BAYESIAN INFERENCE IN EPIDEMIOLOGY-- MODELLING AND PREDICTION ON INFECTIOUS AND CHRONIC DISEASES

YANJI ZHAO

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Applied Mathematics

BAYESIAN INFERENCE IN EPIDEMIOLOGY-- MODELLING AND PREDICTION ON INFECTIOUS AND CHRONIC DISEASES

YANJI ZHAO

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

March 2024

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

_____YANJI ZHAO_____ (Name of student)

Acknowledgements

Appreciate pioneers. The thesis was shaped by my parents, supervisor Dr. He, and scholars' thoughts and wisdom I cited. I'd like to express my sincerest gratitude to my parents and supervisor for their unending love, encouragement, and unwavering belief in my potential. They jointly shaped the thesis, my knowledge and personality.

Appreciate latecomers. Any references and inspiration in the future is crucial to broaden the meaning and life of the research.

Appreciate colleagues. I would like to extend my heartfelt appreciation to my friends. I will always keep the experience of learning and struggle in my heart.

Pursuing PhD degree has been a fantastic journey of episodes. I would also move forward on the road of my life with love and wisdom from all of you.

List of Tables

List of Figures

Catalog

Abstract

Bayesian inference is an effective statistical inference method, which could deal with some uncertain problems and update our beliefs about unknown parameters with the integration of new evidence. However, one disadvantage of Bayesian statistics is that the Bayesian formula is more complex than the frequentist parameter estimation, thus it has become an appealing issue that how to choose an appropriate prior distribution to make Bayesian statistics easier to calculate for specific data. The primary objective of the thesis is to explore the performance of Bayesian inference for different data types in the aspect of epidemiology in three parts. As the representatives of Bayesian inference methods, Markov chains Monte Carlo and integrated nested Laplace approximation are explored for different types of data to focus on some practical issues in epidemiology: 1.) access the common features and variations of infectiousness of infectious disease in different contact settings; 2.) evaluate the effect of immigration on chronic disease mortality in the past and future along with the effects of age, period and cohort; 3.) due to their merits and weaknesses, evaluate the performance of Markov chains Monte Carlo and integrated nested Laplace approximation to determine preferred simulation methodology for different types of data in epidemiology.

The principle and application of one Bayesian inference method—Markov chain Monte Carlo (MCMC) with datasets of COVID-19 were emphasized to explore the estimated reproductive numbers and dispersion parameters, in order to 1) access the common features and variations of infectiousness in different settings, and 2) examine if there exist significant variation among individuals to investigate the association between community spread and superspreading events. MCMC performs satisfactory convergence and estimation as iterations increase for the short-term dataset of infectious disease.

Methodology of Bayesian inference on the long-term and large sample-sized data

of chronic disease was perform in this chapter. The mortality rates of lung, pancreatic, colon, liver, prostate and stomach cancers between locally born residents in Hong Kong and immigrants from mainland China were assessed, and we adopted a MCMC-free Bayesian age-period-cohort (APC) model based on integrated nested Laplace approximation (INLA) to explore the projection of mortality rates for the locally born population and immigrants in Hong Kong, taking into account age, period, and birth cohort effects as well. Compared to MCMC, INLA indicate higher computational efficiency, accuracy and flexibility for long-term data of chronic disease.

With similar data of chronic disease, some criteria, such as Continuous Ranked Probability Score (CRPS) and a calibration test were applied to evaluate the performance of retrospective projections based on MCMC and INLA. Two methods expound approximately significant performance on retrospective projections, and the projections based on INLA indicate less dispersion with observations than those based on MCMC in most of immigrations groups. Some circumstances, such as prostate cancer and stomach cancer, against the conclusion result from lack of data since INLA requires large sample size. The findings underscore the significance for targeted interventions and strict control measures for vulnerable populations to curb the spread of infectious diseases effectively, and we could reach to equal opportunities of optimal healthcare of cancers and other chronic diseases for every individual regardless of culture or background. Furthermore, the research demonstrates that the findings and conclusions can be also applied to other countries and regions with similar methods.

1. Introduction

1.1Background

Bayesian inference is a statistical inference method, combining subjective judgment with extensive calculations. It can effectively deal with some uncertain problems and update our beliefs about unknown parameters with the integration of new evidence. When we consider the potential of application of Bayesian statistics, different people will give different answers to a direct question that how to choose the prior distribution. However, one disadvantage of Bayesian statistics is that the Bayesian formula is more complex than the frequentist parameter estimation [1][2][3], thus it has become an appealing issue that how to choose an appropriate prior distribution to make Bayesian statistics easier to calculate.

Why is Bayesian statistics difficult to calculate? One of the most straightforward difficulties is that it is tough for us to ensure that the posterior distribution can be analytically solved [4]. When the posterior distribution can be analytically solved, it means that the density function of "the posterior distribution has an analytical solution". Another question is why posterior distribution sometimes doesn't have an analytical solution. It obviously does not necessarily have an analytical solution when the denominator of the specific Bayesian formula is an integral [5]. It means that the denominator with no analytical solution can be necessarily equivalent to the posterior distribution with no analytical solution, and the integral of this denominator with analytical solution is attainable as long as a suitable model and prior distribution can be selected. However, Bayesian statistics is an iterative process of continuously collecting data and updating parameter distributions. During the update process using Bayesian to calculate the posterior distribution, the last iteration will be set as the prior distribution of the new iteration each time [6][7]. Therefore, in order to ensure that the posterior distribution of each iteration can obtain an analytical solution, the same parameterized distribution family of the prior distribution and the posterior distribution could be

required, that is, ensure that the forms of prior distribution and posterior.

How to calculate the parameters of the generalized linear models with and without mixed effects models is another challenge, which has always been the focus of research. The random effects in the model are associated with spatial locations, and the number and specific coordinates of spatial locations directly affect the dimension of the spatial effects [8]. Furthermore, it is an essential challenge of computation as both Bayesian estimation and maximum likelihood estimation of parameters are inseparable from high-dimensional integration of spatial effects. Under the Bayesian method, the Metropolis program of random walks was proposed to implement Markov chain Monte Carlo (MCMC) algorithm to obtain the posterior density distribution and posterior quantity of the estimated parameters and values [9]. Additionally, Langevin-Hastings algorithm achieved better computational efficiency than the random walk Metropolis algorithm [10][11]. A subsequent robust version was developed by Christensen is given. In actual operation, the main problems faced by the Markov Chain Monte Carlo algorithm are convergence diagnosis and calculation time [12]. Of course, the algorithm implementation itself is also very important. For end users, most of them may not be good at it programming, so there may be problems in the algorithm implementation process. Therefore, it is also important to seek a good Bayesian inference tool or platform. Currently, models with random effects fitted through MCMC include software such as WinBUGS, OpenBUGS, JAGS, BayesX, MultiBUGS, and Stan [13][14][15][16]. In recent years, some researchers have begun to pay attention to the approximation of high-dimensional integrals, leading to the emergence of a new types of approximate Bayesian inference in Gaussian Malar under the setting of random field approximation of stationary spatial Gaussian process [17]. Laplacian is used to approximate the high-dimensional integral of spatial effects, thus proposing an integrated nested Laplacian algorithm, Lindgren et al. proposed a similar algorithm for parameter estimation of the SGLMM model when the random effect is a skewed distribution[18]. It was affirmed the use of the Laplace approximation method and believed that this type of approximation has sufficient accuracy and can be used for

actual data analysis.

Although it is computationally fast in some aspects, the most severe weakness of the Bayesian method is that it relies on the choices of prior distribution. Christensen also proposed the Monte Carlo maximum likelihood algorithm, which still relies on the MCMC algorithm, but provides likelihood analysis on parameters[19]. Its algorithm implementation is packaged in the R package geoRglm. As an alternative to Monte Carlo likelihood, Hao proposed the Monte Carlo Expectation Maximum algorithm (MCEM), which treats the part of spatial random effects that cannot be directly observed as missing data[20]. Due to its robustness and convenience, Bayesian inference has already been extensively applied on epidemiological modeling and analysis.

As the study of disease patterns with determinants within populations, epidemiology is the science which focuses on the distribution and determinants of diseases and health conditions among specific populations, as well as the research aspect of strategies and interventions to mitigate the spread of diseases and promote healthcare facilities [21]. In recent years, infectious diseases, represented by COVID-19, were widespread among the population around the world and brought great disasters, which lead to more in-depth epidemiological investigations and studies on infectious diseases [22]. Meanwhile, when major infectious diseases are gradually under control, epidemiologists have increasingly focus on researches on some non-communicable diseases, especially chronic diseases, such as cardiovascular and cerebrovascular diseases, malignant tumors, diabetes, and injuries and disabilities [23]. Due to the novel coronavirus SARS-CoV-2, the pandemic has posed unprecedented challenges to public health systems, governments and communities all over the world, which has necessitated a comprehensive understanding of the transmission dynamics of the epidemic, risk factors and effective public health interventions. Epidemiologists swiftly mobilized to understand key aspects of COVID-19, including its transmission modes, incubation period, and clinical manifestations [24]. With rigorous surveillance and contact tracing, epidemiologists have identified respiratory droplets as the primary

mode of transmission, emphasizing the importance of measures such as mask-wearing, physical distancing, and hand hygiene in reducing viral spread [25]. Additionally, evidence has emerged regarding the potential for airborne transmission, particularly in enclosed spaces with poor ventilation.

Bayesian inference has played an essential role in epidemiological modeling, and

Datasets and samples related to infectious and chronic diseases have also gave rise to Bayesian inference techniques. Based on the National Radioactivity Survey data of the Marshall Islands, Diggle et al. recorded the intensity data of 137Cs radiation particles on Rongelap Island in the South Pacific, and established an SGLMM model in which the response variable obeys the Poisson distribution [26]. Under the Bayesian method, Metropolis-Hastings sampling was used to implement the MCMC algorithm, obtain parameter estimates of the SGLMM model, and analyze the spatial distribution of the residual nuclear pollutant concentration [27]. In addition, they also established an SGLMM model with response variables obeying the binomial distribution to analyze the North Spatial distribution of Campylobacter infection among residents in Lanarkshire and South Cumbria. Christensen added non-spatial independent random effects to the model used by Diggle et al. The fitting effect, this non-spatial random effect is often called the nugget effect in geo-statistics. As focused on malaria data from Nyanza Province in Kenya, which combines school and village information. The analysis is a multi-source data, assuming that one of the data is biased. From a non-random survey, the other data is unbiased and comes from a random survey, so a spatial random effect containing two stationary spatial processes is established, and the binomial SGLMM model is estimated using the Monte Carlo maximum likelihood algorithm (MCML) [28][29]. parameters to obtain the spatial distribution of malaria in the province. The second data is malaria data collected from May 2010 to June 2013 in Wawa District, Malawi. Diggle et al., considered the binomial SGLMM model, and they assumed that the time term and the space term are independent, and the nugget effect only depends on time changes [13]. Also based on the MCML algorithm, each parameter of the model is estimated; The third data modeling is based on SGLMM with

nugget effect [30]. It is believed that the response variable should obey a mixed binomial distribution to contain very low infection levels. For example, none of the villages are infected, so zero excess binomial spatial mixed effects model analysis of the third river blindness data set.

When faced with complex high-dimensional integrals, each alternative method, whether taking the route of random simulation or approximation, has a corresponding cost. Resulted from the property of the hierarchical models and random effects stated above, the convergence issues of MCMC leads to the development of Integrated Nested Laplace Approximation (INLA) as an alternative method to fit Bayesian hierarchical models within the latent Gaussian model [31]. The method based on Laplace approximation relies on the selection of initial values, while MCMC are random simulation. The algorithm relies on the adjustment of prior distributions and algorithm parameters, which will have an impact on the final data analysis results [32][33]. The process of adjusting parameters is often full of experience and skills. Although new and complex algorithms and methods are constantly being developed, Bonat and Ribeiro believed that only parameter estimation methods that can be widely used and be relatively straightforward implemented are more general and more reliable to select. Therefore, the inference methods according to different criteria and data types has already been an essential issue in epidemiology [34][35].

1.2 Objectives and significance

Since the choice of an appropriate prior distribution is essential to calculate in Bayesian inference for specific data, the primary objective of the thesis is to explore the performance of Bayesian inference for different data types in the aspect of epidemiology in three parts. As the representatives of Bayesian inference methods,

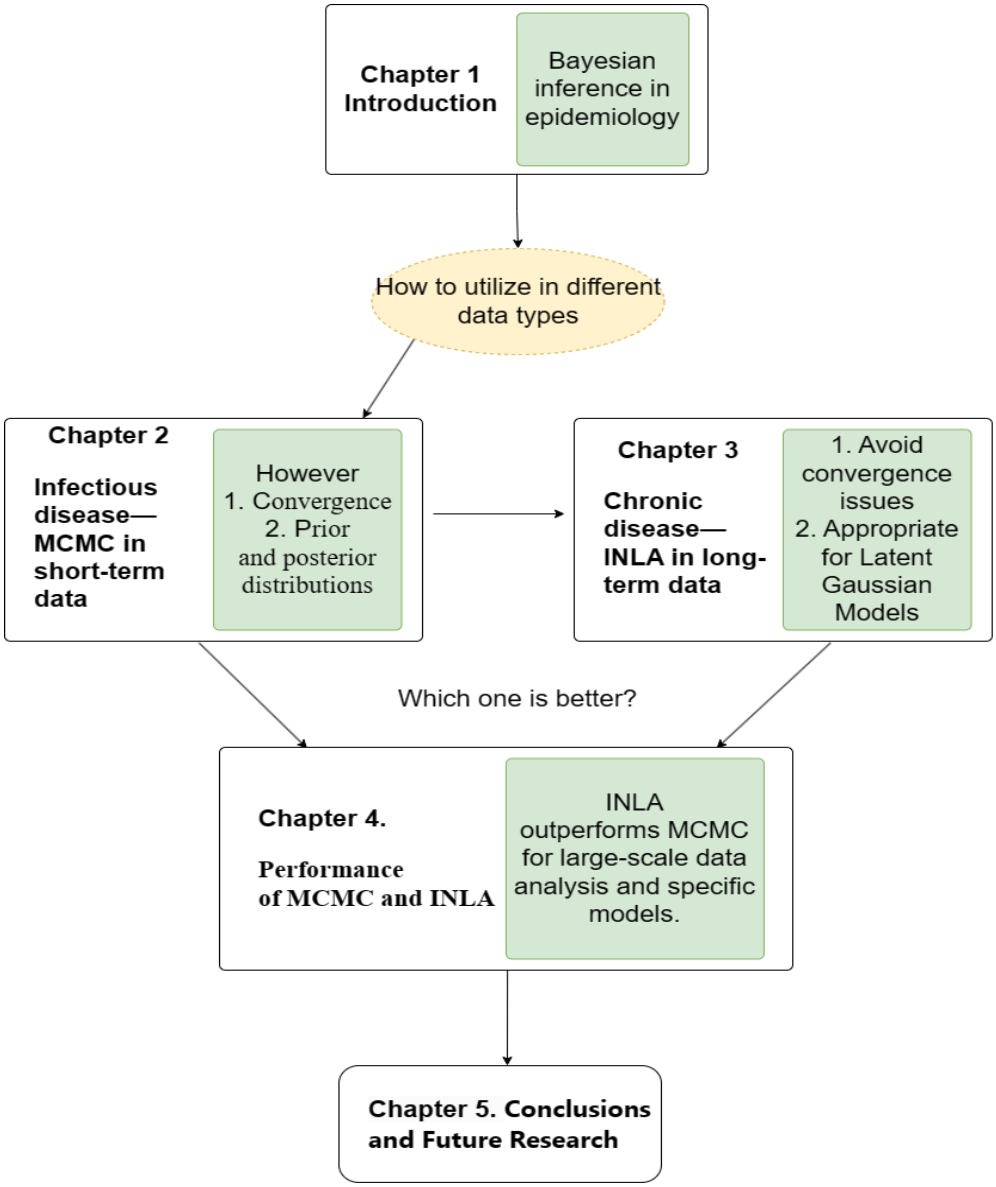
Markov chains Monte Carlo and integrated nested Laplace approximation are explored for different types of data to focus on some practical issues in epidemiology: 1.) access the common features and variations of infectiousness of infectious disease in different contact settings; 2.) evaluate the effect of immigration on chronic disease mortality in the past and future along with the effects of age, period and cohort; 3.) due to their merits and weaknesses, evaluate the performance of Markov chains Monte Carlo and integrated nested Laplace approximation.

1.3 Publications and Contributions

The thesis is mainly arisen from two publications "Differences in the superspreading potentials of COVID-19 across contact settings" and "Age-period-cohort analysis and projection of cancer mortality in Hong Kong, 1998–2030". As the first author of these two papers, I was in charge of methodology, formal analysis, data curation, writing draft and visualization of them.

1.4 Outline

Background, motivations and objectives of this thesis are introduced in chapter 1. In chapter 2, with the reference of one of my published paper "Differences in the superspreading potentials of COVID-19 across contact settings", I sets forth the methodology of Bayesian inference, based on MCMC for infectious disease data. Due to some weaknesses of MCMC, I present a MCMC-free Bayesian APC model for chronic disease data, based on INLA in chapter 3, with the reference of another published paper "Age-period-cohort analysis and projection of cancer mortality in Hong Kong, 1998–2030". Evaluation of the performance related to MCMC and INLA in epidemiology are shown in chapter 4. Discussions and conclusions are presented in chapter 5. The flow chart illustrates the outline in details.


Chapter 1
Introduction
Bayesian inference in epidemiology
How to utilize in different data types
Chapter 2
Infectious disease—MCMC in short-term data
However
1. Convergence
2. Prior and posterior distributions
Chapter 3
Chronic disease—INLA in long-term data
1. Avoid convergence issues
2. Appropriate for Latent Gaussian Models
Which one is better?
Chapter 4.
Performance of MCMC and INLA
INLA outperforms MCMC for large-scale data analysis and specific models.
Chapter 5. Conclusions and Future Research

2. Infectious disease—MCMC in short-term data

Markov chain Monte Carlo (MCMC) is a powerful computational method used in Bayesian inference, and it has been widely applied in epidemiology, which allows sampling from complex probability distributions by constructing a Markov chain that has the desired distribution as its equilibrium distribution. This makes MCMC particularly useful for estimating complex models that are not analytically tractable.

One of the main advantages of MCMC is its flexibility. It can handle a wide range of models, including those with complex hierarchical structures, non-linear relationships, and high-dimensional parameter spaces. This flexibility makes MCMC a valuable tool in epidemiology, where complex models are often needed to capture the intricate relationships between various risk factors and health outcomes. Another advantage of MCMC is that it provides a full posterior distribution of the parameters, rather than just point estimates [23]. This allows for a more comprehensive assessment of uncertainty, which is crucial in epidemiological studies.

Furthermore, maximum likelihood estimation (MLE) is a classical method of statistical estimation that is also used in epidemiology. While MLE has its own advantages, such as simplicity and consistency under certain conditions, it also has several disadvantages compared to Bayesian methods like MCMC. One of the main disadvantages of MLE is that it only provides point estimates of the parameters, without a direct measure of uncertainty [24][25]. While confidence intervals can be constructed, they are based on asymptotic approximations and may not be accurate for small sample sizes or complex models.

Another disadvantage of MLE is that it does not allow for the incorporation of prior knowledge or beliefs. In contrast, Bayesian methods like MCMC allow for the integration of prior information through the use of prior distributions, which can be particularly useful in epidemiological studies where prior information is often available.

The development of MCMC and other Bayesian methods in the academic language has been facilitated by the availability of software packages, such as WinBUGS and JAGS for MCMC [27]. These packages have made Bayesian inference more accessible to epidemiologists and have contributed to the increasing use of Bayesian methods in epidemiology.

In this chapter, I emphasize the principle and application of one Bayesian inference method—MCMC with datasets of infectious disease.

2.1 Introduction

In the past few years, spurred by the increasing burden of Coronavirus disease 2019 (COVID-19) outbreaks, researchers have been concentrating on characterizing superspreading events caused by multiple variants of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [36][37][38].Meanwhile, the aggregation of transmission for some superspreading cases also draw researchers' attention, defined as 20/80 rule [39] in epidemiology, which implies that approximately 80% secondary infected cases and transmissions result from roughly 20% of primary cases. Additionally, cluster infections involved in superspreading events were generally adjudged to be responsible for the epidemic and its rapid evolution[36][40]. Therefore, exploring and summarizing the inherent dynamics in transmission chains of SARS-CoV-2 can conduce to more effective and enhanced interventions and prevention from the epidemic.

During the past few years, the coronavirus disease 2019 (COVID-19) that caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been continuously spreading worldwide, posing a significant threat to public health. A comprehensive understanding on the epidemiological characteristics of COVID-19

underlies the strategic development of region-wide control policies to combat the epidemics. The fundamental biological parameters – basic reproduction number (R_0) and effective reproduction number (R) describe the transmission potential of a typical infectious disease agent, that is, the average number of secondary cases generated by an infectious person in a completely and not completely susceptible population, respectively[36]. While for the COVID-19 epidemics, the differences arose in infectiousness, behavioral patterns and locally implemented public health interventions give rise to heterogeneous individual transmissibility[37][38], which cannot be reflected by a single measurement of R_0 [39].

A superspreading event (SSE) is defined as a transmission event involving an unusual large number of cases, initiated by the super-spreader. The SSE represented a heterogeneous transmission pattern, where the majority of the cases were seeded by a small fraction of super-spreaders [40][41]. As a distinct feature of the transmission dynamics of COVID-19, SSEs played essential roles in aggravating the COVID-19 epidemics. For instance, in early November 2021 in Hong Kong, an outbreak in the community was caused by a few SSEs in entertainment places, which led to a major epidemic wave in the whole city [42]. In South Korea, the SSE seeded by the SARS-CoV-2 Omicron variants occurred in churches and schools, causing the disease to spread widely in the local community [43]. Characterizing the superspreading potential of the epidemics in the context could give policymakers a hint on how to effectively curb the local transmissions [44]. For example, identifying and shutting down the hot-spot contact settings favoring the occurrence SSE (e.g., bars, social parties, and gyms) could timely chop the transmission chains and prevent future large outbreaks.” As a forceful circumstantial evidence of community transmission and SSEs, Furuse et al. exemplified demographic information regarding some clusters of COVID-19 infectors and schematized their features in transmission chains from January to July 2020 in Japan with different contact settings of SSEs [45].

2.2 Objective

As a representative research, it exemplified demographic information regarding some clusters of COVID-19 infectors and schematized their features in transmission chains from January to July 2020 in Japan with different settings of superspreading events [41]. This study sought to explore the estimated reproductive numbers and dispersion parameters of rearranged contact tracing data in transmission chains in Japan from [45]. The objective was to 1) access the common features and variations of infectiousness in different settings, and 2) examine if there exist significant variation among individuals to investigate the association between community spread and superspreading events.

2.3 Data and methods

2.3.1 Data

28 circumstances of transmission clusters of COVID-19 from January to July 2020 in Japan were retrieved [45]. Based on the contact tracing and exposure history of each cases within the transmission clusters, 545 infectee-infector transmission pairs were constructed. We thereafter extracted the number of secondary cases (i.e., infectees) that were directly generated by each infector for further analysis. We excluded the cases that are indirectly linked with the infectors. The identified transmission pairs were further grouped by different contact settings (i.e., community, health care facility, school, household, and workplace) according to where the transmission occurred, and those without detailed information regarding contact settings were also omitted. Counts of cases in three age groups (0–19, 20–59, and 60 or more) were also recorded.

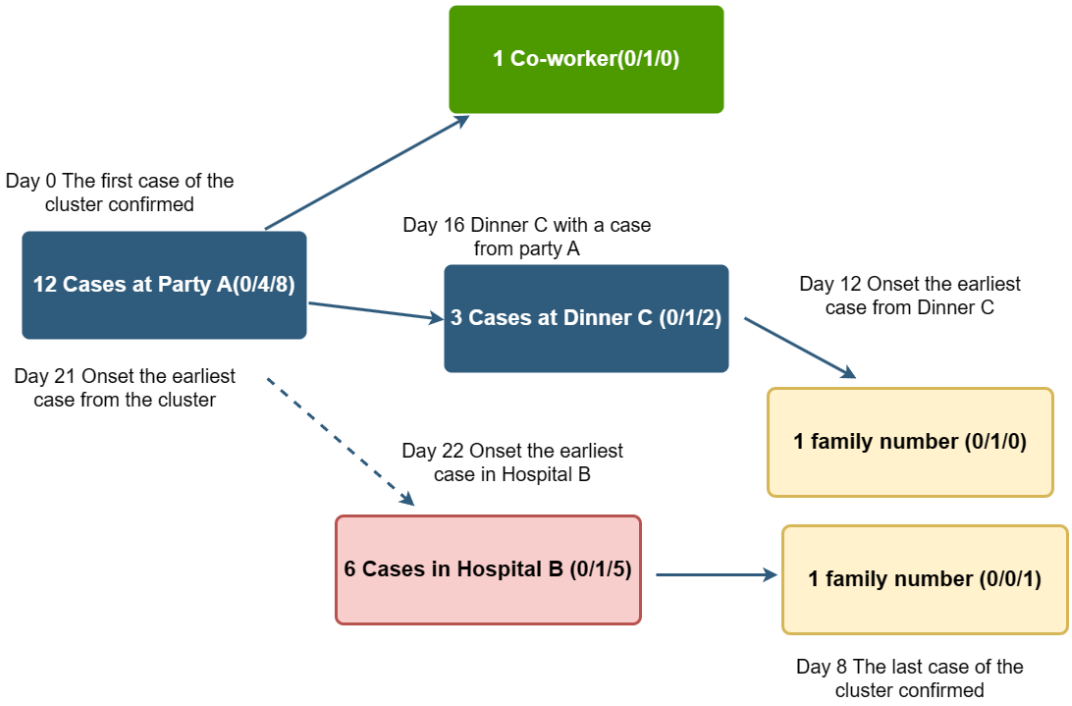


Figure 1. One instance of 28 circumstances of transmission clusters. The blue, green, red, and yellow boxes represent cases at community superspreading events, cases among co-workers, cases at hospitals/care facilities/schools, and cases among family members, respectively. Arrows indicate infector-infectee transmission pairs. A solid and dashed arrow line indicate direct and indirect transmission chains. Values in square brackets denote the number of patients aged 0–19, 20–59, and 60 or more. Arrows indicate infector-infectee transmission pair.

2.3.2 Methods

To quantify the superspreading potential, we assumed the number of secondary cases seeded by each infector following a Negative binomial distribution [6], which

was parameterized by an effective reproduction number (R) as mean and a dispersion parameter (k). The k captured the heterogeneity in the individual transmissibility. A lower value of k indicated a higher transmission heterogeneity, and thereby a higher superspreading potential. The number of offspring cases generated by each seed case was fitted to a negative binomial model. For the model parameter estimation, Markov chain Monte Carlo (MCMC) method was applied to estimate the joint posterior distribution of R and k.

The proportion of the most infectious cases that seeded 80% of the total transmissions was calculated [46]. The probability that a seed case generates a cluster with size 10 or more and the probability of observing SSEs were also computed. In addition to incorporating the expected proportion of infectors generating at least one infected individual and the probability that a seed case generates a cluster with size 10 or more, some intuitive concepts, such as the proportion of the most infectious infectors responsible for 80% of infectees and the expected probability of superspreading events, were also attained based on estimated [47][48][49]. Followed by previous work [41], we defined the threshold of SSEs as the 99-th percentile of the Poisson distribution with the rate at reproduction number. Any transmission event that is seeded by a single infector would be counted as an SSE if the number of secondary cases exceeds the threshold. We thereafter calculated the probability of observing SSEs seeded by a single infector according to the SSE threshold. Subgroup analysis in different contact settings was also conducted in the same procedure to obtain the above estimates.

Secondary case distribution in the context of superspreading

Given the stochastic effect of the transmission events, the transmission dynamics can be modelled by a Poisson process, such that the number of secondary cases Y generated by each infector is described by a Poisson distribution a mean of λ [50]. To characterize the heterogeneity in individual transmissibility, the λ was assumed to

follow a gamma distribution and thereby yield a Negative binomial secondary case distribution parameterized by the reproduction number (R) and a dispersion parameter (k) [41]. Therefore, the probability that an infector generates $y(\geq 0)$ secondary cases is given by

$$f(y) = Pr(Y = y) = \frac{\Gamma(k+y)}{y!\Gamma(k)}\left(\frac{k}{k+R}\right)^{k}\left(\frac{R}{k+R}\right)^{y} \quad (1)$$

where $\Gamma(y)$ is the gamma function satisfying that $y! = \Gamma(y+1)$. When a transmission cluster involves x_i infectors who seeded a total of y_i secondary cases, then the above function is adjusted as [51][52]:

$$r(x_i, y_i) = Pr(Y = y_i; R, k) = \frac{\Gamma(k * x_i + y_i)}{\Gamma(y_i + 1)\Gamma(k * x_i)}\left(\frac{k}{R+k}\right)^{k*x_i}\left(\frac{R}{R+k}\right)^{y_i} \quad (2)$$

Then, the likelihood function L based on the dataset with totally n transmission pairs is

$$L = \prod_{i=1}^{n} r(x_i, y_i) \quad (3)$$

Parameter estimation

The Markov Chain Monte Carlo (MCMC) method was employed to jointly estimate the reproduction number R and dispersion parameter k based on formula (3). Metropolis-Hastings algorithm was adopted and the marginal posterior distributions were obtained from 110 000 MCMC iterations, among which the first 30 000 were discarded as burn-in. Uniform prior distributions were applied for R and k. The 95% credible intervals were drawn from the marginal posterior distributions. Trace plot and

Gelman–Rubin convergence diagnostic were used for checking the convergence of posterior reproduced based on the MCMC [53]. The illustration of convergence diagnostic can be found in **Appendices A1.** Iterations of the whole population and contact settings shows no significant differences among within and between variance. (eFigure 1) All statistical analyses were performed in **R** version 4.2.1 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2013, http://www.R-project.org/).

Measurements of superspreading potentials

Armed with the estimated *R* and *k* values, we deduced the proportion of the most infectious cases responsible for 80% of total transmissions, which was formulated as per [54]:

$$P_{80\%} = 1 - \int_0^Z f(\lfloor z \rfloor) dz \tag{4}$$

where $\lfloor\ \rfloor$ represents the floor function and Z satisfies

$$1 - 0.8 = \frac{1}{R} \int_0^Z \lfloor z \rfloor * f(\lfloor z \rfloor) dz \tag{5}$$

Followed by previous work [54], we defined the threshold of SSE for the COVID-19 as the 99th percentile of the Poisson distribution of the basic reproduction number (R_0). Given that a consensus R_0 estimates were in a range of 2 to 3 [55], the threshold of SSE was determined to be 6 to 8. The threshold was assumed to be 6 in this study. Any transmission event that is directly seeded by a single infector would be counted as an SSE if the number of secondary cases exceeds the threshold (i.e., 6).

Then, the probability of observing an SSE seeded by a single infector is given by

$$P_s = 1 - F(y; R, k)|_{y=5 \text{ or } 7} \quad (6)$$

Here, $F(.)$ is the cumulative probability function of equation (2).Furthermore, based on the methods derived in [54][56], when we refocus on the final cluster size with the assumption that the offspring distribution are independently and identically distributed (iid) negative binomial distribution given by equation (2), the possibility mass function for the final size s of th cluster caused by x initial cases is $c(s,\ x)$ given by:

$$c(s,\ x) = \frac{kx}{ks+s-x}\binom{ks+s-x}{s-x}\left(\frac{k}{k+R}\right)^{ks}\left(\frac{R}{k+R}\right)^{s-x} = \frac{x}{s}r(s,\ s-x) \quad (7)$$

Therefore, the probability of x seed cases resulting in a cluster with size s or more is $1-\sum_{x=1}^{s-1} c(s,x)$.

2.4 Results

A total of 545 transmission pairs were constructed from the reported 28 transmission clusters. Of the settings where the identified transmission pairs occur, 31.1%, 25.6%, 28.7%, 4.0%, and 10.6% belonged to the community, household, health care facility, school and workplace, respectively. Among 1017 identified infectors, 75.0% of them lead to no secondary cases, and 0.8% of them directly generated more than 10 cases. From the observed secondary case distribution and

fitted negative binomial models, we estimated that the overall *R* and *k* were 0.561 (95% CrI: 0.496, 0.640) and 0.221 (95% CrI: 0.186, 0.262), 0.107 (95% CrI: 0.046, 0.331) and 0.004 (95% CrI: 0.002, 0.007) for community setting, 0.137 (95% CrI: 0.110, 0.168) and 0.141 (95% CrI: 0.098, 0.210) for household setting, 0.186 (95% CrI: 0.079, 0.409) and 0.004 (95% CrI: 0.002, 0.006) for healthcare facilities setting, 0.088 (95% CrI: 0.028, 0.295) and 0.002 (95% CrI: 0.001, 0.005) for school setting, 0.080 (95% CrI: 0.052, 0.138) and 0.019 (95% CrI: 0.012, 0.029) for workplace setting, respectively (Table 1). Based on the estimated *R* value, the threshold of SSEs was determined to be 6, and there were 17 out of 500 (3.4%) transmission events identified as SSEs. We inferred that 80% of total transmissions were generated by 13.14% (95% CrI:11.55%, 14.87%) of the most infectious seed cases.

