RESEARCH ARTICLE

Modeling Age-specific Cancer Incidences Using Logistic Growth Equations: Implications for Data Collection

Xing-Rong Shen[1]*, Rui Feng[2], Jing Chai[1], Jing Cheng[1], De-Bin Wang[1,3]

Abstract

Large scale secular registry or surveillance systems have been accumulating vast data that allow mathematical modeling of cancer incidence and mortality rates. Most contemporary models in this regard use time series and APC (age-period-cohort) methods and focus primarily on predicting or analyzing cancer epidemiology with little attention being paid to implications for designing cancer registry, surveillance or evaluation initiatives. This research models age-specific cancer incidence rates using logistic growth equations and explores their performance under different scenarios of data completeness in the hope of deriving clues for reshaping relevant data collection. The study used China Cancer Registry Report 2012 as the data source. It employed 3-parameter logistic growth equations and modeled the age-specific incidence rates of all and the top 10 cancers presented in the registry report. The study performed 3 types of modeling, namely full age-span by fitting, multiple 5-year-segment fitting and single-segment fitting. Measurement of model performance adopted adjusted goodness of fit that combines sum of squred residuals and relative errors. Both model simulation and performance evalation utilized self-developed algorithms programed using C# languade and MS Visual Studio 2008. For models built upon full age-span data, predicted age-specific cancer incidence rates fitted very well with observed values for most (except cervical and breast) cancers with estimated goodness of fit (Rs) being over 0.96. When a given cancer is concerned, the R valuae of the logistic growth model derived using observed data from urban residents was greater than or at least equal to that of the same model built on data from rural people. For models based on multiple-5-year-segment data, the Rs remained fairly high (over 0.89) until 3-fourths of the data segments were excluded. For models using a fixed length single-segment of observed data, the older the age covered by the corresponding data segment, the higher the resulting Rs. Logistic growth models describe age-specific incidence rates perfectly for most cancers and may be used to inform data collection for purposes of monitoring and analyzing cancer epidemic. Helped by appropriate logistic growth equations, the work vomume of contemporary data collection, e.g., cancer registry and surveilance systems, may be reduced substantially.

Keywords: Cancer - incidence - models - logistic growth equations - data collection - China cancer registry

Asian Pac J Cancer Prev, **15 (22)**, 9731-9737

Introduction

Cancers have long been a major cause of human death and diseases-adjusted life year loss (Ma et al., 2006). In 2002, new cancer cases and deaths accounted for 10.86 and 6.73 million respectively worldwide (Parkin et al., 2000). In 2005, over 7.6 million people died to cancer accounting for 13% of total death (Parkin et al., 2005). Predicted cases and deaths will rise to 15 and 10 million by 2020 (Parkin et al., 2001). WHO data showed that malignant tumors worldwide took up 5% of total burden caused by all diseases in 2005 (World Health Organization, 2006). More recent investigations revealed that cancers were the first death cause in cities and higher than cerebrovascular diseases and cardiopathy (China Ministry of Health, 2008). Estimated direct and indirect economic loss due to the disease was 11.323 and 6.000 billion USD respectively representing 4.67% of total medical cost (Wei, 2009).

Escalating cancer threats and harms have attracted tremendous efforts exploring the epidemiology of the diseases via large scale secular registry or surveillance systems (Goss et al., 2014; Ullrich et al., 2014). These efforts have been accumulating vast data that allow for establishment of mathematic models simulating cancer incidence and mortality rates for different age groups, time periods or cohorts. Leung and colleagues proposed a model for analyzing cervical cancer incidence using maximum likelihood and Bayesian methods and data from the Hong Kong Cancer Registry (Leung et al., 2006). Tyson and coworkers established a model incorporating the effects of age, year of diagnosis, and year of birth on incidence trends of renal cell carcinoma using data from

[1]School of Health Service Management, Anhui Medical University, [2]Department of Literature Review and Analysis, Library of Anhui Medical University, [3]Collaboration Center for Cancer Control, First Affiliated Hospital of Anhui Medical University, Hefei, China
**For correspondence: dbwang@vip.sina.com*

United States National Cancer Institute's Surveillance, Epidemiology, and End Results public-use registry (Tyson et al., 2013). The most commonly adopted approaches for modeling cancer rates are time series and APC models (Meira et al., 2013; Ocana-Riola et al., 2013; Wang et al., 2014). Usually, time series models assume a Poisson distribution of cancer counts and include autoregressive error terms and/or time trends (Wingo et al., 1998; Knorr-Held et al., 2001); while APC models generally consists of three components, i.e., age (A), period (P), and cohort (C) (Jurgens et al., 2014). Of all the variables studied so far, age seems to have the highest effect on cancer mortality and incidence rates (Dyzmann-Sroka et al., 2014). So the performance of models depends heavily on how the influence of age on cancers is incorporated. With contemporary models, methods used simulating the relationship between age and cancer rates include mainly linear (Lee et al., 2011), polynomial (Wingo et al., 1998), piecewise linear (Kim et al., 2000), spline, log linear (Du et al., 2014) or power curves (Moller, 2004). Typical S-shaped line graphs of age-specific cancer incidence and mortality rates are clearly observable with almost all cancers worldwide (Bouchbika et al., 2013; Al-Hashimi et al., 2014; Wei et al., 2014). If S-line represents true general pattern, most methods (including linear, log linear, power curves) used in previous models may not fitt well at least for some age ranges (e.g., under 35 or over 75 years). Some of the curves (e.g., piecewise linear and spline curves) may adequately approach any S-line. Yet this requires much more detailed data about observed age-specific cancer counts. Besides, most previous work in this regard focuses mainly on predicting or analyzing cancer epidemic with little attention being paid to informing relevant data collection.

