise whenever mathematical models must be calibrated using data and have been noted in such diverse fields as mathematical epidemiology [43], viral dynamics [146], biochemical modelling [118], plant science, [38] economics [28], and engineering [77].

There are a number of different definitions of identifiability used by different authors, but the most common seems to be the following from [137]. Consider a general model

$$\begin{cases} \dot{x}(t,p) = f(x(t),p) \\ x(0,p) = x_0(p) \end{cases} \tag{2.2}$$

with observations

$$y(t,p) = g(x(t),p) \tag{2.3}$$

where p is a vector of parameters, which may include initial conditions for the model and any parameters required for the observation function. Both x and y may be vector or matrix valued. Then we define

Definition 1. A parameter p_i, $i = 1, \ldots, m$, is *structurally nonidentifiable* if for almost

any p^* there exists no neighbourhood $V(p^*)$ such that

$$p \in V(p^*) \text{ and } y(t,p) = y(t,p^*)\ \forall t\ \Rightarrow p_i = p_i^*.$$

In other words, if a parameter is nonidentifiable, there are many possible values for that parameter which result in identical observed model outcomes. This phenomenon often occurs because multiple parameters have similar effects on the model outcomes that can be observed. Although multiple values of the nonidentifiable parameters result in the same observed outcomes, they do not necessarily result in the same model behaviour. In fact, the behaviour of unobserved parts of the model may be very different. If the model is to be used to draw conclusions about behaviours not captured in the data, care is required around the question of identifiability.

In practice, a system that is nonidentifiable will usually have an entire surface of values for p which give the same observed behaviour $y(t,p)$. This occurs because two or more parameters have the same effect on the observed output. The parameters involved can be thought of as being linked together, a change in one can be counteracted by an opposing change in the others such that no change is observed in the output.

Other authors differ in the details of their definitions, but all capture the same idea: models fail to be identifiable when there are multiple parameter values giving the same observed output.

This definition of identifiability is quite theoretical. It assumes that observations will not be subject to observational error and that there is no randomness in the effects being modelled or modelling errors in the model itself. Further, it assumes that observations will be available for all times. In practice, these assumptions are inappropriate when modelling real world phenomena. Models and measurements are never perfect and data is certainly not available for all times. As a result structural identifiability, as defined above, is only one of the reasons that a model can fail to be identifiable from observed data. For example, a model may fail to be identifiable because of insufficient observation. If only the initial state of the model is observed, any parameters related to how the state changes after the initial time will be unidentifiable.

On the other hand, a model may be nonidentifiable even when observed quantities are available for all t as in the definition of structural identifiability. In this case, the problem is not an insufficient quantity of data, but rather the type of data available does not provide all the necessary information.

In the case of the HIV transmission model described in Section 2.2, the available data determines the observation function y. In particular, y is vector valued with three

components:

$$\begin{aligned} y^{diag}(t) &= \int_{t-1}^{t} \alpha I(\tau) d\tau \\ y^{death}(t) &= \int_{t-1}^{t} d_D D(\tau) d\tau \\ y^{pop}(t) &= S(t) + I(t) + D(t) \end{aligned} \tag{2.4}$$

represent total new diagnosis in the year ending at time t, HIV deaths in the same time period, and total population at time t respectively. Details of the numerical computation of these integrals is given in the Appendix A.1. Of course, the data is only available once each year, so the function y will, in fact, only be observed at discrete times, $\{t_1, t_2, ..., t_n\}$. The data measured at time t_j will be denoted y_j^{data} while the corresponding observation from the model is given by $y(t_j, p)$.

Given this, we will address the question of identifiability on a practical level by asking whether there is a unique value of the parameters $p = p^*$ that minimizes the sum of squared errors

$$SSE(p) = \sum_{i=1}^{n} \left| y(t_j, p) - y_j^{data} \right|^2 \tag{2.5}$$

This type of identifiability is called *local least squares identifiability* in [53]. The sum of squared errors can also be written in matrix form as

$$SSE(p) = (y(p) - y_{data})^T (y(p) - y_{data}) \tag{2.6}$$

where $y(p)$ is the vector valued observation function constructed by concatenating the vectors $y(t_j, p)$ and y^{data} is the vector of measured data constructed similarly, ie.

$$y(p) = \begin{bmatrix} y(t_1, p) \\ y(t_2, p) \\ \vdots \\ y(t_n, p) \end{bmatrix} \text{ and } y^{data} = \begin{bmatrix} y_1^{data} \\ y_2^{data} \\ \vdots \\ y_n^{data} \end{bmatrix}.$$

Note that in the case of the Alberta HIV data these vectors have length $q := 3n$.

The sum of squared errors is an important quantity for all of the parameter fitting methods we will use in this thesis and it will appear again in Sections 2.3.2 and 2.3.3. As a result, this more practical condition addresses the question that we are most interested in: Is it possible to find a (at least locally) unique best fit value for the parameter p from the data that is available?

There are a number of methods available to asses the SSE for the surfaces of minima that result from nonidentifiable parameters. We will use a simple graphical method along with a method based on the local sensitivity matrix and the singular value decomposition.

Graphical method: For the graphical method we begin by using a numerical tool to find some value $p = p^*$ which minimizes the sum of squared errors. We then vary the components of p in pairs, compute the sum of squared errors at each point, and draw a contour or surface plot of the resulting surface. In order to avoid confusion by considering values of the parameters very far from the original fitted values, only a small domain is used when the parameters are varied. The resulting plots can be examined for long flat valleys which indicate that the two parameters being varied are counteracting each other in their effect on the sum of squared errors.

This graphical method has a number of drawbacks. First, as a graphical method, it is somewhat subjective and to complicate matters the scaling of the axes can greatly change the appearance of the plot. Secondly, it can only easily diagnose identifiability problems involving no more than two parameters. If changes in a parameter can be counteracted only by changing multiple parameters at the same time, this behaviour will be missed. Plots involving more parameters may be possible, but they quickly become computationally intensive and difficult to interpret. Another concern with this method is that the values for parameters not being varied must be fixed in advance. This means that identifiability is considered only near a single point. As we will see when these methods are illustrated using the Alberta dataset, the identifiability structure may be quite different at different points in the parameter space. Nonetheless, the graphical method can give a simple illustration of some of the potential identifiability concerns with a particular system.

Local sensitivity: The remaining methods we will use are based on the Jacobian matrix. To understand these methods, consider that if we attempted to analytically find a minimum for equation (2.5) we would begin by differentiating with respect to each of the parameters and setting these derivatives equal to zero,

$$\frac{\partial SSE(p)}{\partial p_j} = 2\sum_{k=1}^{n} \frac{\partial y_k(p)}{\partial p_i}(y_k(p) - y_k^{data}) = 0 \tag{2.7}$$

This can be written in matrix form as

$$.\frac{\partial SSE(p)}{\partial p} = 2J(p)^T(y(p) - y^{data}) = 0 \tag{2.8}$$

where $J(p)$ is the Jacobian matrix with entries $\frac{\partial y_k(p)}{\partial p_i}$. If the matrix J has full rank then this expression describes m equations for the m parameters p_i and it may be possible to find a unique minimum. On the other hand, if the matrix J has rank $r < m$ then there are only r independent equations for the m parameters and the minimum is not, in general, unique. Unfortunately it is usually unwieldy to analytically compute J for the problems we are interested in so we instead rely on numerical tools.

The matrix J is already computed as a side effect of the least squares minimization

routine. The quantities represented in y and p may have very different magnitudes, therefore we continue by defining

$$\begin{aligned} J^{Rel} &= [J_{ki}] \\ J^{Rel}_{ki} &= \left.\frac{\partial y_k(p)}{\partial p_i}\right|_{p^*} \frac{p^*_i}{y_k(p^*)}. \end{aligned} \tag{2.9}$$

This is the relative local sensitivity matrix around the point p^* including the standard scaling that is used to allow comparisons between elements of different sizes. It will appear again in Section 4.3 when we discuss the sensitivity of model outcomes to changes in the parameters.

A special case of the identifiability problem occurs when the observed values $y(t, p)$ are not sensitive to one or more parameters. In other words, changes in a single parameter cause only minor changes in the observed values. In this case, it is a single parameter, rather than a combination of parameters that is non-identifiable. Identifiability problems arising due to a lack of sensitivity can be diagnosed by examining the local sensitivity matrix J^{Rel} as defined in (2.9). If all the elements of a column of J^{Rel} are small, this indicates that none of the observed values are sensitive to the parameter associated with that column. To simplify the process of examining J^{Rel} we will consider only the maximum magnitude element in each column.

Singular value decomposition: For identifiability problems involving more than one parameter the singular value decomposition is useful The singular value decomposition is commonly used to numerically estimate the rank and null space of a matrix and so it is well suited to identifying situations where J has deficient rank.

Because we already have a technique for diagnosing the case when a single parameter has very little effect on the observed values, we remove the impact of this type of identifiability problem by first normalizing the columns of J. This normalization is standard for variance decomposition. Define,

$$\begin{aligned} J^{Norm} &= \left[J^{Norm}_{ki}\right] \\ J^{Norm}_{ki} &= J^{Rel}_{ki}(p^*) \left(\left(\sum_{j=1}^{q} J^{Rel}_{ji}(p^*)^2 \right)^{-1/2} \right). \end{aligned} \tag{2.10}$$

The singular value decomposition is used to write the matrix J^{Norm} as

$$J^{Norm} = USV^T, \tag{2.11}$$

where U and V are orthogonal square matrices, and S is an n by m diagonal matrix with non-increasing elements $\{\lambda_1 \ldots \lambda_m\}$. If these computations were done analytically, any zero elements on the diagonal of S would indicate a deficiency in the rank of J^{Norm}.

However, when the computations are done numerically, we instead look for diagonal elements of S that are small compared to the maximum element. In particular we compute the condition indices

$$\nu_i = \frac{\lambda_1}{\lambda_i}. \tag{2.12}$$

Large values for ν_i indicate deficiencies in the rank of J.

Variance decomposition: In order to determine which parameters are involved in the nonidentifiable surface, a variance decomposition method is used. This method is commonly used to diagnose colinearities in data to be used in linear least squares problems. A discussion of variance decomposition in the linear case can be found in [12]. In order to apply it to the nonlinear problem we begin by linearizing the function $y(p)$ about a predefined point p^*

$$y(p) \approx y(p^*) + J(p^*)(p - p^*). \tag{2.13}$$

Incorporating the scaling (2.10) into the linear approximation for $y(p)$ in (2.13) can be interpreted as a rescaling of the quantities $p - p^*$ and $y(p) - y(p^*)$. The result of this rescaling is

$$\Delta y \approx +J^{Norm}\Delta p \tag{2.14}$$

where $\Delta p_i = \left(\sum_{i=1}^{n} J_{ij}(p^*)^2)\right)^{1/2} p_i^*(p - p^*)$ and $\Delta y_k = y_k(p*)^{-1}\left(y_k(p) - y_k(p^*)\right)$. Now the same techniques usually used for linear least squares allow the calculation of an estimate for Δp

$$\widehat{\Delta p} = ({J^{Norm}}^T J^{Norm})^{-1} {J^{Norm}}^T \Delta y^{data}, \tag{2.15}$$

where $\Delta y^{data} = y_k(p*)^{-1}\left(y_k^{data} - y_k(p^*)\right)$. An estimate for the variance-covariance matrix $\widehat{\Delta p}$ is given by

$$cov(\widehat{\Delta p}) = \sigma^2({J^{Norm}}^T J^{Norm})^{-1}, \tag{2.16}$$

where σ^2 is the variance of the scaled data $y_k(p*)^{-1}y_k^{data}$. This variance-covariance estimate is based on the standard assumptions for least squares problems which will be described in Section 2.3.2. Once again using the singular value decomposition this matrix can be rewritten as

$$\begin{aligned} cov(\widehat{\Delta p}) &= \sigma^2({J^{Norm}}^T J^{Norm})^{-1} \\ &= \sigma^2\left((VS^TU^T)(USV^T)\right)^{-1} \\ &= \sigma^2\left(VS^TSV^T\right)^{-1} \\ &= \sigma^2 VS^{-2}V^T, \end{aligned} \tag{2.17}$$

where S^{-2} is the square $m \times m$ matrix with entries λ_j^{-2} The variances of the elements

of $\widehat{\Delta p}$ are found on the diagonal of the matrix $cov(\widehat{\Delta p})$ and are therefore given by

$$var(\widehat{\Delta p_j}) = \sigma^2 \sum_{i=1}^{n} \frac{{v_{ji}}^2}{{\lambda_i}^2}. \tag{2.18}$$

From this expression we see that $var(\widehat{\Delta p_j})$ consists of a sum of elements corresponding to each of the singular values λ_i. As the singular values appear in the denominator, small values of λ_i will in general correspond to large variances for $\widehat{\Delta p_j}$. Parameters involved in nonidentifiabilities will have most of their variance coming from terms associated with small values of λ_i. In order to identify these parameters we examine the fraction of $var(\widehat{\Delta p_j})$ found in each element of the sum found in (2.18) using the matrix with elements π_{ij}

$$\pi_{ij} = \frac{{v_{ji}}^2}{{\lambda_i}^2} \frac{1}{\sum_{k=1}^{n} \frac{{v_{jk}}^2}{{\lambda_k}^2}}. \tag{2.19}$$

The jth column of this matrix represents the fractions of the variance $var(\widehat{\Delta p_j})$ which are associated with each of the singular values. Note that because this matrix contains fractions of the variances, it is not necessary to maintain the scaling of $\widehat{\Delta p_j}$. The variances of $\widehat{p}$ will have the same fraction associated with each singular value as the variances of $\widehat{\Delta p}$.

As already mentioned, a large fraction of the variance appearing in terms associated with small singular values indicates that the parameter p_j is likely to be involved in the nonidentifiability surface. The ith row of the matrix π represents the contributions to the various variances by terms influenced by λ_i. If λ_i is very small, any of the variances it impacts will be increased.

Since this method uses the numerically computed matrix J evaluated at a single parameter value $p = p^*$, the results will be better the closer the fixed parameter value p^* is to the true value of the parameters. Additionally, this method gives only local information about the nonidentifiability surface. However it is able to diagnose identifiability problems involving multiple parameters.

An analysis of the identifiability structure is performed for the Alberta data and illustrates these methods.

2.3.1.1 Identifiability for Alberta Data

Using a preliminary model fit for the value of p^*, the condition indices from the normalized Jacobian J^{Norm} are given in Table 2.3. The values of the first four condition indices are acceptable. The largest two condition indices are quite large indicating a nearly singular Jacobian matrix and a linkage between some of the parameters. These condition indices lead us to expect that it will be necessary to fix two of the model parameters in order to achieve a fully identifiable model.

The proportions of the variances associated with each of the singular values is used

ν_j	Proportion of Variance					
	$var(\alpha)$	$var(d_I)$	$var(d_D)$	$var(\beta_I)$	$var(a)$	$var(r)$
1.0	5.389e-14	2.560e-10	0.005227	3.429e-14	9.435e-08	3.749e-13
2.3	6.098e-14	7.689e-11	0.6974	1.030e-14	5.002e-08	3.891e-13
4.7	9.330e-13	1.246e-10	0.2667	1.667e-14	2.128e-07	3.619e-11
45.9	6.166e-11	1.799e-07	0.02927	2.423e-11	0.0007494	6.859e-12
3.057e4	4.066e-05	0.8373	1.233e-06	1.796e-05	2.013e-05	4.066e-05
3.150e6	0.9999	0.1627	0.001447	0.9999	0.9992	0.9999

Table 2.3: Condition indices and the proportion of variance in the parameter estimates associated with each. The largest two condition indices account for the majority of the variance of all parameters except d_D.

Parameter	Largest Sensitivity
α	1.10453
d_I	-0.00075
d_D	1.00397
β_I	2.07566
a	0.00001
r	1.00004

Table 2.4: Maximum local sensitivity of the observed quantities with respect to each of the parameters.

to determine which parameters are involved in the nonidentifiability. This information is also found in Table 2.3. The last row of the table indicates that the majority of the variance for parameters α, β_I, a, and r is associated with the smallest singular value while the majority of the variance for the parameter d_I is associated with the smallest two singular values. Taken together, this suggests that all parameters except d_D are involved in the nonidentifiability, resulting in a two dimensional surface of minima in the five dimensional space of all parameters except d_D.

The graphical method confirms that parameters d_I, β_I, and a are involved in the nonidentifiability. The contour plots for these parameters are illustrated in Figure 2.5 and contour plots for some other pairs of parameters are included in Figure 2.6 for comparison.

Examination of the sensitivities of the parameters found in Table 2.4 indicates that the observations have low relative sensitivity to the parameters a and d_I. This is in addition to involvement of these parameters in the nonidentifiability revealed either through the variance decomposition in Table 2.3 or as seen through the graphical method in Figure 2.5. This suggests that the data does not give much information about parameters a and d_I and so these parameters will be fixed to resolve the nonidentifiability problem.

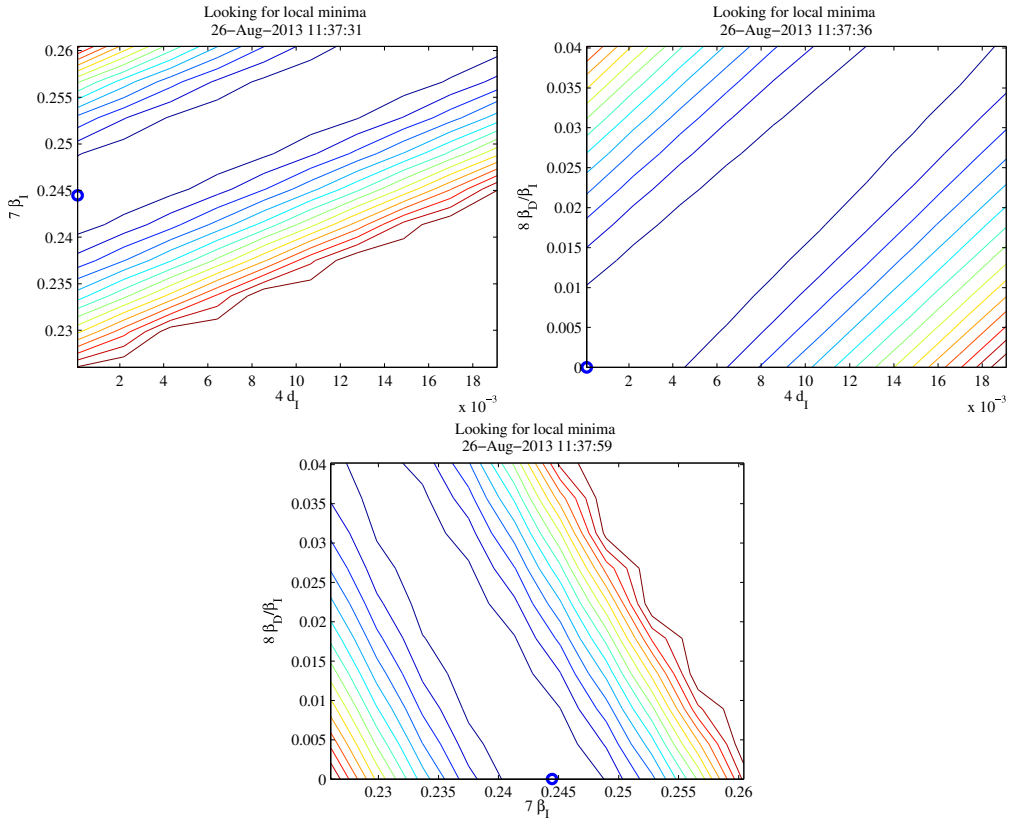


Figure 2.5: Contour plots of the sum of squared errors for pairs of parameters that are not simultaneously identifiable. A whole line of parameter values attain the same minimum.

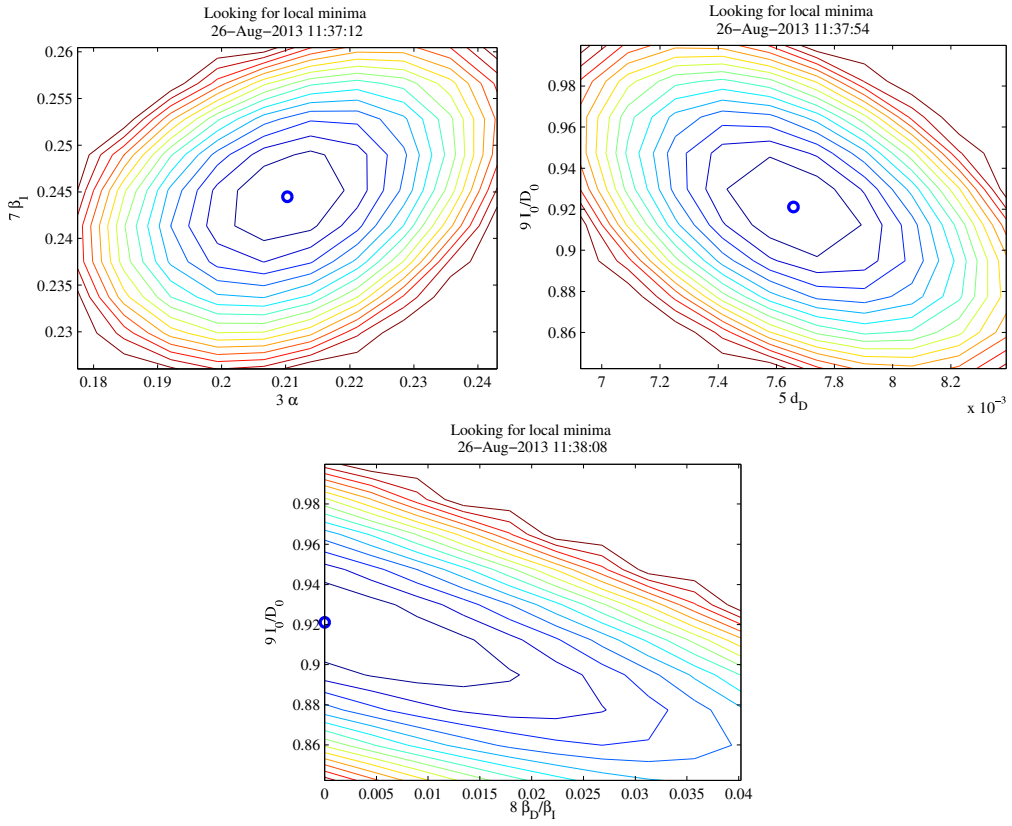


Figure 2.6: Contour plots for pairs of parameters that are simultaneously identifiable. The minimum occurs at a unique point in the parameter space.

Condition Indices	Proportion of Variance			
	$var(\alpha)$	$var(d_D)$	$var(\beta_I)$	$var(r)$
1.0	2.5733e-08	0.024853	1.3062e-08	3.154e-08
1.6	6.9905e-08	0.50241	2.9938e-09	1.0076e-08
2.6	2.3379e-07	0.44261	1.3605e-09	3.4356e-07
6857.1	1	0.030126	1	1

Table 2.5: Condition indices and variance decomposition for the remaining parameters.

A value for the parameter d_I can be fixed using the death rate from untreated HIV as is observed in other contexts. We choose $d_I = \frac{1}{12}$.

Choosing a value for a is more complicated. This parameter represents the fraction of diagnosed HIV patients who are still transmitting the virus, either because they have only recently been diagnosed and are not yet effectively treated, or because they have stopped or failed treatment. The identifiability structure at different values for a appears to be quite different. We consider two cases, $a = 0$, and $a = 0.001$.

Case 1: $a = 0$: An optimistic value would be to choose a to be equal to zero. The preliminary least squares fit that we have been using to select the parameter values chooses the value of a to be extremely small, but not zero. Since a is not identifiable, this value is meaningless. Instead we fix $a = 0$. Now the graphical method suggests that there is a link between the parameters α and β_I. This is problematic as α has a direct impact on the outcomes we are interested in – a change in the rate at which those who are HIV-positive are diagnosed will definitely also change the size of the undiagnosed population. At the same time β_I is extremely difficult to measure experimentally. The fact that we cannot take either of these parameters from some other source means that we will not be able to resolve this identifiability problem, however such a small value of a is almost certainly overly optimistic. The D category does not include only those who are successfully treated, but also those who have not yet started treatment and those who have stopped treatment for any reason. It is unlikely that no one who is diagnosed as HIV positive ever transmits the virus.

Case 2: $a = 0.001$: If a is instead chosen to be somewhat larger at $a = 0.001$, the graphical method suggests that the remaining parameters are all uniquely identifiable. These results are illustrated in Figure 2.7. Examination of the variance decomposition, found in Table 2.5, indicates that the largest condition index is still somewhat large, but greatly improved from those computed before fixing a and d_I. We conclude that the four remaining parameters are sufficiently identifiable. The last row of Table 2.5 indicates the parameters associated with the smallest singular value. It can be seen that the parameters α, β_I, and r are affected. These are all parameters which we are particularly interested in so it is not desirable to fix any of them in the hopes of further improving the identifiability properties of the model.

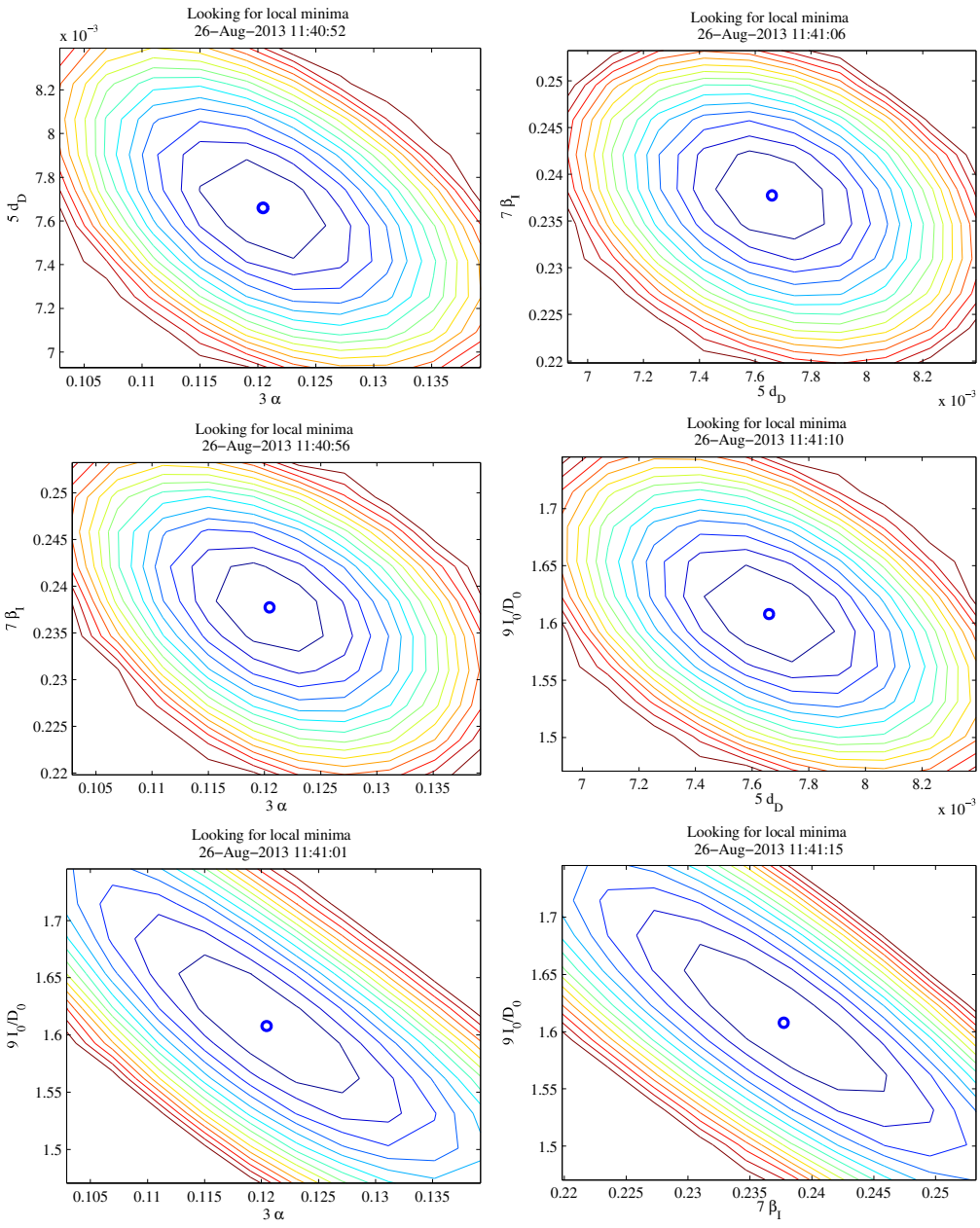


Figure 2.7: Contour plots of the sum of squared errors for all remaining pairs of parameters illustrate that each pair can be simultaneously identified. A unique minimum is seen in each plot.

2.3.2 Nonlinear Least Squares

Nonlinear least squares is a standard method for curve fitting. The basic idea is to find parameter values $p = \widehat{p}$ which minimize the sum of squared errors as defined in 2.5. This corresponds to minimizing the vertical distance between the fitted curve $y(t, \widehat{p})$ and the data y^{data} in the l_2 norm. This is a simple and intuitive concept of what is required for a "good fit" – the fitted curve should be as close as possible to the available data.

In our problem, the sizes of the individual parts of the data are very different. In particular, the data includes the overall population a quantity which is much larger than the number of HIV deaths. This has the result that the least squares fitting routine tends to treat the overall population as more important and a better fit will be found for this component at the expense of a good fit for the disease components that we are most interested in. For this reason, we use a weighted version of the sum of squared errors

$$SSE(p) = \sum_{j=1}^{n} \left| W(y(t_j, p) - y_j^{data}) \right|^2. \tag{2.20}$$

In the case of the HIV data, we choose W to be a 3×3 diagonal matrix where the diagonal elements are given by

$$\begin{aligned} W^{diag} &= \left(\frac{1}{n} \sum_{j=1}^{n} y_j^{diag} \right)^{-1} \\ W^{death} &= \left(\frac{1}{n} \sum_{j=1}^{n} y_j^{death} \right)^{-1} \\ W^{pop} &= \left(\frac{1}{n} \sum_{j=1}^{n} y_j^{pop} \right)^{-1}. \end{aligned} \tag{2.21}$$

That is, each component of the data is weighted by the average value of that data component. The inclusion of weighting improves the results of the fitting procedure by emphasizing all types of data similarly regardless of the size of the quantities involved. The weighted version of the sum of squared errors 2.20 can be reduced to the original unweighted version 2.5 by including the weights in the functions $y_j(p)$ and the data. Therefore the weighting will be omitted in the remainder of the discussion of nonlinear least squares.

If the function $y(t, p)$ is linear in the parameters p, the least squares method reduces to linear least squares and the optimal value for $\widehat{p}$ can be calculated analytically. For nonlinear least squares, the optimal value of $\widehat{p}$ must be estimated numerically. Methods for numerically solving nonlinear least squares problems are included in the standard MATLAB libraries [95]. We use the function `lsqnonlin` extensively.

With the addition of a few more assumptions, it can be seen that the nonlinear

least squares formulation is equivalent to maximum likelihood methods of parameter estimation. In particular, assume that

$$y_{ij}^{data} = y_j(t_i, p) + \epsilon_{ij} \tag{2.22}$$

where $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$ is a random variable describing the error produced when y_{ij} is estimated by $y(t_i, p)$. If the value of $\sigma_{ij} = \sigma$ is the same for all the observations, then the likelihood of observing the data y_{ij} given the parameters p is

$$L(y, p) = P(y|p) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{SSE}{2\sigma^2}} \tag{2.23}$$

Then choosing $p = \widehat{p}$ to maximize the likelihood is the same as choosing it to minimize the sum of squared errors which appears in the exponent.

The likelihood formulation also allows the use of statistical methods for estimating confidence regions for the computed parameters p. One simple method is based on the fact that if X_j is a random variable with the standard normal distribution, $X_j \sim N(0, 1)$ then

$$\sum_{j=1}^{n} X_j^2 \sim \chi_n^2. \tag{2.24}$$

Then if the value of σ is known, a confidence region for p can be determined by finding all the points where the sum of squared errors is less than some threshold value

$$\sum_{ij} (y_{ij} - y_j(t_i, p))^2 < c. \tag{2.25}$$

A threshold value of $c = \sigma\chi_{n,0.95}^2$ will give a confidence region for the 95% confidence level. These confidence sets are related to the plots that we have created using the graphical method for diagnosing identifiability problems. In two dimensions the boundary of the confidence set will be a contour line as seen with the graphical method.

This method of computing confidence regions is simple in theory, but there are a number of drawbacks. First, the value of σ must be constant for all the data. This is not particularly problematic as varying values of σ_{ij} could be removed by appropriate weighting of the least squares problem. However, to implement this, the values of σ_{ij} must be known in advance or estimated from the data. In general these values are unknown, and since we have only one data point for each quantity at each time, it is not possible to estimate the values of σ_{ij} from the data available. As a result, another method must be used to compute interval estimates for the fitted parameters. A Bayesian method is used to compute parameter distributions, and thus interval estimates. This is discussed in Section 2.3.3.

Parameter	Fitted Value
Λ	96345
d_S	0.00579

Table 2.6: Fitted values for population parameters.

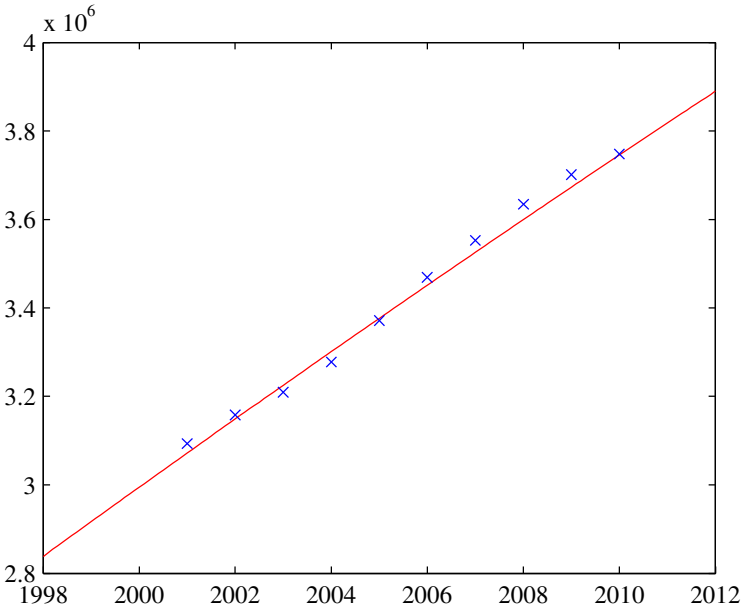


Figure 2.8: Fitted curve for the population model.

2.3.2.1 Fitted Alberta Model

The first step to fitting the Alberta model is to fit the population parameters Λ and d_S. For this purpose we use Statistics Canada population estimates and yearly deaths estimates for the Province of Alberta and the model without any disease. Note that it is necessary to use the yearly deaths estimates for this parameter fitting as the linear population growth model is nonidentifiable if only population data is available.

The results of fitting the population parameters are given in Table 2.6. The resulting population curve is visually a very good fit for the data as shown in Figure 2.8.

In Section 2.3.1.1, we determined that it will only be possible to fit the parameters α, d_D, β_I and $r = \frac{I_0}{D_0}$. The unconstrained nonlinear least squares fitting routine results in the parameter estimates found in Table 2.7. The fitted curves and the data are plotted in Figure 2.9.

While the fitted curves from this initial fitting approximate the data as well as can be expected, there is concern over the appropriateness of the parameter values chosen.

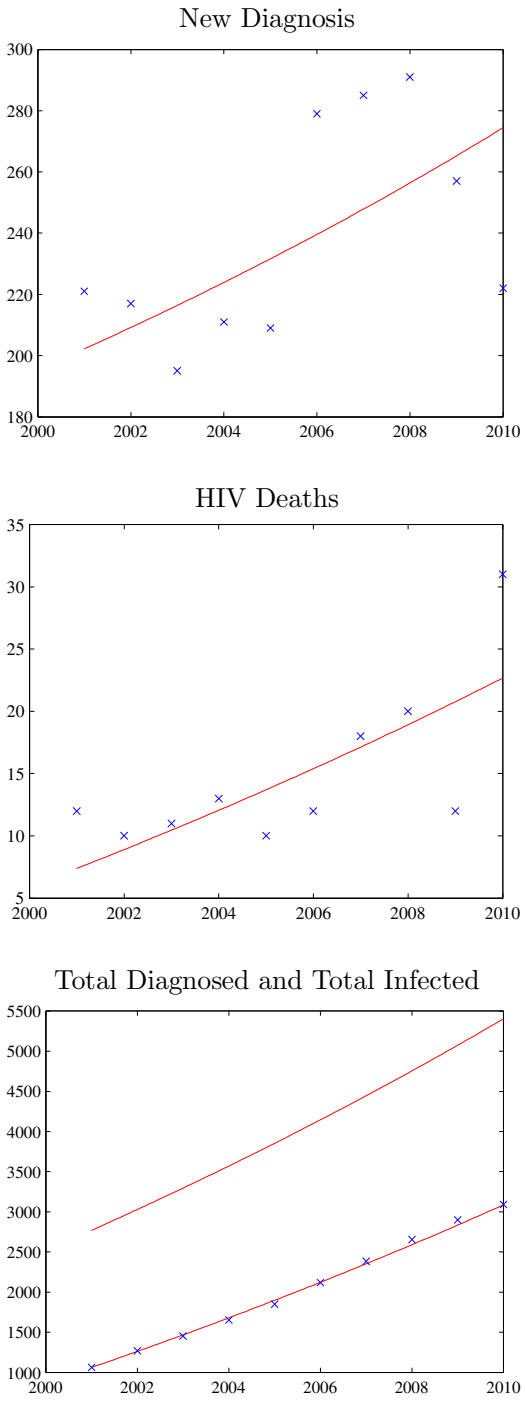


Figure 2.9: Fitted curves for the HIV transmission model.