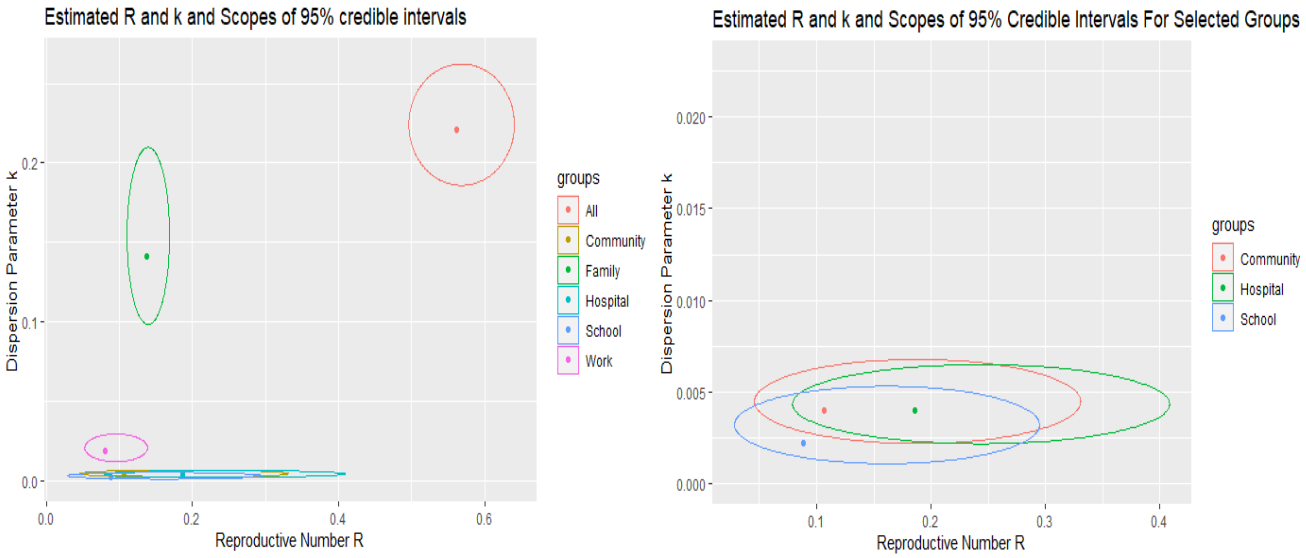


Figure 2. Estimated reproduction numbers R and dispersion parameters k of total population and five types of contact settings with 95% credible intervals. Right figure is the zoom-in of three selected contact settings.

Across all contact settings, the health care facility and household had a higher risk of transmission (larger value of R) whereas school, health care facility, and community had a higher superspreading potential (smaller value of k). The probability that an infector generates at least one secondary case was 24.37% (95% CrI: 21.47, 27.68), where cases in household setting were more likely to generate more than one infectors (9.13%, 95% CrI: 7.11, 11.61). Furthermore, the probability of observing SSEs with a predefined threshold is 1.75% (95% CrI:1.57, 1.99), while probabilities of SSEs in different contact settings are approximately comparable with a higher probability in health facilities (0.33%, 95% CrI: 0.13, 0.94). The probability that a seed case generates a transmission cluster with a size of 10 or greater is 3.87% (95% CrI: 2.94, 5.24). Other epidemiological results for mentioned contact settings are shown in Table 1.

	Total	Community	Household	Healthcare facilities	School	Workplace
Reproduction number (*R*)	0.561 (0.496, 0.640)	0.107 (0.046, 0.331)	0.137 (0.110, 0.168)	0.186 (0.079, 0.409)	0.088 (0.028, 0.295)	0.080 (0.052, 0.138)
Dispersion parameter (*k*)	0.221 (0.186, 0.262)	0.004 (0.002, 0.007)	0.141 (0.098, 0.210)	0.004 (0.002, 0.006)	0.002 (0.001, 0.005)	0.019 (0.012, 0.029)
Probability of 1 infector generating ≥1 infectees	24.37% (21.47, 27.68)	1.32% (0.63, 2.68)	9.13% (7.11, 11.61)	1.53% (0.74, 2.51)	0.76% (0.34, 2.03)	3.09% (1.99, 4.95)
Proportion of infector seeding 80% transmission	13.14% (11.55, 14.87)	0.44% (0.21, 0.76)	6.39% (4.91, 8.24)	0.44% (0.22, 0.66)	0.22% (0.11, 0.55)	1.54% (0.99, 2.43)
Probability of observing SSE	1.75% (1.57, 1.99)	0.49% (0.22, 1.18)	0.07% (0.06, 0.08)	0.67% (0.31, 1.21)	0.33% (0.13, 0.94)	0.32% (0.21, 0.60)
Probability of cluster size ≥10 seeded by 1 infector	3.87% (2.94, 5.24)	0.37% (0.16, 1.00)	0.05% (0.04, 0.06)	0.55% (0.24, 1.04)	0.26% (0.09, 0.80)	0.17% (0.10, 0.38)

Table 1. Summary of the estimated metrics of superspreading potential under different contact settings. The metrics were summarized as 'median estimate (95% CrI)' format.

2.5 Discussion

Characterizing the superspreading potential could provide a better understanding of the transmission potential of the COVID-19 pandemic and help to formulate targeted public health interventions. In this study, using transmission cluster data collected during the early phase of the epidemic in Japan, we assessed the superspreading potential of COVID-19 within different contact settings.

We found that the early epidemics in Japan exhibited a significant superspreading potential (k=0.22), which is in line with another study conducted during a similar study period (k=0.23) [41], but is smaller than an estimate obtained in Hong Kong (k=0.43) [16]. This discrepancy could be attributed to the differences in imposed control policies. In Japan, cluster-based measures that focused on identifying and preventing transmission clusters were adopted to curb the epidemics [45]. On the other hand, a series of social distancing interventions including school closure, work-from-home-policy, and cancellation of mass gatherings were implemented in Hong Kong [57], which may have a greater effect on reducing the potential of societal SSEs [58] and thus resulting in a relatively higher k. It was also concluded in [10] that rare superspreading events in community resulted from infectors from hospitals, healthcare facilities or schools, whereas some cases in hospitals, healthcare facilities or schools were caused by the transmission chains originated from community superspreading events, which may lead to a low dispersion parameter in the distribution of offspring from communities. Meanwhile, the super-aged society in Japan [45] can also be deemed as the underlying cause of the estimates in each setting.

We also found that the risk of transmission and superspreading potentials varied across different contact settings. The higher estimated superspreading potential in school and community is consistent with a study conducted in South Korea, whereby the transmission chains in community and schools were more heterogeneous (smaller

k) than that in the household [59].

In conclusion, the early COVID-19 epidemics in Japan demonstrated a significant potential of superspreading. Particularly, the school, health care facility and community had relatively higher potential of superspreading when compared to other contact settings. The different potential of superspreading in contact settings highlights the need to continuously monitoring the transmissibility accompanied with the dispersion parameter, to timely identify high risk settings favoring the occurrence of SSE.

The findings underscore the significance for targeted interventions and strict control measures of infections for specific circumstances to curb the spread of infectious diseases effectively, including implementing strict healthcare protocols, promoting vaccination campaign, enhancing ventilation systems, ensuring adequate distancing measures, and conducting regular testing to mitigate the risk of outbreaks. Additionally, tailored public health campaigns and education may also be necessary to raise awareness to the infectious diseases and foster a culture of health and safety within these high-risk contact settings, which ultimately safeguards individuals and communities.

There are also some limitations in this study. Firstly, the transmission cluster data used was subjected to any bias (e.g., recall bias) generated during the contact tracing process and thus it is plausible that some cases that are exposed to the clusters were missed. This imperfect case ascertainment may lead to an underestimation of the *R* value but an overestimation of the *k* value [60][61]. Secondly, disproportional attention to infectors who generated infectees or not may have resulted in that infectors generating infectees were more likely to be collected and reported. Besides, the transmission clusters included in our study occurred during the early stage of the COVID-19 epidemics. Finally, more types of contact settings combination can be considered when some places are interconnected through ventilation. Given that the current epidemics are dominated by the SARS-CoV-2 Omicron variants, further study

is warranted to assess the superspreading potential of the emerging variants in Japan to help with formulating control policy.

With the in-depth study and investigation of the algorithm of Markov chains Monte Carlo, we can conclude that MCMC is more universal and general than integrated nested Laplace approximation because MCMC can be applied for all models, theoretically. However, it requires the user to be very proficient in the methodology and hyperparameters of MCMC to avoid misjudgment of prior and posterior distributions so that it would be crucial to select or design the appropriate sampler. When the model is complex, such as spatial model, or the amount of data is huge, the convergence speed would also be extremely slow. Therefore, it can be used in large-scale data analysis, theoretically. Meanwhile, INLA can only be used with Latent Gaussian Models (LGMs) [62]. However, LGMs are very universal and involve with a lot of amounts of models, including many spatial models, time models, space-time models. specifically, INLA can be applied for the inference issues in linear models, generalized linear models, linear mixed models, generalized linear mixed models, generalized additive models and survival models. Furthermore, with lower computation consumptions, the accuracy of INLA also outperforms MCMC for large-scale data analysis and specific models.

3. Chronic disease—INLA in long-term data

INLA is a Bayesian inference method based on Laplace approximation, which approximates the posterior distribution into an analytical form of distribution and avoids the large number of Monte Carlo sampling required in traditional Bayesian inference. INLA transforms the Bayesian inference problem into an approximation problem of solving Gaussian Markov random fields by decomposing parameters into fixed effects and random effects and exploiting the properties of Gaussian Markov random fields.

Compared to some traditional Bayesian inference methods such as MCMC, INLA has the following advantages [63]:

1). High computational efficiency: INLA uses the Laplace approximation method, which avoids the large number of Monte Carlo sampling required in traditional methods, with lower computation consumptions.

2). High accuracy: With high computational efficiency, the posterior distribution is more accurately approximated, so more accurate inference can be obtained.

3). Higher flexibility: It can be applied to a variety of different models, including linear models, generalized linear models, and some nonlinear models, etc.

In this chapter, I perform the methodology of Bayesian inference on the long-term and large sample-sized data of chronic disease.

3.1 Introduction

Several migration waves from mainland China to Hong Kong have occurred over

the past century. These migration waves included a large-scale migration inflow from 1945 to 1950 (the Chinese Civil War) and a few small-scale inflows in the 1950s, 1970s, and 1990s [64][65][66]. In 2016, immigrants from mainland China formed approximately 38% of the population of Hong Kong. These inflows have led to a growing interest in research on the disparity of health conditions between the locals and immigrants.

Cancer has been one of the most common causes of death, as an estimated 19.3 million new cancer cases and 9.9 million new cancer-associated deaths occurred worldwide in 2020 [67]. In Hong Kong, lung cancer is one of the most common causes of cancer deaths [68]. Previous studies suggested that the primary cause of lung cancer is cigarette smoking [69][70][71][72]. Genetic factors, asbestos, radon gas, second-hand smoke, and other forms of air pollution have been proven to influence the risk of lung cancer [73-79]. The overall daily smoking rate in mainland China was approximately 23.2% in 2018 [80], whereas the daily smoking rate in Hong Kong was only 10.2% in 2019 [81]. The leading causes of liver cancer include viral infection, drinking of alcohol and polluted water and food supplies which are also culprits for colon, stomach and pancreatic cancer [82]. Alcohol consumption per capita in Hong Kong has reached 2.37 liters in 2021 [83], compared to 7.0 liters of per capita consumption of alcohol in mainland China in 2018 [84]. As approximately 99% of prostate cancer cases occur after age 50, factors of prostate cancer have been regarded as old age, race, family history and the diet of red meat consumption [85]. In addition to these risk factors, studies have suggested that cancer mortality rates vary depending on migrant status [86-89].

According to data from the Census and Statistics Department of Hong Kong, approximately 81% of immigrants in Hong Kong immigrated from mainland China, Macau, and Taiwan. Immigrants from mainland China account for the bulk of this population. Previous studies have shown that child immigrants in Hong Kong tend to suffer from a higher risk of wheezing disorders and cardiovascular diseases, and immigrant women have higher age-specific mortality rates of breast cancer than locally-

born women in Hong Kong [90][91]. However, to date, few studies have investigated the effect of length of stay in Hong Kong and birthplace on the risk of other types of cancer.

3.2 Objective

In this part, we compared the mortality rates of lung, pancreatic, colon, liver, prostate and stomach cancers between locally born residents in Hong Kong and immigrants from mainland China. Both populations are widely considered as ethnically homogeneous with similar cultures. Nevertheless, due to different early life experiences, immigrants are exposed to more various social economy and lifestyles than locals. Therefore, it's constructive to ascertain whether immigrants from mainland China have a different mortality pattern of cancers from locals to verify the significance of migration status for this health outcome. As Age-period-cohort (APC) analysis plays a vital role in studying time-specific phenomena in epidemiology. in this study, to evaluate the effect of immigration on cancer mortality in the past and future, we developed APC models specified by sex and migrant status to assess the effects of age, period, birth cohort, and of the length of stay in Hong Kong on the mortality risks of cancers. Additionally, we explore the projection of mortality rates for the locally born population and immigrants in Hong Kong who were younger or older than 60 using a predictive model, taking into account age, period, and birth cohort effects as well.

3.3 Data and methods

3.3.1 Data

We obtained the death registry data, related to six types of cancer: lung cancer, colon cancer, liver cancer, stomach cancer, pancreatic cancer for males and females and prostate cancer for males, in Hong Kong between 1998 and 2021 from the Census and Statistics Department of Hong Kong, as the data in 2022 has not been available up to now. The data was extracted from a routine census held by the Hong Kong government as subjective errors caused by resampling can be neglected. The population data were stratified by age, sex, immigration status, and length of stay in Hong Kong. We retrieved six types of cancer cases from the death registry data using ICD codes, such as ICD-9 code 162 and ICD-10 codes C34.0–C34.3, C348, and C349 for lung cancer. To assure comparability among registries, deaths from the age group of 35–85 years were selected, since cases younger than 35 and older than 85 were relatively trivial for lack of statistical interpretability [92]. Cases were also divided into different age groups, such as younger or elder than 60, to explore the projection of mortality rates for the locally born population and immigrants in Hong Kong who were younger or older than 60 using a predictive model, taking into account age, period, and birth cohort effects as well.

Immigration status was classified into three groups: locals born in Hong Kong, immigrants who have lived in Hong Kong for >10 years before death defined as long-stay immigrants, and immigrants who have lived in Hong Kong for ≤10 years before death defined as short-stay immigrants. Notably, much focus was placed on immigrants from mainland China, because approximately 81% of immigrants in Hong Kong came from mainland China, Macau, and Taiwan based on the data from the Census and Statistics Department of Hong Kong. Moreover, few cases recorded from Macau and Taiwan are statistically insignificant in the analysis. Demographics and population projections from 2022 to 2030 were retrieved from the Census and Statistics Department of Hong Kong and estimated with cubic smoothing spline as the

prerequisite of the predictive model. Combined data of all types of cancer was omitted due to inevitable incomparability of results and unbalanced dataset for each type of cancer. For example, more than 20 long-stay male deaths at age 59 for lung cancer in each year, whereas there were no long-stay male deaths at age 59 for prostate cancer for some years. It would lead to merge and mitigate the effect of age and immigration history. They are introduced in detailed in “Methods” below.

3.3.2 Methods

Cubic smoothing spline

Data of demographic census is exposed every five years from Census and Statistics Department of Hong Kong, thus we obtained the data of population and immigrants in Hong Kong for each year until 2030 based on cubic smooth spline. Theoretically, smooth spline regression is a local modeling method, which is a continuous piece-wise polynomial based on certain smoothness. In order to obtain accurate estimates of population for each year, the idea of cubic smooth spline was considered that penalization was introduced to minimize residual sum of squares (RSS) based on ordinary least squares (OLS) such that

$$min\ s(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx \tag{8}$$

where f is the piece-wise cubic spline function and $\lambda \int (f''(x))^2 dx$ is the penalization. The detail of missing data imputation of immigrants population for each year is demonstrated in **Appendices A2.**

Age-period-cohort effects between immigration groups, 1998-2021

We modeled cancer mortality rates in Hong Kong using APC analysis based on log-linear Poisson regression models. The model aimed to disentangle age, period, and cohort effects of time-varying phenomena simultaneously [93][94], given that

$$\begin{aligned}\log(\mu_{apc}) = \beta_0 + (\alpha_L + \pi_L)\left(a - \bar{a}\right) + (\pi_L + \gamma_L)(c - \bar{c}) \\ + \tilde{\alpha}_a + \tilde{\pi}_p + \tilde{\gamma}_c + \log(O_{\text{apc}})\end{aligned} \tag{9}$$

where β_0 is the intercept with the mean age effect of $\bar{a}$ and mean birth cohort effect of $\bar{c}$, α_L , π_L , γ_L represent the combination of linear trend in age, period and cohort effects, $\tilde{\alpha}_a$, $\tilde{\pi}_p$, $\tilde{\gamma}_c$ are deviation (curvature) parameters and capture nonlinear patterns of observed rate. $\log(O_{\text{apc}})$ represents high-order residual. Finally, curvature parameters satisfy the following conditions in order to have identifiability, such that

$$\begin{aligned}\sum_a \tilde{\alpha}_a = \sum_p \tilde{\pi}_p = \sum_c \tilde{\gamma}_c = 0 \quad \text{and} \\ \sum_a \tilde{\alpha}_a(a-\bar{a}) = \sum_p \tilde{\pi}_p(p-\bar{p}) = \sum_c \tilde{\gamma}_c(c-\bar{c}) = 0.\end{aligned} \tag{10}$$

Model (9) parameterization is identifiable. Notably, no prior assumption on the magnitudes of α_L , π_L , γ_L is required. Furthermore, $\alpha_L+\pi_L$ is the so-call "longitudinal age trend" and $\pi_L + \gamma_L$ is called the "net drift parameter", as well as the difference $\alpha_L - \gamma_L$ is called the "cross-sectional age trend". Based on the model (9), we propose to incorporate "age when arrived in Hong Kong" and "Years in Hong Kong" into intercept, age trend and drift as following:

$$\begin{aligned}\log(\mu_{apc}^{(s)}) = (\beta_0 + u_0^{(s)}) + (\alpha_L + \pi_L + u_a^{(s)})\left(a - \bar{a}\right) + (\pi_L + \gamma_L + u_c^{(s)})(c - \bar{c}) \\ + \tilde{\alpha}_a + \tilde{\pi}_p + \tilde{\gamma}_c + \log(O_{apc}^{(s)})\end{aligned} \tag{11}$$

Here, the choice of group would be “immigrant from mainland” and “locally born resident”. We further divided the “immigrant from mainland” into subgroups, such as “age when entered HK less than 12 years old” and “age when entered Hong Kong later than 12 years old”. Every ten years bin was another choice for the “age when entered Hong Kong” subgroups. We fit such a model to cancer mortality in Hong Kong from 1998 to 2021. We aim to gain new insights on the impact of migrant status and forecast the trends of the cancer deaths under the background of demographics changes and inform public health policy making in Hong Kong.

We mainly focused on the contributions of sex and immigration status due to the non-identifiability problem that the effects of these three components are collinear with each other (denoted as period – age = cohort) [95][96]. Birth cohort effect and period effect were assessed with relative risks to evaluate the effect of three components. The median year of birth among cases was regarded as the reference cohort [97][98][99]. Since death cases aged 35–85 years between 1998 and 2021 were selected, the range of birth cohort from 1913 to 1986 covered observations and further projections until 2030. The second and penultimate period effects were constrained to the reference for period. For sex and immigration status, maximum likelihood framework was applied to estimate the relative risks and 95% confidence intervals (CIs) by age groups, calendar period, and birth cohort.

Although migrant status on risk of cancer is not a new topic, migrant status in these mortality data in Hong Kong have not been studied as a risk factors for cancer deaths, except for one of our earlier works, where we examined the impacts of birth place (either mainland China or Hong Kong) on the effects of age, period and cohort effects. Given that the birth place indeed played a role in the age, period and cohort effects of breast cancer deaths for women, it is reasonable to hypothesize that the “age when entered Hong Kong” would likely play a role in some of the cancer deaths. If the early life environment determines the development of cancer in the later life, then the younger the age entered Hong Kong, the weaker impact of early lifestyle would be for immigrants.

Mortality projection with Bayesian APC model, 2022-2030

Several projection approaches for future cancer mortality have been developed, but a Bayesian age-period-cohort (BAPC) model built upon integrated nested Laplace approximations (INLA) [100] yields relatively higher coverage and better performance for all evaluated parameter combinations [101]. To prevent some sampling problems caused by Markov chain Monte Carlo (MCMC), this MCMC-free BAPC approach was applied to predict future cancer mortality within a fully Bayesian inference setting and provide outputs of interest simply, such as projected age-standardized and age-specific rates. Convergence checks are not necessary for this technique [100]. The projections of age-standardized cancer mortality rates for each sex, age group (younger or older than 60 years) and migrant status, taking into account age, period, and birth cohort effects, were performed based on the weights of population age groups from the WHO World Standard population [102], with 95% prediction intervals.'' The Mann-Kendall trend test was applied to verify the projection trend. Friedman's Two-Way Analysis of Variance was applied to test interactions between gender and immigration groups for each year.

All analyses were performed via R version 4.2.1 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2013, http://www.R-project.org/). The APC models were established using the Epi package, and the projections based on Bayesian APC models were performed with the BAPC package.

3.4 Results

The first two figures of Figures 3-8 (i.e. a & b of Figures 3-7 and Figure 8a) illustrate the estimates of age (assessed by cancer mortality), cohort and period effects (assessed by relative risk) based on APC models among three migrant groups for men and women with six types of cancers, respectively. All the mortality rates for each gender and immigration status exhibit notable increasing trends with age. Age, cohort and period effects of six types of cancer for immigrants who stayed in Hong Kong for ≤10 years revealed relatively more pronounced fluctuations and deviations from those effects in the other two immigration groups. Significant increasing trends of age effect occurred in all types of cancer, regardless of gender and immigration status.

Figure 3c-7c & 8b, eFigure 2-6 in **Appendices A3** illustrate the age-standardized mortality rates of six types of cancer from 1998 to 2021 and their projections by sex, immigrant status and age groups from 2022 to 2030, taking into account age, period, and birth cohort effects. Means and standard deviations of predictive mortality rates are shown in Table 2-7. For all ages projection (Figure 3c-8c), as approximately significant interactions between gender and immigration groups emerge for each type of cancer in each year ($p<0.05$), given the projected trends, immigrants for each gender, especially who have stayed in Hong Kong for > 10 years will suffer from higher mortality rates of cancer in each year than locals.

Monotone decreasing trends or plateau of forecasting occur for both genders and all immigration groups in cancers, except for increasing trends for male immigrants who have stayed in Hong Kong for ≤10 years with colon cancer ($p < 0.05$, Avg +0.30 deaths/100,000 per annum) from 15.47 deaths/100,000 (95% CI: 11.28, 19.66) in 2021 to 18.50 deaths/100,000 (95% CI: 2.31, 34.69) in 2030, and male immigrants who have stayed in Hong Kong for > 10 years with pancreatic cancer ($p < 0.05$, Avg +0.72 deaths/100,000 per annum) from 16.30 deaths/100,000 (95% CI: 14.38,17.26) in 2021 to 23.49 deaths/100,000 (95% CI: 12.49, 34.49) in 2030. Results of six types of cancers are introduced as follows

3.4.1 Lung Cancer

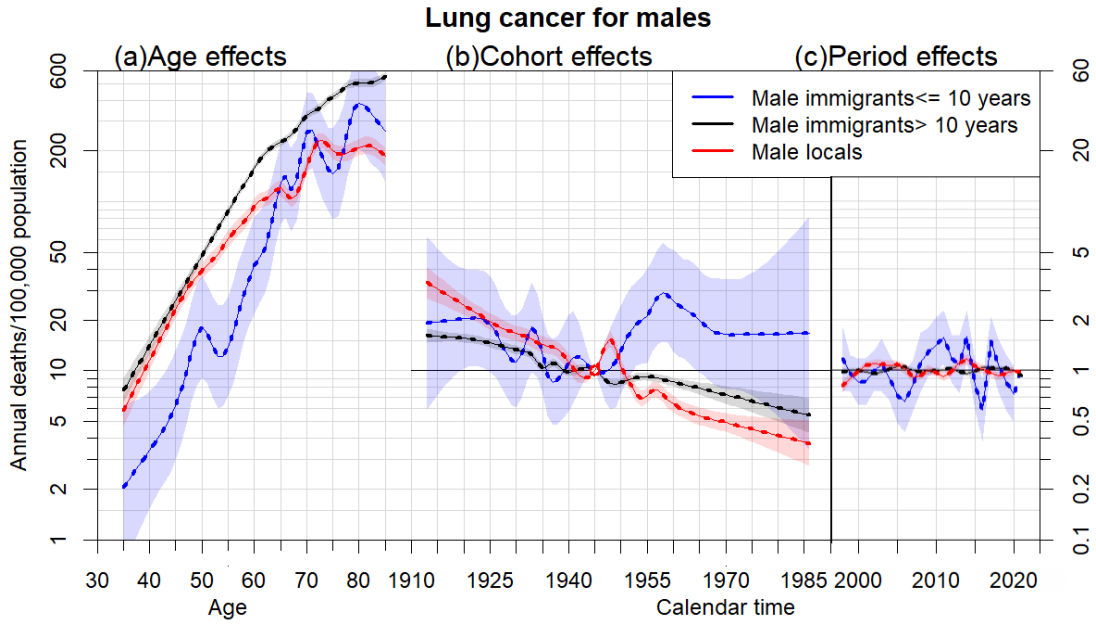


Figure 3a. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of male lung cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

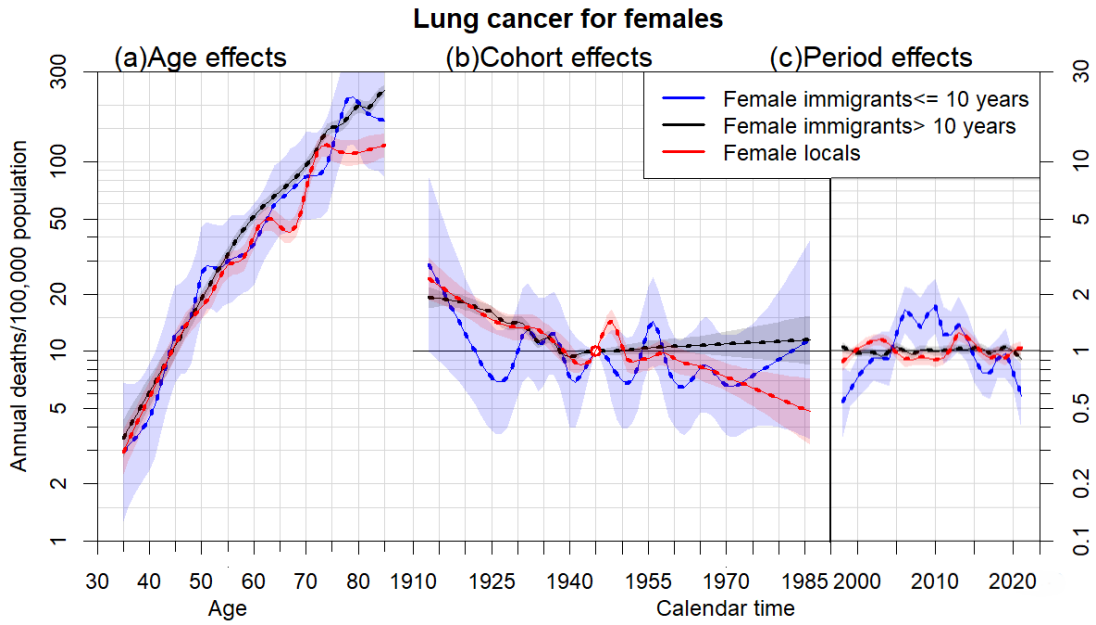


Figure 3b. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of female lung cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

While relatively insignificant differences in lung cancer mortality rates by immigration status among females have performed, male immigrants who remained in Hong Kong for >10 years had higher lung cancer mortality rates at ages above 50 years and those who arrived ≤10 years had lower lung cancer mortality at ages below 62 years compared to local men Figure 3. In addition to compatible dynamics of period effect for locals and long-stay immigrants, similar changes of relative risks by birth cohort for locals and long-stay immigrants in lung occurred before 1945, whereas significant differences of relative risks by birth cohort between these two immigration groups occurred after 1960. Short-stay immigrants who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks affected by period effects before 2020 than

those for locals and long-stay immigrants. Consequently, immigrants for both gender who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks of lung cancer mortality affected by cohort and period effects than locals and immigrants who have stayed in Hong Kong for >10 years.

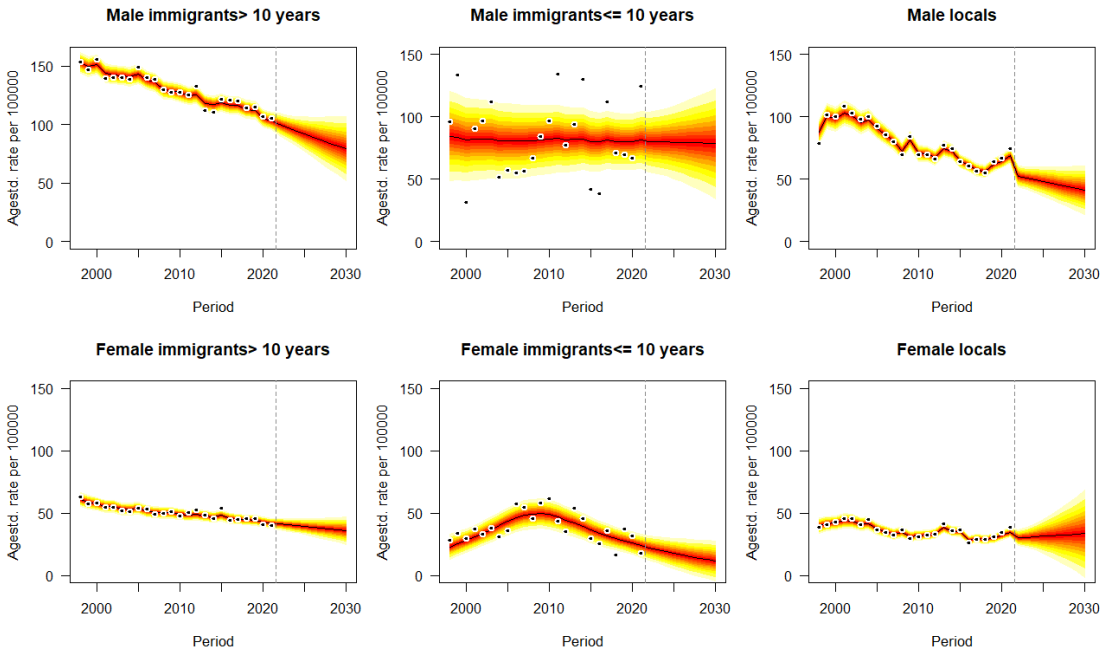


Figure 3c. Projections of lung cancer mortality rates by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

Given the projected trends in Figure 3c, immigrants for each gender, especially who have stayed in Hong Kong for > 10 years will suffer from higher mortality rates of lung cancer in each year than locals. Monotone decreasing trends or plateau of forecasting occur for both genders and all immigration groups, except for increasing trends for female local immigrants ($p < 0.05$, Avg +0.33 deaths/100,000 per annum) from 30.22

deaths/100,000 (95% CI: 30.01, 33.56) in 2022 to 33.55 deaths/100,000 (95% CI: 29.31, 36.11) in 2030. Compared with other immigration groups, male immigrants who have stayed in Hong Kong for >10 years with lung cancer would perform the most significant decline in predictive mean from 102.90 (95% CI: 98.14, 107.66) to 79.55 (95% CI: 47.46, 111.64) deaths per 100,000 population (Avg -2.34 deaths/100,000 per annum) (Table 2). Men would suffer from higher mortality rates of lung cancer in the future than females in the same immigration group. In 2030, the highest mortality rate of lung cancer would be 79.55 deaths/100,000 (95% CI: 70.11, 85.47) for male immigrants who have stayed in Hong Kong for > 10 years.

eFigure 2 in **Appendices** illustrate the age-standardized mortality rates of lung cancer from 1998 to 2021 and their projections by sex, immigrant status and two age groups from 2022 to 2030. Most of predictive trends for younger cases (<60 years) and older cases (≥60 years) reach a consensus with those for all ages population in Figure 3c, except for mortality rates of lung cancer for men immigrants ≤10 that insignificant trend for all ages ($p > 0.05$) vs. decline for younger cases ($p < 0.05$) vs. increase for older cases ($p < 0.05$). It's also reasonable that elders would be at higher risk of death by lung cancer than youngers regardless immigration groups and genders. Male individuals would also suffer from higher mortality rate of liver cancer than females for youngers and elders.

