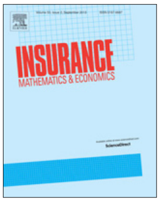
Insurance: Mathematics and Economics 78 (2018) 316–324

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime



Using Taiwan National Health Insurance Database to model cancer incidence and mortality rates[☆]

Jack C. Yue[a,b], Hsin-Chung Wang[c,*], Yin-Yee Leong[a], Wei-Ping Su[a]

[a] *Department of Statistics, National Chengchi University, Taipei, Taiwan, Republic of China*
[b] *Insurance Research Center, National Chengchi University, Taipei, Taiwan, Republic of China*
[c] *Department of Finance and Actuarial Science, Aletheia University, New Taipei City, Taiwan, Republic of China*

ARTICLE INFO

Article history:
Available online 3 October 2017

Keywords:
Cancer insurance
Longevity risk
Big data
Stochastic models
National Health Insurance

ABSTRACT

The increasing cancer incidence and decreasing mortality rates in Taiwan worsened the loss ratio of cancer insurance products and created a financial crisis for insurers. In general, the loss ratio of long-term health products seems to increase with the policy year. In the present study, we used the data from Taiwan National Health Insurance Research Database to evaluate the challenge of designing cancer products. We found that the Lee–Carter and APC models have the smallest estimation errors, and the CBD and Gompertz models are good alternatives to explore the trend of cancer incidence and mortality rates, especially for the elderly people. The loss ratio of Taiwan's cancer products is to grow and this can be deemed as a form of longevity risk. The longevity risk of health products is necessary to face in the future, similar to the annuity products.

1. Preface

People in Taiwan are focusing more on their life planning after retirement because of prolonging life. In addition to economic needs, medical expenses have become a primary focus for the elderly (ages 65 and over) people, mainly due to their high medical demands. On average, the annual medical cost for the elderly is about 5 times that of the ages between 0 and 64 in Taiwan. Moreover, like population aging, age-related death has also become one of the main causes of death in Taiwan. For example, most of the top 10 main causes in 1935 are acute and infectious diseases, while those in 2015 are not (Table 1). Also, the average age at death for the top 10 main causes in 2015 is over 70 and cancer has been the leading cause of death for 34 years in Taiwan, since 1982.

Although only 2% people in Taiwan have cancer, cancer accounts for almost 30% of death. In fact, more elderly people died of cancer, 25% in 2015 compared to 21% in 1996. (Source: Ministry of Health and Welfare) Cancer is also one of the major causes of death in OECD countries and over 1/4 of deaths are cancer related after diseases of the circulatory system in 2013. (Source: OECD Health Statistics, 2015) Cancer is also the second leading cause of death in the U.S. (next to heart disease) and accounts for about 22.5% of death. (Source: National Vital Statistics Reports, 2016)

Curing cancer is expensive and the financial burden of cancer is growing in most of the countries around the globe. In Taiwan, about 6% of national health care expenditure is cancer-related. According to Taiwan's National Health Insurance, the average medical expense per cancer patient is around 3 times the average medical expense per person (Fig. 1). Cancer is also a major cause in the OECD countries and around 5% of the total health cost involves cancer. The direct medical cost (including inpatient stays, outpatient visit, and prescription drugs) for cancer in the U.S. in 2011 was $88.7 billion, which was slightly over 2% of the national health expenditure. (Source: Agency for Healthcare research and Quality)

It is believed that increasing incidence, prolonged survival, and high medical cost are the three main causes of increase in cancer expenditure. The incidence rate of cancer generally increases with age. The financial burden of cancer is likely to increase in Taiwan due to the prolonged life and population aging. According to Taiwan's government agency Financial Supervisory Commission, the loss ratio (i.e., the ratio of claim to premium) of long-term health products seems to increase with the policy year and many insurance companies have 100% or more loss ratios after 15 years. This is especially the case for the cancer products. Table 2 shows the loss ratios of long-term health products in Taiwan over the past three years. (Source: Taiwan Insurance Institute)

Higher loss ratio is not unique in Taiwan and many countries in Asia gained similar experiences. Several reasons can be identified that cause higher than expected loss ratio. First, the economic and insurance markets in Asia developed rapidly since the end of 20th century, but not enough experience rates are available. As a result,

[☆] Paper presented at Longevity 11: The Eleventh International Longevity Risk and Capital Markets Solutions Conference, 7–9 September 2015, Université Lyon 1, Lyon, France.
[*] Corresponding author.
E-mail address: au4369@mail.au.edu.tw (H.-C. Wang).

https://doi.org/10.1016/j.insmatheco.2017.09.016

Table 1
Top 10 main death causes and their % in 1935 and 2015.
Source: Ministry of Interior (Statistical Abstract of Taiwan Province) and Ministry of Health and Welfare

Cause	1	2	3	4	5	6-10
1935	Pneumonia 21.1%	Enteritis 12.2%	Parasites disease 7.1%	Tuberculosis 6.8%	Inherence 6.0%	Others 20.9%
2015	Cancer 29.4%	Heart Diseases 12.0%	Cerebrovascular Diseases 7.0%	Pneumonia 6.8%	Diabetes Mellitus 6.0%	Others 17.8%

Note: The cause "Inherence" includes birth injuries, infections of the newborn, other diseases of infancy, and immaturity (Liu, 2000).

Table 2
The loss ratio of long-term health products in Taiwan (unit: %).

Policy Year	1	2	3	4	5	6	7	8	9	10	11∼14	15+
2013	33	37	37	39	39	48	53	54	52	55	66	76
2014	35	43	44	44	45	45	46	55	57	54	61	79
2015	36	41	45	45	45	46	44	48	53	63	62	76

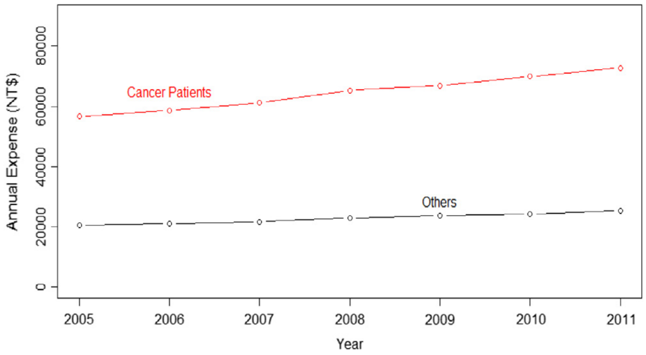


Fig. 1. Annual medical expense of cancer and other patients in Taiwan.

quite a lot of Asian insurance companies use experience rates from the more developed countries (such as Japan). Another reason is that the insurers did not expect longevity risk, like the case of annuity products. Furthermore, the rapid progress in medical technology is also a key factor. People can go to private clinics or health exam centers to check if they have cancers. This creates the possibility of moral hazard. The loss ratios of some one-year term cancer products in Taiwan were 70% or more for the last two years, which is unusual for health insurance products.

