Master Thesis

"How many will be suffered by Cancer in 2020? Forecasting with Lee-Carter model with covariates"

Prepared for: Matin Dribe
Prepared by: Takashi Ando: mas08tan (Name in Japanese: 安藤 高志)
Supervised by: Panagiotis Mantalos

CONTENTS

ABSTRACT

By using Lee-Carter model by sex and the origin of tumor we forecasted the cancer incidence. The usefulness of Lee-Carter methods to the incidence rates was shown. For forecasting the future the time trend parameter of Lee-Carter model behaves very close to Random-walk but it is not exactly Random-walk for most cases. It is better to leave the option open to AR(1) and not restrict ourselves to Random-walk. Another alternative would be including the exogenous information into the model if the information is available. This gives the model more precision and the ability to respond the changes in the information. Proportion of smokers did not have a huge impact on incidence rate while GDP seems to consistently affect the incidence rate. The author's conclusion is that the model which included GDP would be the first choice followed by AR(1) or Random-walk model. Smoking (proportion of smokers in the society) should be considered on individual case basis and there might be better scale to measure smoking other than the proportion of smokers.

INTRODUCTION part

1. Introduction

The number of people who are aged 65 or older is increasing and this phenomenon is observed not only in Japan but among most of the developed countries. According to the report of United Nations Population Division, the number of population who are 80 years of age or older will increase by 8 fold in 2050 compared with 2001, while the total population will increase by two fold. Aging in Japan has been an issue. The total population is expected to go down to 101 million in 2050 compared with 127 million in 2001 but the proportion of people aged 65 years old or older is expected to increase to 33% in 2050. What is behind the growing proportion of elderly to the total population is decreasing fertility rate on the one hand and decreasing mortality on the other. People live longer and elderly gets majority.

"People live longer" causes what it is called "Longevity risk" in the calculation of pension system. Not only in pension but also health care system will be affected by how mortality (or life expectancy) is going. Mortality has been important in future system design, policy making and political decision making. In describing mortality's movement over the years, dynamic model which captures the change of calendar year on mortality has been studied. One of the most widely used in mortality forecast is Lee-Carter method (Lee and Carter 1992). The application of Lee-Carter model (and its extensions) to the mortality data have been seen in empirical analyses. However application of Lee-Carter model to morbidity (disease incidence) has not been found. The reason would be the age-specific disease incidence data in most cases is hard to obtain and if it existed the series would cover too short period of time (calendar years) to apply Lee-Carter model. Fortunately National Cancer Center (NCC) in Japan has the data of age-specific, site of origin specific incidence rate by sex during the years between 1975 and 2004 (30 observations for each age categories, some of the data are good approximate of incidence), which gives us an opportunity to apply Lee-Carter method to incidence data.

The overall construction of the thesis is as follow. The motivation to apply Lee-Carter model to incidence rate data and its meaning are touched now. Small summary of Japanese health care system is also touched in Chapter 1. In Chapter 2 the overall trend of cancer incidence from 1975 onwards is summarized. Methods are written in Chapter 3. From Chapter 4 to the rest of the chapters calculation and comparison of different models were made. In those chapters we can see which model would fit the data well and attempt to predict 2020's picture of cancer incidence rate. The calculation

was done separately by sex and by different site of origin (“All sites together”, “Oral cavity and pharynx", “Esophagus", “Stomach", “Colon", “Rectum", “Liver", “Gallbladder and bile ducts", “Pancreas", “Larynx", “Lung Trachea", “Skin", “Breast", “Uterus (incl. epithelial carcinoma)", “Cervix uteri", “Corpus uteri", “Ovary", “Prostate", “Bladder", “Kidney and other urinary organs", “Brain, nervous system", “Thyroid", “Malignant lymphoma", “Multiple myeloma", “Leukemia", and “Colon and Rectum (lower digestive organ)"). Then after having looked at the results discussion and conclusion will be made about which point is similar and which point is different between sexes or between different organs.

1.1. Research Question

I would like to apply Lee-Carter model to the cancer incidence data in Japan and discuss about its strength and limitation.

1.2. The meaning of forecasting incidence rate

When it comes to health care expenditure, how does aging affect the expenditure of health care? Empirically acute care expenditure is affected mainly by the current GDP level while long-term care is affected not only by GDP but also by societies aging. Health care provision has been difficult subject for system designers. Disease profiles is changing in different period of time and the treatment has progressed. DPC (Diagnostic Procedure Combination) is becoming widely used and the trend has been going toward basket payment system. In basket payment system the principle for payment to the health care provider would be that fixed amount of fee for a certain disease which is decided by the negotiation between health care providers and health care insurance (or government), would be paid for the care of the patient with that diagnosis. Theoretically if the number of patients for a certain disease could be forecasted one could calculate the good estimate of payment or total health care expenditures.

