國立臺灣大學公共衛生學院
流行病學與預防醫學研究所
碩士論文
Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University
Master Thesis

使用動態貝氏網路建立傳染病個體化模擬模型：以肺結核介入政策為例
Using Dynamic Bayesian Networks for Agent-Based Modelling: Application in Tuberculosis Control

辜鉅璋
Chu-Chang Ku

指導教授：林先和博士
Advisor: Hsien-Ho Lin, Sc.D.

中華民國 103 年 7 月
July, 2014

致謝

硕士論文是一個漫長如破蛹的挑戰。過程中，在方法學及應用間的擺盪帶來許多苦水但也有不少甘霖。這路上陪伴及同行的每個人我都萬分感謝。

故事開始於去年初，蕭老師鼓勵我多走一些不同的路，才讓方法背景的我有機會沾上較偏向具體應用的題目。在林老師的研究室裡，一開始是怕怕的，因為我感覺我好像只能當個工具、不知道自己能有什麼樣的定位。對臨床的不熟悉以及面對本來就複雜的肺結核，好幾度讓我不知如何前進。在和林老師一次又一次的討論後，才慢慢感覺到自己可以做的事及可以怎麼做。感謝老師忍受我這個麻煩；體諒我的難控制；在如家常便飯的拔河中，讓我在站穩在稜線上。也請老師原諒我常常一不小心就用和朋友的模式和老師對話。

感謝我原生 LAB 的家長蕭老師，為我的研究能力打了不少的基礎。蕭老師一路上一直為我的未來所學所用作設想。儘管轉籍了，也願意讓我約時間討論。如果一直持在同一間研究室也許我可以因著專長完成一篇更有力的論文，但這意味著這一年看到的景色不會出現在我生命裡，我將更難離開我的舒適圈。

謝謝方老師願意讓我這個外人旁聽 Group meeting，在聽同學報告及老師講解的過程中有很多的學習。這讓過去較少修臨床、實務相關課程的我可以快速對傳染病有深刻的認識，也讓我試著去思考在一個政策裡各種角色上的人可能會有什麼樣的聲音。感謝方老師願意讓我對知識衝撞也認真的回應我的衝撞並使我的視野擴大。

身邊的伙伴們我更是感謝。和我互利共生 (雖然我好像佔比較多便宜) 的傳涵學姐，在好幾個的碩研室的日子裡無數的討論及有助身心的閒聊；威利學長、LULU 學姐、亭君學姐、偉成學長、陳博士常給我各種的問題與建議；小伍、洛彤、政由與嘉珍同為傳染病領域的學伴們，一起學習、互相打氣；當然還有與我同甘共苦的藏鏡人，藍婷。最後感謝上帝的陪伴，不管是起是落都不離不棄的同在。聖經裡的文字不只是為我打氣還為我的方法有些啟發。

然而，碩士論文只是個引信，綻放的時刻在未來。

中文摘要

傳染病模擬模型在流行病學中被用來探索一些在現實中難以探究的問題。其中，個體化模擬模型 (Agent-based model) 利用在電腦中的虛擬個體模擬由複雜行為組成的系統。近年來，由於電腦運算技術的進步，個體化模擬模型有許多的應用產生，然而對於如何擬合與校正個體化模擬模型的研究甚少。本研究利用連續時間貝氏網路 (Continuous-time Bayesian Networks) 發展了一組具有統計界面的傳染病個體化模擬模型，並進一步以過去的擬和架構為基礎，發展出一套擬合程序。我們成功將遺傳演算法中的數值點突變 (Numerical mutation) 及參數分組策略 (Blocking strategy) 應用於序列蒙地卡羅法 (Sequential Monte Carlo) 中，使擬合程序可以處理大量參數且來源各異的資料。最後，我們以肺結核的接觸者追蹤政策為例，使用易感受 -感染者 -復原者模型 (Susceptible-Infectious-Recovery model) 來演示我們為個體化模擬模型從模型建構、估計到預測所發展的實證架構。

關鍵字：個體化模擬模型、傳染病數理模型、動態貝氏網路、數值突變、肺結核、接觸者追蹤

Abstract

The simulation models in epidemiology were developed to answer the questions which were not easy to solve by observational studies in the real world. In particular, Agent-based models (ABMs) were usually employed to deal with the complex system of disease transmission by simulating computational agents in the virtual world. However, the fitting scheme of ABMs is less developed than the applications.. With the aim of investigating disease dynamics and creating an interface for statistical analysis, we proposed a class of ABMs with Continuous-time Bayesian network, a temporal multivariate probability model. While retaining the strength of existing procedure for simulation model fitting based on sequential Monte Carlo, we set up an improved framework for fitting ABMs. We further synthesized the numerical mutation in genetic algorithm and the parameters augmentation in blocking Gibbs sampling in order to overcome the challenges of multidimensional parameters and multi-sources data. Using an example of Susceptible-Infectious-Recovery model for contact tracing in tuberculosis control, we briefly presented the properties of our proposed model and demonstrated its potential applications in the future. By including model construction, fitting, and forecasting, we formalized an empirical scheme for individual based models in simulating disease dynamics.

Key words: Agent-based model, Mathematical model for Infectious Disease, Continuous-time Bayesian Networks, Numerical mutation, Blocking Gibbs sampling, Tuberculosis, Contact tracing

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Modern challenge of infectious disease control

Infectious diseases occupy an important position in human history. In developing countries, infectious diseases cause the majority of deaths each year. On the other hand, infectious diseases were not a threat to developed countries any more due to social progress and medical development. However, many new challenges of infectious diseases have arisen in recent decades. First of all, the appearance of emerging infectious diseases such as severe acute respiratory syndromes (SARS) and avian influenza revealed a potential impact on global epidemic from fast transmission(Meltzer et al., 1999). Second, drug use and physical activities were known as the risk factors of non-communicable disease but were proved to have the communicable feature. The individual-based interventions can have better effect if the dependence of risk factors among the population have been considered. It showed that the evaluation for interventions based on group level should be considered and the statistical method based on independent and identically distributed (i.i.d.) might be malfunctioned. The change of human contact may open a new way for disease spreading while the traditional epidemiological studies, field works and laboratory experiments were not enough to overcome these challenges. Therefore, upgrading the methods for assessing interventions of complex infectious diseases were in demand.

1.2 Why simulation model?

In the last few decades, epidemiologists have used dynamic models to simulate the infectious diseases under different settings, in order to compensate the disadvantages of observational studies(Grassly and Fraser, 2008). Simulation models (also known as mathematical models) of infectious diseases create virtual scenarios to capture the phenomenon of disease transmission in the real world. They are able to deal with the difficulties from time cost, ethical issues, and limited budget in observational studies. Moreover, even we have the scientific knowledge from epidemiological research, the information for policy analysis is often insufficient(Pearl, 2014). In such situation, simulation models were more tractable because they compared different policies through counter-factual experiments. Ideally, simulation model could integrate the knowledge from many sources and assumed the practical scenarios to help researchers and decision makers get valid inference.

1.2.1 Agent-based models (ABMs)

Agent-based models transformed the elements in a real system into the agents and let agents have individual behaviours in the virtual world(Macal and North, 2005). Agent-based models are bottom-up constructed simulation model which form the system from atomic attributions, behaviour, agents, and interactions to the whole population. The behaviours were depicted in terms of scripts, describing the actions by specific order or at specific impulse. ABMs simulated the real world system both on individual and population level and the properties of interest could be observed directly in the model. Because ABMs provided a flexible way to model a system and provide a new scheme for thinking questions, the applications of ABM have grown exponentially in many fields and disciplines, including ecology(Doran, 2001, Grimm and Railsback, 2013, Janssen, 2002), cytology(Deisboeck et al., 2011), public health(Auchincloss and Roux, 2008, Galea et al., 2010, Maglio and Mabry, 2011), and social science(Epstein, 1999, Farmer and Foley, 2009). In fact, ABM-based inference was more intuitive and it appeared earlier than other simulation models. In an early stage, the popularity of ABM was limited by the lack of standard and simple processes for application. Recently, the studies with ABM emerged because of the demands for dealing with complex issues and the upgrade of computational techniques. Generally, the most important advantage of ABM is its flexibility in model construction. With this advantage, the structure of ABM should sophistically reflect the complexity of study question(Parunak et al., 1998). The dynamics of an infectious disease usually involve the behaviours at different levels; and the knowledge and assumptions of disease transmission are also described in different hierarchy. For example, the disease progression is usually discussed at the individual level; the contact process happens at the sub-group level; and the intervention policy is designed at the population level. While the single process of disease transmission can be investigated by a simple epidemiological survey, the underlying disease dynamics were difficult to capture. ABMs were able to demonstrate the complex epidemic and examine the fitness of model in different levels. Moreover, as Farmer and Foley (2009) mentioned, the traditional equation-based models were not suitable when the dynamic was out of equilibrium. Instead, the construction of ABM for infectious diseases can help us find out the hidden information behind the compartmental models.

1.3 Challenges in ABMs construction for epidemiologist

Although ABMs could occupy a special niche in epidemiology and could be applied in many sub-fields of epidemiology, epidemiologists are still not familiar with them. In epidemiology the top-down thinking is the main stream. The population is usually stratified and the specific behaviours are observed in each stratum or sub-group in order to identify specific risk factors for disease occurrence. It is different from the construction of ABMs. In addition, studies of ABMs are usually conducted with language for researcher of computer science background (i.e. UML (unified modelling language)) which is not included in the formal training of epidemiology. The gap kept the ABMs for disease from judgments and checking by others. That is, people can read the input and output of an ABM study but don't know how the model works. Fortunately, Bayesian networks (BNs) (Pearl, 1988)and directed acyclic graph (causal diagram) are unified languages for both computer science and epidemiology. BNs are introduced to epidemiology from machine learning in 1999 (Greenland et al., 1999). It clarified the definitions of many elements epidemiology, such as confounding and selection bias. Intuitively, the simulation models formed by BNs could be mapped to the epidemiological system easily.

1.4 Fitting scheme and research gap

Bayesian approach was widely used in simulation models because the integration of information from different sources were essential in epidemic modelling(Jewell et al., 2009). In Bayesian approach, the estimation of posterior distribution was based on two principle factors, the distance between observed data and model (i.e. likelihood) and knowledge of the parameters (i.e. prior distribution). However, the likelihood function in simulation models was relatively difficult to obtain. Approximate Bayesian computation has overcome this challenge by generating a pseudo likelihood to approximate the exact likelihood function(Toni et al., 2009). On the other hand, pattern-orients methods obtained the posterior distribution by pattern matching without exact likelihoods(Grimm et al., 2005).

The method proposed by previous studies have provided general procedures for most of simulation models either in deterministic or in stochastic models(Hartig et al., 2011). However, the general fitting procedure for simulation models might be insufficient for ABMs. With the rich properties in describing disease transmission at different levels, the statistical procedure of ABMs faced special challenges.

The first challenge was the multidimensional parameters applied in ABMs. The behaviours or state transition were composed of a set of parameters. Therefore, the estimation given less data usually introduced non-identifiability (spurious correlation). Second, the input data were multidimensional and came from various sources with different reliability based on study designs. Epidemiological knowledge can be collected from randomized control trails, cohort studies and case-control studies; the prevalence data were observed from cross-sectional surveys; the specific descriptions of behaviour were gathered form case studies. By improving the limitation of dealing with multidimensional parameters and data, ABMs can be a promising approach in simulating the epidemic under complex settings (Blower and Go, 2011).

1.5 Example: Tuberculosis control in Taiwan

Tuberculosis (TB) is an endemic infectious disease which has great impacts on public health and human productivity in Taiwan. The latent TB population is a walking bomb that is still a big problem. In 2006, Taiwan Centre for Disease Control conducted a project with the goal of reducing the incidence by half in 2015. The main intervention of this project was DOTS (directly observed therapy, short-course) originally. However, it seems that DOTS alone was not sufficient to achieve the goal of TB control, and Taiwan CDC has been trying to accelerate the decline of TB through active case finding.