Predictive mean of age-standardized mortality rates of lung cancer per 100,000 population										
Year	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Female immigrants >10	41.80 (1.27)	41.34 (1.86)	40.58 (2.27)	39.87 (2.75)	39.19 (3.28)	38.53 (3.86)	37.89 (4.46)	37.26 (5.09)	36.65 (5.74)	36.04 (6.4)
Female immigrants ≤ 10	23.92 (4.00)	22.22 (4.67)	20.56 (5.38)	19.01 (6.10)	17.57 (6.80)	16.24 (7.45)	15.00 (8.04)	13.85 (8.56)	12.79 (9.01)	11.81 (9.39)
Female locals	34.67 (1.76)	30.22 (3.54)	30.63 (4.77)	31.05 (6.38)	31.48 (8.29)	31.9 (10.47)	32.32 (12.87)	32.73 (15.48)	33.15 (18.31)	33.55 (21.33)
Male immigrants >10	102.90 (2.43)	100.18 (4.18)	97.18 (5.33)	94.34 (6.72)	91.71 (8.24)	89.15 (9.84)	86.66 (11.47)	84.19 (13.11)	81.81 (14.74)	79.55 (16.37)
Male immigrants ≤10	81.26 (9.21)	79.90 (10.41)	79.81 (11.82)	79.72 (13.42)	79.62 (15.19)	79.50 (17.09)	79.32 (19.09)	79.08 (21.18)	78.78 (23.32)	78.41 (25.53)
Male locals	60.96 (2.82)	52.27 (4.86)	50.83 (5.39)	49.56 (6.13)	48.18 (6.97)	46.64 (7.84)	45.13 (8.76)	43.83 (9.76)	42.67 (10.8)	41.43 (11.8)
Female immigrants>10(<60y)	15.51 (1.12)	14.51 (1.50)	13.90 (1.76)	13.29 (2.04)	12.71 (2.33)	12.13 (2.62)	11.57 (2.91)	11.02 (3.18)	10.49 (3.43)	9.98 (3.68)
Female immigrants ≤ 10(<60y)	8.14 (1.91)	7.79 (1.95)	7.18(2.23)	6.62(2.53)	6.10(2.81)	5.63(3.08)	5.19(3.32)	4.79 (3.53)	4.42 (3.72)	4.09 (3.88)
Female locals(<60y)	10.25 (0.77)	9.48 (0.89)	9.17(1.02)	8.87(1.16)	8.57(1.32)	8.27(1.49)	7.97(1.65)	7.68 (1.82)	7.38 (1.98)	7.09 (2.13)
Male immigrants>10(<60y)	27.81 (2.10)	26.36 (3.58)	24.96 (3.94)	23.64 (4.35)	22.38 (4.79)	21.17 (5.23)	20.03 (5.67)	18.96 (6.10)	17.96 (6.51)	17.03 (6.90)
Male immigrants ≤ 10(<60y)	15.01 (2.98)	13.38 (3.71)	12.02 (4.17)	10.79 (4.59)	9.68 (4.95)	8.69 (5.24)	7.79 (5.46)	6.98 (5.61)	6.25 (5.69)	5.59 (5.72)

Male locals(<60y)	15.19 (0.78)	14.45 (1.15)	14.03 (1.29)	13.61 (1.46)	13.14 (1.64)	12.65 (1.82)	12.13 (2.01)	11.55 (2.17)	10.93 (2.31)	10.26 (2.43)
Female immigrants >10(≥ 60y)	108.85 (4.80)	107.21 (5.17)	106.26 (6.24)	105.52 (7.54)	104.94 (9.04)	104.51 (10.72)	104.21 (12.57)	104.07 (14.61)	104.06 (16.78)	104.16 (19.14)
Female immigrants ≤ 10(≥60y)	66.16 (13.25)	63.84 (15.72)	59.88 (17.50)	56.14 (19.31)	52.60 (21.03)	49.27 (22.66)	46.14 (24.16)	43.20 (25.52)	40.44 (26.74)	37.85 (27.81)
Female locals(≥60y)	77.33 (9.40)	76.53 (10.11)	76.22 (10.85)	75.94 (11.79)	75.69 (12.94)	75.49 (14.28)	75.32 (15.80)	75.19 (17.48)	75.10 (19.33)	75.03 (21.32)
Male immigrants>10(≥60y)	293.56 (9.13)	289.8 (11.7)	286.6 (15.19)	284.28 (19.51)	282.78 (24.49)	281.99 (30.07)	281.88 (36.31)	282.31 (43.15)	283.37 (50.66)	285.03 (58.86)
Male immigrants ≤ 10(≥60y)	244.88 (30.29)	247.01 (36.85)	251.24 (42.94)	255.62 (50.06)	260.14 (58.14)	264.82 (67.14)	269.61 (77.01)	274.52 (87.75)	279.55 (99.34)	284.69 (111.81)
Male locals(≥60y)	150.75 (16.22)	146.29 (18.46)	143.54 (20.58)	141.84 (23.97)	140.07 (28.24)	138.14 (33.39)	136.65 (39.82)	136.49 (47.87)	137.24 (57.47)	138.26 (68.52)

Table 2. Predictive means and standard deviations (in brackets) of age-standardized mortality rates of lung cancer per 100,000 population for each gender, age group (less than, greater or equal to 60 years old) and immigrant status from 2022 to 2030. Reported means and standard deviations (in brackets) of age-standardized mortality rates in 2021 are also listed.

3.4.2 Colon Cancer

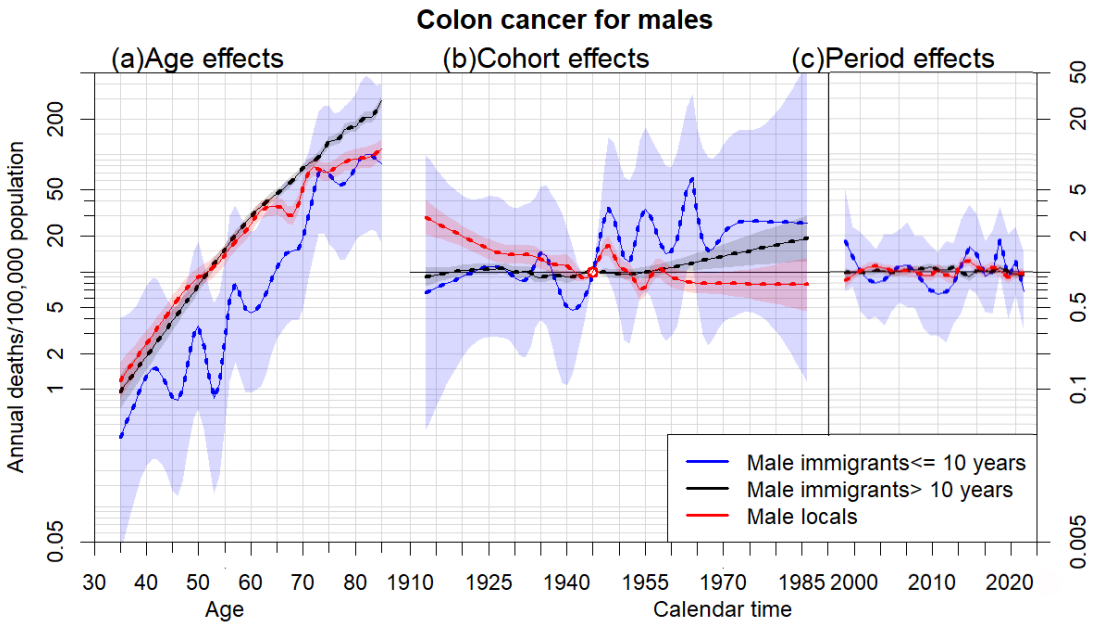


Figure 4a. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of male colon cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

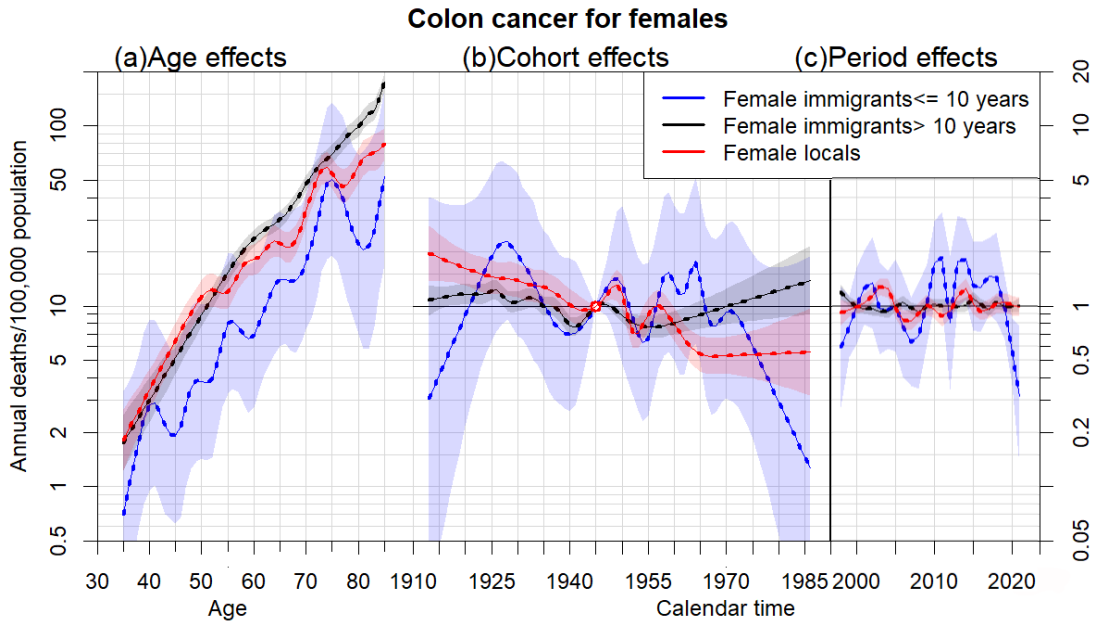


Figure 4b. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of female colon cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

Although relatively insignificant differences in colon cancer mortality rates for both genders between locals and long-stay immigrants have performed, immigrants who remained in Hong Kong for <10 years had lower colon cancer mortality rates at each age compared to locals and immigrants who remained in Hong Kong for >10 years in Figure 4. In addition to compatible dynamics of period effect for locals and long-stay immigrants, similar changes of relative risks by birth cohort for locals and long-stay immigrants in colon occurred before 1955, whereas significant differences of relative risks by birth cohort between these two immigration groups occurred after 1960. Short-stay immigrants who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks affected by period effects before 2020 than those for locals and long-stay

immigrants. Consequently, immigrants for both gender who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks of colon cancer mortality affected by cohort and period effects than locals and immigrants who have stayed in Hong Kong for >10 years.

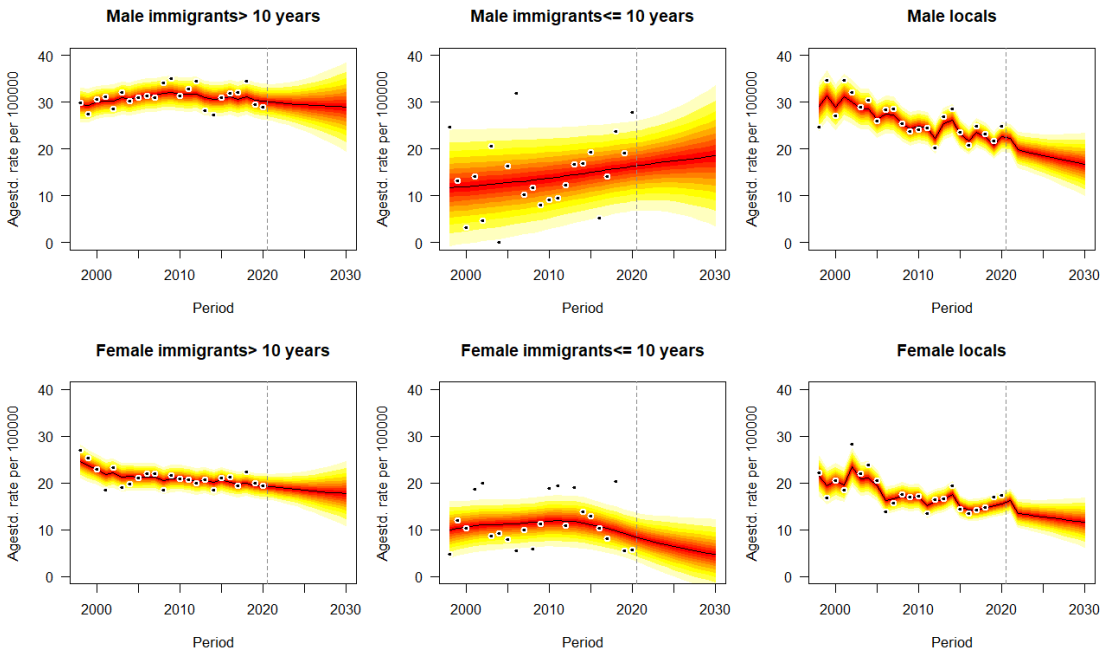


Figure 4c. Projections of colon cancer mortality rates by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

Given the projected trends in Figure 4c, immigrants for each gender, especially who have stayed in Hong Kong for > 10 years will suffer from higher mortality rates of colon cancer in each year than locals. Men would suffer from higher mortality rates of colon cancer in the future than females in the same immigration group. Monotone decreasing trends or plateau of forecasting occur for both genders and all immigration groups, except for increasing trends for male immigrants who have stayed in Hong

Kong for < 10 years ($p < 0.05$, Avg +0.30 deaths/100,000 per annum) from 15.47 deaths/100,000 (95% CI: 11.09, 18.26) in 2021 to 18.50 deaths/100,000 (95% CI: 15.44, 21.11) in 2030. Compared with other immigration groups, male locals with colon cancer would perform the most significant decline in predictive mean from 21.28 (95% CI: 18.14, 23.17) to 16.71 (95% CI: 10.46, 19.25) deaths per 100,000 population ($p < 0.05$, Avg -0.45 deaths/100,000 per annum) (Table 3). In 2030, the highest mortality rate of colon cancer would be 28.98 deaths/100,000 (95% CI: 26.53, 31.47) for male immigrants who have stayed in Hong Kong for > 10 years, while the lowest mortality rate of liver cancer would be 4.71 deaths/100,000 (95% CI: 2.77, 7.12) for female immigrants who have stayed in Hong Kong for < 10 years.

eFigure 3 in **Appendices** illustrate the age-standardized mortality rates of colon cancer from 1998 to 2021 and their projections by sex, immigrant status and two age groups from 2022 to 2030. Most of predictive trends for younger cases (<60 years) and older cases (≥60 years) reach a consensus with those for all ages population in Figure 4c, except for mortality rates of colon cancer for women immigrants >10 that insignificant trend for all ages ($p > 0.05$) vs. increase for younger cases ($p < 0.05$) vs. insignificant trend for older cases ($p > 0.05$). It's also reasonable that elders would be at higher risk of death by colon cancer than youngers regardless immigration groups and genders. Male individuals would also suffer from higher mortality rate of liver cancer than females for youngers and elders.

Predictive mean of age-standardized mortality rates of colon cancer per 100,000 population										
Year	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Female immigrants >10	20.03 (0.95)	18.95 (1.13)	18.77 (1.37)	18.59 (1.66)	18.42 (1.98)	18.27 (2.33)	18.12 (2.71)	17.98 (3.11)	17.85 (3.53)	17.73 (3.96)
Female immigrants ≤ 10	8.11 (2.19)	7.70 (2.51)	7.25 (2.81)	6.82 (3.11)	6.42 (3.37)	6.03 (3.61)	5.67 (3.83)	5.33 (4.01)	5.01 (4.17)	4.71 (4.31)
Female locals	13.77 (1.30)	13.47 (1.61)	13.24 (1.72)	13.01 (1.87)	12.77 (2.04)	12.53 (2.24)	12.29 (2.46)	12.06 (2.68)	11.82 (2.92)	11.59 (3.16)
Male immigrants >10	31.22 (1.28)	29.82 (1.46)	29.66 (1.79)	29.52 (2.19)	29.41 (2.63)	29.30 (3.11)	29.21 (3.64)	29.14 (4.19)	29.06 (4.78)	28.98 (5.39)
Male immigrants ≤10	15.47 (2.14)	16.77 (3.77)	17.02 (4.18)	17.23 (4.64)	17.45 (5.14)	17.67 (5.69)	17.88 (6.27)	18.09 (6.91)	18.31 (7.56)	18.50 (8.26)
Male locals	21.28 (1.38)	19.81 (2.07)	19.39 (2.22)	18.97 (2.42)	18.57 (2.61)	18.18 (2.85)	17.81 (3.12)	17.43 (3.40)	17.06 (3.71)	16.71 (4.03)
Female immigrants >10(<60y)	7.09 (0.99)	7.36 (1.12)	7.46 (1.28)	7.56 (1.46)	7.65 (1.68)	7.74 (1.92)	7.83 (2.19)	7.92 (2.48)	8.01 (2.79)	8.09 (3.13)
Female immigrants ≤ 10(<60y)	3.11 (0.67)	2.82 (0.86)	2.65 (0.91)	2.51 (0.97)	2.36 (1.02)	2.22 (1.07)	2.08 (1.11)	1.95 (1.14)	1.83 (1.18)	1.72 (1.22)
Female locals(<60y)	4.10 (0.41)	3.87 (0.50)	3.73 (0.54)	3.61 (0.59)	3.47 (0.65)	3.34 (0.70)	3.22 (0.76)	3.11 (0.82)	2.99 (0.88)	2.88 (0.94)
Male immigrants >10(<60y)	8.29 (0.91)	7.98 (1.17)	7.85 (1.38)	7.71 (1.60)	7.54 (1.83)	7.36 (2.08)	7.17(2.32)	6.97(2.57)	6.76(2.81)	6.55(3.05)
Male immigrants ≤ 10(<60y)	5.03 (1.44)	5.18 (1.58)	5.22 (1.75)	5.26 (1.93)	5.30 (2.14)	5.34 (2.36)	5.38(2.59)	5.43(2.84)	5.47(3.11)	5.51(3.38)

Male locals(<60y)	5.14 (0.43)	4.88 (0.63)	4.66 (0.79)	4.46 (0.96)	4.26 (1.13)	4.08 (1.31)	3.91(1.48)	3.73(1.65)	3.57(1.82)	3.42(1.97)
Female immigrants >10(≥60y)	52.16 (2.59)	49.21 (2.99)	48.70 (3.56)	48.26 (4.26)	47.87 (5.05)	47.54 (5.94)	47.26 (6.90)	47.05 (7.94)	46.91 (9.06)	46.81 (10.26)
Female immigrants ≤ 10(≥60y)	24.01 (5.83)	22.44 (6.56)	21.69 (6.96)	20.95 (7.38)	20.23 (7.80)	19.52 (8.23)	18.84 (8.66)	18.17 (9.08)	17.51 (9.49)	16.86 (9.90)
Female locals(≥60y)	37.42 (5.31)	36.69 (5.74)	36.29 (6.06)	35.87 (6.46)	35.46 (6.95)	35.04 (7.5)	34.61 (8.12)	34.19 (8.79)	33.77 (9.51)	33.34 (10.27)
Male immigrants >10(≥60y)	84.17 (3.55)	82.72 (4.09)	82.16 (4.95)	81.64 (5.97)	81.19 (7.12)	80.81 (8.39)	80.47 (9.77)	80.15 (11.24)	79.85 (12.81)	79.56 (14.45)
Male immigrants ≤ 10(≥60y)	43.25 (11.07)	44.93 (13.09)	45.62 (14.52)	46.30 (16.09)	46.96 (17.80)	47.61 (19.64)	48.25 (21.62)	48.88 (23.73)	49.51 (25.97)	50.13 (28.34)
Male locals(≥60y)	55.79 (6.86)	54.89 (7.65)	53.75 (8.03)	52.63 (8.52)	51.54 (9.12)	50.47 (9.8)	49.43 (10.55)	48.42 (11.37)	47.42 (12.25)	46.44 (13.16)

Table 3. Predictive means and standard deviations (in brackets) of age-standardized mortality rates of colon cancer per 100,000 population for each gender, age group (less than, greater or equal to 60 years old) and immigrant status from 2022 to 2030. Reported means and standard deviations (in brackets) of age-standardized mortality rates in 2021 are also listed.

3.4.3 Liver Cancer

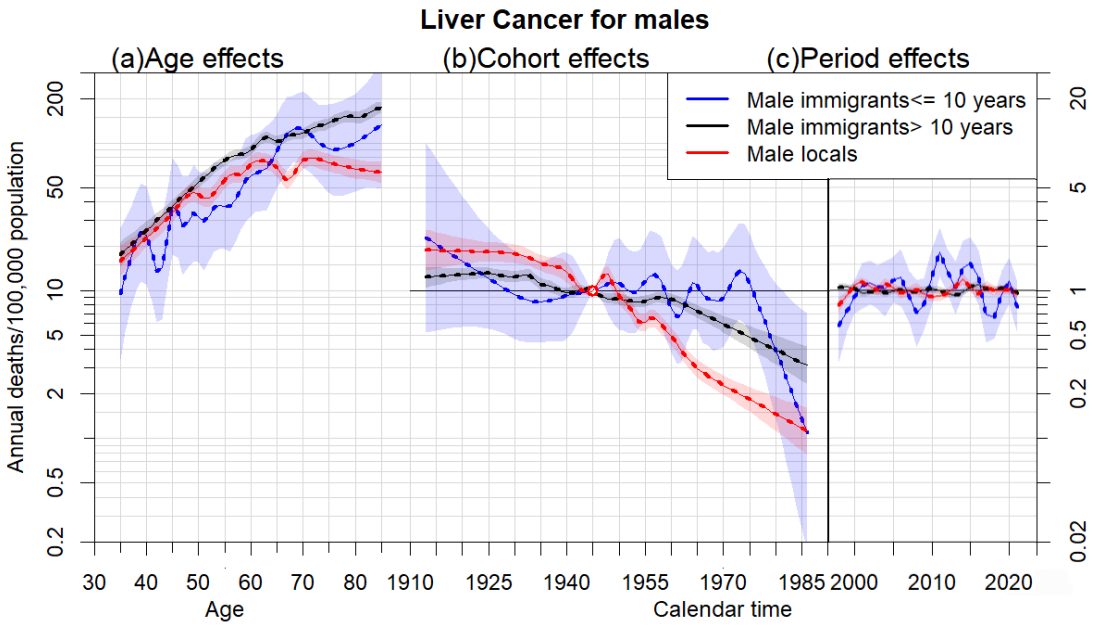


Figure 5a. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of male liver cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

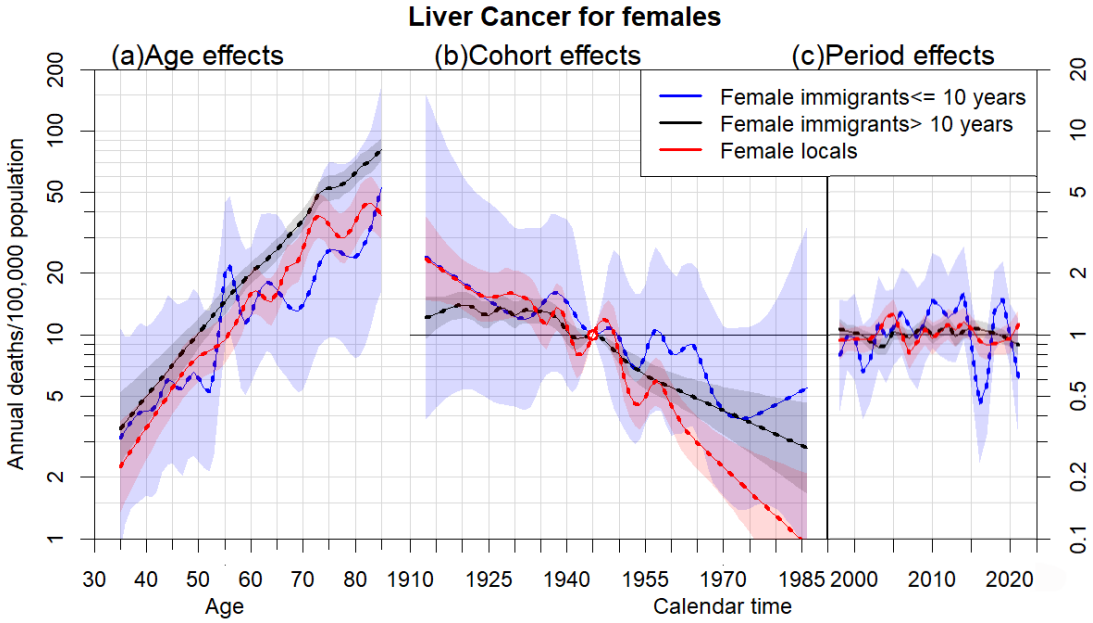


Figure 5b. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of female liver cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

Relatively insignificant differences in liver cancer mortality rates for both genders between locals and long-stay immigrants have performed, even though immigrants aged younger than 55 and older than 65 who remained in Hong Kong for <10 years had lower colon cancer mortality rates at each age compared to locals and immigrants who remained in Hong Kong for >10 years in Figure 5. In addition to compatible dynamics of period effect for locals and long-stay immigrants, similar changes of relative risks by birth cohort for locals and long-stay immigrants in liver occurred before 1955, whereas significant differences of relative risks by birth cohort between these two immigration groups occurred after 1960. Short-stay immigrants who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks affected by period effects before 2020 than those for locals and long-stay immigrants. Consequently, immigrants for both gender who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks of liver cancer mortality affected by cohort and period effects than locals and immigrants who have stayed in Hong Kong for >10 years.

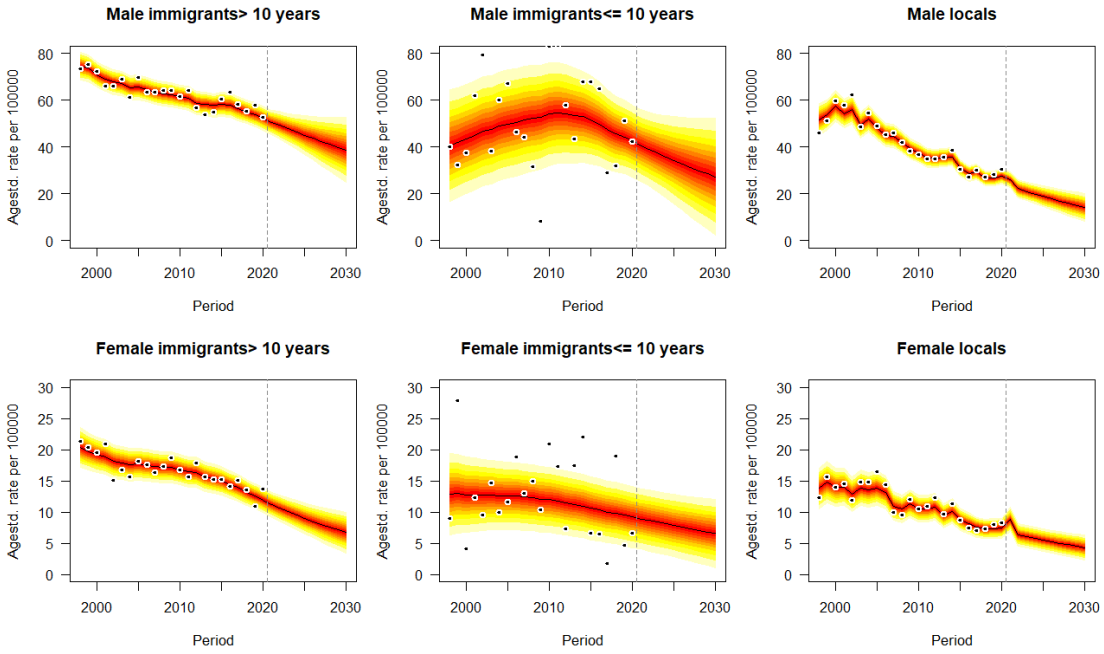


Figure 5c. Projections of liver cancer mortality rates by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

Given the projected trends in Figure 5c, immigrants for each gender, especially who have stayed in Hong Kong for > 10 years will suffer from higher mortality rates of liver cancer in each year than locals. Men would suffer from higher mortality rates of liver cancer in the future than females in the same immigration group. Monotone decreasing trends ($p < 0.05$) or plateau ($p > 0.05$) of forecasting occur for both genders and all immigration groups. Compared with other immigration groups, male immigrants who have stayed in Hong Kong for < 10 years with liver cancer would perform the most significant decline in predictive mean from 42.33 (95% CI: 38.94, 44.25) to 24.25 (95% CI: 19.46, 28.77) deaths per 100,000 population ($p < 0.05$, Avg -1.51 deaths/100,000 per annum) (Table 4). In 2030, the highest mortality rate of liver cancer would be 38.71 deaths/100,000 (95% CI: 36.53, 41.92) for male immigrants who have stayed in Hong Kong for > 10 years, while the lowest mortality rate of liver cancer would be 4.30

deaths/100,000 (95% CI: 1.77, 7.42) for female local.

eFigure 4 in **Appendices** illustrate the age-standardized mortality rates of liver cancer from 1998 to 2021 and their projections by sex, immigrant status and two age groups from 2022 to 2030. Most of predictive trends for younger cases (<60 years) and older cases (≥60 years) reach a consensus with those for all ages population in Figure 5c, except for mortality rates of liver cancer for men immigrants >10 that decline for all ages ($p < 0.05$) vs. decline for younger cases ($p < 0.05$) vs. insignificant trend for older cases ($p > 0.05$). It's also reasonable that elders would be at higher risk of death by liver cancer than youngers regardless immigration groups and genders. Male individuals would also suffer from higher mortality rate of liver cancer than females for youngers and elders.

Predictive mean of age-standardized mortality rates of liver cancer per 100,000 population										
Year	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Female immigrants >10	11.34 (0.66)	10.68 (0.71)	10.09 (0.85)	9.54 (1.01)	9.01 (1.16)	8.50 (1.31)	8.02(1.45)	7.57(1.59)	7.14(1.72)	6.74(1.83)
Female immigrants ≤10	9.15 (1.55)	8.66 (1.82)	8.38 (1.95)	8.11 (2.08)	7.84 (2.22)	7.58 (2.36)	7.32(2.49)	7.07(2.63)	6.82(2.76)	6.58(2.88)
Female locals	6.72 (0.69)	6.36 (0.88)	6.08 (0.90)	5.81 (0.93)	5.53 (0.97)	5.26 (1.01)	5.01(1.06)	4.77(1.11)	4.53(1.15)	4.3(1.21)
Male immigrants >10	52.17 (1.78)	49.22 (2.36)	47.76 (2.93)	46.35 (3.59)	45.01 (4.31)	43.67 (5.05)	42.37 (5.81)	41.1(6.56)	39.89 (7.33)	38.71 (8.08)
Male immigrants ≤10	42.33 (5.87)	39.03 (6.49)	37.39 (7.47)	35.81 (8.51)	34.26 (9.58)	32.76 (10.63)	31.31 (11.65)	29.91 (12.62)	28.56 (13.54)	27.25 (14.40)
Male locals	24.22 (1.77)	22.16 (2.09)	21.02 (2.22)	19.91 (2.39)	18.85 (2.58)	17.83 (2.79)	16.85 (3.03)	15.92 (3.21)	15.03 (3.40)	14.18 (3.59)
Female immigrants >10(<60y)	3.62 (0.45)	3.39 (0.52)	3.29 (0.57)	3.20 (0.63)	3.12 (0.69)	3.04 (0.75)	2.96(0.82)	2.89(0.89)	2.82(0.96)	2.75(1.03)
Female immigrants ≤ 10(<60y)	4.10 (0.79)	3.81 (0.91)	3.69 (0.96)	3.57 (1.02)	3.46 (1.08)	3.36 (1.15)	3.25(1.22)	3.15(1.29)	3.06(1.36)	2.97(1.43)
Female locals(<60y)	1.50 (0.13)	1.37 (0.2)	1.29 (0.21)	1.22 (0.23)	1.16 (0.24)	1.10 (0.26)	1.04(0.27)	0.99(0.29)	0.94(0.30)	0.89(0.31)
Male immigrants >10(<60y)	26.32 (2.11)	24.04 (2.35)	23.02 (2.63)	22.05 (2.94)	21.13 (3.27)	20.25 (3.61)	19.41 (3.95)	18.62 (4.30)	17.86 (4.64)	17.14 (4.98)
Male immigrants ≤ 10(<60y)	25.52 (2.99)	22.56 (3.96)	21.71 (4.44)	20.87 (4.94)	20.04 (5.45)	19.22 (5.95)	18.42 (6.45)	17.63 (6.91)	16.86 (7.36)	16.11 (7.78)

Male locals(<60y)	8.25 (0.69)	7.47 (0.74)	6.97 (0.79)	6.52 (0.86)	6.11 (0.93)	5.73 (1.01)	5.38(1.08)	5.04(1.15)	4.73(1.21)	4.44(1.27)
Female immigrants >10(≥ 60y)	33.67 (1.88)	29.63 (2.01)	27.99 (2.36)	26.42 (2.75)	24.92 (3.14)	23.49 (3.52)	22.13 (3.88)	20.85 (4.23)	19.64 (4.55)	18.50 (4.85)
Female immigrants ≤10(≥ 60y)	21.72 (5.11)	19.08 (5.81)	18.38 (6.14)	17.71 (6.48)	17.03 (6.83)	16.39 (7.16)	15.76 (7.49)	15.16 (7.80)	14.57 (8.11)	14.01 (8.39)
Female locals(≥60y)	20.63 (3.03)	18.41 (3.23)	17.55 (3.26)	16.72 (3.32)	15.91 (3.40)	15.11 (3.49)	14.34 (3.59)	13.59 (3.69)	12.87 (3.81)	12.17 (3.93)
Male immigrants >10(≥ 60y)	115.39 (4.54)	113.96 (5.95)	113.43 (7.65)	113.17 (9.70)	113.16 (12.04)	113.37 (14.66)	113.79 (17.56)	114.39 (20.73)	115.19 (24.18)	116.17 (27.91)
Male immigrants ≤10(≥ 60y)	88.61 (15.58)	85.14 (18.85)	82.59 (20.6)	80.02 (22.44)	77.42 (24.34)	74.83 (26.24)	72.23 (28.12)	69.64 (29.94)	67.07 (31.70)	64.52 (33.38)
Male locals(≥60y)	62.88 (5.97)	58.95 (7.91)	56.51 (8.20)	54.14 (8.61)	51.84 (9.12)	49.61 (9.70)	47.46 (10.33)	45.38 (11.01)	43.38 (11.68)	41.45 (12.36)

Table 4. Predictive means and standard deviations (in brackets) of age-standardized mortality rates of liver cancer per 100,000 population for each gender, age group (less than, greater or equal to 60 years old) and immigrant status from 2022 to 2030. Reported means and standard deviations (in brackets) of age-standardized mortality rates in 2021 are also listed.