This paper models age-specific cancer incidence rates using logistic growth equations. Although there are evidences that such equations describe well cancer cell proliferation under various conditions (Fory's et al., 2003), publications linking them with age-specific cancer rates are limited. In particular, the paper explores the performance of logistic growth models under different scenarios of data completeness. This may reveal clues for reshaping contemporary data collection, e.g., cancer registry, surveillance or evaluation initiatives. Given the huge amount of scarce resources invested annually on these initiatives (Hutchison et al., 1997), they merit continuous scrutinize and refinement.

Materials and Methods

Data source

All source data used in this study came from China Cancer Registry Report 2012, the latest available annual report of the kind by far (He et al., 2012). It draws from data collected in 2009 by 72 sites throughout China covering 85.47 million urban and rural Chinese residents and provides incidence and mortality rates of all and over 20 specific cancers by age, gender and registry sties. A sample datasets was given in our previous paper (Chen et al., 2014) and detailed charateristics of the data will be described seperately.

Formulae used

Based on empirical observations of patterns with the reported cancer incidence rates along different ages for all and specific cancers, the study adopted a three parameter logistic growth equation (Formula 1). In this formula, t stands for age; and p_t, cancer incidence rate for a given age t; p_{max}, the highest cancer incidence rate for all ages; k, growth rate; while b serves as a baseline growth rate that determines the location of "the rapidly growing phase" of a S-curve along the age spectrum. In addition, the study used Formula 2 in identifying the most optimal model from a set of potential models for a given type of cancer and in evaluating the performance of the models selected. In Formula 2, R represents goodness of fit of the model under concern; while p_{ot} and p_{st} stand for observed and simulated (or predicted) cancer incidence rate for a given age t resperctively.

$$p_t = \frac{p_{max}}{1+e^{b-kt}} \quad (1)$$

$$R = 1 - \sqrt{\frac{\sum(p_{ot}-p_{st})^2}{\sum {p_{ot}}^2}} \quad (2)$$

Selection of input data

In terms of how input data were selected, the study performed 3 types of modeling, namely full age-span fitting, multiple 5-year-segment fitting and single-segment fitting. Full age-span fitting utilized the registered cancer incidence data covering all the ages (i.e., from age 0 through to age 85); while the other two types of fitting, only part of the data. The multiple 5-year-segment fitting divided the whole age-span into segments consisting of 5 consecutive ages (e.g., ages 0-4, ages 5-9 etc.) first and then enter the corresponding observed age-specific cancer incidence rates for every other (e.g., ages 0-4, ages 10-14, …, ages 70-74, ages 80-84), every other 2, every other 3, every other 4 and every other 5 segments into modeling respectively. With regard to single-segment fitting, it selected only one segment of ages and entered corresponding incidence rates into modeling. These segments covered 15, 30, 45, 60 and 75 consecutive ages respectively. Considering that location along the age-span covered by a same length of data segment may result in different model parameters, the study set beginning, middle and end as 3 crieteria for selecting single segment data. For example, for the segment consisting of 15 ages, the study established 3 different logistic growth equations based on registered cancer incidence rates for ages 0-14 (beginning segment), ages 38-52 (middle segment) and ages 70-84 (end segment) respectively.

Algorithms for model building

Given that available statistical software does not allow for logistic growth modeling using segmental input data. The study employed a self-developed mini-program to perform all the computation. Written in C# language, the program runs on a webpage built with Microsoft Visual Studio 2008. For each model building, the webpage accepts an intended age set (e.g., {61; 62; 63; 64; 65})