Contact tracing (CT) is a case finding strategy in infectious disease control. When an index case is notified to the Centre of Disease Control (CDC), the neighbours of the index case will be informed to do some screening of a specific disease. This strategy could be effective in two main aspects. First, it helps the neighbours of the index case with pathogen but no significant syndrome yet to seek health care. Second, it enhances the intention of traced people with sickness to take medicine in order to prevent further infection. Additionally, contact tracing is sometimes discussed with mass screening, which screens the risky workspaces or communities. Compared to the latter, contact tracing might have more effectiveness and less external cost in lower prevalence setting. However, the mechanism of contact tracing would malfunction when social networks are poorly identified. The contact tracing would sometimes be costly because of personal privacy and casual contact history which can't be traced. In such a condition, the screening which targets at specific work place or community might become a proper choice(Begun et al., 2013). Result from these complex factors, the general effectiveness remains unclearKranzer et al. (2013).

The source of Tuberculosis cases mainly come from different processes so the interventions need to be targeted at different aspects. Cohen et al. (2007) shows the intention to use contact tracing in high population density or high incidence setting based on social networks model. Armbruster and Brandeau (2007b) advocated the importance of order in contact tracing policy. Armbruster and Brandeau (2007a) provided a cost effectiveness (QALY) maximization framework of contact tracing in a small world model. Although abstract frame have been well developed, the concrete procedure of contact tracing for tuberculosis still needs to be investigated. Thus, we fallow the paper of Armbruster and Brandeau (2007b) and develop a TB model extended from our abstract model. The purpose of this case study is to understand the potential effectiveness of contact tracing policy in TB control. ABM was suitable for this task because it naturally incorporates the contact structure of disease transmission(Guzzetta et al., 2011). The model was constructed to represent the general population in Taiwan and incorporated the age-specific contact pattern based on a recent survey. We then evaluated the impact of the proposed procedure on the reduction of TB incidence.

1.6 Objective and outline

The purpose of this thesis was to propose an improved method for fitting agent-based models with dynamic Bayesian networks to provide an interface to statistical procedure which can integrate the information of epidemiological studies. Apart from the general methods in dynamic Bayesian networks for structure and parameters learning, we developed an extensive procedure to deal with multidimensional data for agent-based inference. Moreover, we simulated the model in continuous time, allowing the various designs in transition rules between states. Using continuous time also saved the computational time compared to discrete-time sampling.

In chapter two, we will describe the procedure of constructing ABMs for infectious disease, by using basic SIR model as an example. Chapter three will provided an efficient learning scheme derived from “Origin of species” for our proposed model. Chapter four demonstrates how the proposed a class of ABMs can be used to model a complex disease setting by an example of tuberculosis control. In chapter five, using the constructed model, we will draw some policy implications and information for individual health decision strategies. Finally, we will discuss the strengths, weaknesses and future direction of thte proposed study framework.

Chapter 2

Model Construction

2.1 An Agent-Based Model with Bayesian networks for modelling infectious diseases

In this section, we will describe the model from the basic concept to the detailed processes of implementation. We will also demonstrate the model construction and parameterization by a simple Susceptible-Infectious-Recovered (SIR) model.

2.1.1 Conceptual model: Disease Triangle Model

We modelled an infectious disease based on the disease triangle model which views the origin of an infectious disease from the interaction among pathogen, host and environment. The pathogen and the environment determine the mode of transmission and exposure to infection. Taking the natural history of TB as an example, the susceptible hosts can be infected through the airborne route in the environment with TB mycobacteria. Because most of the disease pathogenesis could be deducted to the interaction among the elements of pathogen, host and environment, our agent-based model followed the framework of disease triangle model and made the elements into three types of agents. Each type of agents has its own properties, behaviours, and rules for interacting with the others. In our model, we focused on the disease outcomes of host (human) agents.

2.2 Dynamic Bayesian Networks (DBNs)

We implemented the agent-based model with dynamic Bayesian networks(Dean and Kanazawa, 1989). In order to create an interface between the model and statistical inference, we incorporated the model with Bayesian networks (Pearl, 1988) originated from artificial intelligence.

Having the similar inference structure with causal diagram, Bayesian networks have been widely applied in epidemiology(Greenland et al., 1999). Bayesian network (BNs) is a probabilistic graphical model composed of a directed acyclic graph (DAG) and a set of conditional probability tables (CPTs). DAGs are the abstract description of the modelled system. The nodes in DAGs represent the attributions of agents with the form of random variables; the edges in DAGs indicate the relationships among nodes. In addition, CPTs provide a concrete route for nodes to affect each other. With DAGs and CPTs, BNs could be performed to deal with interaction between many stochastic processes (nodes). The complete procedure of BNs consists of structure learning (analysing the relationships between variables), parameter learning (estimating the parameters), and inference (answering the questions in a specific situation)(Charniak, 1991, Uusitalo, 2007). These features of BNs match the requirements for constructing ABMs. By integrating BNs into ABMs, we could apply the statistical procedure such as Gibbs sampling and model selection to ABM.

Specifically, we used the Continuous-Time Bayesian networks (CTBNs) (Nodelman et al., 2002a) in our model to deal with the non-synchronized transition time, i.e. the nodes in the same model updating in different frequencies. Compared with discrete-time modelling, CTBNs are able to improve structure learning, avoid the inflation of both parameters(Nodelman et al., 2002b) and state space, handle the complex durations from various transition rules(Nodelman et al., 2012) and require a lower computational time(Nodelman et al., 2012). CTBNs are formed with a set of dependent continuous-time stochastic processes, according to a directed graph and conditional transition matrices extended from CPTs. CTBNs have been applied in pathogenesis(Gatti et al., 2012), reliability analysis(Boudali and Bechta Dugan, 2006), user pattern analysis(Nodelman and Horvitz, 2003), and etc. In our study, we used continuous-time models to capture the interactions among nodes but observed the model in discrete-time for clearer presentation.

2.2.1 Dynamic model for infectious diseases

We propose a class of agent-based model with the elements above. Figure 2.1 shows the conceptual processes of the basic model structure.

For each edge, the node in arrow head would be affected by the action of node in arrow tail. The exogenous variables are stochastic processes or manipulable processes which are not affected by disease dynamics. The summary nodes are variables which collect the information of the disease model and compute some statistics such as disability adjusted life years (DALY). Figure 2.2 shows an example of model extension with social networks, medicine use and symptom. Each process in Figure 2.1 is a form of many nodes in Figure 2.2. If we model the system in compartmental model or Markov model, the number of states would be hundreds or even thousands (product of state space size of each node).

We describe the general infectious disease model in the following.

Host Agents (human)

The main type of host agents in our model is human agent. A human agent includes the following nodes: basic information, family, work, contact rules, information about immunity, and a container for pathogens.

The basic information includes sex and age which affects the birth, aging, and death process. Information about family and work decides the everyday schedule of a human agent and contact structure.

Environment Agents

Environment agents are used to provide places for contact, implement health intervention, and introduce exogenous variables (e.g. policy, season, time effect). Contact spots generally include houses, work spaces, hospitals, and communities. The host agents with the same family ID would appear in the same house; The host agents with the same work ID (for workers) or class ID (for students) would appear in the same work space or school class; The host agents who have clinical symptoms would go to hospitals and seek for diagnosis and treatments. House and workspace provide spatial location for contact within social networks The community agents allow for casual contact. The hospital agents provide the medical service and allow for high risk contact. To model intervention of interest, we would create the policy maker agents (e.g. Centre of Disease Control agents). General background environments can be set as parameters in model, and time-varying environments (e.g. season) can be modelled by environment agents.

Pathogen agents

Each pathogen agents represents a strain of a particular pathogen. After infection, the clone of pathogen would interact with disease and immunity nodes in the host. Each pathogen would carry the information about transmission mode and transmission parameters.

Transmission process: hosts-environments interaction

In our model, contact and transmission are mediated by the container. The host agent and the contact agent (e.g., house and work space) both have a container for pathogens. At the start of every time slice, host agents with active infectious disease would throw the pathogen into the container of the current contact spot. In end of the time slice, the pathogen in the container of the contact spot would try to infect the susceptible host agents. The selected susceptible host agents would be infected if the pathogen be accepted by immunity node. The formalized transmission process proceeds as Algorithm 2.4

2.3 An illustrative example: SIR model

In this section, we constructed an agent-based model with CTBNs to demonstrate the proposed method. Based on the general knowledge of infectious diseases, we built a SIR model for a hypothetical infectious disease and assigned the parameter values with assumptions of an acute respiratory infectious disease. We note that this generic model was only for illustrative purpose. We will demonstrate the disease dynamics and the model performance in analysing complex policies with short computational time. This generic model can be applied to specific infectious diseases in different situations and can be subject to sophisticated model fitting (Chapter 3).

2.3.1 Collect information

The first step of model construction was to identify the knowledge necessary for input information. The settings for the example model are listed as follow:

1. There are three groups of population in SIR model, including the susceptible, infectious and recovered population.

2. Pathogen can cause the disease only when they stay in the body of the host and the impact of pathogen will not vary with the ages of the host.

3. The modelled disease in our example is an acute infectious disease transmitted through airborne route.

4. As the disease onsets, the hosts would show symptoms but would not have fatal events. In addition, we assumed the illness would discount the quality of life by half.

5. The transmission of the disease can be interrupted by personal protective equipment (PPE) such as masks.

6. Medical interventions can reduce the symptoms of diseases but can not improve the recovery rate.

In this simplified example, the parameter values of the model were assumed to be collected from literature review with some level of uncertainty (probability distribution) (Table 2.1). As will be discussed in Chapter 3, some (or essentially all) parameter values can also be fitted to the observed data using the proposed fitting algorithm.

2.3.2 Identify nodes and form the agent

The disease outcomes, symptoms, P.P.E. and drug use were identified as the nodes of a host agent. Embedded in the transmission history, the pathogen agent is presented as different strain ID numbers. The pathogen agent can connect with the disease node in the host agents if the host is infected. The agents of environment contain the community and the CDC (Table 2.2).

2.3.3 Set the interaction between nodes

To describe the interactions between nodes in disease transmission, we set the rules for nodes to interact within and between agents (Figure 2.3) . The rules of interaction can be categorized into two major processes. First, the assumptions for contact and transmission pattern were as the following: 1.)The infectious hosts can spread the pathogen to the environments. 2.)The susceptible hosts can be infected in an environment with contagious pathogens. 3.)The susceptible hosts with PPE such as masks have a lower probability of getting infection. 4.)The infectious hosts equipped with masks show a reduced chance of spreading the pathogens. On the other hand, the rules for describing disease process were illustrated as following: 1.)Events of infection start the disease process within hosts. 2.)The symptoms occur after the infected host shows infectiousness. 3.)The symptoms of the disease would bring the negative effect on the quality of life. 4.)The behaviours of health care seeking and wearing masks are induced by the occurrence of symptoms. 5.)The medical treatment can improve the symptoms of the infected hosts. 6.)Once the disease incidence (notification) exceeds the threshold, CDC will issue an alert to the public and recommended people to wear masks.

2.3.4 Set the initial states

After defining the model, we created a set of agents. Initially, we specified all the host agents to be susceptible agents. Then, we linked a pathogen host to a randomly selected host agent and start the simulation of disease transmission.