Fitted Parameters	
Parameter	Value
α	0.12045
d_D	0.00766
β_I	0.23774
$r = \frac{I_0}{D_0}$	1.60778

All Parameters	
Parameter	Value
Λ	96345
d_S	0.00579
d_I	0.08333
d_D	0.00766
α	0.12045
β_I	0.23774
β_D	0.00024
S_0	3071704.74193
I_0	1707.45882
D_0	1062.00000

Table 2.7: Fitted parameter values for the HIV transmission model.

In particular, the fitted value of $\alpha = 0.12045$ seems much to small for the Province of Alberta context resulting in an excessively long average time to diagnosis. This value indicates that the average HIV-positive person spends $1/\alpha = 8.3$ years infected with HIV but undiagnosed. This same concern is also reflected in the fitted value of the initial condition $I_0 = 1707.46$. This is about 61.6% of the total HIV-positive population and, while the exact value of I_0 is unknown, national estimates for the percentage of the HIV-positive population that is undiagnosed are available for the year 2002 [19]. These national estimates indicate 29% (21.5% to 36.1%) of the HIV-positive population was undiagnosed in 2002. While the Province of Alberta may differ somewhat from the national population, this estimate does suggest that the initial fitted parameters are inappropriate.

In order to address the concerns raised by these fitted parameter values, the impact of the initial condition parameter $r = \frac{I_0}{D_0}$ on the fitting results will be considered in more detail. The national estimates of the size of the HIV-positive population which was undiagnosed in 2002 correspond to a range for r of $(0.266, 0.564)$ [19] while the preliminary least squares indicated a value of $r = 1.608$. To investigate these results further, fifty different values of r are considered from the range $(0.1, 2)$. For each of these values of r, the least squares estimate of the other parameters is calculated, holding r fixed.

The changes in the fitted parameter values and SSE resulting from changes in the fixed value of r are displayed in Figure 2.10. This figure illustrates that, decreasing the value of r from its least squares fit increases the SSE. This increase is small but begins increasing more quickly below $r = 0.5$. At the same time, the best fit values of α and β_I also increase. Figure 2.10 also illustrates the range estimated for r by [19] and the best fit value for the parameters when r is constrained to remain in this range. We conclude that while such a constraint does not allow the model to attain the best possible

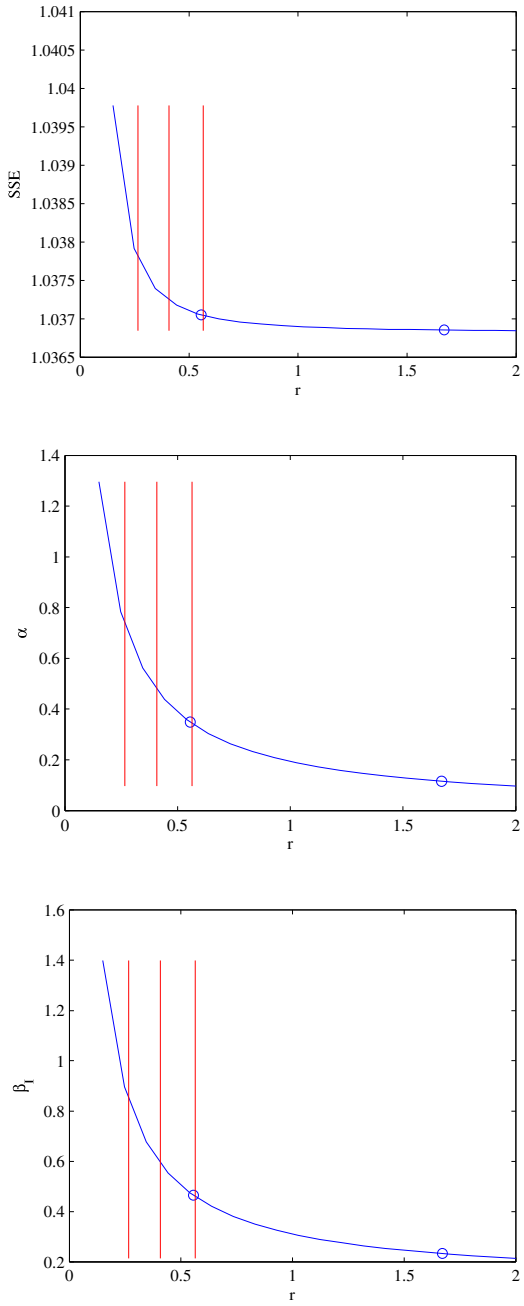


Figure 2.10: Impact of varying the parameter r on the sum of squared errors and the fitted values for β_I and d_I. The dashed lines indicate the estimates for r found in [19] while the solid line indicates the best fit values for the sum of squared errors, β_I, or d_I. The circles indicate the best constrained and unconstrained fits.

Fitted Parameters

Parameter	Value
α	0.34873
d_D	0.00766
β_I	0.46504
$r = \frac{I_0}{D_0}$	0.55554

All Parameters

Parameter	Value
Λ	95632.09068
d_S	0.00581
d_I	0.08333
d_D	0.00766
α	0.34873
β_I	0.46504
β_D	0.00047
S_0	3064748.53601
I_0	589.98075
D_0	1062.00000

Table 2.8: Fitted parameter values for the HIV model using constraints on the value of r

fit, the increase in the SSE from the unconstrained model is small. Additionally, the constrained model results in a more reasonable fitted value for α. This will be taken as the final model fitted by the least squares method. The parameter values determined by this model fit are an estimate of the appropriate parameters for the model in the context of the Province of Alberta. However, such a simple least squares estimate does not provide all the desired information about these quantities. In particular, computing a confidence region for the parameter values requires a number of further assumptions which cannot be easily justified. Therefore model fitting will continue using Bayesian techniques.

2.3.3 Bayesian Parameter Estimation

The Bayesian method assumes that the parameter values will be random variables from some probability distribution. Initially, these distributions are set using whatever outside knowledge is available about the possible parameter values. These assumed distributions are known as the prior distributions or simply "the priors". The data is then used to update the priors to get the posterior distributions.

Bayesian techniques are discussed at length by others [29, 50] but briefly, Bayesian methods rely on computation of the posterior distribution:

$$P(p|y_j) = \frac{P(y_j|p)P(p)}{\int P(y_j|p)P(p)dp} \tag{2.26}$$

where $P(y_j|p)$ is the likelihood of observing the data given a particular value of p and $P(p)$ is the prior distribution of the parameter p. Once the posterior distribution 2.26 is computed, point estimates for the parameters p can be obtained as the mode of $P(p|y_j)$ and interval estimates can be obtained as percentile ranges.

Computing the likelihood portion $P(y_j|p)$ of the Bayesian posterior requires some

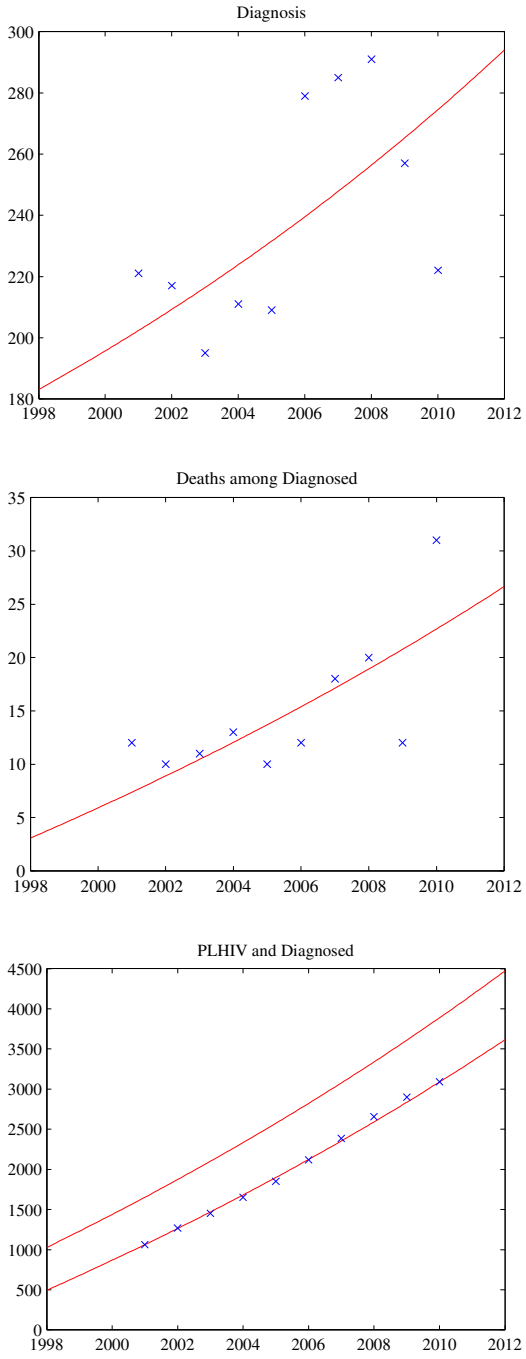


Figure 2.11: Updated fitting results for the HIV model using constraints on the value of r.

additional assumptions. Commonly it is assumed that the errors introduced in collecting data are normally distributed. That is,

$$y_j = y(t_j, p) + \varepsilon_j.$$

Where ε_j is a random variable with $\varepsilon_j \sim N(0, \sigma^2)$. It is also usual to assume that σ^2 is a random variable itself with σ following an inverse gamma distribution, $\sigma^2 \sim \text{Inv}\,\Gamma(\eta, \beta)$. Using these assumptions, the likelihood portion of the posterior can be written as

$$\begin{aligned} P(y_j|p) &= \int P(y_j|p, \sigma^2) P(\sigma^2) d\sigma^2 \\ &= \int \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{SSE(p)}{2\sigma^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{-2(\alpha-1)} e^{-\frac{\beta}{\sigma^2}} d\sigma^2 \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \frac{\Gamma(\frac{N}{2} + \alpha)}{\left(\frac{SSE(p)+2\beta}{2}\right)^{(\frac{N}{2}+\alpha)}}, \end{aligned}$$

where

$$SSE(p) = \sum_j |y(t_j, p) - y_j^{data}|^2. \tag{2.27}$$

When the prior, $P(p)$, and the scaling factor, $\int P(y_j|p)P(p)dp$, are included, the expression for the posterior, $P(p|y_j)$, can become quite complicated. For this reason we do not usually hope to compute it analytically. Instead, we use Markov chain Monte Carlo (MCMC) methods to approximate the distribution by drawing a sample from it. In particular, we use the Metropolis-Hastings algorithm [32] as implemented in MATLAB. A brief introduction to the algorithm is given in Appendix A.2.1.

Once a sample from the posterior distribution has been computed, it is necessary to analyze further in order to find point and interval estimates for the parameters. For the point estimate, we wish to use the maximum posterior density point (the mode) of the distribution. For interval estimates, we will project the highest posterior density (HPD) region onto the various axes. In order to find the maximum posterior density point and the highest posterior density region, we first use a kernel method to estimate the posterior density function from the sample. An introduction to kernel density estimation is found in [117].

2.3.3.1 Nonidentifiable Parameters

In theory, the Bayesian method is applicable even in the presence of nonidentifiable parameters. However, when some of the parameters are nonidentifiable, the choice of prior distribution becomes extremely important [48] and convergence for the numerical sampling methods may be extremely slow [41]. As we do not have good information with which to choose a prior distribution, it will be important to fix some of the parameters

in order to use an identifiable model as we did for the least squares fitting.

However, this means that the Bayesian method will not provide interval estimates for the parameters which are fixed in advance. In order to produce interval estimates for all parameters, we use an iterated method.

Iterated MCMC Method In order to compute interval estimates for all parameters, including those which are nonidentifiable, we propose the following iterated Markov chain Monte Carlo (iMCMC) method for a model with m parameters, $p_1, p_2, \ldots, p_m$.

1. Determine the identifiability structure of the model and use this information to fix some of the model parameters so that the remaining parameters are all identifiable. We suppose that there are r identifiable parameters, $p_1, p_2, \ldots p_r$, and $m - r$ fixed parameters $p_{r+1}, p_{r+2}, \ldots, p_m$.

2. Use the method to find point and interval estimates for the r identifiable parameters, $p_1, p_2, \ldots p_r$. Use the fitted point estimates to fix the values of all parameters.

3. Allow p_{r+1}, which was fixed in Step 1, to vary and repeat the data fitting step to find an interval estimate for this parameter.

4. Repeat Step 3 as necessary to find interval estimates for $p_{r+2}, \ldots, p_m$.

The interval estimates obtained by this method are credible intervals from the joint distributions of the parameters being fit given that the parameters not being fit are fixed at predefined values. This is true regardless of what step of the method is creating the interval, however intervals created at different steps will have different combinations of fitted and fixed values. As a result, the order in which non-identifiable groups of parameters are fitted will make a difference in the reported interval estimates. Interval estimates computed when many parameters are being fit to data and only a few are fixed will tend to be wider than those computed with only a few other parameters being fit.

2.3.3.2 A Toy Problem

In order to illustrate the iterated method on as simple a model as possible, we will consider the linear growth model where the compartment population is observed.

$$\dot{x} = r + dx$$
$$y = x$$

In the absence of disease, the disease transmission models that we have been considering reduce to this model. In our disease models, we have relabeled $r = \Lambda$ and $d = d_S$. We have already mentioned that this model is nonidentifiable if the data is the population

Parameter	Fixed Quantities	Point Estimate	Interval Estimate
d	$r = 2.2$	3.2526	$(1.2491, 7.7719)$
r	$d = 3.2526$	2.3145	$(0.62387, 4.4544)$

Table 2.9: Interval estimates from iMCMC method for a simple population growth model.

of the compartment. The two parameters d and r are linked. For the Alberta case, we avoided this problem by providing deaths data in addition to the population data.

When fitting the disease data, we do not wish to add additional complication by finding interval estimates for the population parameters. The point estimates computed in section 2.3.2 are sufficient for our purposes. However this simple model makes a good illustration of the iterated MCMC method. Here, we use this simple model to create some data and then use the model and the created data to compute interval estimates using the MCMC method.

The data for this illustration is created by setting $r = 2$, $d = 3$, and computing $y_i = dx(t_i)$ for $t_i = 29.0, 29.1, 29.2, \ldots, 29.9, 30.0$ with $x(29.0) = 0.5$. To make the problem a little more realistic, some noise is added with mean 0 and standard deviation 0.01.

Since the parameters are nonidentifiable, this data is used to illustrate the iMCMC method. First, the value of r is fixed allowing the parameter d to be identified. Although we could chose the correct value for r which was used to create the data, this is not usually possible in practice so we instead choose $r = 2.2$. That is, we perturb the true value of r by 10%. A Metropolis-Hastings sample for d is used to estimate the Bayesian posterior. We calculate a point estimate for d by selecting the mode of the estimated posterior distribution. A 95% credible interval is selected using the 2.5% and 97.5% percentiles of the sample. The posterior distribution for d is illustrated in Figure 2.12.

In order to compute an interval estimate for r, the procedure is repeated. This time, d is fixed at the previously computed point estimate and a sample is selected from the Bayesian posterior for r. As before, a 95% credible interval is selected using the 2.5% and 97.5% percentiles of the sample. The posterior distribution for r is illustrated in Figure 2.12, and the results of the iterated procedure are summarized in Table 2.9.

Notice that the intervals listed for d are conditional on r taking the fixed value $r = 2.2$ as defined before the first MCMC run. The interval for r, on the other hand, is conditional on d taking the value $d = 3.2526$ as estimated in the first MCMC run and fixed for the second MCMC run.

In order to confirm these results, we perform both steps of the procedure a second time fixing the parameters in the opposite order. That is, we begin by fixing $d = 3.3$ and compute the distribution for r. Then we fix r at the point estimate and compute the distribution for d. The results are found in Figure 2.13 and Table 2.10

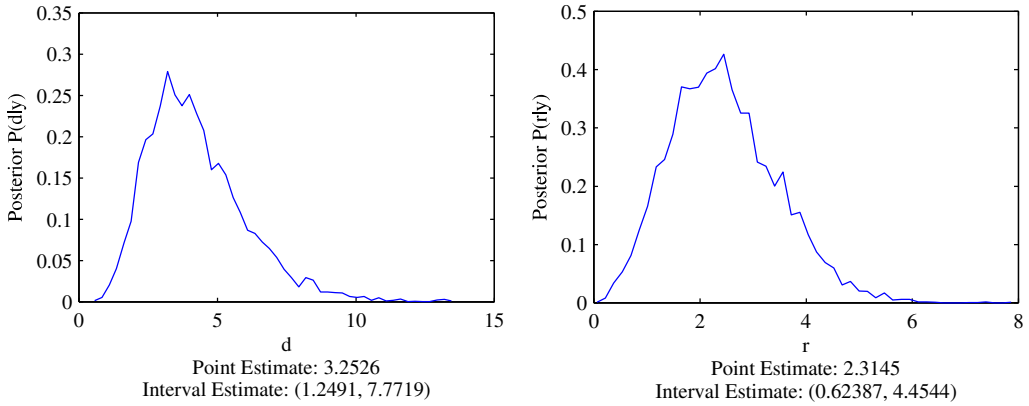


Figure 2.12: Results of iMCMC method for a simple population growth model.

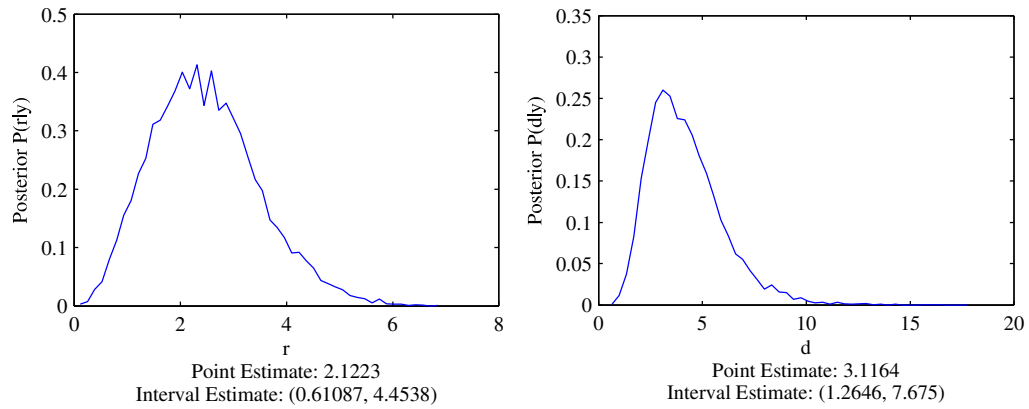


Figure 2.13: Results of iMCMC method for a simple population growth model.

Parameter	Fixed Quantities	Point Estimate	Interval Estimate
r	$d = 3.3$	2.1223	$(0.61087, 4.4538)$
d	$r = 2.1223$	3.1164	$(1.2646, 7.675)$

Table 2.10: Parameter values resulting from iMCMC method for a simple population growth model.

The results for this simple model show a good correspondence between the fitting results and the true parameter values. The difference between the point estimate of the parameter d and the true value is the result of the influence of the prior distribution along with the choice of fixed value for r. The prior distribution and the fixed value for r were chosen as if the true values of the parameters were unknown as this information would realistically not be available. A more careful choice of prior distribution for d and fixed value for r would give better results. The same is true for r, the difference between the fitted estimates and the true value can be attributed to the prior distribution and the difference in the fitted value of d. The results from the second fitting procedure, when r was fit first followed by d are similar indicating that the order in which these parameters are fitted has only a minor impact on the fitted point and interval estimates.

2.3.3.3 Alberta Bayesian Results

The priors used for the parameter values in the Alberta model are given in Table 2.11. Using uniform prior distributions for most of the priors requires us to specify only a plausible range for the parameter values. These priors provide no additional information about the parameter value and avoid influencing the posterior distribution with unjustified assumptions. Most of the parameter ranges allowed under the assumed uniform priors are generous – they do not impose serious restrictions on the parameter values. In contrast, the prior for r is chosen using estimates of the undiagnosed population in 2002 the by the Public Health Agency of Canada (PHAC) [19]. A triangular distribution with mode at the estimated value and maximum and minimum covering the estimated range is used. This distribution influences the resulting posterior by adding additional weight to PHAC's estimated value as well as constraining the support of the resulting posterior. These assumptions are similar to the constraints placed on the parameter values in the final least squares fitting using the PHAC estimates.

The results of the fitting procedure are given in Table 2.12. While the point estimates are similar to the estimates calculated by the least squares method, the fitted value of r has been reduced somewhat by the triangular prior adding additional weight to the PHAC estimate. This in turn, increases the value of α and β as expected by our previous examination of the impact of changing the value of r.

Parameter	Description	Prior Distribution
Λ	Population growth	Fixed $\Lambda = 100570$
d_S	Removal rate for S	Fixed $d_S = 0.005$
d_I	Removal rate for I	Uni$(0, 0.2)$
d_D	Removal rate for D	Uni$(0.005, 0.0833)$
α	Diagnosis rate	$Uni(0, 2)$
β_I	Transmission coefficient for I	Uni$(0, 2)$
β_D	Transmission coefficient for D	$\beta_D = a\beta_I$
a	Transmission remaining for D	Uni$(0, 0.1)$
S_0	Number in S in 2001	Fixed $S_0 = 2932300$
I_0	Number in I population in 2001	$I_0 = rD_0$
r	I_0 as a fraction of D_0	Tri$(0.266, 0.408, 0.564)$[19]
D_0	Number in D in 2001	Fixed $D_0 = 1062$

Table 2.11: Description and prior distributions for parameter values in the HIV model.

Parameter	Point Estimate	Interval Estimate
β_I	0.56548	$(0.2637, 1.019)$
α	0.42813	$(0.024503, 1.0984)$
d_D	0.0080529	$(0.0057914, 0.017563)$
r	0.41542	$(0.29525, 0.5632)$
a	0.00071391	$(0.000030036, 0.064314)$
d_I	0.090582	$(0.014496, 0.19983)$

Table 2.12: Results of iMCMC method for the HIV model parameters.

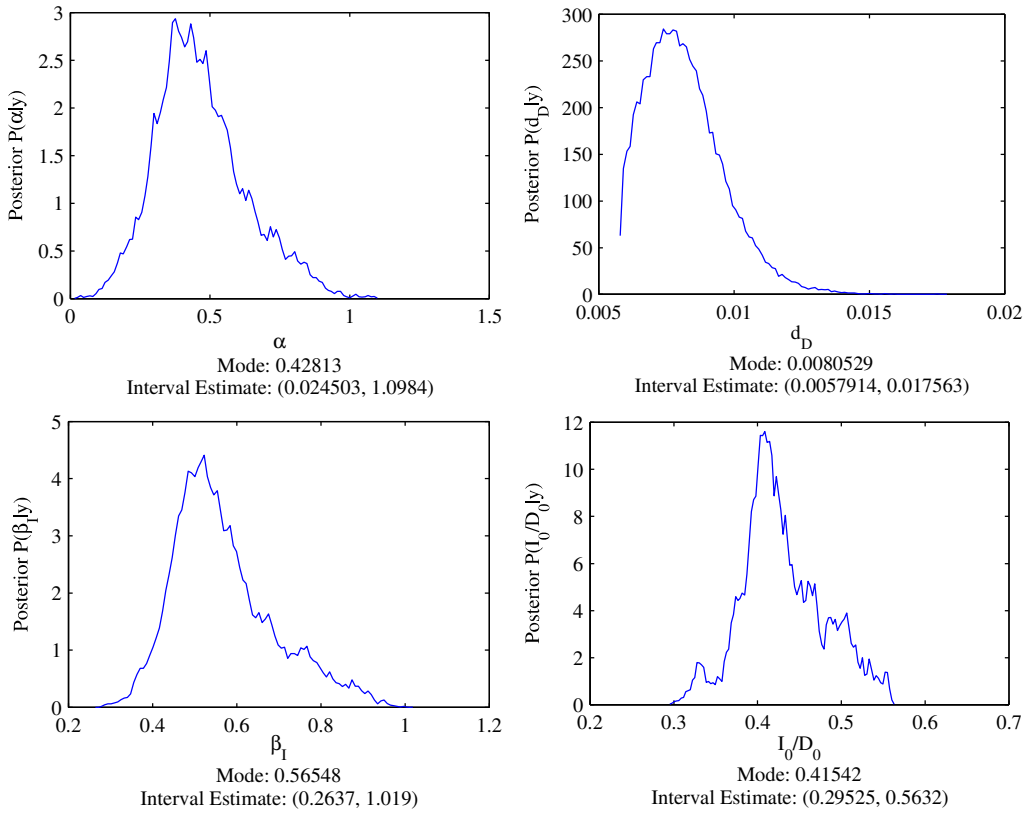


Figure 2.14: Posterior distributions from first MCMC step.

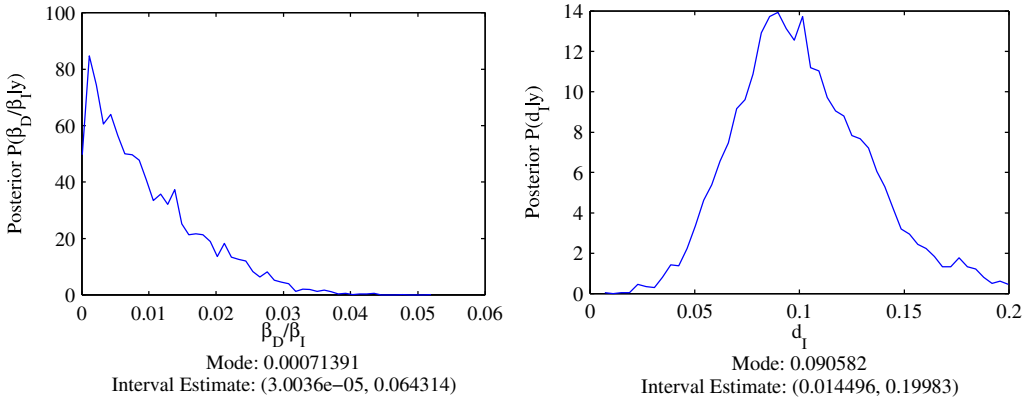


Figure 2.15: Posterior distributions from second and third MCMC step.

Chapter 3

Validating the Model

A model should be validated to to determine how well it performs before being used to make predictions. Validation usually involves comparing model predictions to data. In this chapter, methods for validating models will be discussed and the Bayesian model from Chapter 2 will be validated.

A model that has been calibrated using data will usually have a good correspondence with the data that was used in calibration. Comparing model predictions to the data used for calibration gives little information on model performance in predicting other outcomes. Therefore, it is important to use data not included in the model fitting process to validate the model. In the case of the HIV model from Chapter 2, there are two possible sources of validation data. First, parts of the original data set may be reserved for validation. This is straightforward but has a number of drawbacks. The data sets we are using are limited and it may be undesirable to limit them further by reserving data for validation. Furthermore, the data that we have available gives only indirect information about some quantities of interest. This concern may be addressed by a second source of validation data: independent estimates of model outcomes or parameters. In general, this type of validation data may not be available but, if available, it can provide a broader picture of the performance of the model.

3.1 Validation Methods

Intuitively, when comparing model predictions to validation data, one wishes to determine if the model produces results that are "close" to the validation data. But how close is close enough? We begin our discussion of model validation by introducing the measures of closeness that we will use.

As the the final model created in Chapter 2 is based on Bayesian principles, we will use Bayesian methods to evaluate its performance in predicting validation data. In particular, we will follow the Bayesian hypothesis testing approach outlined in [112, 88]. This method begins by testing a null hypothesis, H_0, against an alternative, H_A, using

the Bayes factor and the posterior null probability. For the purposes of model validation, the hypotheses will be

- H_0: The model is "correct".
- H_A: The model is "incorrect".

In order to be useful, the hypotheses must be formulated more carefully. In fact, there is more than one different type of "correctness" that we may wish to evaluate. Different formulations of the hypotheses will be useful for evaluating these different types of "correctness". Regardless of the specific hypotheses chosen, the Bayes factor is defined as

$$B = \frac{P(y_v|H_0)}{P(y_v|H_A)}$$

where y_v is the validation data. A Bayes factor that is greater than one indicates that the observed validation data is more probable under the null hypothesis than under the alternative hypothesis. Thus with $B > 1$, the validation data gives more support to the null hypothesis than the alternative hypothesis. Similarly, a Bayes factor less than one gives more support to the alternative hypothesis.

The null posterior probability is $P(H_0|y_v)$. In order to compute this quantity we assume prior probabilities $P(H_0)$ and $P(H_A)$ and use Bayes' theorem to write

$$\frac{P(H_0|y_v)}{P(H_A|y_v)} = \frac{P(y_v|H_0)}{P(y_v|H_A)}\frac{P(H_0)}{P(H_A)} = B\frac{P(H_0)}{P(H_A)}.$$

Using the fact that the null and alternative hypotheses are assumed to be exhaustive, this becomes

$$\frac{P(H_0|y_v)}{1 - P(H_0|y_v)} = B\frac{P(H_0)}{1 - P(H_0)}$$

which can be rearranged to give

$$P(H_0|y_v) = \frac{BP(H_0)}{BP(H_0) + 1 - P(H_0)}.$$

This quantity gives a measure of the strength of the evidence for the null hypothesis given by the data. If the prior probabilities of $P(H_0)$ and $P(H_A)$ are assumed to both equal 0.5, this quantity will simplify to

$$P(H_0|y_v) = \frac{B}{B+1}.$$

3.1.1 Types of Hypotheses

Hypotheses about Distributions: To validate the estimated distributions for the model outcomes we choose the null and alternative hypotheses as

- H_0: The model gives the true distribution for θ

- H_A: The true distribution for θ is something else

where θ is the estimated quantity whose distribution is to be validated. In order to complete the computation of the Bayes factor, we will need to specify a distribution to be used for H_A. In general we will choose a minimally informative distribution such as a uniform or triangular distribution for H_A. Using a uniform distribution requires that bounds are placed on the possible values of θ, while using a triangular distribution also requires that the mode of the distribution be specified. Once the distribution to be used in H_A is specified, the computation of the Bayes factor and posterior null probability are straightforward.

Whatever the source of the validation data, it must be assumed that there is some measurement error associated with the data collection. In most cases we will assume

$$y_v = y + \varepsilon$$

where y is the true value of the quantity being measured and ε is a random variable with $\varepsilon \sim N(0, \sigma^2)$. Knowledge about the source of validation data can guide us in selecting an appropriate value for σ. Incorporating the presence of measurement error requires an integration:

$$P(y_v|H_0) = \int f(y_v|y) f_{H_0}(y) dy$$

where $f(y_v|y)$ is the probability density function for the measured data given that the true value is y, and $f_{H_0}(y)$ is the probability density function for y under the assumption that H_0 is true. The quantity $P(y_v|H_A)$ for the alternate hypothesis is computed in exactly the same way.

These quantities are all based on a Bayesian framework. In a classical hypothesis testing framework, a p-value is commonly used. The p-value is defined as

$$p = P(T \geq T(y_v)|H_0) \tag{3.1}$$

where T is some test statistic whose distribution can be computed assuming the null hypothesis is true and $T(y_v)$ is the actual observed value of the test statistic. The p-value is a measure of how unusual the observed data is under the null hypothesis. If the observed data is very unusual, this provides evidence that the null hypothesis is unreasonable. While p-values are usually used in a classical framework and take into account only the distribution of the measurement error in determining the distribution for T, computing $p = P(T \geq T(y_v)|H_0)$ is also possible in the context of validating distributions in the Bayesian framework. While the "true" value y is assumed to be fixed in the classical framework, it is assumed to be a random variable with a distribution on the Bayesian context. The measured value y_v is assumed to have a distribution in both

the classical and Bayesian contexts. In the Bayesian context, we can compute

$$p = \int P(T \geq T(y)|H_0) f(y_v|y) dy.$$

This quantity will behave like a p-value, a familiar quantity to researchers in many applied fields. In particular, small values of p indicate evidence against the null hypothesis.

There are a number of different test statistics that could be used to produce p-values. For our purposes, we will use

$$T(y) = (y - \bar{y}) C^{-1} (y - \bar{y})^T$$

where $\bar{y}$ is the mean and C is the variance-covariance matrix for y under the null hypothesis. This test statistic can be interpreted as the squared distance from y to $\bar{y}$ using a distance measure that takes into account the covariance structure of y.

Hypotheses about Regions for Parameters: Rather than validating the estimated distributions for the model outcomes, we may instead wish to validate the interval or region estimates for the parameters. In this case we will formulate the null and alternative hypotheses differently:

- H_0: $\theta \in I$.
- H_A: $\theta \notin I$.

In these expressions, θ is again the quantity to be validated. It is a scalar or a vector consisting of parameter values. The interval I may be a one dimensional interval or a region in a higher dimensional space. Computing the probabilities $P(y_v|H_0)$ and $P(y_v|H_A)$, we may write B as

$$B = \frac{\int_I f(y_v|y(\theta)) f(\theta) d\theta}{\int_{I^C} f(y_v|\theta) f(\theta) d\theta}$$

where I^C is the complement of I and $f(\theta)$ is the probability density function for θ and $f(y_v|y(\theta))$ is the probability density function for the validation data accounting for measurement error. Notice that the value of B must be larger if a larger set I is used. In other words, if there is sufficient evidence to support $\theta \in I_1$ then the same evidence also supports $\theta \in I_2$ if $I_1 \subset I_2$.

In Sections 3.2 and 3.3, validation is performed using both reserved data and results from independent studies. In each case, two sets of hypotheses are considered.

- First, the distributions for the model outcomes obtained from the distributions for the parameter values found in Section 2.3.3 will be validated by comparing the modelled distribution to a triangular distribution with mode at the measured

Validating Distributions			
	B	$P(H_0\|y_v)$	p-value
2011 and 2012 Diagnoses	8.4214	0.8939	0.6485
1999 and 2000 Diagnoses	1.7659	0.6405	0.8371
1999 and 2000 Deaths	0.0013	0.0013	0.0002

Validating 50% HPD Region		
	B	$P(H_0\|y_v)$
2011 and 2012 Diagnoses	1.5186	0.6030
1999 and 2000 Diagnoses	1.3910	0.5818
1999 and 2000 Deaths	0.0032	0.0032

Table 3.1: Validation results with reserved data

value of the validation data and maximum and minimum defined by a 99 percentile range of the modelled distribution.

- Secondly, the confidence set obtained as a 50% highest posterior density (HPD) region for the parameters is validated. Validation of this region also acts as validation of any larger region – for example the 95% confidence region or the joint 95% confidence intervals obtained by projecting this region onto the parameter axes.

3.2 Validation with Reserved Data

A validation dataset can be created simply by reserving some of the data for validation before fitting begins. Given the small sizes of the datasets that are available, it is tempting to include all available data in the fitting routine. However, as we have already mentioned this may cause the model to appear better than it really is. As we wish to use the model to project into the future as well as simply describing the present, it is vital to have a validation data set which was not used in the fitting procedure. The data for the Province of Alberta model was originally accessed in 2011. At that time, diagnosis and deaths numbers were only complete to the end of 2010. Since that time, further data has become available. Number of new diagnosis for 2011 and 2012 will be the primary validation data. Additionally, our modelling dataset begins in 2001 even though data is available from 1999 onward. This allows us to additionally validate the model using diagnosis and deaths data from 1999 and 2000.

The validation results are summarized in Table 3.1. Bayes factors, $B > 1$ indicate that the additional diagnosis data for both the 2011-2012 and 1999-2000 time periods provide support for the modelled results. The posterior null probabilities, $P(H_0|y_v) > 0.5$ indicate the validation data is more likely to come from the modelled distributions

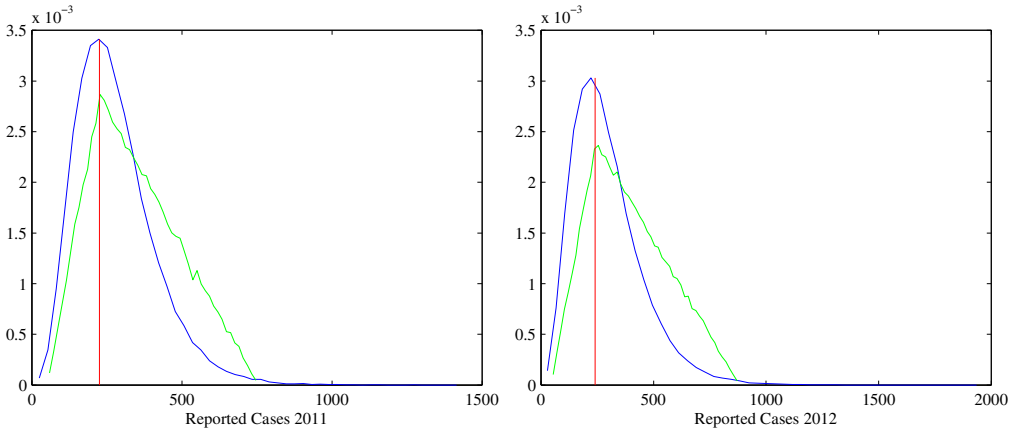


Figure 3.1: Marginal distributions for the number of reported cases in 2011 and 2012 given the null hypothesis that the sampled distribution is correct (blue) and the alternative hypothesis that these quantities follow independent triangular distributions (green), along with the measured validation data (red).

for these quantities than from the triangular distributions to which they were compared.

To further illustrate these results we plot the distributions for the reported cases in 2011 and 2012 under the null and alternative hypotheses. As this validation data is two dimensional, we can view either the marginal distributions for reported cases in each year or the two dimensional joint distribution of both outcomes. Figure 3.1 displays the marginal distributions and shows that the modelled density is higher than the alternative density near the measured data value, indicating a good validation result. Figure 3.2 shows a contour plot of the two dimensional joint distributions. The solid black dot indicates the validation data and again falls in a region where the modelled density is higher than the alternative density.

Figures 3.3 and 3.4 show the same distributions for the reported cases in 1999 and 2000.

The 50% HPD regions being validated in the second type of validation are regions for the fitted parameters. Figure 3.5 shows the estimated marginal distribution for the number of reported cases in 2011 and 2012 under the null hypothesis that the parameters are in the 50% HPD region and under the alternative hypothesis that the parameters are outside of this region. Figure 3.6 shows the joint distributions. As before, the data is more probable under the null hypothesis than under the alternative hypothesis.