and a corresponding set of observed cancer incidence rates (e.g., {351; 362;373;384; 395}, in 1/100000) as input and then produces a best-fit parameter set (e.g., {p_{max}=200, b=10.5, k=0.17}) and a goodness of fit (e.g., R=0.98). This computation proceeds in 5 steps. Step 1 assumes a proper value range for each of the 3 parameters included in Formula 1 (i.e., $0 \le p_{max} \le 5000, 0 \le b \le 20, 0 \le k \le 0.8$). Step 2 sets a small enough incremental value for each of the 3 parameters, i.e., 1 for p_{max}, 0.1 for b and 0.01 for k, respectively and divides the ranges of p_{max}, b and k into 3 serial parameter sets, i.e., {0, 1, 2, …, 1×i, …, 4999, 5000}, {0, 0.1, 0.2, 0.3, …, 0.1×j, …, 19.8, 19.9, 20.0} and {0, 0.01, 0.02, …, 0.01×n, …, 0.78, 0.79, 0.80}; here i=1, 2, 3, …, 5000 (i.e., 5000/1), j=1, 2, 3, …, 200 (i.e., 20/0.1) and n=1, 2, 3, …, 80 (i.e., 0.8/0.01). Step 3 selects one element from each of the 3 serial parameter sets and generates a complete set (5000×200×80 elements in total) of potential parameter combinations, i.e., {{p_{max}=1, b=0, k=0}, {p_{max}=0, b=0, k=0.01}, …, {p_{max}=5000, b=20, k=0.79}, {p_{max}=5000, b=20, k=0.80}}. Step 4 uses Formula 2 and compares the goodness of fit between the registered cancer incidence set entered via the webpage and that predicted by Formula 1 using each of the potential parameter combinations. Step 5 outputs the parameter combination that has the largest R.

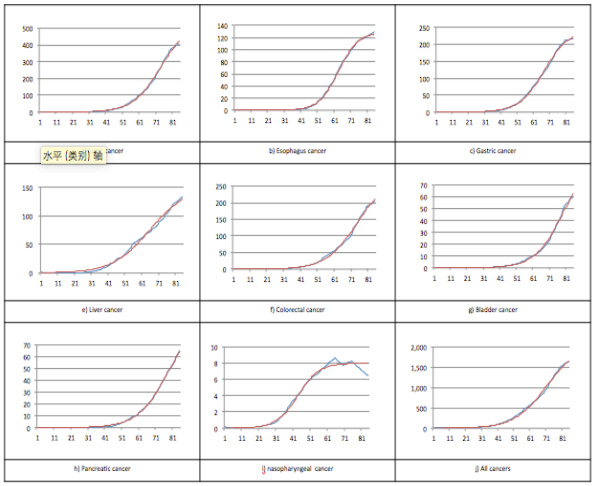

Figure 1. Predicted *vs* Registered Age-specific Cancer Incidence Rates. Red lines represent predicted incidence rates and blue lines, actual incidence rates; Y-Axis represents cancer incidence rate in 1/100000 and X-Axis, age; Data source came from China cancer registry report 2012

Results

Full age-span fitting models

As shown in Table1 and Figure 1, the majority of curves representing observed age-specific cancer incidence rates fit very well with predictions by logistic growth models with estimated goodness of fit (R) being over 0.96. Yet the R values for some types of cancer, e.g., cervical cancer and breast cancer, were quite low, ranged from 0.25 to 0.92. The 3 parameters defining the logistic growth models showed substantial variations. P_{max} ranged from 8 to 2248; b, from 5.50 to 14.80; and k, from 0.08 to 0.44.

Multiple-5-year-segment fitting models

Table 2 and Figure 2 displays findings from multiple-5-year-segment modeling. Goodness of fit (R) decreased as the number of data segments being left out increased. Yet, it remained fairly high even when only one fourth of segments of observed data were entered into modeling. And this phenomenon applied to all types of cancers. The differences between the R values of models for different cancer types (excluding cervical and breast cancer) built upon "every other segment" of observed data (Table 2, column 5) and that built upon "every other 4 segments" of

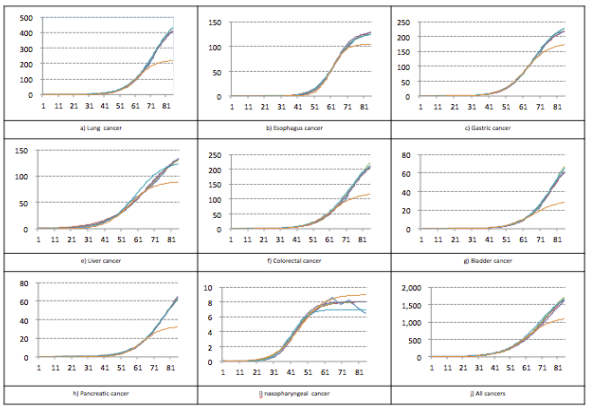

Figure 2. Predicted *vs* Registered Age-specific Cancer Incidence Rates Using Difference Segment of Input Data. Blue lines represent actual incidence rates; and red, green, purple, light blue, brown lines represent predicted incidence rates using every other 1, 2, 3, 4, 5 segment of input data respectively, Y-Axis represents cancer incidence rate in 1/100000 and X-Axis, age; Data source came from China cancer registry report 2012

Table 1. Parameters and Goodness of Fit of Logistic Growth Models Based on Full-age-span Cancer Incidence Data