2.3.5 Simulation

We simulated the disease process in continuous-time and the contact process per day. The total length of simulation was set as 60 days and we observed the outcome every day. We simply demonstrated four scenarios to address the basic abilities of our model. The first scenario is a baseline condition without any interventions. The second scenario is to model the self-driven behaviour of mask wearing among the sick host. This individual behaviour could block the pathogen spreading and impact the overall epidemic. Third, we demonstrated a real-time policy that CDC announced the epidemic alert and the susceptible host agents put on masks for personal protection. Instead of changing the duration of host states, we modelled the heterogeneity of behaviour by adding one node that showed the condition affected by the alarm. Our proposed model can overcome the challenges of increasing complexity of the study scenarios, by using Bayesian networks to reduce the duplicative assumptions. Last, we carried out the scenario by one of the summary nodes, the QOL node. It collected the information of agents and their related node transitions. As a useful index, quality of life(QOL) can combine the impacts of different factors through simple and efficient calculations. Besides, the probability of outbreak can be analysed in ABMs which were unavailable in compartmental models. All the above experiments of the agent-based model were conducted by JAVA programming language SE7 and the epidemic curves were estimated in the size of 1,000 in Monte Carlo simulation (Table 2.3 & Figure 2.3.6).

2.3.6 Sub-model inference: basic reproductive numbers (R_0)

Basic reproductive numbers (R_0) is an important criterion for measuring the potential power of transmission of an infectious disease. It is an expected number of secondary cases generated by an infectious host in a susceptible population. In the analysis of surveillance data, R_0 can be estimated by the case number in the logarithmic period of an outbreak. In a basic equation-based SIR model, R_0 can be calculated by multiplying the transmission parameter and the duration of infectiousness. In agent based model, we conducted a Monte Carlo experiment with $1,000$ times of simulation to assess the basic reproductive number.

In each of the simulations, we assigned a person with infectious pathogen and the person was the only source of infection in the whole population. Then, we simulated the contact and the disease process in the population and ended the simulation till no events happen any more. During the simulation, all the recorded events of infection can directly provide the information for the calculation of R_0 and its credit interval. Consequently, the mean value and uncertainty range of R_0 can be obtained from the results of the 1,000 simulations. The formalized protocol for the R_0 estimation are shown in algorithm 2.5.

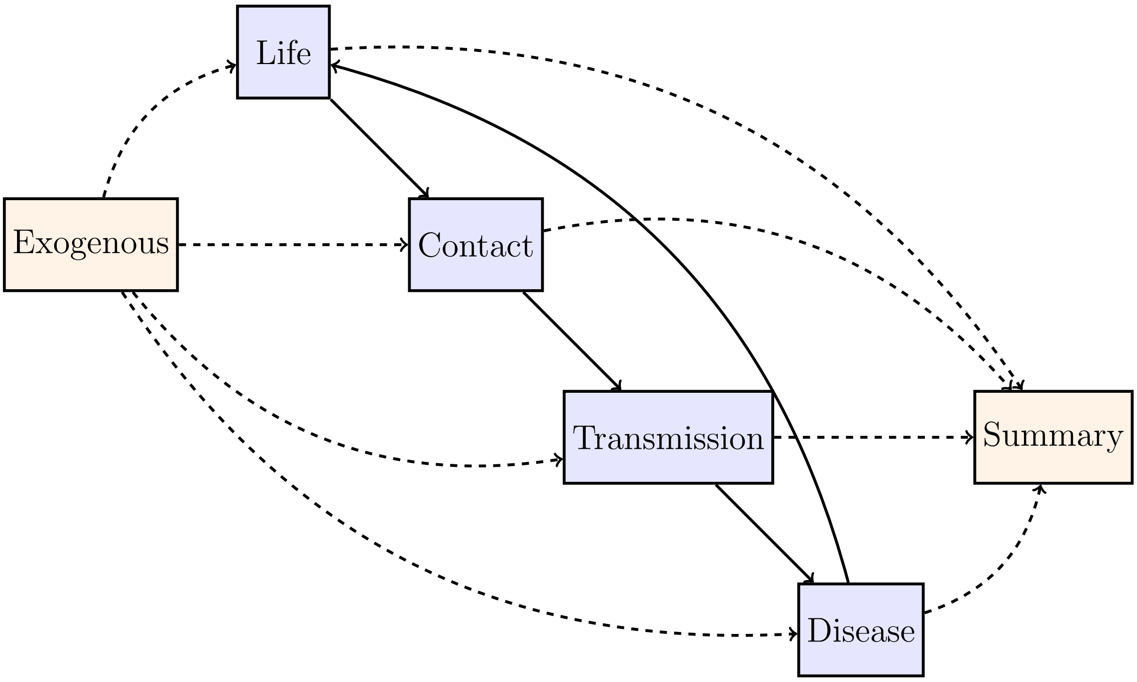
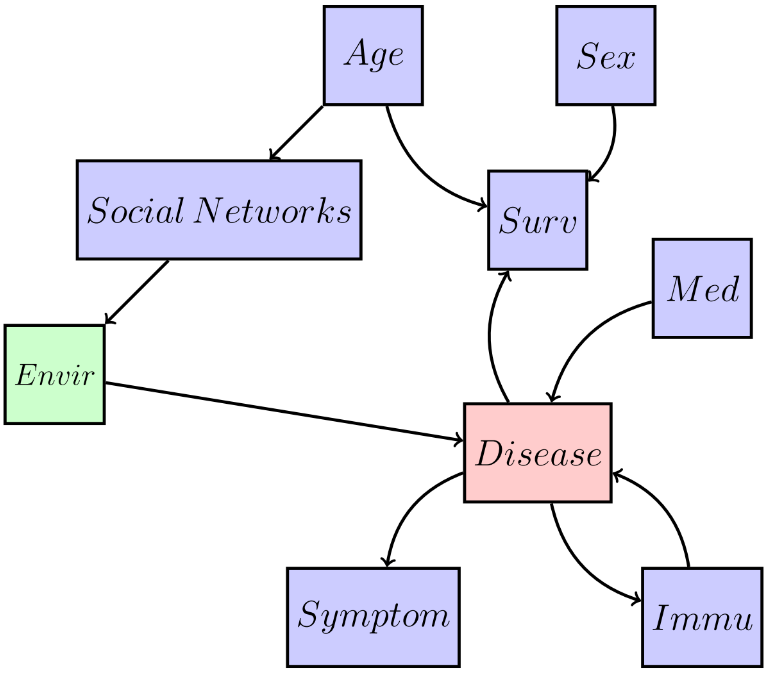


Figure 2.1: Abstract model and main processes in proposed model



Symbols	Illustrations
A	**A** is a node (attribution) for host agents
B	**B** is a node (attribution) for pathogen agents
C	**C** is a node (attribution) for environment agents
A → D	The dynamic of **D** is depended on **A**

Figure 2.2: An example of model extension

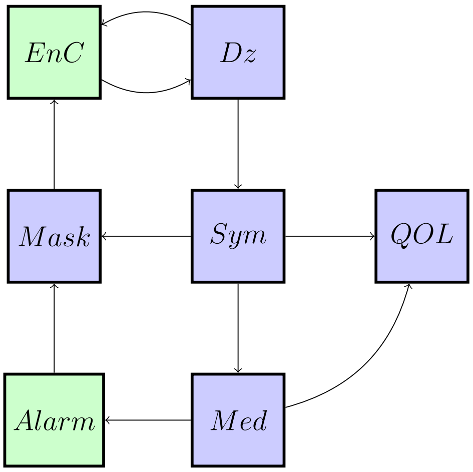


Figure 2.3: Directed graph of the example.

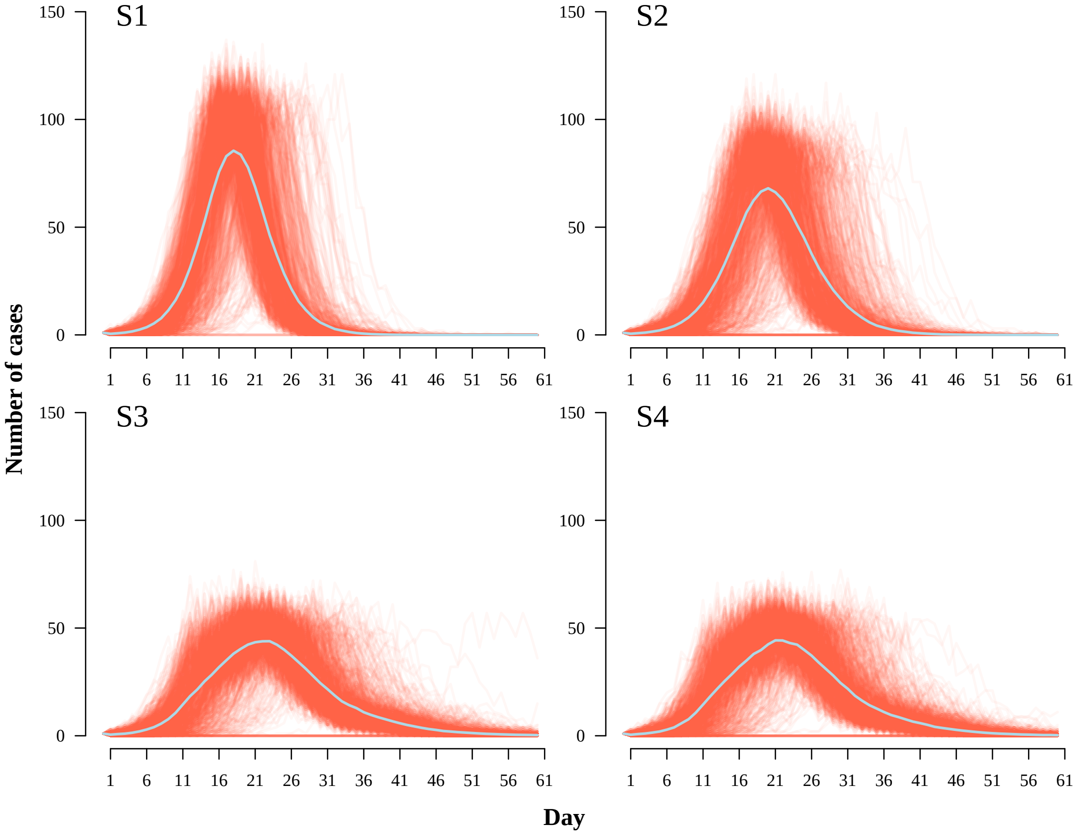


Figure 2.4: Epidemic curves of SIR example

S1: no intervention; S2: mask wearing as symptom occurrence; S3: S2 + alarm; S4: S3 + medicine use; Red line: the epidemic curves of simulations; light blue: average epidemic curves

Table 2.1: Parameter table for example of SIR model

Process	Parameter	Value	Unit	Discription
Population	N	1000	People	Population size
Disease	α	10	Day	The disease duration follow a $Gamma(\alpha, \beta)$
	β	0.1		i.e. $E(DurationRecovery) = 10$ days,
Contact	$delta$	2	Day	Symptom delay from infection
	m	100	People/Day	Number of contacts per person per day
	θ	0.005		Transmission probability
Intervention	$\grave{\theta}$	0.003		Transmission probability with mask
	Cp	50	%	Compliance of policy
	K	20	People	Threshold incident cases of alarm
Quality of Life	Q	1	Day	Quality of life for healthy people
	Q_D	0.5	Day	Quality of life for sick people
	Q_{med}	0.8	Day	Quality of life for sick people with medicine

Table 2.2: Nodes and state spaces for example of SIR model

Agent	Nodes	State-space	Paraent	Discription
Host	Dz	$[S = 0, I = 1, R = 2]$	EnC	Disease state (immunity embedded)
	Sym	$[Good = 0, Sick = 1]$	Dz, Med	Symptom state
	Med	$[NoDrug = 0, Drug = 1]$	Syn	Use of medicine
	Mask	$[NoMask = 0, Mask = 1]$	Syn, Alarm	Use of mask
	QOL	numeric ranges $[0, 1]$	Syn	Quality of life; large value indicates better quality
Pathogen	Patho	None	EnC	A carriage for information of transmission pathway
Environment	EnC	Number of Pathogens	Dz of all hosts	Sites for contact and disease transmission
	Alarm	$[No = 0, Yes = 1]$	Incidence of disease	A indicator for real-time policy