1.3. Japanese Health Care system[1]

Japan is one of welfare states which have universal health care insurance system. Payments for care are on the combination of fee for service basis and basket payment system. Medical payment fee is determined by MHWL (Ministry of Health, Welfare and Labor in Japan) and is revised every two years. There is a book called “payment fee book” in which we can find payment fee for the examination or the procedure. For example suppose a patient goes to a hospital and is examined. The hospital calculates how much of medical services they have given to the patient according to the “payment fee book”. This is the total cost. The patient has to pay the 30% of the total cost to the hospital on site and hospital invoices the insurer the rest of 70%. We are still looking for the ideal combination of fee-for-service and fixed-fee payment systems however basket payment system has been on increase recently.

More details about Japanese health care system can be found in Ando (2009)[2] and in Hirose et. al. (2003).

[1] Some of the explanation of Japanese health care system (in English) refers to Hirose et. al.(2003)
[2] T. Ando Master thesis (2009), Lund.

2. The trend of incidence of cancer over the year (1975-)

Matsuda (2008) calculated cancer incidence based on cancer registries in 11 areas in Japan.

‘’’The Japan Cancer Surveillance Research Group estimated the number of cancer incidences in Japan in 2002 as a part of Monitoring of Cancer Incidence in Japan (MCIJ) on the basis of data collected from 11 population-based cancer registries: Miyagi, Yamagata, Kanagawa, Niigata, Fukui, Shiga, Osaka, Tottori, Okayama, Saga and Nagasaki.’’’[3]

Age specific Incidence for cancer by sex for 1975, 1985, 1990 and 2004 are overlaid in Figure 1 to Figure 47 in appendix figure file in order to see the general trend in the last three decades. The general pattern could be interpreted as that the incidence are highest in 2004, 1995 is in the middle followed by 1985 and it is lowest in 1975. The shape of the curve is quite similar between years for almost all cases. The shape of the curve is unique for each cancer site. But most of the cases the age specific incidence gets higher as time goes by. The incidence has been increasing for both sexes but males' incidence has grown up more than that of females. However the age of takeoff (the age where the incidence goes up sharply) depends on sites. Cancers like uterus cancer or ovary cancer take off at the early 20s. Leukemia's trend is quite similar as the mortality trend (U shaped, not low in very young, stays at low level in the middle and rises in the older age groups again). Prostate cancer takes off at around the age of 50.

METHODS part

3. How to analyze?

The analysis step is **1) to apply Lee-Carter model to cancer incidence data** and **2) to use the estimated coefficients to forecast the incidence rate for 2020**.

Cancer is the name of the disease which cells grow uncontrollably. The nature of cancer is different depending on primary cancer sites. I analyze separately by sex (male and female) and origin sites (“All sites together”, “Oral cavity and pharynx", “Esophagus", “Stomach", “Colon", “Rectum", “Liver", “Gallbladder and bile ducts", “Pancreas", “Larynx", “Lung Trachea", “Skin", “Breast", “Uterus (incl. epithelial carcinoma)", “Cervix uteri", “Corpus uteri", “Ovary", “Prostate", “Bladder", “Kidney and other urinary organs", “Brain, nervous system", “Thyroid", “Malignant lymphoma", “Multiple myeloma", “Leukemia", and “Colon and Rectum (lower digestive organ)"). There is one note regarding breast cancer. Breast cancer is not only for female but it happens for male too. However the number is small and the data does not provide the male breast cancer so that we cannot (do not) run any models for male breast cancer.

All calculations and analyses are performed by combination of SAS, EViews and EXCEL.

3.1. Lee-Carter model

Modeling the relationship between age and mortality has been done by many researchers. Gompertz (1825) assumed the linear increase in log age specific mortality after age 20. Similarly many different expressions have been proposed (According to Mantalos (supervisor) lecture notes of the course “Demographic Forecasting” different extensions can be found in Makeham (1867), Thiele (1872), Wittstein (1883), Pearson (1895)). These mathematical models would be useful but expressing mortality law by one mathematical model is impossible. Heligman and Pollard (1980) proposed applying three different mortality laws in different age periods. Lee-Carter model was developed by Lee and Carter (1992) and has been used in forecasting mortality. This model is a non-parametric model which does not assume any

[3] Matsuda (2008) p.641

mortality laws. The model description is explained as follows.