Type of cancer	Total				Males				Females				Urban				Rural			
	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R
Lung cancer	532	8.7	0.12	0.97	779	8.8	0.12	0.97	339	8.7	0.12	0.96	593	8.8	0.12	0.97	366	8.9	0.13	0.95
Esophagus cancer	129	10.6	0.17	0.98	200	8.3	0.13	0.93	97	9.9	0.15	0.96	86	9	0.14	0.96	261	9.5	0.15	0.95
Gastric cancer	245	8.6	0.13	0.97	338	9.1	0.14	0.97	195	7.3	0.1	0.97	244	7.8	0.11	0.96	309	9.39	0.15	0.96
Liver cancer	151	5.8	0.09	0.94	194	5.5	0.09	0.93	128	7.3	0.1	0.98	167	5.6	0.08	0.94	152	5.9	0.1	0.94
Colorectal cancer	272	8	0.11	0.96	333	8.1	0.11	0.96	244	7.4	0.1	0.96	336	8.1	0.11	0.96	125	7.5	0.11	0.96
Bladder cancer	91	9.3	0.12	0.96	160	9.5	0.12	0.97	49	9.5	0.12	0.94	107	9.3	0.12	0.96	56	8.7	0.11	0.96
Cervical cancer	10	13.5	0.39	0.73	NA	NA	NA	NA	20	9.9	0.29	0.7	10	14.8	0.44	0.67	11	14.6	0.39	0.77
Breast cancer	44	11.6	0.28	0.9	3	5.9	0.08	0.86	85	9.8	0.24	0.87	53	9.9	0.24	0.92	23	9.7	0.25	0.79
Pancreatic cancer	95	8.5	0.11	0.98	112	8.5	0.11	0.96	74	9.1	0.12	0.98	108	8.6	0.11	0.98	54	9.2	0.13	0.96
Nasopharyngeal cancer	8	6.8	0.16	0.92	12	6.9	0.16	0.92	4	6.7	0.17	0.92	9	6.7	0.16	0.89	6	6.6	0.15	0.88
All cancers	2064	7	0.1	0.96	2710	7.8	0.11	0.97	1459	6.1	0.09	0.94	2424	6.6	0.09	0.96	1578	7.1	0.11	0.96

*Note: Source data came from age-specific incidence rates of top ten and all cancers from China cancer registry report 2012; P_{max} , b and k represents the parameters in the logistic equation, $y_t=P_{max}/(1+e^{b-kt})$, where t stands for age and yt, incidence rate for age t; R stands for goodness of fit between predicted and observed age-specific cancer incidence rates; NA stands for not applicable

Table 2. Parameters and Goodness of Fit of Logistic Growth Models Based on Multiple 5-age-segment Cancer Incidence Data

Type of cancer	Every other segment				Every other 2 segments				Every other 3 segments				Every other 4 segment				Every other 5 segments			
	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R
Lung cancer	485	9.2	0.13	0.96	517	9.3	0.13	0.96	484	9.2	0.13	0.96	530	8.6	0.12	0.95	224	10.4	0.17	0.63
Esophagus cancer	132	10	0.16	0.98	126	11.1	0.18	0.98	130	11.8	0.19	0.96	127	10.5	0.17	0.98	105	14.4	0.24	0.88
Gastric cancer	238	8.5	0.13	0.97	250	8.6	0.13	0.97	233	9.1	0.14	0.97	251	8.6	0.13	0.96	178	9.2	0.15	0.84
Liver cancer	167	5.4	0.08	0.94	156	5.9	0.09	0.94	158	5.9	0.09	0.94	129	7	0.12	0.89	90	7.6	0.14	0.78
Colorectal cancer	248	8.5	0.12	0.95	291	8.1	0.11	0.95	269	8	0.11	0.96	262	7.8	0.11	0.93	120	8.5	0.14	0.65
Bladder cancer	90	9.3	0.12	0.96	105	9.5	0.12	0.94	81	9.8	0.13	0.95	95	9.3	0.12	0.94	31	8.5	0.13	0.55
Cervical cancer	10	18.5	0.54	0.73	10	14.8	0.43	0.73	10	19.8	0.67	0.67	10	14.7	0.44	0.73	10	9.5	0.27	0.72
Breast cancer	44	12.1	0.29	0.9	43	10.9	0.27	0.9	42	11.7	0.28	0.89	44	16.5	0.47	0.74	46	9.2	0.22	0.89
Pancreatic cancer	95	8.5	0.11	0.98	80	8.9	0.12	0.97	83	9	0.12	0.97	90	8.4	0.11	0.97	33	10.7	0.17	0.62
Nasopharyngeal cancer	8	6.8	0.16	0.92	8	7.2	0.17	0.92	8	8.5	0.2	0.91	7	8.4	0.21	0.89	9	6.3	0.14	0.87
All cancers	2248	6.5	0.09	0.97	2180	7.1	0.1	0.96	2229	6.6	0.09	0.97	1989	6.8	0.1	0.94	1151	7.2	0.12	0.75

*Note: Source data came from age-specific incidence rates of top ten and all cancers from China cancer registry report 2012; P_{max}, b and k represents the parameters in the logistic equation, $y_t=P_{max}/(1+e^{b-kt})$, where t stands for age and yt, incidence rate for age t; R stands for goodness of fit between predicted and observed age-specific cancer incidence rates