3.4.4 Pancreatic Cancer

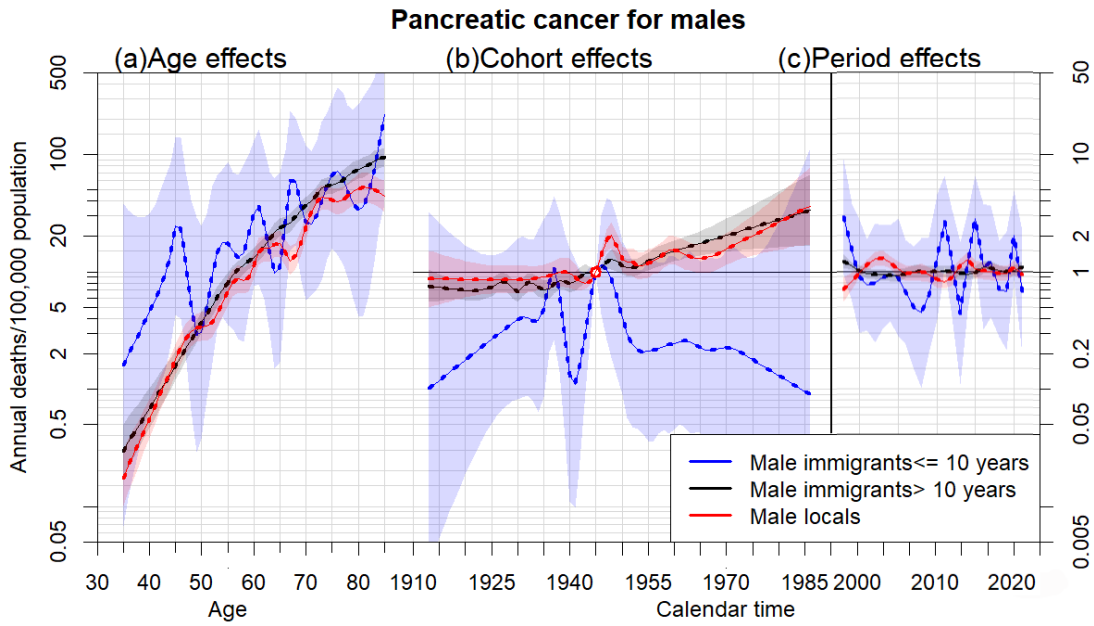


Figure 6a. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of male pancreatic cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

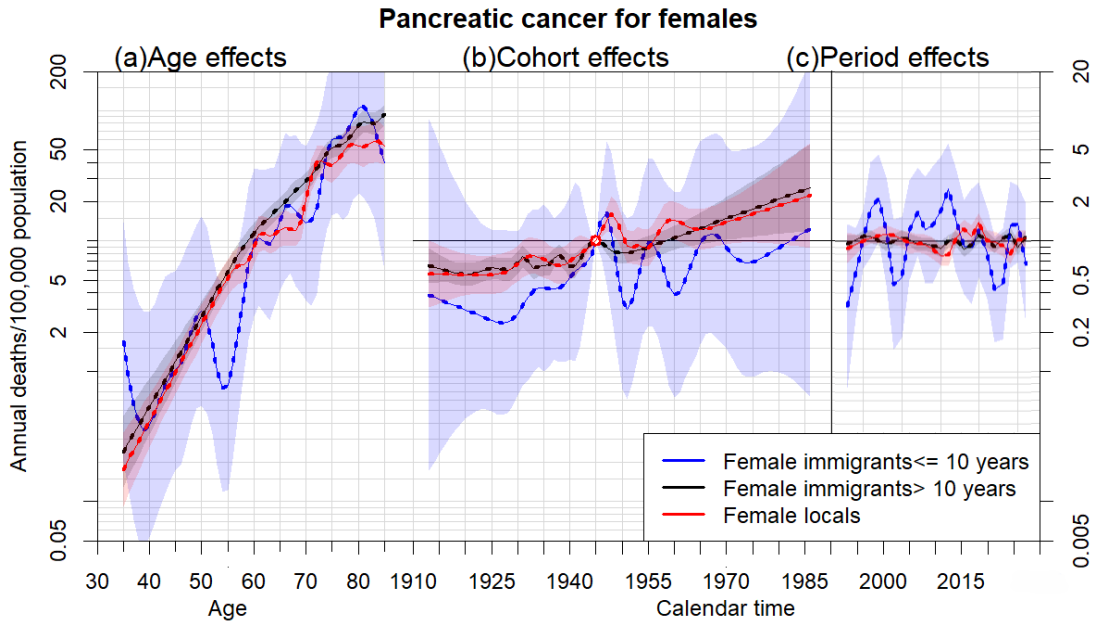


Figure 6b. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of female pancreatic cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

Relatively insignificant differences in pancreatic cancer mortality rates for both genders between locals and long-stay immigrants have performed, even though female immigrants aged younger than 50 and older than 75 who remained in Hong Kong for <10 years had lower pancreatic cancer mortality rates at each age compared to locals and immigrants who remained in Hong Kong for >10 years, while male immigrants who have stayed in Hong Kong for <10 years had significantly higher pancreatic cancer mortality rates than other two immigration status for each age group in Figure 6. In addition to compatible dynamics of period effect for locals and long-stay immigrants, similar changes of relative risks by birth cohort for locals and long-stay immigrants in pancreatic occurred before 1955, whereas significant differences of relative risks by

birth cohort between these two immigration groups occurred after 1960. Short-stay immigrants who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks affected by period effects before 2020 than those for locals and long-stay immigrants. Consequently, immigrants for both gender who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks of pancreatic cancer mortality affected by cohort and period effects than locals and immigrants who have stayed in Hong Kong for >10 years, and relative risks of pancreatic cancer mortality affected by cohort and period for locals and long-stay immigrants perform significant consistency.

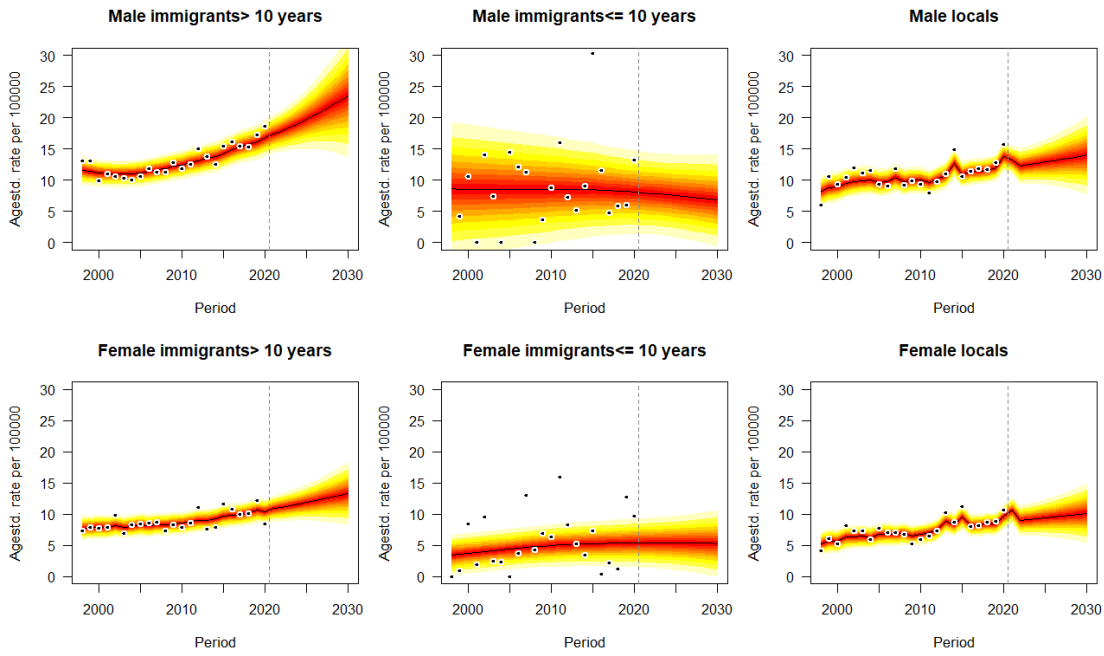


Figure 6c. Projections of pancreatic cancer mortality rates by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

Given the projected trends in Figure 6c, immigrants for each gender, especially who have stayed in Hong Kong for > 10 years will suffer from higher mortality rates of pancreatic cancer in each year than locals. Men would suffer from higher mortality rates of pancreatic cancer in the future than females in the same immigration group. Unlike the trends of age standardized mortality rates depict above, monotone increasing trends ($p < 0.05$) or plateau ($p > 0.05$) of forecasting occur for both genders and all immigration groups, except for decreasing trends for male immigrants who have stayed in Hong Kong for < 10 years ($p < 0.05$, Avg -0.12 deaths/100,000 per annum) from 8.10 deaths/100,000 (95% CI: 6.04, 11.13) in 2021 to 6.81 deaths/100,000 (95% CI: 4.47, 9.74) in 2030. Compared with other immigration groups, male immigrants who have stayed in Hong Kong for < 10 years with pancreatic cancer would perform the most significant uptrend in predictive mean from 16.30 (95% CI: 15.88, 17.74) to 23.49 (95% CI: 19.96, 26.59) deaths per 100,000 population ($p < 0.05$, Avg +0.72 deaths/100,000 per annum) (Table 5). In 2030, the highest mortality rate of pancreatic cancer would be 23.49 deaths/100,000 (95% CI: 19.96, 26.59) for male immigrants who have stayed in Hong Kong for > 10 years, while the lowest mortality rate of pancreatic cancer would be 5.31 deaths/100,000 (95% CI: 3.22, 7.92) for female who have stayed in Hong Kong for < 10 years.

eFigure 5 in **Appendices** illustrate the age-standardized mortality rates of pancreatic cancer from 1998 to 2021 and their projections by sex, immigrant status and two age groups from 2022 to 2030. Most of predictive trends for younger cases (<60 years) and older cases (≥60 years) reach a consensus with those for all ages population in Figure 6c, except for mortality rates of pancreatic cancer for men immigrants <10 that decline for all ages ($p < 0.05$) vs. decline for younger cases ($p < 0.05$) vs. insignificant trend for older cases ($p > 0.05$). It's also reasonable that elders would be at higher risk of death by pancreatic cancer than youngers regardless immigration groups and genders. Male individuals would also suffer from higher mortality rate of liver cancer than females for youngers and elders.

Predictive mean of age-standardized mortality rates of pancreatic cancer per 100,000 population										
Year	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Female immigrants >10	10.89 (0.62)	11.11 (0.75)	11.36 (0.91)	11.61 (1.09)	11.87 (1.31)	12.14 (1.56)	12.42 (1.84)	12.71 (2.15)	13.01 (2.48)	13.3(2.85)
Female immigrants ≤ 10	5.51 (1.44)	5.44 (1.56)	5.44 (1.69)	5.43(1.84)	5.42(1.99)	5.41(2.15)	5.39(2.32)	5.36(2.49)	5.34(2.66)	5.31(2.84)
Female locals	8.79 (1.10)	9.01 (1.22)	9.15 (1.34)	9.29(1.48)	9.43(1.64)	9.57(1.83)	9.71(2.05)	9.85(2.28)	9.99(2.54)	10.14 (2.83)
Male immigrants >10	16.30 (0.98)	17.87 (1.19)	18.48 (1.49)	19.11 (1.87)	19.78 (2.32)	20.47 (2.83)	21.18 (3.42)	21.92 (4.07)	22.69 (4.81)	23.49 (5.61)
Male immigrants ≤10	8.10 (2.02)	7.87 (2.37)	7.76 (2.53)	7.64(2.70)	7.51(2.87)	7.38(3.05)	7.24(3.23)	7.09(3.41)	6.95(3.58)	6.81(3.75)
Male locals	11.97 (1.26)	12.29 (1.49)	12.49 (1.64)	12.69 (1.83)	12.91 (2.06)	13.11 (2.33)	13.33 (2.63)	13.55 (2.97)	13.78 (3.34)	14.02 (3.74)
Female immigrants >10(<60y)	3.47 (0.33)	3.62 (0.57)	3.74 (0.66)	3.87(0.77)	4.01(0.89)	4.14(1.02)	4.28(1.18)	4.42(1.34)	4.57(1.53)	4.72(1.73)
Female immigrants ≤ 10(<60y)	1.12 (0.33)	1.21 (0.48)	1.22 (0.52)	1.23(0.56)	1.24(0.61)	1.25(0.66)	1.26(0.71)	1.26(0.77)	1.27(0.83)	1.28(0.89)
Female locals(<60y)	2.76 (0.27)	2.88 (0.36)	2.91 (0.41)	2.93(0.48)	2.96(0.55)	2.99(0.63)	3.02(0.71)	3.04(0.81)	3.07(0.90)	3.10(1.01)
Male immigrants >10(<60y)	6.88 (0.98)	7.05 (1.11)	7.24 (1.32)	7.43(1.56)	7.62(1.84)	7.82(2.16)	8.01(2.50)	8.21(2.88)	8.40(3.30)	8.61(3.75)
Male immigrants ≤ 10(<60y)	2.20 (0.71)	2.01 (0.85)	1.95 (0.91)	1.9(0.94)	1.84(0.99)	1.79(1.04)	1.74(1.09)	1.69(1.14)	1.64(1.19)	1.60(1.24)

Male locals(<60y)	4.16 (0.35)	4.33 (0.48)	4.41 (0.57)	4.46(0.68)	4.53(0.81)	4.61(0.94)	4.69(1.09)	4.77(1.26)	4.85(1.44)	4.93(1.63)
Female immigrants >10(≥60y)	28.58 (1.83)	29.45 (2.11)	29.91 (2.54)	30.38 (3.06)	30.85 (3.66)	31.33 (4.33)	31.81 (5.08)	32.29 (5.91)	32.78 (6.79)	33.27 (7.74)
Female immigrants ≤ 10(≥60y)	16.79 (5.29)	15.65 (6.08)	15.49 (6.71)	15.33 (7.36)	15.16 (8.03)	14.97 (8.73)	14.79 (9.43)	14.59 (10.14)	14.39 (10.86)	14.19 (11.58)
Female locals(≥60y)	22.80 (4.23)	23.85 (4.46)	24.21 (4.81)	24.56 (5.23)	24.91 (5.73)	25.25 (6.30)	25.58 (6.95)	25.90 (7.67)	26.22 (8.47)	26.54 (9.34)
Male immigrants >10(≥60y)	42.70 (2.55)	44.36 (3.02)	45.85 (3.76)	47.41 (4.69)	49.04 (5.78)	50.73 (7.05)	52.48 (8.50)	54.28 (10.13)	56.16 (11.95)	58.11 (13.98)
Male immigrants ≤ 10(≥60y)	24.68 (8.21)	23.96 (9.01)	23.87 (9.74)	23.75 (10.52)	23.61 (11.33)	23.45 (12.17)	23.28 (13.04)	23.09 (13.93)	22.89 (14.83)	22.68 (15.75)
Male locals(≥60y)	30.10 (4.68)	31.17 (5.22)	31.55 (5.63)	31.93 (6.14)	32.30 (6.75)	32.66 (7.45)	33.01 (8.23)	33.35 (9.11)	33.69 (10.08)	34.03 (11.12)

Table 5. Predictive means and standard deviations (in brackets) of age-standardized mortality rates of pancreatic cancer per 100,000 population for each gender, age group (less than, greater or equal to 60 years old) and immigrant status from 2022 to 2030. Reported means and standard deviations (in brackets) of age-standardized mortality rates in 2021 are also listed.

3.4.5 Stomach Cancer

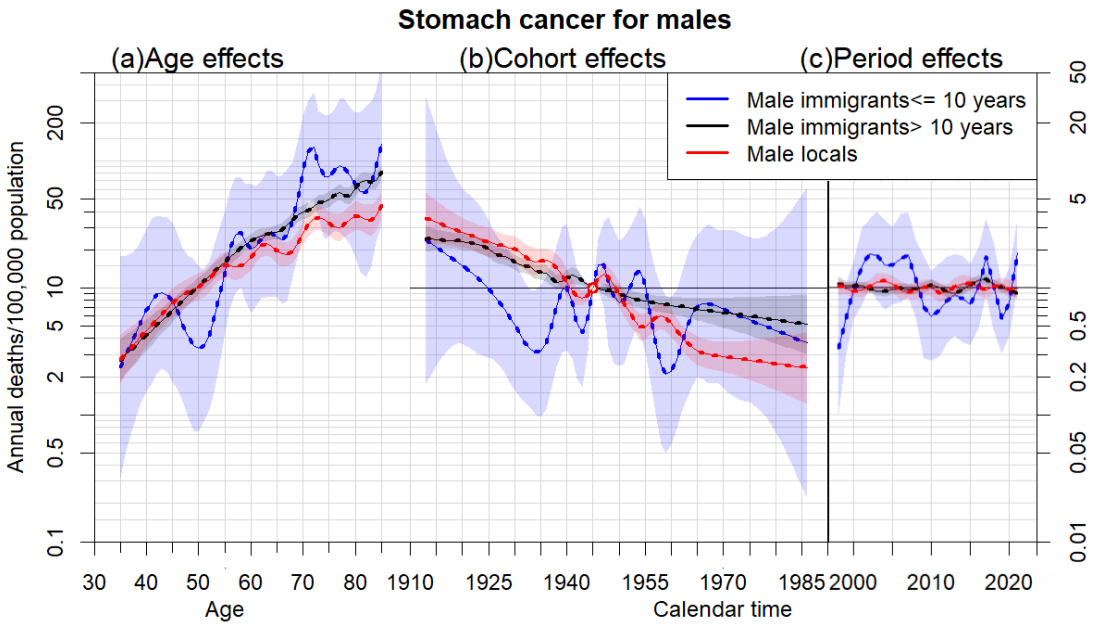


Figure 7a. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of male stomach cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

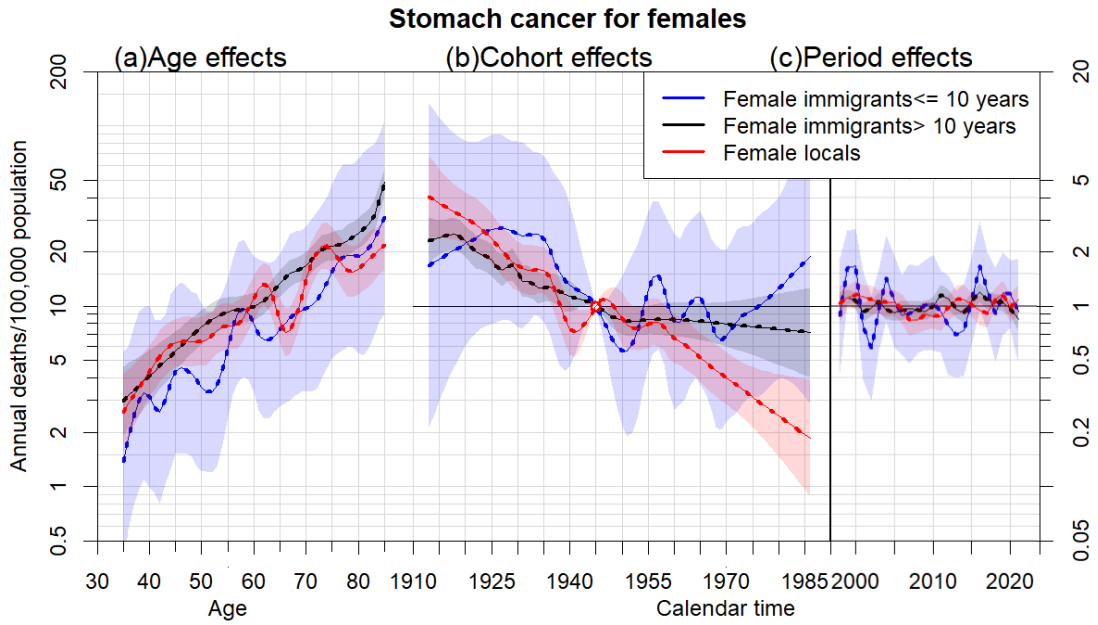


Figure 7b. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of female stomach cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

Compared to the age effect of other types of cancer, relatively insignificant differences in stomach cancer mortality rates for both genders between locals and long-stay immigrants have performed before 55, and stomach cancer mortality rates for long-stay immigrants are lower than those of locals after 55 for both genders. Short-stay male immigrants who was aged younger than 50 suffered lower mortality rate, while those aged older than 70 suffered higher mortality rate than other two immigration groups, and female short-stay immigrants were at lower mortality risk of stomach cancer than others in Figure 7. In addition to compatible dynamics of period effect for locals and long-stay immigrants, similar changes of relative risks by birth cohort for locals and long-stay immigrants in stomach occurred before 1955, whereas significant differences of relative risks by birth cohort between these two immigration groups occurred after

1960. Short-stay immigrants who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks affected by period effects before 2020 than those for locals and long-stay immigrants. Consequently, immigrants for both gender who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks of stomach cancer mortality affected by cohort and period effects than locals and immigrants who have stayed in Hong Kong for >10 years, and relative risks of stomach cancer mortality affected by period for locals and long-stay immigrants perform significant consistency.

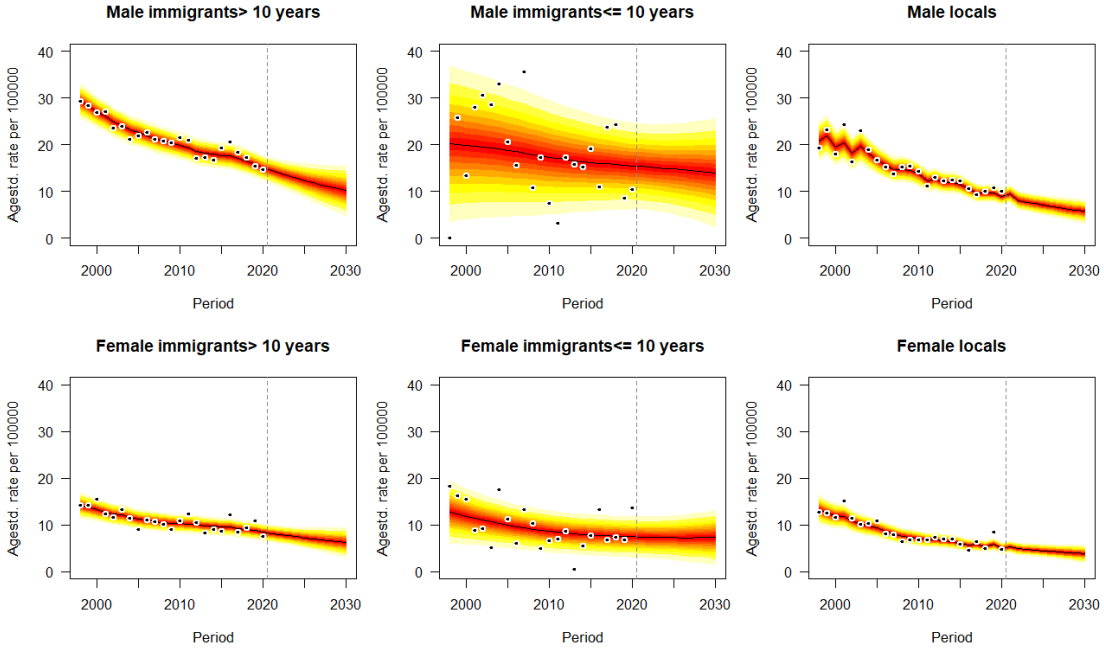


Figure 7c. Projections of stomach cancer mortality rates by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

Given the projected trends in Figure 7c, immigrants for each gender, especially who have stayed in Hong Kong for < 10 years will suffer from higher mortality rates of stomach cancer in each year than locals after 2021. Men would suffer from higher mortality rates of liver cancer in the future than females in the same immigration group. Monotone decreasing trends ($p < 0.05$) or plateau ($p > 0.05$) of forecasting occur for both genders and all immigration groups. Compared with other immigration groups, male immigrants who have stayed in Hong Kong for < 10 years with stomach cancer would perform the most significant decline in predictive mean from 15.22 (95% CI: 12.91, 19.20) to 10.15 (95% CI: 6.41, 18.27) deaths per 100,000 population ($p < 0.05$, Avg -0.51 deaths/100,000 per annum) (Table 6). In 2030, the highest mortality rate of stomach cancer would be 14.03 deaths/100,000 (95% CI: 10.53, 19.22) for male immigrants who have stayed in Hong Kong for < 10 years, while the lowest mortality rate of stomach cancer would be 3.83 deaths/100,000 (95% CI: 0.57, 7.92) for female local.

eFigure 6 in **Appendices** illustrate the age-standardized mortality rates of stomach cancer from 1998 to 2021 and their projections by sex, immigrant status and two age groups from 2022 to 2030. Most of predictive trends for younger cases (<60 years) and older cases (≥60 years) reach a consensus with those for all ages population in Figure 7c, except for mortality rates of stomach cancer for female immigrants >10 that decline for all ages ($p < 0.05$) vs. insignificant trend for younger cases ($p > 0.05$) vs. insignificant trend for older cases ($p > 0.05$). It's also reasonable that elders would be at higher risk of death by stomach cancer than youngers regardless immigration groups and genders. Male individuals would also suffer from higher mortality rate of liver cancer than females for youngers and elders.

Predictive mean of age-standardized mortality rates of stomach cancer per 100,000 population										
Year	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Female immigrants >10	8.20 (0.55)	7.95 (0.62)	7.71 (0.74)	7.47 (0.87)	7.25 (1.01)	7.03 (1.15)	6.83 (1.29)	6.62 (1.43)	6.43 (1.57)	6.24 (1.71)
Female immigrants ≤ 10	7.51 (1.44)	7.36 (1.56)	7.33 (1.69)	7.30 (1.85)	7.28 (2.01)	7.27 (2.20)	7.27 (2.40)	7.28 (2.61)	7.31 (2.84)	7.33 (3.09)
Female locals	5.26 (0.40)	4.91 (0.52)	4.75 (0.57)	4.61 (0.63)	4.47 (0.71)	4.34 (0.77)	4.21 (0.84)	4.08 (0.91)	3.95 (0.99)	3.83 (1.06)
Male immigrants >10	15.22 (0.64)	13.89 (0.97)	13.34(1.21)	12.81 (1.46)	12.31 (1.73)	11.83 (1.99)	11.38 (2.26)	10.95 (2.51)	10.54 (2.76)	10.15 (3.01)
Male immigrants ≤10	15.83 (3.04)	15.21 (3.38)	15.07 (3.67)	14.93 (3.98)	14.79 (4.31)	14.64 (4.65)	14.51 (5.02)	14.35 (5.39)	14.19 (5.78)	14.03 (6.17)
Male locals	8.14 (0.89)	8.07 (0.99)	7.73 (1.03)	7.41(1.07)	7.10 (1.13)	6.81 (1.19)	6.51 (1.26)	6.23 (1.33)	5.97 (1.39)	5.71 (1.46)
Female immigrants >10(<60y)	4.81 (0.56)	4.69 (0.79)	4.62 (0.87)	4.55 (0.96)	4.47 (1.07)	4.39 (1.17)	4.31 (1.29)	4.22 (1.41)	4.13 (1.52)	4.03 (1.64)
Female immigrants ≤ 10(<60y)	3.89 (0.80)	4.08 (0.93)	4.10 (1.03)	4.13 (1.14)	4.17 (1.27)	4.21 (1.41)	4.24 (1.55)	4.28 (1.70)	4.32 (1.87)	4.36 (2.05)
Female locals(<60y)	2.28 (0.21)	2.08 (0.27)	1.98 (0.29)	1.88 (0.32)	1.79 (0.35)	1.71 (0.37)	1.61 (0.41)	1.53 (0.43)	1.44 (0.45)	1.37 (0.47)
Male immigrants >10(<60y)	4.94 (0.57)	4.71 (0.79)	4.55 (0.89)	4.41 (0.99)	4.25 (1.10)	4.12 (1.21)	3.98 (1.32)	3.86 (1.43)	3.74 (1.54)	3.63 (1.65)
Male immigrants ≤ 10(<60y)	4.81 (1.31)	4.70 (1.42)	4.66 (1.55)	4.63 (1.69)	4.59 (1.83)	4.55 (1.99)	4.52 (2.15)	4.48 (2.32)	4.44 (2.50)	4.41 (2.68)

Male locals(<60y)	2.48 (0.21)	2.37 (0.29)	2.28 (0.32)	2.21 (0.35)	2.12 (0.38)	2.04 (0.42)	1.97 (0.45)	1.91 (0.49)	1.83 (0.52)	1.77(0.55)
Female immigrants >10(≥60y)	17.80 (1.04)	16.23 (1.26)	15.65 (1.47)	15.08 (1.70)	14.55 (1.94)	14.03 (2.18)	13.54 (2.43)	13.07 (2.68)	12.62 (2.92)	12.19 (3.16)
Female immigrants ≤ 10(≥60y)	14.72 (4.29)	13.01 (4.83)	12.52 (5.11)	12.03 (5.37)	11.55 (5.63)	11.08 (5.88)	10.63 (6.12)	10.19 (6.35)	9.76(6.56)	9.34 (6.75)
Female locals(≥60y)	12.20 (1.66)	11.86 (1.84)	11.67 (1.98)	11.49 (2.15)	11.33 (2.35)	11.18 (2.58)	11.04 (2.84)	10.91 (3.11)	10.79(3.4)	10.68 (3.71)
Male immigrants >10(≥60y)	37.23 (2.29)	36.59 (2.56)	35.17(3.18)	33.82 (3.86)	32.55 (4.57)	31.34 (5.28)	30.19 (6.01)	29.08 (6.70)	28.02 (7.40)	27.01 (8.07)
Male immigrants ≤ 10(≥60y)	42.30 (10.88)	41.43 (11.78)	41.03 (12.71)	40.61 (13.70)	40.17 (14.75)	39.71 (15.85)	39.24 (16.99)	38.75 (18.16)	38.23 (19.35)	37.71 (20.57)
Male locals(≥60y)	23.04 (3.29)	22.69 (3.56)	22.37(4.07)	22.16(4.84)	21.89 (5.86)	21.61 (7.22)	21.52 (9.02)	21.74 (11.29)	22.17 (14.03)	22.73 (17.28)

Table 6. Predictive means and standard deviations (in brackets) of age-standardized mortality rates of stomach cancer per 100,000 population for each gender, age group (less than, greater or equal to 60 years old) and immigrant status from 2022 to 2030. Reported means and standard deviations (in brackets) of age-standardized mortality rates in 2021 are also listed.