In contrast, the data on HIV/AIDS deaths in 1999 and 2000 does not validate our model results. The Bayes factor, B, for this data found in Table 3.1 is much less than one and the posterior null probability is likewise small. This is not unexpected as improvements in treatment have decreased the rate of HIV/AIDS deaths since 2000. Furthermore, the diagnosed HIV-positive population has changed due to an increased availability of early testing further decreasing the death rate. Indeed, the model predicts

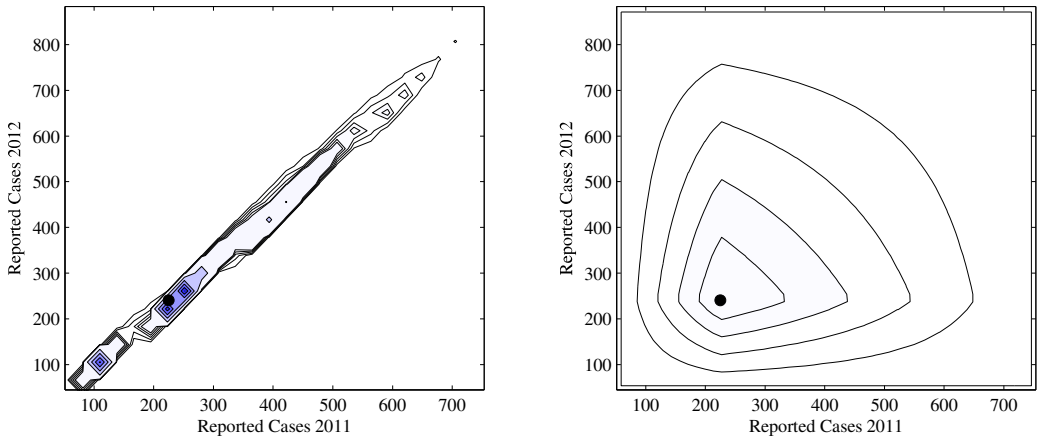


Figure 3.2: Joint distribution for the number of reported cases in 2011 and 2012 given the null hypothesis that the sampled distribution is correct (left) and the alternative hypothesis that these quantities follow independent triangular distributions (right). The black dot indicates the measured validation data value.

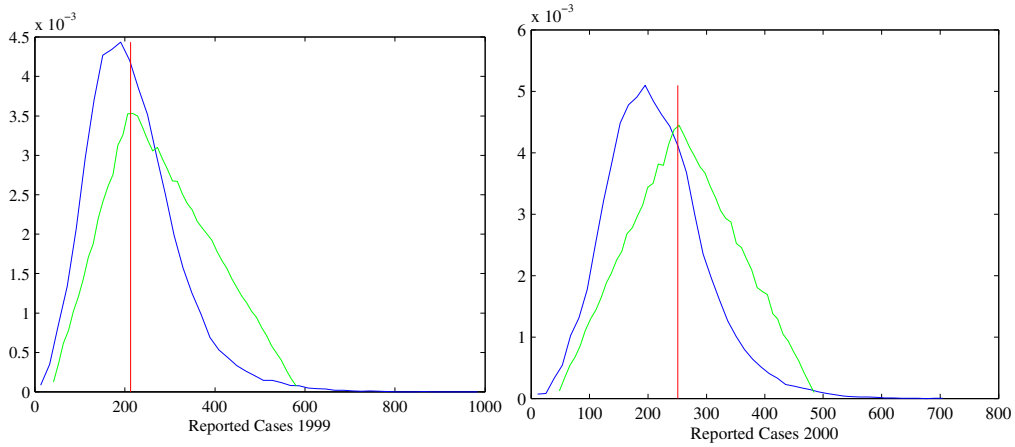


Figure 3.3: Marginal distributions for the number of reported cases in 1999 and 2000 given the null hypothesis that the sampled distribution is correct (blue) and the alternative hypothesis that these quantities follow independent triangular distributions (green), along with the measured validation data (red).

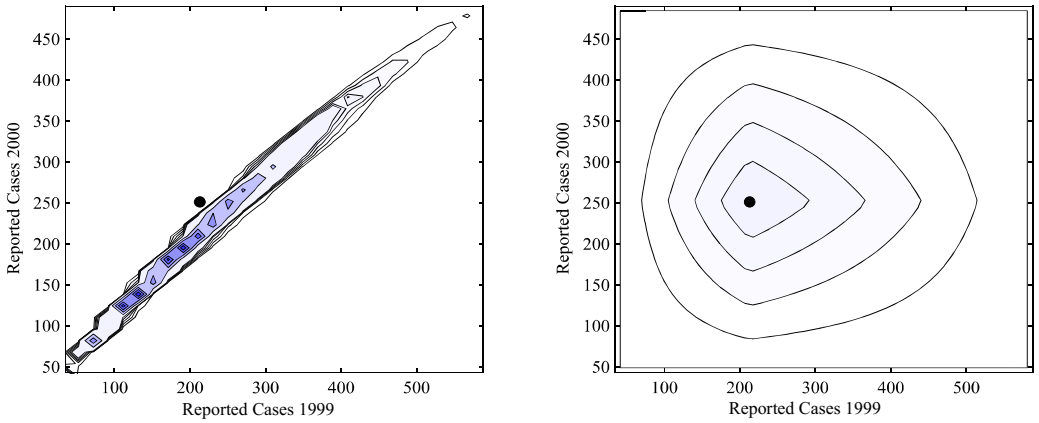


Figure 3.4: Joint distribution for the number of reported cases in 1999 and 2000 given the null hypothesis that the sampled distribution is correct (left) and the alternative hypothesis that these quantities follow independent triangular distributions (right). The black dot indicates the measured validation data value.

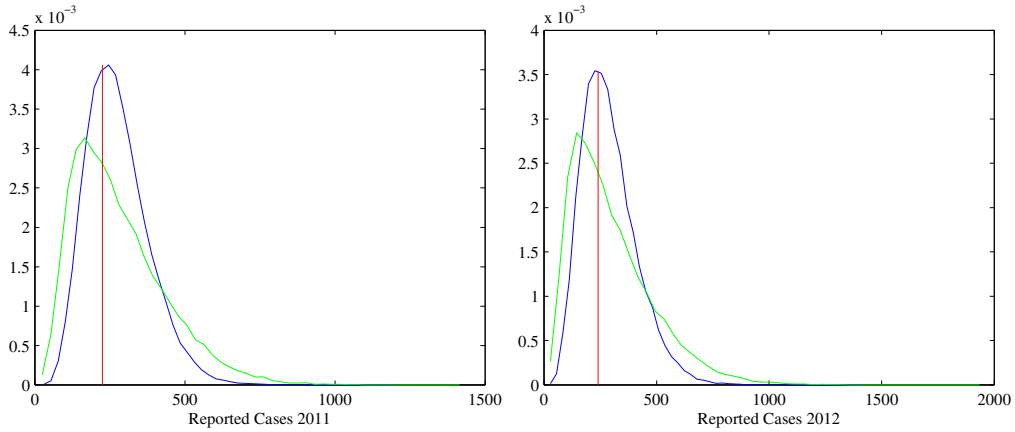


Figure 3.5: Marginal distributions for the number of reported cases in 2011 and 2012 given the null hypothesis that the parameters are in the 50% credible region around the point estimate (blue), and the alternative hypothesis that the parameters are outside of this region (green), along with the measured validation data (red).

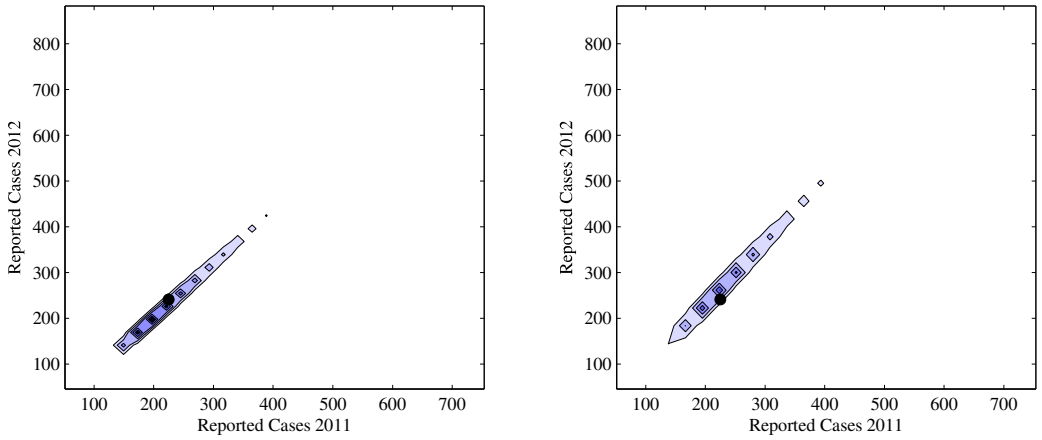


Figure 3.6: Joint distribution for the number of reported cases in 2011 and 2012 given the null hypothesis that the parameters are in the 50% credible region around the point estimate (left) and the alternative hypothesis that that the parameters are outside of this region (right). The black dot indicates the measured validation data value.

far fewer deaths for the 1999-2000 period than were actually reported for the same time period. Distributions under both the null and alternative hypotheses are illustrated in Figure 3.7. It is clear that the observed validation data, which occurs at the mode of the alternative distribution, is much larger than predicted by the modelled distribution.

Overall, the validation using reserved data is either successful, providing evidence that our model is appropriate, or fails in predictable ways, indicating limitations of our model that are not surprising.

3.3 Validation with Independent Results

The results of independent studies of HIV incidence, prevalence, and diagnosis can also be used to validate the model. These may include studies estimating the prevalence of HIV or evaluating the success of intervention programs. Any quantities computed in other studies that can also be estimated using the model developed in this thesis can be used for validation. The validation results obtained using this type of data are limited by the amount of similarity between the population included in the independent study and the general population of Alberta which is described by the model. However, this type of data can allow us to evaluate how well the model corresponds with reality for a wider variety of model outcomes. The model for the Province of Alberta is validated using

- 2005, 2008, and 2011 PHAC estimates of the percentage of the HIV-positive population who are undiagnosed [19, 149, 31].
- 1998 and 2006 emergency department sentinel surveillance studies [71, 70].

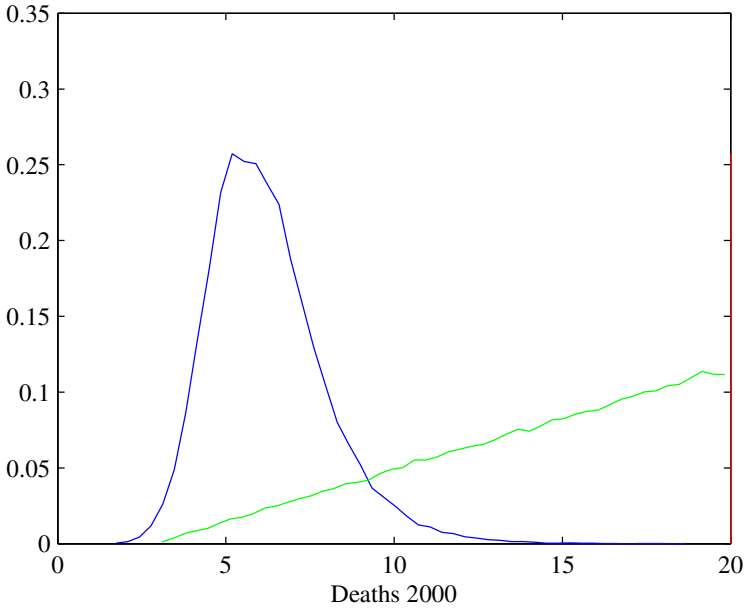


Figure 3.7: Distribution of the number of HIV deaths in the year 2000 given the null hypothesis that sampled modelled is correct (blue), and the alternative hypothesis that this quantity follows a triangular distribution (green). The observed validation data value for this quantity is 20.

- Results of 2002-2004 prenatal screening [106].

Each of these populations differs from the general population of the province of Alberta which the model describes. The PHAC estimates include the entire Canadian population which may differ somewhat from the population in the Province of Alberta. The emergency department studies considered only those undergoing blood tests in emergency departments. This group is likely to be at higher risk of HIV than the general population. Similarly, those undergoing prenatal HIV screening are also a specific subset of the population of Alberta. This group is likely to be at lower risk of HIV than the general population.

Nonetheless, comparison to these quantities will act as a validation of the model results. For some of these studies, the population involved is similar enough to the general population of Alberta that we expect the validation to be successful. For others of these studies, the population is thought to be different enough from the general Alberta population that it will be a greater cause for concern if the model agrees with the data than if it does not.

The validation results for these independent studies found in Table 3.2 have the expected outcome. In particular, the PHAC estimates validate the modelled results with Bayes factors greater than one for both for the distribution and for the 50% HPD Region. This is the case when considering either the most recent two estimates or when combining the data from the years 2002, 2005, 2008, and 2011. The resulting

Validating Distributions

	B	$P(H_0\|y_v)$	p-value
2008 and 2011 PHAC Estimates	1.4387	0.5899	0.2410
2002 – 2011 PHAC Estimates	3.5006	0.7778	0.1829
1998 Emergency Study	0.0049	0.0049	0.0000
2006 Emergency Study	0.0000	0.0000	0.0000
Prenatal Screening	0.0133	0.0131	0.0013

Validating 50% HPD Region

	B	$P(H_0\|y_v)$
2008 and 2011 PHAC Estimates	1.1395	0.5326
2002 – 2011 PHAC Estimates	1.1744	0.5401
Prenatal Screening	0.4091	0.2903

Table 3.2: Validation using independent results

four-dimensional distribution validates easily with $B = 3.5006$ and $P(H_0|y_v) = 0.7778$.

The model behaves only moderately well compared to each of the PHAC estimates individually. In each estimate, the model somewhat underestimates the fraction of the HIV-positive population that is undiagnosed. This is illustrated for the years 2008 and 2011 in Figure 3.8. In this figure, it can be seen that the majority of the modelled distribution falls below the validation data values and the alternative distribution has higher density near the validation data. Nonetheless the model predicts the correct trend in the fraction of the HIV-positive population that is undiagnosed. Correctly predicting this trend increases confidence in the null hypothesis that the modelled distributions are correct. The joint distribution is illustrated in Figure 3.9.

When validating the 50% HPD Region using PHAC estimates of the fraction of the HIV-positive population who are undiagnosed, the validation is always successful. Bayes factors greater than one in Table 3.2 indicate that validation data is more likely to be produced by parameter values inside the 50% HPD Region than by parameter values outside this region. The distributions under the null and alternative hypotheses for this test are found in Figures 3.10 and 3.11.

On the other hand, the emergency room studies and the prenatal screening data do not validate the model. The Bayes factors for these quantities, in Table 3.2 are all less than one, and some of them are quite small. As already mentioned, this is to be expected. The emergency room study considers only a very high risk population and the model estimates a much lower HIV prevalence than measured in the emergency room setting. On the other hand prenatal screening applies only to a population at lower risk than the general population and the model estimates a somewhat higher HIV prevalence than is measured for this population.

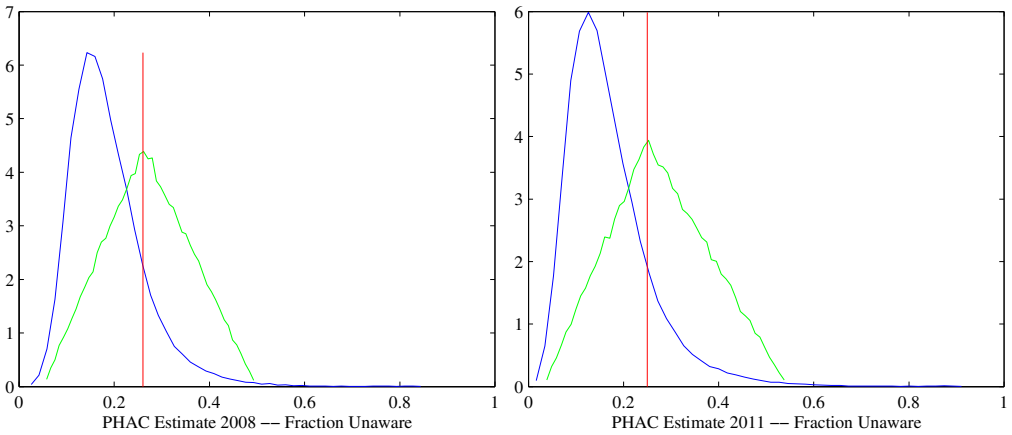


Figure 3.8: Marginal distributions for the fraction of the HIV-positive population undiagnosed in 2008 and 2011 under the null hypothesis that the sampled distribution is correct (blue) and the alternative hypothesis that these quantities follow triangular distributions (green). The size of the undiagnosed HIV-positive population as estimated by PHAC is shown in red.

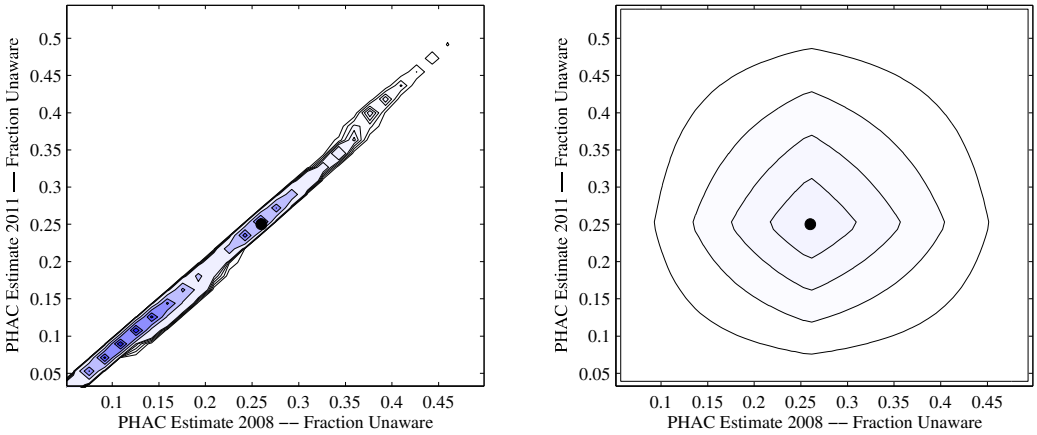


Figure 3.9: Joint distributions of the fraction of the HIV-positive population undiagnosed in 2008 and 2011 under the null hypothesis that the sampled distribution is correct (left) and the alternative hypothesis that these quantities follow independent triangular distributions (right). The validation data is indicated by the black dot.

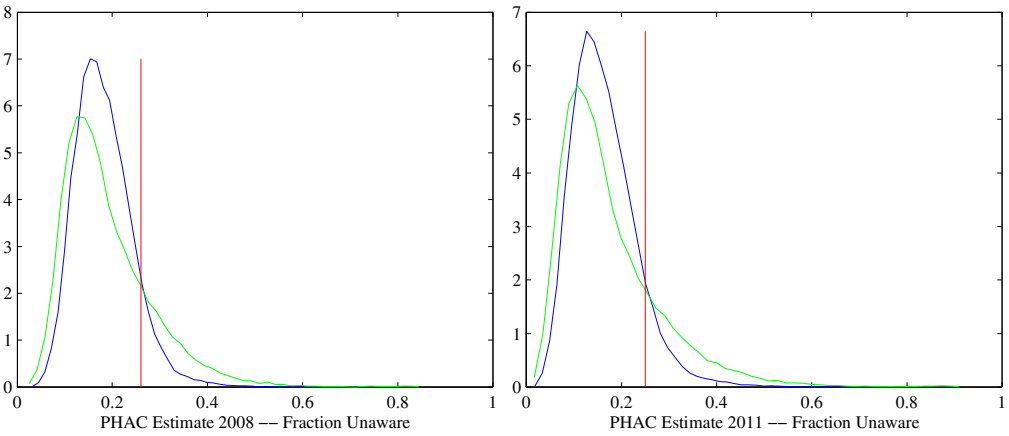


Figure 3.10: Marginal distributions for the fraction of the HIV-positive population undiagnosed in 2008 and 2011 under the null hypothesis that the parameter values are with their 50% credible region (blue) and the alternative hypothesis that the parameter values are outside this region (green). The PHAC estimate for this quantity is shown in red.

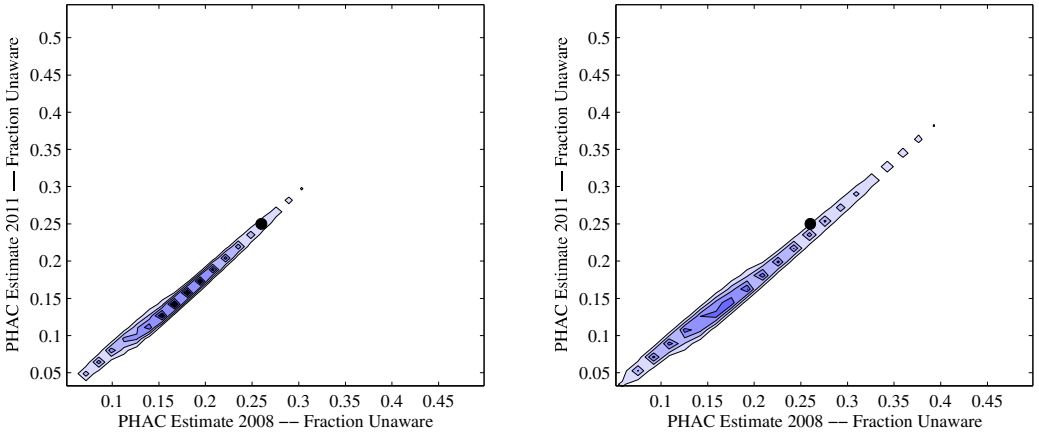


Figure 3.11: Joint distributions for the fraction of the HIV-positive population undiagnosed in 2008 and 2011 under the null hypothesis that parameter values are within their 50 % credible region (left) and the alternative hypothesis that the parameter values are outside this region (right). The PHAC estimate for these quantities is shown indicated with a black dot.

Overall, the validation with independent results is successful. The PHAC estimates provide some support for our model, especially when taken together. The emergency room studies and prenatal data do not support the model results. However, this is to be expected as the populations in these studies are not directly comparable to the general population of the province modelled by the current project.

Chapter 4

Using the Model

Since it includes both transmission and diagnosis dynamics, the model that has been created in Chapter 2 and validated in Chapter 3 allows the description of a variety of features of disease transmission that are not easily studied by other methods. In this chapter, we estimate the total size of the population living with HIV, the time to diagnosis, and HIV incidence. The model is also used to predict the future course of the disease and the potential outcome of changes to public health programs.

In order to produce these estimates it is necessary to describe how uncertainty in the parameter values will be accounted for. Some uncertainty remains in the parameter estimates and is described by the distributions and interval estimates computed. The result of this uncertainty in the parameter values is that the model results for the quantities of interest will also be uncertain.

To determine which parameters might have important effects on the outcomes of interest, we employ sensitivity analysis. Both relative local sensitivity coefficients and global sensitivity coefficients based on partial rank correlations will be used. Each of these methods of sensitivity analysis gives slightly different information about how the parameter values relate to the outcomes.

4.1 Uncertainty Analysis

It is important to take into consideration the fact that the parameter values we have calculated for the model are uncertain. In fact, we have calculated a number of different possible parameter values for the model. The least squares fit gives a single parameter value that best fits the data, but the data itself may contain some random variation and perhaps even errors. Therefore, we have also used Bayesian methods to fit the model resulting in not only a single point estimate for the parameters, but also interval estimates and estimates of distributions for the parameters.

There are a number of different methods available for uncertainty analysis [35, 64]. We will use a method based on a random sample of the uncertain quantities. The core

of this method is to use the distributions for the uncertain quantities (the parameters) and the model equations to create distributions for the outcomes of interest – size of the HIV-positive population in 2010 and 2015 for example. The Bayesian model fitting that was carried out in Chapter 2 resulted in a sample from the joint distribution of the parameters. This is the sample we will use in the uncertainty analysis.

If a sample was not already available for the uncertain quantities, we would need to create one. A common technique is to use a Latin hypercube sample as in Section 5.1, but any sampling technique appropriate to the desired parameter distribution can be used [16, 63]. A brief introduction to Latin hypercube sampling is given in Appendix A.2.2.

Transforming the sample of parameter values to create a sample for the model outcomes of interest requires computing the model outcomes for each element of the parameter sample and thus creating a sample from the distribution of model outcomes. Once this sample is created, it must be analyzed and appropriate conclusions drawn from it. This analysis may take a number of different forms but is likely to include the computation of summary statistics and result in point and interval estimates for the outcomes of interest.

4.1.1 Reporting Uncertainty Results

One common method of reporting the results of an uncertainty analysis is to use a box plot. Examples of box plots can be found in Figures 4.1, 4.3, and several other figures throughout this chapter. This plot displays the median, interquartile range, and 95th percentile range of the sampled outputs. In the plot style used here, the median is indicated by a circle, the interquartile range by a narrow filled box, and the 95th percentile range by vertical lines. This information is often included for the same model output calculated at several different time points and displayed on the same plot. This type of plot can give a simplified view of how the distribution of a quantity is changing over time.

However, this style of plot has a significant drawback for reporting results from the differential equation based models that are used in this project. In particular, the box plot discards all information about the relationships between model outcomes that may have been contained in the sample. While a box plot gives no information about the relationship between quantities, the outcomes we are interested in are related to each other. A box plot may give an incorrect impression of these relationships.

For example, consider the box plots in Figure 4.1. This figure shows box plots (A, B, and C) for three different samples. Sample paths for these samples are also illustrated (D, E, and F). In each of these samples the box plots are identical at the points 1 and 2. A viewer may be tempted to conclude that the value of the underlying quantities does not change between these two points. This is the case in Figure 4.1 A and D. However, in Figure 4.1 B and E the sign of the quantity changes between these two points. In Figure

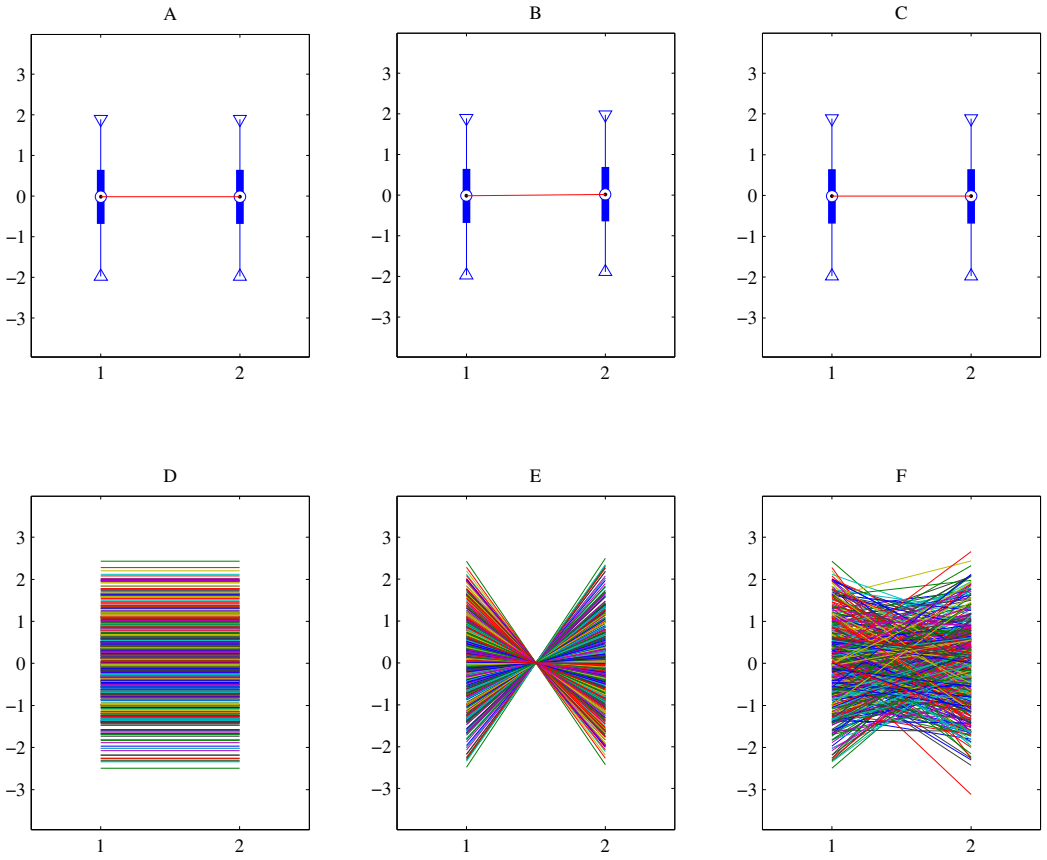


Figure 4.1: Identical boxplots for three different samples. In the first (A, D), the quantity is equal at t_1 and t_2. In the second (B, E), the quantity at t_2 is negative that at t_1. In the third (C, F), the two quantities are statistically independent.

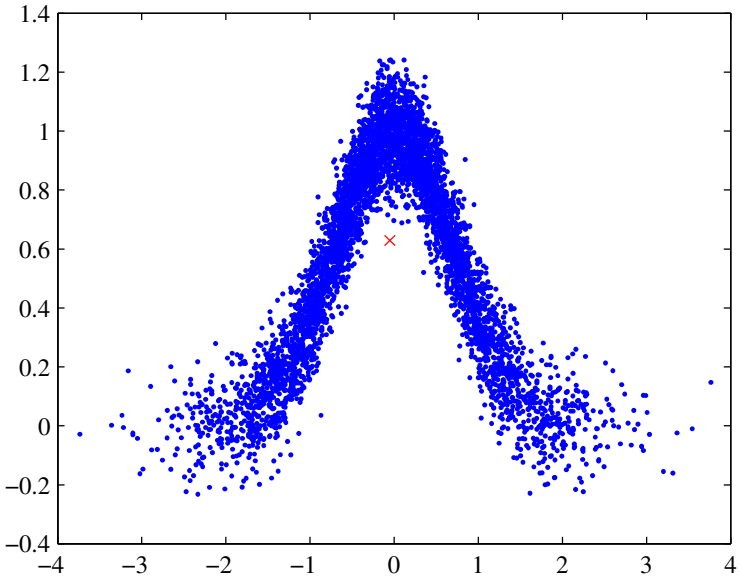


Figure 4.2: The two components of this data have a nonlinear relationship and illustrate how the use of marginal medians can be nonrepresentative of nonlinear data.

4.1 C and F the value of the quantities at the two points are statistically independent. The box plot cannot distinguish between these three situations. This concern may be partially remedied by also displaying the sample paths in the underlying data, however showing data in such a raw form can be somewhat difficult to interpret.

The medians highlighted by the box plot may also give an incomplete impression. Consider the data illustrated in Figure 4.2. The two components of this data have a nonlinear relationship with each other. The medians of each component of the data however do not follow this relationship. Reporting only the marginal medians, as indicated by the $\times$ in the figure, may suggest that it is possible to have elements of the sample near the point in the plane identified by the pair of medians.

Despite these limitations of the box plot, it is a commonly used method of summarizing some types of uncertainty results and it continues to be a useful tool for describing the possible range of many model outcomes simultaneously. However, it has already been noted that model outcomes of interest may have important relationships with each other. Figure 3.4 on page 50 illustrated a strong relationship between the number of cases reported in two subsequent years. To avoid discarding information about the relationships between quantities of interest, we develop a new way of summarizing uncertainty results which can augment the results often summarized in the box plot.

The method we will use to summarize uncertainty results is based on a simple idea: whenever we are interested in the relationships between model outcomes, they should not be analyzed one at a time but instead multivariate methods should be used to consider several model outcomes simultaneously.

There are several multivariate methods that could be appropriate to summarize uncertainty results. A simple method could use the mean vector and the variance-covariance matrix to create point estimates and prediction regions for the multivariate model outcome. However, determining point estimates and prediction regions from these quantities requires assumptions about the joint distribution for the model outcomes. In particular, this method works well if the outcome distribution is a multivariate normal distribution. Since the distributions we are considering are not in general normal, the point estimates and prediction regions given by this method may not be appropriate.

The correlation matrix can be used to quantify the strength of a linear relationship between model outcomes. However, as these relationships may be nonlinear this may fail to capture some of the relationships that we are interested in. Therefore, for the purposes of this thesis, we will use a method based on a density estimate. A discussion of density estimation can be found in [117]. The method is as follows:

- A kernel density estimator is computed for the multivariate model outcome under consideration.
- The maximum density sample point is taken as the point estimate for the multivariate model outcome. This point may be interpreted as a "representative" outcome or the "most likely" outcome.
- A prediction region for the model outcome is defined using the 50% or 95% highest density region. The 50% region corresponds to the interquartile range often shown in a box plot, while the 95% region is a more common size for a prediction region. This region can be projected onto the model outcome axes to create joint prediction intervals the for individual outcomes.

When considering only two model outcomes, contour plots of the 50% and 95% highest density regions along with the maximum density point can be produced. These plots may be considered to be a two dimensional version of the box plot.

This method has the advantage of preserving any relationships between the components of the multivariate model outcome that is used. Additionally, since an element of the sample will be chosen as the point estimate, it is guaranteed that the behaviour illustrated by the point estimate is a possible behaviour of the model. One limitation in using this method is that it quickly becomes unwieldy when attempting to analyze many model outcomes simultaneously. A great deal of data is required to adequately resolve a high dimensional density estimate and results in higher dimensions are difficult to visualize and report. Therefore, we will use this method to investigate relationships between only two or three model outcomes simultaneously.

Year	Median	25	75	2.5	97.5
2000	1319	1269	1365	1146	1452
2005	2372	2234	2514	1961	2809
2010	3558	3099	4088	2369	5437
2015	4933	3909	6328	2602	10995
2020	6582	4699	9569	2726	22844

Table 4.1: Median and percentiles of model results for the estimated total number of people living with HIV in selected years.

4.2 Model Results

The model from Chapter 2 can be used to determine the values of a variety of model outcomes that are not easily determined by other methods. For example, the fitted model can be used to estimate the total size of the HIV-positive population – including those who have not been diagnosed. Additionally, the model can be used to estimate the fraction of the HIV-positive population who are undiagnosed, and HIV incidence. All of these quantities can be computed for any time when the model is believed to be appropriate. For the Province of Alberta model, results will be given for the years 2000 to 2020. Recall that the model was fitted on data from 2001 to 2010 and validated using data from 1999 to 2012. Whether or not the model will remain valid until 2020 depends on the amount of change in HIV transmission behaviours, diagnosis patterns, and treatment effectiveness that occurs in this time period.

4.2.1 People Living with HIV

A box plot and a selection of sample plots for the total number of people living with HIV for years 2000-2020 are illustrated in Figure 4.3. In these plots the median number of people living with HIV is seen to increase steadily over this time period. At the same time, the amount of uncertainty in this quantity, captured by the interquartile and percentile ranges, increases rapidly. Most of the increase in uncertainty occurs at the top of the range. The lower end of the 95 percentile range is mostly constant after 2010. The median, interquartile range, and 95 percentile range are given in Table 4.1.

Examining the sample included in Figure 4.3, it can be seen that sample elements with a large HIV-positive population in earlier years, usually also have a large HIV-positive population in later years. This suggests that there are relationships among the yearly total number of HIV-positive people. To further investigate this possibility we consider the joint distributions created using the number of HIV-positive people in two different years. Figure 4.4 displays a contour plot of the estimated density. In these plots the filled region indicates a 50% prediction region and the outlined region indicates a 95% prediction region. The maximum density point is also indicated. The density contour plot reveals a moderate association between the number of HIV-positive people in 2005

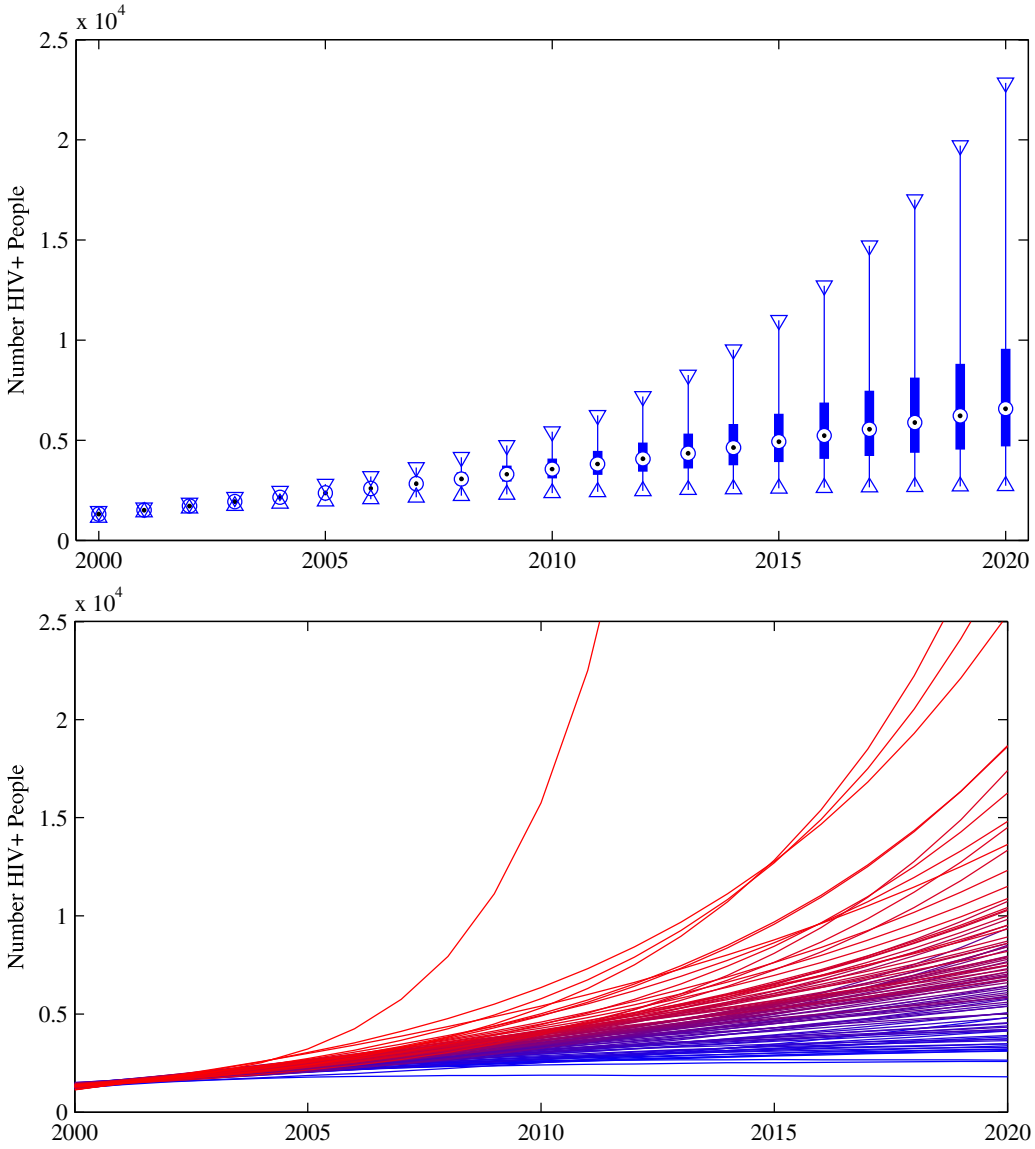


Figure 4.3: Boxplot and a sample of model results for the estimated total number of people living with HIV.