Table 3. Parameters and goodness of fit of Logistic Growth Models Based on Single-segment Cancer Incidence Data

Type of cancer	15 ages				30 ages				45 ages				60 ages				75 ages			
	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R	P_{max}	b	k	R
All cancers																				
-Beginning	20	0.1	0.01	0.71	278	3.7	0.04	0.72	561	6.7	0.13	0.91	965	7.4	0.13	0.97	2099	6.4	0.09	0.96
-Middle	476	7.3	0.15	0.98	966	7.4	0.13	0.98	971	7.4	0.13	0.98	1384	6.9	0.11	0.96	2067	7	0.1	0.96
-End	2026	7.7	0.11	0.98	1985	6.9	0.1	0.96	1988	6.9	0.1	0.96	2064	7	0.1	0.96	2064	7	0.1	0.96
Lung cancer																				
-Beginning	1	2.9	0.01	0.55	7	7.1	0.18	0.9	49	10.5	0.22	0.97	185	10	0.17	0.98	530	8.7	0.12	0.97
-Middle	70	9.7	0.19	0.96	184	10	0.17	0.98	216	9.7	0.16	0.99	397	8.9	0.13	0.96	565	8.8	0.12	0.97
-End	433	14	0.2	0.99	532	8.7	0.12	0.97	488	9.2	0.13	0.97	488	9.2	0.13	0.96	532	8.7	0.12	0.97
Esophagus cancers																				
-Beginning	1	2.9	0.01	0.24	1	3.1	0.01	0.26	34	13	0.25	0.94	77	12.5	0.22	0.98	127	9.9	0.16	0.97
-Middle	51	13	0.24	0.96	71	12.9	0.23	0.98	115	11.5	0.19	0.97	120	11	0.18	0.98	128	9.9	0.16	0.97
-End	133	9.4	0.15	0.99	133	9.4	0.15	0.98	131	10	0.16	0.98	129	10.6	0.17	0.98	129	9.9	0.16	0.97
Gastric cancer																				
-Beginning	1	4.7	0.19	0.83	25	8.6	0.2	0.94	84	8.8	0.16	0.96	177	9.2	0.15	0.98	234	8.5	0.13	0.97
-Middle	76	8.7	0.16	0.97	206	9.4	0.15	0.98	177	9.2	0.15	0.99	207	8.9	0.14	0.98	263	8.1	0.12	0.97
-End	224	14.8	0.22	0.99	253	8	0.12	0.97	245	8.6	0.13	0.97	245	8.6	0.13	0.97	245	8.6	0.13	0.97
Liver cancer																				
-Beginning	1	0.7	0.01	0.4	96	8.1	0.14	0.63	39	9	0.21	0.96	80	7.7	0.15	0.97	115	6.4	0.11	0.95
-Middle	47	8.1	0.18	0.98	85	7.3	0.14	0.96	81	7.7	0.15	0.97	103	6.7	0.12	0.96	144	5.7	0.09	0.94
-End	165	6.9	0.1	0.99	288	4.3	0.05	0.99	191	5	0.07	0.96	167	5.4	0.08	0.94	151	5.8	0.09	0.94
Colorectal cancer																				
-Beginning	1	5.5	0.21	0.86	3	7.2	0.25	0.94	26	7.8	0.17	0.98	124	8.5	0.14	0.97	320	7.6	0.1	0.95
-Middle	83	7.6	0.13	0.97	205	8.6	0.13	0.97	115	8.4	0.14	0.98	208	7.6	0.11	0.95	342	7.7	0.1	0.94
-End	216	14	0.2	0.99	302	7.5	0.1	0.96	302	7.5	0.1	0.96	272	8	0.11	0.96	272	8	0.11	0.96
Bladder cancer																				
-Beginning	1	4.1	0.01	0.47	81	10.3	0.16	0.88	26	8.7	0.14	0.96	25	8.7	0.14	0.97	118	9	0.11	0.95
-Middle	8	7.8	0.15	0.96	45	8.9	0.13	0.97	21	9	0.15	0.98	191	8.9	0.1	0.96	94	8.5	0.11	0.97
-End	68	13.9	0.19	0.98	84	9.9	0.13	0.96	91	9.3	0.12	0.96	91	9.3	0.12	0.96	95	8.5	0.11	0.96
Pancreatic cancers																				
-Beginning	0	0	0	0	1	6.7	0.18	0.87	96	11.2	0.16	0.94	17	11.1	0.2	0.98	95	8.5	0.11	0.97
-Middle	43	11.2	0.18	0.97	15	11.4	0.21	0.98	29	9.8	0.16	0.97	68	8.7	0.12	0.96	94	8.5	0.11	0.98
-End	95	8.5	0.11	0.99	95	8.5	0.11	0.99	95	8.5	0.11	0.98	95	8.5	0.11	0.98	95	8.5	0.11	0.98
Nasopharyngeal cancer																				
-Beginning	9	5.3	0.01	0.5	3	5.6	0.15	0.89	6	7.7	0.2	0.96	8	6.8	0.16	0.97	8	6.8	0.16	0.96
-Middle	9	5.3	0.12	0.98	8	6.4	0.15	0.97	9	5.8	0.13	0.96	8	6.8	0.16	0.95	8	6.8	0.16	0.96
-End	8	0.2	0.04	0.91	8	14.6	0.3	0.92	8	6.8	0.16	0.93	8	6.8	0.16	0.92	8	6.8	0.16	0.92