The insurers can stop selling, renewing design, or refining the experience rate for the insurance products with high loss ratios. However, this does not work for long-term health insurance products and stopping of selling can only act like a form of damage control. In Taiwan and many Asian countries, whole-life health insurance products are popular and longevity issue needs to be considered to prevent occurrence of financial burden to the insurer in future. Annuity products can be considered to handle the longevity risk, and to this end, a reliable data set and sound stochastic models are required. Taiwan's National Health Insurance (NHI) has been effective since 1995 and cancer is one of the focal diseases. Currently, there are about half million patients recorded with malignant tumors in the NHI database, which have been used in the current study.

In a sense, it seems that the longevity risk exists not only for the annuity products but also for the long-term health insurance products. Thus, predicting future rates of health products are important and we can adopt the experience of mortality rates. Most of the early cancer studies focused on the estimation of incidence and mortality/survival. For example, for the study of breast cancer, Shek and Godolphin (1988) used Cox's proportional hazards model for breast cancer survival; Rosner and Colditz (1996) proposed nonlinear regression methods to model breast cancer incidence; Uddin et al. (2010) utilized the logistic regression method for the incidence of breast cancer. Other models for cancer study include Beta distribution (Pompei and Wilson, 2001), incidence–prevalence–mortality (IPM) model (Kruijshaar et al., 2002), compound-Poisson distribution (Moger et al., 2004), survival model (Hoggart et al., 2012), and functional data analysis (Pokhrel and Tsokos, 2015).

Mortality models became popular and famous mortality models have been applied in cancer studies, especially for the purpose of projection. For example, Arbeev et al. (2005) compared several models for cancer incidence; Robertson and Boyle (1997) and Shibuya et al. (2005) projected the breast/lung cancer mortality by Age–Period–Cohort model; Di Cesare and Murphy (2009) used the Lee–Carter, Booth–Maindonald–Smith, Age–Period–Cohort, and Bayesian models to analyze trends and forecast mortality rates of three major death causes. One of the advantages of using these models is their ease of use in the mortality projection. In the current study, similar to dealing with longevity risk in annuity, we have considered the Generalized Age–Period–Cohort (GAPC) stochastic mortality models (Villegas et al., 2016), which include most of the frequently used mortality models from the past studies, and applied them for projections. Our aim was to verify which GAPC model is suitable for describing the cancer incidence and mortality rates.

The current paper has been arranged as follows. Section 2 gives a brief introduction of Taiwan's National Health Insurance Research Databases (NHIRD) and describes the idea of handling big data, including the exploratory data analysis for cancer incidence and mortality rates. The introduction of proposed models and the evaluation of modeling cancer incidence and mortality rates are given in Section 3. The applications of projecting cancer trend and their consequences with respect to the loss ratio of cancer insurance products are given in Section 4. In the final section, we provided a discussion and suggestions on dealing with the potential risks in cancer insurance.

2. Exploratory data analysis of the big data

Taiwan launched the NHI program on March 1, 1995, and about 99.68% Taiwan residents were enrolled in the program at the end of 2015. The data from the NHI program, including registration files and original claim data for reimbursement, were collected by the NHI Bureau. Based on the principle of privacy protection, personal identification numbers were de-identified by using the scrambling program twice. After the scrambling, the data were sent to the National Health Research Institutes (NHRI) for storage. Scholars from research institutes and universities can apply to the NHRI for the NHI data, by paying a fee based on the size of the data requested.

Table 3
Discrepancy between Databases.

Year	HV		HV_CD	
	#patients	#records	#patients	#visits
2007	1,103,431	1,453,483	649,106	17,946,211
2008	1,164,465	1,529,866	678,544	19,173,919
2009	1,276,315	1,733,251	712,828	20,357,173
2010	1,350,786	1,863,254	746,746	21,619,442
2011	1,401,449	1,933,455	779,179	22,861,178

Cancer is one of the Catastrophic Illnesses (CI) recorded in the NHI database. The CI is one of the key features in Taiwan's NHI, and to ease some financial burden, the government provides some medical privilege and a co-payment waiver to people diagnosed with CI. In 2015, the CI patients were about 4% of Taiwan population (about 0.9 million) but about 27% of total NHI expenditure was spent. The cancer patients spent approximately 35% of the total CI expenditures, which was around 65 billion NT dollars (or 2 billion US dollars).

In this study, we used all the records of the CI patients to avoid the possibility of biased selection for cancer data. The size of the CI related databases is huge, around 228 GB, and the data size of cancer is approximately half of the CI data, about half million Taiwan people with cancer in 2016. Datasets considered in this study include the registry for beneficiaries (ID), registry for CI patients (HV), and CI patients' original claim data extracted from the CD (ambulatory care expenditures by visits) data file (HV_CD). We used database software (SQL) and applied big data techniques for handling data analysis, since it is impossible to use a regular statistical software to perform a data analysis.

Considering the data quality, we used the data only during the period 2002–2011. The size and data quality complicate the analysis of big data, and the NHIRD showed no differences. The data codebook was used to check whether there was any problem in the data content. However, we can still find problems by cross-checking different databases. For example, the numbers of HV patients are different in the databases of HV and HV_CD (Table 3). The numbers of patients in HV_CD are close to (but slightly larger than) the official records from Ministry of Health and Welfare (MHW). It is possible that patients can have two or more than two records in the HV_CD database if they have more than one CIs. On average, about 5% of CI patients have more than one CIs. On the other hand, the numbers of patients in the HV database are too large, and it is likely that the HV records might not have been updated (e.g. CI patients passed away or recovered but are still in the list). We need to remove this kind of content errors before conducting any data analysis.

The process of analyzing big data is tedious and time-consuming. Therefore, we used the example of judging whether a CI patient is alive as a demonstration. There is a data column ("death note") in the HV_CD database showing the status (i.e. alive or dead) of each patient. However, the number of death recorded in this data column is too small. For example, the number of cancer deaths is more than 40,000 annually but only 7,000–8,000 deaths (e.g. 7,517 deaths in 2007) are shown in this column, meaning under-reported deaths from the death note. Therefore, to judge whether a patient is dead or not, certain criteria were needed to be set.