Year	Median	25	75	2.5	97.5
2000	0.343	0.317	0.372	0.271	0.430
2005	0.207	0.171	0.248	0.115	0.355
2010	0.157	0.115	0.213	0.058	0.393
2015	0.134	0.090	0.196	0.036	0.413
2020	0.122	0.077	0.187	0.025	0.422

Table 4.2: Median and percentiles of model results for the estimated fraction of the HIV-positive population who are undiagnosed in selected years.

and 2015. The association between the number of HIV-positive people in 2010 and 2020 is slightly stronger. In particular, a low HIV-positive population in 2005 or in 2010 tends to occur for the same parameter values as a low HIV-positive population a decade later. The low density region in the lower right indicates that significant decreases in the HIV-positive population are unlikely. Nonetheless, there are some parameter values which exhibit a decline in the size of the HIV-positive population from 2010 to 2020. In general, these occur in cases where the HIV-positive population in 2010 is already relatively small. At the same time, the low density region in the upper left reveals that very large increases in the size of HIV-positive population are also unlikely.

4.2.2 Undiagnosed HIV-Positive Population

The faction of the HIV-positive population that is undiagnosed is investigated similarly. A box plot and a selection of sample plots for this fraction in the years 2000-2020 are illustrated in Figure 4.5. In these plots the median fraction is seen to decrease from 2001 to 2010 and appears to level off slightly above 0.1. The decrease may partially be the result of the total number of people with HIV increasing while the number who have not been diagnosed remains relatively constant. At the same time, the amount of uncertainty in this quantity, captured by the interquartile and percentile ranges, increases from 2000 to 2010 but does not change substantially from 2010 to 2020. The median, interquartile range, and 95 percentile range are given in Table 4.2.

Once again, the sample included in Figure 4.5 suggests that there is a relationship between the outcomes for subsequent years. To further investigate this possibility we consider the joint distributions created using the fractions of the HIV-positive population who are undiagnosed in two different years. Figure 4.6 displays the contour plot of the estimated density. As before, the filled region indicates a 50% prediction region and the outlined region indicates a 95% prediction region. The maximum density point is also indicated while the dashed line indicates no change over the decade. These plots reveal a strong association between the fractions of the HIV-positive population who are undiagnosed in 2005 and 2015 and an even stronger association between the fractions of the HIV-positive population who are undiagnosed in 2010 and 2020. For the majority of sampled parameter values this quantity declines over the decades from 2005 to 2015

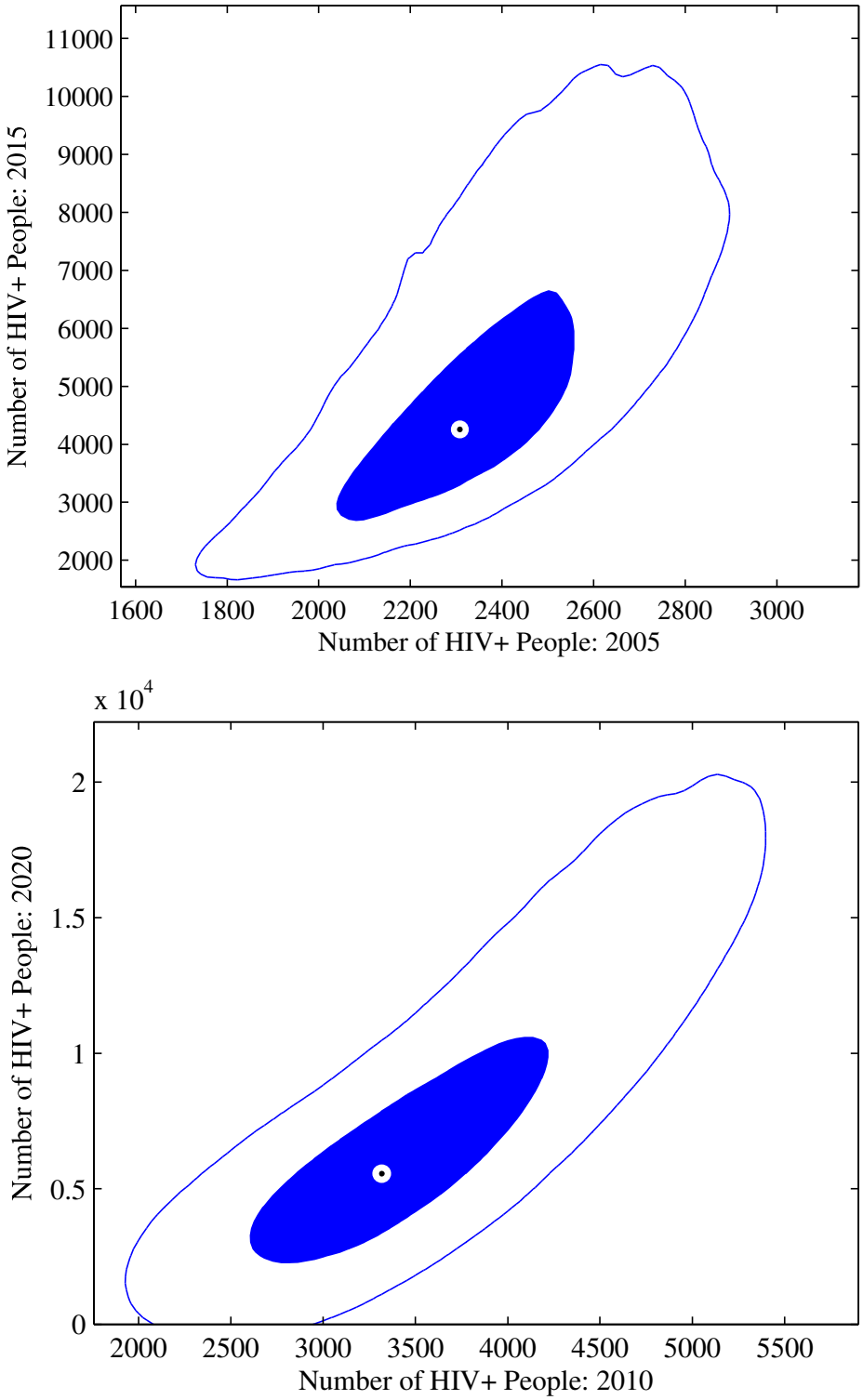


Figure 4.4: Contour plots of the two dimensional distributions of the estimated total number of people living with HIV in 2005/2015 and in 2010/2020.

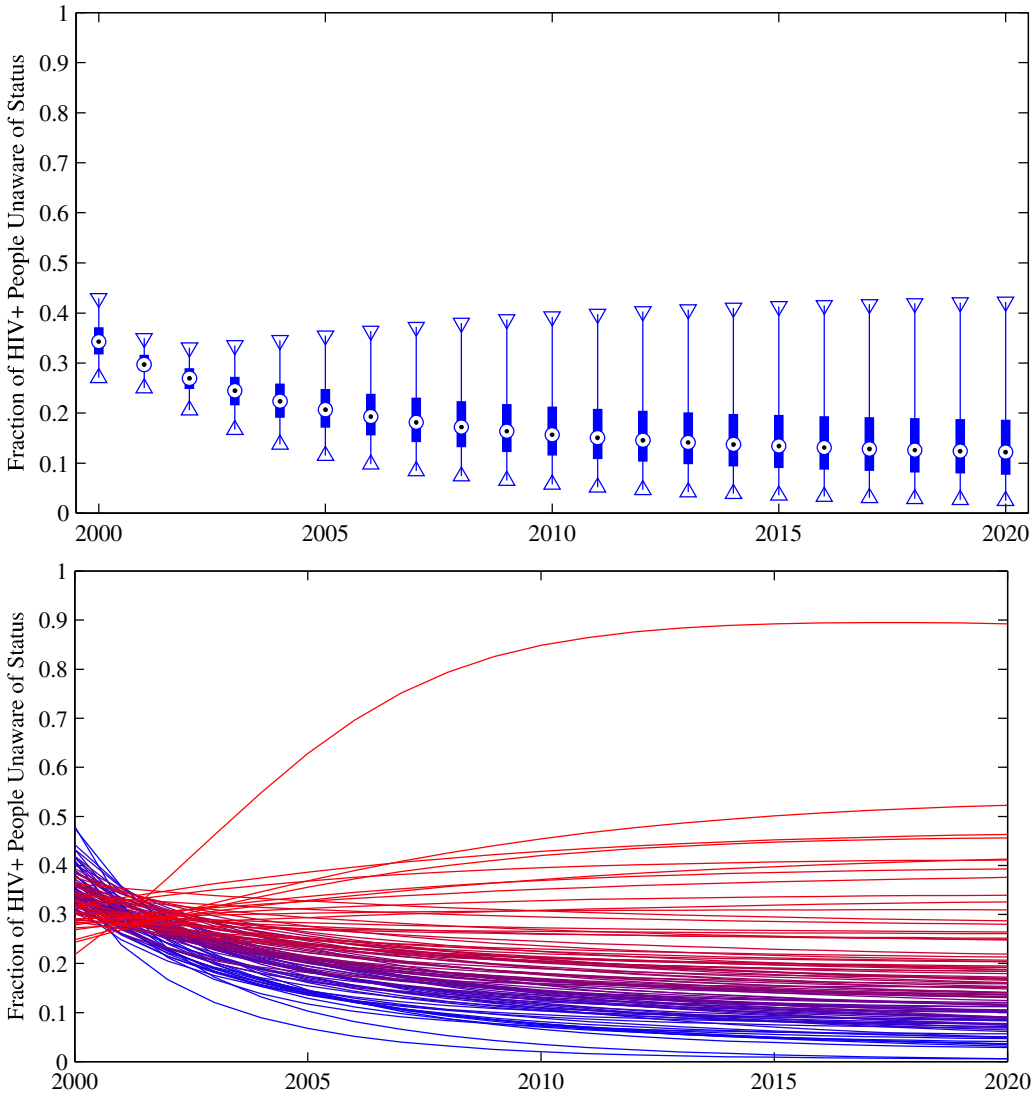


Figure 4.5: Boxplot and a sample of model results for the estimated fraction of the HIV-positive population who are undiagnosed.

Year	Median	25	75	2.5	97.5
2000	8.49	7.26	9.88	5.18	13.48
2005	8.36	7.09	9.68	4.68	12.59
2010	8.73	6.09	12.16	2.67	22.73
2015	9.55	5.51	16.01	1.62	43.86
2020	10.80	5.22	21.49	1.08	85.88

Table 4.3: Median and percentiles of model results for the estimated number of new HIV cases / 100 000 population in selected years.

and from 2010 to 2020. These declines are generally small resulting in prediction regions consisting of a narrow band in the output space. Declines in the fraction of the HIV-positive population who are undiagnosed are greater for small values of this quantity in 2005 and 2010. For some very large values of this quantity, a small increase is observed over a decade.

4.2.3 Incidence

The same methods are used to investigate HIV incidence. A box plot and a selection of sample plots for the number of new HIV cases per 100000 population for years 2000-2020 are illustrated in Figure 4.7. In these plots, the median HIV incidence is seen to remain mostly constant over this time period. However, the amount of uncertainty in this quantity, captured by the interquartile and percentile ranges, increases rapidly for the second half of the time period. The median, interquartile range, and 95 percentile range are given in Table 4.3.

As before the sample included in Figure 4.7 suggests that there are relationships among the incidences of HIV in subsequent years. Particularly after 2005, a the samples with high incidence remain high in future years. To further investigate this possibility we consider the joint distributions created using the incidence of HIV in two different years. These plots are found in Figure 4.8. As before, these plots are a two dimensional version of the box plot with the filled region indicating a 50% prediction region and the outlined region indicating a 95% prediction region. These plots reveal a moderate association between the incidence of HIV in 2005 and 2015 and a somewhat weaker association between the incidence of HIV in 2010 and 2020. The maximum density point exhibits very little change from 2005 to 2015 and even less change from 2010 to 2020.

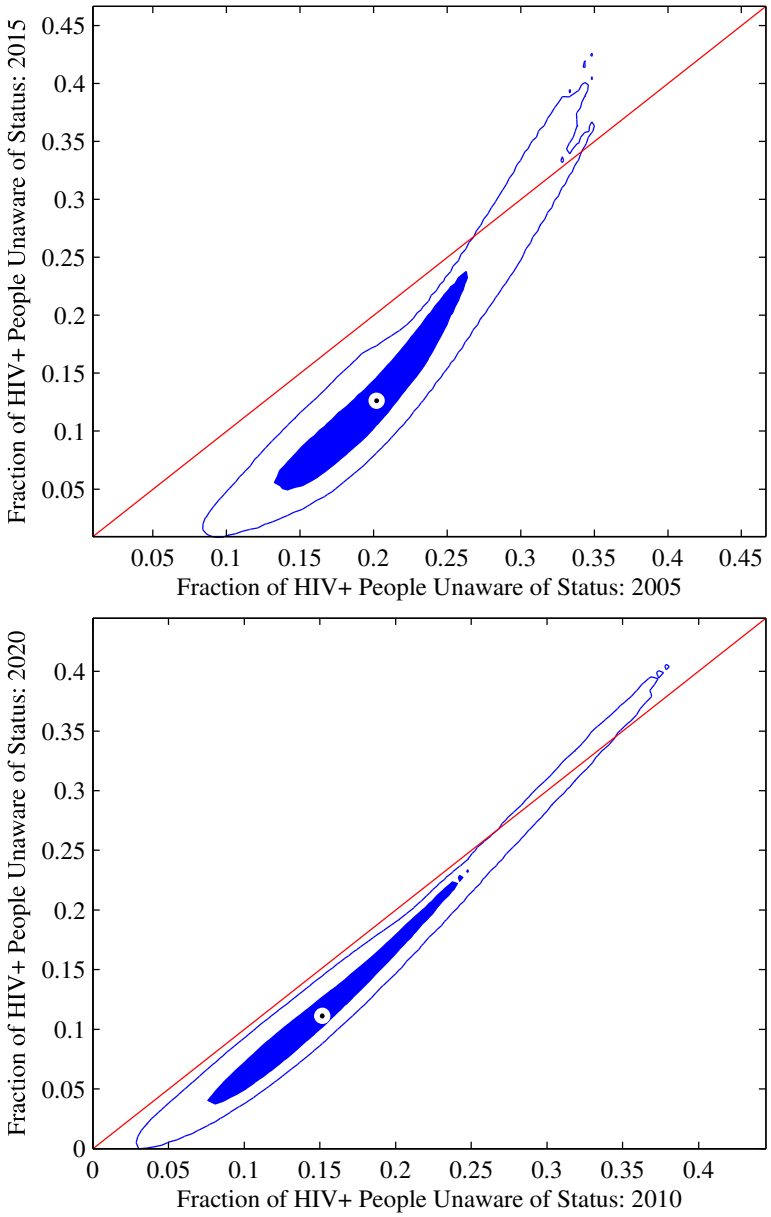


Figure 4.6: Contour plots of the two dimensional distributions of the estimated fraction of the HIV-positive population who are undiagnosed in 2005/2015 and in 2010/2020.

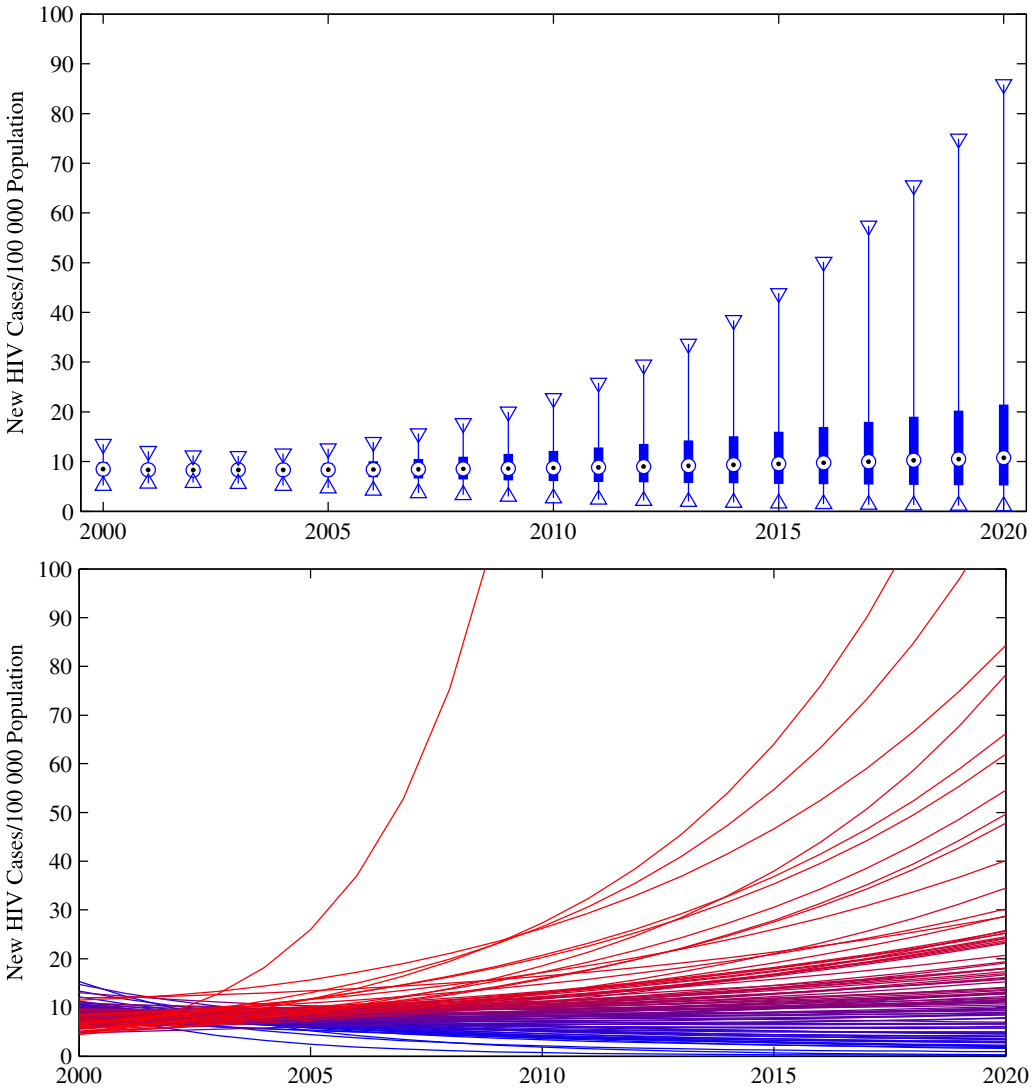


Figure 4.7: Boxplot and a sample of model results for the estimated HIV incidence.

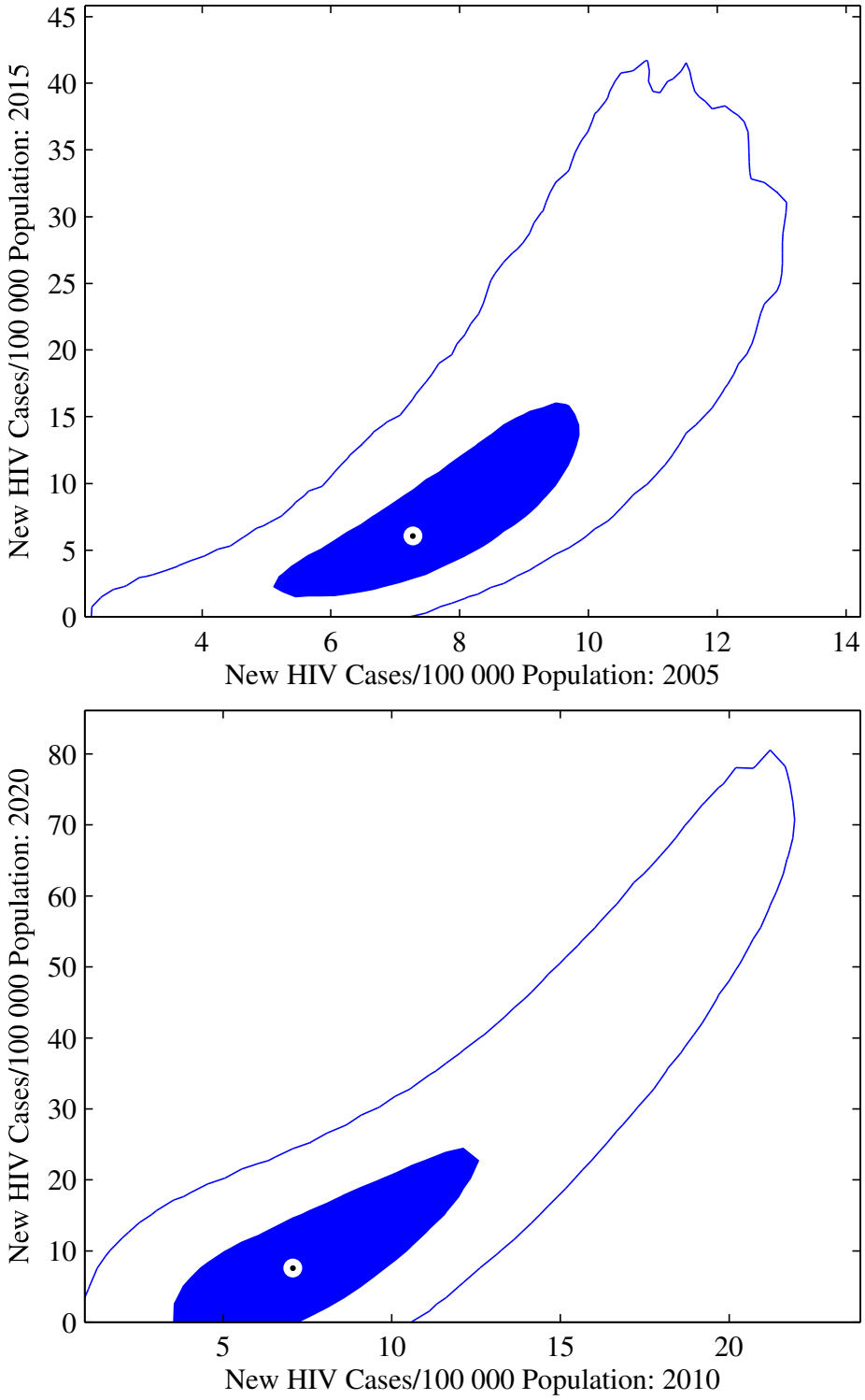


Figure 4.8: Contour plots of the two dimensional distributions of the estimated HIV incidence in 2005/2015 and in 2010/2020.

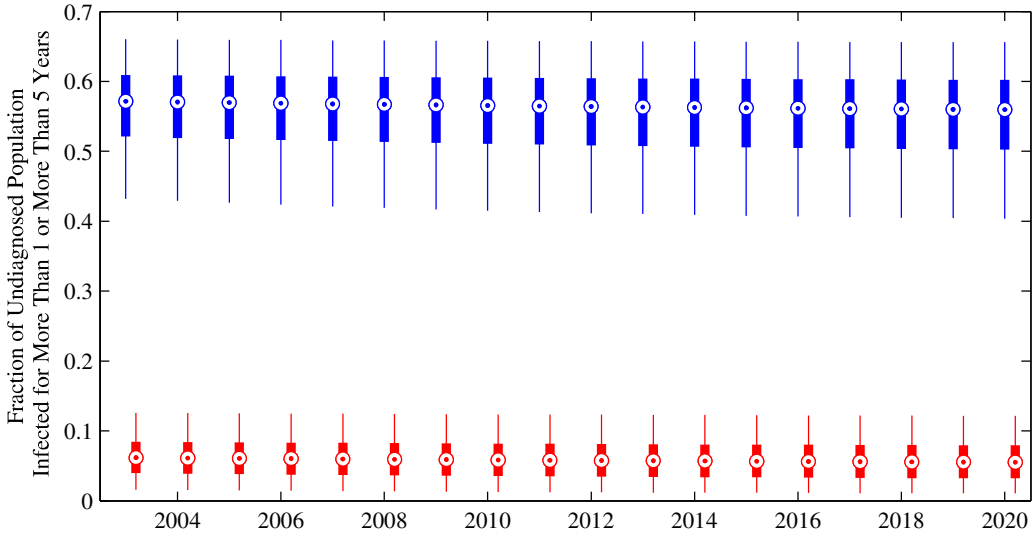


Figure 4.9: The fraction of the undiagnosed population who have been infected for more than 1 year (blue) and more than 5 years (red).

4.2.4 Time to Diagnosis

The average time that individuals being diagnosed have been infected is given by $\frac{1}{\alpha}$. The Bayesian fitting results summarized in Table 2.12 give a point estimate of $\alpha = 0.42813$ with an interval estimate for α of $(0.024503, 1.0984)$. This translates to point and interval estimates for the average time undiagnosed of $\frac{1}{\alpha} = 2.33$ years with an interval estimate of $(0.910, 40.8)$ years. While the top end of this range is not realistic, only a small number of sample points are near the top end.

The fitted model also allows the time between infection and diagnosis to be explored in more detail. In particular the fraction of the undiagnosed population at time t that has been infected for longer than time τ, $\Phi_\tau(t)$, can be calculated. This is accomplished by considering how many of those who were in the undiagnosed population at time $t-\tau$ remain in the undiagnosed population at time t. For the constant death and diagnosis rates that have been used for the Province of Alberta model, this fraction is given by

$$\Phi_\tau(t) = \frac{I(t-\tau)e^{-(\alpha+d_I)\tau}}{I(t)}.$$

Box plots for the resulting fractions $\Phi_1(t)$ and $\Phi_5(t)$ are illustrated in Figure 4.9. These quantities give the fraction of the undiagnosed population that has been infected for more than a year or more than five years respectively. Both of these fractions can be seen to be nearly constant over the time frame simulated. This is due to the fact that constant parameters are used and the population of the compartment I changes very little over the time frame simulated.

The median, interquartile range, and 95% range for the $\Phi_1(2015)$ and $\Phi_5(2015)$ are

	Median	25	75	2.5	97.5
$\Phi_1(2015)$	0.5622	0.5057	0.6036	0.3823	0.6752
$\Phi_5(2015)$	0.0568	0.0337	0.0806	0.0085	0.1407

Table 4.4: The median, interquartile range, and 95% range for fraction of the undiagnosed population who have been infected for at least one year and at least five years in 2015.

found in Table 4.4.

4.3 Sensitivity Analysis

Sensitivity analysis is used to quantify how model outcomes change when model inputs such as parameters change. This idea has multiple uses in a modelling project. In Section 2.3.1, the local relative sensitivity of the model outcomes given in the data with respect to the parameters was used as part of the discussion of parameter identifiability. It was noted that if all model outcomes for which there is data are insensitive to the parameters, it will not be possible to choose appropriate parameter values.

Sensitivity results are sometimes used to augment uncertainty analysis. These results can indicate how much model outcomes could vary as a result of uncertainty in parameter values. As sensitivity indicates what the potential impact of changes to the parameters might be, this information may also be used to suggest potential interventions – parameter changes to achieve a desired effect on model outcomes.

4.3.1 Local Sensitivity

The term *local sensitivity* refers to sensitivity calculated at a single point in the parameter space. This is the type of sensitivity that we have already encountered. The sensitivity of a model outcome $y(p,t)$ to changes in a parameter p_i is usually computed via the partial derivative. Recall from (2.9) on page 20

$$\begin{aligned} J_{Rel} &= [J_{ij}] \\ J_{ij} &= \left.\frac{\partial y_i}{\partial p_j}\right|_{p^*} \frac{p_j^*}{y_i(p^*)}. \end{aligned} \tag{4.1}$$

where p^* is the point at which sensitivities are to be calculated – usually a fitted point estimate. This sensitivity is a relative sensitivity – it is scaled by the size of p_j^* and the size of $y_i(p^*)$ to allow for comparisons between the sensitivities to parameters and outcomes of different sizes [74]. This scaling should not be used in cases where $y_i(p^*)$ is close to zero.

While relative local sensitivities were used as part of the investigation of identifiability prior to parameter fitting, the model outcomes that we are interested in here are different from those previously considered. Previously, sensitivity of the model outcomes

for which data is available was considered. Now, sensitivity will be calculated for other outcomes of interest such as the total size of the HIV-positive population, the fraction of the HIV-positive population who are undiagnosed, and the HIV incidence.

Local sensitivities are straightforward to interpret – they indicate the relative change in the model outcomes caused by a small change in a single parameter near a given point. However, this reliance on a single point also means that they are not easily applied in the case of uncertain parameter values.

4.3.2 Global Sensitivity

Global sensitivity attempts to quantify the sensitivity of model outcomes to parameters over an extended region rather than at a single point. There are several methods available for global sensitivity analysis, but we will consider only one – a statistical method using a sample of the parameter space and partial rank correlation coefficients [16, 93].

The sample from the parameter space required for this method can be the results of a Bayesian fitting routine or a Latin hypercube sample chosen specifically for uncertainty and sensitivity analysis. Using the sample from the parameter space, the model outcomes of interest are computed for each element of the sample as was done for uncertainty analysis. Then the partial rank correlations between the parameters and each of the model outcomes are computed. We used the MATLAB [95] routine `partialcorr` with the type option `'Spearman'` to produce rank correlations.

Rank correlations assess the strength of the monotonic relationship between two quantities while the use of partial correlations also controls for the effects of the other parameters. The result is sensitivity coefficients that are between -1 and 1. A value near 1 indicates that the model outcome y always increases when the parameter p_i increases while a value near -1 indicates that y always decreases when p_i increases. A value near 0 indicates that there is no monotonic relationship between y and p_i – y may increase, decrease or remain fixed when p_i increases. These coefficients do not give any indication of how large an effect changes in parameter values may have. Instead, high values of the correlation coefficients indicate that the association between y and p_i is reliable, occurring for a wide range of parameter values.

Despite the fact that global sensitivities are somewhat more difficult to interpret, they are a useful supplement to local sensitivities indicating whether the behaviour observed using a local sensitivity can be expected to occur over a range of parameter values.

4.3.3 Sensitivity Results

Sensitivities will be calculated for the same quantities that were used in the investigation of uncertainty: Total size of the HIV positive population, fraction of the HIV positive population that is undiagnosed, and annual HIV incidence per 100 000 population. The

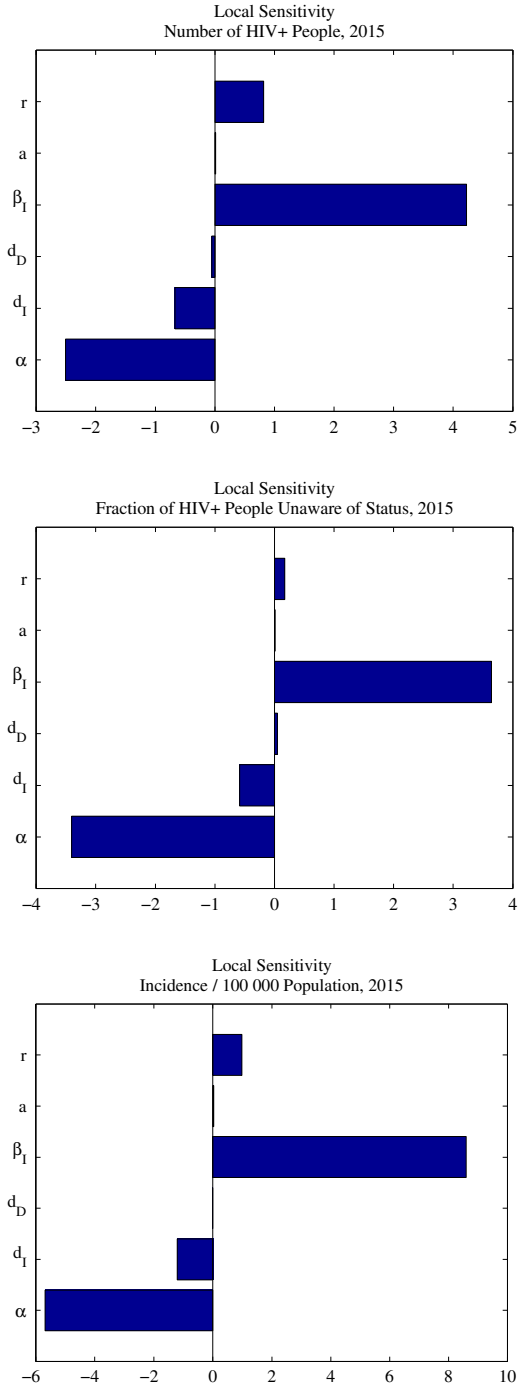


Figure 4.10: Local sensitivity results for the year 2015

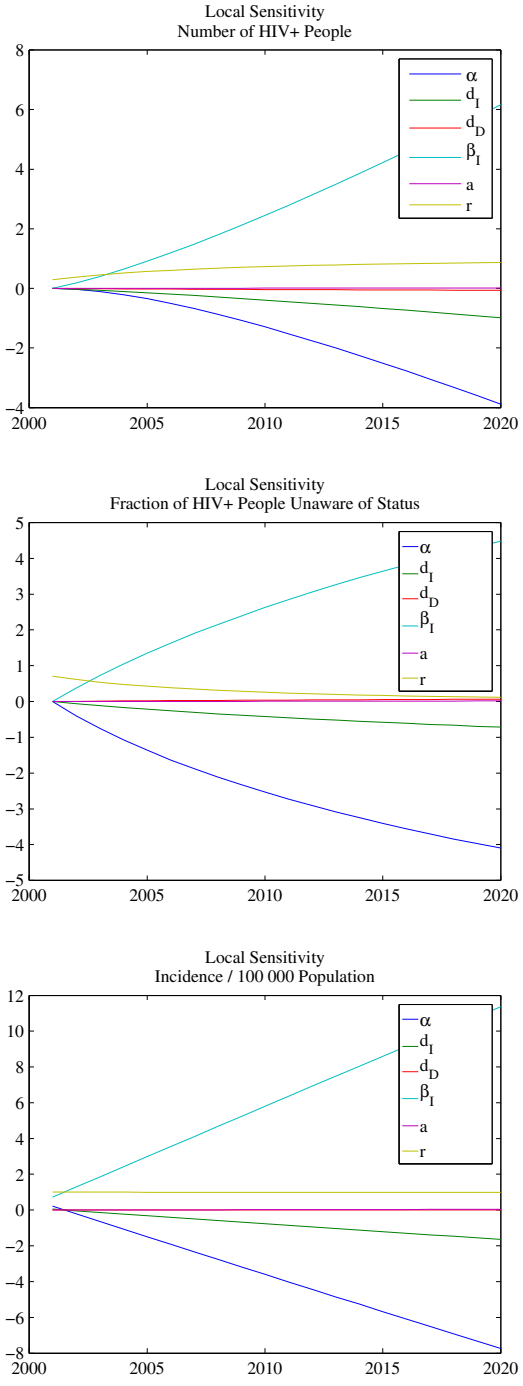


Figure 4.11: Local sensitivity results for the years 2001 to 2020

results for the year 2015 are highlighted in Figures 4.10 and 4.12, while Figures 4.11 and 4.13 illustrate how the sensitivities change over time. Local sensitivities are calculated around the point estimate from the Bayesian fitting routine from Table 2.12. Global sensitivities are calculated using the sample from the parameter space selected during the Bayesian fitting routine.

4.3.3.1 Local Sensitivity Results

The local sensitivity results for the year 2015 are illustrated in Figure 4.10. For all three of the quantities we are interested in, the largest sensitivities are to the parameters β_I and α. The sensitivities to β_I are positive, indicating that an increase in β_I will lead to increases in the total size of the HIV positive population, the fraction unaware of their HIV positive status, and HIV incidence.

HIV incidence is the most sensitive of these quantities to changes in β_I. A local sensitivity of nearly 9 indicates that if β_I is increased by 1%, the HIV incidence will increase by almost 9%. For the same 1% increase in β_I, the fraction unaware of their HIV positive status will increase by 3.6% and the total size of the HIV positive population increases by 4.4%.

Similarly, changes in α have a large impact on the quantities we are interested in. A 1% increase in α results in a 5.5% decrease in incidence, a 3.4% decrease in the fraction unaware of their HIV positive status, and a 2.5% decrease in the total size of the HIV positive population.

Considering the changes in the local sensitivities over time illustrated in Figure 4.11, it can be seen that in the year 2001, the only parameter that has any impact on the total size of the HIV positive population and the fraction that is unaware of their status is r. These two model outcomes have zero sensitivity to the remainder of the parameters as this is the initial time and r is the only parameter that appears in the initial condition. The sensitivities to the other parameters increase as time passes.

4.3.3.2 Global Sensitivity Results

Global sensitivity results indicate that the parameters β_I and α also have the most consistent effects on the model outcomes. Results for the year 2015 are found in Figure 4.12.

The parameter β_I has partial rank correlations of 0.79 with HIV incidence, 0.80 with the fraction unaware of their HIV positive status, and 0.79 with the total size of the HIV positive population. These correlations indicate that an increase in β_I is consistently associated with an increase in these model outcomes across the parameter space.