Age specific mortality (incidence) of age x in year (calendar year) t while m_{xt} will be expressed as the sum of mean age specific mortality a_x and $b_x * k_t$.

$$\ln(m_{xt}) = a_x + b_x * k_t + e_{xt} \qquad (1)$$

The interpretation of k_t and b_x are "time trend" and "how much it changes from the mean mortality for age x when k_t changes", respectively. In order to have unique estimates we need the following condition.

$$\sum b_x = 1, \qquad \sum k_t = 0$$

Estimation can be done by Singular Value Decomposition, Maximum Likelihood Estimation or Weighted Least Square. In Lee-Carter model calendar year effect on mortality is described only by k_t and in order for us to forecast the mortality what we need is the expected k_t and use of those forecasted k_t we would be able to calculate the mortality by combining a_x, b_x which are independent of calendar time. Original Lee-Carter model assumes that k_t series is random walk (unit root process).

Lee-Carter model can be interpreted as extracting the primary component in principal component analysis. One can extend the Lee-Carter model by putting more than one principal component.

$$\ln(m_{xt}) = a_x + \sum_{i=1}^{p} b_x{}^{(i)} * k_t{}^{(i)} + e_{xt} \qquad (2)$$

Other extension can be assuming non-normal distribution to the error terms. Kogure (2005) analyzed Japanese mortality by Lee-Carter model with Poisson distribution to deal with the skewed error distribution.

As we see age specific incidence rate by sex and the site of origin, the incidence rate is taking off differently between the origins of tumors (ex. Prostate cancer starts to appear in the age groups 50s or 60s while uterus cancer comes up in age 20s). If incidence rate is zero, Lee-carter model cannot be applied because incidence cannot be logged. In the first step which age groups should be considered meaningful or practically doable is checked and decided. In other words for some cancer I would apply Lee-carter model only to those who are 30 yrs old or older, for the other cancer only 40 yrs old or older would be chosen. It depends on which type of cancer is analyzed. For example, prostate cancer starts to take off around age 50 yrs old so that it would not be inappropriate to say zero as the forecasted incidence rate in 2020 for those who are younger than 50 yrs old. Also I checked whether or not it seems reasonable to take into account only the first principal component (i=1 in the formula (2)). For the purpose of checking the validity to use of only first principal component, I put tables on the appendix file in which we can see which age group was used and how much of variability explained by first and second principal component from Lee-Carter model (Table 1a and Table 1b).

3.2. How we can estimate the incidence for 2020?

Parameters a_x, b_x in Lee-Careter model do not have any information about time and they are the same for 2020. Once we get $\hat{k}_{2020}$ we can forecast the incidence rate for 2020 by using the formula (1). So what we need is $\hat{k}_{2020}$. In other words question would be what kind of model would be fit for k_t. Original Lee-Carter model assumes random walk for k_t series. k_t seems to be Random-walk based on series of unit root tests for any type of cancer. However doubt must be casted on assuming just random walk. First reason is that sample size (n=30) is small. Unit root tests must be rejected to be concluded as stationary and small sample size tends to fail rejecting the null. It might be the case where the series is quite close to random walk but not exactly random walk. Therefore the strategy would be applying different

models to k_t and compare them. Three types of models are considered.

3.2.1. ***Random walk: (Model 1)***

If k_t series is random walk it is the same as original Lee-Carter model.

The model would be

$$\begin{aligned} k_t &= \alpha_0 + k_{t-1} + e_t \\ e_t &\sim N(0, \sigma^2) \end{aligned} \qquad (3)$$

I call the model "Random walk" or "model 1" in the following text.

3.2.2. ***AR(1) (Auto Regressive 1): (Model 2)***

Although it looks unit root, k_t might be Auto regressive one (AR(1)) with the coefficient close to one. One could think of AR(1) as an alternative model to "Random Walk" which would be

$$\begin{aligned} k_t &= \alpha_0 + \alpha_1 * k_{t-1} + e_t \\ e_t &\sim N(0, \sigma^2) \end{aligned} \qquad (4)$$

I call the model "AR(1) model" or "model 2" in the following text.

3.2.3. ***COVARIATE model: (Model 3)***

If we had exogenous variables which may explain the time trend k_t, it would be better to use those information to forecast. Following two factors were considered in the thesis. One is GDP per capita and the other is smoking, so that the model would be

$$k_t = f(GDP, Smoking) + error$$

Here are the explanations for GDP and smoking.