The cancer patients recorded in the HV and HV_CD database are those with malignant tumors between stage 1 and stage 4 who require medical treatments. Thus, we can use the nature of cancer patients' medical visits to judge if they are still alive. We think that, if cancer patients stop visiting doctors, especially after a series of regular treatments, it is likely that health conditions of the patients are worsening, is similar to Lee et al. (2011) research results.

After a few trial-and-errors, we determined to use the condition "no outpatient records for two consecutive years" (Condition 1) for cancer patients to judge if a patient is dead, since the percentage of misjudgment was smaller than 5%. We used the International Classification of Diseases (ICD) codes in the medical records to identify whether cancer patients were receiving medical treatments. In addition, we also included the condition, whether the cancer patients stop visiting doctors suddenly, or more precisely, whether there are 3 or more outpatient visits for the month of last outpatient visit (Condition 2). We used the official records of cancer deaths to verify the above two conditions.

In this part, the cancer incidence rates and cancer mortality rates derived from the analysis of NHIRD have been shown. First, we defined the cancer incidence rate and mortality rate of cancer patients, the same way as that defined by the Ministry of Health and Welfare. Let the age-specific cancer incidence rate of ith age group is I_i/n_i, where i indicates the age groups 0–14, 15–19, 20–24, …, 85–89, I_i is the number of new cancer patients diagnosed in the age group i, n_i is the number of persons who do not have cancer in the age group i in the beginning of the year. The mortality rate of cancer patients for the ith age group is the ratio between the death number of cancer patients and the number of cancer patients in the age group i.

The age-specific incidence cancer rates are shown in Fig. 2. Since there were a large volume of data, we only selected the results of 2002, 2005, 2008, and 2011 for demonstration. As expected, men showed a higher incidence rate than women for all the age groups. To note, the incidence rate did not show big changes from 2002 to 2011 but only showed small increments between two consecutive years. However, still the increments did indicate potential longevity risk. Next, we applied stochastic models to the incidence rates.

Since our goal is to design cancer insurance products, we computed the mortality rates for those who had cancer. These mortality rates were different from the official records where the cancer mortality rate is defined as the number of cancer deaths divided by all Taiwan residents. Moreover, since we used "no outpatient records for two consecutive years" (Condition 1) to determine if a patient is alive, we lost two years of data, i.e. could not determine the status of patients for 2010 and 2011. Thus, following the similar format as that in Fig. 2, Fig. 3 shows the age-specific mortality rates for cancer patients in 2002, 2005, 2008, and 2009. It seems that the mortality rate reduced slowly, and similar to those in Fig. 2, the change in mortality rate between two consecutive years was not much.

Since the incidence rate increased slowly and the mortality rate reduced gradually, it is reasonable to expect that the number of Taiwan cancer patients (or the cancer prevalence rate) will increase. This is exactly the case in Taiwan and the number of Taiwan cancer patients increased about 30% from 2002 to 2011. Together with the population aging in Taiwan, the number of cancer patients will continue to increase in the future. On the other hand, increasing incidence rate and decreasing mortality rate indicated a higher expenditure for cancer insurance, coinciding with the increasing loss ratio for cancer insurance in recent years. In the next section, we will detail the use of stochastic models to explore the possibility of including longevity risk in cancer products.

3. Methodology

Most of the stochastic models considered in the literature of mortality study can be categorized as the family of Generalized Age–Period–Cohort (GAPC) stochastic mortality models (Villegas et al., 2016). For instance, Age–Period–Cohort model (Cairns et al., 2009), the Lee–Carter (LC) model (Lee and Carter, 1992), the Renshaw–Haberman (RH) model (Renshaw and Haberman, 2006),

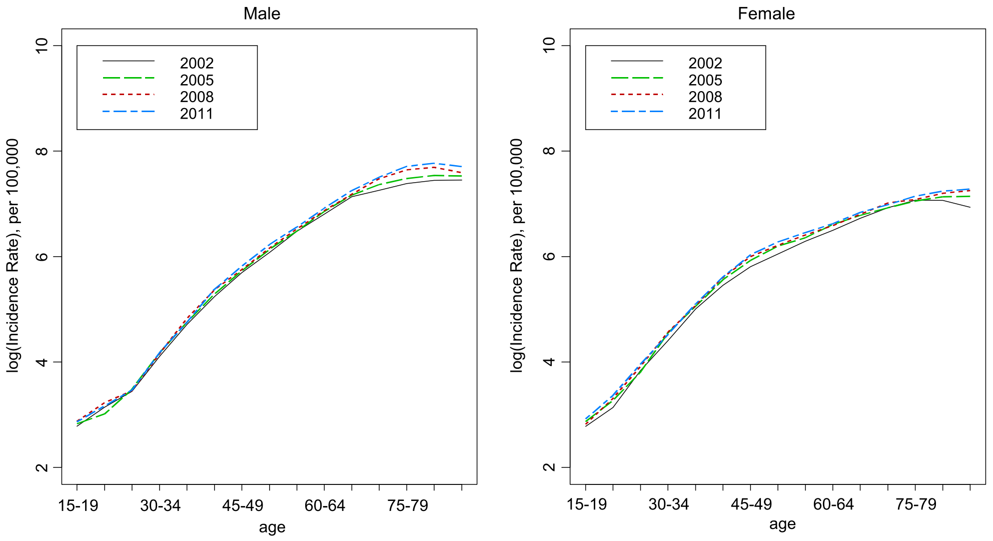


Fig. 2. Taiwan age-specific cancer incidence rates (estimated).

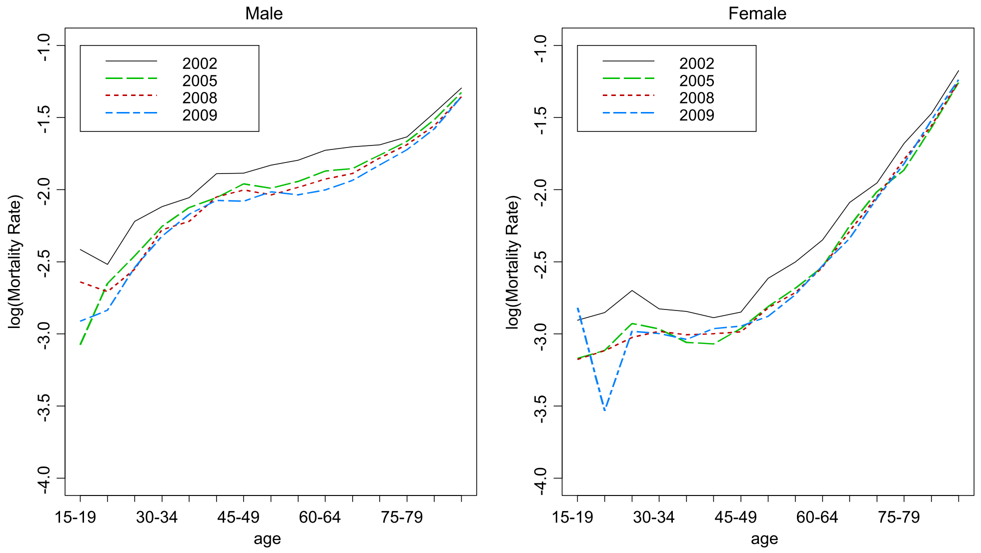


Fig. 3. Taiwan age-specific cancer patients' mortality rates (estimated).

the Cairns–Blake–Dowd (CBD) model (Cairns et al., 2009), and Plat model (Plat, 2009) are some well-known examples.