Table 2.3: Epidemic statistics in the SIR example

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Mask in I. with symptom	.	O	O	O
Alarm	.	.	O	O
Medicine use	.	.	.	O
Peak Size	794(619,969)	695(427,964)	468(290,647)	468(284,652)
Peak Time	24(17,31)	26(15,37)	27(16,39)	27(15,38)
Mean(QOL) per person-time	0.951(0.934,0.969)	0.954(0.928,0.979)	0.959(0.937,0.982)	0.994(0.984,1.003)
Pr(Outbreak)	0.989	0.965	0.971	0.969

Note: Outbreak: Peak Size > 5
Peak Size: maximum incidence during the outbreak
Peak Time: time when the outbreak reach epidemic peak
QOL: Population-wide mean of QOL during the outbreak
$Pr(Outbreak)$: Probability of Outbreak

Table 2.4: Algorithm: Transmission and contact

```
Input: $En$: a contact spot agents;
       $Hs$: a set of host agents in $En$
Output: $\grave{H}s$ host agents after contact

for Host $H$ in $Hs$ do
    $H$ spread the pathogen to $En$ if it can
end for

for Pathogen $P$ in $En$ do
    for Host $H$ in $Hs$ do
        if $H$ cannot resist the P then
            $P$ infect $H$
        end if
    end for
end for
Clear the pathogen in $En$
Return $Hs$
```

Table 2.5: Algorithm: Simulate stochastic Basic Reproductive Numbers

```
Input: $M$: target ABM for infectious disease; $P$: target pathogen agent; $num$: size of
    boot strapping
Output: Expect $R_0$ with credit interval
    Initial $R_0s$: a collection for $R_0$
    while Size of $R_0s < num$ do
        $r0 = 0$
        $N$ = size of host population in $M$
        for each Host $host$ in $M$ do
            Initial $M$
            Add $P$ to $h$ forcedly
            while true do
                Simulate contact process
                Simulate disease process
                if h is recovered then
                    if No more host in model is in latent state then
                        Break while;
                    end if
                end if
            end while
            $r0 = r0 + \#$ of hosts experimented disease reproduced by $h$
        end for
        $r0 = r0/N$
        Append $r0$ to $R_0s$
    end while
    return mean and 95% credit interval of $R_0s$
```

Chapter 3

Fitting Scheme

To connect the model to issue of interest, modeller would fit some pattern of simulation model to the data. The class of models we proposed are based on a well-defined probabilistic model, continuous-time Bayesian networks. Sub model inference is an elegant property of CTBNs. Based on the properties of our ABMs for disease dynamics, the interface of ABMs to statistical procedure would be more accessible. Based on sequential Monte Carlo, we proposed a fitting scheme for our agent-based model. The fitting scheme synthesized the numerical mutation in genetic algorithm and the parameters augmentation in blocking Gibbs sampling in order to overcome the challenges of multidimensional and multi-source data and parameters.

3.1 General fitting scheme

3.1.1 Bayesian approach: Sequential Monte Carlo

In Bayesian approach, the aim of estimation is to obtain the posterior distribution by integrating the prior information and the available data. In practice, the posterior distributions are easy to obtain but difficult to apply in subsequent analysis and make inference directly.

Sampling-based inference, such as Gibbs sampling (Casella and George, 1992), is a method aimed to get the sampled variables from the posterior joint probability density function (j.p.d.f.). The sampled variables can be further applied to approximate the estimator. Sequential Monte Carlo (SMC) methodsDoucet et al. (2001), Gordon et al. (1993) are a set of Monte Carlo simulations which select the variables in a given domain to depict the posterior j.p.d.f. There are three steps in conducting SMC. First, we began SMC with generating a set of variables in a given domain. Second, the weight was assigned to each variable based on the distribution of posterior probability. Third, the variables with higher weight would experience a higher chance of selection and then propose new variables for next iteration. After repeating the second and third steps till the convergence of posterior distribution, we would get samples which follow the posterior j.p.d.f.

Although the likelihood function is difficult to obtain in complex models, the procedure of Approximate Bayesian Computation (ABC)(Beaumont et al., 2002) has been well-developed in fitting dynamic models with the combination of SMC(Toni et al., 2009) (algorithm 3.1). . The method of ABC is used to attain the approximate likelihood function by evaluating the distance between model and data. For an instance, we can calculate the Euclidean distance as the root sum square of the difference between simulated incidence and observed incidence (i.e. $\sqrt{\sum_i (Incidence_i^{data} - Incidence_i^{model})^2}$). Sometimes, the exact likelihood function is available to calculate in complex models even though the computation process takes a longer time. In the class of proposed models, we could obtain the exact likelihood because the underlying model was a well-defined probabilistic model. Moreover, the exact likelihood function can be used to examine the sufficiency of the statistics, which indicates whether the statistics contain all the important information. The negation of exact likelihood is a good sufficient statistic in ABC-SMC or in generic SMC methods. Thus, the estimation of our model was equipped with reliability.

3.1.2 Frequentist approach: Genetic algorithm

Duboz et al. (2010) employed the genetic algorithm (GA)(Fraser, 1960) to propose a fitting scheme for agent-based model in ecology study. GA has been adopted to shorten the computation time and deal with multidimensional parameters. Generally, GA is a greedy optimization algorithm extended from the generation modification model proposed

Table 3.1: Algorithm: ABC-SMC algorithm

```
Input: $\pi(\Theta)$: prior distribution of parameter set; $l(data \mid \Theta)$: likelihood function; $q(\theta_2 \mid \Theta_1)$: proposal distribution for parameter set; $T$: targeted iteration; $N$: number of each iteration
Output: $P(\Theta \mid data)$: empirical posterior distribution of parameters
    Initial collection of $\Theta$ with N random $\Theta$ : $S^0$
        i.e. $S^0 = \{\Theta^0 = \Theta_1^0, ..., \Theta_N^0\}$
    Initial $\{\epsilon_0, ..., \epsilon_T\}$
    Assign equal weight $\{w_1^0, ..., w_N^0\}$ to each parameter set in $S$
    Normalize weight to sum one
    for $t$ in 1 to $T$ do
        Initial empty collection of $\Theta$: $S^t$
        while (Size of parameter sets in $S^t$) $< N$ do
            Sample $\Theta^t$ from last iteration by weight
            Propose $\tilde{\Theta}^t \sim q(\tilde{\Theta}^t \mid \Theta^t)$
            if $(\pi(\tilde{\Theta}^t) < 0)$ or $((data \mid \tilde{\Theta}^t) > \epsilon_t)$ then
                Continue
            else
                Assign weight $w^t = \frac{\pi(\tilde{\theta}^t)}{\Sigma\pi(\theta_n^{t-1})q(\tilde{\theta}^t \mid \theta_n^{t-1})}$ to $\tilde{\theta}^t$
                Append $\tilde{\theta}^t$ to parameter sets in $t$
            end if
        end while
        Normalize weight in iteration $t$ to sum one
    end for
    Return $S^T$
```

by Darwin(Darwin, 1859). GA possesses the advantage of fast convergence even when it was applied to multidimensional parameter sets. In addition, GA can also perform in non-differentiable or complex functions(Fogel, 1994). Although GA could not ensure the solution to be the global maximum, the solution has a lower possibility of being the local maximum. The concept of reproduction, mutation, crossover, and natural selection process are embedded in GA(Algorithm 3.2). First, the parameters of targeted function are converted into binary base value and treated as chromosome. The targeted function works as environment, evaluating the fitness of each value of parameters. The parameters with higher fitness would have high chance to reproduce new parameters in next generation(iteration). New parameters would mutate and crossover before the selection in new generation. After generation by generation, the parameters would converge to solution. Besides the targeted function, the speed of convergence and accuracy are dependent on the mutation rate and crossover rate. In Frequentist approach, model fitting is by the find the optimal value of parameters to maximize the likelihood function. Using GA to maximize the complex likelihood function of ABMs is an appropriate option. However, when applying GA to the approach of Frequentist inference, the prior knowledge has to be set as hard evidence which cannot update by data.

Table 3.2: Algorithm: Genetic algorithm(values are in binary base)

```
**Input:** $f(x)$: target function of parameters;
           $L, U$: lower and upper bounds of x;
           $N$: generation size
**Output:** $\hat{x}$: a parameter vectors which can maximize the $f(x)$
Initial generation $t = 0$
Initial $X^t = x_1^t, ..., x_N^t$ at random within $(L, U)$
**while**  Series of $X^0, ..., X^t$ not converge **do**
    $t = t + 1$
    $X^t$ a empty set of parameters
    **while**  Size of $X^t) < N$ **do**
        Pick a pair of $(x_{n1}^{t-1}, x_{n2}^{t-1})$ from $X^{t-1}$ according to $f(x^{t-1})$
        Generation $x^t$ form a pair of $(x_{n1}^{t-1}, x_{n2}^{t-1})$ by crossover
        Mutate the $x^t$
        **if** $x^t$ outside the $(L, U)$ **then**
            rule out
        **else**
            Append $x^t$ to $X^t$
        **end if**
    **end while**
**end while**
Return $\hat{x}$ =mean of $X^t$
```

3.2 Fitting scheme for single dataset

ABC-SMC have stable frameworks in model fitting and GA is fast in searching through the domain of parameters. Both of these algorithm contain proposing process(reproduction, mutation, and crossover) and filtering process(nature selection) but are applied in difference approaching(Bayesian and Frequentist). We combined these two algorithm to form a variant fitting scheme.

In this section, we are going to overcome the challenges of multidimensional parameters.

The random walk process is usually applied to propose new variables in SMC. However, when the posterior distribution is complex (e.g. with many local maximum, many parameters, non-continuous, and etc.), the methods might be improper because random walks are easily trapped with local maximum and the difference of suitable random walk size for each parameter might differ. Thus, we employed numerical mutation of GA to propose new parameters value through the iterations.

3.2.1 Numerical mutation

We employed the numerical mutation in GA to propose the new variables in SMC frameworks. Numerical mutation can be seen as a probability density function with a flat tail and a high probability of generating numbers around the base number. In the following part, we demonstrated the procedure of numerical mutation with an example.

1. Identify the variables.

   $\Rightarrow X = 100, Y = 45$

2. Convert the variables into binary value (like DNA sequences) and append them together.

   $\Rightarrow X_{(2)} = 0001100100, Y_{(2)} = 0000101101$

   $\Rightarrow XY_{(2)} = 00011001000000101101$

3. Sample the sites to mutate.

   $\Rightarrow Sites = 01000001000000010000$

4. Mutate

   $\Rightarrow XY_{(2)} = 00011001000000101101$

   $\Rightarrow X\grave{Y}_{(2)} = 01011000000000111101$

5. Restore the variables

   $\Rightarrow \grave{X}_{(2)} = 0101100000, \grave{Y}_{(2)} = 0000111101$

   $\Rightarrow \grave{X} = 351, \grave{Y} = 61$

As the example showed, the probability of mutation and the number of sites which can be mutated required parametrization. In our fitting scheme, the probability of proposed values (i.e. $Q(\theta_1 \mid \theta_0)$) can be calculated by the binomial distribution (X =difference of two sets of variables, N = sites can be mutated, $prob$ = mutation probability) and the probability showed the symmetric property (i.e. $Q(\theta_1 \mid \theta_0) = Q(\theta_0 \mid \theta_1)$).

3.2.2 Fitting scheme

In sub-model analysis of CTBNs, the unit of fitting can be a single agent, a group of agents or the whole model. Thus, we could place the relationship between data and model in the proper situation according to the fitting unit. Each fitting unit carried a set of variables and simultaneously these variables also selected the fitting unit. Apart from numerical mutation, the concept of our scheme to fit the model to the single data was almost the same as generic SMC methods. In each iteration, the proposed units were selected twice by the prior distribution at initiation and by the fitting performance (likelihoods) during simulations. Then, the survival units with higher connection to the real data were more likely to reproduce new generations. After generation by generation, the variables in survival units would follow the posterior distribution.