3.4.6 Prostate Cancer

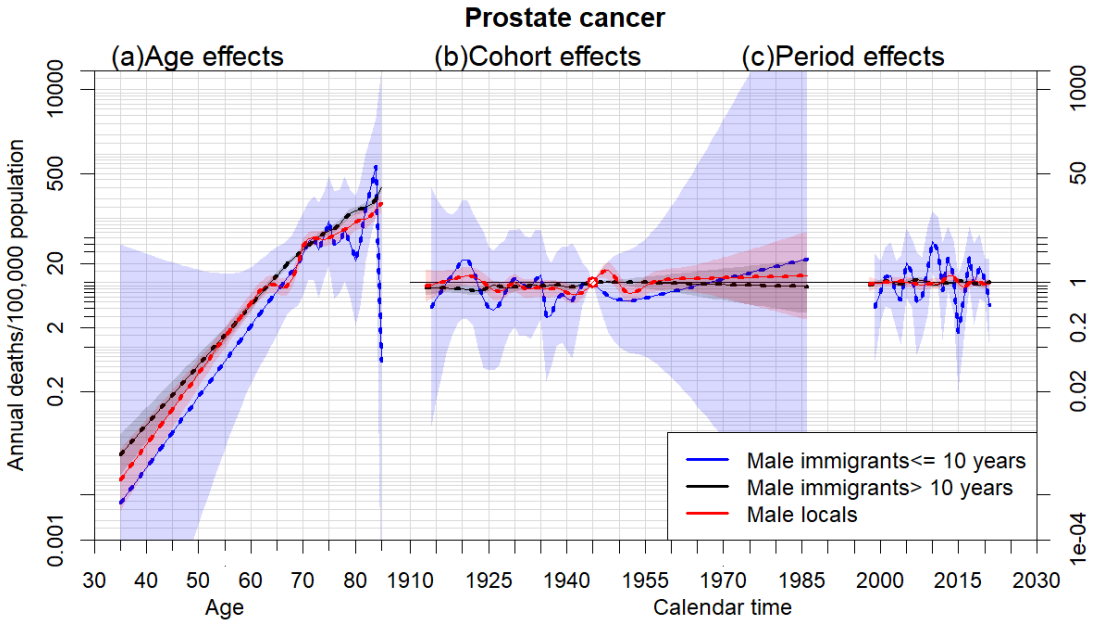


Figure 8a. Parameter estimates of age (a), cohort (b) and period (c) effects based on an age-period-cohort model of male prostate cancer mortality rates by immigrant groups: locals, immigrants stay in Hong Kong for more than 10 years and immigrants stay in Hong Kong for less than or equal to 10 years. Age effect was assessed by mortality (left axis). Cohort and period effects were assessed by relative risk (right axis), 95% confidence intervals are shown as shaded bands.

Unlike other types of cancer, only the APC effects of prostate cancer for males illustrate in Figure 8. Compared to the age effect of other types of cancer, relatively insignificant differences in prostate cancer mortality rates between locals and long-stay immigrants have performed. Short-stay male immigrants who was aged younger than 65 suffered lower mortality rate, while those aged older than 80 suffered higher mortality rate than other two immigration groups in Figure 7. A sharp decline of mortality by age for short-stay immigrants is performed resulted from lack of data of deaths aged 80 or older. In addition to compatible dynamics of period effect for locals

and long-stay immigrants, similar changes of relative risks by birth cohort for locals and long-stay immigrants in prostate occurred before 1955, whereas significant differences of relative risks by birth cohort between these two immigration groups occurred after 1960. Short-stay immigrants who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks affected by period effects than those for locals and long-stay immigrants. Consequently, immigrants for both gender who have stayed in Hong Kong for ≤10 years had more fluctuating relative risks and more broad confidence interval of stomach cancer mortality affected by cohort and period effects than locals and immigrants who have stayed in Hong Kong for >10 years, and relative risks of prostate cancer mortality affected by period for locals and long-stay immigrants perform significant consistency.

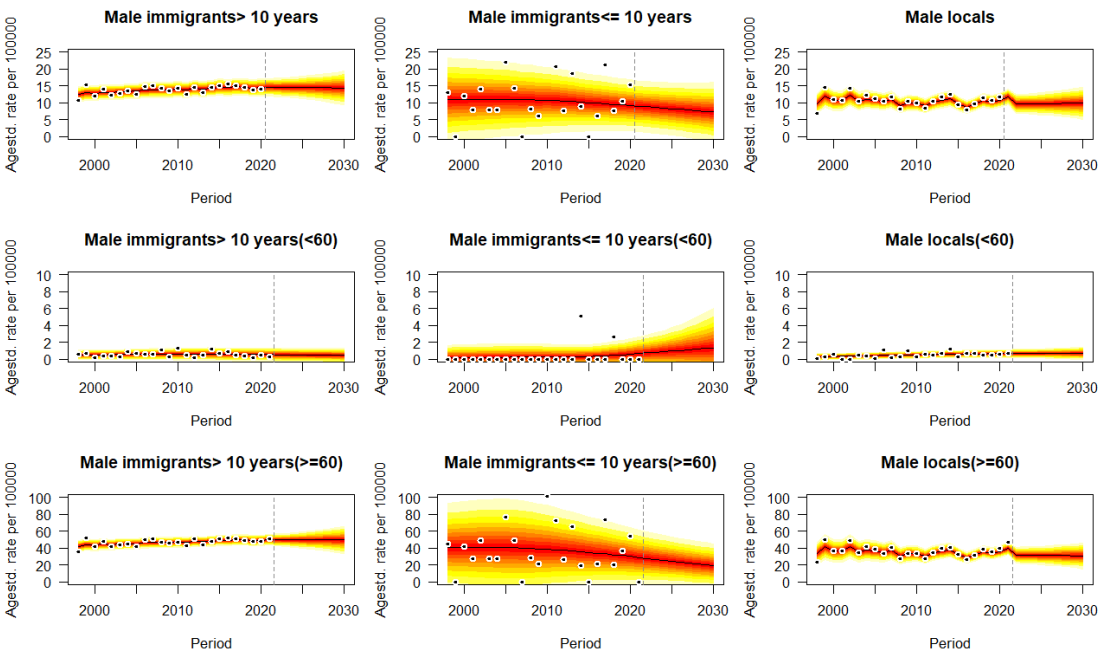


Figure 8b. Projections of prostate cancer mortality rates for males by immigrant status and age groups (less than, greater than or equal to 60 years old) from 2022 to 2030. Observations are shown as dots

with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

Given the projected trends in Figure 8b, male immigrants, especially who have stayed in Hong Kong for > 10 years will suffer from higher mortality rates of prostate cancer in each year than locals after 2021. Monotone decreasing trends ($p < 0.05$) or plateau ($p > 0.05$) of forecasting occur for all immigration groups. Compared with other immigration groups, male immigrants who have stayed in Hong Kong for < 10 years with prostate cancer would perform the most significant decline in predictive mean from 9.03 (95% CI: 5.91, 17.01) to 7.27 (95% CI: 2.91, 12.12) deaths per 100,000 population ($p < 0.05$, Avg -0.18 deaths/100,000 per annum) (Table 7). In 2030, the highest mortality rate of prostate cancer would be 14.38 deaths/100,000 (95% CI: 11.13, 20.62) for male immigrants who have stayed in Hong Kong for > 10 years, while the lowest mortality rate of prostate cancer would be 7.27 deaths/100,000 (95% CI: 2.91, 12.12) for male immigrants who have stayed in Hong Kong for < 10 years.

Figure 8b also illustrate the age-standardized mortality rates of prostate cancer from 1998 to 2021 and their projections by sex, immigrant status and two age groups from 2022 to 2030. Most of predictive trends for younger cases (<60 years) and older cases (≥60 years) reach a consensus with those for all ages population, except for mortality rates of prostate cancer for male immigrants <10 that decline for all ages ($p < 0.05$) vs. increasing trend for younger cases ($p < 0.05$) vs. decline for older cases ($p < 0.05$). It's also reasonable that elders would be at higher risk of death by prostate cancer than youngers regardless immigration groups and genders. Some particular cases occur in the projection of prostate cancer that young long-stay male immigrants (0.44 deaths/100,000, 95% CI: 0, 1.05) aged less than 60 will be at lower mortality rate than locals (0.69 deaths/100,000, 95% CI: 0, 1.42) in 2030 (Table 7).

Predictive mean of age-standardized mortality rates of prostate cancer per 100,000 population										
Year	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**
Male immigrants >10	14.81 (0.61)	14.59 (0.79)	14.57 (0.96)	14.56 (1.15)	14.54 (1.37)	14.51 (1.61)	14.48 (1.86)	14.45 (2.13)	14.42 (2.42)	14.38 (2.72)
Male immigrants ≤10	9.03 (2.95)	8.78 (3.11)	8.58 (3.29)	8.39 (3.49)	8.19(3.69)	8.10(3.89)	7.82(4.11)	7.63(4.31)	7.45(4.51)	7.27(4.72)
Male locals	9.54 (1.40)	9.66 (1.57)	9.67 (1.66)	9.69 (1.77)	9.72(1.91)	9.75(2.06)	9.78(2.23)	9.82(2.43)	9.86(2.64)	9.9(2.88)
Male immigrants >10(<60y)	0.57 (0.12)	0.52 (0.17)	0.51 (0.19)	0.50 (0.21)	0.49(0.22)	0.48(0.24)	0.47(0.25)	0.46(0.27)	0.45(0.29)	0.44(0.31)
Male immigrants ≤ 10(<60y)	0.65 (0.59)	0.73 (0.77)	0.81 (0.93)	0.87 (1.10)	0.94(1.31)	1.01(1.51)	1.09(1.75)	1.16(2.02)	1.24(2.32)	1.33(2.64)
Male locals(<60y)	0.63 (0.12)	0.66 (0.14)	0.66 (0.16)	0.66 (0.19)	0.67(0.21)	0.67(0.24)	0.67(0.27)	0.68(0.31)	0.68(0.33)	0.69(0.37)
Male immigrants >10(≥ 60y)	49.43 (2.59)	49.61 (2.73)	49.63 (3.29)	49.64 (3.94)	49.64 (4.68)	49.64 (5.51)	49.63 (6.38)	49.62 (7.32)	49.61 (8.32)	49.58(9.37)
Male immigrants ≤ 10(≥60y)	28.29 (9.15)	27.66 (9.78)	26.53 (10.21)	25.4 (10.63)	24.28 (11.03)	23.16 (11.41)	22.07 (11.76)	21.01 (12.09)	19.96 (12.38)	18.95(12.63)
Male locals(≥60y)	31.57 (5.17)	31.48 (5.49)	31.40 (5.76)	31.32 (6.09)	31.24 (6.48)	31.15 (6.94)	31.06 (7.44)	30.96 (8.01)	30.86 (8.61)	30.74(9.26)

Table 7. Predictive means and standard deviations (in brackets) of age-standardized mortality rates of prostate cancer per 100,000 population for each age group (less than, greater or equal to 60 years old) and immigrant status from 2022 to 2030. Reported means and standard deviations (in brackets) of age-standardized mortality rates in 2021 are also listed.

3.5 Discussion

Early detection of cancer is positive and instructive for increasing chances of cure. Nevertheless, the high mortality rate of cancer results from late diagnosis among most patients after progression to more advanced or severe stages. Individuals at high risk of cancer, such as smokers, alcoholics or those who are frequently exposed to susceptible circumstances, should be screened for early detections to increase opportunities for cure [103]. Therefore, the differences in mortality rates among immigration groups are synonymous with detection means, therapies, and social history in disparate periods and areas.

While the changes in mortality rates by age for long-stay immigrants reached approximate harmony with those for locals, the changes in mortality rates by age for short-stay immigrants revealed clear differences with those for the other two populations. The group of long-stay immigrants had a higher risk of death from lung, colon and liver cancers than the other two immigration groups after the age of 60 years. Short-stay male immigrants were less likely to die from lung cancer before the age of 65 years. The contrast in age effects among the immigration groups was partially consistent with studies [88][104] that highlighted the age effects for locals and immigrants on breast cancer mortality in Hong Kong and lung cancer incidence in Sweden, as they both showed similar trends and magnitudes between locals and immigrants before the age of 60 years. They are also compatible with the results in [105] that diagnosis of liver cancer is the most frequent among populations at 55 to 65 years old. According to these trends, young individuals, especially new young immigrant men, who have benefited from all-rounded development in mainland China and Hong Kong, are more likely to seek early detection and be treated for cancers using more advanced treatments [106]. Differences in birth cohort effects among immigrant groups partially comply with the interpretation above.

We observed significant trends of cohort effects among locals and immigrants. These findings are partially consistent but subtly different from previous findings, regarding the effect of immigration status on cancers. Zhao et al. [88] described multiple peaks of cohort effects on breast cancer mortality between locals and immigrants in Hong Kong, as well as a significant decline in cohort effects after 1950. In contrast, Sung et al. [107] investigated the difference in breast cancer incidence between Chinese Americans and non-Hispanic whites in the U.S. and emphasized that Chinese Americans were at lower risk of breast cancer than non-Hispanic whites born in the same year. Here, we interpret the cohort-driven trends resulting from the intricacy of social history and lifestyle. Compared to a relatively stable social development in Hong Kong, representing downward trends of relative risks for locals, wars and social instability in mainland China resulted in several immigration waves from mainland China to Hong Kong before 1950. Additionally, remarkable increasing trends were recorded for new immigrants after 1950, which corresponded to the economic downturn after wars and famine between 1959 and 1961 during their youth[108].

The increasing trends for new immigrants and similar trends for locals and long-stay immigrants were consistent with the finding that nutrient deficiency contributes to a higher risk of severe mortality rates of cancers [109]. Furthermore, we speculate that these trends, especially those for locals and long-stay immigrants, are most likely attributed to social development and personal behaviors, such as daily habits, occupational history, different diagnoses and treatments, and domestic environmental exposures. Notably, short-stay immigrants suffered from a lower risk of death from colon cancer for all ages (Figure 4c). As locals and immigrants in Hong Kong transitioned to more westernized lifestyles, higher consumption of meat was associated with a higher risk of these types of cancer, whereas consumption of vegetables had a strong protective effect against pancreatic cancer, and moderate consumption of coffee appeared to be beneficial against lung cancer [109][110].

Further studies on potential risk factors are required.

Short-stay immigrants had more fluctuating and non-stationary but inconspicuous relative risks by period effects before 2021 than locals and long-stay immigrants. Cumulatively, an arch pattern and fluctuating curve depicting period effects externally resulted in an arch pattern of age-standardized mortality rates for short-stay immigrant women and irregular rates for short-stay immigrant men before 2021. The external performance of different period effects on mortality rates could be most likely attributed to the higher effect of different lifestyles and social development on new immigrants than on long-stay immigrants and locals in Hong Kong. For the age-standardized mortality rates and projections, consistent with previous findings [111][112], we predict that the mortality rates of cancer in Hong Kong after 2021 will continue to decline or remain relatively stable, consistent with the trends before 2020, except for male immigrants who have stayed in Hong Kong for ≤10 years with colon cancer and male immigrants who have stayed in Hong Kong for >10 years with pancreatic cancer. Men will be at higher risk of mortality rates of cancer than women, regardless of immigration status. They are also compatible with the results in [113] that men suffer from a higher risk of these types of cancer than women, excluding prostate cancer. Furthermore, new immigrant women will be at lower risk than local women, even though long-stay immigrants will suffer from higher mortality rates than locals in the future. Potential interpretations could be consistent with those for birth cohort effects, as age and period effects are considered as confounders of cohort effects.

In the past few decades, spurred by an increasing burden of high incidence and mortality rates of cancer, several studies focused on the inherent identification dilemma of three effects in the APC model. Further, complicated population distribution and immigration status in Hong Kong, one of the areas with the highest population density and migration frequency in the world, have intricate causes and inherent dynamics of cancer and other diseases. To our knowledge, few studies have assessed the relationship between immigration status and cancer mortality. Therefore,

this study is original to examine the effect of the length of stay in Hong Kong and origins of previous residence on cancer deaths, which is instructive for further immigration policy-making and targeted strategies of disease detection and intervention. However, this study had several limitations. Given the non-identifiability problem in age-period-cohort models, we could only depict trends and variations among different immigration and sex groups, as illustrated in figures, and insufficiently perform the estimates of the contributions of three effects or subgroups to mortality rates. Furthermore, we adopted a cubic smoothing spline to estimate populations of immigrants and locals due to the large proportion of unspecified immigration status from official demographic projections. A few acceptable cases resulted in a limited type of cancer so that some common cancers, such as the ovary and cervix, were discarded. Since the issue of quantification, the future perspective of cancer therapies and techniques have not been considered in the model of projection.

To explore the relationship between immigration groups and cancer mortality, this study aimed to explore age, period, birth cohort effects and effects across genders and immigration groups on mortality rates of lung, pancreatic, colon, liver, prostate and stomach cancers and their projections. Death registry data in Hong Kong between 1998 and 2021, which were stratified by age, sex and immigration status. Immigration status was classified into three groups: locals born in Hong Kong, long-stay immigrants and short-stay immigrants. Age-period-cohort analysis was used to examine age, period, and birth cohort effects for genders and immigration groups from 1998 to 2021. Bayesian age-period-cohort models were applied to predict the mortality rates from 2022 to 2030. Short-stay immigrants revealed pronounced fluctuations of mortality rates by age and of relative risks by cohort and period effects for six types of cancers than those of long-stay immigrants and locals. Immigrants for each type of cancer and gender will be at a higher mortality risk than locals. After 2021, decreasing trends ($p<0.05$) or plateau ($p>0.05$) of forecasting mortality rates of cancers occur for all immigration groups, except for increasing trends for short-stay male immigrants with colon cancer ($p<0.05$, Avg +0.30 deaths/100,000 per annum from 15.47 to 18.50

deaths/100,000) and long-stay male immigrants with pancreatic cancer ($p<0.05$, Avg +0.72 deaths/100,000 per annum from 16.30 to 23.49 deaths/100,000).

We conclude that immigrants, especially short-stay immigrants, had more pronounced fluctuations of mortality rates by age and of relative risks by cohort and period effects for six types of cancers than those of long-stay immigrants and locals. Male immigrants who have stayed in Hong Kong for ≤10 years with colon cancer and male immigrants who have stayed in Hong Kong for >10 years with pancreatic cancer would perform significant uptrend in the future, while other immigration groups for each type of cancer would continue to decline or remain relatively stable. Immigrants for each gender in Hong Kong would suffer from higher mortality risks of cancers than locals in the future.

The conclusions that immigrants, males, and elders would face higher mortality risks from cancer compared to local populations carries profound implications for public health interventions and healthcare planning moving forward. Specific interventions are required to reduce disparities in cancer outcomes among these vulnerable groups, involved with tailored screening programs to facilitate early detection, culturally sensitive health education campaigns to promote awareness to precaution, and improved access to quality healthcare services for immigrants, males, and older individuals. Addressing underlying social determinants of health, such as language barriers, socioeconomic status, and healthcare access disparities, will be crucial in mitigating the projected higher mortality rates from cancer in these populations. By prioritizing specific strategies and fostering inclusivity in healthcare delivery, we could reach to equal opportunities of optimal healthcare of cancers and other chronic diseases for every individual regardless of culture or background.

However, we could also realize some weaknesses of INLA as the Bayesian APC model based on INLA was applied to the data of chronic disease. Although INLA has brought out high efficiency and accuracy when processing complex models, there are still some limitations [22]. For example:

1). Strict model assumptions: assumptions of INLA about the model are strict, which requires the properties of model should be parallel with those of Gaussian Markov random fields, so that it is not suitable for all types of models.

2). Complexity on parameter selection: There are many parameters which need to be selected in INLA, including improper selection of grid size, number of nodes, etc. They may lead to inaccurate results.

3). Requirements for data size: INLA has high requirements for data size. If the sample size is too small, the inference results may be unreliable.

Therefore, the performance of MCMC and INLA in epidemiology has triggered our curiosity and it's expounded in detail in chapter 4.

4. Performance of MCMC and INLA

4.1 Introduction

Methods derived from Bayesian inference have been increasingly worth heeding on parameter estimation and modeling in epidemiology. As two representative methods of Bayesian inference, MCMC and INLA perform different effectiveness and accuracy on model fitting for different types of data, and also reveal different advantages and disadvantages. With various circumstances, which method or assumption is more appropriate to adopt to obtain robust and reliable results has already been a worthy topic in modeling of epidemiology.

Unlike traditional uniform sampling, Markov chain Monte Carlo sampling adjusts the proposed sampling distribution function to approach the objective function, which reveals that sampling from the proposed distribution is equivalent to sampling the objective function. MCMC is an attractive method for Bayesian inference due to its flexibility and versatility. Any target distribution can be handled without the need for analysis or numerical integration, as long as the density or likelihood function can be evaluated. Furthermore, MCMC is robust and consistent, and does not rely on assumptions or approximations about the target distribution. Additionally, given enough samples and time, convergence to the true posterior distribution is guaranteed. Additionally, MCMC is easy to implement and use, and there are many packages and libraries available for different languages and platforms. It can also be customized and expanded to suit your individual needs and preferences.

However, MCMC is not a foolproof solution for Bayesian inference and has some challenges and limitations that should be addressed [23]. For example, it can be slow and inefficient, taking a long time and many samples to reach convergence and obtain accurate estimates. It can also be tricky and deceptive, producing misleading or

erroneous results if the right algorithm, parameters, or samples are not chosen. Additionally, MCMC can be complex and confusing, involving many technical details that are not intuitive or familiar [27]. To combat these issues, we may need to optimize your algorithm, use parallel computing, monitor and diagnose your MCMC, use multiple chains, compare different algorithms, learn the theory and practice of MCMC, and understand the specificities of your problem and model.

Compared to MCMC, INLA transforms the Bayesian inference problem into an approximation problem of solving Gaussian Markov random fields by decomposing parameters into fixed effects and random effects and exploiting the properties of Gaussian Markov random fields. Compared to some traditional Bayesian inference methods such as MCMC, INLA has emerged high computational efficiency, since INLA uses the Laplace approximation method, which avoids the large number of Monte Carlo sampling required in traditional methods, with lower computation consumptions [66]. Furthermore, with high computational efficiency, the posterior distribution is more accurately approximated, so more accurate inference can be obtained. Additionally, INLA can also be applied to a variety of different models, including linear models, generalized linear models, and some nonlinear models, etc.

However, we could also realize some weaknesses of INLA as the Bayesian APC model based on INLA was applied to the data of chronic disease in chapter 3. Although INLA has brought out high efficiency and accuracy when processing complex models, there are still some limitations. For example [63]:

1). Strict model assumptions: assumptions of INLA about the model are strict, which requires the properties of model should be parallel with those of Gaussian Markov random fields, so that it is not suitable for all types of models.

2). Complexity on parameter selection: There are many parameters which need to be selected in INLA, including improper selection of grid size, number of nodes, etc. They may lead to inaccurate results.

3). Requirements for data volume: INLA has high requirements for data size. If the

sample size is too small, the inference results may be unreliable.

4.2 Objective

With data of chronic diseases in last chapter, projections of cancer mortality based on MCMC Bayesian APC model are performed. As what have been mentioned in chapters above regarding strengths and weaknesses of MCMC and INLA with different types of data, the contrast of performance of them is demonstrated in this chapter to determine the superior method based on different criteria by evaluating their performance of 10-year retrospective projections.

4.3 Data and methods

4.3.1 Data

Similar death registry data from the Census and Statistics Department of Hong Kong was applied again, related to six types of cancer: lung cancer, colon cancer, liver cancer, stomach cancer, pancreatic cancer for males and females and prostate cancer for males, to adopt MCMC sampling in this chapter. Beside the introduction of dataset in last chapter, more preprocessing of data was considered in this chapter for MCMC. The predictive means of age-standardized cancer mortality rates for each sex and migrant status, taking into account age, period, and birth cohort effects, were calculated based on the weights of population age groups from the WHO World Standard population [102]. With the demography of mortality rates of six types of cancer from

1998 to 2021, the evaluation of projection performance between MCMC and INLA in Bayesian APC model is to compare the quantities of retrospective projection of different genders, immigration groups and types of cancer in 10 years from 2012 to 2021.

With data of transmission clusters of COVID-19 in Chapter 2, estimations of transmission potential and heterogeneity based on INLA are excluded in the main body due to unreliable performances. One potential cause is lack of samples for INLA estimation as 545 infectee-infector transmission pairs were constructed. The posterior distributions of the effective reproduction numbers and dispersion parameters for each contact setting can be approximated with integrated nested Laplace approximations by applied the INLA package. More details can be found in **Appendices A4.**

4.3.2 Methods

Recall Equation 9 in chapter 3.3.2, a linearized APC model can be rewritten as

$$Y_{apc} = \log(\mu_{apc}) = \beta_0 + \alpha(a - a^*) + \pi(p - p^*) + \gamma(k - k^*) + \tilde{\alpha}_a + \tilde{\pi}_p + \tilde{\gamma}_c + \varepsilon_{apc} \quad (12)$$

Beside other definition of notations in formula 10 and 11, to set up the assumption of identification issue in APC model, we assume that

$$\alpha^* = \alpha + v \quad \pi^* = \pi - v \quad \gamma^* = \gamma + v \quad (13)$$

where v denoted as undetermined scale to combine with true unknow linear effects α, π and γ. This scale can be considered in maximum likelihood estimation in APC model. Assume $a = 1, ..., A$, $p = 1, ..., P$ and $c = 1, ...C$ as age, period and cohort group, respectively, and the asterisk a^*, p^* and c^* are the midpoint of the

range of age, period and cohort group. Similarly, a th age effect can be represented as an overall linear age effect with a th non-linear age effect as $\alpha_a = (a - a^*)\alpha + \tilde{\alpha}_a$ with $a^* = (A+1)/2$ [114]. Other notations of period and cohort effects are similar above. Furthermore, the model and prior distribution for estimation in Bayesian APC can be specified from equation 12 as follows.

$$\begin{aligned} &Y_{apc} \sim N\left(\hat{Y}_{apc}, \sigma^2\right) \\ &\hat{Y}_{apc} = \beta_0 + \alpha\Lambda_L + \pi\Pi_L + \gamma\Upsilon_L + \sum_{2}^{A-1}\tilde{\alpha}_a\Lambda_a + \sum_{2}^{P-1}\tilde{\pi}_p\Pi_p + \sum_{2}^{C-1}\tilde{\gamma}_c\Upsilon_c \end{aligned} \tag{14}$$

$$\beta_0 \sim N(\mu, \sigma_\mu^2) \tag{15}$$

$$\begin{aligned} &\alpha \sim Uniform(a_\alpha, b_\alpha) \\ &\pi \sim Uniform(a_\pi, b_\pi) \\ &\gamma \sim Uniform(a_\gamma, b_\gamma) \end{aligned} \tag{16}$$

$$\begin{aligned} &\tilde{\alpha}_a \sim N(\mu_{\tilde{\alpha}_a}, \sigma^2_{\tilde{\alpha}_a}) \ \ if\ a = 2, ..., l-1 \\ &\tilde{\pi}_p \sim N(\mu_{\tilde{\pi}_p}, \sigma^2_{\tilde{\pi}_p}) \ \ if\ p = 2, ..., m-1 \\ &\tilde{\gamma}_c \sim N\left(\mu_{\tilde{\gamma}_c}, \sigma^2_{\tilde{\gamma}_c}\right) \ \ if\ c = 2, ..., n-1 \end{aligned} \tag{17}$$

$$\begin{aligned} &\tilde{\alpha}_a \sim Laplace(\mu_{\tilde{\alpha}_a}, \sigma^2_{\tilde{\alpha}_a}) \ \ if\ a = 2, ..., A-1 \\ &\tilde{\pi}_p \sim Laplace(\mu_{\tilde{\pi}_p}, \sigma^2_{\tilde{\pi}_p}) \ \ if\ p = 2, ..., P-1 \\ &\tilde{\gamma}_c \sim Laplace\left(\mu_{\tilde{\gamma}_c}, \sigma^2_{\tilde{\gamma}_c}\right) \ \ if\ c = 2, ..., C-1 \end{aligned} \tag{18}$$

In Equation 14, Λ_L, Π_L and Υ_L are regarded as linear effects and Λ_a, Π_p and Υ are non-linear. Estimation in Bayesian inference, such as MCMC, can be adopted with the likelihood model (Equation 14) and the prior distributions of model parameters (Equation 15-18) [115]. Laplace distribution is regarded as the prior distribution of high degree polynomials, since there is a spiked concentration closed to zero. It's also equivalent to the constraining of sum of zero (Equation 10). The estimations of the Bayesian APC model based on MCMC can be attained with the

package BAMP [116] and those based on INLA can be attained with the BAPC package [100]. The main idea of performance contrast is to build up 10-year retrospective projections and true value from 2012 to 2021 with criteria expounded below.

Proper scoring rule

Proper score rule is adopted to evaluate the probabilistic forecast performance of two methods. The scoring rule is often used to score the effectiveness of prediction of an event. The proper scoring rule is defined that, when designing the scoring rule, we generally expect that the model of forecast can get higher score only when it honestly produces true view of the event [117]. In other words, for a proper score, the forecaster maximizes the expected reward if forecasts are consistent with the true distribution. It's often assumed that all forecasters own the prior knowledge of the predicted event in real life, and they can obtain more information and thus have corresponding posterior knowledge with one more step of forecast [117]. Intuitively, the rule designer would want people to be able to obtain a posterior knowledge and report it honestly. However, for some common scoring rules, there are some scenarios where whether the answer is a posteriori knowledge with little impact on the final score. Therefore, the predictors have no motivation for one more step and are only satisfied with the answers based on a prior distribution.

Continuous Ranked Probability Score (CRPS) is the criterion, quantifying the contrast between the observed value and theoretical value of continuous probability distribution. It often be applied as the loss function or the criteria of evaluating function of a probabilistic model, which can be adopted in real-life problems such as probabilistic weather forecasting and error analysis [118]. As an evaluation function,

the results obtained by evaluating the probability model based on CRPS are equivalent to the results from evaluating the expectations of the probability model based on mean absolute error [118][119]. As the description of data in chapter 4.3.1 and 3.3.1, we assume y_{ij} as observed counts of deaths aged at i in year j, then CPRS for the ij th forecast is

$$\begin{aligned} CRPS_{ij} &= \sigma_{ij}\left\{\tilde{y}_{ij}\left[2F\left(\tilde{y}_{ij}\right)-1\right]+2f\left(\tilde{y}_{ij}\right)-1/\sqrt{\pi}\right\} \\ \tilde{y}_{ij} &= \left(y_{ij}-\mu_{ij}\right)/\sigma_{ij} \end{aligned} \tag{19}$$

where μ_{ij} and σ_{ij} are the mean and standard deviation of predictive distribution, and $F(\cdot)$ and $f(\cdot)$ are the distribution and density function of normal distribution, respectively. CRPS can be regarded as the integral of the square of the difference between the cumulative distribution function and the step function in the real number domain, so it's equivalent to the generalization of the mean absolute error (MAE) on the continuous probability distribution [118]. That is, if $\sigma_{ij}=0$, CRPS can reduce to absolute error (AE) as

$$AE_{ij} = CRPS_{ij} = E|Y_{ij}-y_{ij}| - \frac{1}{2}E|Y_{ij}-Y'_{ij}| \tag{20}$$

To evaluate the performance of retrospective projections, we applied $\overline{CRPS}$ and MAE $\overline{AE}$, which are the mean $CRPS$ and AE of age standardized projection for all periods, respectively. Smaller values of them indicate less differences between theoretical values and observed values and better performance of retrospective projection.

Another calibration test based on CRPS was adopted with the test statistic [120]

$$z = \frac{\overline{CRPS} - E_0\left(\overline{CRPS}\right)}{\sqrt{Var_0\left(\overline{CRPS}\right)}} \tag{21}$$

where E_0 and Var_0 are the mean and variance of CRPS under the null hypothesis of expected calibration. For computation convenience, the retrospective projections are approximately independent among different age groups, so that the statistic is standard normally distributed. Lower p-value indicate more significant dispersion among projection and the observed value.

4.4 Results

Figure 9 and eFigure 7-11 in **Appendices A5** illustrate the contrast of performance on retrospective projections of six types of cancer based on INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021. Additionally, Table 8 and eTable 2-6 state the contrast criteria, such as mean absolute error, mean of CRPS, statistic and corresponding p-value of retrospective projections performance between inference methods and observations. Two methods expound significant projection performance (p-values >0.05) on both genders and most of immigration groups, except for: 1. Lung cancer: MCMC for short-stay males (p-value=0.01); 2. Colon cancer : MCMC for short-stay females (p-value=0.04), INLA and MCMC for short-stay males (p-value=0.01/0.01), and MCMC for local males (p-value=0.02); 3. Liver cancer: INLA for short-stay females (p-value=0.03); 4. Stomach cancer: INLA and MCMC for short-stay males (p-value< 0.01 and =0.01).

Furthermore, most of 10-year retrospective projections based on INLA indicate less dispersions with observations than those based on MCMC and outperform in most of types of cancer and immigration groups, except for some circumstances such as 1. Liver cancer (eFigure 8, eTable 3): short-stay females (p-value: INLA 0.03 vs. MCMC 0.17); 2. Stomach cancer (eFigure 10, eTable 5): short-stay females (p-value: INLA 0.14 vs. MCMC 0.18) and short-stay males (p-value: INLA <0.01 vs. MCMC 0.01); 3. Prostate cancer (eFigure 11, eTable 6): long-stay males (p-value: INLA 0.80

vs. MCMC 0.86) and short-stay males (p-value: INLA <0.44 vs. MCMC 0.52). Lack of data can be potential reason since INLA requires large sample size.

INLA estimates for short-stay males with stomach cancer indicates the most significant dispersions with observations and performs the worst projections (p-value< 0.01), while MCMC for long-stay males with prostate cancer performs the best simulation (p-value = 0.86). Both techniques for long-stay immigrants state better performance on the retrospective projection than those for short-stay immigrants, regardless of types of cancer and genders.