First, an introduction to the Gompertz model and some of the GAPC stochastic mortality models will be provided. Originally, the Gompertz model was for modeling the mortality rates of higher ages. It is believed that the force of mortality μ_x at age x satisfies

$$\mu_x = BC^x, \tag{1}$$

where $B > 0$ and $C > 1$ are model parameters. Eq. (1) can be converted to the form of central death rate m_x by $m_x = BC^x$. The Gompertz model has been applied to situations such as fertility and morbidity. Additionally, Strehler and Mildvan (1960) used the Gompertz mode to fit cancer mortality rate.

If m_{xt} denote the central death rate or incidence rate for a person aged x at time t. The LC model assumes that

$$\log(m_{xt}) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \varepsilon_{x,t}, \tag{2}$$

with $\sum_x \beta_x^{(2)} = 1$ and $\sum_t \kappa_t^{(2)} = 0$, $\beta_x^{(i)}$ are age related parameters $(i = 1,\ 2)$, and $\kappa_t^{(2)}$ represents the time related parameter. Note that $\beta_x^{(1)}$ is the general mortality level, $\beta_x^{(2)}$ is the decline in mortality at age x, and $\kappa_t^{(2)}$ is usually a linear function in time. The term $\varepsilon_{x,t}$ denotes the deviation of the model and is assumed to be white noise, with 0 mean and relatively small variance.

The residuals after fitting the LC model are often not random (Debón et al., 2008), and adding extra time or cohort component to the LC model is one of the possible modifications. The RH model can be treated as a version of LC model with an extra cohort component,

$$\ln(m_{xt}) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}, \tag{3}$$

where $\sum_x \beta_x^{(2)} = 1,\ \sum_t \kappa_t^{(2)} = 0,\ \sum_x \beta_x^{(3)} = 1,\ \sum_{x,t} \gamma_{t-x}^{(3)} = 0$, and the parameter $\beta_x^{(i)}$ denotes the average age-specific mortality, $\kappa_t^{(2)}$ represents the general mortality level, and $\gamma_{t-x}^{(3)}$ reflects the cohort-related effect.

The CBD model was designed to model mortality rates of higher ages and to deal with the longevity risk in pensions and annuities. For the CBD model, it assumes that the mortality rates satisfy

$$\operatorname{logit}(m_{xt}) = \log\frac{m_{xt}}{1 - m_{xt}} = \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}, \tag{4}$$

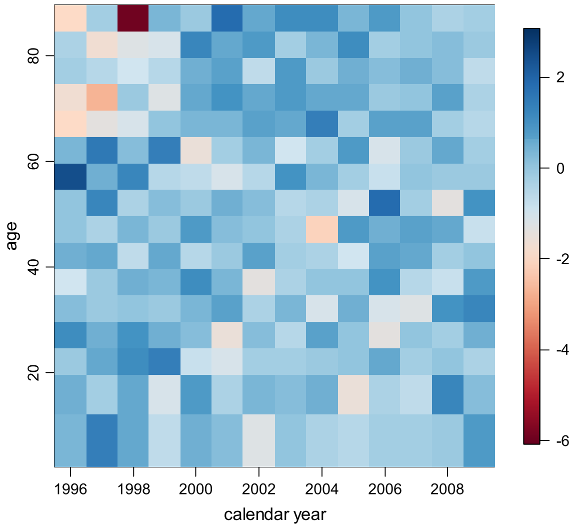


Fig. 4. Residuals of the LC model (Cancer, male).

where the parameters are $\beta_x^{(i)}$ and $\kappa_t^{(i)}$ $(i = 1, 2)$ denote the average age-specific mortality and the general mortality levels. If we assume $\beta_x^{(1)} = 1$ and $\beta_x^{(2)} = x - \bar{x}$, then the model has a simple parametric form:

$$\text{logit}\,(m_{xt}) = \kappa_t^{(1)} + \kappa_t^{(2)}\,(x - \bar{x})\,. \tag{5}$$

The Age–Period–Cohort (APC) model is a popular tool for modeling disease incidence and mortality in epidemiology. Heuristically speaking, if we consider the notion of Analysis of Variance, the LC model considers the effects of Age and Age×Period (Interaction), while the APC model considers three main effects, Age, Period, and Cohort:

$$\ln\,(m_{xt}) = \alpha_x + \kappa_t + \gamma_{t-x}, \tag{6}$$

where $\sum_{c=t-x}\gamma_c = 0,\ \sum_c c\gamma_c = 0\,.$

Three criteria are used to evaluate the models: mean absolute percentage error (MAPE), Akaike Information Criteria (AIC), and Bayesian Information Criterion (BIC). The MAPE is defined as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\varepsilon_i|}{Y_i} \times 100\%,$$

where Y_i and ε_i are the observed value and residual of observation $i, i = 1, 2, \ldots, n$. The AIC and BIC are defined as

$$AIC = -2\log(L) + 2k,$$
$$BIC = -2\log(L) + k\log(n)\,,$$

where L is the likelihood of the data, k is the number of parameters in the model, and n is the number of observations. A model with smaller AIC or BIC value is treated as a better model.

For the following discussion, we have chosen four GAPC mortality models (LC, APC, RH, and CBD) and used the R package StMoMo to explore these models. There were two reasons for choosing these four GAPC models. First, the LC model generally has the best model fit for all ages among the GAPC family models. The RH and CBD models provide fine modifications to the LC model under certain conditions. The RH model is more appropriate with obvious cohort effects, while the CBD model is a widespread alternative to LC model for the elderly people. On the other hand, the APC model is a popular choice in modeling disease incidence and mortality, and it is easy to provide interpretation.