3.3 Fitting for multiple datasets

In simulation models, parameters, process(model), and data have a three hierarchical relationship. Every process are simulated based on specific parameters and could project the information connected to the data. The connections of the elements between and within each hierarchy are a complex network. According to the manipulation of Bayesian networks, we could identify the relationship of the processes in practice. With these connections, we proposed a elegant scheme for fitting complex model to multiple datasets. We would illustrate the fitting scheme by the SIR model in chapter 2. We assume that we know the model structure and have some reasonable range of parameters but do not know the exact parameter values. The survival data of disease duration, time series data of epidemic curve, questionnaire of quality of life are available for model fitting.

3.3.1 Blocking strategy

To map the parameter to the data, we employed the blocking strategy(Jensen et al., 1995, Liu et al., 1994) in our analysis. Blocking strategy partitions the variable set into many blocks and variables in a block are proposed jointly in iteration. It is a procedure for improving efficacy of sampling-based inference (especially in Gibbs sampler) in increasing the acceptance rate of proposed variable. In addition, sampling in blocks is suitable for variables with special constraint or non-identifiable attributions. In simulation model, dynamics of each process are dominated to specific parameters. That is, some parameters act jointly in the model. In our fitting scheme, we classified the parameters into process-specific blocks. The parameters of model are designed for dominating specific process as possible. For instance, the parameters of within agent process should not be updated by data of social networks.

3.3.2 Reducing problem of spurious correlation

Spurious correlation (collider bias)(Greenland et al., 1999, Pearson, 1896, VanderWeele and Robins, 2007) of two variables is introduced when our inference is conditioned on strata of variables which are common children of both variables. The problem of spurious correlation in dynamic model would raise in the posterior distribution of parameters given data. If possible, we should group the parameters with spurious correlation into a block in our inference. Although some contribution of parameters are non-identifiable, the data could help us rule out some spurious correlations. Thus, with appropriate data, the spurious correlation between blocks should be low.

3.3.3 Identifying the order of processes

The marginal posterior distributions of parameters have different speed in convergence. If a parameter A have much influence on many process, the convergence of the other parameters would lag to the convergence of A. However, if the process dominated by A would not be affected by other per-fitting process, we could fitting it first without simulating other processes. Thus, the simulation time and would be shorter than fitting to whole model. To reduce the simulation time in fitting procedure, we develop a reduced procedure. In Bayesian networks, the variables can be ordered in its directed acyclic graph(Pearl, 1988). The root nodes (i.e. exogenous variables, nodes with no parent nodes) have highest order. We say a variables A have higher order than a variables B, if A is ascendant variables of B. In dynamic Bayesian networks, the nodes are dominated by the nodes in the past. Each nodes in diagram of DBNs is a series of random variables. The variables in past have higher order than variables in current time in definition. The order of variables in different nodes are assigned as Bayesian networks. However, the diagrams of Bayesian networks are acyclic so the nodes with loop relation need to be pooled into a new nodes. The formalized order procedure proceeds as following and the example is the SIR example (No intervention scenario) in last chapter(Figure 2.3 & 3.1):

1. Identify the nodes of interest.

   EnC: Environment agents for contact.

   Dz: Disease state.

   Sym: Symptoms.

   Cur: Disease cure process (Within host process).

   QOL: Quality of life accounting process

2. Find the parent nodes of each node.

   EnC: Number of contact, transmission probability, Dz.

   Dz: Cur, EnC.

   Sym: Dz, symptom delay.

   Cur: disease duration.

   QOL: EnC, Sym.

3. Collapse the loops

   Dz and EnC have relationship of positive feedback.

4. Sort the processes

   $P1$**:** $Cur \rightarrow Dz$

   $P2$**:** $Cur \rightarrow Dz \leftrightarrow EnC$

   $P3$**:** $Dz \rightarrow Sym$

   $P4$**:** $(EnC, Sym) \rightarrow QOL$

5. Identify the demands for data

   $P1$**:** Cohort, RCT and Survey for time to cure ($Data1$).

   $P2, P3$**:** Epidemic curve ($Data2$)

   $P4$**:** Survey for quality of life and panic during outbreaks ($Data3$)

After sorting the processes, the mapping of parameter block, sub-model, and data could be rearranged. The parameters block and data can be assigned order respectively.

3.3.4 Fitting scheme

After sorting the processes, the mapping of parameter block, sub-model, and data could be rearranged. Then the fitting scheme for single dataset could be applied to data from high to low order. Based on the example above, we could complete the fitting procedure.

1. Fitting for disease duration

   Data: $Data1$

   Para: Parameters and distribution of disease duration

   Sub-model: A group of host agents infected with pathogen $P1$

2. Fitting for disease dynamics

   Data: $Data2$

Para: Parameters about transmission and symptom delay

Sub-model: Whole model without QOL node $P2, P3$

Given: Host agents with trained disease process $P1, Data1$

3. Fitting for QOL with respect to disease states

   Data: $Data3$

   Para: Disability weight

   Sub-model: Whole model $P4$

   Given: $P1, P2, P3, Data1, Data2$

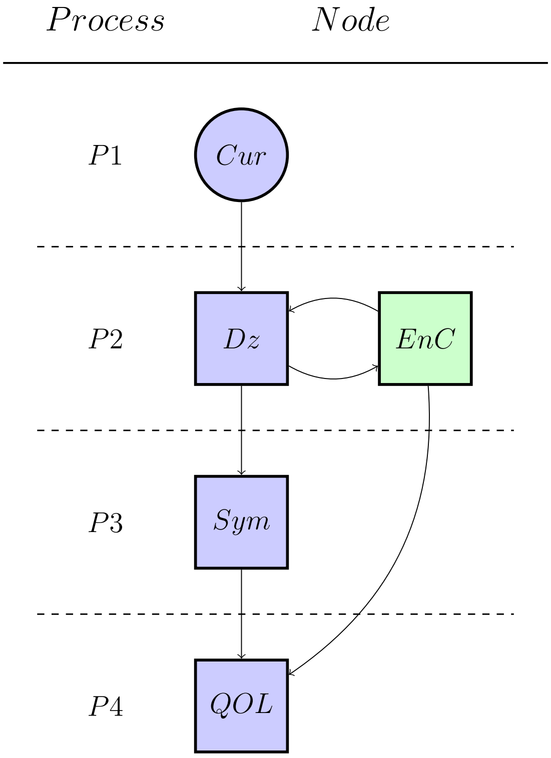


Figure 3.1: Ordered directed graph of the example with no intervention.

Chapter 4

Tuberculosis Dynamic Model

4.1 Tuberculosis control model

The directed graph of our TB model are shown in figure 4.1. The blue boxes indicate the node in human agents; the green boxes indicate the node about environments; the red box represents the pathogen and disease process.

4.1.1 Environments agents

There are three kinds of environment agents we modelled. Environments for contact including house and workspace agents support spreading pathogens and contact to pathogens; Hospital agent provides health care for host with symptom; CDC agent is the agent conducting policy in the model. Workspace are randomly assigned according age and updated every three years. House are assigned according the family tree of each human agents. We model complex family process and form various household type in Taiwan because the family contact are highly correlated to TB in our assumption. Beside the birth and death, we simply model the marriage and divorce events and each adult human agents can choose if it want to leave with its parents.

4.1.2 Human agents

There are seven nodes in a human agent: sex, age, immunity, disease, BCG(Bacillus Calmette-Guérin), Anti-TB drug, symptom, and DALY. Age and sex would affect the disease-free mortality rate; some disease states would stimulate the appearance of symptom; immunity records the activation and infection events of the host; DALY (Disability-adjusted life year) is summary node records the uncomfortable bring by TB.

4.1.3 Tuberculosis agents

Each TB agents represents a strain of bacteria. The TB agents and nodes in the host are connected while infection succeed. There are there states embedded in tuberculosis agents: latent, activated, and controlled. These three states are corresponded to the latent, infectious, recovered respectively in compartmental model. The progression from latent to activated follows a exponential distribution and slow down after duration larger than two year. As a strain of TB activates, the type of TB would be labelled in terms of smear positive or negative and the corresponded fatality rate would be assigned. The active TB can be controlled by host immunity or by Anti-TB drug. The detail parameters of TB process are shown in table 4.1. Multiple infection are allowed in our model. When a strain activating, the immunity would be notified and the progression of other strain of TB would be ceased until its host recovered. Controlled strain sometimes would relapse, and make the recurrent disease. As partial immunity hold by activation, human host would have a barrier for further transmission and the progression rate for the other strain of TB would be lower.

4.1.4 Health care seeking

Health care seeking is stimulated by the symptom of host agents. The time from disease onset to diagnosis of TB is defined as total delay. Total delay combines the delay form disease onset(patient delay) to first time health seeking and the delay during diagnosis(system delay). In our model we make the assumption as follows

1. Average total delay without active case finding is six month.

2. Diagnosis immediately(no system delay).

3. Prefect diagnosis(100% specificity and sensitivity).

4. No contact in hospital.

The people with activated TB would be prescribed Anti-TB medicine and be notified to CDC. The latent infection would also be notified to CDC and CDC would place the intervention. The average traced people per index case is about 10 in 2012 in Taiwan.

4.1.5 Intervention model

Besides active case finding, we model the three kinds of intervention for TB: BCG vaccination program, anti-TB drug for host with active TB, preventive therapy for latent TB infection. Host agents are vaccinated with BCG by CEC at birth. BCG could prevent host agents from disease progression. Anti-TB could eliminate the symptom and reduce the infectiousness through two weeks. preventive therapy could reduce TB activation(progression progression, reactivation, and relapse). In our model, we model the contact tracing policy. As a patient be notified, CDC would identify the neighbours of the patient. Then the neighbours would be sort by their relationship from family members to colleagues or classmates. The neighbours which have experienced latent TB infection treatments would be rule out. Finally, CDC would trace neighbours until reaching target number and ask traced people to go to hospital, doing the TB screening.

4.2 Model fitting

We calibrated the model to the data in Taiwan, and used the model to identify the most effective contact investigation algorithm based on individualized information. The likelihood function are connected the model to dependency ratio, and household type structure and TB notification form 2006 to 2013. We applied to fitting scheme in section 3 to our model (algorithm 4.2). Household structure are considered in contact process, and work or school contact are assumed. The results of model fitting are show in figure 4.2 and figure 4.3.

Based on the model we constructed and calibrated in the previous chapter, we tried to do some inference by counter-factual experiments and draw some policy implication.

4.3 Forecasting: Policy analysis

Our analysis was aimed at figuring out the contribution and synergism of components of CT policy in Taiwan. We assumed the policy was started in 2006. Before TB patient would go to hospital only when symptoms occurrence. Our experiments are adopted from effectiveness analysis and rank scheme in Armbruster and Brandeau (2007b). We consider three scenarios: CT for active TB, CT for latent TB, and CT for both active and latent TB. We first simulated many models to equilibrium. We then cloned each model three times to form an experiment set. Each clone was assigned an experiment setting in later simulation. Lastly, we simulated forward for 15 years. The targeted numbers of traced contacts pre index case was 10 in the experiments. The Contact tracing policy were conducted as follows:

1. Accept the notification of an active TB case.
2. Identify the visible contacts of the case.
3. Assign order to the contacts via its connection to the case (From family to co-worker)
4. For the contact C with highest order:
   (a) If C has not received the latent TB infection treatment
   $\rightarrow$ Suggest him/her to receive the screen in hospital.
   (b) Remove C from the list.
5. Repeat step 4 until targeted number reached or no contact can be traced.