Figure 9. Contrast of retrospective projections of lung cancer mortality between INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021

Contrast of retrospective projections performance of lung cancer between INLA and MCMC								
Methods / Immigration	INLA				MCMC			
	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$
Female immigrants >10	14.86	13.46	0.66	0.51	18.97	14.22	0.92	0.32
Female immigrants ≤10	20.25	18.81	-1.33	0.27	21.27	18.20	1.77	0.08
Female locals	21.21	17.22	0.84	0.40	27.88	20.52	1.42	0.16
Male immigrants >10	20.70	17.52	1.44	0.15	24.33	17.27	1.62	0.10
Male immigrants ≤10	31.51	24.72	1.76	0.08	41.02	32.33	-2.56	0.01
Male locals	25.67	17.88	1.52	0.11	27.04	20.84	1.94	0.05

Table 8. Contrast of retrospective projections performance of lung cancer between INLA and MCMC for different immigration groups and genders. Mean absolute error, mean of CRPS, statistic and corresponding p-value are listed.

4.5 Discussion

Through the simulation study on the contrast of MCMC and INLA, the performance of INLA on retrospective projection is approximately equivalent to, or even more significant than that of MCMC, with less MAE, less mean CRPS and higher p-value of calibration test in most of immigration groups for each gender and type of cancer. Firstly, as two methods expound significant projection performance (p-values >0.05) on both genders and most of immigration groups, the reason can be the merits of them. The flexibility of MCMC and INLA allows sampling from complex probability distributions by constructing Markov chains or Gaussian Markov random fields, handling a wide range of models.

Secondly, most of 10-year retrospective projections based on INLA indicate less dispersions with observations than those based on MCMC and outperform in most of types of cancer and immigration groups. It consistent with the superiority of INLA of its higher computational efficiency and accuracy. As it approximates the posterior distribution into an analytical form of distribution and avoids the large number of Monte Carlo sampling required in traditional Bayesian inference, the posterior distribution is more accurately approximated, so more accurate inference can be obtained. With lower computation consumptions, the accuracy of INLA also outperforms MCMC for large-scale data analysis and specific models. The choice of posterior distributions and convergence issues from MCMC can result in higher dispersions of projections than those obtained based on INLA.

In general, MCMC is more universal for all types of data and can theoretically work on all models. However, it requires researchers to be very proficient in MCMC theorems and hyperparameters. It’s crucial to select and design appropriate sampler, since the effectiveness of convergence would be extremely low with some complex models, such as spatial models, and extremely huge data size. Therefore, MCMC can only be adopted theoretically in large-scale data analysis. Even though INLA can only be applied for Latent Gaussian models, Latent Gaussian models also cover many

aspects of models and have been considered as vast quantities of inference techniques, such as spatial models, time models and spatial-time models, as substitutes of MCMC. The efficiency of computation and inference accuracy of INLA could also be more satisfactory than those of MCMC. However, some weaknesses and unexpected performance on projection based on INLA have also been revealed.

On the other hand, a minority of projections have also presented significant dispersions with the observed mortality rate, and retrospective projections based on INLA have also indicate higher dispersions with observations than those based on MCMC for some specific circumstances. The potential causes could be the complexity on parameter selection and requirements for data size for INLA. There are many parameters which need to be selected in INLA, including improper selection of grid size, number of nodes, etc. They may lead to inaccurate results, and INLA has high requirements for data size. If the sample size is too small, the inference results may be unreliable. They are comparable with the outcome that most of insignificant projections and higher dispersions from INLA occur in short-stay immigration group which sample size is relatively less than the other two groups.

5. Conclusion and Future Research

Epidemiological studies have shed light on the risk factors associated with the outcomes of infectious and chronic diseases, such as severe COVID-19 infections and cancer mortality. Advanced age and the presence of underlying health conditions, such as immigration status and coordination of other diseases, have been consistently identified as significant risk factors for severe illness and mortality. We have also recognized the disproportionate impact of COVID-19 on marginalized communities, different mobile index and contact settings, who often face social and economic disparities that contribute to increased vulnerability.

To estimate the spread of COVID-19 and inform public health responses, we have employed Bayesian models based on MCMC that incorporate various factors, such as population demographics, mobility patterns, and contact settings. These models have helped guide decision-making regarding the implementation of non-pharmaceutical interventions like lockdowns, travel restrictions, and school closures. Furthermore, epidemiological modeling has played a critical role in predicting healthcare system capacity requirements and evaluating the potential impact of vaccination campaigns. The modeling technique proved the flexibility of MCMC for short-term infectious data. In conclusion, the early COVID-19 epidemics in Japan demonstrated a significant potential of superspreading. Particularly, the school, health care facility and community had relatively higher potential of superspreading when compared to other contact settings. The different potential of superspreading in contact settings highlights the need to continuously monitoring the transmissibility accompanied with the dispersion parameter, to timely identify high risk settings favoring the occurrence of SSE.

INLA is widely used in Bayesian inference problems in various fields. Compared to MCMC, we emphasize the performance of INLA based on a MCMC-free Bayesian APC prediction model to assess the effect of immigration on cancer mortality, as well

as the effects of age, period and cohort, for long-term chronic disease data. We conclude that immigrants, especially short-stay immigrants, had more pronounced fluctuations of mortality rates by age and of relative risks by cohort and period effects for six types of cancers than those of long-stay immigrants and locals. Male immigrants who have stayed in Hong Kong for ≤10 years with colon cancer and male immigrants who have stayed in Hong Kong for >10 years with pancreatic cancer would perform significant uptrend in the future, while other immigration groups for each type of cancer would continue to decline or remain relatively stable. Immigrants for each gender in Hong Kong would suffer from higher mortality risks of cancers than locals in the future.

Through the simulation study on the contrast of MCMC and INLA, the performance of INLA on retrospective projection is approximately equivalent to, or even more robust than that of MCMC. Generally, MCMC is more universal for all types of data and can theoretically work on all models. However, it requires researchers to be very proficient in MCMC theorems and hyperparameters. It's crucial to select and design appropriate sampler, since the effectiveness of convergence would be extremely low with some complex models, such as spatial models, and extremely huge data size. Therefore, MCMC can only be adopted theoretically in large-scale data analysis. Even though INLA can only be applied for Latent Gaussian models, Latent Gaussian models also cover many aspects of models and have been considered as vast quantities of inference techniques, such as spatial models, time models and spatial-time models, as substitutes of MCMC. The efficiency of computation and inference accuracy of INLA could also be more satisfactory than those of MCMC. However, some weaknesses and unexpected performance on projection based on INLA have also been revealed. Higher dispersions and worse performance of INLA in some special cases are consistent with the conclusion of lack of samples.

To estimate the spread of COVID-19 and inform public health responses, we have employed Bayesian models based on MCMC with contact tracing data in Japan. Compared to MCMC, we emphasize the performance of INLA based on a MCMC-free

Bayesian APC prediction model to assess the effect of immigration on cancer mortality with death registry data in Hong Kong. The research findings and conclusions can be extended to other countries and regions with similar methods. What we may concentrate more on could be certain circumstances, and we would adjust our models of different population. For instance, we are able to observe the superspreading and heterogeneity of infectious diseases in China, and effect of immigration on cancer mortality in the US after rearranging datasets and adjusting assumptions of models for different populations. Meanwhile, it would be also logical that discussions of different culture, demography and history are essential, since we may figure out different results between Japan, China, Hong Kong and the US.

The development of the Bayesian method in recent years is inseparable from the contribution of modern computer techniques, as the computing power have been more and more powerful. The Monte Carlo method and INLA have presented in majority of software and program libraries, especially Stan. The latest version of Stan has been already expressive with large-scale parallelism and computation based on MCMC, which is very helpful to promote Bayesian theory and applications [121]. Therefore, more researches of Bayesian inference on Stan could be potential appealing issue to obtain more robust and accurate analysis in the future.

Issues of measurement errors and missing data have been arousing our interests on the development of some frameworks of probabilistic programming, such as Stan, and the application of them for Bayesian inference. As what have been demonstrated by the authors of WinBUGS [13], Bayesian methods are eligible and universal in dealing with issues and data from uncertain sources. Particularly, some popular multiple interpolation methods for missing data have been developed within Bayesian paradigm, and they can be regarded as approximation for full Bayesian analysis. Therefore, the improvement of accuracy and speed of Bayesian inference, such as more accurate interpolation methods for missing data within Bayesian paradigm, could be prospective in future work.

References

[1]. Box GE, Tiao GC. Bayesian inference in statistical analysis. John Wiley & Sons; 2011 Jan 25.

[2]. Dempster AP. A generalization of Bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological). 1968 Jul;30(2):205-32.

[3]. Friston KJ, Glaser DE, Henson RN, Kiebel S, Phillips C, Ashburner J. Classical and Bayesian inference in neuroimaging: applications. Neuroimage. 2002 Jun 1;16(2):484-512.

[4]. Boussinesq M, Gardon J, Kamgno J, Pion SD, Gardon-Wendel N, Chippaux JP. Relationships between the prevalence and intensity of Loa loa infection in the Central province of Cameroon. Annals of Tropical Medicine & Parasitology. 2001 Jul 1;95(5):495-507.

[5]. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. Journal of statistical software. 2017;76.

[6]. Bonat WH, Ribeiro Jr PJ. Practical likelihood analysis for spatial generalized linear mixed models. Environmetrics. 2016 Mar;27(2):83-9.

[7]. Christensen OF. Monte Carlo maximum likelihood in model-based geostatistics. Journal of computational and graphical statistics. 2004 Sep 1;13(3):702-18.

[8]. Christensen OF, Roberts GO, Sköld M. Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. Journal of Computational and Graphical Statistics. 2006 Mar 1;15(1):1-7.

[9]. Cressie, Noel A. C. 1993. Statistics for Spatial Data. Revised. London: John Wiley; Sons Inc.

[10]. Diggle PJ, Giorgi E. Model-based geostatistics for prevalence mapping in low-resource settings. Journal of the American Statistical Association. 2016 Jul 2;111(515):1096-120.

[11]. Diggle P, Moyeed R, Rowlingson B, Thomson M. Childhood malaria in the Gambia: a case-study in model-based geostatistics. Journal of the Royal Statistical Society Series C: Applied Statistics. 2002 Oct;51(4):493-506.

[12]. Hengl T, Minasny B, Gould M. A geostatistical analysis of geostatistics. Scientometrics. 2009 Aug;80:491-514.

[13]. Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. Journal of the Royal Statistical Society Series C: Applied Statistics. 1998 Sep;47(3):299-350.

[14]. Diggle PJ, Thomson MC, Christensen OF, Rowlingson B, Obsomer V, Gardon J, Wanji S, Takougang I, Enyong P, Kamgno J, Remme JH. Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty. Annals of Tropical Medicine & Parasitology. 2007 Sep 1;101(6):499-509.

[15]. Gardon J, Gardon-Wendel N, Kamgno J, Chippaux JP, Boussinesq M. Serious reactions after mass treatment of onchocerciasis with ivermectin in an area endemic for Loa loa infection. The Lancet. 1997 Jul 5;350(9070):18-22.

[16]. Krige DG. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy. 1951 Dec 1;52(6):119-39.
[17]. Lindgren F, Rue H, Lindström J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2011 Sep;73(4):423-98.
[18]. Ribeiro Jr PJ, Diggle PJ, Ribeiro Jr MP, Suggests MA. The geoR package. R news. 2007 Oct 4;1(2):14-8.
[19]. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2009 Apr;71(2):319-92.
[20]. Schlüter, D.K., Ndeffo-Mbah, M.L., Takougang, I., Ukety, T., Wanji, S., Galvani, A.P. and Diggle, P.J., 2016. Using community-level prevalence of Loa loa infection to predict the proportion of highly-infected individuals: statistical modelling to support lymphatic filariasis and onchocerciasis elimination programs. *PLoS neglected tropical diseases*, *10*(12), p.e0005157.
[21]. Takougang I, Meremikwu M, Wandji S, Yenshu EV, Aripko B, Lamlenn SB, Eka BL, Enyong P, Meli J, Kale O, Remme JH. Rapid assessment method for prevalence and intensity of Loa loa infection. Bulletin of the World Health Organization. 2002 Nov;80(11):852-8.
[22]. Zhang H. On estimation and prediction for spatial generalized linear mixed models. Biometrics. 2002 Mar;58(1):129-36.
[23]. Cotter SL, Roberts GO, Stuart AM, White D. MCMC methods for functions: modifying old algorithms to make them faster.
[24]. Chib S. Markov chain Monte Carlo methods: computation and inference. Handbook of econometrics. 2001 Jan 1;5:3569-649.
[25]. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences. 2003 Dec 23;100(26):15324-8.
[26]. Gelman A, Rubin DB. Markov chain Monte Carlo methods in biostatistics. Statistical methods in medical research. 1996 Dec;5(4):339-55.
[27]. Spall JC. Estimation via markov chain monte carlo. IEEE Control Systems Magazine. 2003 Mar 26;23(2):34-45.
[28]. Schrödle B, Held L, Riebler A, Danuser J. Using integrated nested Laplace approximations for the evaluation of veterinary surveillance data from Switzerland: a case-study. Journal of the Royal Statistical Society Series C: Applied Statistics. 2011 Mar;60(2):261-79.
[29]. Beguin J, Martino S, Rue H, Cumming SG. Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. Methods in Ecology and Evolution. 2012 Oct;3(5):921-9.
[30]. Gómez-Rubio V, Bivand RS, Rue H. Estimating spatial econometrics models with integrated nested Laplace approximation. Mathematics. 2021 Aug

25;9(17):2044.
[31]. Diekmann O, Heesterbeek JA. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. John Wiley & Sons; 2000 Apr 7.
[32]. Li MY, Muldowney JS. Global stability for the SEIR model in epidemiology. Mathematical biosciences. 1995 Feb 1;125(2):155-64.
[33]. Hethcote HW. Three basic epidemiological models. InApplied mathematical ecology 1989 (pp. 119-144). Berlin, Heidelberg: Springer Berlin Heidelberg.
[34]. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. International journal of epidemiology. 1999 Oct 1;28(5):964-74.
[35]. Liu WM, Hethcote HW, Levin SA. Dynamical behavior of epidemiological models with nonlinear incidence rates. Journal of mathematical biology. 1987 Sep;25:359-80.
[36]. Vynnycky, E., & White, R. (2010). An introduction to infectious disease modelling. OUP oxford.
[37]. Chen, P. Z., Koopmans, M., Fisman, D. N., & Gu, F. X. (2021). Understanding why superspreading drives the COVID-19 pandemic but not the H1N1 pandemic. The Lancet Infectious Diseases, 21(9), 1203-1204.
[38]. Frieden, T. R., & Lee, C. T. (2020, June). Identifying and interrupting superspreading events-implications for control of severe acute respiratory syndrome coronavirus 2. Emerging infectious diseases. Retrieved August 1, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7258476/
[39]. Bauch, C. T. (2021). Estimating the COVID-19 R number: a bargain with the devil?. The Lancet Infectious Diseases, 21(2), 151-153.
[40]. Galvani, A. P., & May, R. M. (2005). Dimensions of superspreading. Nature, 438(7066), 293-295.
[41]. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. Nature, 438(7066), 355-359.
[42]. Westbrook L. The dance club scene behind Hong Kong's biggest coronavirus cluster. South China Morning Post 2020. https://www.scmp.com/news/hong-kong/society/article/3111507/dance-niche-hong-kong-social-scene-behind-citys-biggest
[43]. Kim, D., Ali, S. T., Kim, S., Jo, J., Lim, J. S., Lee, S., & Ryu, S. (2022). Estimation of serial interval and reproduction number to quantify the transmissibility of SARS-CoV-2 omicron variant in South Korea. Viruses, 14(3), 533.
[44]. Lewis, D. (2021, February 23). Superspreading drives the COVID pandemic - and could help to tame it. Nature News. Retrieved August 1, 2022, from https://www.nature.com/articles/d41586-021-00460-x#:~:text=23%20February%202021-,Superspreading%20drives%20the%20COVID%20pandemic%20%E2%80%94%20and%20could%20help%20to%20tame,best%20to%20target%20control%20measures.&text=Dyani%20Lewis%20is%20a%

20freelance%20science%20journalist%20in%20Melbourne%2C%20Australia.
[45]. Furuse, Y., Tsuchiya, N., Miyahara, R., Yasuda, I., Sando, E., Ko, Y. K., ... & Oshitani, H. (2022). COVID-19 case-clusters and transmission chains in the communities in Japan. Journal of Infection, 84(2), 248-288.
[46]. Endo, A., Abbott, S., Kucharski, A. J., & Funk, S. (2020). Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome open research, 5.
[47]. Kucharski, A. J., & Althaus, C. L. (2015). The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance, 20(25), 21167.
[48]. Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. International journal of infectious diseases, 93, 201-204.
[49]. Blumberg, S., Funk, S., & Pulliam, J. R. (2014). Detecting differential transmissibilities that affect the size of self-limited outbreaks. PLoS pathogens, 10(10), e1004452.
[50]. Karlis, D., & Xekalaki, E. (2000). A simulation comparison of several procedures for testing the Poisson assumption. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 355-382.
[51]. Adam, D. C., Wu, P., Wong, J. Y., Lau, E. H., Tsang, T. K., Cauchemez, S., ... & Cowling, B. J. (2020). Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine*, *26*(11), 1714-1719.
[52]. Zhang, Y., Li, Y., Wang, L., Li, M., & Zhou, X. (2020). Evaluating transmission heterogeneity and super-spreading event of COVID-19 in a metropolis of China. *International journal of environmental research and public health*, *17*(10), 3705.
[53]. Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., & Huth, A. (2011). Statistical inference for stochastic simulation models–theory and application. *Ecology letters*, *14*(8), 816-827.
[54]. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.
[55]. Galvani, A. P., & May, R. M. (2005). Dimensions of superspreading. *Nature*, *438*(7066), 293-295.
[56]. Endo, A. (2020). Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome open research*, *5*.
[57]. Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven

analysis. *International journal of infectious diseases*, *93*, 201-204.

[58]. Blumberg, S., Funk, S., & Pulliam, J. R. (2014). Detecting differential transmissibilities that affect the size of self-limited outbreaks. *PLoS pathogens*, *10*(10), e1004452.

[59]. Ko, Y. K., Furuse, Y., Ninomiya, K., Otani, K., Akaba, H., Miyahara, R., ... & Oshitani, H. (2022). Secondary transmission of SARS-CoV-2 during the first two waves in Japan: demographic characteristics and overdispersion. International Journal of Infectious Diseases, 116, 365-373.

[60]. Adam, D. C., Wu, P., Wong, J. Y., Lau, E. H., Tsang, T. K., Cauchemez, S., ... & Cowling, B. J. (2020). Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. Nature Medicine, 26(11), 1714-1719.

[61]. Oshitani, H., & The Expert Members of The National COVID-19 Cluster Taskforce at The Ministry of Health, L. and W. (2020, November 24). Cluster-based approach to coronavirus disease 2019 (COVID-19) response in Japan, from February to April 2020. Japanese Journal of Infectious Diseases. Retrieved August 1, 2022, from https://www.jstage.jst.go.jp/article/yoken/73/6/73_JJID.2020.363/_article

[62]. Cowling, B. J., Ali, S. T., Ng, T. W., Tsang, T. K., Li, J. C., Fong, M. W., ... & Leung, G. M. (2020). Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. The Lancet Public Health, 5(5), e279-e288.

[63]. Majra, D., Benson, J., Pitts, J., & Stebbing, J. (2020, November 25). SARS-COV-2 (COVID-19) superspreader events. Journal of Infection. Retrieved August 1, 2022, from https://www.sciencedirect.com/science/article/pii/S0163445320307179

[64]. Fan S-C. The population projection of Hong Kong. Southeast Asian Journal of Social Science. 1974;2(1/2):105-17.

[65]. Department CaS. Hong Kong Statistics 1947-1967 (Report). https://www.statistics.gov.hk/pub/hist/1961_1970/B10100031967AN67E0100.pdf, Accessed 4th May 2019.

[66]. Department CaS. Demographic Trends in Hong Kong 1981-2011 (Report). http://www.statistics.gov.hk/pub/B1120017032012XXXXB0100.pdf, Accessed 4th May 2019.

[67]. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021 2021/05/01;71(3):209-49. doi: https://doi.org/10.3322/caac.21660.

[68]. Wang XR, Chiu YL, Qiu H, Au JSK, Yu ITS. The roles of smoking and cooking emissions in lung cancer risk among Chinese women in Hong Kong. Annals of Oncology. 2009 2009/04/01/;20(4):746-51. doi: https://doi.org/10.1093/annonc/mdn699.

[69]. Chiu Y-L, Wang X-R, Qiu H, Yu IT-S. Risk factors for lung cancer: a case–control study in Hong Kong women. Cancer Causes & Control. 2010 2010/05/01;21(5):777-85. doi: 10.1007/s10552-010-9506-9.

[70]. Office on S, Health. Publications and Reports of the Surgeon General. Women and Smoking: A Report of the Surgeon General. Atlanta (GA): Centers for Disease Control and Prevention (US); 2001.
[71]. Escobedo LG, Peddicord JP. Smoking prevalence in US birth cohorts: the influence of gender and education. American Journal of Public Health. 1996 1996/02/01;86(2):231-6. doi: 10.2105/AJPH.86.2.231.
[72]. Husten CG, Shelton DM, Chrismon JH, Lin YC, Mowery P, Powell FA. Cigarette smoking and smoking cessation among older adults: United States, 1965-94. Tobacco Control. 1997;6(3):175. doi: 10.1136/tc.6.3.175.
[73]. Bolego C, Poli A, Paoletti R. Smoking and gender. Cardiovascular Research. 2002;53(3):568-76. doi: 10.1016/S0008-6363(01)00520-X.
[74]. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits; a preliminary report. Br Med J. 1954;1(4877):1451-5. PMID: 13160495. doi: 10.1136/bmj.1.4877.1451.
[75]. Ramada Rodilla JM, Calvo Cerrada B, Serra Pujadas C, Delclos GL, Benavides FG. Fiber burden and asbestos-related diseases: an umbrella review. Gaceta Sanitaria. 2021 2021/06/11/. doi: https://doi.org/10.1016/j.gaceta.2021.04.001.
[76]. Collishaw NE, Kirkbride J, Wigle DT. Tobacco smoke in the workplace: an occupational health hazard. Can Med Assoc J. 1984;131(10):1199-204. PMID: 6498670.
[77]. Dresler CM, Fratelli C, Babb J, Everley L, Evans AA, Clapper ML. Gender differences in genetic susceptibility for lung cancer. Lung Cancer. 2000 2000/12/01/;30(3):153-60. doi: https://doi.org/10.1016/S0169-5002(00)00163-X.
[78]. Alexandrov K, Cascorbi I, Rojas M, Bouvier G, Kriek E, Bartsch H. CYP1A1 and GSTM1 genotypes affect benzo[a]pyrene DNA adducts in smokers' lung: comparison with aromatic/hydrophobic adduct formation. Carcinogenesis. 2002;23(12):1969-77. doi: 10.1093/carcin/23.12.1969.
[79]. Samet JM. Radon and Lung Cancer. JNCI: Journal of the National Cancer Institute. 1989;81(10):745-58. doi: 10.1093/jnci/81.10.745.
[80]. Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, et al. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ. 2005;330(7485):223. doi: 10.1136/bmj.38308.477650.63.
[81]. Raaschou-Nielsen O, Andersen ZJ, Beelen R, Samoli E, Stafoggia M, Weinmayr G, et al. Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). The Lancet Oncology. 2013 2013/08/01/;14(9):813-22. doi: https://doi.org/10.1016/S1470-2045(13)70279-1.
[82]. 2018 Summary-20190719. Retrieved August 26, 2022, from https://www.who.int/docs/default-source/wpro---documents/countries/china/2018-gats-china-factsheet-cn-en.pdf?sfvrsn=3f4e2da9_2
[83]. Thematic household survey.Retrieved August 26, 2022, from

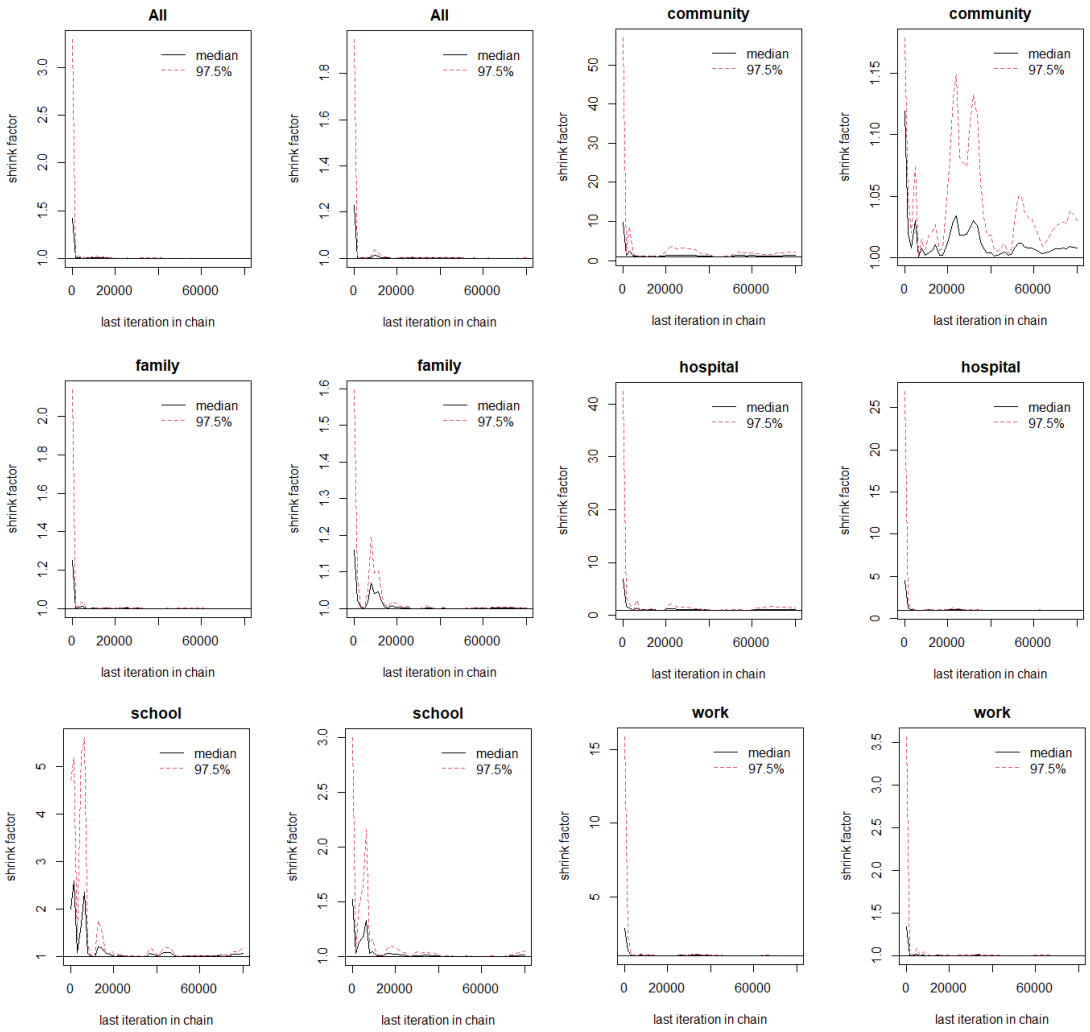
https://www.censtatd.gov.hk/en/data/stat_report/product/B1130201/att/B11302702020XXXXB0100.pdf
[84]. Abubakar II, Tillmann T, Banerjee A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015 Jan 10;385(9963):117-71.
[85]. Estimated alcohol consumption per capita in Hong Kong. Change4Health. (n.d.). Retrieved December 1, 2022, from https://www.change4health.gov.hk/en/alcohol_aware/figures/alcohol_consumption/index.html
[86]. World Health Organization. Global status report on alcohol and health 2018. World Health Organization; 2019 Feb 14.
[87]. Wild C. World cancer report 2014. Wild CP, Stewart BW, editors. Geneva, Switzerland: World Health Organization; 2014.
[88]. Zhao S, Dong H, Qin J, Liu H, Li Y, Chen Y, et al. Breast cancer mortality in Chinese women: does migrant status play a role? Annals of Epidemiology. 2019 2019/12/01/;40:28-34.e2. doi: https://doi.org/10.1016/j.annepidem.2019.10.006.
[89]. Gomez SL, Yang J, Lin S-W, McCusker M, Sandler A, Cheng I, et al. Incidence trends of lung cancer by immigration status among Chinese Americans. Cancer Epidemiol Biomarkers Prev. 2015;24(8):1157-64. PMID: 25990553. doi: 10.1158/1055-9965.EPI-15-0123.
[90]. Hemminki K, Li X, Czene K. Cancer risks in first-generation immigrants to Sweden. International Journal of Cancer. 2002 2002/05/10;99(2):218-28. doi: https://doi.org/10.1002/ijc.10322.
[91]. Vanthomme K, Rosskamp M, De Schutter H, Vandenheede H. Lung cancer incidence differences in migrant men in Belgium, 2004–2013: histology-specific analyses. BMC Cancer. 2021 2021/03/30;21(1):328. doi: 10.1186/s12885-021-08038-6.
[92]. Schooling M, Leung GM, Janus ED, Ho SY, Hedley AJ, Lam TH. Childhood migration and cardiovascular risk. International Journal of Epidemiology. 2004;33(6):1219-26. doi: 10.1093/ije/dyh221.
[93]. Leung JYY, Li AM, Leung GM, Schooling CM. Mode of delivery and childhood hospitalizations for asthma and other wheezing disorders. Clinical & Experimental Allergy. 2015 2015/06/01;45(6):1109-17. doi: https://doi.org/10.1111/cea.12548.
[94]. Baker A, Bray I. Bayesian projections: what are the effects of excluding data from younger age groups?. American Journal of Epidemiology. 2005 Oct 15;162(8):798-805.
[95]. Rosenberg PS, Anderson WF. Age-Period-Cohort Models in Cancer Surveillance Research: Ready for Prime Time? APC Models. Cancer Epidemiology, Biomarkers & Prevention. 2011 Jul 1;20(7):1263-8.
[96]. Holford T. Analyzing the effects of age, period and cohort on incidence and mortality rates. Stat Meth Med Res. 1992;1:317-37.
[97]. Brookmeyer R, Stroup DF, editors. Monitoring the health of populations:

statistical principles and methods for public health surveillance. Oxford University Press; 2004.
[98]. Yang Y, Land KC. Age-period-cohort analysis: New models, methods, and empirical applications. Taylor & Francis; 2013.
[99]. Robertson C, Gandini S, Boyle P. Age-period-cohort models: a comparative study of available methodologies. Journal of clinical epidemiology. 1999 Jun 1;52(6):569-83.
[100]. Riebler A, Held L. Projecting the future burden of cancer: Bayesian age–period–cohort analysis with integrated nested Laplace approximations. Biometrical Journal. 2017 May;59(3):531-49.
[101]. Knoll M, Furkel J, Debus J, Abdollahi A, Karch A, Stock C. An R package for an integrated evaluation of statistical approaches to cancer incidence projection. BMC medical research methodology. 2020 Dec;20(1):1-1.
[102]. Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJ, Lozano R, Inoue M. Age standardization of rates: a new WHO standard. Geneva: World Health Organization. 2001 Jan;9(10):1-4.
[103]. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021 May;71(3):209-49.
[104]. Mousavi SM, Fallah M, Sundquist K, Hemminki K. Age-and time-dependent changes in cancer incidence among immigrants to Sweden: colorectal, lung, breast and prostate cancers. International journal of cancer. 2012 Jul 15;131(2):E122-8.
[105]. National Cancer Institute. SEER stat fact sheets: liver and intrahepatic bile duct cancer.
[106]. Wu X, Chung VC, Hui EP, Ziea ET, Ng BF, Ho RS, Tsoi KK, Wong S, Wu JC. Effectiveness of acupuncture and related therapies for palliative care of cancer: overview of systematic reviews. Scientific reports. 2015 Nov 26;5(1):1-5.
[107]. Sung H, Rosenberg PS, Chen WQ, Hartman M, Lim WY, Chia KS, Wai-Kong Mang O, Tse L, Anderson WF, Yang XR. The impact of breast cancer-specific birth cohort effects among younger and older C hinese populations. International journal of cancer. 2016 Aug 1;139(3):527-34.
[108]. The world economy volume 1: a millennial perspective, 2, Historical statistics: Academic Foundation, Gurgaon, India (2007)
[109]. Elias SG, Peeters PH, Grobbee DE, van Noord PA. The 1944-1945 Dutch famine and subsequent overall cancer incidence. Cancer Epidemiology Biomarkers & Prevention. 2005 Aug;14(8):1981-5.
[110]. Chiu YL, Wang XR, Qiu H, Yu IT. Risk factors for lung cancer: a case–control study in Hong Kong women. Cancer Causes & Control. 2010 May;21(5):777-85.
[111]. Li J, Lam AS, Yau ST, Yiu KK, Tsoi KK. Antihypertensive treatments and risks of lung Cancer: A large population-based cohort study in Hong Kong. BMC cancer. 2021 Dec;21(1):1-9.
[112]. Du J, Sun H, Sun Y, Du J, Cao W, Sun S. Assessment of age, period, and cohort effects of lung cancer incidence in Hong Kong and projection up to 2030 based on

changing demographics. American Journal of Cancer Research. 2021;11(12):5902.
[113]. Centre for Health Protection, Department of Health - Lung Cancer. Centre for Health Protection. Retrieved August 10, 2022, from https://www.chp.gov.hk/en/healthtopics/content/25/49.html
[114] Diouf I, Charles MA, Ducimetière P, Basdevant A, Eschwege E, Heude B. Evolution of obesity prevalence in France. An age-period-cohort analysis. Obésité. 2010 Dec;5:109-16.
[115] Ananth CV, Keyes KM, Hamilton A, Gissler M, Wu C, Liu S, Luque-Fernandez MA, Skjaerven R, Williams MA, Tikkanen M, Cnattingius S. An international contrast of rates of placental abruption: an age-period-cohort analysis. PloS one. 2015 May 27;10(5):e0125246.
[116] Schmid VJ, Held L. Bayesian age-period-cohort modeling and prediction-BAMP. Journal of Statistical Software. 2007 Oct 16;21:1-5.
[117] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association. 2007 Mar 1;102(477):359-78.
[118] Hersbach H. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting. 2000 Oct 1;15(5):559-70.
[119] Grimit, E.P., Gneiting, T., Berrocal, V.J. and Johnson, N.A., 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *132*(621C), pp.2925-2942.
[120] Zamo M, Naveau P. Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. Mathematical Geosciences. 2018 Feb;50(2):209-34.
[121] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of statistical software, 76.