Our previous studies showed that the other GAPC family models did not produce good fitting results. They either gave much larger fitting/prediction errors or showed no convergence during estimation. The goodness-of-fit of stochastic models can be verified by inspecting the residuals, e.g. as the form of heat map (Fig. 4). If there are system errors in the residuals, then spatial patterns (such as hot spots) can be detected. Debón et al. (2008) found that there is spatial autocorrelation for using the LC model to fit mortality rates. However, in the present study, we did not find clusters or autocorrelation for using the LC model to fit the cancer data shown in Fig. 4. The cohort effect was treated as a cluster of 45° and this indicated that the RH model, i.e. a cohort modification of LC model, might have larger fitting errors for modeling cancer incidence and mortality rates than using the LC model.

We evaluated the Gompertz model and the above-mentioned four mortality models from the GAPC models, using the R package StMoMo, based on the cancer incidence and mortality rates. The age-specific rates were in the format of 5-age group, except for the younger groups since their cancer incidence rate and the number of cancer patients for younger groups were very small. Therefore, the data were divided into ages 0–14, 15–19, 20–24,$\cdots$, and 85–89. In addition, we also considered the model evaluation for the case of higher ages, viz. ages 50–54, 55–59, …, and 85–89.

For the Gompertz model, we used the weighted least squares to obtain the parameter estimates of B and C. As for the GAPC family models, we have only shown the results under the log-Poisson assumption because the estimation results were about the same as those of the log-Poisson and logit-Binomial assumptions. We do not recommend the estimation under the normal assumption since it would produce unstable results when the population sizes are small (populations smaller than 50,000, according to our experience).

Next, we compared the model fits of Gompertz's model and the GAPC family models. Table 4 shows the MAPEs for fitting the incidence rates and mortality rates of cancer patients. For both males and females, the LC and APC models showed the smallest fitting errors. The MAPEs of LC and APC models were very small, indicating that the trend of incidence and mortality rates were well captured. For higher ages, i.e. ages 50–89, the Gompertz and CBD models turned to be good alternatives to the LC and APC models, especially for the mortality rates.

We also considered the criteria of AIC and BIC to avoid overparameterization in the models. Since the LC, APC, and CBD models had the smallest MAPEs, we have only shown the AIC and BIC values for these three models in Table 5, where smaller AIC and BIC values are preferred. Since the number of parameters were about the same for these three models, the results of AIC and BIC were similar to those of MAPE. In general, the LC and APC models are preferred and the CBD model is also a possible choice for higher ages.

We also examined the differences between observed and estimated rates (or residuals) to double check the estimation results. Fig. 5 shows the observed cancer incidence rates and the estimated rates for four GAPC family models, for the cases of Taiwan male in 2002, 2005, 2008, and 2011. Apparently, the fitting curves of LC and APC models almost overlapped with those of the observed values. The RH and CBD models seemed to produce overbiased estimates for the younger ages and sometimes overbiased estimates for the elderly as well. The results for mortality rates estimation were similar and thus the details are not shown.

It seems that both the LC model and APC model are good candidates to model the cancer incidence and mortality rates. We should focus our discussions on the LC model, since we only need to consider the time-related parameter $\kappa_t^{(2)}$ of LC model for prediction. Note that $\kappa_t^{(2)}$ is often a linear function of time and we can derive rough estimates of annual increment/reduction

Table 4
MAPEs of cancer patients mortality and incidence.

			Gompertz	LC	APC	RH	CBD
Incidence Rates	ages 0–89	Male	47.8	2.4	2.8	45.4	50.1
		Female	69.8	2.4	2.9	10.3	63.1
	ages 50–89	Male	16.1	1.6	2.0	55.8	15.6
		Female	7.6	1.4	1.9	31.0	7.3
Mortality Rates	ages 0–89	Male	11.3	4.3	4.3	62.0	11.4
		Female	21.5	6.5	6.6	108.6	21.0
	ages 50–89	Male	5.2	2.7	2.0	25.3	5.2
		Female	4.4	4.2	2.8	---	4.0

Note: "---" means that the model fitting didn't converge.

Table 5
AIC and BIC for LC, APC and CBD models.

			LC		APC		CBD	
			AIC	BIC	AIC	BIC	AIC	BIC
Incidence Rates	Ages 0–89	Male	1,705	1,828	1,778	1,926	33,071	33,132
		Female	1,659	1,782	1,744	1,892	33,771	33,833
	Ages 50–89	Male	954	1,011	1,012	1,088	8,191	8,239
		Female	892	949	954	1,030	2,111	2,159
Mortality Rates	Ages 0–89	Male	2,146	2,296	2,157	2,348	3,219	3,315
		Female	2,154	2,304	2,120	2,311	5,147	5,243
	Ages 50–89	Male	1,246	1,322	1,219	1,328	1,747	1,823
		Female	1,245	1,321	1,180	1,289	1,289	1,365

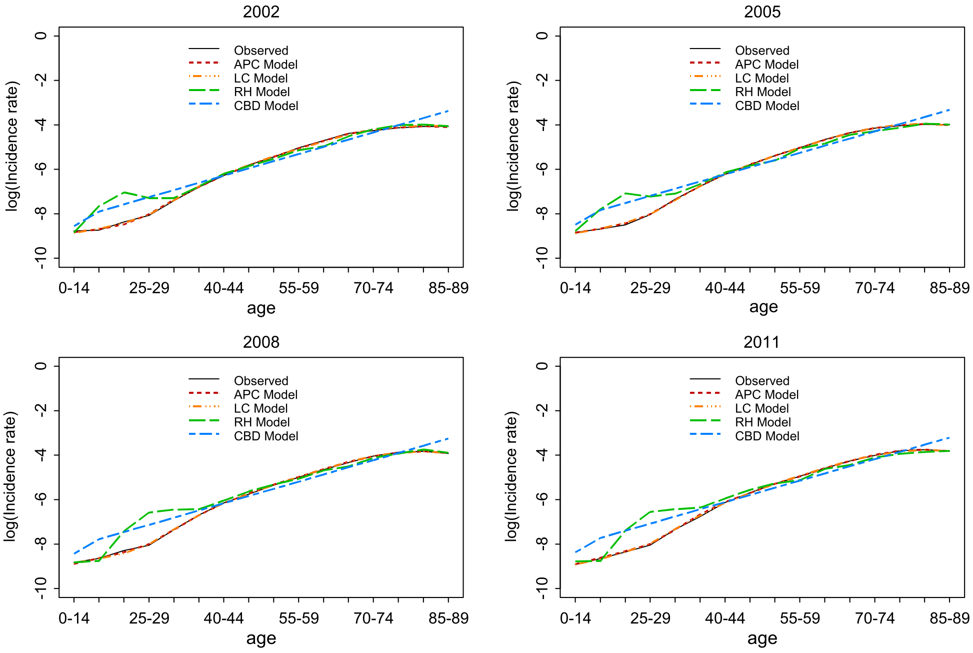


Fig. 5. Observed vs. estimated cancer incidence rates (male).

in incidence/mortality rates. If $\kappa_t^{(2)} = a + bt$ then the annual increment/reduction rate at age x is $\beta_x^{(2)} \times b$. Fig. 6 shows the annual increment of cancer incidence (left panel) and the annual reduction of cancer mortality (right panel). Older ages showed the largest increment in cancer incidence and younger ages displayed the largest reduction in cancer mortality.