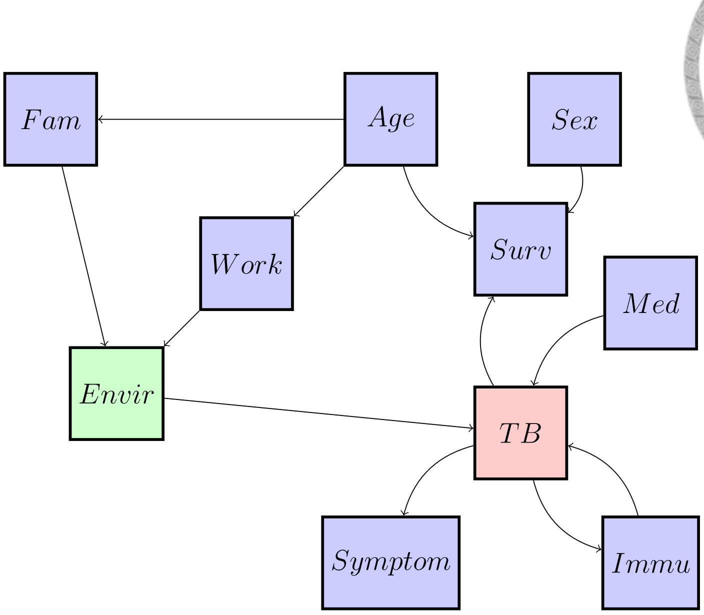


Figure 4.1: Directed Graph of Tuberculosis model

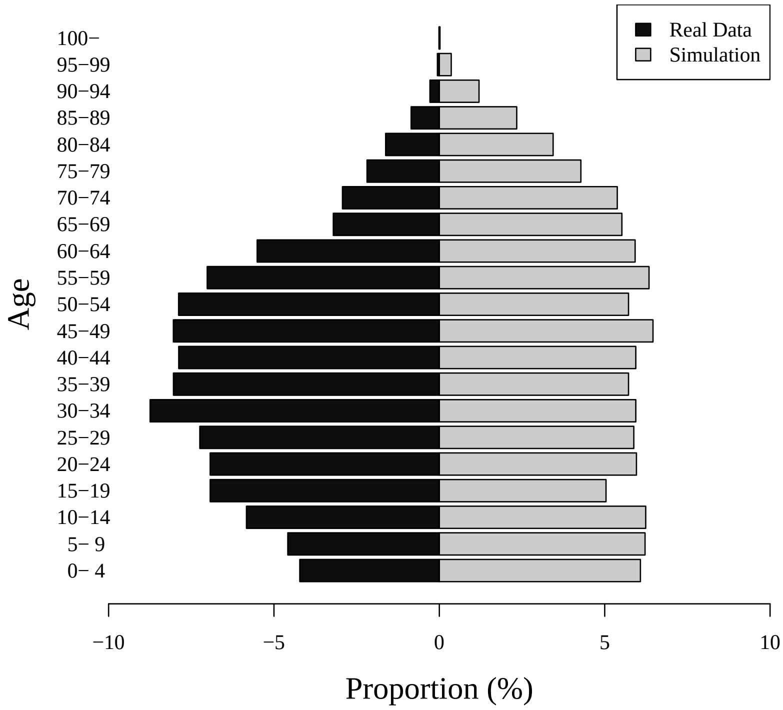


Figure 4.2: Population pyramid of model and real data in 2013

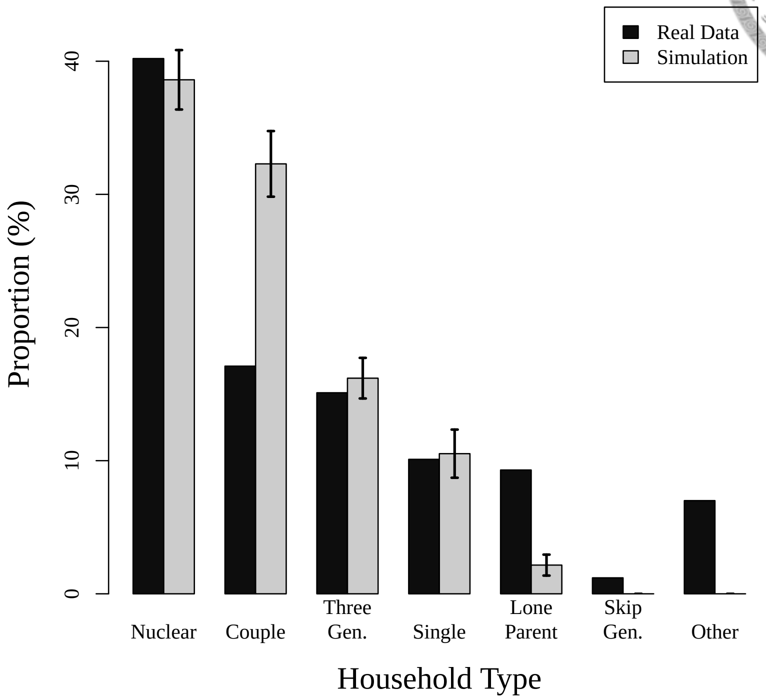


Figure 4.3: Distribution of household type of model and real data in 2011

Table 4.1: Parameter table for TB model

	Value	Unit	Source
Primary progression rate	0.07	yr^{-1}	Vynnycky and Fine (1997)
Time to late latent	5	yr	Vynnycky and Fine (1999)
Reactivation Rate	0.00073	yr^{-1}	O'Shea et al. (2014)
Recovery rate(Sp)	0.1	yr^{-1}	Tiemersma et al. (2011)
Recovery rate(Sn)	0.27	yr^{-1}	Tiemersma et al. (2011)
Sn / Sp	0.5	none	
Fatality rate(Sp)	0.23	yr^{-1}	Tiemersma et al. (2011)
Fatality rate(Sn)	0.07	yr^{-1}	Tiemersma et al. (2011)
Relapse rate	0.038	yr^{-1}	Taiwan CDC survey
Reinfection Barrier	79	%	Tostmann et al. (2008)

Table 4.2: Algorithm: Fitting scheme for TB model

```
**Input:** $\pi(model)$: prior p.d.f. of model;
    $l(data \mid model)$: likelihood function;
    $D$: Data;
    $Model$: Model with $L$(ife), $C$(ontact), $D$(isease) process;
    $pr(L, C, D) = pr(D \mid L, C)pr(C \mid L)pr(L)$
**Output:** Trained ABMs follow $p(model \mid data)$ (posterior j.p.d.f)
Initial empty collection of ABM: $M$
Put N ABMs with random parameters into $M$
Block parameter sets by $L, C, D$
Train model in $M$ with $L$ by data of $L$
Train model in $M$ with $C$ given $L$ by data of $C$
Train model in $M$ with $D$ given $C, L$ by data of $D$
Return $M$
```

Chapter 5

Results for TB Model

5.1 Population-level inference

Figure 5.1 shows the impulse-response function of the policy with CT for different experiments setting. The response are in terms of prevalence aversion by the policy. Prevalence aversion of intervention i at year t are defined as follows:

$$PrvAvrt_t^i = \frac{Prv_t^{NoIntervention} - Prv_t^i}{Prv^{NoIntervention}} \times 100\%$$

which Prv_T^I denotes the prevalence of intervention I at year T. It indicate that the effectiveness of the policy are concentrated at early year. If there really are no active case finding before 2006 as well as our model, the policy would give TB prevalence a shock. And then the policy would gradually take the epidemic to another steady state (i.e. diminishing return to policy duration). In fact, active case finding including contact tracing started earlier than our setting. In Taiwan CDC surveillance, the incidence (notified) of TB have no great impulse at 2006. If the results hold in reality, the shock has happened in the pass so the current contact tracing policy are effectiveness in sustain the epidemics. The results for effectiveness of contact tracing only for latent TB or active TB indicate that the overall effectiveness are mainly attributed to the former. The effectiveness of contact tracing for active TB are around zero. The possible mechanism would be discussed in the discussion section.

We analysed the record of traced host agents in the scenario with contact tracing for both active and latent TB (Table 5.1). The results indicated that, compared with the general population, the close contacts of an index TB case are with higher proportion of active and latent TB comparing to the general population of the model. The TB patients found by the policy are about 2 percent. It illustrated the low effectiveness of contact tracing only for active TB. The repeat traced, the proportion of people who have been traced in the past among the traced people, contact in all traced contact revealed the importance of clustering in TB and the efficiency loss. For traced TB cases, we found that the delays from disease onset to diagnosis were about 40 percent off in average.

5.2 Individual-level inference

After drawing the inference in group level, we want to have some implication in individual levels. We generated two groups of host agents in 20 years old and infect them with Tuberculosis. First group were started with latent disease and the other were started with active disease. Then, we simulated survival and disease processes of these agents until them died. Figure 5.2 examines the effect of varying Total Delay (TD), the time between TB onset and getting anti-TB medicine.

The rational TD ranges from 3 month to 2 years in Taiwan. Infinite TD denotes no-medicine scenario; reference group is no-TB scenario. The left side sub-graph is the outcome of experiments on latent host agents and the other side is on infectious host. Figure 5.2A and 5.2B shows downward trends of life expectancy in age 20 as the TD increases. The life expectancy in 20 years old has no significant change in small TD improvements (within one year). The insignificance life expectancy change between normal people and people with latent infection may be due to the low life time risk of TB activation. The scenario of one day TD indicates that the anti-TB medicine could prolong 8 year life time expectedly based on collected knowledge. Figure 5.2C and 5.2D shows increasing years lived with disability (YLD) from age 20 as we longed the TD. The disability weight of TB patients we use is 0.313 life years loss by disability. YLD indicates the length of infectious period. Compared to life expectancy, the YLD would have higher elasticity to TD varying. Figure 5.2E and 5.2F are graphs of the Disable-Adjusted Life Year (DALY). Higher value of DALY indicates the more harmful events might bring by TB. If we halve the TD from six month, a TB patient should take back about 0.4 life time loss in average. If we have more estimated sex-specific information, we could adjust for that. Overall, figure 5.2 suggests that the intervention of shortening the total delay have limited impacts on TB epidemics from individual level inference.