Appendices

A1. Convergence diagnostic



eFigure 1. Gelman–Rubin convergence diagnostic with 100,000 iterations and 20,000 burn-in

A2. Missing data imputation of immigrants population for each year

Objective

Obtain hk population stratified by age (1-100+), gender and birthplace (hk/mainland)

Source

1. https://www.bycensus2016.gov.hk/tc/bc-mt.html?search=A118
2. https://www.censtatd.gov.hk/tc/EIndexbySubject.html?scode=170&pcode=D5211101
3. https://www.censtatd.gov.hk/tc/web_table.html?id=1B#

Method

Step 1

Firstly, we combined data from three sources. For year 2001, 2006 ,2011, 2016, We have complete data for age groups ranged 0-100+ [1,2]. As for other years we used data for the age group ranged 0-85+ [3].

```
rm(list=ls())
sample.year <- c(2001, 2006, 2011, 2016)
all.year <- c(1995:2020)
spar <- 0.125
```

Male group

```
matrix_male <- read.csv("age_matrix_male.csv", header =T)
rownames(matrix_male) <- matrix_male[,1]
matrix_male <- matrix_male[,-1]
matrix_male[,-c(7,12,17,22)] <- matrix_male[,-c(7,12,17,22)] * 1000
colnames(matrix_male) = paste("year.", all.year, sep = "")
rownames(matrix_male) = paste("age.", c(0:100), sep = "")
```

Female group

```
matrix_female <- read.csv("age_matrix_female.csv", header =T)
rownames(matrix_female) <- matrix_female[,1]
matrix_female <- matrix_female[,-1]
matrix_female[,-c(7,12,17,22)] <- matrix_female[,-c(7,12,17,22)] * 1000
colnames(matrix_female) = paste("year.", all.year, sep = "")
rownames(matrix_female) = paste("age.", c(0:100), sep = "")
```

Step 2

Then we use cubic smoothing spline to fill the missing values. (refer to year 2001, 2006 ,2011, 2016)

```
spline.fun <- function(input){
  smooth.func <- smooth.spline(x = sample.year,
                               y = input[c(7,12,17,22)] %>% log(), spar = spar)
  age.pred <- predict(smooth.func, all.year)$y %>% exp() %>% round()
  age.pred[which(age.pred<0)] <- 0
  return(age.pred)
}
matrix_female[86:101,] <- apply(matrix_female[86:101,], 1, spline.fun) %>%
  t() %>% as.data.frame()
matrix_male[86:101,] <- apply(matrix_male[86:101,], 1, spline.fun) %>%
  t() %>% as.data.frame()
```

Step 3

In this step, we are going to extend the data to year 1995-2020 and age ranged 0-100+. This part of code is largely consistent with the original code in the `preprocess` file.

Read in data (population stratified by birthplace, age and gender).

```
## read data
hk.source.data.1996 = read_excel("birth_place_age_gender_1996.xlsx")
hk.source.data.2001 = read_excel("birth_place_age_gender_2001.xlsx")
hk.source.data.2006 = read_excel("birth_place_age_gender_2006.xlsx")
hk.source.data.2011 = read_excel("birth_place_age_gender_2011.xlsx")
hk.source.data.2016 = read_excel("birth_place_age_gender_2016.xlsx")
```

Define the extending function and sample years

```
trans.age.range = function(source.data = the.data, year = year){
  # Female: 4-5
  if(gender=="F"){
    age.info = matrix_female
    this.data = source.data[,c(4:5)] %>% as.data.frame
    #print("F")
  }
  # Male: 2-3
  if(gender=="M"){
    age.info = matrix_male
    this.data = source.data[,c(2:3)] %>% as.data.frame
    #print("M")
  }
```

```
  this.age.info = c(age.info[,(year-1995+1)])
  the.length = dim(source.data)[1]
  ##
  temp.index = NULL
  temp.list = NULL
  temp.sum = NULL
  temp.prop = NULL
  temp.data = NULL
  data.matrix = matrix(data = NA, nrow = length(this.age.info),
                       ncol = dim(this.data)[2])
  for(j in 1:the.length){
    temp.index = c((j*5-4):(j*5))
    if(j == the.length)
      temp.index = c((j*5-4):(101))
    #
    temp.list = this.age.info[temp.index]
    temp.list = c(unlist(temp.list))
    temp.sum = sum(temp.list)
    temp.prop = temp.list/temp.sum

    #
    temp.data = this.data[j,1]*temp.prop
    data.matrix[temp.index,1] = temp.data
    temp.data = this.data[j,2]*temp.prop
    data.matrix[temp.index,2] = temp.data
  }
  colnames(data.matrix) = paste(c("HongKong.", "OtherCN."), year, sep = "")
  rownames(data.matrix) = paste("age.", c(0:100), sep = "")
  return(data.matrix)
}

sample.year <- c(1996, 2001, 2006, 2011, 2016)
```

Extended female group first

```
gender = "F"
```

Extend population to all age groups (0-100+), referring to the age proportion provided by the `matrix_female` data.

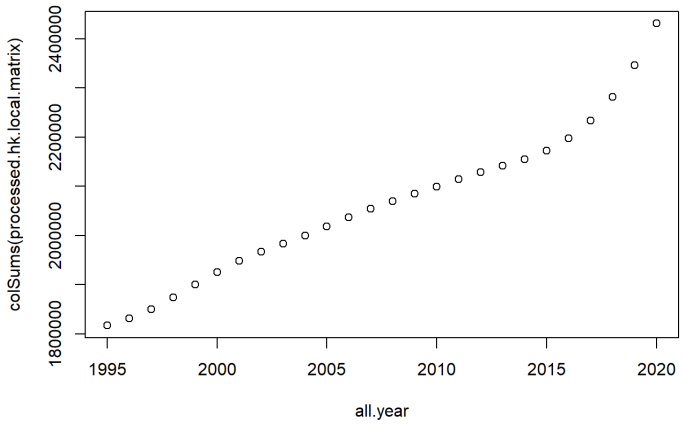
```
## make the matrix
hk.local.matrix = matrix(data = NA, nrow = 101, ncol = length(sample.year))
hk.immigrant.matrix = matrix(data = NA, nrow = 101, ncol = length(sample.year))

# 1996
```

```
temp.matrix = trans.age.range(source.data = hk.source.data.1996,year = 1996)
hk.local.matrix[,1] = temp.matrix[,1]
hk.immigrant.matrix[,1] = temp.matrix[,2]

# 2001
temp.matrix = trans.age.range(source.data = hk.source.data.2001, year = 2001)
hk.local.matrix[,2] = temp.matrix[,1]
hk.immigrant.matrix[,2] = temp.matrix[,2]

# 2006
temp.matrix = trans.age.range(source.data = hk.source.data.2006, year = 2006)
hk.local.matrix[,3] = temp.matrix[,1]
hk.immigrant.matrix[,3] = temp.matrix[,2]

# 2011
temp.matrix = trans.age.range(source.data = hk.source.data.2011, year = 2011)
hk.local.matrix[,4] = temp.matrix[,1]
hk.immigrant.matrix[,4] = temp.matrix[,2]

# 2016
temp.matrix = trans.age.range(source.data = hk.source.data.2016, year = 2016)
hk.local.matrix[,5] = temp.matrix[,1]
hk.immigrant.matrix[,5] = temp.matrix[,2]
```

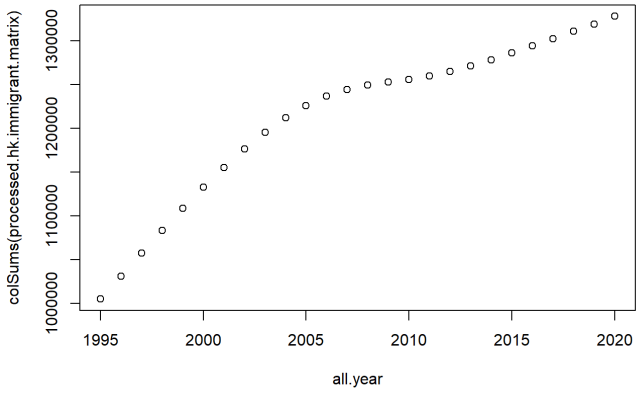
Extend population to all years (1995-2020), using cubic smoothing spline.

```
## start loop for the local
temp.age.data = NULL
processed.hk.local.matrix = matrix(data = NA, nrow = 101, ncol = length(all.year))
#dim(processed.hk.local.matrix)
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(hk.local.matrix[i,c(1:5)])))
  temp.smooth.func = smooth.spline(x = sample.year,
                                   y = temp.age.data %>% log(), spar = spar)
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% exp() %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  processed.hk.local.matrix[i,] = temp.age.pred
}
colnames(processed.hk.local.matrix) = paste0("y", all.year)
rownames(processed.hk.local.matrix) = c(1:101)
processed.hk.local.matrix = as.data.frame(processed.hk.local.matrix)
#  check plot
```

```r
plot(all.year, colSums(processed.hk.local.matrix))
```




```r
## for the other Chinese
#
temp.age.data = NULL
processed.hk.immigrant.matrix = matrix(data = NA, nrow = 101, ncol = length(all.year))
#dim(processed.hk.immigrant.matrix)
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(hk.immigrant.matrix[i,c(1:5)])))
  temp.smooth.func = smooth.spline(x = sample.year, y = temp.age.data, spar = spar)
  #temp.smooth.func$y
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  processed.hk.immigrant.matrix[i,] = temp.age.pred
}
colnames(processed.hk.immigrant.matrix) = paste0("y", all.year)
rownames(processed.hk.immigrant.matrix) = c(1:101)
processed.hk.immigrant.matrix = as.data.frame(processed.hk.immigrant.matrix)
#  check plot
plot(all.year, colSums(processed.hk.immigrant.matrix))
```




```
##
female.hk <- processed.hk.local.matrix
female.cn <- processed.hk.immigrant.matrix
```

Repeated for male group

```
gender = "M"
```

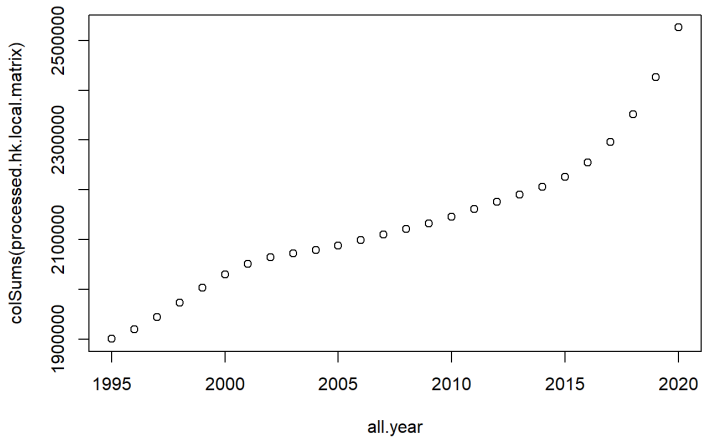
Extend population to all age groups (0-100+)

```
## make the matrix
hk.local.matrix = matrix(data = NA, nrow = 101, ncol = length(sample.year))
hk.immigrant.matrix = matrix(data = NA, nrow = 101, ncol = length(sample.year))

# 1996
temp.matrix = trans.age.range(source.data = hk.source.data.1996,year = 1996)
hk.local.matrix[,1] = temp.matrix[,1]
hk.immigrant.matrix[,1] = temp.matrix[,2]

# 2001
temp.matrix = trans.age.range(source.data = hk.source.data.2001, year = 2001)
hk.local.matrix[,2] = temp.matrix[,1]
hk.immigrant.matrix[,2] = temp.matrix[,2]

# 2006
temp.matrix = trans.age.range(source.data = hk.source.data.2006, year = 2006)
hk.local.matrix[,3] = temp.matrix[,1]
hk.immigrant.matrix[,3] = temp.matrix[,2]

# 2011
temp.matrix = trans.age.range(source.data = hk.source.data.2011, year = 2011)
hk.local.matrix[,4] = temp.matrix[,1]
```

```r
hk.immigrant.matrix[,4] = temp.matrix[,2]


# 2016
temp.matrix = trans.age.range(source.data = hk.source.data.2016, year = 2016)
hk.local.matrix[,5] = temp.matrix[,1]
hk.immigrant.matrix[,5] = temp.matrix[,2]
```

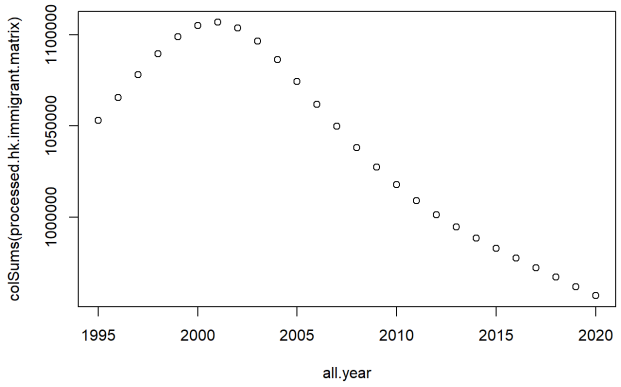
Extend population to all years (1995-2020)

```r
## start loop for the local
temp.age.data = NULL
processed.hk.local.matrix = matrix(data = NA, nrow = 101, ncol = length(all.year))
#dim(processed.hk.local.matrix)
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(hk.local.matrix[i,c(1:5)])))
  temp.smooth.func = smooth.spline(x = sample.year,
                                   y = temp.age.data %>% log(), spar = spar)
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% exp() %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  processed.hk.local.matrix[i,] = temp.age.pred
}
colnames(processed.hk.local.matrix) = paste0("y", all.year)
rownames(processed.hk.local.matrix) = c(1:101)
processed.hk.local.matrix = as.data.frame(processed.hk.local.matrix)
#  check plot
plot(all.year, colSums(processed.hk.local.matrix))
```




```r
##
```

```r
## for the other chinese
#
temp.age.data = NULL
processed.hk.immigrant.matrix = matrix(data = NA, nrow = 101, ncol = length(all.year))
#dim(processed.hk.immigrant.matrix)
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(hk.immigrant.matrix[i,c(1:5)])))
  temp.smooth.func = smooth.spline(x = sample.year, y = temp.age.data, spar = spar)
  #temp.smooth.func$y
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  processed.hk.immigrant.matrix[i,] = temp.age.pred
}
colnames(processed.hk.immigrant.matrix) = paste0("y", all.year)
rownames(processed.hk.immigrant.matrix) = c(1:101)
processed.hk.immigrant.matrix = as.data.frame(processed.hk.immigrant.matrix)
#  check plot
plot(all.year, colSums(processed.hk.immigrant.matrix))
```




```r
##
male.hk <- processed.hk.local.matrix
male.cn <- processed.hk.immigrant.matrix
```

Step 4 create subgroup

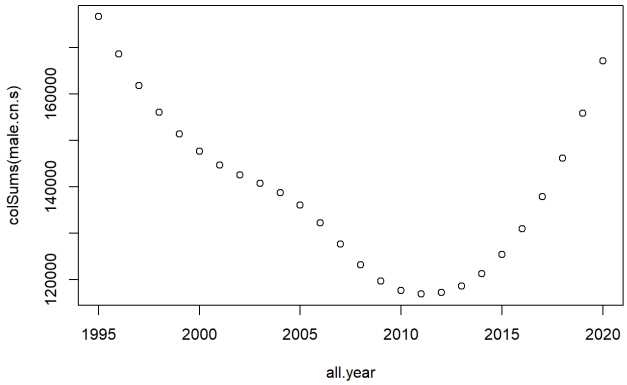
```r
subgroup_ref <- read_xlsx("processed_length_group.xlsx")
subgroup_ref$num[which(subgroup_ref$num=="-")] <- 0
subgroup_ref$num <- as.numeric(subgroup_ref$num)
```

```
for(sex in c("male", "female")){
  for(group in c("s", "l")){
    assign(paste0("cn.",sex,".",group,".matrix"), NULL)
    for(i in seq(2001,2016,5)){
      #ref_set <- get(paste0(sex,".cn"))
      ref_set <- get(paste0("matrix_", sex))

      start <- seq(0,85,5)+1
      end <- c(seq(4,84,5), 100)+1
      duration <- end-start+1
      if(i==2011 & group=="s"){
        duration[1] <- sum(duration[1:3])
        duration <- duration[-c(2:3)]
      }

      data.frame(num=ref_set[,grepl(i,names(ref_set))],
             ind=rep(1:length(duration), times = duration)) %>%
        group_by(ind) %>%
        summarise(n=sum(num)) -> temp
      if(i==2011 & group=="l") temp <- temp[-c(1,2), ]

      subgroup_ref %>%
        filter(gender==sex, grepl(group, length), year==i) -> temp_group

      temp_pop <- temp_group$num/temp$n
      if(i==2011 & group=="l") temp_pop <- c(0,0,temp_pop)
      rep(temp_pop, times = duration) *
        ref_set[,grepl(i, names(ref_set))] -> temp_pop

      assign(paste0("cn.",sex,".",group,".matrix"),
            cbind(get(paste0("cn.",sex,".",group,".matrix")),
                  temp_pop))
    }
  }
}
```

Extend population to all years (1995-2020)

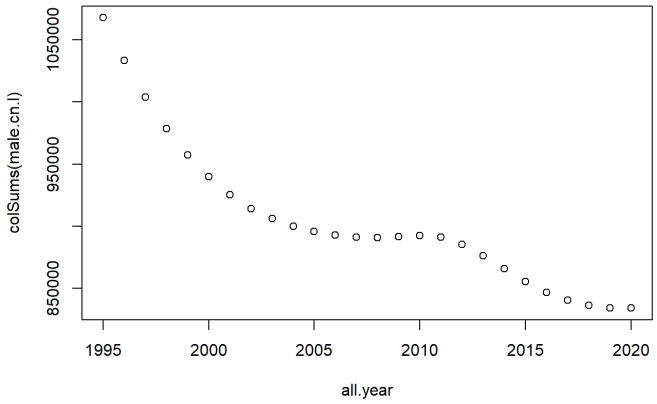
```
sample.year <- seq(2001, 2016, 5)
## start loop for the local
temp.age.data = NULL
male.cn.s = matrix(data = NA, nrow = 101, ncol = length(all.year))
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(cn.male.s.matrix[i,c(1:4)])))
```

```
  temp.smooth.func = smooth.spline(x = sample.year,
                                   y = temp.age.data %>% log1p(), spar = spar)
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% expm1() %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  male.cn.s[i,] = temp.age.pred
}
colnames(male.cn.s) = paste0("y", all.year)
rownames(male.cn.s) = c(1:101)
male.cn.s = as.data.frame(male.cn.s)
#  check plot
plot(all.year, colSums(male.cn.s))
```

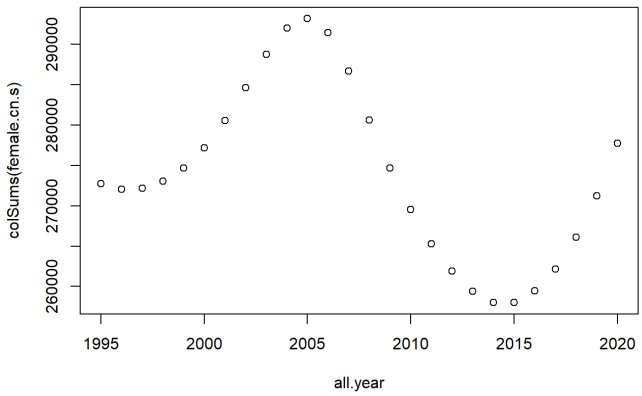



```
##


temp.age.data = NULL
male.cn.l = matrix(data = NA, nrow = 101, ncol = length(all.year))
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(cn.male.l.matrix[i,c(1:4)])))
  temp.smooth.func = smooth.spline(x = sample.year,
                                   y = temp.age.data %>% log1p(), spar = spar)
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% expm1() %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  male.cn.l[i,] = temp.age.pred
}
colnames(male.cn.l) = paste0("y", all.year)
rownames(male.cn.l) = c(1:101)
male.cn.l = as.data.frame(male.cn.l)
#  check plot
```

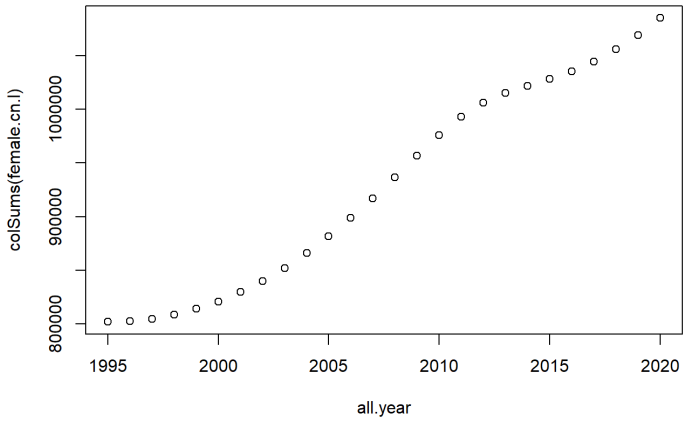
```r
plot(all.year, colSums(male.cn.l))
```




```r
##

### Female

## start loop for the local
temp.age.data = NULL
female.cn.s = matrix(data = NA, nrow = 101, ncol = length(all.year))
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(cn.female.s.matrix[i,c(1:4)])))
  temp.smooth.func = smooth.spline(x = sample.year,
                                   y = temp.age.data %>% log1p(), spar = spar)
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% expm1() %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  female.cn.s[i,] = temp.age.pred
}
colnames(female.cn.s) = paste0("y", all.year)
rownames(female.cn.s) = c(1:101)
female.cn.s = as.data.frame(female.cn.s)
#  check plot
plot(all.year, colSums(female.cn.s))
```

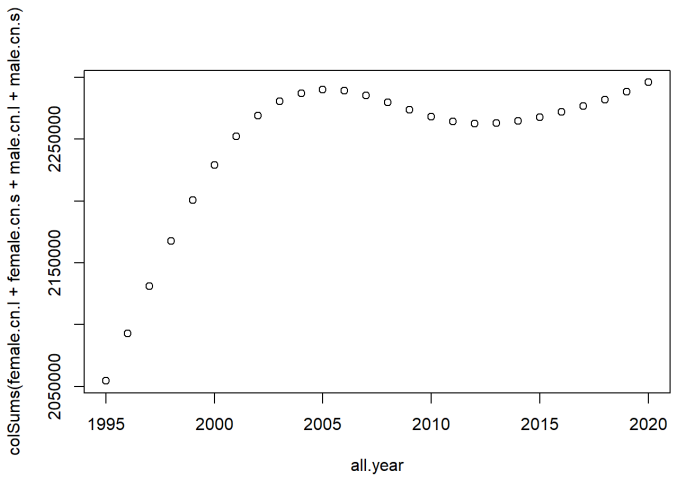



```
##


temp.age.data = NULL
female.cn.l = matrix(data = NA, nrow = 101, ncol = length(all.year))
#     i = 2
for(i in 1:101){
  temp.age.data = c((unlist(cn.female.l.matrix[i,c(1:4)])))
  temp.smooth.func = smooth.spline(x = sample.year,
                                   y = temp.age.data %>% log1p(), spar = spar)
  temp.age.pred = predict(temp.smooth.func, all.year)$y %>% expm1() %>% round()
  temp.age.pred[which(temp.age.pred<0)] <- 0
  #
  female.cn.l[i,] = temp.age.pred
}
colnames(female.cn.l) = paste0("y", all.year)
rownames(female.cn.l) = c(1:101)
female.cn.l = as.data.frame(female.cn.l)
#  check plot
plot(all.year, colSums(female.cn.l))
```

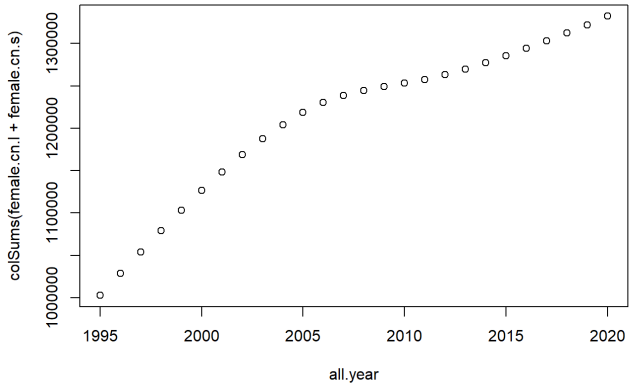


Since we do not have data in 1996, we need to corrected the spline by previous results `male.cn` and `female.cn`.

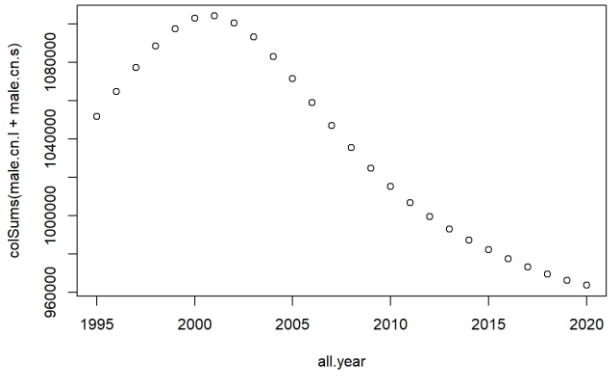
```
male.cn.l <- round(male.cn.l/(male.cn.l+male.cn.s)*male.cn)
female.cn.l <- round(female.cn.l/(female.cn.l+female.cn.s)*female.cn)
male.cn.s <- round(male.cn.s/(male.cn.l+male.cn.s)*male.cn)
female.cn.s <- round(female.cn.s/(female.cn.l+female.cn.s)*female.cn)

plot(all.year, colSums(female.cn.l+female.cn.s+male.cn.l+male.cn.s))
```



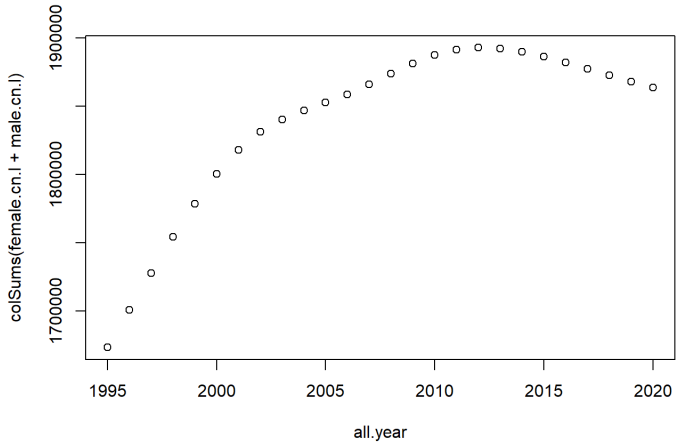

```
plot(all.year, colSums(female.cn.l+female.cn.s))
```



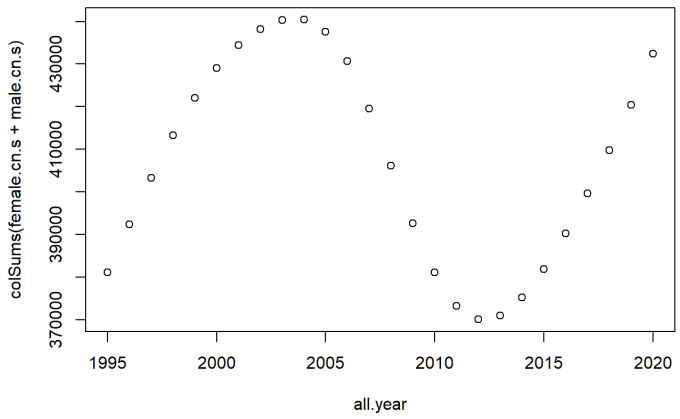

```
plot(all.year, colSums(male.cn.l+male.cn.s))
```



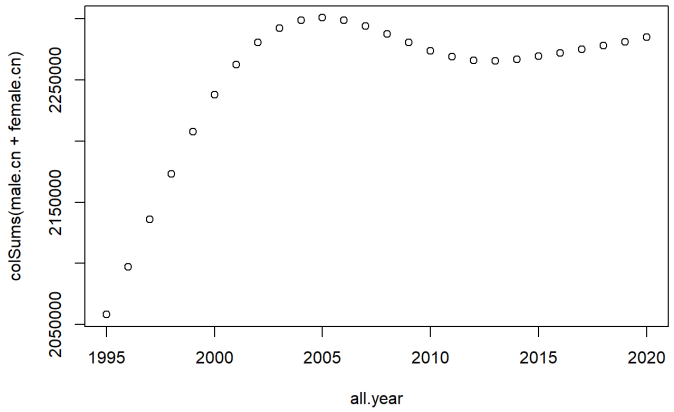

```
plot(all.year, colSums(female.cn.l+male.cn.l))
```



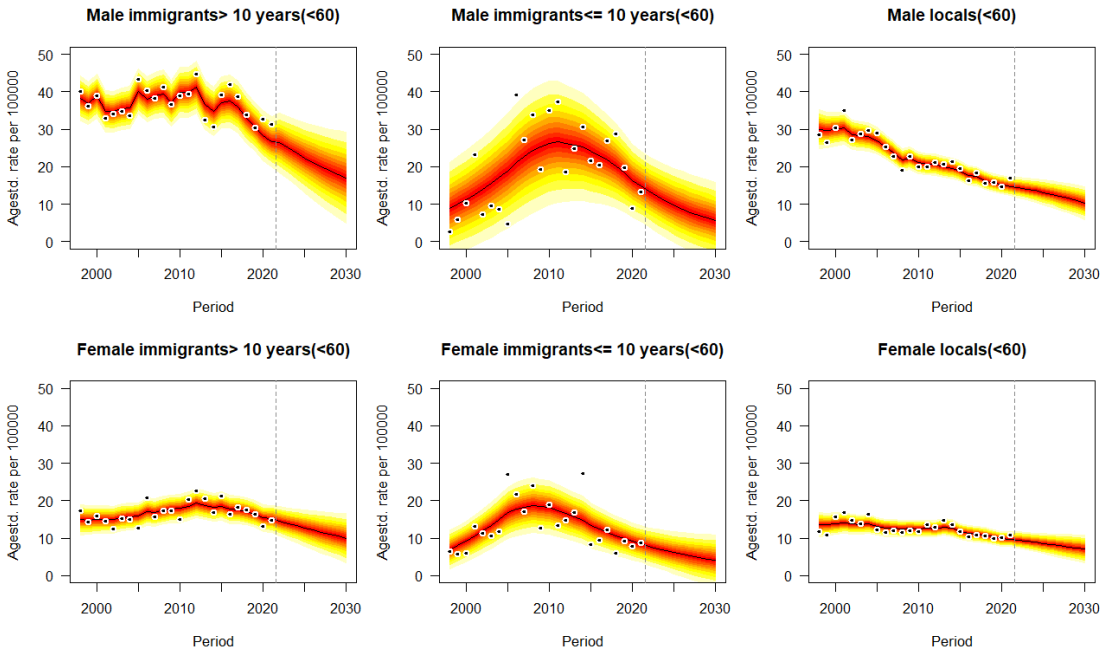

```
plot(all.year, colSums(female.cn.s+male.cn.s))
```