*Note: Source data came from age-specific incidence rates of top ten and all cancers from China cancer registry report 2012; P_{max}, b and k represents the parameters in the logistic equation, $y_t=P_{max}/(1+e^{b-kt})$, where t stands for age and yt, incidence rate for age t; R stands for goodness of fit between predicted and observed age-specific cancer incidence rates

observed data (Table 2, column 17) ranged from only 0 to 0.05. However, starting from the column of "every 5 other segments", the R values reduced dramatically. Similarly, although all the 3 parameters of the simulated logistic growth equations varied as the number of segments of data entered for modeling changed, most of these variations remained to a minimum extent (less than 10%) until the column of "every other 4 segments" and did not show clear decreasing or increasing trend.

Single-segment fitting models

Table 3 and Figure 3 resulted from single-segment fittings. Goodness of fit (R) increased as the length of the data segment increased and this increase was dependent on the location of the segment of data entered for fitting. For a same length of segment (e.g., 15 ages), the older the age covered by the corresponding data segment, the higher the resulting R. As for the segment covering the oldest part of age-span, all the R values turned out to be very high. The

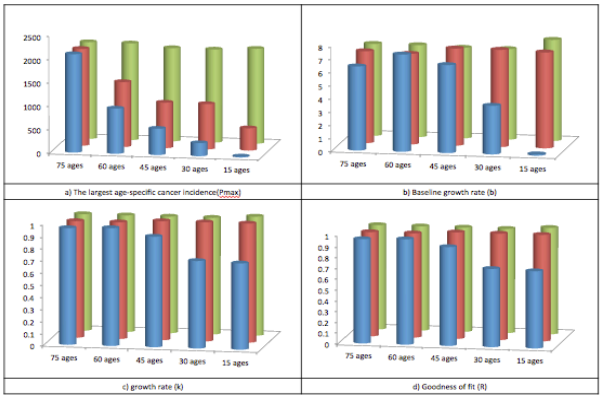

Figure 3. Indicators of Single Segment-fitting Models for All Cancers. Blue, red and green histograms represent parameters of models built using data segment covering the beginning, middle and end part of age-span respectively

modeled parameters were also linked to the length and age-range covered by the data segment. For data segment covering beginning ages, all the 3 parameters increased as the length changed from 15 ages to 75 ages; while for data segment covering the end ages, P_{max} increased yet b and k decreased as the length increased.

Discussion

Although typical S-shaped line graphs of age-specific cancer incidence rates are clearly observable with almost all cancer registry and other relevant epidemological reports worldwide, their relations with logistic growth equations have not been fully addressed. The current study demonstrated that logistic growth models perfectly describe the incidence rates along different age groups for most type of cancers. This may be explained by: a) onset of clinically detectable cancers results from the counteraction between cancer cell occurrence and removal (Baker et al., 2013); b) cancer cell occurs after a normal somatic cell has experienced multiple times (say n times) of damages due to exposure to same or different risk factors (Shaukat et al., 2013); c) a certain level of risk exposure defines a corresponding chance (q) for a normal somatic cell to get one time damage and hence the chance (q^n) for an innate cell to mutate into cancer cell in an unit time period; d) given c, as time (t) passes by and somatic cell gets damaged for more and more times, its chance (p) for becoming malignant increases exponentially ($p \cong q^{n \cdot qt}$); e) level of life spectrum exposure to cancer risk factors starts relatively low at birth, increases during childhood and adolescence (due initiation of unhealthy or unprotected behaviors), remains the highest in adulthood and begins to decrease gradually in late lifetime (due to reduced smoking, drinking etc.) (Katulanda et al., 2014; Chockalingam et al., 2013); f) cancer cell removal or immunity manifests similar lifetime trend as risk exposure (Wu et al., 2012). Therefore, the early low and relatively stable phase of the S-shaped age-specific cancer rates may reflect the combined effect of low cancer cell occurrence vs. high immunity; while the rapidly growing part, exponentially increasing occurrence vs. high and stable immunity; and the late high and relatively stable stage, diminishing occurrence due to reduced risk exposure vs. downward immunity.

Linking logistic growth law with age-specific cancer rates leads to a plausible thinking that description of cancer incidence or mortality rates along the whole age span is to estimate the parameters involved in the equations rather than uncover counts for each of the ages. Such a shift of focus may result in great resource reduction, since logistic equations generally involve only a few parameters (e.g., 3 parameters in our cases) and estimation of these requires much less data than what have usually been collected. This is of particular significance to cancer registry. As suggested by our simulations (Table 2 and Figure 2), the work volume of current China national cancer registry could be reduced by 3 fourths without severely damage its capacity in producing age-specific cancer incidence rates. This should also apply to other registries. Given that over fifty countries have large scale operating cancer registry systems that consume huge amount of scarce resources year by year (Izquierdo et al., 2000; Tangka et al., 2010), a growth model-guided rethinking merits special attention. Even though segmental cancer registry may sound unacceptable to some, the findings suggest priority age groups for monitoring and controlling data quality of registry systems.