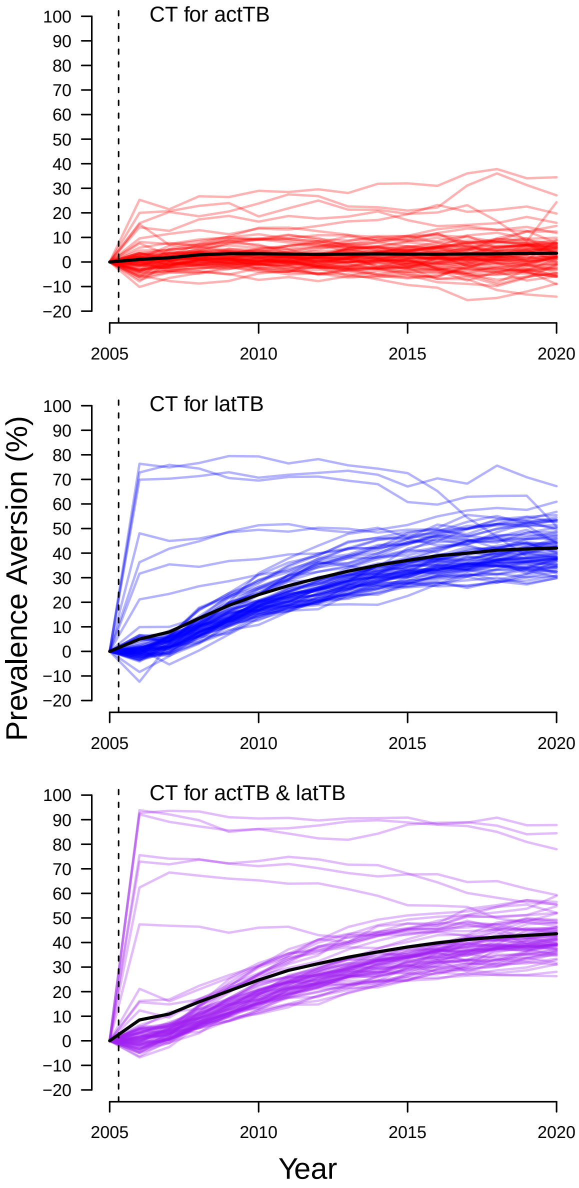


Figure 5.1: TB prevalence aversions for experiments

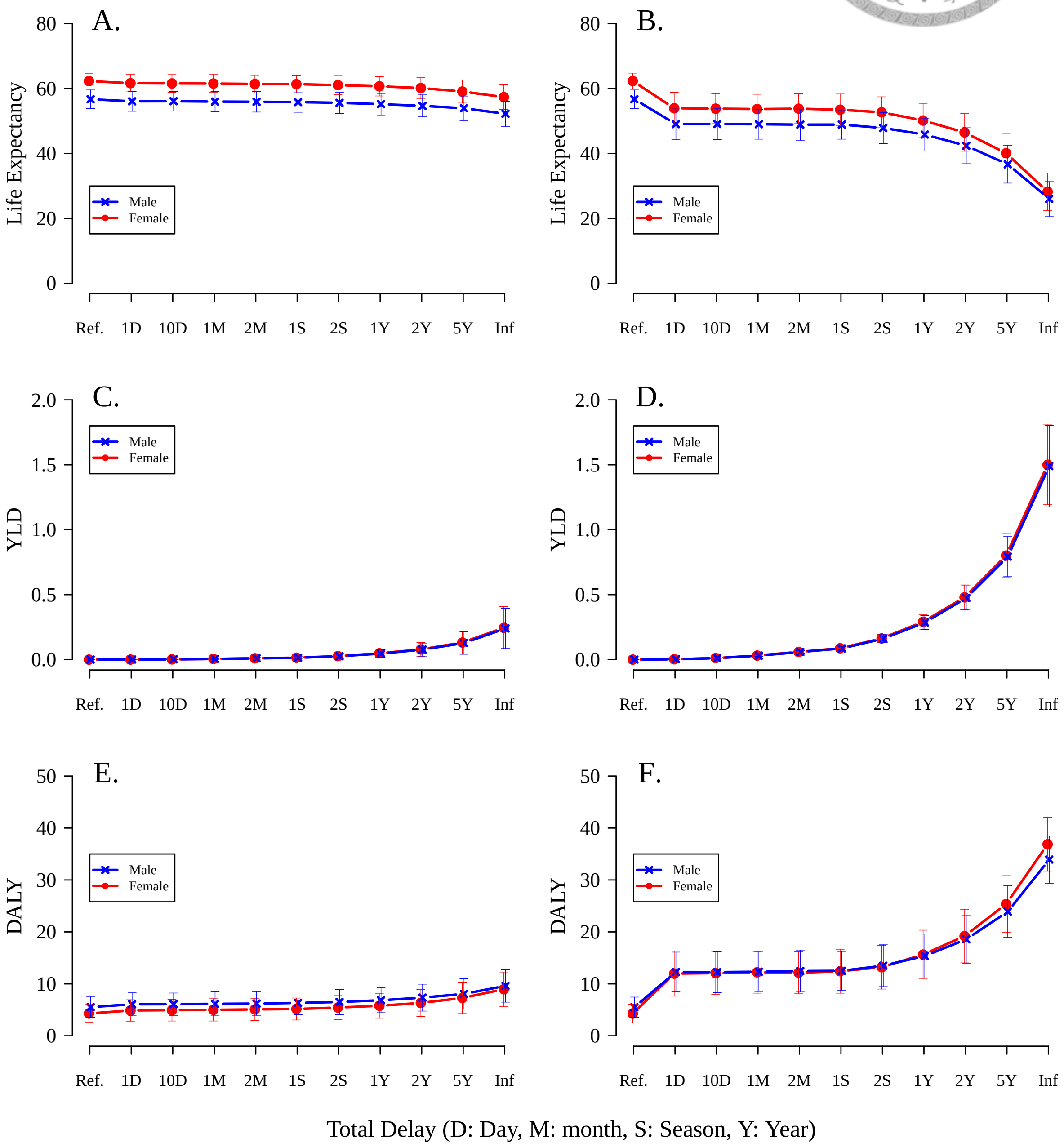


Figure 5.2: Sub-model simulation

Table 5.1: Summary of contact tracing record in model

Population	Traced		General	
Year	2011	2016	2011	2016
Actual traced contact (#)	9.71 (9.15, 10)	8.97 (0, 10)		
Smear Positive (%)	53 (0, 100)	0.48 (0, 100)	0.5	0.5
Latent TB (%)	73 (44, 84)	69 (0, 85)	62.02 (54.96, 67.95)	60.87 (52.69, 67.64)
Active TB (%)	2 (0, 3)	1 (0, 3)	0.64 (0.02,1.10)	0.65 (0.05,1.07)
Repeat traced (%)	49 (21, 61)	53 (24, 64)		
Patient delay (Day)	115.66 (0, 222.17)	95.34 (0, 235.46)	187.87 (33.26, 669.82)	177.43 (34.59, 534.16)

Chapter 6

Discussion

We proposed a class of agent-based model with Continuous-time Bayesian network, a temporal multivariate probability model, for investigating disease dynamics and creating an interface to statistical analysis. While retaining the strength of existing procedure for simulation model fitting based on sequential Monte Carlo, we propose an improved framework for fitting agent-based model for disease dynamics. We synthesized the numerical mutation in genetic algorithm and the parameters augmentation in blocking Gibbs sampling in order to overcome the challenges of multidimensional parameters and multi-source data. Moreover, we simulated the model in continuous time, allowing the various designs in transition rules between states. Using continuous time also saved the computational time compared to discrete-time sampling.

6.1 Connection of the proposed agent-based model with epidemiological studies

As a temporal probabilistic model, dynamic Bayesian Networks in our model can facilitate the general epidemiological studies and make the statistical models more informative. However, the process of constructing an agent-based model is usually seen as a black box. Causal diagrams, an extension of Bayesian networks, have been employed in epidemiology for many years. Graphical inference might be more familiar to people who study the epidemiology than state space inference. As shown in SIR example, the continuous-time Bayesian networks model state changed the processes by time-to-events distribution but not by the probability of transition. It corresponded intuitively the thinking in epidemiology. With continuous-time Bayesian networks, we could summarize the detailed settings of the model by three tables, a directed graph, and a list of special behaviours included. The first table shows the node and it possible state we model (e.g., Table reftab:sirnodes); the second table describes the conditional probability matrices for time to next events (e.g., Subsection ref s:Sirint); the third table lists the parameters in conditional probability matrices. The directed graph visualize the relationship between nodes (e.g., Figure reffig:sirdg). Our model makes members in epidemiology to judge agent-based model easier. It is still not easy to rebuild the model by these information. Although the information in model has been clarified by these tables, the actual implementation demands the use of programming languages. In the future, we should build a graphical user interface for generating an agent-based model with the basic information.

This model also provided a statistical interface for integrating data and information of epidemiological studies. The fitting scheme we proposed combines numeric calculations in genetic algorithm, blocking strategy in blocked Gibbs sampling, and sequential Monte Carlo. With the assistance of the improved fitting scheme, our model could incorporate data from multiple sources. If we are able to well-identify the relationships between different data, the fitting process could be shortened. As the sources of data increase, the advantages of blocking scheme will be more significant. Because the blocking strategy can also help in clarifying how to deal with heterogeneity of agents and hierarchical inference. We could made each agent in our model to carry a block of parameters. Host agents in the same family or in the same sub-groups carry the same block. Hence, the effects of family, workspace and other nodes in the model are fixed.

6.2 Advantage of using continuous-time sampling

The method of continuous-time sampling possesses many good properties for application in agent-based models. First, the flexibility of continuous-time sampling allows the time-to-event of each node to be drawn from any non-negative distributions in contrast to exponential distribution in most modelling studies. Unlike the restrictions in Markovian models that the current states were dependent on the previous states, we can select a distribution to determine the next states in the model and modify the range of time slice to capture the characteristics of a specific behaviour. For example, a disease with three phases of progression stages could be modelled by phase-type distribution(Gopalratnam et al., 2005, O'cinneide, 1999). Applying the continuous-time sampling method to the agent-level simulation, we can obtain the output data which can be further compared with observational studies. The property of flexibility offered a convenient way not only in observing output data but also in inputting model parameters. The parameters could be placed in the right position without much transformation. Next, continuous-time sampling can skip the time slice without the occurrence of events. In discrete-time sampling, the state of each node should be checked and the next state should be re-sampled at every time slice. With frequent checks of states, the discrete-time sampling method wasted more time in computation compared to the continuous-time sampling method(Engel and Etzion, 2011). Moreover, continuous-time Bayesian networks are proposed to solve the problem that different nodes switch the states at different time points in a single model. For instance, the progression of acute infectious disease happens every few days while the contact structure changes at the time scale of months or even years.

Although continuous-time sampling has brought many benefits to the process of model construction, the obtained results are different from the population-level real data which are usually measured in discrete time. Therefore, the exact likelihood function was difficult to estimate in the model. To solve the problem, we can perform approximate methods such as Approximate Bayesian Computation or assume probability functions according to the observed data types. Normal distribution was frequently applied for continuous or aggregate data while Dirichlet distribution was taken for categorical data. Another concern of CTBNs is the inability to manage the continuous data such as CD4 count, viral load and blood pressure which are frequently seen in disease models. Further studies could explore the stochastic differential equation nodes in agent-based model for nodes with continuous values.

6.3 Computation time saving

The consumption of computation time is an important issue in performing agent-based model. The long computation time creates a barrier to applying sampling-based inference such as Monte Carlo experiments. In order to shorten the simulation time of disease dynamic model, we made some efforts as follows. First, the application of continuous-time sampling can accelerate the simulation. CTBNs use event-driven sampling strategy which update the model just at the target action happened (Engel and Etzion, 2011). Second, the simulation of pathogen agents we used reduced the time of contact process. Comparing with the pairwise contact method which checked every edge of contact network, we separated the contact process into two steps, involving the spread and the receipt of pathogens. The contact process could filter out the unnecessary contact histories such as contacts between susceptible. With the above rules, the computation time of our contact process could be reduced form $O(n^2)$ to $O(n)$ under low burden settings. The higher bound of our algorithm is the paired-wise contact. This contact process made the occurrence of contacts within specific sites. This assumption is more likely to hold for air-borne, water-borne or diseases with geometrical effects. On the contrast, it didn’ t work for social network models. For diseases requiring dense connections such as sexual-transmitted diseases, the identification of the contacts with high risk might be a better choice for strategy design when the disease outbreaks. In our model, we took the summary nodes to gather and compute the information in the agents. It was a just-in-time strategy(Hutchins, 1999) and it ensured the synchronization of information input and the model simulation. It let the model drop the used information as soon as possible and released the memory storage for the following complex computation. In addition, summary nodes is able to create the possibility for modelling the real time decisions at an individual level after further extension. Combining the summary nodes with economic knowledge, broader information such as the compliance and the cost-effectiveness of health policies can be appropriately evaluated. In practice, the timing for initiation and termination of an intervention can be well planned. Consequently, the largest benefits can be generated while the least negative impact.

6.4 Numerical mutation

Numerical mutation is a non-classical probability mass function. It has many properties correspond to the demand in agent-based model fitting. It is parametrized by parent number, upper and lower bound, and mutation rate. Agent-based model are composed of a large set of non-linear behaviours so the distribution of observation are usually with high complexity. Numerical mutation can sample the number around parent number and the number with longer distance to parent number can be easily selected, comparing with normal distribution and uniform distribution. The distribution of numerical mutation has a flat tail, but its variance, unlike Cauchy distribution, can converge as the sample size enlarges. The convergence of variance ensures the convergence of Monte Carlo inference. The setting of mutation rate is a trade-off between the speed of convergence and the accuracy of algorithm. Higher frequency of mutation would bring larger variance so the searching space would be more complete with respect to prior distribution, which means that the posterior distribution could be depicted more precisely during the convergence. However, more generations for convergence are required for the algorithm. Numerical mutation is suitable for multiple parameters. We could append all parameters in binary base to a sequence and mutation at one time. With numerical mutation, we could increase the efficacy of proposing new parameter block and of generating new agents in posterior distribution. Further discussions about parameters in algorithm should be clarified.