Similarly, the parameter α has partial rank correlations of -0.86 with HIV incidence, -0.93 with the fraction unaware of their HIV positive status, and -0.83 with the total size of the HIV positive population. These correlations indicate that an increase in α is

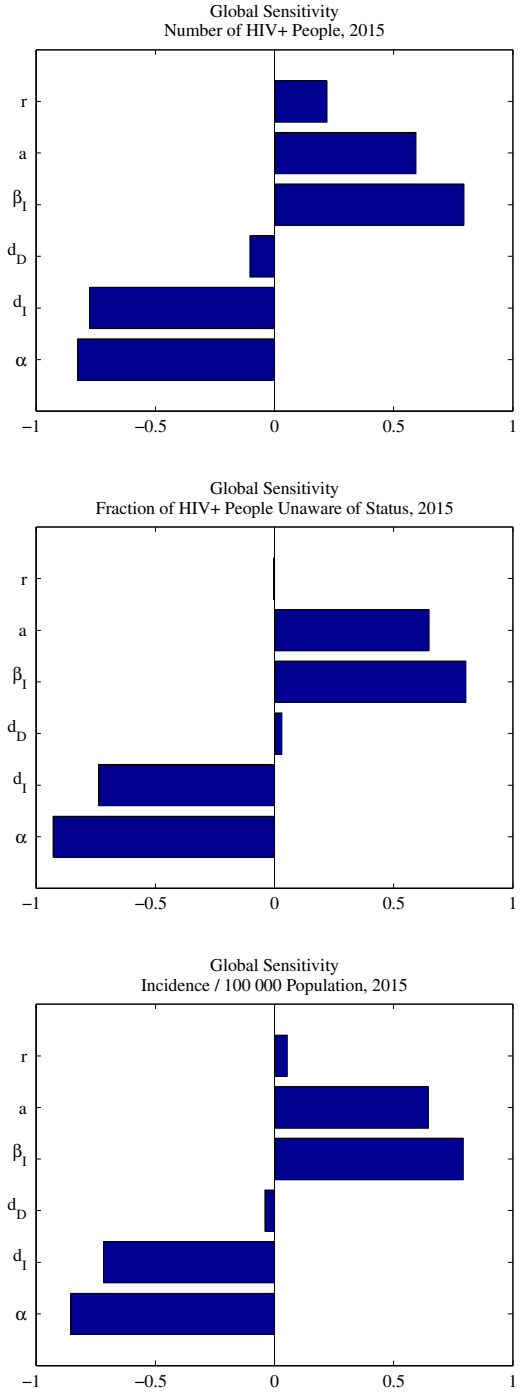


Figure 4.12: Global sensitivity results for the year 2015

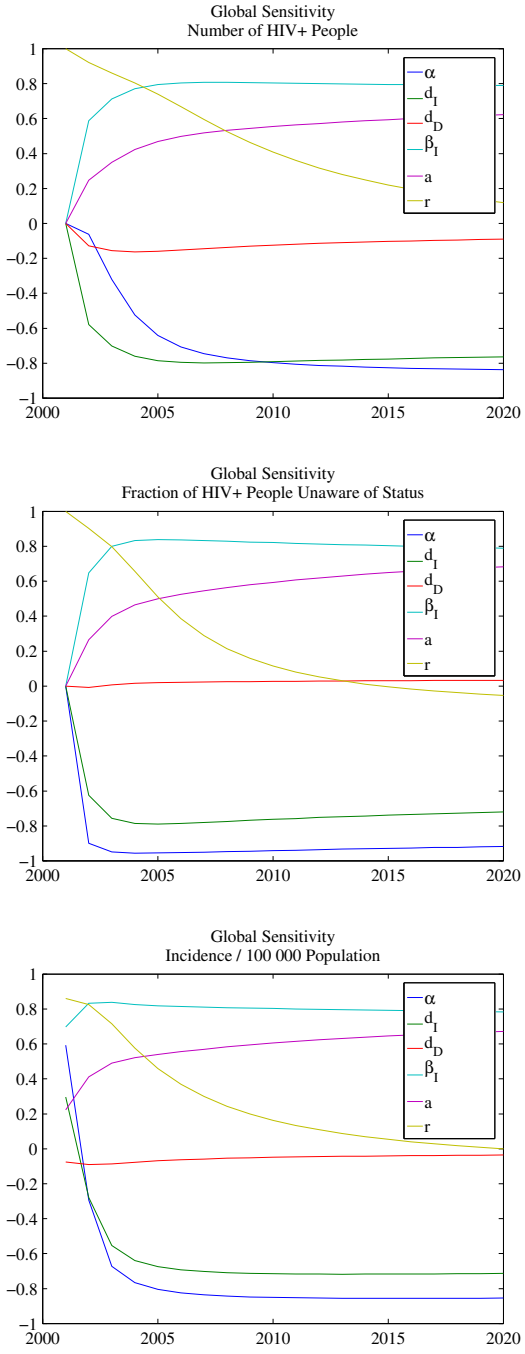


Figure 4.13: Global sensitivity results for the years 2001 to 2020

consistently associated with a decrease in these model outcomes across the parameter space.

The parameters a and d_I also have moderately high partial rank correlations although the local sensitivity to these parameters was small. This indicates that while the effect of these parameters is not large, it is somewhat consistent. An increase in a usually results in an increase in the model outcomes while an increase in d_I usually results in a decrease.

The changes of the global sensitivities over time, displayed in Figure 4.13, show several shifts in importance. For example, all three outcomes begin with a high global sensitivity to the initial condition r and this sensitivity declines over time. For the fraction of the HIV positive population unaware of their status, this sensitivity changes from positive to negative around the year 2013. Recall that r represents the number undiagnosed at the initial time as a fraction of those diagnosed. An increase in this quantity will naturally increase the fraction unaware of their HIV positive status near the initial time. As more of those who where undiagnosed at the initial time become diagnosed, they will begin to increase the denominator of the fraction $\frac{I}{I+D}$ decreasing the overall fraction unaware of their HIV positive status.

4.4 Interventions

One of the advantages of mathematical disease models is the ability to test a variety of interventions. In this section we consider several theoretical intervention programs.

The sensitivity results show that when considering the number of people living with HIV, the fraction of those who are undiagnosed, and the HIV incidence rate, the largest change is produced by varying the diagnosis rate α and the transmission coefficient β_I. Public health programs aimed at increasing testing could increase the parameter α while programs aimed at decreasing risky behaviours could decrease the parameter β_I. In this section, the potential results of implementing such programs will be explored. These hypothetical intervention programs will begin in 2015 and two different scenarios will be investigated: a long term program where parameter values are modified for the entire period from 2015 to 2020, and a short term program where parameter values are modified for only a single year in 2015 before returning to their preintervention values. In both cases large changes are made to parameter values which may not be possible to implement in practice.

As each of these intervention programs is represented mathematically by modifying the value of a single parameter, they do not include some of the potential side effects of these types of programs. For example, programs aimed at increasing diagnosis may have additional effects on the parameters a and d_D as the population that has been diagnosed changes.

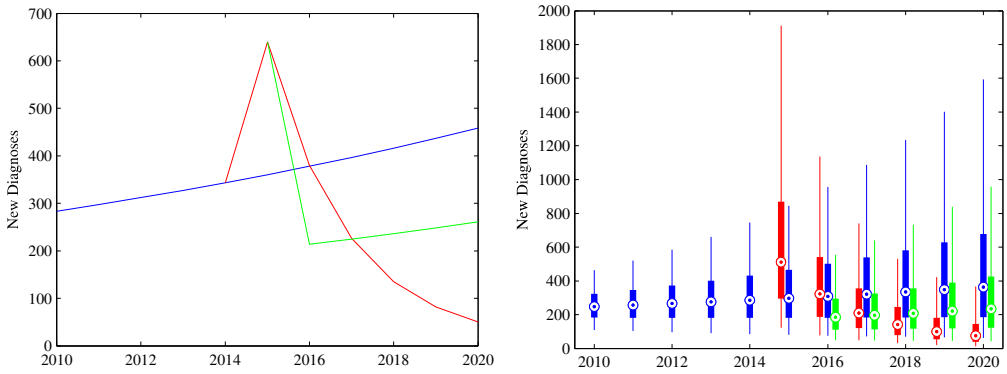


Figure 4.14: The effect on number of new diagnoses of short term (green) or long term (red) interventions to reduce diagnosis delay. The single simulation shown on the left uses the mode value of the parameters as a starting point, while the plot on the right shows the results for the entire uncertainty sample.

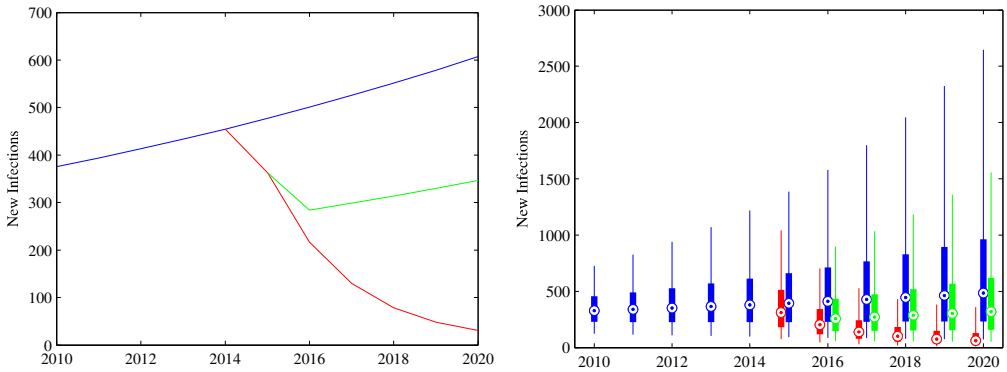


Figure 4.15: The effect on number of new infections of short term (green) or long term (red) interventions to reduce diagnosis delay. The single simulation shown on the left uses the mode value of the parameters as a starting point, while the plot on the right shows the results for the entire uncertainty sample.

4.4.1 Reducing Diagnosis Delay

To investigate the potential impact of reducing diagnosis delay, we consider the impact of reducing the average time from infection to diagnosis to one year. This is accomplished by setting $\alpha^{Int} = 1$. As already mentioned, the intervention begins in 2015 and two cases are considered: a long term program in which the increased diagnosis is maintained through 2020 and a short term program in which α returns to the preintervention value after one year.

For both the short and long term intervention, the result is a temporary spike in new diagnoses . For the short term, program this increase lasts only a single year. Once the program ends, the yearly number of diagnoses falls below the number simulated by the model with no intervention. This occurs because the size of the undiagnosed population has been reduced. For the long term program, the number of new diagnoses

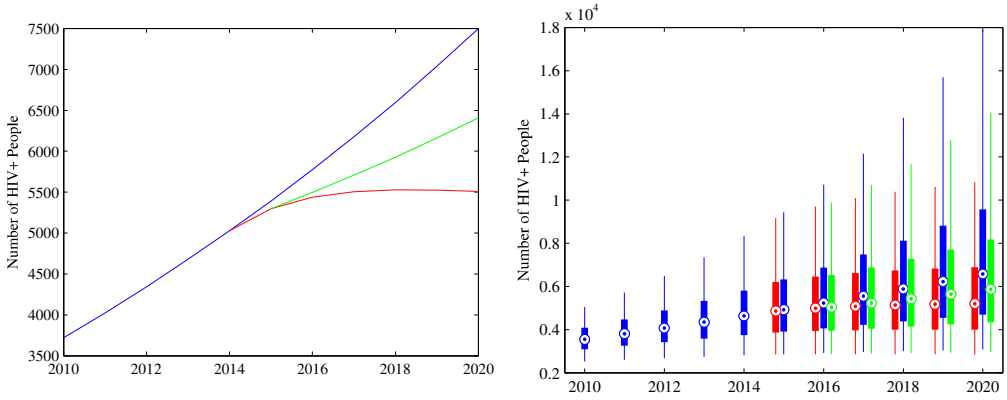


Figure 4.16: The effect on the total number of people living with HIV of short term (green) or long term (red) interventions to reduce diagnosis delay. The single simulation shown on the left uses the mode value of the parameters as a starting point, while the plot on the right shows the results for the entire uncertainty sample.

remains high for slightly longer than a year, but eventually also falls as the undiagnosed population is depleted. The effect of the interventions on number of new diagnoses is illustrated in Figure 4.14.

At the same time, both programs result in reductions in the number of new infections. For the short term program, new infections fall during the intervention and continue to fall for the next year due to the reduction in the undiagnosed population. Within two years of the end of the program, new infections begin to increase slightly. For the long term program, new infections continue to decline resulting in dramatic reductions in the number of new HIV infections by the year 2020. Even the short term program has a lasting impact on the number of new infections. These effects are illustrated in Figure 4.15

This dramatic decrease in new infections results in the size of the HIV-positive population levelling off by 2020 with the long term program. With the short term program the size of the HIV-positive population continues to increase but the rate of increase is slower than predicted with no intervention. Figure 4.16 displays these results.

Figures 4.14, 4.15 and 4.16 also include uncertainty analysis for the effect of the intervention programs. As diagnosis delay is an important source of uncertainty in the model outcomes, specifying the value of α for the duration of the intervention has the additional effect of reducing uncertainty in the outcomes.

4.4.2 Reducing Transmission

We also consider the impact of an intervention which reduces HIV transmission by half. This is accomplished by setting $\beta_I^{Int} = 0.5\beta_I^{NonInt}$. Since the value of β_D is given by $a\beta_I$, this is a reduction in transmission from all sources. As already mentioned the intervention begins in 2015 and two cases are considered: a long term program in which

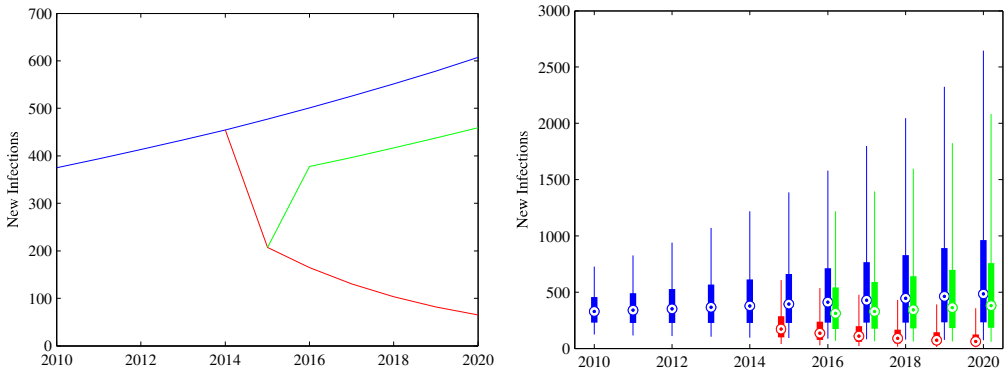


Figure 4.17: The effect on number of new infections of short term (green) or long term (red) interventions to reduce HIV transmission. The single simulation shown on the left uses the mode value of the parameters as a starting point, while the plot on the right shows the results for the entire uncertainty sample.

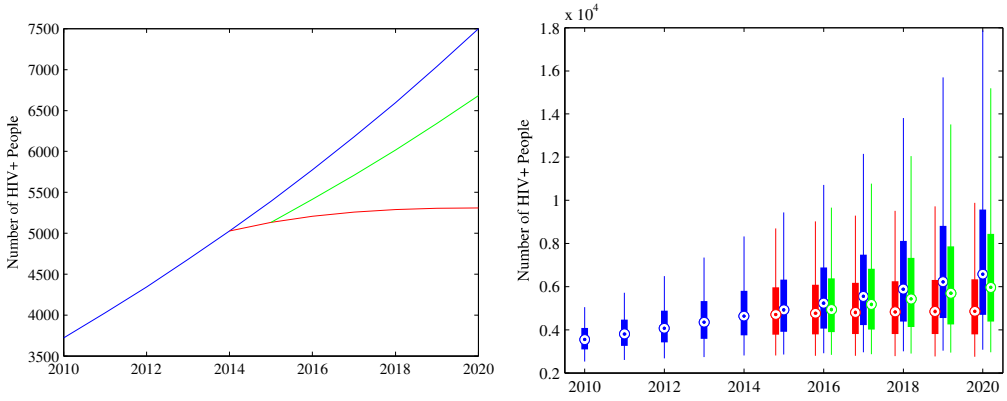


Figure 4.18: The effect on the total number of people living with HIV of short term (green) or long term (red) interventions to reduce HIV transmission. The single simulation shown on the left uses the mode value of the parameters as a starting point, while the plot on the right shows the results for the entire uncertainty sample.

the decreased transmission is maintained through 2020 and a short term program in which β_I returns to the preintervention value after one year.

For both the short and long term programs, the result is a decline in infections in 2015. For the short term, program some of this decline is lost after the end of the program, but the number of new infections remains below the level predicted with no intervention. For the long term program new infections continue to decline resulting in dramatic reductions in the number of new HIV infections by the year 2020. The effects of the intervention on new infections is illustrated in Figure 4.17.

This dramatic decrease in new infections results in the number of total size of the HIV-positive population levelling off almost immediately. With the long term program the size of the HIV-positive population nearly constant after 2015. With the short term program the size of the HIV-positive population resumes increasing after the program

is ended. These results are illustrated in Figure 4.18

Figures 4.17 and 4.18 also provide uncertainty results for the outcome of these intervention programs.

Comparing the results for increased diagnosis intervention to those for the decreased transmission intervention, we see that the two types of interventions have similar impacts on the number of new infections and the number of HIV-positive people if the program is continued over the long term. However, the short term program to increase diagnosis results in a reduction in HIV incidence that persists even after the program has ended. Much of the decrease in HIV incidence observed for the transmission prevention program is lost at the end of the program. This can be seen by comparing the green curves in Figures 4.15 and 4.17.

Chapter 5

Modelling in the Absence of Data

While the main focus of this thesis is HIV modelling using disease surveillance data to compute parameter values, appropriate data is not always available. In this case, model behaviours may be investigated by fixing the values of all parameters manually using values found in medical literature. This procedure bypasses the model fitting step and replaces it with careful interpretation of the model parameters. As quantities reported in medical literature may not correspond exactly with the required parameters, there may also be a number of assumptions which must be made. This process is greatly aided by medical and epidemiological experts who can provide insights and intuition into appropriate parameter values.

Alternatively, it may also be possible to theoretically determine some of the behaviours of the model without specifying parameter values. This is often only possible for simple models and most often provides information about the long term behaviours of the system rather than the short to medium term results that have been considered thus far.

This chapter contains two smaller projects involving modelling that do not rely on parameter fitting from data. A study of transmitted drug resistant HIV strains is found in Section 5.1. For this project, model behaviours are investigated using parameter values found in medical literature. Section 5.2 contains an investigation of long term model behaviour for a simple model without specifying parameter values.

5.1 Drug Resistance

Treatment with combination antiretroviral therapy (cART) reduces viral load and could result in reduced transmission [99]. This effect has been observed on an individual level as the risk of transmission in discordant couples is strongly correlated with viral load [110, 90, 33] and on a community level as increased treatment coverage is associated with a decline in transmission [45, 143, 10].

A recent modelling study considered the impact of universal testing and prompt

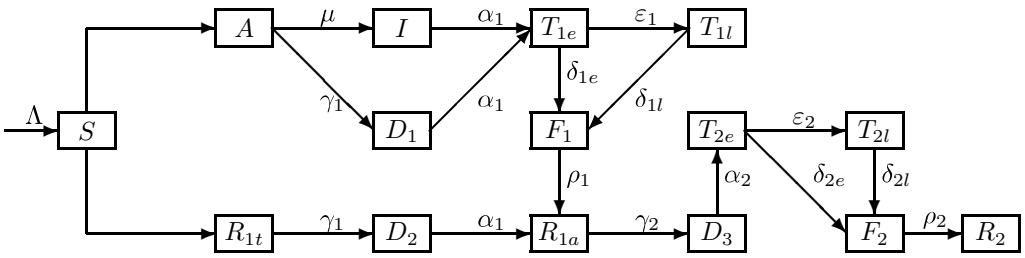


Figure 5.1: Model diagram for an HIV model including both acquired and transmitted drug resistance.

treatment on HIV transmission in South Africa [51]. The model predicted a major impact on transmission including elimination of HIV transmission within 50 years. However, the model considered acquired drug resistance only to note that second line drugs could increase the cost of the program. Transmission of drug resistance was not considered at all.

In this section, a mathematical model is developed to examine HIV transmission under conditions of universal testing and treatment. The model includes treatment failure and the potential for drug resistant strains to develop and be transmitted. Population parameters appropriate for the country of Botswana are used for this model and the results are intended to apply to a context in which HIV is endemic and spread largely through heterosexual contact.

5.1.1 Methods

5.1.1.1 The Model

The model used in this project was developed as part of an interdisciplinary collaboration lead by Dr. S. Houston of the Faculty of Medicine and Department of Public Health and Dr. M. Li of the Department of Mathematical and Statistical Sciences. The model is illustrated in Figure 5.1. Descriptions of the compartments and parameters are given in Tables 5.1 and 5.2 respectively. As in the Province of Alberta model, this is a compartmental disease transmission model described by a system of differential equations. However, due to the complexity of tracking treatment failure and drug resistant viral strains this model is substantially more complicated than the one used for the Province

of Alberta data. The full system of 15 equations is

$$\begin{aligned}
S' &= \Lambda - (\beta_A A + \beta_I I + \beta_{D_1} D_1 + \beta_{T_{1e}} T_{1e} + \beta_{T_{1l}} T_{1l} + \beta_{F_1} F_1) S \\
&\quad - (\beta_{T_{2e}} T_{2e} + \beta_{T_{2l}} T_{2l} + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{R_{1t}} R_{1t} + \beta_{R_{1a}} R_{1a} + \beta_{F_2} F_2) S - d_S S, \\
A' &= (\beta_A A + \beta_I I + \beta_{D_1} D_1 + \beta_{T_{1e}} T_{1e} + \beta_{T_{1l}} T_{1l} + \beta_{F_1} F_1) S \\
&\quad - \mu A - \gamma_1 A - d_A A, \\
R_{1t}' &= (\beta_{T_{2e}} T_{2e} + \beta_{T_{2l}} T_{2l} + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{R_{1t}} R_{1t} + \beta_{R_{1a}} R_{1a} + \beta_{F_2} F_2) S \\
&\quad - \gamma_2 R_{1t} - d_{51} R_{1t}, \\
I' &= \mu A - \alpha_1 I - d_I I, \\
D_1' &= \gamma_1 A - \alpha_1 D_1 - d_{D_1} D_1, \\
D_2' &= \gamma_2 R_{1t} - \alpha_1 D_2 - d_{D_2} D_2, \\
D_3' &= \gamma_3 R_{1a} - \alpha_2 D_3 - d_{D_3} D_3, \\
T_{1e}' &= \alpha_1 I + \alpha_1 D_1 - \delta_1 T_{1e} - \varepsilon_1 T_{1e} - d_{T_{1e}} T_{1e}, \\
T_{1l}' &= \varepsilon_1 T_{1e} - \delta_2 T_{1l} - d_{T_{1l}} T_{1l}, \\
T_{2e}' &= \alpha_2 D_3 - \delta_3 T_{2e} - \varepsilon_2 T_{2e} - d_{T_{2e}} T_{2e}, \\
T_{2l}' &= \varepsilon_2 T_{2e} - \delta_4 T_{2l} - d_{T_{2l}} T_{2l}, \\
F_2' &= \delta_{2e} T_{2e} + \delta_{2l} T_{2l} - d_{F_2} F_2, \\
F_1' &= \delta_{1e} T_{1e} + \delta_{1l} T_{1l} - \rho F_1 - d_{F_1} F_1, \\
R_{1a}' &= \rho_1 F_1 + \alpha_1 D_2 - \gamma_3 R_{1a} - d_{R_{1a}} R_{1a}, \\
R_2' &= \rho_2 F_2 - d_{R_2} R_2.
\end{aligned} \tag{5.1}$$

Those who are infected with a non-resistant viral strain but have never been treated may be either asymptomatic (compartment A) or symptomatic (compartment I). As the parameters chosen correspond to an intensive program of testing and treatment the population in the symptomatic compartment will decline to insignificant levels within the first year modelled. Upon diagnosis, (compartment D_1) those who are HIV-positive immediately receive treatment and move into compartment T_{1e} as the treatment suppresses their viral load. First line treatment is divided into two compartments representing early and late treatment, (compartments T_{1e} and T_{1l}) in order to capture a higher failure rate within the first year of treatment. Treatment failures (compartment F_1) may result in acquired drug resistance (compartment R_{1a}).

At the same time, drug resistance may be transmitted and those who have never been treated but are nonetheless infected with drug resistant strains (compartment R_{1t}). This population also benefits from the intensive testing program. Since the model includes no resistance at the initial time, a symptomatic but undiagnosed compartment is not necessary for the drug resistant population. While this population is diagnosed at the same rate as the non-resistant population, treatment with first line drugs does

Compartment	Description
S	Susceptible Population
Viral Strains	Without Drug Resistance
A	Infected, Treatment Naive, Asymptomatic Population
I	Symptomatic Population
D_1	Diagnosed Population
T_{1e}	Population in Early First Line Treatment
T_{1l}	Population in Late First Line Treatment
F_1	Population with First Line Treatment Failure
Viral Strains	With Drug Resistance
R_{1t}	Infected, Drug Resistant, Treatment Naive Population
D_2	Drug Resistant, Diagnosed Population
R_{1a}	Previously Treated, Drug Resistant Population
D_3	Population Diagnosed as Drug Resistant
T_{2e}	Population in Early Second Line Treatment
T_{2l}	Population in Late Second Line Treatment
F_2	Population with Second Line Treatment Failure
R_2	Population with Resistance to Second Line Treatment

Table 5.1: Description of model compartments in the drug resistance model

not reduce their viral load and they are incorporated into compartment R_{1a} until their resistant status can be diagnosed.

Upon diagnosis of resistance, second line treatments are available. Once again, these have a higher failure rate within the first year of beginning the treatment (compartments T_{2e} and T_{2l}).

As this model is intended to model a context in which resources are limited, it is assumed that upon failure of second line treatments, no further treatments are available.

5.1.1.2 Parameters and Uncertainty

A number of assumptions are imposed on the parameter values. Of particular interest are the transmission coefficients β. The transmission coefficient for the compartment A is used as a baseline and all the other transmission coefficients are expressed in terms of this value. In particular, it is assumed that the resistant virus will not be as transmissible as the non-resistant virus. The coefficient a_R is introduced to capture this reduction. Similarly, it is assumed that diagnosis will have an impact on transmission through changes in behaviour. This effect is captured through the introduction of the coefficient a_B. Finally, it is assumed that those who are effectively treated transmit very little. These relationships are summarized in Table 5.3.

The population parameters Λ and d_S are chosen to ensure that population growth and death rates are appropriate for the country of Botswana. This was done by fixing d_S using estimates of life expectancy and fitting Λ using the population growth curve

from 1981-1991 before any significant impact from HIV. The transmission coefficient β_A was likewise fitted using the model without diagnosis or treatment and HIV prevalence data for Botswana from 1991-2001 when there were few treatment programs available.

The remaining parameter values were determined through reference to a variety of sources. For each parameter, an estimate and a range were chosen. These estimates were used to numerically compute the model results over a 50 year period. The ranges were used to perform an uncertainty analysis on the model. Parameters were assumed to be independent and follow triangular distributions with support given by the range chosen and mode at the point estimate. A Latin hypercube sample was drawn from this parameter distribution and model outcomes of interest were computed at each sampled point.

5.1.2 Results

During the first 5 years of the testing and treatment program, the incidence of HIV falls dramatically. After this time, the incidence levels off at a greatly reduced level where it remains, increasing only slightly over the next several decades. At the same time, the uncertainty increases. The 95th percentile of the sample indicates that for some sampled parameter values, the majority of the initial decrease in incidence is eventually lost. A box plot of the incidence per 100 000 population is found in Figure 5.2.

Figure 5.3 shows the prevalence of drug resistant strains as a fraction of the total HIV-positive population. This includes both acquired and transmitted resistance and is seen to increase steadily. Even the parameter values with the least drug resistance have over half of the HIV-positive population infected with resistant strains after 50 years of the testing and treatment program. Parameter values leading to high resistance may have as much as 90% of the HIV-positive population infected with strains resistant to some treatments.

The roles of parameters a_R, δ_{1e}, and a_B are now considered in more detail. All parameters are set to their baseline estimated values and the parameters a_R, a_B, and δ_{1e} are varied one at a time. In Figures 5.4 and 5.5, three levels are displayed for each of the varied parameters: a low level representing an optimistic scenario, a moderate level, and a high level representing a pessimistic scenario. This procedure of varying parameters one at a time is related to local sensitivity analysis in that both procedures investigate the impact of varying individual parameters separately. While local sensitivity analysis uses a partial derivative to describe the effect of varying a parameter near a single point, the procedure used in this section provides a bigger picture by displaying the results of varying parameters by predetermined amounts. Results are displayed for the incidence of HIV per 100 000 population and for the prevalence of resistance in the HIV-positive population.

Parameter	Symbol	Estimate	Range
Influx of susceptible	Λ	6.4×10^4	$(5.8, 7) \times 10^4$
Life expectancy at age fifteen	$\frac{1}{d_S}$	40	$(35, 50)$ [30]
Life expectancy of I	$\frac{1}{d_I}$	4 years	$(2, 5)$ years [100]
Life expectancy of R_2	$\frac{1}{d_{R_2}}$	2 years	$(1, 4)$ years
Life expectancy of A, R_{1t}	$\frac{1}{d_A}$, $\frac{1}{d_{R_{1t}}}$	5 years less than $\frac{1}{d_S}$	
Life expectancy of D_1, D_2	$\frac{1}{d_{D_1}}$, $\frac{1}{d_{D_2}}$	5 years less than $\frac{1}{d_S}$	
Life expectancy of T_{1e}, T_{1l}	$\frac{1}{d_{T_{1e}}}$, $\frac{1}{d_{T_{1l}}}$	6 years less than $\frac{1}{d_S}$	
Life expectancy of T_{2e}, T_{2l}	$\frac{1}{d_{T_{2e}}}$, $\frac{1}{d_{T_{2l}}}$	8 years less than $\frac{1}{d_S}$	
First-line failure rate (≤ 1 year)	δ_{1e}	10%	$(3\%, 30\%)$ [10, 56]
First-line failure rate (> 1 year)	δ_{1l}	5%	$(1.5\%, 15\%)$ [10, 56]
Second-line failure rate (≤ 1 year)	δ_{2e}	20%	$(6\%, 40\%)$
Second-line failure rate (> 1 year)	δ_{2l}	7.5%	$(5\%, 20\%)$
Rate of acquired resistance among failures	ρ_1	50%	$(30\%, 70\%)$ [10, 104, 136, 67]
Annual diagnosis rate	γ_1	100%	Assumed [51]
Diagnosis time of resistance	$\frac{1}{\gamma_2}$	2.5 years	$(1, 5)$ years
HIV progression	$\frac{1}{\mu}$	10 years	$(5, 15)$ years [100]
Length of treatment	$\frac{1}{\alpha_1}$	0.167 year	$(0.125, 0.2)$ year
Early stage for ART treatments	$\frac{1}{\epsilon_1}$, $\frac{1}{\epsilon_2}$	1 year	Assumed
Disease progression after failure	$\frac{1}{\rho_2}$	8 years	$(5, 10)$ years
Impact of behavioral changes	a_B	75%	$(30\%, 100\%)$ [26, 42]
Fitness of resistant strains	a_R	30%	$(10\%, 50\%)$ [34, 128, 37, 86]

Table 5.2: Parameter values for the drug resistance model

Compartment	Transmission Coefficients
T_{1e}, T_{1l}	$\beta_{T_{1e}} = \beta_{T_{1l}} = 0.001\beta_A$
T_{2e}, T_{2l}	$\beta_{T_{2e}} = \beta_{T_{2l}} = 0.001\beta_A$
R_{1t}	$\beta_{R_{1t}} = a_R\beta_A$
I, D_1, F_1	$\beta_I = \beta_{D_1} = \beta_{F_1} = a_B\beta_A$
D_2, R_{1a}	$\beta_{D_2} = \beta_{R_{1a}} = a_B a_R\beta_A$
F_2, R_2, D_3	$\beta_{F_2} = \beta_{R_2} = \beta_{D_3} = a_B a_R\beta_A$

Table 5.3: Relationships between the transmission coefficiencts for the drug resistance model.

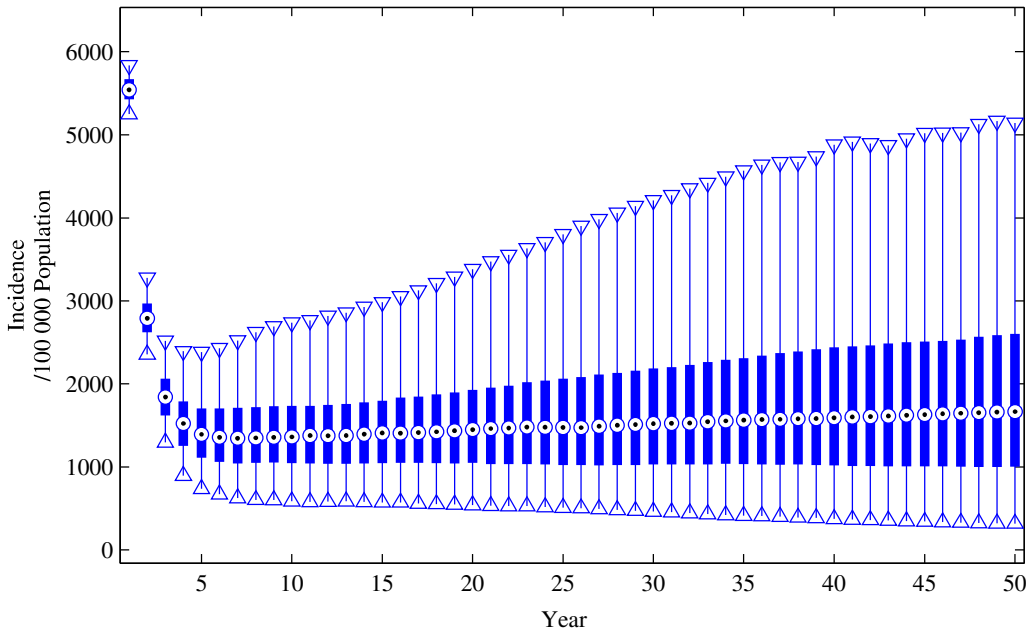


Figure 5.2: Boxplot of HIV incidence simulated by the drug resistance model.

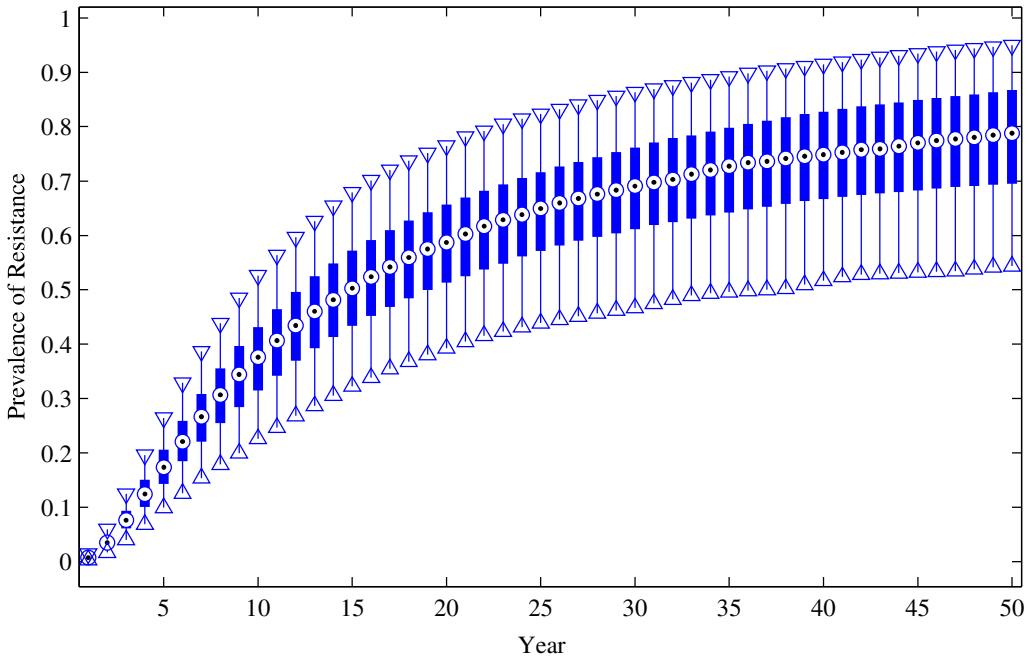


Figure 5.3: Boxplot of simulaton results for the prevalence of drug resistant strains as a fraction of the total HIV-positive population.

Transmitted Resistance: The parameter a_R, representing the relative transmissibility of the resistant strain, is one of the most important parameters for overall HIV incidence. In Figure 5.4 A, B, and C, three levels of this parameter are considered. In the optimistic case, a relative transmissibility of 0.1 indicates that the resistant virus is only 10% as likely to be transmitted as the non-resistant strain. In this case the incidence is seen to level off at around 750 cases per 100 000 population with transmitted resistant strains only a small part of that number. For the moderate case, a relative transmissibility of 0.3 is used, resulting in incidence levelling off at around 1000 cases per 100 000 population. In this case, transmitted resistance eventually rises to become over half of all new cases. In the pessimistic case, a relative transmissibility of 0.5 is used indicating that the drug resistant strain is half as transmissible as the non-resistant strain. In this case HIV incidence initially drops to around 1000 cases per 100 000 population, but subsequently increases again to over 2000 cases per 100 000 population. Much of this increase is driven by an increase in transmitted drug resistant strains.

The impact of the parameter a_R on the prevalence of the resistant strain in the HIV-positive population is illustrated in Figure 5.5 Parts A, B, and C. The optimistic value of the parameter results in a fraction of the HIV-positive population with drug resistance of under 0.5 for the entire modelled duration. As indicated by the incidence results, much of this fraction will be due to treatment failures rather than transmission of resistant strains. The moderate value of a_R results in the fraction of the HIV-positive population

with drug resistance reaching 0.5 after about 20 years and continuing to increase slowly. The pessimistic value results in the fraction of the HIV-positive population with drug resistance continuing to increase steadily and reaching about 0.9 after 50 years. The incidence results previously discussed suggest that much of this fraction will be due to transmission of drug resistant strains.