Logistic growth analysis may also inform data collection for intervention or hypothesis assessments. As shown in Table 3 and Figure 3, for a same length of data segment, the older the age covered by the data, the higher the goodness of fit of the resulting model. This suggests that, for studies evaluating the effect of an intervention or an influencing factor on cancer rates using limited age groups, backward sampling (i.e., start to choose from the oldest age group backward to younger ones) may work better than forward selection (from age 0-5 to 6-10 and then to 11-15 etc.). For studies that have yielded data showing differences in cancer rates between two groups (say, intervention *vs* control) of middle ages (say, ages 30-59), simulated logistic growth equations may be used to measure extended difference (say for ages 60-69, or even 60 and over) between the two groups. However, the goodness of fit of models based on data covering middle segment of ages is only moderate.

In addition, logistic growth equations may help assessing data collections biases and/or errors under certain circumstances. If there are sufficient evidences to believe that certain age-specific cancer rates follow logistic growth law, then the goodness of fit estimations (Rs) can also be viewed as a quality indicator of the observed cancer counts. Of the ten cancers included in Table 1, cervical and breast cancers showed clear deviations from logistic equations. By excluding these two cancers, all the cancer-specific pairs of Rs (Table1, column 17 *vs* 21) showed a consistent trend, i.e., for any given cancer, the R of the model built upon observed data from urban residences was higher than or at least equal to that from rural people. This may indicate better cancer registry in urban than in rural China. The Rs for models of different cancers witnessed much greater variations ranging from 0.92 for nasopharyngeal cancer to 0.98 for esophagus and pancreatic cancers (Table 1, column 5). This suggests a need for tailored data quality control or

improvement with special attention being paid to cancers with the lowest Rs. The varied biases and errors in the rates for different cancers in our case may be attributed to a whole range of reasons including number of cases registered (e.g., too few for nasopharyngeal cancer), physical symptoms and sings, easiness to get cancer tissues for pathologic diagnosis, availability of auxiliary examination techniques etc.

Finally, readers are cautioned about a number of issues. First, this study used only most simple logistic growth equations and they do not fit very well with the observed data for some cancers, e.g., cervical and breast cancers. Such problems can be solved by adding more parameters and introducing more sophisticated growth equations. Second, parameters presented in this paper were all average estimates derived from pooled cancer counts reported by 72 CNCR sites in 2009. Age-specific incidence and mortality bands with means and 95% confidence intervals rather than single mean estimates may be produced by building similar set of logistic growth equations using the data from each CNCR sites (72 sets in total) and then performing bootstrap re-sampling and jackknife-correction (Dexter et al., 2013; Yu et al., 2013). Third, this paper focuses primarily on implications for data collection without any attention being paid to identifying trends and components with the cancer rates. Forth, apart from goodness of fit, this paper did not provide other performance indicators (sensitivity, specificity etc.) of the models used due to space limit. Most of these will be addressed separately in a forthcoming paper titled "modeling age-specific cancer incidence using logistic growth equations: jackknife-corrected bootstrap estimates".

Acknowledgements

This paper was funded by the Natural Science Foundation of China (Grant Number 81172201). All authors declare no conflicts of interest.

References

Baker K, Rath T, Flak MB, et al (2013). Neonatal Fc receptor expression in dendritic cells mediates protective immunity against colorectal cancer. *Immunity*, **39**, 1095-107.

Bouchbika Z, Haddad H, Benchakroun N, et al (2013). Cancer incidence in Morocco: report from Casablanca registry 2005-2007. *Pan Afr Med J*, **16**, 31.

Chen PL, Zhao T, Feng R, et al (2014). Patterns and trends with cancer incidence and mortality rates reported by the China National Cancer Registry. *Asian Pac J Cancer Prev*, **15**, 6327-32.

Chockalingam K, Vedhachalam C, Rangasamy S (2013). Prevalence of tobacco use in urban, semi urban and rural areas in and around Chennai City, India. *PLoS One*, **8**, 76005.

China Ministry of Health (2008). China third national death cause survey. *China Cancer*, **5**, 344.

Dyzmann-Sroka A, Malicki J (2014). Cancer incidence and mortality in the greater poland region-analysis of the year 2010 and future trends. *Rep Pract Oncol Radiother*, **19**, 296-300.

Du LB, Li HZ, Wang XH, et al (2014). Analysis of cancer incidence in Zhejiang cancer registry in China during 2000 to 2009. *Asian Pac J Cancer Prev*, **15**, 5839-43.

Dexter TA, Kowalewski M (2013). Jackknife-corrected parametric bootstrap estimates of growth rates in bivalve mollusks using nearest living relatives. *Theor Popul Biol*, **90**, 36-48.