6.5 Limitation

One of the limitations of our proposed method is the ability to deal with the confounders and exogenous variables. Although the models were assumed the prefect worlds, the data were not. In fact, unmeasured exogenous variables and confounders might bias the estimation even our fitting scheme perfectly work. Pearl (2014) uncover the importance of confounder in policy analysis. It suggest that the scientific knowledge could not help the policy analysis because the policy is not placed in the world with adjusted confounders as in trials. If we ignored the underlying confounders, the results of model study might be useless. In theoryl, confounders introduce the non-identifiable process into model, making the true association between factors difficult to disentangle. Our model, like the general epidemiological studies, is unable to adjust the unmeasured confounders automatically. Besides, if we have more information of the possible range or the potential mechanism of confounding, we can narrow the uncertainty in model simulation and provide the implication with more robustness. Fortunately, based on Wold's theorem(Box et al., 1976), the impact of confounder can be transformed into the impact of historical data if the impact functions are the same through the duration of interest. In other words, given vector time series data of the visible process, the confounding would bias the estimator but has no influence on forecasting. This property remains when the model is stationary, i.e. in steady-state. It is important because the identification of confounder could affect the robustness of simulation. Sometimes, the agent-based model could have better forecasting of behaviours than equation-based model even the equilibrium is not reached. The difference of model performance in forecasting might be non-significant because of ill-identification.

Our results are based on limited Monte Carlo size and population size. The endemic of TB may be broken due to stochastic impulse. Theoretically, enlarge population size of sampling could increase the stability of model. However, the population size of holding an endemics is less accessed. Increasing the population size for endemics might mislead the model from reality. The further study should focus on the stability of the system and find the potential risk of random outbreaks with respect to different population size setting.

6.6 Tuberculosis control policy

The outcomes of the posterior model showed reasonable properties compatible with current knowledge and disease measurements. In the policy analysis, we found that the contact tracing policy in Taiwan were mainly effective in reducing latency. The works of finding undetected active TB cases shown limited effectiveness. Low incidence of TB and long serial interval might bring the hardness of finding active TB cases around index cases. Our results suggested that reducing the individual DALY by shortening the patient delay may not be sufficient for TB control.

The results of our model showed similar characteristics of contact tracing with a modelling study conducted in a middle-burden setting Kasaie et al. (2014). The maintenance of contact tracing led the epidemic to a new steady state but not a constant decreasing trend. As the policy of contact tracing was removed, the epidemic reverted to the initial level. In Taiwan, the majority of the elder TB cases might be partially attributed to the infection in the early time (remote infection). If the current policy is able to hold the epidemic at a low equilibrium, the vanishing of the cohort with high latent infection would increase the possibility of TB elimination. And then, the potential power of TB transmission might be reduce as the policy holds. However, the timing of removing the policy needs investigations. To inform policies for TB control, further simulation studies can investigate the relationship between the intensity of contact tracing level and the level of incidence at equilibrium. Moreover, with the improvement of computation time in our model, Monte Carlo studies could be applied in testing the probability of elimination in different combinations of control strategies.

6.7 Conclusion

To sum up, we proposed a procedure that could be used to generate generalized ABMs for infectious disease dynamics. The improved techniques in model fitting and computation increase the accessibility of ABM inference. Although future studies for statistical properties of the proposed method are still needed, the framework could be readily applied to assess complex questions of infectious diseases, for example, contract tracing and mixed-strain infection.

Bibliography

Armbruster, B. and Brandeau, M. L. (2007a). Contact tracing to control infectious disease: when enough is enough. *Health care management science*, 10(4):341–355.

Armbruster, B. and Brandeau, M. L. (2007b). Who do you know? a simulation study of infectious disease control through contact tracing. In *Proceedings of the 2007 Western Multiconference on Computer Simulation*, pages 79–85. Citeseer.

Auchincloss, A. H. and Roux, A. V. D. (2008). A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American journal of epidemiology*, 168(1):1–8.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Begun, M., Newall, A. T., Marks, G. B., and Wood, J. G. (2013). Contact tracing of tuberculosis: a systematic review of transmission modelling studies. *PloS one*, 8(9):e72470.

Blower, S. and Go, M.-H. (2011). The importance of including dynamic social networks when modeling epidemics of airborne infections: does increasing complexity increase accuracy? *BMC medicine*, 9(1):88.

Boudali, H. and Bechta Dugan, J. (2006). A continuous-time bayesian network reliability modeling, and analysis framework. *Reliability, IEEE Transactions on*, 55(1):86–97.

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (1976). *Time series analysis: forecasting and control*. John Wiley & Sons.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Charniak, E. (1991). Bayesian networks without tears. *AI magazine*, 12(4):50.

Cohen, T., Colijn, C., Finklea, B., and Murray, M. (2007). Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission. *Journal of The Royal Society Interface*, 4(14):523–531.

Darwin, C. (1859). *ON THE ORIGIN OF SPECIES-6TH*. John Murray, London.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational intelligence*, 5(2):142–150.

Deisboeck, T. S., Wang, Z., Macklin, P., and Cristini, V. (2011). Multiscale cancer modeling. *Annual review of biomedical engineering*, 13.

Doran, J. (2001). Agent-based modelling of ecosystems for sustainable resource management. In *Multi-Agent Systems and Applications*, pages 383–403. Springer.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer.

Duboz, R., Versmisse, D., Travers, M., Ramat, E., and Shin, Y.-J. (2010). Application of an evolutionary algorithm to the inverse parameter estimation of an individual-based model. *Ecological modelling*, 221(5):840–849.

Engel, Y. and Etzion, O. (2011). Towards proactive event-driven computing. In *Proceedings of the 5th ACM international conference on Distributed event-based system*, pages 125–136. ACM.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Generative Social Science: Studies in Agent-Based Computational Modeling*, 4(5):4–46.

Farmer, J. D. and Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256):685–686.

Fogel, D. B. (1994). An introduction to simulated evolutionary optimization. *Neural Networks, IEEE Transactions on*, 5(1):3–14.

Fraser, A. S. (1960). Simulation of genetic systems by automatic digital computers vi. epistasis. *Australian Journal of Biological Sciences*, 13(2):150–162.

Galea, S., Riddle, M., and Kaplan, G. A. (2010). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology*, 39(1):97–106.

Gatti, E., Luciani, D., and Stella, F. (2012). A continuous time bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*, 24(4):496–515.

Gopalratnam, K., Kautz, H., and Weld, D. S. (2005). Extending continuous time bayesian networks. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 981. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET.

Grassly, N. C. and Fraser, C. (2008). Mathematical models of infectious disease transmission. *Nature Reviews Microbiology*, 6(6):477–487.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48.

Grimm, V. and Railsback, S. F. (2013). *Individual-based modeling and ecology*. Princeton university press.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., and DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *science*, 310(5750):987–991.

Guzzetta, G., Ajelli, M., Yang, Z., Merler, S., Furlanello, C., and Kirschner, D. (2011). Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting. *Journal of theoretical biology*, 289:197–205.

Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., and Huth, A. (2011). Statistical inference for stochastic simulation models–theory and application. *Ecology Letters*, 14(8):816–827.

Hutchins, D. (1999). *Just in time*. Gower Publishing, Ltd.

Janssen, M. (2002). *Complexity and ecosystem management: the theory and practice of multi-agent systems*. Edward Elgar Publishing.

Jensen, C. S., Kjærulff, U., and Kong, A. (1995). Blocking gibbs sampling in very large probabilistic expert systems. *International Journal of Human-Computer Studies*, 42(6): 647–666.

Jewell, C. P., Kypraios, T., Neal, P., Roberts, G. O., et al. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4(3):465–496.

Kasaie, P., Andrews, J. R., Kelton, W. D., and Dowdy, D. W. (2014). Timing of tuberculosis transmission and the impact of household contact tracing. an agent-based simulation model. *American journal of respiratory and critical care medicine*, 189(7):845–852.

Kranzer, K., Afnan-Holmes, H., Tomlin, K., Golub, J., Shapiro, A., Schaap, A., Corbett, E., Lonnroth, K., and Glynn, J. (2013). The benefits to communities and individuals of screening for active tuberculosis disease: a systematic review. *Int J Tuberc Lung Dis*, 17(4):432–46.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

Macal, C. M. and North, M. J. (2005). Tutorial on agent-based modeling and simulation. In *Proceedings of the 37th conference on Winter simulation*, pages 2–15. Winter Simulation Conference.

Maglio, P. P. and Mabry, P. L. (2011). Agent-based models and systems science approaches to public health. *American journal of preventive medicine*, 40(3):392.

Meltzer, M. I., Cox, N. J., Fukuda, K., et al. (1999). The economic impact of pandemic influenza in the united states: priorities for intervention. *Emerging infectious diseases*, 5:659–671.

Nodelman, U. and Horvitz, E. (2003). Continuous time bayesian networks for inferring users' presence and activities with extensions for modeling and evaluation. *Microsoft Research, July-August*.

Nodelman, U., Koller, D., and Shelton, C. R. (2012). Expectation propagation for continuous time bayesian networks. *arXiv preprint arXiv:1207.1401*.

Nodelman, U., Shelton, C. R., and Koller, D. (2002a). Continuous time bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 378–387. Morgan Kaufmann Publishers Inc.

Nodelman, U., Shelton, C. R., and Koller, D. (2002b). Learning continuous time bayesian networks. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 451–458. Morgan Kaufmann Publishers Inc.

O'cinneide, C. A. (1999). Phase-type distributions: open problems and a few properties. *Stochastic Models*, 15(4):731–757.

O'Shea, M. K., Koh, G. C., Munang, M., Smith, G., Banerjee, A., and Dedicoat, M.

(2014). Time-to-detection in culture predicts risk of mycobacterium tuberculosis transmission: A cohort study. *Clinical Infectious Diseases*, page ciu244.

Parunak, H. V. D., Savit, R., and Riolo, R. L. (1998). Agent-based modeling vs. equation-based modeling: A case study and users' guide. In *Multi-agent systems and agent-based simulation*, pages 10–25. Springer.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (2014). Is scientific knowledge useful for policy analysis? a peculiar theorem says: No. *Journal of Causal Inference J. Causal Infer.*, 2(1):109–112.

Pearson, K. (1896). Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the royal society of london*, 60(359-367):489–498.

Tiemersma, E. W., van der Werf, M. J., Borgdorff, M. W., Williams, B. G., and Nagelkerke, N. J. (2011). Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in hiv negative patients: a systematic review. *PloS one*, 6(4):e17601.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Tostmann, A., Kik, S. V., Kalisvaart, N. A., Sebek, M. M., Verver, S., Boeree, M. J., and van Soolingen, D. (2008). Tuberculosis transmission by patients with smear-negative pulmonary tuberculosis in a large cohort in the netherlands. *Clinical infectious diseases*, 47(9):1135–1142.

Uusitalo, L. (2007). Advantages and challenges of bayesian networks in environmental modelling. *Ecological modelling*, 203(3):312–318.

VanderWeele, T. J. and Robins, J. M. (2007). Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *American journal of epidemiology*, 166(9):1096–1104.

Vynnycky, E. and Fine, P. (1997). The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiology and infection*, 119(02):183–201.

Vynnycky, E. and Fine, P. (1999). Interpreting the decline in tuberculosis: the role of secular trends in effective contact. *International journal of epidemiology*, 28(2):327–334.