Treatment Failure: The parameter δ_{1e} representing the rate of treatment failure during early first line treatment was considered at the optimistic level of 0, the moderate level of 0.15, and the pessimistic level of 0.3. The results for HIV incidence and fraction with resistant virus are found in Figures 5.4 and 5.5 Parts D, E, and F. As late first line treatment was constrained to have a failure rate of half that of early first line treatment, the optimistic case resulted in no failure of first line treatment. This was the only case in which HIV incidence could be completely eliminated with no resistance developing. In the moderate and pessimistic scenarios, treatment failure has only a small impact on the HIV incidence after 50 years, but did have an impact on how quickly incidence declined. Similarly, treatment failure has only a small impact on the fraction of the HIV-positive population with drug resistance after 50 years, but does effect the rate at which this fraction increases.

Behavioural Change: The final parameter for which several scenarios are considered is a_B. This parameter captures changes in behaviour due to diagnosis. The optimistic scenario assumes that transmission can be reduced by 50% for those who are aware of their status. The moderate scenario uses a transmission reduction of 25% while the pessimistic scenario considers the possibility that those who are diagnosed will not make any changes their behaviour to reduce transmission. The incidence results for these three scenarios are found in Figure 5.4 Parts G, H, and I. The optimistic scenario results in an HIV incidence that continues to decline slowly for the entire time computed by the model and results in an incidence of about 500 per 100 000 after 50 years. The moderate scenario does not see this continued decline. Instead the HIV incidence levels off at around 1000 per 100 000 after about 5 years and remain at this level for the remainder of the time modelled. The pessimistic scenario of no behavioural changes results in the HIV incidence beginning to increase slightly after 5 years and produces an incidence of around 1750 per 100 000 population after 50 years.

Behavioural changes have very little impact on the fraction of the HIV-positive population with drug resistance. This is because behavioural changes have no impact on the probability of acquiring drug resistance through treatment failures, and although changes in behaviour can reduce HIV incidence, drug resistant and non-drug resistant viral types are affected equally. This is illustrated in Figure 5.5 Parts G, H, and I.

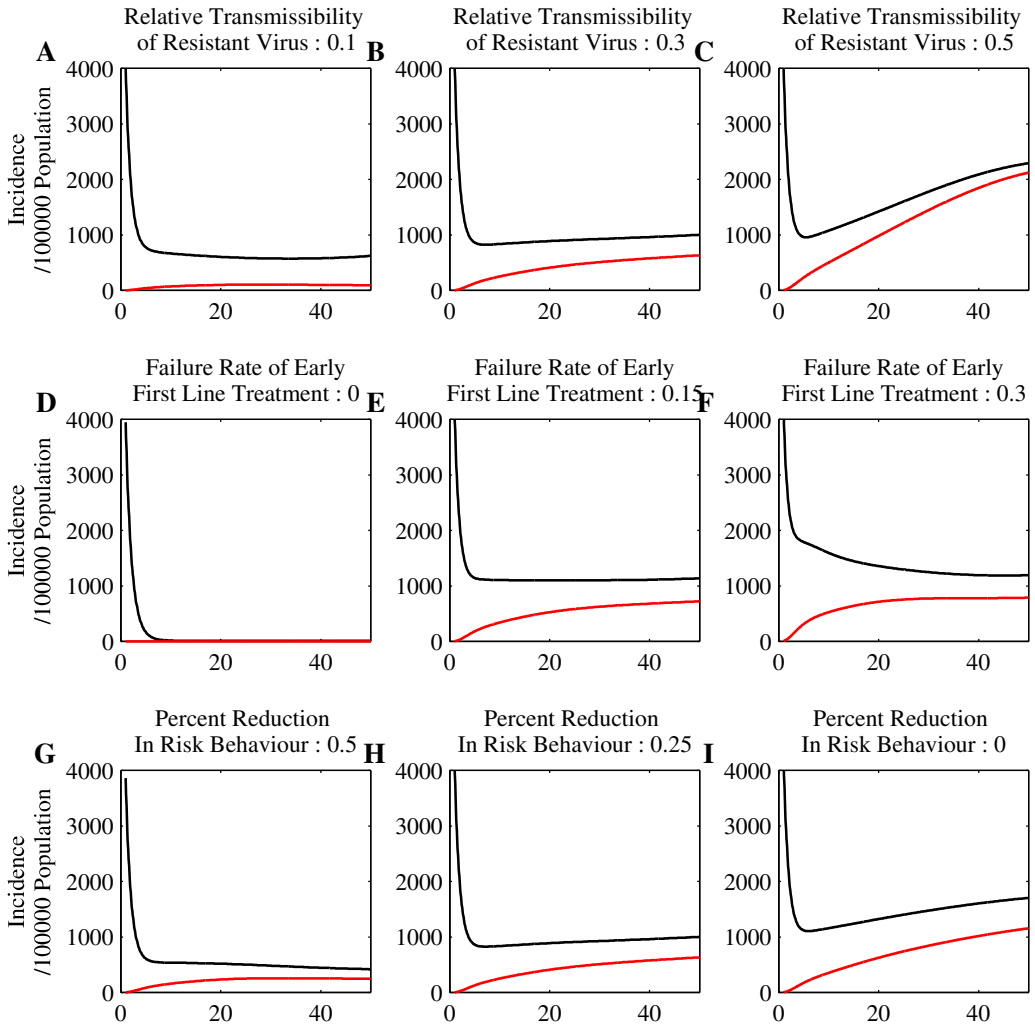


Figure 5.4: The impact of a variety of factors on the incidence of HIV (black) and the incidence of drug resistant HIV strains (red). In Parts A, B, and C the transmissibility of the drug resistant strain, a_R, is varied. In Parts D, E, and F the treatment failure rate, d_{1e}, for first line treatment is varied. In parts G, H, and I the reduction in risk behaviour, a_B, is varied.

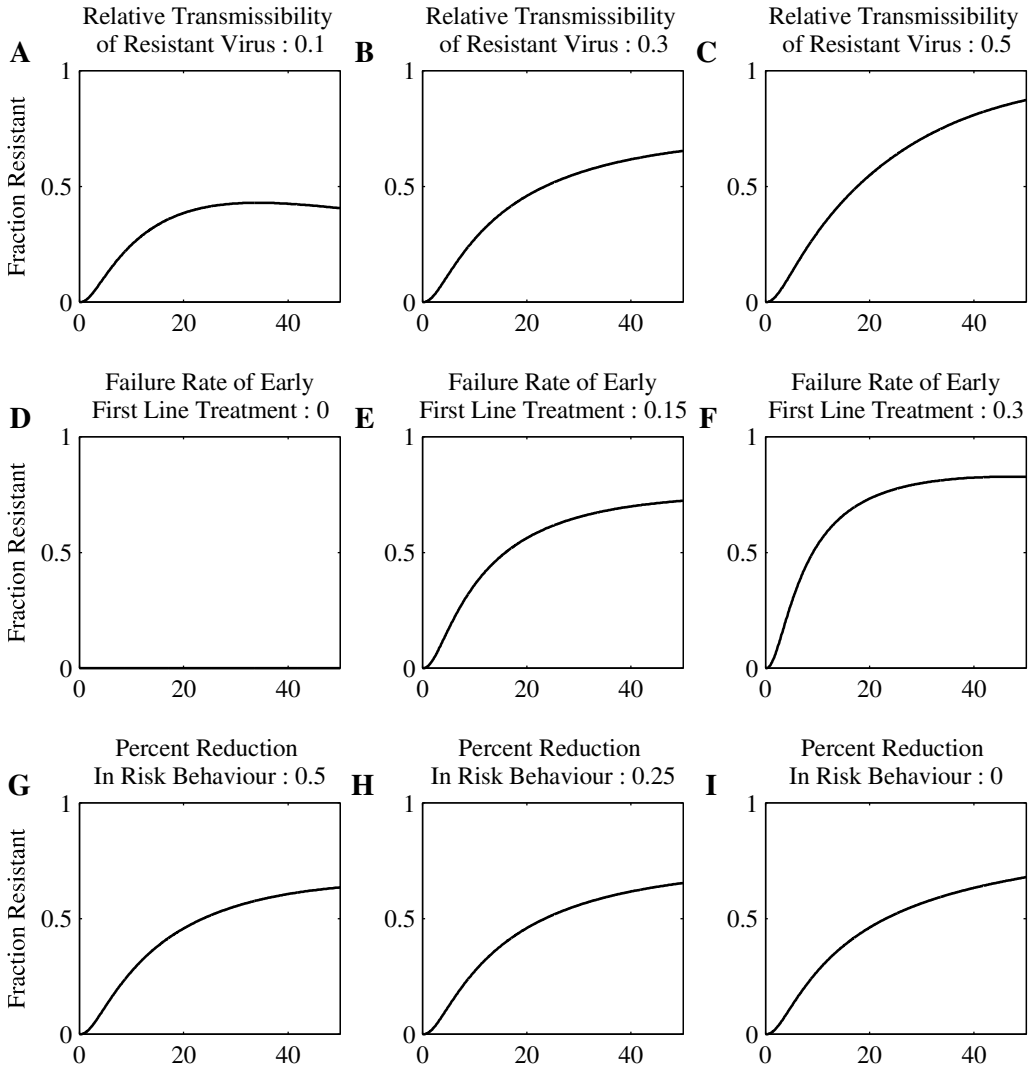


Figure 5.5: The impact of a variety of factors on the fraction of HIV cases with drug resistance. In Parts A, B, and C the transmissibility of the drug resistant strain, a_R, is varied. In Parts D, E, and F the treatment failure rate, d_{1e}, for first line treatment is varied. In parts G, H, and I the reduction in risk behaviour, a_B, is varied.

5.1.3 Conclusion

The results of this model suggest that a program of universal testing and prompt treatment could have a significant impact on HIV incidence in a context where the virus is endemic. However, the model results highlight the potential concern arising from acquired and transmitted drug resistance. The increase in resistance that is found in all cases indicates that universal testing and treatment programs should include provision for management of both acquired and transmitted drug resistant viral strains.

5.2 Population Dependent Transmission

This section has been previously published [36] and appears here with only minor changes. The mathematical analysis and the majority of the manuscript composition for this section are my own work. My supervisor, Dr. M. Li, assisted with problem formulation and some aspects of manuscript composition.

In this section a simple mathematical disease model is used to investigate the effects of population dependent transmission on long term model behaviour. The model used in this section bears some resemblance to the model we have used Chapters 2 through 4 to describe HIV transmission in the Province of Alberta, however, there are important differences as well and the results in this section are not directly applicable to the model used previously. In particular, rather than having a diagnosed population D the model described in this section has a recovered population R which never transmits the disease. Secondly, the HIV model in the previous chapters uses a standard proportionate term to model disease transmission while in this section we consider the potential effects of including other types population dependence in disease transmission. Population dependent transmission may be thought of as capturing social or behavioural changes that may occur due to changes in the total population size.

It should be noted that the use of the word *incidence* differs slightly in this section from the definition most commonly used in epidemiology and in the other chapters of this thesis. In both cases, incidence refers to new infections. In other sections of this thesis, we have taken incidence to mean the number of new cases in some fixed time period (usually one year). Incidence is often reported as the number of new cases per population in one year. In this section we take a more mathematical perspective and use the word *incidence* to refer to an *instantaneous* rate of disease transmission.

5.2.1 Introduction

For infectious diseases, incidence describes the rate at which new infections occur. To model the transmission dynamics of an infectious disease, accurately describing and estimating disease incidence are of utmost importance. In classical epidemic models such as those of Kermack-McKendric [21, 80], infection occurs through horizontal transmission,

and the disease incidence is customarily modelled by a bilinear form, $\beta I(t)S(t)$, which is in proportion to the size or density of the sub-population of susceptible hosts $S(t)$ and that of the sub-population of infectious hosts $I(t)$. The parameter β represents the transmission coefficient and is dependent on the frequency of host-host contact and the probability of a contact being infectious [2]. For infections that spread through sexual contacts, as in HIV and other STD transmissions, the incidence is typically modelled by a proportionate form $\beta\frac{I(t)S(t)}{N(t)}$, where $N(t)$ is the total host population and the constant β describes the effective contact rate among hosts [21]. When the host population is predominantly susceptible or infectious, saturated incidence forms, $\frac{\beta I(t)S(t)}{a+S(t)}$ or $\frac{\beta S(t)I(t)}{b+I(t)}$ respectively, have been used to account for the saturation of contacts [21, 125]. Nonlinear incidence of the form $\beta I(t)^p S(t)^q$ is used in [89] and shown to lead to complicated behaviours such as multiple endemic equilibria and existence of periodic oscillations. Other general forms of incidence terms have been used and further studied in more recent work [87, 147].

Among these common incidence forms, the proportionate incidence $\beta\frac{I(t)S(t)}{N(t)}$ depends explicitly on the total population size $N(t)$, and is said to be density dependent. If $N(t)$ is a constant, it can be combined with β so that the proportionate incidence is equivalent to the bilinear one. However, due to density dependence, the proportionate incidence and the bilinear incidence differ when $N(t)$ varies with time. More general forms of density dependence can be incorporated into the incidence term as $\beta I(t)S(t)f(N(t))$, for certain classes of functions f. A typical example is $f(N) = N^{-\alpha},\ \alpha \geq 0$. When $\alpha = 0$, we obtain the bilinear incidence, and when $\alpha = 1$, we arrive at the proportionate incidence. It is shown in [54, 55] that, if $0 \leq \alpha \leq 1$, the incidence form $\beta\frac{I(t)S(t)}{N(t)^\alpha}$ typically leads to standard threshold behaviour in simple epidemic models: the disease dies out if the basic reproduction number $R_0 < 1$ and the disease becomes endemic and persists at a unique endemic equilibrium when $R_0 > 1$. In Section 5.2, we show that if $f(N)$ takes more general and yet biologically plausible forms, density-dependent incidence $\beta I(t)S(t)f(N(t))$ can lead to complicated dynamics even in the simplest SIR models.

To introduce and interpret density dependence in disease incidence, we first recall that the disease incidence can be calculated as

$$\lambda(N(t))\frac{S(t)}{N(t)}I(t).$$

In this expression, the contact rate, $\lambda(N)$, is the average number of effective contacts made by a single infectious individual in one unit of time. Such a contact is made with a susceptible individual, and therefore produces an new infection, with probability $\frac{S(t)}{N(t)}$. Multiplied by the total number $I(t)$ of infectious hosts, the expression gives the total number of new infections per unit time. Density dependence will be introduced through $\lambda(N)$. For bilinear incidence, we have $\lambda(N) = \beta N$, which assumes the number of contacts is linearly proportional to the size of the population. This may be plausible

for populations of large urban centres where, because of limited living space, an increase in population size is likely to increase population density and the frequency of contact. In proportionate incidence, we have that $\lambda(N) = \beta$ and is independent of population size N. This is plausible for populations in rural areas where an increase in population size does not necessarily increase population density and the frequency of contact. In a more general incidence $\beta I(t)S(t)f(N(t))$, we have the contact rate equal to $\lambda(N) = \beta N f(N)$.

The first class of $f(N)$ considered is $f(N) = BN^{-\alpha},\ \alpha > 1$. In this case, the contact rate $\lambda(N) = \beta B N^{1-\alpha}$. When $\alpha > 1$, an increase in N leads to a decrease of contact. This is plausible for diseases which require a significant level of contact to be transmitted and a culture where people in large urban centres tend to have less contact with their neighbours when population density increases. We demonstrate that, with this incidence form, backward bifurcations can occur in a simple SIR model, namely, multiple endemic equilibria exist when $R_0 < 1$. Backward bifurcations have been investigated in a variety epidemic models and are known to lead to catastrophic effects in terms of disease control [4, 15, 20, 39, 52, 58, 57, 72, 82, 94, 109, 116, 131, 138]; when backward bifurcation occurs, the outcome of a disease outbreak not only depends on the parameter values, but also critically depends on the initial conditions. Some known biological mechanisms that may lead to backward bifurcations are imperfect immunity [52], a vaccine that provides only partial protection [4, 82], and behavioural responses to perceived disease risk [58]. In many of these models, backward bifurcation occurs due to asymmetry among different contact groups or multiple routes for transmission. Our result shows that backward bifurcation can result solely from density dependence in the incidence form.

A second class of $f(N)$ investigated is $f(N) = aN^2 + bN + c,\ a > 0$. In this case, the contact rate $\lambda(N) = \beta N\left(aN^2 + bN + c\right)$ is a cubic function that increases for small or large values of N, and may decrease for intermediate values of N. Such a region of decrease can be the result of adjustment of social behaviours as population size increases. For this class of f, we show that multiple endemic equilibria can exist when $R_0 > 1$. Furthermore, one of the endemic equilibria can undergo stability change, and a Hopf bifurcation occurs, producing stable periodic oscillations. Such dynamical outcomes have been observed in SEIR models of constant total population with nonlinear incidence [89]. Our result shows that the same phenomenon can also explained through density dependence in the disease incidence.

We also consider the effect of adding multiple infectious stages to the model with these classes of $f(N)$. Our result shows that endemic equilibria for the multi-stage model can be identified using the same analysis that was used in the single stage model. Furthermore, we determine that the existence of a backward bifurcation depends on the number of infectious stages as well as the model parameters.

Incidence forms that incorporate social behaviour changes have been investigated in epidemic models using nonlinear incidence forms [58, 116]. A piecewise incidence function incorporating a sharp change in social behaviours is studied in [5] and shown

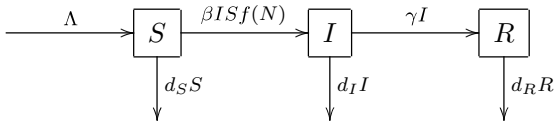


Figure 5.6: The transfer diagram for model (5.2).

to lead to periodic behaviour. While incidence forms in our study may be considered as social behaviours related, they are different from those in previous studies in that we incorporate social behaviours through nonlinear density dependence, dependence on the total population size N, rather than explicit nonlinear dependence on I or S.

5.2.2 The Model and Preliminaries

We consider a simple SIR epidemic model with population dependant incidence. The transfer diagram is depicted in Figure 5.6. Here, S is the susceptible population, I is the infectious population, and R is the recovered or removed population. The total population $N = S + I + R$. The model is described by a system of differential equations:

$$\begin{aligned} \dot{S} &= \Lambda - \beta ISf(N) - d_S S \\ \dot{I} &= \beta ISf(N) - \gamma I - d_I I \\ \dot{R} &= \gamma I - d_R R. \end{aligned} \tag{5.2}$$

The average per capita contact rate is given by $\lambda(N) = \beta N f(N)$. The parameter Λ indicates the influx of susceptibles, and γ denotes the rate constant for recovery. The mean infectious period is given by $\frac{1}{\gamma}$. We assume that the disease can be fatal and the death rate d_I for the I compartment may contain both natural and disease-related death. That is, we assume that the death rates d_S, d_I and d_R satisfy

$$0 < d_S \leq d_R, \quad \text{and} \quad d_S < d_I. \tag{5.3}$$

This assumption is sufficient for the total population N to be time dependent. In contrast, if $d_S = d_I = d_R = d$, then $N'(t) = \Lambda - dN(t)$, and $N(t) \to \Lambda/d$ as $t \to \infty$ and model (5.2) can be replaced by a limiting system with $N = \Lambda/d$.

It can be verified that solutions to model (5.2) with nonnegative initial conditions remain nonnegative and bounded for all $t \geq 0$. Furthermore, the compact region

$$\Gamma = \{(S, I, R) \in \mathbb{R}^3_+ \mid S + I + R \leq \Lambda/d_S\}$$

is positively invariant for model (5.2) and globally attracting. It suffices to investigate the global dynamics of model (5.2) in Γ.

For all nonnegative parameter values, the model has a disease-free equilibrium $P_0 = (\bar{S}, 0, 0)$, with $\bar{S} = \Lambda/d_S$. The Jacobian matrix of of model (5.2) at P_0 is given by

$$\begin{bmatrix} -d_S & -\beta\bar{S}f(\bar{S}) & 0 \\ 0 & \beta\bar{S}f(\bar{S}) - (d_I + \gamma) & 0 \\ 0 & \gamma & -d_R \end{bmatrix}$$

and has eigenvalues $\lambda_1 = -d_S < 0$, $\lambda_2 = -d_R < 0$, and $\lambda_3 = \beta\bar{S}f(\bar{S}) - (d_I + \gamma)$. Therefore, the disease free equilibrium P_0 is asymptotically stable if and only if $\beta\bar{S}f(\bar{S}) < d_I + \gamma$. Let

$$R_0 = \frac{\beta}{d_I + \gamma}\bar{S}f(\bar{S}). \tag{5.4}$$

This is the basic reproduction number for model (5.2) [2, 65]. When $f(N) = 1$, $R_0 = \frac{\beta}{d_I+\gamma}\bar{S}$, which agrees with the basic reproduction number for SIR models with bilinear incidence. When $f(N) = 1/N$, $R_0 = \frac{\beta}{d_I+\gamma}$, which agrees with the basic reproduction number of SIR model with proportionate incidence. Note that $R_0 < 1$ if and only if $\beta\bar{S}f(\bar{S}) < d_I + \gamma$. The following threshold result is standard.

Proposition 1. *If $R_0 < 1$, then the disease-free equilibrium P_0 is asymptotically stable. If $R_0 > 1$, then P_0 is unstable, model* (5.2) *is uniformly persistent, and an endemic equilibrium exists in the interior of* Γ.

Endemic equilibria, (S^*, I^*, R^*) with $I^* > 0$, of model (5.2) are determined by

$$\begin{aligned} 0 =& \Lambda - \beta I^* S^* f(N^*) - d_S S^* \\ 0 =& \beta S^* f(N^*) - \gamma - d_I \\ 0 =& \gamma I^* - d_R R^*. \end{aligned} \tag{5.5}$$

From the second equation in (5.5) we obtain

$$\beta S^* f(N^*) = \gamma + d_I. \tag{5.6}$$

Denoting

$$p = \frac{1}{d_I + \gamma} + \frac{1}{d_R}\frac{\gamma}{d_I + \gamma} \quad \text{and} \quad \sigma = \frac{\beta}{d_I + \gamma} \tag{5.7}$$

and using the remaining equations in (5.5), we obtain

$$N^* = (1 - pd_S)S^* + p\Lambda. \tag{5.8}$$

The parameter p is the mean life expectancy of those who become infected. The parameter σ is a measure of total transmissibility over the mean infectious period $\frac{1}{d_I+\gamma}$. In the case where $f(N) = \frac{1}{N}$, σ is the average number of effective contacts made by an infective in its mean infectious period, in other words the *contact number* [65]. Assumption (5.3) implies that $pd_S < 1$.

Define

$$g(S) = Sf(N(S)) \tag{5.9}$$

with

$$N(S) = (1 - pd_S)\, S + p\Lambda.$$

Then, from equation (5.6), an equilibrium S^* must satisfy

$$g(S^*) = \frac{1}{\sigma}, \quad S^* \in (0, \bar{S}]. \tag{5.10}$$

In the following two sections, we will show that for different classes of $f(N)$, it is possible for model (5.2) to have multiple endemic equilibria, which in turn can lead to complicated dynamics.

5.2.3 Backward Bifurcations

In this section, we assume that $f(N) = BN^{-\alpha}$. For $\alpha \leq 1$, it is known that the traditional threshold result holds for model (5.2) [54, 55]: if $R_0 \leq 1$ then the disease-free equilibrium P_0 is globally stable; if $R_0 > 1$, then a unique endemic equilibrium exists and is globally stable in the interior of the feasible region.

In the rest of the section, we assume that $\alpha > 1$. The function $g(S)$ defined in (5.9) is written as

$$g(S) = \frac{BS}{[(1 - pd_S)S + p\Lambda]^{\alpha}}. \tag{5.11}$$

Note that $g(0) = 0$, $g(\bar{S}) = \left(\frac{\Lambda}{d_S}\right)^{1-\alpha}$, and $g(S)$ is continuous and positive for $S > 0$. Differentiating $g(S)$ in (5.11) gives

$$g'(S) = \frac{B\left[S(1 - pd_S)(1 - \alpha) + p\Lambda\right]}{[(1 - pd_S)S + p\Lambda]^{\alpha+1}},$$

and $g(S)$ has a single critical point

$$S_{crit} = \frac{p\Lambda}{(1 - pd_S)(\alpha - 1)}.$$

When $\alpha \leq \frac{1}{1-pd_S}$, $g(S)$ is monotone in the interval $(0, \bar{S})$, and multiple endemic equilibria can not occur. When $\frac{1}{1-pd_S} < \alpha$, $S_{crit} \in (0, \bar{S})$, and equation (5.10) can have two solutions in $(0, \bar{S})$.

The value of g at the critical point is given by

$$g(S_{crit}) = \frac{g(\bar{S})}{(1 - pd_S)(pd_S)^{\alpha-1}} \frac{(\alpha - 1)^{\alpha-1}}{\alpha^{\alpha}}. \tag{5.12}$$

Since the function $h(x) = (1 - x)(x)^{\alpha-1}$ has its maximum value of $\frac{(\alpha-1)^{\alpha-1}}{\alpha^{\alpha}}$ attained at

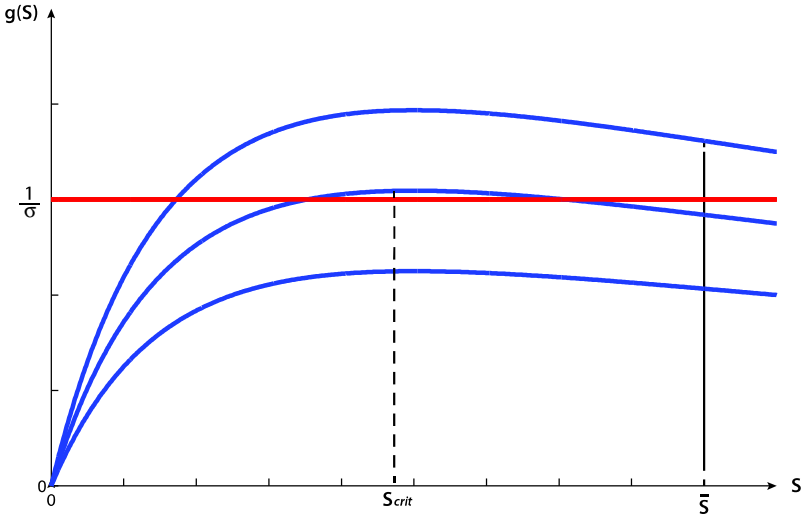


Figure 5.7: Possible solutions of equation $g(S) = \frac{1}{\sigma}$ with $f(N) = \frac{1}{N^\alpha}$.

$\frac{1}{1-x} = \alpha$, we see that, if $\frac{1}{1-pd_S} < \alpha$, then

$$(1 - pd_S)(pd_S)^{\alpha-1} < \frac{(\alpha-1)^{\alpha-1}}{\alpha^\alpha}. \tag{5.13}$$

As a consequence, $g(S_{crit}) > g(\bar{S})$ and $g(S_{crit})$ is a maximum. We can draw conclusions about the number of solutions of (5.10) based on the location of $g(S_{crit})$ and $g(\bar{S})$ relative to $\frac{1}{\sigma}$, as summarized in the next proposition. Figure 5.7 illustrates the possibilities.

Proposition 2. *Let $f(N) = BN^{-\alpha}$ and assume that $\alpha > \frac{1}{1-pd_S}$. Then system (5.2) has no endemic equilibria if $g(S_{crit}) < \frac{1}{\sigma}$, exactly one endemic equilibria if $g(\bar{S}) > \frac{1}{\sigma}$, and two endemic equilibria if $g(\bar{S}) < \frac{1}{\sigma} < g(S_{crit})$.*

We note that $g(\bar{S}) = \bar{S}f(\bar{S})$, and thus $\sigma g(\bar{S}) = 1$ if and only if $R_0 = 1$. Similarly, from (5.12) we know $\sigma g(S_{crit}) = 1$ if and only if R_0 agrees with

$$\overline{R}_0 := \frac{\alpha^\alpha}{(\alpha-1)^{\alpha-1}}(1 - pd_S)(pd_S)^{\alpha-1}. \tag{5.14}$$

We see from (5.13) that $\overline{R}_0 < 1$.

Proposition 2 implies existence of two endemic equilibria when R_0 is in the range $\overline{R}_0 < R_0 < 1$. When this happens, system (5.2) is said to undergo a *backward bifurcation* at $R_0 = 1$ [72]. To complete the bifurcation diagram, we discuss the stability of endemic equilibria. We can show that, in the case when $d_S = d_R$, if two endemic equilibria $P^* = (S^*, I^*, R^*)$ and $P_* = (S_*, I_*, R_*)$ exist, with $S^* < S_*$, then P^* is unstable and P_* is asymptotically stable. See Proposition 4 in Appendix A.3 for proof. The result is summarized in the following theorem, and the bifurcation diagram using either B or R_0

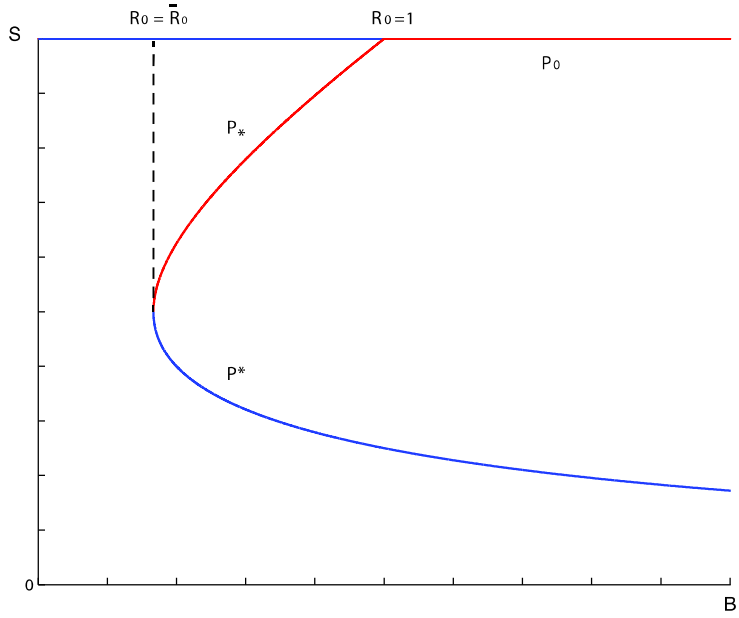


Figure 5.8: Backward bifurcation at $R_0 = 1$ for model (5.2) when $f(N) = \frac{1}{N^\alpha}$.

as a bifurcation parameter is illustrated in Figure 5.8.

Theorem 1. *Let* $f(N) = BN^{-\alpha}$, $\alpha > \frac{1}{1-pd_S}$. *Assume that* $d_S = d_R$.

(1) *If* $0 < R_0 < \overline{R}_0$, *then the disease-free equilibrium* P_0 *is the only equilibrium in* Γ *and it is asymptotically stable.*

(2) *If* $\overline{R}_0 < R_0 < 1$, *then there exist two endemic equilibria* $P^* = (S^*, I^*, R^*)$ *and* $P_* = (S_*, I_*, R_*)$ *with* $S^* < S_*$. *Equilibrium* P^* *is unstable while* P_0 *and* P_* *are asymptotically stable.*

(3) *If* $R_0 > 1$, *then* P_0 *is unstable, and a unique endemic equilibrium* $P_* = (S_*, I_*, R_*)$ *exists and is asymptotically stable.*

5.2.4 Hopf bifurcations

In this section, we consider the class of functions

$$f(N) = B(N^2 + aN + b).$$

We will assume that $R_0 = \sigma\, g(\bar{S}) = \frac{\beta}{d_I+\gamma}\bar{S}f(\bar{S}) > 1$ and investigate possibilities of multiple endemic equilibria. As shown in Section 2, an endemic equilibrium $P^* = (S^*, I^*, R^*)$ satisfies equation (5.10). In this case, the function $g(S)$ is a cubic function

$$g(S) = B_1\, S\, (S^2 + a_1 S + b_1), \tag{5.15}$$

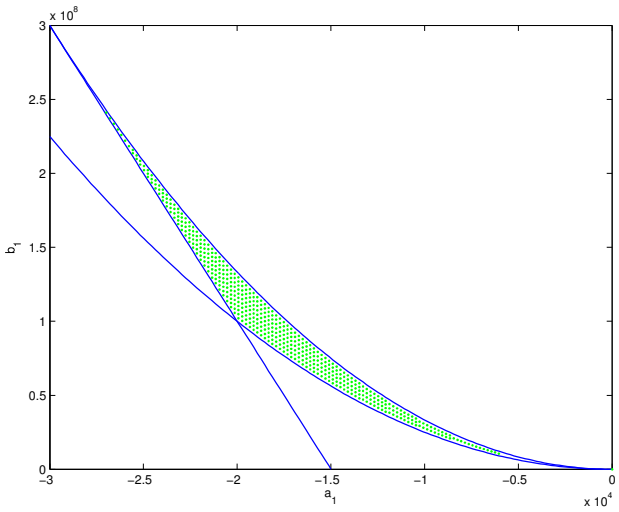


Figure 5.9: Shaded region indicates parameter values (a_1, b_1) for which three endemic equilibria are possible with a quadratic $f(N)$.

with

$$B_1 = B(1 - d_S p)^2,\ a_1 = \frac{a + 2p\Lambda}{1 - pd_S}, \quad \text{and} \quad b_1 = \frac{(p\Lambda)^2 + ap\Lambda + b}{(1 - pd_S)^2}. \tag{5.16}$$

We require that $g(S) = 0$ if and only if $S = 0$. This is the case when coefficients in $g(S)$ satisfy relation

$$a_1{}^2 < 4b_1. \tag{5.17}$$

Function $g(S)$ in (5.15) has two critical points

$$S_c^{\pm} = \frac{-a_1 \pm \sqrt{a_1{}^2 - 3b_1}}{3}. \tag{5.18}$$

When these critical points are real, distinct, and located in the interval $(0, \bar{S})$, equation (5.10) will have three solutions for appropriate values of B_1, resulting in three endemic equilibria. For both critical points to be real, distinct, and positive, it is necessary that

$$a_1 < 0, \text{ and } 3b_1 < a_1{}^2. \tag{5.19}$$

Since $g(\bar{S}) > \frac{1}{\sigma}$, for all solutions of $g(S) = \frac{1}{\sigma}$ to belong to $(0, \bar{S})$, it is necessary and sufficient that $g'(\bar{S}) > 0$ and $g''(\bar{S}) > 0$, namely that

$$b_1 > -2a_1\bar{S} - 3(\bar{S})^2 \quad \text{and} \quad a_1 > -3\bar{S}. \tag{5.20}$$

Figure 5.9 illustrates the nonempty region in the (a_1, b_1) parameter space defined by conditions (5.17), (5.19) and (5.20), where three endemic equilibria are possible.

Using

$$a_1 = -\frac{3}{2}(S_c^+ + S_c^-) \quad \text{and} \quad b_1 = 3S_c^+ S_c^-,$$

we can rewrite $g(S)$ as

$$g(S) = B_1 S\Big[S^2 - \frac{3}{2}(S_c^+ + S_c^-)S + 3S_c^+ S_c^-\Big].$$

Then

$$g(S_c^-) = \frac{B_1}{2} {S_c^-}^2 (3S_c^+ - S_c^-) \quad \text{and} \quad g(S_c^+) = \frac{B_1}{2} {S_c^+}^2 (3S_c^- - S_c^+). \tag{5.21}$$

The point S_c^- is a local maximum for the function $g(S)$ while the point S_c^+ is a local minimum. Whether three endemic equilibria occur is determined by relative relations among $g(S_c^\pm)$, $g(\bar{S})$ and $\frac{1}{\sigma}$.

Proposition 3. *Let $f(N) = B(N^2 + aN + b)$. Assume that $R_0 = \sigma g(\bar{S}) > 1$, $g'(\bar{S}) > 0$, and $g''(\bar{S}) > 0$.*

(1) *If $\frac{1}{\sigma} < g(S_c^+)$, then model (5.2) has exactly one endemic equilibrium.*

(2) *If $g(S_c^+) = \frac{1}{\sigma}$, then model (5.2) has two endemic equilibria.*

(3) *If $g(S_c^+) < \frac{1}{\sigma} < g(S_c^-)$, then model (5.2) has three endemic equilibria.*

(4) *If $g(S_c^-) = \frac{1}{\sigma}$, then model (5.2) has two endemic equilibria.*

(5) *If $g(S_c^+) < \frac{1}{\sigma}$, then model (5.2) has exactly one endemic equilibrium.*

Proposition 3 is illustrated in Figure 5.10. Using (5.16) and (5.21) we can rewrite conditions in Proposition 3 in terms of ranges for parameter B_1. For instance, the range of B_1 for model (5.2) to have three endemic equilibria is

$$\frac{2}{{S_c^-}^2(3S_c^+ - S_c^-)} < B_1 < \frac{2}{{S_c^+}^2(3S_c^- - S_c^+)}.$$

To investigate secondary bifurcations when multiple endemic equilibria exist, we examine the stability of endemic equilibria. Routh-Hurwitz conditions for all eigenvalues of the Jacobian matrix F at an endemic equilibrium to have negative real parts are

1. $T = \text{trace}(F) < 0$,
2. $D = \det(F) < 0$, and
3. $C = TM - D < 0$,

where M is the sum of the second-order principle minors of F. It can be verified that

$$T = -If(N) - d_S - d_R < 0,$$

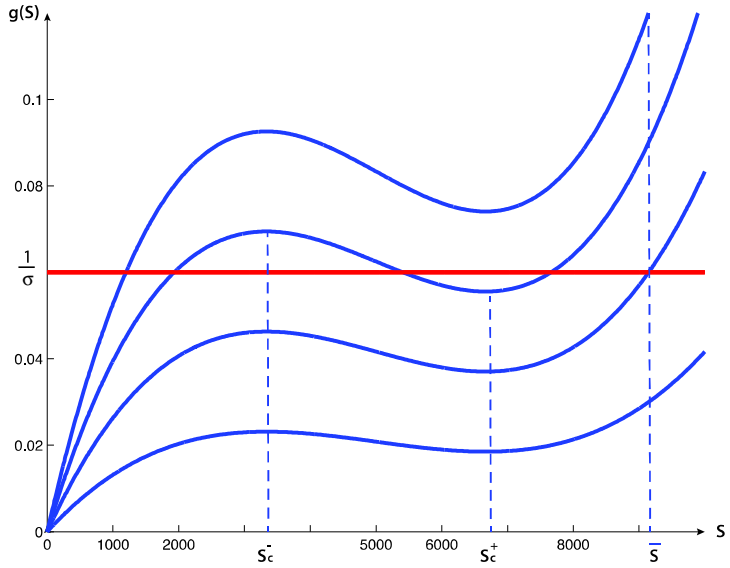


Figure 5.10: Number of solutions of equation of $g(S) = \frac{1}{\sigma}$ in $[0, \bar{S})$ with a quadratic $f(N)$.

and that

$$\begin{aligned} D &= -\frac{d_R}{\beta} If(N) + ISf'(N)[d_R d_S + \gamma d_S - d_R(d_I + \gamma)] \\ &= -d_R(d_I + \gamma)[If(N) + ISf'(N)(1 - pd_S)] \\ &= -d_R(d_I + \gamma)\frac{dg(S)}{dS}. \end{aligned}$$

Then, whenever $\frac{dg(S)}{dS} < 0$, D is positive and the equilibrium will be unstable. As shown in Figure 5.10, when there are three endemic equilibria, the equilibrium with intermediate S^* value will always be unstable. When $\frac{dg(S)}{dS} > 0$, the stability is determined by the sign of $C = TM - D$. This allows the possibility for one of the branches to undergo stability change, and possibility for Hopf bifurcation.