The changes in the incidence rates were positive which indicated an increasing trend, while the changes in the mortality rates were negative, thereby indicating a decreasing trend. There are double risks of insurers in cancer product. The first is that the increasing cancer incidence rate should enhance the benefit paid when the insured is diagnosed with cancer. The second risk

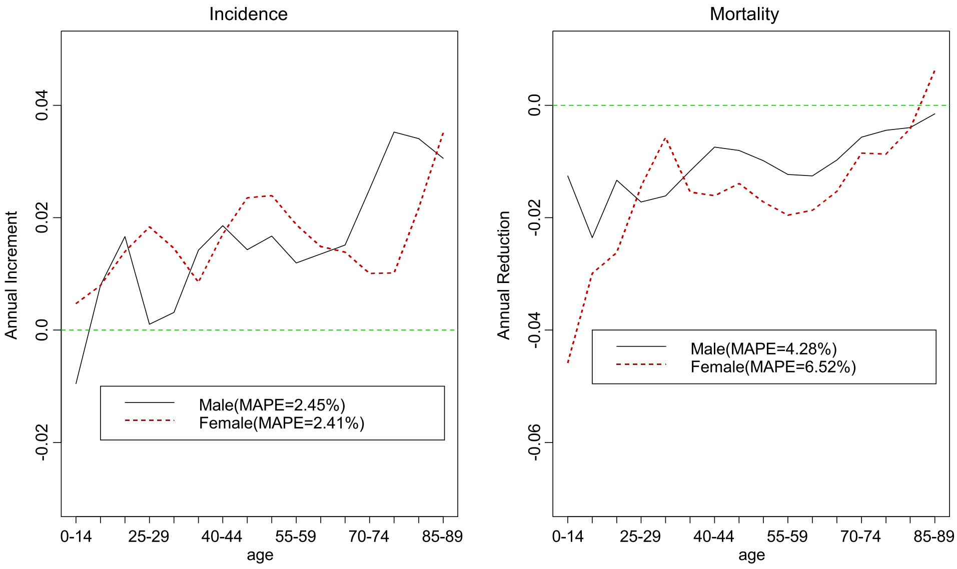


Fig. 6. Annual increment/reduction of cancer data (LC model).

is that improvement in cancer patients' mortality rates would add medical benefits burden and annuity, which is similar to the longevity risk for annuity products. Note that, the annual reductions and increments are not a smooth function of age (due to smaller population sizes) and in practice, we can use graduation methods to obtain smoother values (e.g. Wang et al., 2016a).

4. Applications

We evaluated the impact of increasing cancer incidence rates and decreasing mortality rates of cancer patients via the pricing of cancer products. In Taiwan, if the insured is diagnosed with cancer, cancer insurance products usually include clinical visit benefit and the death benefit. Thus, the increasing incidence rate and the decreasing mortality rate of cancer patients are expected to worsen the loss ratio of cancer products. For simplicity, we considered two types of whole-life cancer products: the first is that a (lump-sum) benefit is paid when the insured was diagnosed with cancer for the first time; the other product is that the annual annuity benefit will be paid after the insured is diagnosed with cancer and still alive.

The first type of cancer product is similar to the whole-life insurance products and the contract is terminated after the benefit is paid. We used the years 2002–2009 as the base-line years and compared the differences in pure premium whether the increasing cancer incidence rates were considered or not. The insured ages for the cancer insurance products were 30–60, with a premium payment period of 20 years. Also, the highest attained age was 110 years old and the interest rate was 2.25%. Since the results for different base-line years are similar, we only show the case of 2009 as a demonstration. Table 6 shows the pure premiums per $1,000 for the whole-life cancer products, where "cohort" and "period" indicate the cases with/without the increment in cancer incidence, respectively.

Fig. 7 shows the percentages of pure premium undercounts if the increment of cancer incidence rate is not considered. The differences in pure premium were very significant, especially for the male and the younger ages, for all base-line years. This can be used as an evidence to explain why the loss ratios of cancer insurance products in Taiwan is increasing in recent years. Also, the differences in pure premium seem to be a decreasing function of time for both genders and the differences in 2002 are the largest.

Table 6
Pure premium of 20-year payment whole-life cancer (per $1,000, interest rate: 2.25%).

Age	Male		Female	
	Period	Cohort	Period	Cohort
30	10.58	17.66	9.72	14.53
35	11.88	19.04	10.71	15.22
40	13.27	20.47	11.68	15.80
45	14.74	21.92	12.49	16.17
50	16.31	23.44	13.15	16.30
55	18.03	25.08	13.65	16.33
60	19.98	27.03	14.12	16.33

Table 7
Pure premium of 20-year payment whole-life annuity cancer (per $1000, interest rate: 2.25%).

Age	Male		Female	
	Period	Cohort	Period	Cohort
30	1,048	1,074	1,222	1,320
35	1,042	1,065	1,192	1,278
40	1,038	1,059	1,157	1,225
45	1,034	1,052	1,119	1,168
50	1,029	1,041	1,082	1,113
55	1,022	1,030	1,050	1,067
60	1,016	1,020	1,026	1,034

The second type of cancer product is similar to the whole-life annuity product and the annuity is paid while the insured is alive, after diagnosed with cancer. Again, the year 2009 was used as the base-line year to evaluate the differences of pure premium if the reduction of cancer patient mortality rates is considered. The settings of age, interest rate, and highest attained age were the same as the first type. Table 7 shows the detailed numbers of pure premiums per $1,000 for the whole-life annuity products. Fig. 8 shows the percentage of pure premium undercount when the reduction of cancer mortality was not considered. Unlike in Fig. 7, the differences in pure premium with respect to mortality reduction were not very significant and the cases of female and younger ages were the largest.