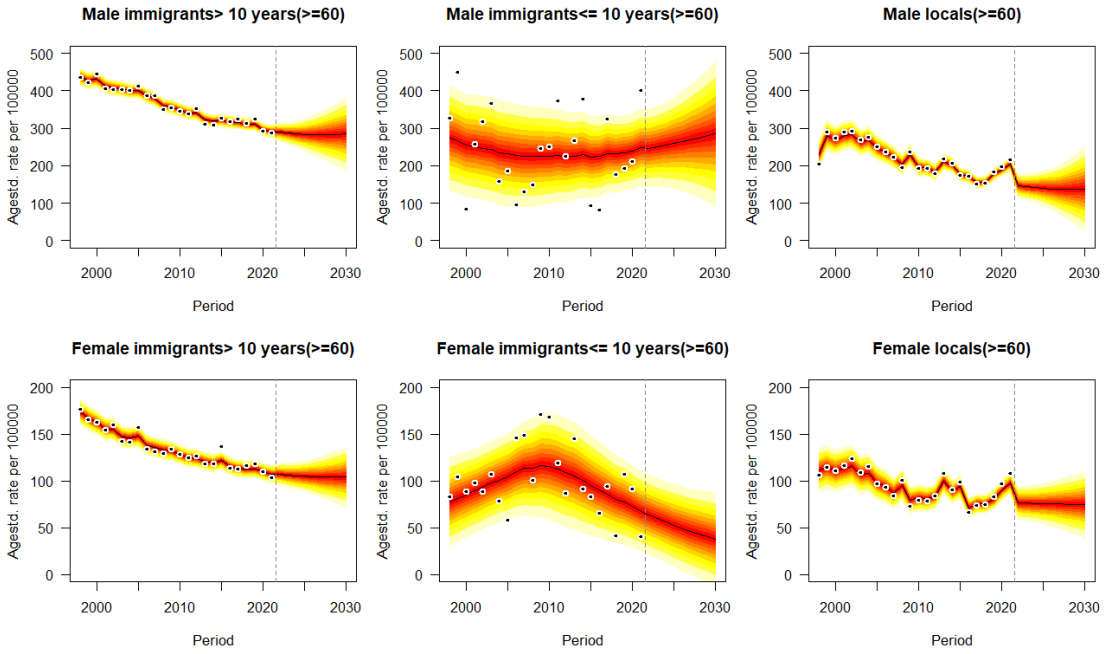

```
plot(all.year, colSums(male.cn+female.cn))
```



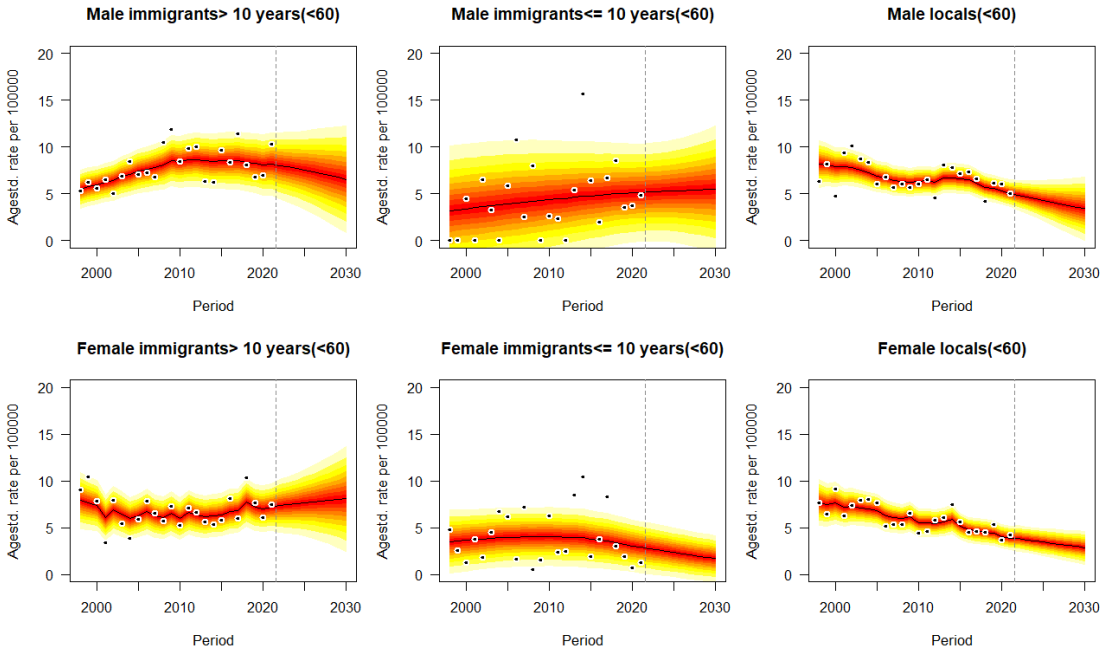
A3. Projections of cancer mortality rates for the population by age strata



eFigure 2(a). Projections of lung cancer mortality rates for the population younger than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

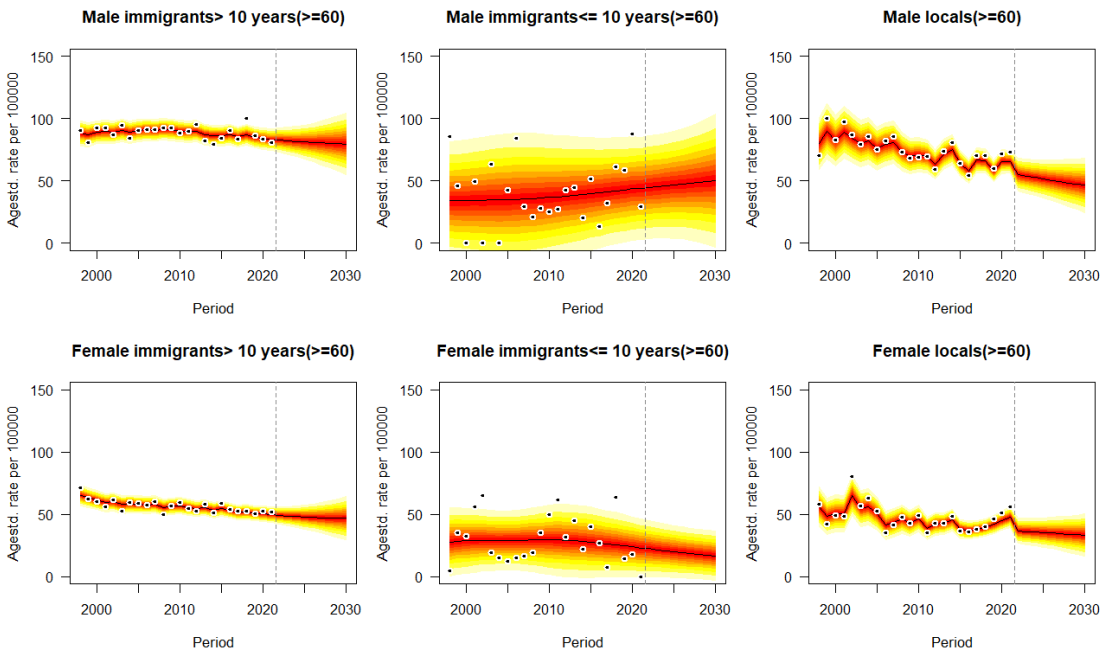


eFigure 2(b). Projections of lung cancer mortality rates for the population older than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

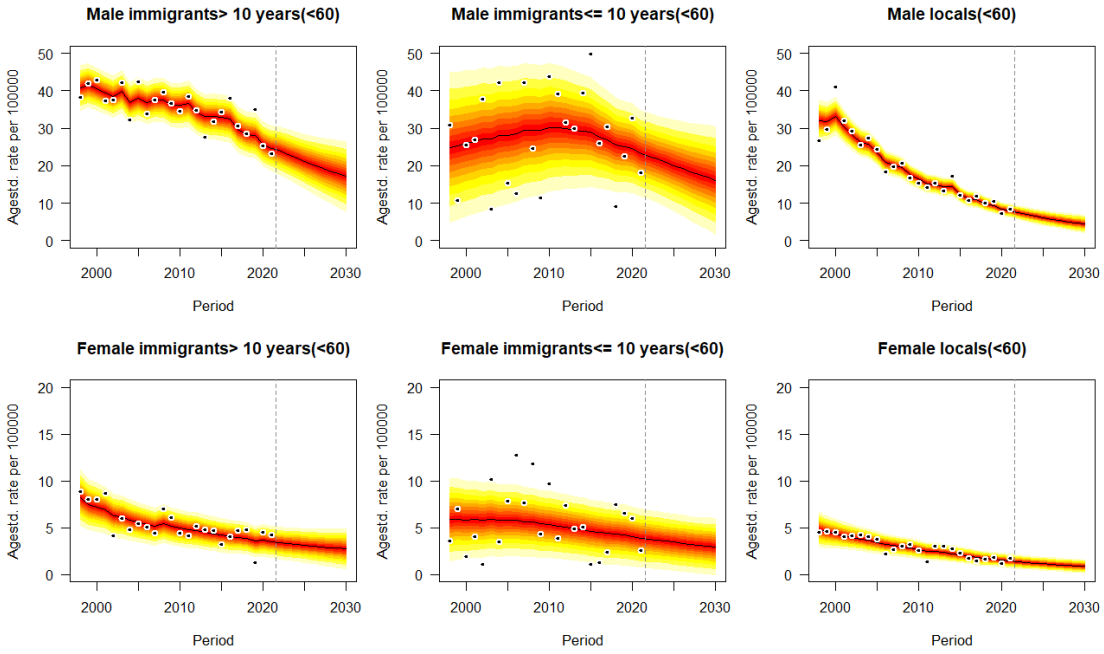


eFigure 3(a). Projections of colon cancer mortality rates for the population younger than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an

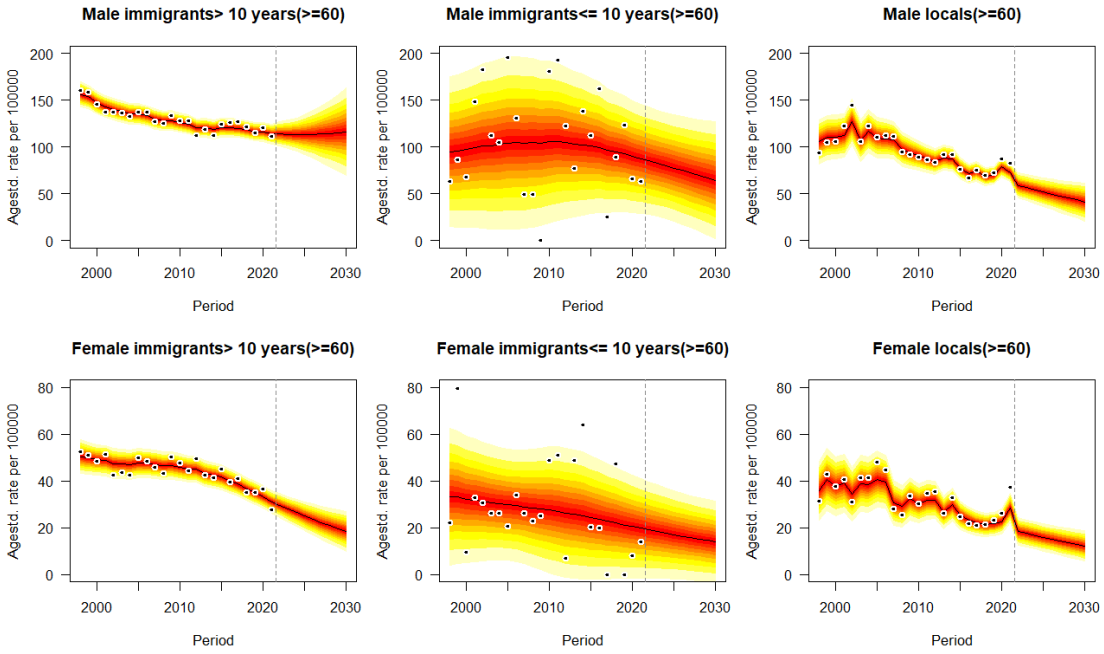
additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.



eFigure 3(b). Projections of colon cancer mortality rates for the population older than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

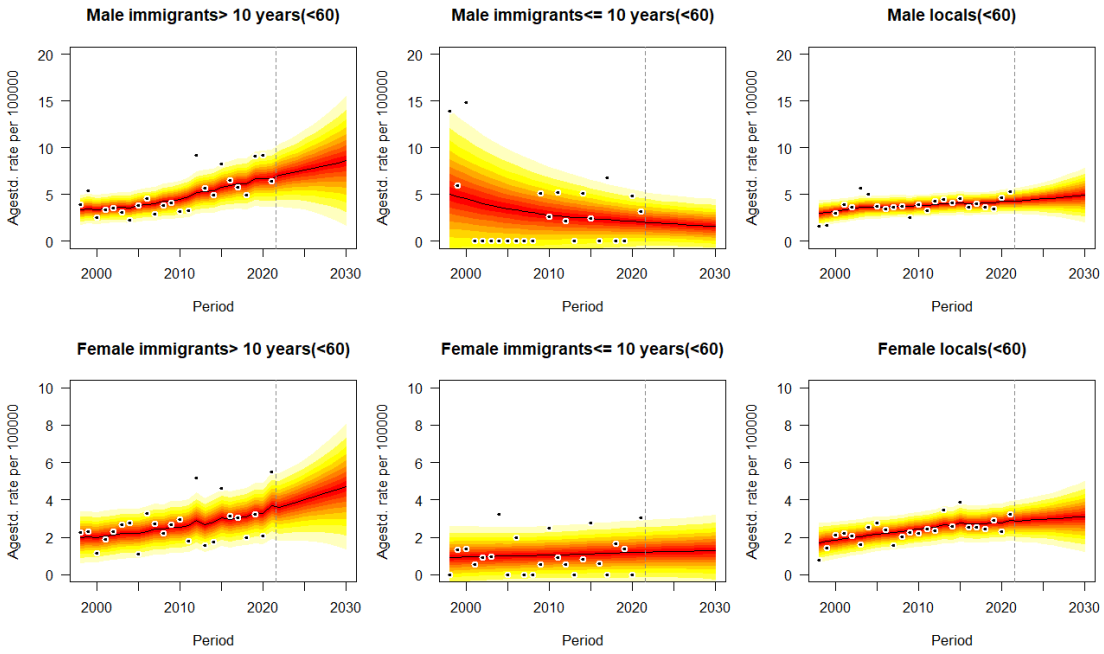


eFigure 4(a). Projections of liver cancer mortality rates for the population younger than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

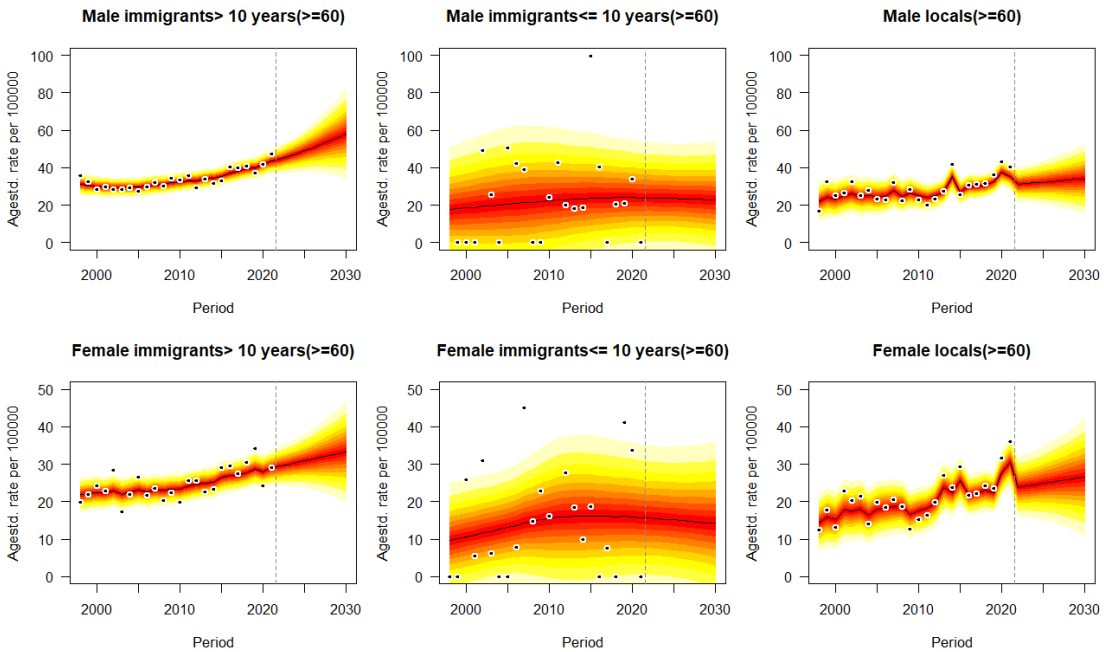


eFigure 4(b). Projections of liver cancer mortality rates for the population older than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10%

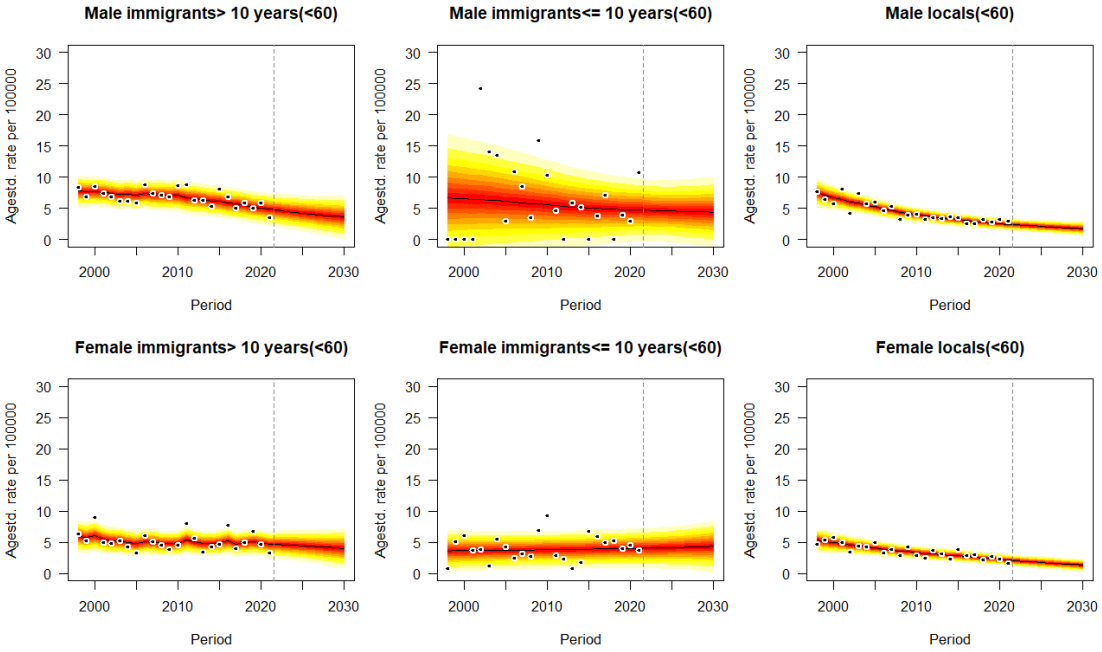
predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.



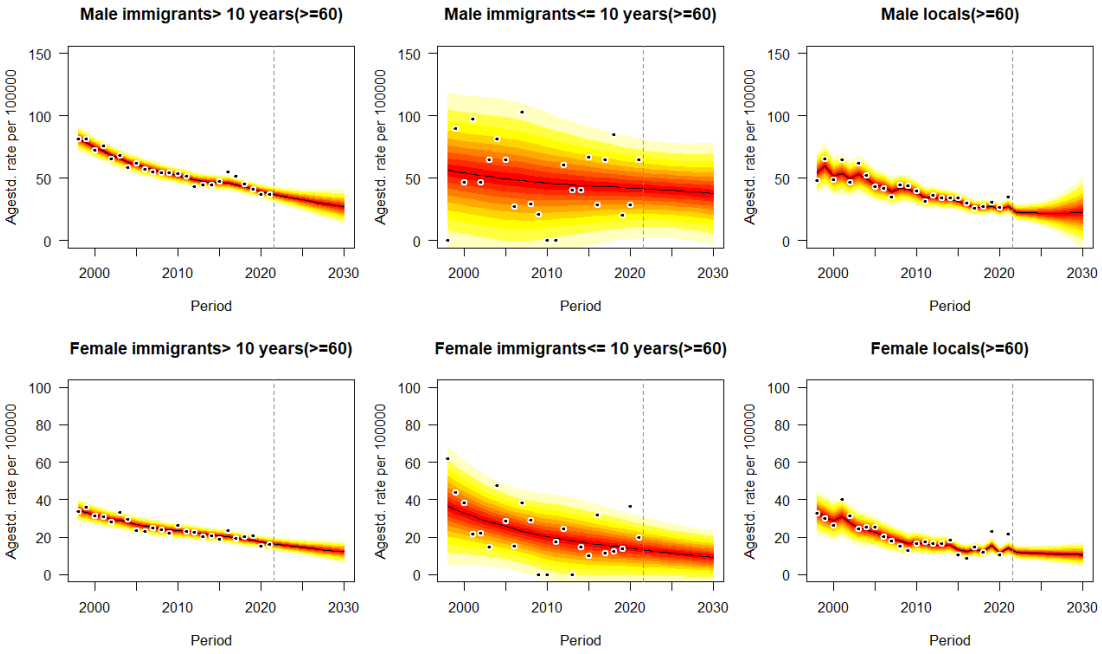
eFigure 5(a). Projections of pancreatic cancer mortality rates for the population younger than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.



eFigure 5(b). Projections of pancreatic cancer mortality rates for the population older than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.



eFigure 6(a). Projections of stoamch cancer mortality rates for the population younger than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.



eFigure 6(b). Projections of stomach cancer mortality rates for the population older than 60 by gender and immigrant status from 2022 to 2030. Observations are shown as dots with the predictive distribution between the 5% and 95% quantile, whereby each lighter shade of red represents an additional 10% predictive CI. The predictive mean is shown as black solid line and the vertical dashed line indicates where prediction started.

A4. Contrast of effective reproduction number and dispersion parameter between INLA and MCMC

	Total	Community	Household	Healthcare facilities	School	Workplace
Reproduction number (*R*) with MCMC	0.561 (0.496, 0.640)	0.107 (0.046, 0.331)	0.137 (0.110, 0.168)	0.186 (0.079, 0.409)	0.088 (0.028, 0.295)	0.080 (0.052, 0.138)
Dispersion parameter (*k*) with MCMC	0.221 (0.186, 0.262)	0.004 (0.002, 0.007)	0.141 (0.098, 0.210)	0.004 (0.002, 0.006)	0.002 (0.001, 0.005)	0.019 (0.012, 0.029)
Reproduction number (*R*) with INLA	5.440 (2.712, 9.343)	3.389 (1.121, 4.897)	0.051 (0.017, 0.278)	1.702 (0.832, 3.880)	0.064 (0.020, 0.337)	0.238 (0.092, 0.671)
Dispersion parameter (*k*) with INLA	0.574 (0.106, 0.753)	0.021 (0.009, 0.033)	0.107 (0.068, 0.334)	0.316 (0.102, 0.948)	0.017 (0.001, 0.083)	0.033 (0.002, 0.071)

eTable 1. Contrast of effective reproduction number and dispersion parameter between INLA and MCMC under different contact settings. The metrics were summarized as 'median estimate (95% CrI)' format.

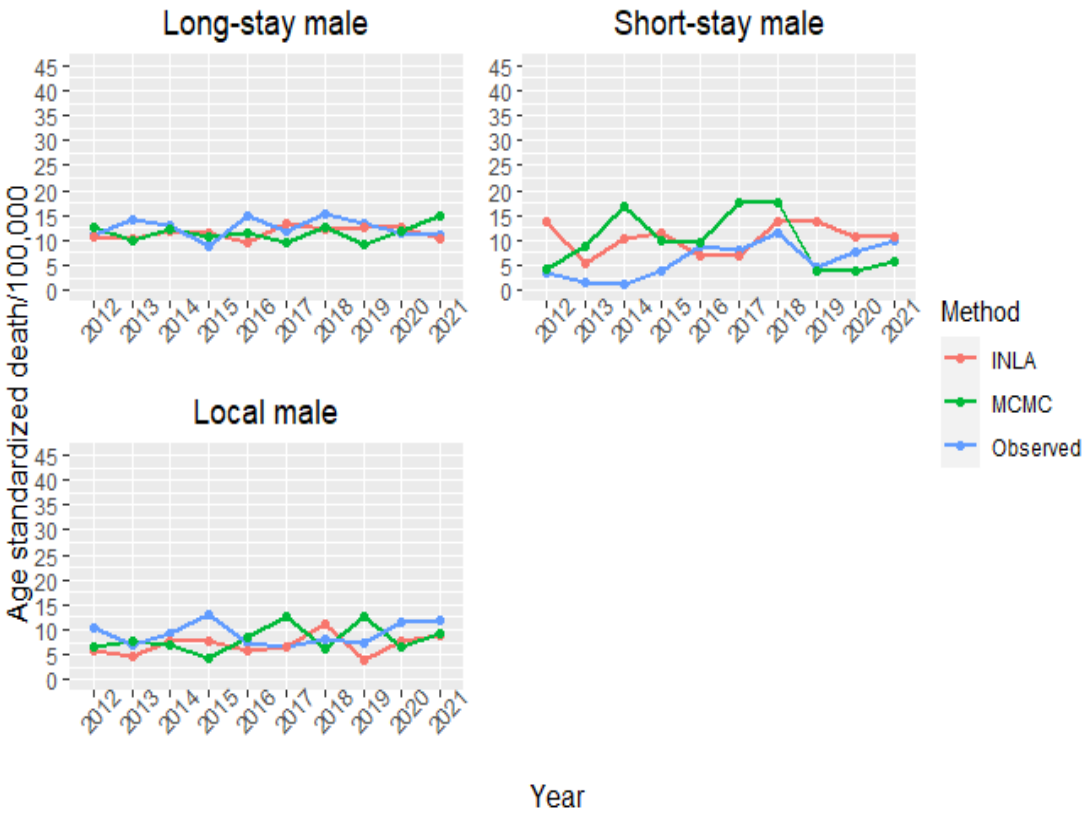
A5. Contrast of retrospective projections of cancer mortality between INLA and MCMC

eFigure 7. Contrast of retrospective projections of colon cancer mortality between INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021

eFigure 8. Contrast of retrospective projections of liver cancer mortality between INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021

eFigure 9. Contrast of retrospective projections of pancreatic cancer mortality between INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021

eFigure 10. Contrast of retrospective projections of stomach cancer mortality between INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021



eFigure 11. Contrast of retrospective projections of prostate cancer mortality between INLA and MCMC as well as observed deaths for immigration groups and genders from 2012 to 2021

Contrast of retrospective projections performance of colon cancer between INLA and MCMC								
Methods / Immigration	INLA				MCMC			
	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$
Female immigrants >10	6.45	5.04	0.33	0.74	8.67	7.88	1.06	0.29
Female immigrants ≤10	11.30	8.22	1.42	0.16	10.26	9.07	1.99	0.04
Female locals	8.05	7.87	0.64	0.52	8.80	8.22	0.68	0.50
Male immigrants >10	9.77	8.91	1.04	0.30	13.73	8.96	1.70	0.09
Male immigrants ≤10	12.51	10.72	-2.50	0.01	21.83	15.48	-2.77	0.01
Male locals	13.67	10.17	1.66	0.10	20.04	18.21	2.33	0.02

eTable 2. Contrast of retrospective projections performance of colon cancer between INLA and MCMC for different immigration groups and genders. Mean absolute error, mean of CRPS, statistic and corresponding p-value are listed.

Contrast of retrospective projections performance of liver cancer between INLA and MCMC								
Methods / Immigration	INLA				MCMC			
	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$
Female immigrants >10	5.77	5.82	0.87	0.38	7.94	7.07	0.92	0.35
Female immigrants ≤10	10.75	8.44	2.14	0.03	13.26	7.56	1.38	0.17
Female locals	8.71	8.11	0.74	0.46	11.18	10.92	1.42	0.16
Male immigrants >10	8.07	7.31	0.88	0.38	7.87	6.21	1.04	0.30
Male immigrants ≤10	12.32	9.68	1.75	0.08	21.45	15.27	1.72	0.08
Male locals	5.33	3.48	0.45	0.65	6.89	6.04	0.88	0.38

eTable 3. Contrast of retrospective projections performance of liver cancer between INLA and MCMC for different immigration groups and genders. Mean absolute error, mean of CRPS, statistic and corresponding p-value are listed.

Contrast of retrospective projections performance of pancreatic cancer between INLA and MCMC								
Methods / Immigration	INLA				MCMC			
	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$
Female immigrants >10	4.05	4.00	0.24	0.81	7.28	6.85	0.59	0.55
Female immigrants ≤10	7.25	7.11	0.58	0.55	9.10	8.42	1.19	0.23
Female locals	5.75	4.25	0.44	0.66	8.73	8.02	1.45	0.15
Male immigrants >10	5.12	3.89	0.38	0.70	10.44	7.68	1.94	0.05
Male immigrants ≤10	7.50	6.35	1.05	0.30	11.27	9.88	1.52	0.14
Male locals	5.93	3.98	-0.45	0.65	7.26	6.04	-0.99	0.32

eTable 4. Contrast of retrospective projections performance of pancreatic cancer between INLA and MCMC for different immigration groups and genders. Mean absolute error, mean of CRPS, statistic and corresponding p-value are listed.

Contrast of retrospective projections performance of stomach cancer between INLA and MCMC								
Methods / Immigration	INLA				MCMC			
	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$
Female immigrants >10	2.91	2.33	0.19	0.85	7.72	5.49	0.77	0.44
Female immigrants ≤10	9.83	6.52	1.47	0.14	9.19	8.03	1.33	0.18
Female locals	5.15	4.67	0.25	0.80	8.45	7.98	0.58	0.56
Male immigrants >10	2.03	0.77	0.16	0.87	3.95	2.67	0.42	0.67
Male immigrants ≤10	12.72	10.38	-2.75	<0.01	12.52	9.37	-2.54	0.01
Male locals	6.93	3.63	0.46	0.65	7.82	6.18	0.87	0.38

eTable 5. Contrast of retrospective projections performance of stomach cancer between INLA and MCMC for different immigration groups and genders. Mean absolute error, mean of CRPS, statistic and corresponding p-value are listed.

Contrast of retrospective projections performance of prostate cancer between INLA and MCMC								
Methods / Immigration	INLA				MCMC			
	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$	$\overline{AE}$	$\overline{CRPS}$	z	$pvalue$
Male immigrants >10	2.59	1.21	0.23	0.80	2.68	2.33	0.18	0.86
Male immigrants ≤10	6.20	4.39	0.89	0.38	6.98	5.01	1.02	0.30
Male locals	5.18	3.92	0.77	0.44	4.90	2.65	0.63	0.52

eTable 6. Contrast of retrospective projections performance of prostate cancer between INLA and MCMC for different male immigration groups. Mean absolute error, mean of CRPS, statistic and corresponding p-value are listed.

List of Publications

1. **Zhao Y**, Zhuang Z, Yang L, He D. Age-period-cohort analysis and projection of cancer mortality in Hong Kong, 1998–2030. **BMJ open**. 2023 Oct 1;13(10):e072751.
2. Lin L, **Zhao Y**, Chen B, He D. Multiple COVID-19 waves and vaccination effectiveness in the United States. International journal of environmental research and public health. 2022 Feb 17;19(4):2282.
3. **Zhao Y**, Zhao S, Guo Z, Yuan Z, Ran J, Wu L, Yu L, Li H, Shi Y, He D. Differences in the superspreading potentials of COVID-19 across contact settings. **BMC infectious diseases**. 2022 Dec 12;22(1):936.
4. Liu Y, Yu Y, **Zhao Y**, He D. Reduction in the infection fatality rate of Omicron variant compared with previous variants in South Africa. **International Journal of Infectious Diseases.** 2022 Jul 1;120:146-9.
5. Chen B, **Zhao Y**, Jin Z, He D, Li H. Twice evasions of Omicron variants explain the temporal patterns in six Asian and Oceanic countries. **BMC Infectious Diseases**. 2023 Jan 13;23(1):25.
6. Zeng T, Lu Y, **Zhao Y**, Guo Z, Sun S, Teng Z, Tian M, Wang J, Li S, Fan X, Wang W. Effectiveness of the booster dose of inactivated COVID-19 vaccine against Omicron BA. 5 infection: a matched cohort study of adult close contacts. **Respiratory Research**. 2023 Oct 12;24(1):246.
7. AVILOV K, WEN L, **Zhao Y**, Wang W, Stone L, He D. The Effectiveness of the COVID-19 Vaccination Campaign in 2021: Inconsistency in Key Studies. **Available at SSRN 4751241**. 2024 Mar 7.
8. Wei H, Musa SS, **Zhao Y**, He D. Modelling of waning of immunity and reinfection induced antibody boosting of SARS-CoV-2 in Manaus, Brazil. **International journal of environmental research and public health**. 2022 Feb 2;19(3):1729.

CURRICULUM VITAE

Academic qualifications of the thesis author, Mr. ZHAO Yanji:

- Received the degree of Bachelor from University of Minnesota, Twin Cities, in December 2017.
- Received the degree of Master from University of Pittsburgh, in December 2020.