Numerical computation using MATLAB reveals that the sign of C at the endemic equilibrium with the largest value of S^* can change as values of parameters change. Since $\frac{dg(S^*)}{dS} > 0$ at this equilibrium, a sign change in C indicates that a pair of complex eigenvalues cross the imaginary axis and the occurrence of a Hopf bifurcation.

Numerical bifurcation analysis using XPPAUTO confirms that the system has a Hopf bifurcation at the endemic equilibria with the largest value of S^*. We show in the bifurcation diagram in Figure 5.11 that, as a bifurcation parameter B decreases, the equilibrium changes from being stable to unstable, and a supercritical Hopf bifurcation occurs, creating a stable periodic orbit. MATLAB simulations also confirm that model (5.2) has a stable periodic orbit in the range of B given in the bifurcation diagram. In Figure 5.12a, a solution of (5.2) is shown to converge to a stable periodic solution when

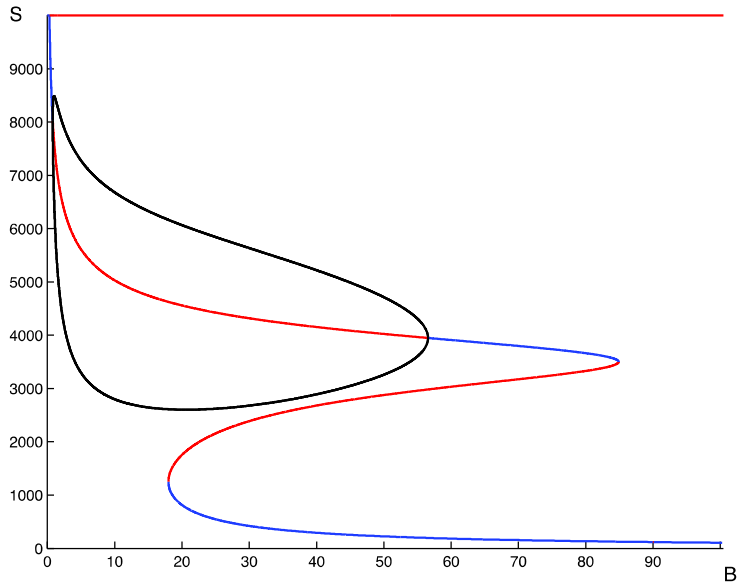
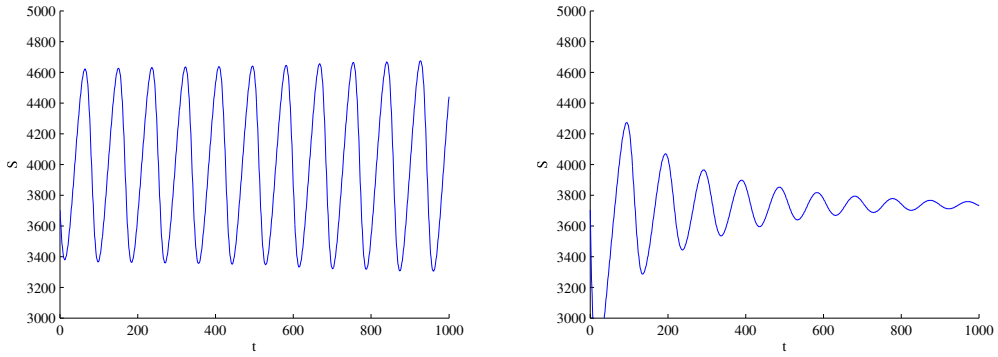


Figure 5.11: Bifurcation diagram when $f(N)$ is quadratic. A supercritical Hopf bifurcation occurs when B decreases through $B = 55$. The dark closed loop indicates stable periodic solutions, and they exist for B in the range $(0.5, 55)$.

$B = 50$. In Figure 5.12b, we show that solutions converge to an endemic equilibrium at a $B = 75$, before the Hopf bifurcation occurs. We remark that, since the incidence form is $\beta ISB(N^2 + aN + b)$, bifurcation observed as we vary parameter B can also be observed if we vary parameter β.

5.2.5 Multiple infectious stages

Similar results and analysis in Sections 3 and 4 can be carried out for epidemic models with more complex structures than the simple SIR model. We consider a multiple-stage model that describes the transmission dynamics of infectious diseases progressing through a long infectious period such as HIV/AIDS [126, 54, 55, 73, 75, 76, 96]. The n-stage model as depicted in Figure 5.13 is a generalization of the single-stage SIR model (5.2). The infectious period is partitioned into n distinct stages with $I_j(t)$ individuals in the j-th infectious stage. Individuals in the j-th infectious stage are assumed to have a transmission coefficient β_j and the transfer rate from the j-th stage to the next is given by γ_j, $j = 1, \cdots, n$. We assume that all parameter values are nonnegative and, as in the preceding sections, that $0 < d_S \leq d_R$, and $d_S < d_{I_j}$, $j = 1, \cdots, n$. The model is



(a) A solution converges to a stable periodic solution when $B = 50$.

(b) A solution converges to an endemic equilibrium when $B = 75$.

Figure 5.12: MATLAB simulations for $B = 50$ and $B = 75$.

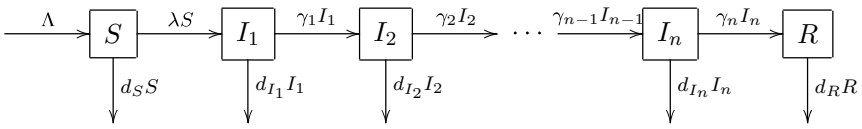


Figure 5.13: Transfer diagram for the n-stage model (5.22). Incidence term is given by $\lambda S = \sum_{j=1}^{n} \beta_j I_j f(N) S$.

described by a system of $n+2$ differential equations

$$\begin{aligned} S' &= \Lambda - \bar{\lambda}S - d_S S, \\ I_1' &= \bar{\lambda}S - \gamma_1 I_1 - d_{I_1} I, \\ I_i' &= \gamma_{i-1} I_{i-1} - \gamma_i I - d_{I_i} I, \quad \text{for } \ i = 2 \cdots n, \\ R' &= \gamma_n I_n - d_R R, \end{aligned} \tag{5.22}$$

where the force of infection is given by

$$\bar{\lambda} = \sum_{j=1}^{n} \beta_j I_j f(N). \tag{5.23}$$

Adding the equations in (5.22) we obtain

$$N^{'} = \Lambda - d_S S - d_{I_1} I_1 - \cdots - d_{I_n} I_n - \gamma_n I_n \leq \Lambda - d_S N.$$

It follows that $\limsup_{t\to\infty} N(t) \leq \Lambda/d_S$. The global dynamics of model (5.22) can be investigated in the positively invariant compact subset of $\mathbb{R}_+^{n+2}$

$$\Delta = \{(S, I_1, \cdots, I_n, R) \in \mathbb{R}_+^{n+2} \ : \ 0 \leq S + I_1 + \cdots + I_n \leq \Lambda/d_S\}.$$

Let

$$A = \begin{bmatrix} d_{I_1} + \gamma_1 & 0 & 0 & \cdots & & \\ -\gamma_1 & d_{I_2} + \gamma_2 & 0 & \cdots & & \\ 0 & -\gamma_2 & d_{I_3} + \gamma_3 & \cdots & & \\ \vdots & \vdots & \vdots & \ddots & & \\ & & & & d_{I_n} + \gamma_n & 0 \\ & & & & -\gamma_n & d_R \end{bmatrix}. \tag{5.24}$$

Then A is a lower triangular matrix with non-zero entries on the diagnonal. Therefore, A is invertible. In particular we can define,

$$\sigma_n = (\beta_1, \cdots, \beta_n, 0) A^{-1} (1, 0, \cdots, 0)^T, \tag{5.25}$$

and

$$p_n = (1, \cdots, 1) A^{-1} (1, 0, \cdots, 0)^T, \tag{5.26}$$

where the superscript T denotes the matrix transposition. The inverse A^{-1} of the triangular and bidiagonal matrix A can be computed to give the explicit expressions for σ_n and p_n

$$\sigma_n = \sum_{j=1}^{n} \frac{\beta_j}{d_{I_j} + \gamma_j} \delta_{j-1} > 0, \tag{5.27}$$

and

$$p_n = \Big(\sum_{j=1}^{n} \frac{1}{d_{I_j}+\gamma_j}\delta_{j-1}\Big) + \frac{1}{d_R}\delta_n > 0, \tag{5.28}$$

where δ_k is given by

$$\delta_k = \prod_{j=1}^{k} \frac{\gamma_j}{d_{I_j}+\gamma_j} > 0.$$

Notice that when $n = 1$, the values given in (5.27) and (5.28) agree with σ and p in Section 2. Thus σ_n can be regarded as the total transmisibility for the n-staged model (5.22) and p_n can be interpreted as the mean remaining life expectancy of those who become infected. The quantity δ_k represents the proportion of those who become infected who survive to reach the $(k+1)$-th stage of infection.

The basic reproduction number of (5.22) is derived in [54] as

$$R_0 = \sigma_n \, \bar{S} \, f(\bar{S}), \tag{5.29}$$

where $\bar{S} = \frac{\Lambda}{d_S}$, using the method of next generation matrix [132]. We see that, when $n = 1$, the expression for R_0 in (5.29) reduces to that for the single stage model in Section 2. For $f(N) = N^{-\alpha}$ and $0 < \alpha \leq 1$, it is also established in [54] that if $R_0 \leq 1$ then the disease-free equilibrium $P_0 = (\bar{S}, 0, \cdots, 0)$ is globally asymptotically stable in Δ; if $R_0 > 1$ then P_0 is unstable and the model (5.22) is uniformly persistent. As a consequence, an endemic equilibrium $P^* = (S^*, I_1^*, \cdots, I_n^*, R^*)$ exists in the interior $\overset{\circ}{\Delta}$ of Δ. Furthermore, it is shown in [54] that P^* is unique and globally asymptotically stable in $\overset{\circ}{\Delta}$ when $R_0 > 1$. We show in this section that when $f(N)$ is chosen from a wider class functions, more complicated dynamics are possible.

In the following, we investigate the number of endemic equilibria following the presentation of [54]. An endemic equilibrium $(S^*, I_1^*, \cdots, I_n^*, R^*)$ of (5.22) satisfies

$$\begin{aligned}
0 &= \Lambda - d_S S^* - \bar{\lambda}^* S^*, \\
0 &= \bar{\lambda}^* S^* - (d_{I_1} + \gamma_1) I_1^*, \\
0 &= \gamma_{i-1} I_{i-1}^* - (d_{I_i} + \gamma_i) I_i^*, \quad i = 2, \ldots, n-1, \\
0 &= \gamma_{n-1} I_{n-1}^* - (d_{I_n} + \gamma_n) I_n^*, \\
0 &= \gamma_n I_n^* - d_R R^*
\end{aligned} \tag{5.30}$$

where

$$\bar{\lambda}^* = \sum_{j=1}^{n} \beta_j I_j^* f(N^*) \tag{5.31}$$

is the force of infection at an endemic equilibrium P^*. We write the equations in (5.30) for $I_1^*, \cdots, I_n^*$ and R^* in the form

$$(I_1^*, \cdots, I_n^*, R^*)^t = \bar{\lambda}^* \, S^* A^{-1} (1, 0, \cdots, 0)^t, \tag{5.32}$$

where A is the matrix in (5.24). Multiplying the row vector $(\beta_1, \cdots, \beta_n, 0)$ to (5.32) and using (5.23) and (5.25), we obtain

$$\sum_{i=1}^{n} \beta_i I_i^* = (\beta_1, \cdots, \beta_n, 0)\,(I_1^*, \cdots, I_n^*, R^*)^t = (\beta_1, \cdots, \beta_n, 0) A^{-1} (1, 0, \cdots, 0)^t \bar{\lambda}^* S^*$$
$$= \sigma_n \bar{\lambda}^*\, S^* = \sigma_n f(N^*) S^* \sum_{j=1}^{n} \beta_j I_j^*.$$

Since $\sum_{i=1}^{n} \beta_i I_i^* \neq 0$ at an endemic equilibrium, it follows that

$$\sigma_n\, S^*\, f(N^*) = 1. \tag{5.33}$$

Similarly, multiplying row vector $(1, \cdots, 1)$ to (5.32) and applying (5.26) we have

$$\sum_{i=1}^{n} I_i^* + R^* = (1, \cdots, 1)(I_1^*, \cdots, I_n^*, R^*)^t = p_n f(N^*) S^* \sum_{j=1}^{n} \beta_j I_j^*, \tag{5.34}$$

where $p_n > 0$ is defined in (5.26). From the first equation of (5.30) we get

$$f(N^*) S^* \sum_{j=1}^{n} \beta_j I_j^* \;=\; \Lambda - d_S S^*,$$

which, together with (5.34), implies

$$\sum_{i=1}^{n} I_i^* + R^* \;=\; p_n(\Lambda - d_S S^*),$$

and thus

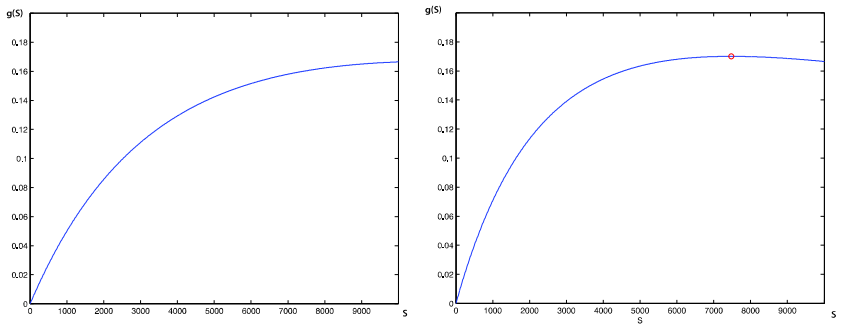
$$N^* = S^* + \sum_{i=1}^{n} I_i^* + R^* \;=\; S^* + p_n(\Lambda - d_S S^*) \;=\; (1 - p_n d_S) S^* + p_n \Lambda. \tag{5.35}$$

Substituting (5.35) into (5.33) we obtain the following equation for an endemic equilibrium $P^* = (S^*, I_1^*, \cdots, I_n^*, R^*)$

$$g_n(S^*) := S^* f((1 - p_n d_S) S^* + p_n \Lambda) \;=\; \frac{1}{\sigma_n}. \tag{5.36}$$

Comparing equation (5.36) to the equilibrium equation (5.10) for the single stage model, we see that the number of endemic equilibria for the n-stage model (5.22) can be determined in exactly the same way as in Sections 2-4. As a consequence, results in Sections 3 and 4, Theorem 1 and Proposition 3 in particular, hold for the n-stage model, with p and σ replaced by p_n and σ_n, respectively.

We now consider the effects on the dynamical outcomes described in the preceding



(a) Function $g_n(S)$ with $n = 1$ (b) Function $g_n(S)$ with $n = 5$

Figure 5.14: Graphs of function $g_n(S)$ for $n = 1$ and $n = 5$.

sections of adding additional infective stages to an SIR model. To simplify the discussion, we assume that $\beta_j = \beta$, $d_{I_j} = d_I$, and $\gamma_j = n\gamma$ for all j. In other words, we chose all the stages of the disease to be identical and assume that individuals move through the stages at a constant rate. The choice of $\gamma_j = n\gamma$ means that adding stages will not change the average length of infection for those who recover.

In such a case, the expression for p_n is simplified to

$$\begin{aligned} p_n &= \frac{1}{d_I}\Big[1 - \left(\frac{n\gamma}{d_I + n\gamma}\right)^n\Big] + \frac{1}{d_R}\left(\frac{n\gamma}{d_I + n\gamma}\right)^n \\ &= \frac{1}{d_I} + \left(\frac{1}{d_R} - \frac{1}{d_I}\right)\left(\frac{n\gamma}{d_I + n\gamma}\right)^n \end{aligned}$$

The quantity p_n depends on n only through the term $\delta_n = \left(\frac{n\gamma}{d_I+n\gamma}\right)^n$, which is a decreasing function of n. Since we assume that $d_I > d_R$, we know p_n and $\frac{1}{1-d_S p_n}$ decrease as n increases. As a result, the range of α values for which backward bifurcation occurs (Theorem 1) becomes larger as n increases; adding more infective stages to a simple SIR model with $f(N) = N^{-\alpha}$, $\alpha > 1$, will increase the chance for backward bifurcation and the associated catastrophic behaviours.

For example, Figures 5.14a and 5.14b show the function $g_n(S)$ for $n = 1$ and $n = 5$, respectively. With only a single stage, $g_1(S)$ is monotonic and has no critical points; only one sub-threshold endemic equilibrium is possible. With five stages, $g_5(S)$ has a critical point in the feasible region; two sub-threshold endemic equilibria and backward bifurcation are possible for suitable range of parameter values.

5.2.6 Conclusions

It is known that complicated dynamics can occur through backward bifurcation or Hopf bifurcation in epidemic models with nonlinear incidence [66, 89], or complex group structures [58, 72], or with time delays [130]. In this Section, we have shown that nonlinear

density dependence in disease incidence can also give rise to backward bifurcations and Hopf bifurcations, in simple models of SIR type.

For incidence functions of the form $\frac{\beta IS}{N^\alpha}$, $\alpha > 1$, we proved that backward bifurcation and bi-stability can occur for $R_0 < 1$. We have also shown that incidence functions of the form $\beta ISf(N)$ with $f(N)$ being quadratic can lead to periodic oscillations through Hopf bifurcation. On the one hand, our results provide a new mechanism for complicated dynamics to occur in simple epidemic models. On the other, they indicate that, by restricting incidence terms to traditional bilinear form (βIS) or standard form ($\frac{\beta IS}{N}$), we may have unintentionally eliminated the possibility of many complicated but interesting dynamics.

Chapter 6

Conclusion

6.1 Summary of Results

In this thesis we have focused mostly on the problem of estimating HIV incidence, the total size of the HIV-positive population, and the size of the population that is HIV-positive but undiagnosed. Methods for estimating these quantities are developed in Chapters 2, 3, and 4 and the methods are demonstrated using the Province of Alberta as an example. Although the focus of this thesis has been on HIV in the Province of Alberta, the techniques used in this thesis can apply to a wide variety of different modelling contexts including HIV in other regions or other transmissible diseases. The project on antiretroviral drug resistant strains of HIV in Section 5.1 uses several of the same methods. Section 5.2 explores the asymptotic behaviour of a simple disease model with population dependent transmission.

6.1.1 HIV in Alberta

In the province of Alberta, newly diagnosed HIV cases are reported to Alberta Health for disease surveillance. Using this type of data to estimate the total size of the HIV-positive population is complicated by the fact that there may be a long delay in diagnosis of HIV and a significant number of those who are HIV-positive may remain undiagnosed. In this thesis we have addressed this problem by creating a mathematical model describing HIV transmission and diagnosis. The model is calibrated using the available case report data and validated using both reserved data and data collected for other studies.

Calibration of the model was carried out using two methods. Nonlinear least squares fit was used for preliminary exploration while a Bayesian fit was employed to include uncertainty in the fitted parameters and was carried out using an MCMC method. Fitting the model requires special care as some of the model parameters are nonidentifiable using the available data. In order to detect the presence of nonidentifiable parameters the local sensitivity matrix was examined using singular value decomposition. Variance decomposition was used to indicate which parameters are not simultaneously identifi-

able. These results were further verified by examining two-dimensional contour plots of the sum of squared errors.

Model validation used a method based on Bayesian hypothesis testing. In this method, the distribution for the validation data estimated by the fitted model was compared to an alternative distribution. Reserved data from 1999, 2000, 2011, and 2012 and PHAC estimates of the fraction undiagnosed were shown to support the model results.

The fitted model was used to produce estimates of HIV incidence and the size of the total HIV-positive population including those who are have not yet been diagnosed. The model predicts that the total size of the HIV-positive population in 2015 will be around 4900 people with about 13% of this population undiagnosed. The predicted incidence rate for 2015 is around 9.5 new cases per 100 000 population. These estimates were projected into the future to give information on how these quantities can be expected to change over the next several years. The amount of uncertainty in these projections was also quantified.

A sensitivity analysis indicated that the two parameters which have the greatest effect on the outcomes of interest are the transmission coefficient β_I and the diagnosis rate α. The theoretical interventions represented by modification of these parameters for future years were tested. The results of these simulations suggest that a significant improvement in disease incidence can be achieved by modifying either of these parameters over the long term. If only a short term modification is possible, temporarily increasing the diagnosis rate α has a more lasting effect than a temporary intervention targetting the transmission coefficient β_I.

6.1.2 Antiretroviral drug resistance

It has recently been accepted that widespread HIV testing and antiretroviral drug treatment can greatly reduce the transmission of the virus by reducing the viral load of those who are successfully treated. In Section 5.1 a mathematical model was used to investigate the potential impact of this type of "treatment as prevention" program on the development of antiretroviral drug resistant viral strains. The parameters for the model are selected by referring to a variety of sources in medical literature. This project utilizes many of the sensitivity and uncertainty techniques discussed in the context of the Province of Alberta project in Chapter 4. The results indicate that drug resistant viral strains may have an important effect on the success of a program to reduce HIV transmission through increased treatment.

6.1.3 Population dependent transmission

While much of this thesis considered only short to medium term model outcomes, Section 5.2 considered the long term behaviour of a simple model with population dependent incidence. Two different types of population dependence were used in order to demon-

strate that even simple models can undergo complicated behaviour. In particular, both backward bifurcations and Hopf bifurcations can occur if the appropriate type of population dependence is included in disease transmission terms.

6.2 Future Possibilities

There are a number of possibilities for future work on this project. These can be divided into two main categories: extensions to the model and additional modelling tools.

6.2.1 Extending the HIV model

The HIV model used in Chapters 2 to 4 can be used to answer some questions about HIV incidence and prevalence. However the model is quite simple and there are many questions which are beyond its scope. In this section, some of the most obvious extensions of the model are discussed. In some cases, adding more detail to the model may improve the model fit or make it more realistic for longer term projection. At the same time, using a more detailed model may allow new outcomes to be measured. However, using a more detailed model usually also requires determining values for more parameters, a task that will most often require additional data along with additional computation time at all stages in the modelling process.

6.2.1.1 Risk groups

One possibility for adding more detail and realism to the model is to include a number of different risk groups. This can be used to account for the fact that different parts of the population have differing risk behaviours. Some groups, such as intravenous drug users and men who have sex with men, are known to be at much higher risk of infection and may also have differing rates of diagnosis, treatment, and death. In fact, we have already seen how one group differs from the total population modelled using the single group version of the model. In Section 3.3 we noted that data from prenatal HIV testing did not provide validation for the single group model. This suggests that males and females are also important groups to consider.

Including all the the risk groups already mentioned, a model with five different groups could be constructed as illustrated in Figure 6.1. Each of the risk groups illustrated follows the simple model from Chapter 2, found in equations (2.1) on 12. The arrows indicate disease transmission between and within groups. Individuals may also move between risk groups as they change their risk behaviours.

Useful results may also be available by including only some of these populations. A two-group model that divides the population into males and females could provide information on sex differences in HIV risk behaviours. A three-group model that additionally includes a group of men who have sex with men could be used to investigate the

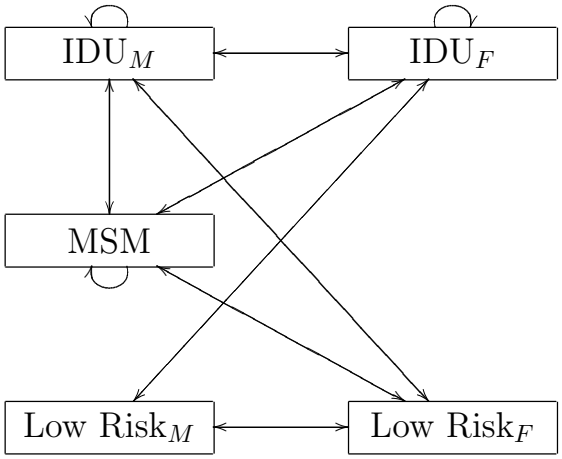


Figure 6.1: A model with five risk groups. Arrows indicate contact between groups. Each risk group follows the simple model.

role of homosexual and bisexual men in HIV transmission. A four-group model dividing the population by sex and level of risk could be used to add additional detail.

Increasing the number of risk groups increases the number of model compartments and the number of parameter values required to completely specify the model. Since case report data stratified by risk group is available, the increase in required parameters is partially compensated by an increase in the data available. However, unlike for the simple one group model, there is not much data available on the total population sizes for several of the risk groups because it is difficult to know how many intravenous drug users or men who have sex with men there are. On the other hand, if the necessary parameters can be fit using the case report data, it may be possible to produce estimates of the sizes of the risk populations over time as a byproduct of the disease model fitting. Whether or not this is possible will depend on the identifiability of the initial size, growth, and death rates for the unknown populations.

6.2.1.2 Treatment and disease stages

Another possibility for including more detail into the model would be to include additional compartments in the base model. These could be used to describe treated populations or to provide more detailed disease progression through the inclusion of multiple disease stages.

For example, a model including treatment is illustrated in Figure 6.2. In this model, a treated compartment T is added to the simple transmission and diagnosis model. The population represented by this compartment is assumed to be effectively treated with viral loads undetectable or nearly so. The treated population is therefore assumed not to transmit the virus to others and their death rate is improved over those who are merely

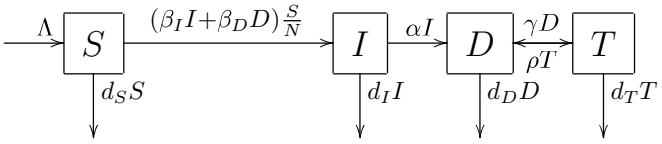


Figure 6.2: Model diagram for an HIV model including transmission, diagnosis, and treatment.

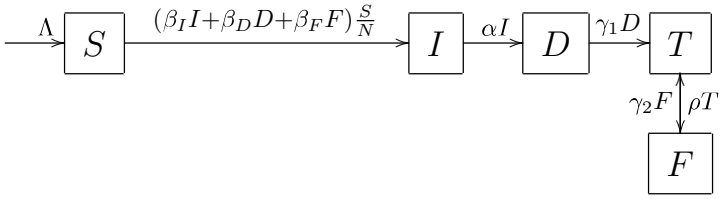


Figure 6.3: Model diagram for an HIV model including transmission, diagnosis, treatment and treatment stoppage.

diagnosed. As some patients may stop and restart treatment for a variety of reasons, the model allows those who have been treated to return to the diagnosed compartment resuming transmission until the are once again effectively treated.

This model includes only a few additional parameters not included in the original model (2.1). However, without some data on HIV treatment, these parameters are non-identifiable. This model could be fit using data giving the number of patients currently under effective treatment or data for the number of individuals starting or restarting a treatment program.

If data was only available for the number of individuals beginning treatment for the first time, it would probably be necessary to modify the model somewhat to utilize this data. See, for example, Figure 6.3. In this model, those who stop treatment do not return to the diagnosed compartment, but instead remain separate so that treatment initiation for those who have never been treated can be distinguished from treatment resumption for those whose treatment has been interrupted.

Another possible extension of the model would be to include multiple disease stages. This may be done for two different reasons. The first possibility is that additional stages may add detail about realistic disease stages such as acute infection or AIDS. Alternatively, additional stages may be introduced in order to modify the survival time or time to diagnosis distributions to more closely approximate those commonly used by statisticians and epidemiologists such as the Weibull distribution. In either case, some additional data or assumptions on disease progression are required to fit the additional parameters introduced by the added compartments. A model with two disease stages is

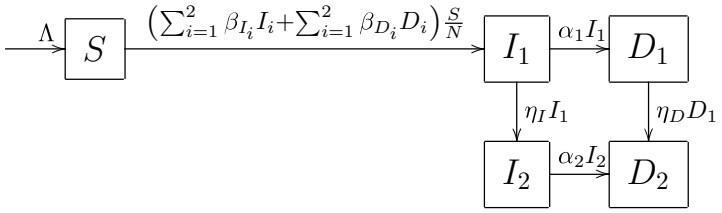


Figure 6.4: Model diagram for an HIV model including transmission, diagnosis, and two disease stages.

illustrated in Figure 6.4

6.2.1.3 Economic analysis

A third possibility for extending the model is to include an economic analysis. Model outcomes can be reinterpreted to incorporate the costs of testing and treatment. This would allow the potential costs and benefits of various intervention programs to be quantified allowing different possibilities to be directly compared.

6.2.2 Additional tools

In addition to extensions to the model itself, additional modelling tools may be useful. In particular, tools for comparing models could assist in deciding which model to use when several variants are available. Tools for parameter reduction may become vital as the number of parameters increases since computational resources may be limited.

6.2.2.1 Comparing models

When multiple models can be fit using the same data, it is natural to ask how they compare. Models may be compared using measures of how closely the model reproduces the data, but this may not be immediately obvious. One model may fit better on some of the data while another improves the fit on other data. Furthermore, goodness of fit may not be the only factor that is relevant in choosing a model. The number of parameters to be estimated, or the ease of computing particular model outcomes of interest may also be relevant considerations. Developing tools for comparing models will require determining what properties are desirable in a model and creating methods of quantifying these properties.

It is possible that our Bayesian model validation techniques may be adapted to compare models using either fitting or validation data. But this technique is generally quite computationally intensive and as such may not be appropriate for a preliminary investigation used to select an appropriate model. Other possibilities should be considered.

6.2.2.2 Parameter reduction

Another challenge raised by the potential expansion of the HIV model is the increasing number of parameters. While we have already highlighted identifiability concerns that may arise as the number of parameters increases, there is also the question of how many parameters can be practically estimated computationally. In particular, our Bayesian methods including MCMC sampling, hypothesis testing for validation, and uncertainty analysis tend to be somewhat computationally intensive, especially when the number of parameters is large. In order to draw a representative sample in a higher dimensional parameter space, the sample size must be quite large and computing model outcomes for every element of the sample can require significant computation time. Some of these concerns may be alleviated by utilizing more efficient algorithms and more powerful computers, however methods that reduce the number of parameters should also be considered.

Reducing the number of parameters may involve using the singular value decomposition and variance decomposition to identify not only those parameters which are nonidentifiable, but also the parameter combinations that are least important to the model fit while still remaining identifiable.

The number of parameters may be reduced by simply fixing the values of one or more parameters, but other methods involving reparameterization of the model to write some of the parameters in terms of the others may also be possible. Deciding how many parameters can be removed without serious impact on the model performance may require the use of tools to compare models as previously discussed.

Bibliography

[1] A. E. Ades and A. J. Sutton, *Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches*, Journal of the Royal Statistical Society A, 169 (2005), pp. 5–35.

[2] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, 1992.

[3] C. G. Arandi, I. Loya, J. Jacobson, F. Arana, and S. M. Miranda, *P3. 398 sentinel surveillance and prevention of sexually transmitted infections among female sex workers in Guatemala: first findings from VICITS*, Sexually Transmitted Infections, 89 (2013), pp. A273–A274.

[4] J. Arino, K. Cooke, P. Van Den Driessche, and J. Velasco-Hernández, *An epidemiology model that includes a leaky vaccine with a general waning function*, Discrete and Continuous Dynamical Systems Series B, 4 (2004), pp. 479–495.

[5] J. Arino and C. C. McCluskey, *Effect of a sharp change of the incidence function on the dynamics of a simple disease*, Journal of Biological Dynamics, 4 (2010), pp. 490–505.

[6] L. Bao, *A new infectious disease model for estimating and projecting HIV/AIDS epidemics*, Sexually Transmitted Infections, 88 (2012), pp. i58–i64.

[7] L. Bao, J. A. Salomon, T. Brown, A. E. Raftery, and D. R. Hogan, *Modelling national HIV/AIDS epidemics: revised approach in the UNAIDS Estimation and Projection Package 2011*, Sexually Transmitted Infections, 88 (2012), pp. i3–i10.

[8] T. Bärnighausen, J. Bor, S. Wandira-Kazibwe, and D. Canning, *Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models*, Epidemiology, 22 (2011), pp. 27–35.

[9] T. Bärnighausen, T. A. McWalter, Z. Rosner, M.-L. Newell, and A. Welte, *HIV incidence estimation using the BED capture enzyme immunoassay: systematic review and sensitivity analysis*, Epidemiology, 21 (2010), pp. 685–697.

[10] R. E. Barth, M. F. S. van der Loeff, R. Schuurman, A. I. Hoepelman, and A. M. Wensing, *Virological follow-up of adult patients in antiretroviral treatment programmes in sub-Saharan Africa: a systematic review*, The Lancet Infectious Diseases, 10 (2010), pp. 155–166.

[11] N. G. Becker, J. J. C. Lewis, Z. Li, and A. McDonald, *Age-specific back-projection of HIV diagnosis data*, Statistics in Medicine, 22 (2003), pp. 2177–2190.

[12] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*, vol. 571, John Wiley & Sons, 2005.

[13] E. Bendavid, M. L. Brandeau, R. Wood, and D. K. Owens, *Comparative effectiveness of HIV testing and treatment in highly endemic regions*, Archives of Internal Medicine, 170 (2010), pp. 1347–1354.

[14] D. Bezemer, F. de Wolf, M. C. Boerlijst, A. van Sighem, T. D. Hollingsworth, M. Prins, R. B. Geskus, L. Gras, R. A. Coutinho, and C. Fraser, *A resurgent HIV-1 epidemic among men who have sex with men in the era of potent antiretroviral therapy*, AIDS, 22 (2008), pp. 1071–1077.

[15] K. W. Blayneh, A. B. Gumel, S. Lenhart, and T. Clayton, *Backward bifurcation and optimal control in transmission dynamics of West Nile virus*, Bulletin of Mathematical Biology, 72 (2010), pp. 1006–1028.

[16] S. M. Blower and H. Dowlatabadi, *Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example*, International Statistical Review, 62 (1994), pp. 229–243.

[17] S. M. Blower, H. B. Gershengorn, and R. M. Grant, *A tale of two futures: HIV and antiretroviral therapy in San Francisco*, Science, 287 (2000), pp. 650–654.

[18] J. T. Boerma, P. D. Ghys, and N. Walker, *Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard*, The Lancet, 362 (2003), pp. 1929–1931.

[19] D. Boulos, P. Yan, D. Schanzer, R. Remis, and C. Archibald, *Estimates of HIV prevalence and incidence in Canada, 2005*, Canada Communicable Disease Report, 32 (2006), p. 165.

[20] F. Brauer, *Backward bifurcations in simple vaccination models*, Journal of Mathematical Analysis and Applications, 298 (2004), pp. 418–431.

[21] F. Brauer and C. Castillo-Châavez, *Mathematical models in population biology and epidemiology*, Springer, 2001.

[22] R. Brookmeyer, *Measuring the HIV/AIDS epidemic: approaches and challenges*, Epidemiologic Reviews, 32 (2010), pp. 26–37.

[23] T. Brown, N. C. Grassly, G. Garnett, and K. Stanecki, *Improving projections at the country level: the UNAIDS Estimation and Projection Package 2005*, Sexually Transmitted Infections, 82 (2006), pp. iii34–iii40.

[24] T. Brown, A. E. Raftery, J. A. Salomon, R. F. Baggaley, J. Stover, and P. Gerland, *Modelling HIV epidemics in the antiretroviral era: the UNAIDS Estimation and Projection package 2009*, Sexually Transmitted Infections, 86 (2010), pp. ii3–ii10.

[25] T. Brown, J. A. Salomon, A. E. Raftery, and E. Gouws, *Progress and challenges in modelling country-level HIV/AIDS epidemics: the UNAIDS Estimation and Projection Package 2007*, Sexually Transmitted Infections, 84 (2008), pp. i5–i10.

[26] R. Bunnell, J. P. Ekwaru, P. Solberg, N. Wamai, W. Bikaako-Kajura, W. Were, A. Coutinho, C. Liechty, E. Madraa, G. Rutherford, et al., *Changes in sexual behavior and risk of HIV transmission after antiretroviral therapy and prevention interventions in rural uganda*, AIDS, 20 (2006), pp. 85–92.

[27] M. P. Busch, C. D. Pilcher, T. D. Mastro, J. Kaldor, G. Vercauteren, W. Rodriguez, C. Rousseau, T. M. Rehle, A. Welte, M. D. Averill, and J. M. Garcia-Calleja, *Beyond detuning: 10 years of progress and new challenges in the development and application of assays for HIV incidence estimation*, AIDS, 24 (2010), pp. 2763–2771.