5. Discussion and conclusions

The loss ratios of cancer insurance products in Taiwan worsened in recent years and the longevity risk seemed to exist in long-term health insurance products as well. Using the databases from

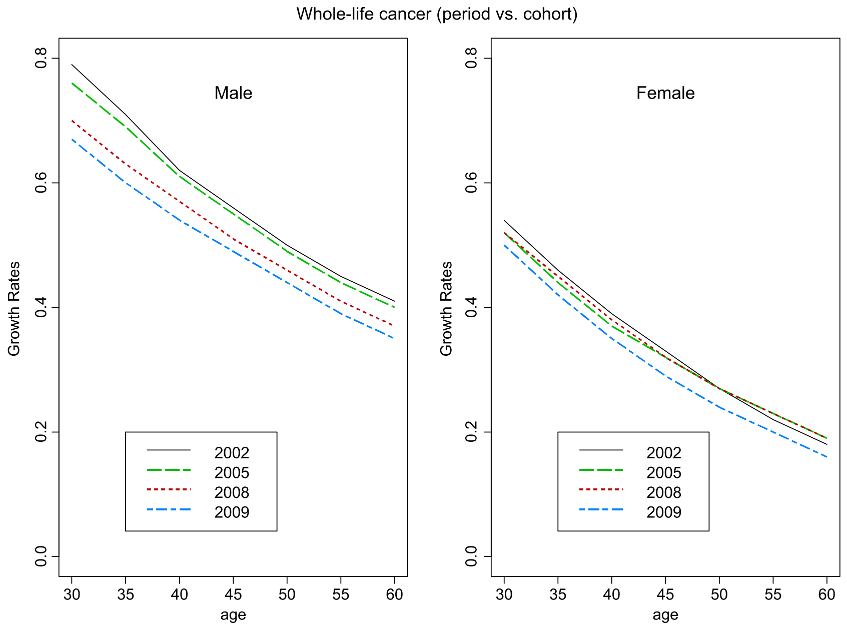


Fig. 7. % difference in pure premium whole-life cancer (LC model). Note: Cohort effect means cancer growth rate of incidence is considered.

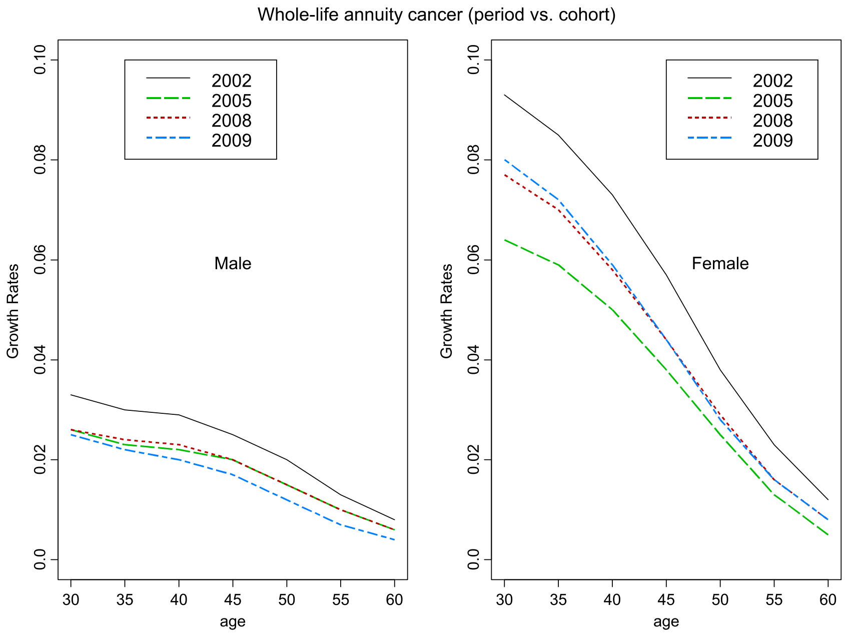


Fig. 8. % difference in pure premium whole-life annuity cancer (LC model).

Taiwan's National Health Insurance (NHI), we found that the incidence rates slightly increased and the mortality rates gradually decreased over the years in Taiwan. This was in accord with the increasing trend of the loss ratios for the cancer insurance products. This suggested that the longevity risk should be included in long-term health products. We also evaluated whether the frequently used mortality models can be employed to cope with the longevity risk or not. The LC and APC models were the top choices for modeling the cancer incidence and mortality rates. The CBD and Gompertz models were possible alternatives for the higher ages.

Taiwan is not the only country with increasing loss ratios, many Asian countries have similar experiences for long-term health products. For example, the number of those diagnosed with cancer in South Korea increased annually by 2.9% between 1999 and 2007, catapulting the loss ratio to around 120% (Source: Korea Insurance Development Institute). Lack of relevant experience data and use of foreign experience rates are the main reasons of increasing loss ratio. Therefore, we used Taiwan's population data (i.e. NHIRD) to acquire the cancer incidence and mortality rates. However, this was still not enough to meet our goal to design long-term cancer insurance products. To this end, the notion of longevity risk should be considered. We think that the influence of mortality improvement is bigger than that of the interest in insurance products, and this became more obvious when we looked at long-term and whole-life products. In fact, the mortality improvement is not always a plus to the insurers. In Taiwan, "whole-life" and "return principal" are the two common attributes in most life insurance policies, and we found that the decreasing mortality rates would cause an increase in the insurance premium (Yue and Huang, 2011).

Using the stochastic models to capture the future trend is one of the possibilities for dealing the increasing loss ratios. However, there still remains potential risk, if the changes of incidence and mortality rates are larger than expected. Natural hedge is one possible alternative, and we can bundle the long-term cancer insurance products with annuity products. Intuitively speaking, cancer patients often have shorter life expectancy and thus receive fewer payments in annuity. Another possibility is to transfer the financial burden of medical claim risk to the capital market, and the experience of the longevity bond can provide useful guidelines.

Furthermore, moral hazards also needs to be considered in cancer insurance. As mentioned in the first section, private clinics or health exam centers can help people to check if they have cancers without showing the exam results in the medical records. This increases the possibilities of adverse selection. There are quite a lot of possible remedies, such as extending the waiting period, reducing the coverage for the first policy year and coverage of the actual expenditures only. For example, a Taiwan's insurance company returns the premium paid to the insured persons if they are diagnosed with cancer in the first policy year, and returns 200% of the total premium paid if they are diagnosed with cancer in the second policy year.