Fory's U, Marciniak CA (2003). Logistic equations in tumor growth modeling. *Int J Appl Math Comput Sci*, **13**, 317-25.

Goss PE, Strasser-Weippl K, Lee-Bychkovsky BL, et al (2014). Challenges to effective cancer control in China, India, and Russia. *Lancet Oncol*, **15**, 489-538.

He J, Chen WQ (2012). Chinese cancer registry annual report. *Chin J Cancer Res*, **24**, 171-80.

Hutchison C, Roffers S, Fritz A (1997). Cancer registry management: principles and practice. lenexa, kan: kendall/ hunt publishing Co.

Izquierdo JN, Schoenbach VJ (2000). The potential and limitations of data from population-based state cancer registries. *Am J Public Health*, **90**, 695-8.

Jürgens V, Ess S, Cerny T, Vounatsou P (2014). A Bayesian generalized age-period-cohort power model for cancer projections. *Stat Med*, **33**, 4627-36.

Katulanda P, Ranasinghe C, Rathnapala A, et al (2014). Prevalence, patterns and correlates of alcohol consumption and its' association with tobacco smoking among Sri Lankan adults: a cross-sectional study. *BMC Public Health*, **14**, 612.

Knorr-Held L, Rainer E (2001). Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics*, **2**, 109-29.

Kim HJ, Fay MP, Feuer EJ, Midthune DN (2000). Permutation tests for join point regression with applications to cancer rates. *Stat Med*, **19**, 335-51.

Lee TC, Dean CB, Semenciw R (2011). Short-term cancer mortality projections: a comparative study of prediction methods. *Stat Med*, **30**, 3387-402.

Leung GM, Woo PP, McGhee SM, et al (2006). Age-period-cohort analysis of cervical cancer incidence in Hong Kong from 1972 to 2001 using maximum likelihood and Bayesian methods. *J Epidemiol Community Health*, **60**, 712-20.

Moller B (2004). Prediction of cancer incidence-methodological considerations and trends in the Nordic countries 1958-2022. phd thesis, faculty of medicine, university of oslo, unipuc AS, oslo.

Meira KC, Silva GA, Silva CM, Valente JG (2013). Age-period-cohort effect on mortality from cervical cancer. *Rev Saude Publica*, **47**, 274-82.

Ma X, Yu H (2006). Global burden of cancer. *Yale J Biol Med*, **79**, 85-94.

Ocana-Riola R, Mayoral-Cortes JM, Blanco-Reina E (2013). Age-period-cohort effect on lung cancer mortality in southern Spain. *Eur J Cancer Prev*, **22**, 549-57.

Parkin DM, Bray F, Ferlay J, Pisani P (2005).Global cancer statistics, 2002, *CA Cancer J Clin*, **55**, 74-108.

Parkin DM, Bray F, Ferlay J, Pisani P (2001).Estimating the world cancer burden: GLOBOACN 2000. *Int J Cancer*, **94**, 153-6.

Parkin DM (2001). Global cancer statistics in the year 2000. *Lancet Oncol*, **2**, 533-43.

Shaukat U, Ismail M, Mehmood N (2013). Epidemiology, major risk factors and genetic predisposition for breast cancer in the Pakistani population. *Asian Pac J Cancer Prev*, **14**, 5625-9.

Tyson MD, Humphreys MR, Parker AS, et al (2013). Age-period-cohort analysis of renal cell carcinoma in United States adults. *Urology*, **82**, 43-7.

Tangka F, Subramanian S, Beebe MC, Trebino D, Michaud F (2010). Economic assessment of central cancer registry operations, Part III: Results from 5 programs. *J Registry*

Manag, **37**, 152-5.
Ullrich A, Miller A (2014). Global response to the burden of cancer: the WHO approach. *Am Soc Clin Oncol Educ Book*, 311-5.
Wang P, Xu C, Yu C (2014). Age-period-cohort analysis on the cancer mortality in rural China: 1990-2010. *Int J Equity Health*, **13**, 1.
Wei KR, Yu X, Zheng RS, et al (2014). Incidence and mortality of liver cancer in China, 2010. *Chin J Cancer*, **33**, 388-94.
Wu J, Li W, Liu Z, et al (2012). Ageing-associated changes in cellular immunity based on the SENIEUR protocol. *Scand J Immunol*, **75**, 641-6.
Wei QL (2009). Malignant disease burden research. MD thesis. Xiamen university.
World Health Organization. World health statistics 2006. Geneva WHO.
Wingo PA, Landis S, Parker S, et al (1998). Using cancer registry and vital statistics data to estimate the number of new cancer cases and deaths in the United States for the upcoming year. *J Registry Management*, **25**, 43-51.
Yu LY, Chen ZZ, Zheng FQ, et al (2013). Demographic analysis, a comparison of the jackknife and bootstrap methods, and predation projection: a case study of *Chrysopa pallens* (Neuroptera: Chrysopidae). *J Econ Entomol*, **106**,1-9.