[28] F. Canova and L. Sala, *Back to square one: identification issues in DSGE models*, Journal of Monetary Economics, 56 (2009), pp. 431–449.

[29] B. P. Carlin and T. A. Louis, *Bayes and empirical Bayes methods for data analysis*, Statistics and Computing, 7 (1997), pp. 153–154.

[30] Central Statistical Office, *1971, 1981, 1991, 2001 census demographic indicators: Botswana*, tech. rep., Republic of Botswana, 2002.

[31] Centre for Communicable Diseases and Infection Control, *Summary: estimates of HIV prevalence and incidence in Canada 2011*, tech. rep., Public Health Agency of Canada, 2012.

[32] S. Chib and E. Greenberg, *Understanding the Metropolis-Hastings algorithm*, The American Statistician, 49 (1995), pp. 327–335.

[33] M. S. Cohen, Y. Q. Chen, M. McCauley, T. Gamble, M. C. Hosseinipour, N. Kumarasamy, J. G. Hakim, J. Kumwenda, B. Grinsztejn,

J. H. Pilotto, et al., *Prevention of HIV-1 infection with early antiretroviral therapy*, New England Journal of Medicine, 365 (2011), pp. 493–505.

[34] S. Corvasce, M. Violin, L. Romano, F. Razzolini, I. Vicenti, A. Galli, P. Duca, I. Caramma, C. Balotta, and M. Zazzi, *Evidence of differential selection of HIV-1 variants carrying drug-resistant mutations in seroconverters*, Antiviral Therapy, 11 (2006), p. 329.

[35] D. C. Cox and P. Baybutt, *Methods for uncertainty analysis: a comparative survey*, Risk Analysis, 1 (1981), pp. 251–258.

[36] R. de Boer and M. Li, *Density dependence in disease incidence and its impacts on transmission dynamics*, Canadian Applied Mathematics Quarterly, 19 (2011), pp. 195–218.

[37] C. de Mendoza, C. Rodriguez, A. Corral, J. del Romero, O. Gallego, and V. Soriano, *Evidence for differences in the sexual transmission efficiency of HIV strains with distinct drug resistance genotypes*, Clinical Infectious Diseases, 39 (2004), pp. 1231–1238.

[38] D. De Pauw, K. Steppe, and B. De Baets, *Identifiability analysis and improvement of a tree water flow and storage model*, Mathematical Biosciences, 211 (2008), pp. 314–332.

[39] J. Dushoff, W. Huang, and C. Castillo-Chavez, *Backwards bifurcations and catastrophe in simple models of fatal diseases*, Journal of Mathematical Biology, 36 (1998), pp. 227–248.

[40] J. W. Eaton, L. F. Johnson, J. A. Salomon, T. Bärnighausen, E. Bendavid, A. Bershteyn, D. E. Bloom, V. Cambiano, C. Fraser, J. A. Hontelez, et al., *HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa*, PLoS medicine, 9 (2012), p. e1001245.

[41] L. E. Eberly and B. P. Carlin, *Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models*, Statistics in Medicine, 19 (2000), pp. 2279–2294.

[42] T. P. Eisele, C. Mathews, M. Chopra, M. N. Lurie, L. Brown, S. Dewing, and C. Kendall, *Changes in risk behavior among HIV-positive patients during their first year of antiretroviral therapy in cape town south africa*, AIDS and Behavior, 13 (2009), pp. 1097–1105.

[43] M. C. Eisenberg, S. L. Robertson, and J. H. Tien, *Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease*, Journal of Theoretical Biology, (2013).

[44] W. M. El-Sadr, B. J. Coburn, and S. Blower, *Modeling the impact on the HIV epidemic of treating discordant couples with antiretrovirals to prevent transmission*, AIDS, 25 (2011), pp. 2295–2299.

[45] C.-T. Fang, H.-M. Hsu, S.-J. Twu, M.-Y. Chen, Y.-Y. Chang, J.-S. Hwang, J.-D. Wang, C.-Y. Chuang, et al., *Decreased HIV transmission after a policy of providing free access to highly active antiretroviral therapy in Taiwan*, Journal of Infectious Diseases, 190 (2004), pp. 879–885.

[46] J. M. Garcia-Callega, E. Gouws, and P. D. Ghys, *National population based HIV prevalence surveys in sub-Saharan Africa: results and implications for HIV and AIDS estimates*, Sexually Transmitted Infection, 82 (2006), pp. iii64–iii70.

[47] G. Garnett, S. Cousens, T. B. Hallett, R. Steketee, and N. Walker, *Mathematical models in the evaluation of health programmes*, The Lancet, 378 (2011), pp. 515–525.

[48] A. E. Gelfand and S. K. Sahu, *Identifiability, improper priors, and Gibbs sampling for generalized linear models*, Journal of the American Statistical Association, 94 (1999), pp. 247–253.

[49] P. D. Ghys, N. C. Grassly, G. Garnett, K. A. Stanecki, J. Stover, and N. Walker, *The UNAIDS Estimationg and Projection Package: a software package to estimate and project national HIV epidemics*, Sexually Transmitted Infections, 80 (2004), pp. i5–i9.

[50] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov chain Monte Carlo in practice*, vol. 2, CRC press, 1996.

[51] R. M. Granich, C. F. Gilks, C. Dye, K. M. De Cock, and B. G. Williams, *Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model*, The Lancet, 373 (2009), pp. 48–57.

[52] D. Greenhalgh, O. Diekmann, and M. de Jong, *Subcritical endemic steady states in mathematical models for animal infections with incomplete immunity*, Mathematical Biosciences, 165 (2000), pp. 1–25.

[53] M. Grewal and K. Glover, *Identifiability of linear and nonlinear dynamical systems*, Automatic Control, IEEE Transactions on, 21 (1976), pp. 833–837.

[54] H. Guo and M. Y. Li, *Global dynamics of a staged progression model for infectious diseases*, Mathematical Biosciences and Engineering, 3 (2006), p. 513.

[55] ——, *Global dynamics of a staged-progression model with amelioration for infectious diseases*, Journal of Biological Dynamics, 2 (2008), pp. 154–168.

[56] R. Gupta, A. Hill, A. W. Sawyer, and D. Pillay, *Emergence of drug resistance in HIV type 1-infected patients after receipt of first-line highly active antiretroviral therapy: a systematic review of clinical trials*, Clinical Infectious Diseases, 47 (2008), pp. 712–722.

[57] K. Hadeler and P. Van den Driessche, *Backward bifurcation in epidemic control*, Mathematical Biosciences, 146 (1997), pp. 15–35.

[58] K. P. Hadeler and C. Castillo-Chavez, *A core group model for disease transmission*, Mathematical Biosciences, 128 (1995), pp. 41–55.

[59] H. I. Hall, R. Song, P. Rhodes, J. Prejean, Q. An, L. M. Lee, J. Karon, R. Brookmeyer, E. H. Kaplan, M. T. McKenna, and J. R. S., *Estimation of HIV in the United States*, Journal of the American Medical Association, 300 (2008), pp. 520–529.

[60] I. H. Hall, T. A. Green, R. J. Wolitski, D. R. Holtgrave, P. Rhodes, J. S. Lehman, T. Durden, K. A. Fenton, and J. H. Mermin, *Estimated future HIV prevalence, incidence, and potential infections averted in the United States: a multiple scenario analysis*, Journal of Acquired Immune Deficiency Syndromes, 55 (2010), pp. 271–276.

[61] T. B. Hallett, *Estimating the HIV incidence rate – recent and future developments*, Current Opinion in HIV and AIDS, 6 (2011), pp. 102–107.

[62] T. B. Hallett, J. Stover, V. Mishra, P. D. Ghys, S. Gregson, and T. Boerma, *Estimates of HIV incidence from household-based prevalence surveys*, AIDS, 24 (2010), pp. 147–152.

[63] J. C. Helton and F. J. Davis, *Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems*, Reliability Engineering and System Safety, 81 (2003), pp. 23–69.

[64] J. C. Helton, J. D. Johnson, C. J. Sallaberry, and C. B. Storlie, *Survey of sampling-based methods for uncertainty and sensitivity analysis*, Reliability Engineering and System Safety, 91 (2006), pp. 1175–1209.

[65] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Review, 42 (2000), pp. 599–653.

[66] H. W. Hethcote and P. Van den Driessche, *Some epidemiological models with nonlinear incidence*, Journal of Mathematical Biology, 29 (1991), pp. 271–287.

[67] C. J. Hoffmann, S. Charalambous, J. Sim, J. Ledwaba, G. Schwikkard, R. E. Chaisson, K. L. Fielding, G. J. Churchyard, L. Morris, and A. D. Grant, *Viremia, resuppression, and time to resistance in human immunodeficiency virus (hiv) subtype c during first-line antiretroviral therapy*, Clinical Infectious Diseases, 49 (2009), pp. 1928–1935.

[68] L. F. Hohnson, T. B. Hallett, T. M. Rehle, and R. E. Dorrington, *The effect of changes in condom usage and antiretroviral treatment coverage on human immunodeficiency virus incidence in South Africa: a model-based analysis*, Journal of the Royal Society Interface, 9 (2012), pp. 1544–1554.

[69] J. A. C. Hontelez, S. J. de Vlas, F. Tanser, R. Bakker, T. Bärnighausen, M.-L. Newell, R. Baltussen, and M. N. Lurie, *The impact of the new WHO antiretroviral treatment guildelines on HIV epidemic dynamics and cost in South Africa*, PLoS One, 6 (2011), p. e21919.

[70] S. Houston, L. Mashinter, B. Rowe, M. Joffe, J. K. Preiksaitis, and G. Jhangri, *An anonymous unlinked seroprevalence study of HIV in urban Canadian emergency departments (abstract)*, Canadian Journal of Infectious Diseases & Medical Microbiology, 21 (2010), p. 62.

[71] S. Houston, B. H. Rowe, L. Mashinter, J. Preiksaitis, M. Joffe, D. Mackey, J. Galbraith, and N. Wiebe, *Sentinel surveillance of HIV and hepatitis C virus in two urban emergency departments.*, Canadian Journal of Emergency Medicine, 6 (2004), p. 89.

[72] W. Huang, K. L. Cooke, and C. Castillo-Chavez, *Stability and bifurcation for a multiple-group model for the dynamics of HIV/AIDS transmission*, SIAM Journal on Applied Mathematics, 52 (1992), pp. 835–854.

[73] J. M. Hyman, J. Li, and E. A. Stanley, *The differential infectivity and staged progression models for the transmission of HIV*, Mathematical Biosciences, 155 (1999), pp. 77–109.

[74] B. Ingalls, *Sensitivity analysis: from model parameters to system behaviour*, Essays in Biochemistry, 45 (2008), pp. 177–194.

[75] J. A. Jacquez, C. P. Simon, J. Koopman, L. Sattenspiel, and T. Perry, *Modeling and analyzing HIV transmission: the effect of contact patterns*, Mathematical Biosciences, 92 (1988), pp. 119–199.

[76] J. A. Jacquez, C. P. Simon, and J. S. Koopman, *The reproduction number in deterministic models of contagious diseases*, Current Topics in Theoretical Biology, 2 (1991), pp. 159–209.

[77] B. R. Jayasankar, A. Ben-Zvi, and B. Huang, *Identifiability and estimability study for a dynamic solid oxide fuel cell model*, Computers & Chemical Engineering, 33 (2009), pp. 484–492.

[78] R. Jewkes, M. Nduna, J. Levin, N. Jama, K. Dunkle, A. Puren, and N. Duvvury, *Impact of stepping stones on incidence of HIV and HSV-2 and sexual behaviour in rural South Africa: cluster randomized controlled trial*, BMJ, 337 (2008), p. a506.

[79] J. M. Karon, R. Song, R. Brookmeyer, H. Kaplan, Edward, and H. I. Hall, *Estimating HIV incidence in the united states from HIV/AIDS surveillance data and biomarker HIV test results*, Statistics in Medicine, 27 (2008), pp. 4617–4633.

[80] W. Kermack and A. McKendrick, *Contributions to the mathematical theory of epidemics–I*, Bulletin of Mathematical Biology, 53 (1991), pp. 33–55.

[81] A. A. Kim, T. Hallet, J. Stover, G. Eleanor, J. Muinguzi, P. K. Mureithi, R. Bunnell, J. Hargrove, J. Mermin, R. K. Kaiser, A. Barsigo, and P. D. Ghys, *Estimating HIV incidence amoung adults in Kenya and Uganda: a systematic comparison of multiple methods*, PLoS One, 6 (2011), p. e17535.

[82] C. M. Kribs-Zaleta and M. Martcheva, *Vaccination strategies and backward bifurcation in an age-since-infection structured model*, Mathematical Biosciences, 177 (2002), pp. 317–332.

[83] Y. V. Kruglov, Y. V. Kobyshcha, T. Salyuk, O. Varetska, A. Shakarishvili, and V. P. Saldanha, *The most severe HIV epidemic in Europe: Ukraine's national HIV prevalence estimates for 2007*, Sexually Transmitted Infection, 84 (2008), pp. i37–i41.

[84] S. Kumta, M. Lurie, S. Weitzen, H. Jerajani, A. Gogate, A. Rowkavi, V. Anand, H. Makadon, and K. H. Mayer, *Bisexuality, sexual risk taking, and HIV prevalence among men who have sex with men accessing voluntary counseling and testing services in Mumbai, India*, Journal of Acquired Immune Deficiency Syndromes, 53 (2010), p. 227.

[85] T. Lane, H. F. Raymond, S. Dladla, J. Rasethe, H. Struthers, W. McFarland, and J. McIntyre, *High HIV prevalence among men who have sex with men in Soweto, South Africa: results from the Soweto men's study*, AIDS and Behavior, 15 (2011), pp. 626–634.

[86] A. J. Leigh Brown, S. D. Frost, W. C. Mathews, K. Dawson, N. S. Hellmann, E. S. Daar, D. D. Richman, and S. J. Little, *Transmission*

fitness of drug-resistant human immunodeficiency virus and the prevalence of resistance in the antiretroviral-treated population, Journal of Infectious Diseases, 187 (2003), pp. 683–686.

[87] G. Li and W. Wang, *Bifurcation analysis of an epidemic model with nonlinear incidence*, Applied Mathematics and Computation, 214 (2009), pp. 411–423.

[88] Y. Ling and S. Mahadevan, *Quantitative model validation techniques: new insights*, Reliability Engineering & System Safety, (2012).

[89] W.-m. Liu, H. W. Hethcote, and S. A. Levin, *Dynamical behavior of epidemiological models with nonlinear incidence rates*, Journal of Mathematical Biology, 25 (1987), pp. 359–380.

[90] W. Lu, G. Zeng, L. Jing, S. Duo, G. Xing, D. Guo-wei, Z. Jian-ping, H. Wen-sheng, and W. Ning, *HIV transmission risk among serodiscordant couples: a retrospective study of former plasma donors in Henan, China*, Journal of Acquired Immune Deficiency Syndromes, 55 (2010), p. 232.

[91] R. Lyerla, E. Gouws, J. M. Garcia-Calleja, and E. Zaniewski, *The 2005 workbook: an improved tool for estimating HIV prevalence in countries with low level and concentrated epidemics*, Sexually Transmitted Infection, 82 (2006), pp. iii41–iii44.

[92] K.-A. Mallitt, D. P. Wilson, and H. Wand, *Is back-projection methodology still relevant for estimating incidence from national surveillance data*, The Open AIDS Journal, 6 (2012), pp. 108–111.

[93] S. Marino, I. B. Bogue, C. J. Ray, and D. E. Kirschner, *A methodology for performing global uncertainty and sensitivity analysis in systems biology*, Journal of Theoretical Biology, 254 (2008), pp. 178–196.

[94] M. Martcheva and H. R. Thieme, *Progression age enhanced backward bifurcation in an epidemic model with super-infection*, Journal of Mathematical Biology, 46 (2003), pp. 385–424.

[95] MATLAB, *version 7.7.0.471 (R2008b)*, The MathWorks Inc., Natick, Massachusetts, 2008.

[96] C. C. McCluskey, *A model of HIV/AIDS with staged progression and amelioration*, Mathematical Biosciences, 181 (2003), pp. 1–16.

[97] C. A. McGarrigle, S. Cliffe, A. J. Copas, C. H. Mercer, D. DeAngelis, K. A. Fenton, B. G. Evans, A. M. Johnson, and O. N. Gill, *Estimating adult HIV prevalence in the uk in 2003: the direct method of estimation*, Sexually Transmitted Infection, 82 (2006), pp. iii78–iii86.

[98] V. Mishra, M. Vaessen, J. T. Boerma, F. Arnold, A. Way, B. Barrere, A. Cross, R. Hong, and J. Sangha, *HIV testing in national population-based surveys: experience from the demographic and health surveys*, Bulletin of the World Heath Organisation, 84 (2006), pp. 537–545.

[99] J. S. Montaner, V. D. Lima, R. Barrios, B. Yip, E. Wood, T. Kerr, K. Shannon, P. R. Harrigan, R. S. Hogg, P. Daly, et al., *Association of highly active antiretroviral therapy coverage, population viral load, and yearly new HIV diagnoses in British Columbia, Canada: a population-based study*, The Lancet, 376 (2010), pp. 532–539.

[100] D. Morgan, C. Mahe, B. Mayanja, J. M. Okongo, R. Lubega, and J. A. Whitworth, *HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries?*, AIDS, 16 (2002), pp. 597–603.

[101] J. D. A. Ndawinz, D. Costagliola, and V. Supervie, *New method for estimating HIV incidence and time from infection to diagnosis using HIV surveillance data*, AIDS, 25 (2011), pp. 1905–1913.

[102] T. Q. Nguyen, C. R. Gwynn, S. E. Kellerman, E. Begier, R. K. Garg, M. R. Pfeiffer, K. J. Konty, L. Torian, T. R. Frieden, and L. E. Thorpe, *Population prevalence of reported and unreported HIV and related behaviors amoung the household adult population in New York City, 2004*, AIDS, 22 (2008), pp. 281–287.

[103] L. M. Niccolai, V. Toussova Olga, S. V. Verevochkin, R. Borbour, R. Heimer, and A. P. Kozlov, *High HIV prevalence, suboptimal HIV testing and low knowledge of HIV-positive serostatus among injection drug users in St. Petersburg, Russia*, AIDS Behavior, 14 (2010), pp. 932–941.

[104] C. Orrell, R. P. Walensky, E. Losina, J. Pitt, K. A. Freedberg, and R. Wood, *HIV-1 clade c resistance genotypes in naïve patients and after first virological failure in a large community art programme*, Antiviral Therapy, 14 (2009), p. 523.

[105] B. S. Parekh, D. L. Hanson, J. Hargrove, B. Branson, T. Green, T. Dobbs, N. Constantine, J. Overbaugh, and S. McDugal, *Determination of mean recency period for estimation of HIV type 1 incidence with the BED-Capture EIA in persons infected with diverse subtypes*, AIDS Research and Human Retroviruses, 27 (2011), pp. 265–273.

[106] S. S. Plitt, A. E. Singh, B. E. Lee, and J. K. Preiksaitis, *HIV seroprevalence among women opting out of prenatal HIV screening in Alberta, Canada: 2002–2004*, Clinical Infectious Diseases, 45 (2007), pp. 1640–1643.

[107] J. Prejean, R. Song, A. Hernandez, R. Ziebell, T. Green, F. Walker, L. S. Lin, Q. An, J. Mermin, A. Lansky, and H. I. Hall, *Estimated HIV incidence in the United States, 2006–2009*, PLoS One, 6 (2011), p. e17502.

[108] A. M. Presanis, O. N. Gill, T. R. Chadborn, C. Hill, V. Hope, L. Locan, B. D. Rice, V. C. Delpech, A. E. Ades, and D. De Angelis, *Insights into the rise in HIV infections, 2001 to 2008: a Bayesian synthesis of prevalence evidence*, AIDS, 24 (2010), pp. 2849–2858.

[109] R. Qesmi, J. Wu, J. Wu, and J. M. Heffernan, *Influence of backward bifurcation in a model of hepatitis B and C viruses*, Mathematical Biosciences, 224 (2010), pp. 118–125.

[110] T. C. Quinn, M. J. Wawer, N. Sewankambo, D. Serwadda, C. Li, F. Wabwire-Mangen, M. O. Meehan, T. Lutalo, and R. H. Gray, *Viral load and heterosexual transmission of human immunodeficiency virus type 1*, New England Journal of Medicine, 342 (2000), pp. 921–929.

[111] B. M. Ramesh, S. Moses, R. Washington, S. Isac, B. Mohapatra, S. B. Mahagaonkar, R. Adhikary, G. N. Brahmam, R. S. Paranjape, T. Subramanian, et al., *Determinants of HIV prevalence among female sex workers in four south Indian states: analysis of cross-sectional surveys in twenty-three districts*, AIDS, 22 (2008), pp. S35–S44.

[112] R. Rebba and S. Mahadevan, *Computational methods for model reliability assessment*, Reliability Engineering & System Safety, 93 (2008), pp. 1197–1207.

[113] T. Rehle, O. Shisana, V. Pillay, K. Zuma, A. Puren, and W. Parker, *National HIV incidence measures – new insights into the South African epidemic*, South African Medical Journal, 97 (2007), pp. 194–199.

[114] T. M. Rehle, T. B. Hallett, O. Shisana, V. Pillay-van Wyk, K. Zuma, H. Carrara, and S. Jooste, *A decline in new HIV infections in South Africa: estimating HIV incidence from three national HIV surveys in 2002, 2005, and 2008*, PLoS One, 5 (2010), p. e11094.

[115] D. N. Shaffer, I. K. Ngetich, C. T. Bautista, F. K. Sawe, P. O. Renzullo, P. T. Scott, R. M. Kibaya, K. O. Imbuki, N. L. Michael, D. L. Birx, M. K. Wasunna, and M. L. Robb, *HIV-1 incidence rates and risk factors in agricultural workers and dependents in rural Kenya: 36-month follow-up of the Kericho HIV cohort study*, Journal of Acquired Immune Deficiency Syndromes, 53 (2010), pp. 514–521.

[116] O. Sharomi, C. Podder, A. Gumel, E. Elbasha, and J. Watmough, *Role of incidence function in vaccine-induced backward bifurcation in some HIV models*, Mathematical Biosciences, 210 (2007), pp. 436–463.

[117] B. W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hall, 1986.

[118] S. Srinath and R. Gunawan, *Parameter identifiability of power-law biochemical system models*, Journal of Biotechnology, 149 (2010), pp. 132–140.

[119] Statistics Canada, *Table 051-0001 – estimates of population, by age group and sex for July 1, Canada, provinces and territories*. CANSIM (database). (accessed: 2013-06-27).

[120] ——, *Table 051-0002 – estimates of deaths, by sex and age group, Canada, provinces and territories, annual (persons)*. CANSIM (database). (accessed: 2013-06-27).

[121] J. Stover, T. Brown, and M. Marston, *Updates to the Spectrum/Estimation and Projection Package (EPP) model to estimate HIV trends for adults and children*, Sexually Transmitted Infections, 88 (2012), pp. i11–i16.

[122] J. Stover, P. Johnson, T. Hallett, M. Marston, R. Becquet, and I. M. Timaeus, *The Spectrum projection package: improvements to estimating incidence by age and sex, mother-to-child transmission, HIV progression in children and double orphans*, Sexually Transmitted Infections, 86 (2010), pp. ii6–ii21.

[123] M. J. Sweeting, D. De Angelis, and O. O. Aalen, *Bayesian back-calculation using a multi-state model with application to HIV*, Statistics in Medicine, 24 (2005), pp. 3991–4007.

[124] F. Tanser, T. Bärnighausen, L. Hund, G. P. Garnett, N. McGrath, and M.-L. Newell, *Effect of concurrent sexual partnerships on a rate of new HIV infections in a high-prevalence, rural South African population: a cohort study*, The Lancet, 378 (2011), pp. 247–255.

[125] H. R. Thieme, *Mathematics in Population Biology*, Princeton University Press, 2003.

[126] H. R. Thieme and Z. Feng, *Endemic models with arbitrarily distributed periods of infection i: fundamental properties of the model*, SIAM Journal on Applied Mathematics, 61 (2000), pp. 803–833.

[127] L. Topp, C. A. Day, J. Iversen, H. Wand, L. Maher, C. of Australian NSPs, et al., *Fifteen years of HIV surveillance among people who inject*

drugs: the Australian Needle and Syringe Program Survey 1995–2009, AIDS, 25 (2011), pp. 835–842.

[128] D. Turner, B. Brenner, J.-P. Routy, D. Moisi, Z. Rosberger, M. Roger, and M. A. Wainberg, *Diminished representation of HIV-1 variants containing select drug resistance-conferring mutations in primary HIV-1 infection*, Journal of Acquired Immune Deficiency Syndromes, 37 (2004), pp. 1627–1631.

[129] UNAIDS Reference Group on Estimates, Modelling and Projections, *Improved methods and assumptions for estimation of the HIV/AIDS epidemic and its impact: recommendations of the UNAIDS Reference Group on Estimates, Modelling and Projections*, AIDS, 16 (2002), pp. W1–W14.

[130] P. Van den Driessche, *Time delay in epidemic models*, IMA Volumes in Mathematics and its Applications, 125 (2002), pp. 119–128.

[131] P. Van den Driessche and J. Watmough, *A simple SIS epidemic model with a backward bifurcation*, Journal of Mathematical Biology, 40 (2000), pp. 525–540.

[132] P. Van den Driessche and J. Watmough, *Reproduction numbers and subthreshold endemic equilibria for compartmental models of disease transmission*, Mathematical Biosciences, 180 (2002), pp. 29–48.

[133] M. G. van Veen, A. M. Presanis, S. Conti, M. Xiridou, A. R. Stengaard, M. C. Donoghoe, A. I. van Sighem, M. A. ven der Sande, and D. De Angelis, *National estimate of HIV prevalence in the Netherlands: comparison and applicability of different estimation tools*, AIDS, 25 (2011), pp. 229–237.

[134] P. Vickerman, F. Ndowa, N. O'Farrell, R. Steen, M. Alary, and S. Delany-Moretlwe, *Using mathematical modelling to estimate the impact of periodic presumptive treatment on the transmission of sexually transmitted infections and HIV among female sex workers*, Sexually Transmitted Infections, 86 (2010), pp. 163–168.

[135] R. P. Walensky, A. D. Paltiel, E. Losina, B. L. Morris, C. A. Scott, E. R. Rhone, G. R. Seage, and K. A. Freedberg, *Test and treat DC: forecasting the impact of a comprehensive HIV strategy in Washington DC*, Clinical Infectious Diseases, 51 (2010).

[136] C. L. Wallis, J. W. Mellors, W. D. Venter, I. Sanne, and W. Stevens, *Varied patterns of HIV-1 drug resistance on failing first-line antiretroviral therapy in south africa*, Journal of Acquired Immune Deficiency Syndromes, 53 (2010), pp. 480–484.

[137] E. Walter, *Identifiability of State Space Models*, Springer-Verlag, 1982.

[138] H. Wan and H. Zhu, *The backward bifurcation in compartmental models for West Nile virus*, Mathematical Biosciences, 227 (2010), pp. 20–28.

[139] H. Wand, P. Yan, D. Wilson, A. McDonald, M. Middleton, J. Kaldor, and M. Law, *Increasing HIV transmission through male homosexual and heterosexual contact in Australia: results from an extended back-projection approach*, HIV Medicine, 11 (2010), pp. 395–403.

[140] N. Wang, L. Wang, Z. Wo, W. Gou, X. Sun, K. Poundstone, and Y. Wang, *Estimating the number of people living with HIV/AIDS in China: 2003–2009*, International Journal of Epidemiology, 39 (2010), pp. ii21–ii28.

[141] A. Welte, T. A. McWalter, O. Laeyendecker, and T. B. Hallet, *Using tests for recent infection to estimate incidence: problems and prospects for HIV*, Euro Surveillance, 15 (2010).

[142] D. P. Wilson, A. Hoare, D. G. Regan, and M. G. Law, *Importance of promoting HIV testing for preventing secondary transmissions: modelling the Australian HIV epidemic among men who have sex with men*, Sexual Health, 6 (2009), pp. 19–33.

[143] E. Wood, T. Kerr, B. D. Marshall, K. Li, R. Zhang, R. S. Hogg, P. R. Harrigan, and J. S. Montaner, *Longitudinal community plasma HIV-1 RNA concentrations and incidence of HIV-1 among injecting drug users: prospective cohort study*, BMJ, 338 (2009).

[144] Working Group on Estimation of HIV Prevalence in Europe, *HIV in hiding: methods and data requirements for the estimation of the number of people living with undiagnosed HIV*, AIDS, 25 (2011), pp. 1017–1023.

[145] World Health Organisation, *Guidelines for second generation HIV surveillance: an update: know your epidemic.*, tech. rep., World Health Organisation, 2013.

[146] H. Wu, H. Zhu, H. Miao, and A. S. Perelson, *Parameter identifiability and estimation of HIV/AIDS dynamic models*, Bulletin of Mathematical Biology, 70 (2008), pp. 785–799.

[147] D. Xiao and S. Ruan, *Global analysis of an epidemic model with nonmonotone incidence rate*, Mathematical Biosciences, 208 (2007), pp. 419–429.

[148] P. Yan, F. Zhang, and H. Wand, *Using HIV diagnostic data to estimate HIV incidence: method and simulation*, Statistical Communications in Infectious Diseases, 3 (2011).

[149] Q. Yang, D. Boulos, F. Zhang, R. S. Remis, D. Schanzer, and C. P. Archibald, *Estimates of the number of prevalent and incident human immunodeficiency virus (HIV) infections in Canada, 2008*, Canadian Journal of Public Health, 101 (2010), pp. 486–490.

Appendix A

Implementation Details

A.1 Tracking New Cases and Deaths

Computing the integrals in equation (2.4) numerically can be greatly simplified by adding some additional compartments to the model. Recall that the observation function consists of three components:

$$\begin{aligned} y_{diag}(t) &= \int_{t-1}^{t} \alpha I(\tau) d\tau \\ y_{death}(t) &= \int_{t-1}^{t} d_D D(\tau) + d_T T(\tau) d\tau \\ y_{pop}(t) &= S(t) + I(t) + D(t) + T(t). \end{aligned} \tag{A.1}$$

As written, it would be necessary to perform a numerical integration to determine the values of $y_{diag}(t)$ and $y_{death}(t)$. However, if quantities $z_{diag}(t)$ and $z_{death}(t)$ are defined as:

$$\begin{aligned} z_{diag}(t) &= \int_{t_0}^{t} \alpha I(\tau) d\tau \\ z_{death}(t) &= \int_{t_0}^{t} d_D D(\tau) + d_T T(\tau) d\tau \end{aligned} \tag{A.2}$$

The fundamental theorem of calculus gives

$$\begin{aligned} \frac{dz_{diag}(t)}{dt} &= \alpha I(t) \\ \frac{dz_{death}(t)}{dt} &= d_D D(t). \end{aligned} \tag{A.3}$$

These equations can be included when the model is solved numerically, computing $z_{diag}(t)$ and $z_{death}(t)$ simultaneously with the other modelled compartments. The implementation used for this project includes only the compartment for those who have been removed from the diagnosed population through death, $z_{death}(t)$. Setting the initial condition $z_{death}(t_0) = 0$, this quantity tracks the total number of people who have

died while in compartment D since the initial time. Then $z_{diag}(t) = D(t) + z_{death}(t)$ is the total number of people who have ever been diagnosed since the initial time. Finally, $y_{diag}(t) = z_{diag}(t) - z_{diag}(t-1)$ gives the number of diagnoses during year t and $y_{death}(t) = z_{death}(t) - z_{death}(t-1)$ gives the number of deaths during year t as required.

A.2 Sampling

A.2.1 Metropolis - Hastings Sample

The Bayesian parameter estimation used in Chapter 2 of this thesis is implemented using the Metropolis - Hastings method to draw a sample from the posterior distribution. The routines used are included in MATLAB. In this section, we give a brief introduction to the Metropolis-Hastings method.

The Metropolis - Hastings algorithm is a Markov chain Monte Carlo (MCMC) method. Using a specified starting point p_0, a new sample point p_1 is selected from a specified *proposal distribution*. This new sample point is either accepted or rejected based on both the proposal distribution and the posterior distribution that is being sampled. Subsequent sample points p_i are selected using p_{i-1} as the starting point.

The acceptance ratio used to make the decision to accept or reject a new sample point is

$$A(p_0, p_1) = \frac{P(p_1|y)f(p_0|p_1)}{P(p_0|y)f(p_1|p_0)} \tag{A.4}$$

where $P(p|y)$ is the posterior distribution to be sampled, and $f(p_1|p_0)$ is the proposal distribution given the starting point p_0. The point p_1 is accepted as an element of the sample with probability

$$min\left(1, A(p_0, p_1)\right).$$

If the proposal distribution $f(p_1|p_0)$ is symmetric so that $f(p_1|p_0) = f(p_0|p_1)$, then the acceptance ratio simplifies to

$$A(p_0, p_1) = \frac{P(p_1|y)}{P(p_0|y)}.$$

In this case, if the point p_1 has a higher posterior density than the point p_0, it will always be accepted. If it has a lower posterior density, it will be accepted with probability $\frac{P(p_1|y)}{P(p_0|y)}$.

It is convenient that any factors in $P(p|y)$ that do not depend on p will be cancelled out in expression (A.4). This property allows computationally intensive normalizing factors to be omitted from the expressions used for $P(p|y)$.

Implementing the Metropolis - Hastings algorithm requires that the proposal distribution $f(p_1|p_0)$ be specified. This distribution controls how the size of the jumps that the algorithm will take between sample points. If the size of the jumps is very small, most of the proposed sample points will be accepted, but it will take a long time for

the algorithm to adequately sample the entire space. On the other hand, if the size of the jumps is too large, points with very low posterior density will often be proposed and very few of them will be accepted. The proposal distribution should be chosen to balance these concerns.

To reduce the amount of correlation between subsequent points, a Metropolis - Hastings sample is often "thinned" by sampling T times the number of points desired in the final sample and retaining only every Tth point. Furthermore, the first sample points chosen using an MCMC method may need to be discarded, as it may take some time for the method to converge to the desired distribution.

A.2.2 Latin Hypercube Sample

Latin hypercube samples are useful for taking samples from relatively simple distributions. This type of sample is used in Section 5.1 where distributions for parameters are specified using medical or public health literature. In this case, the specified distributions are usually independent for each variable with relationships between the parameters enforced through reparameterization. Routines are included in the standard MATLAB distribution to take Latin hypercube samples for independent normal and uniform distributions. To use other distributions, such as the triangular distribution often used in this thesis, customized code must be written.

Latin hypercube sampling is usually less computationally intensive to produce than the Metropolis-Hastings samples described in the previous section. Latin hypercube samples often give good coverage of the distribution with relatively few sample points. However, Latin hypercube sampling is not able to sample from distributions where the relationships between the variables are complicated or not well understood such as Bayesian posteriors for which a Metropolis-Hastings sample can be used.

A Latin hypercube sample is chosen by stratified random sampling without replacement. A description of the method is given in [16]. The method to produce a sample of size n is as follows

1. For each variable to be sampled
   (a) Specify the distribution.
   (b) Divide the distribution into n equiprobable intervals and sample one point from each of the intervals.
   (c) Randomly reorder the sample.
2. Join the samples for the individual variables into a single multidimensional sample.

When writing custom MATLAB code to create such a sample, integration is often required in step 1(b). For a simple cases such as the triangular distribution this can be done analytically. For more complicated distributions numerical integration may be required.

A.3 Stability of Endemic Equilibria

We consider the special case when the recovered class suffers no disease-related fatality (full recovery), namely $d_S = d_R$.

Proposition 4. *Assume that $d_S = d_R = d$. Let $P^* = (S^*, I^*, R^*)$ be an endemic equilibrium of (5.2). Then P^* is asymptotically stable if $\frac{dg(S^*)}{dS} > 0$, and unstable if $\frac{dg(S^*)}{dS} < 0$.*

Proof. The Jacobian matrix J at P^* is given by

$$\begin{bmatrix} -\beta If(N) - \beta SIf'(N) - d & -\beta Sf(N) - \beta SIf'(N) & -\beta ISf'(N) \\ \beta If(N) + \beta SIf'(N) & \beta Sf(N) + \beta SIf'(N) - (d_I + \gamma) & \beta ISf'(N) \\ 0 & \gamma & -d \end{bmatrix}$$

with superscripts suppressed. Since $\beta S^* f(N^*) = d_I + \gamma$

$$J = \begin{bmatrix} -\beta If(N) - \beta SIf'(N) - d & -(d_I + \gamma) - \beta SIf'(N) & -\beta ISf'(N) \\ \beta If(N) + \beta SIf'(N) & \beta SIf'(N) & \beta ISf'(N) \\ 0 & \gamma & -d \end{bmatrix}.$$

Therefore

$$\begin{aligned} &\det(J - \mu I) \\ &= (-d - \mu)[\mu^2 + (d + \beta If(N))\mu + \beta(If(N)(d_I + \gamma) + (d_I - d)ISf'(N))]. \end{aligned}$$

One eigenvalue is $\mu_1 = -d < 0$, while the remaining two eigenvalues are the solutions of the quadratic equation

$$\mu^2 + \Big(d + \beta If(N)\Big)\mu + \beta If(N)(d_I + \gamma) + (d_I - d)\beta ISf'(N) = 0.$$

Since $d + \beta If(N) > 0$, a necessary and sufficient condition for both solutions to have negative real parts is that

$$\beta If(N) + \sigma(d_I - d)ISf'(N) > 0.$$

Using the facts that $\sigma(d_I - d) = \beta(1 - pd)$ and

$$\frac{dg(S)}{dS} = f(N(S)) + Sf'(N(S))(1 - pd),$$

we obtain

$$\beta If(N) + \sigma(d_I - d)ISf'(N) = \beta I\frac{dg(S)}{dS}.$$

Therefore, if $\frac{dg(S)}{dS} > 0$, then all the eigenvalues have negative real parts, and the endemic equilibrium is stable. If $\frac{dg(S)}{dS} < 0$, then one of the eigenvalues is positive, and the endemic equilibrium is unstable.

□