The issue of big data also appeared in this study. We would expect more data to be available in future for the insurance companies, and not restricted to their own experience data. The insurance companies need to invest more on the manpower, such as organizing big data teams and training data scientists, in order to analyze the big data. It is neither cost effective nor feasible to hire outsiders to handle the data. For example, in order to handle the NHI databases (e.g. HV and HV_CD), our big data team has been operating for more than 10 years. Analysis of big data is a process of knowledge accumulation and trial-and-error method. In this regard, we keep updating the standard operating procedure for handling the NHI data over the years.

As for the stochastic models, we considered the GAPC family models in this study. Like in many previous studies, the LC and APC models showed better model fits. It seemed that the RH model (i.e. cohort modification of the LC model) does not fit well. It can be a result of shorter data period (fewer than 15 years) which makes observation of the cohort effect difficult. For model development, we can also consider other models for cancer incidence and mortality rates. Possible candidates include the discount sequence method and its extension (Wang and Yue, 2011; Wang et al., 2016b) and spatial modification of LC model. For example, Debón et al. (2008) considered adding a term of autocorrelation when the residuals of LC model showed systematic errors. We can add clusters or other first degree moment (i.e. mean shift) to the LC model (Wang and Yue, 2013).

Acknowledgments

This research was supported in part by a grant from the Ministry of Science and Technology in Taiwan, NSC 103-2410-H-004-093 and MOST 106-2410-H-156-006. We greatly appreciate the insightful comments from the editor and two anonymous reviewers, which helped us to clarify the context of our work.

References

Arbeev, K.G., Ukraintseva, S.V., Arbeeva, L.S., Yashin, A.I., 2005. Mathematical models for human cancer incidence rates. Demographic Res. 12 (10), 237–260.

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., Balevich, I., 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. N. Am. Actuar. J. 13 (1), 1–35.

Di Cesare, M., Murphy, M., 2009. Forecasting mortality, different approaches for different causes of death? The case of lung cancer, influenza, pneumonia, and bronchitis and motor vehicle accidents. British Actuar. J. 15 (suppl), 185–211.

Debón, A., Montes, F., Mateu, J., Porcu, E., Bevilacqua, M., 2008. Modelling residuals dependence in dynamic life tables: A geostatistical approach. Comput. Statist. Data Anal. 52, 3128–3147.

Hoggart, C., Brennan, P., Tjonneland, A., et al., 2012. A risk model for lung cancer incidence. Cancer Preven. Res. 5, 834–846.

Kruijshaar, M.E., Barendregt, J.J., Hoeymans, N., 2002. The use of models in the estimation of disease epidemiology. Bull World Health Organ 80, 622–628.

Lee, R.D., Carter, L.R., 1992. Modeling and forecasting US mortality. J. Amer. Statist. Assoc. 87 (419), 659–671.

Lee, T., Yang, C., Wang, T., 2011. Population aging and NHI expenditures in Taiwan. Population Studies 43, 1–35.

Liu, S.Y., 2000. Epidemiological transition in colonial taiwan and its explanation. In: Conference Paper for the Conference of Disease History. Academia Sinica, Taipei, Taiwan, (in Chinese). http://www.ihp.sinica.edu.tw/~medicine/conference/disease/shihyong.PDF.

Moger, T.A., Aalen, O.O., Halvorsen, T.O., Storm, H.H., Tretli, S., 2004. Frailty modelling of testicular cancer incidence using Scandinavian data. Biostataistics 5, 1–14.

National Vital Statistics Reports. 2016. Vol. 64 No. 2.

OECD, 2015. Health at a Glance 2015: OECD Indicators. OECD Publishing, Paris.

Plat, R., 2009. On stochastic mortality modeling. Insurance Math. Econom. 45 (3), 393–404.

Pokhrel, K.P., Tsokos, C.P., 2015. Forecasting age-specific brain cancer mortality rates using functional data analysis models. Adv. Epidemiol. http://dx.doi.org/10.1155/2015/721592.

Pompei, F., Wilson, R., 2001. The age distribution of cancer: the turnover at old age. Health Environ. Risk Assess 7, 1619–1650.

Renshaw, A.E., Haberman, S., 2006. A cohort-based extension to the Lee–Carter model for mortality reduction factors. Insurance Math. Econom. 38 (3), 556–570.

Robertson, C., Boyle, P., 1997. Statistical modelling of breast cancer incidence and mortality rates in Scotland. Br. J. Cancer 76, 1248–1252.

Rosner, B., Colditz, G., 1996. Nurses' health study: Log-incidence mathematical model of breast cancer incidence. J. Natl. Cancer Inst. 88, 359–364.

Shek, L.L., Godolphin, W., 1988. Model for breast cancer survival: relative prognostic roles of axillary nodal status, TNM stage, estrogen receptor concentration, and tumor necrosis. Cancer Res. 48 (19), 5565–5569.

Shibuya, K., Inoue, M., Lopez, A.D., 2005. Statistical modeling and projections of lung cancer mortality in 4 industrialized countries. Int. J. Cancer 117 (3), 476–485.

Statistical Abstract of Taiwan Province for the Past Fifty-One Years, 1946 first edition and 1994 reprinted (in Chinese). http://twstudy.iis.sinica.edu.tw/twstatistic50/.

Strehler, B.L., Mildvan, A.S., 1960. General theory of mortality and aging. Science 132, 14–21.

Uddin, S., Ullah, A., Najma, ., Iqbal, M., 2010. Statistical modeling of the incidence of breast cancer in NWFP. Pakistan. J. App. Quant Methods Med. 5, 159–165.

Villegas, A.M., Millossovich, P., Kaishev, V.K., 2016. StMoMo: An R Package for Stochastic Mortality Modelling. R package version 0.3.1. URL http://CRAN.R-project.org/package=StMoMo.

Wang, H.C., Yue, C.J., 2011. Using regular discount sequence to model elderly mortality. J. Popul. Stud. 43, 37–70 (in Chinese).

Wang, H.C., Yue, C.J., Tsai, Y.H., 2016a. Marital status as a risk factor in life insurance: An empirical study in Taiwan. ASTIN Bull. 46 (2), 487–505.

Wang, H.C., Yue, C.J., Chen, Y.X., 2016b. A study of elderly mortality models. J. Popul. Stud. 52, 1–42 (in Chinese).

Wang, T., Yue, C.J., 2013. Spatial clusters in a global-dependence model. Spatial Spatio-temporal Epidemiol. 5, 39–50.

Yue, C.J., Huang, H.C., 2011. A study of incidence experience for Taiwan life insurance. Geneva Papers Risk Insur. Issues Practice 36, 718–733.