① GDP per capita

GDP is a basic measure of the country's overall economic output. It is an annual market value of all final goods and services made within the country. It is often used as a proxy of living standard and has been regarded as exogenous factors. For example Murry (1997d) included GDP as one of explanatory variables in order to calculate the burden of disease in his series of Global Burden of Disease Study (GBD) (Murry 1997a, 1997b, 1998c). However this is the cross country comparison setting. In that setting GDP means how much difference in countries is observed economically. In my thesis it is longitudinal data in one country and different meaning is put on GDP. That is how GDP changes in a country affect the cancer incidence in the long run. We can think of two ways of explanation about GDP to cancer incidence. The first one is GDP as an indicator of longevity. Cancer incidence is related to aging and there is a strong relationship between the life expectancy at birth and GDP level. GDP data is available from 1950 onward in our dataset and we can lag 25 years for GDP data if necessary. The second explanation is GDP as a material of policy discussion. OECD (2006)[4] decomposed growth in public health spending using data between 1981 and 2002 and showed that the growth rate in health spending among OECD countries is 3.6%, out of which GDP accounts for 2.3%, aging accounts for 0.3% and the residual (technological

[4] P.32, "Table 2.1 Decomposing growth in public health spending, 1981-2002"

change or inflation) accounts for 1.0%. If GDP is going up the new screening measure might be implemented to detect new cancer. In that case incidence rate goes up. If the economy is not booming the policy will be in the direction of containing health care expenditure and that may affect the cancer detection. For new policy to be implemented it would take about five years (having looked at the current economic situation and catching what needs to be done. It takes a year or more for the bills to put into act). It is known the high correlation is observed between lagged GDP growth (or the average of GDP growth for the last five yrs) and health care expenditures (ex. Kenjo (2005)[5]). I could draw the following implications for GDP to the incidence. Lag might be essential and just taking one year is not enough. It has to be average for a certain length of time because most of the countries health care expenditures are controlled by public (one could imagine the money paid by the public to the health care sector including money for research and new technology as well as health care providers may affect the incidence of cancer) .Which time span should be included? As I mentioned 25 years is the maximum to calculate in the dataset. However the emphasis must be put on the closer past. So I would take the average GDP for the past 5 years as Getzen (1995)[6] did and use it as a potential explanatory variable. In the following text I call it just GDP. But its meaning is the latest five years' average of GDP.

② Smoking

Smoking has been proved to be the cause of cancer and smoking cessation programs are everywhere in the developed countries. For individual level smoking should have an effect on cancer. However its effect on the macro levels might be a different story. What kind of indicator could it be suitable? One could think of price of tobacco. However cigarette is cheap in Japan. So tobacco's total sales in each year may not capture the effect of tobacco on cancer incidence. Besides, the sales data which ranged wide enough for the analysis could not be found. Therefore I chose the smoker's rate (proportion of smokers in the society) instead. Age-specific smoker's rate and total smoker's rate by sex are obtainable from two organizations, Ministry of Health, Welfare and Labor (MHWL) and Japan Tobacco Inc. (JT). Both parties conducted surveys independently. The year range and categorization of age is different between organizations. I am using the data from JT website. Smoking rates are different between male and female significantly but smoking rates are similar between different age groups in each sex. So I use (calendar) year specific smoking rates for all male in the analysis of male cancers and female data for the analysis of female cancer. Smoking is controlled in Global Burden of Disease Study as well in Murray (1997c). The data from JT ranges from 1965 to 2004. Smoking is said to have life-long effects. However putting more than one annual smokers' proportion into the regression would cause colinearity problem. In epidemiological studies pack year (calculated by packs smoked per day multiplied by years as a smoker) is used for smoking. What could be the closest to the concept of pack year? In the thesis smoking rate data is available from 1965 (the incidence data is from 1975). Among the options available average of smoking rate for the previous 10 years will be the most close to the pack year concept and is used in the analysis. In the following text I call it just smoking. But it is the last ten years' average of the proportion of people who are smokers. Its meaning is how much proportion of people is at higher risk of cancer. So its meaning would be different from what we would think of smoking as the individual person's risk of cancer.

[5] Kenjo (2005) p.194 Fig 1. and p.195 Fig 2. Original source is Getzen (1995) "Macroeconomics and Health Care Spending" in J.M. Pogodzinski ed., *Readings in Public Policy*, Oxford: Blackwell

[6] Getzen (1995) s tables which the author referred to can be found in Kenjo (2005).

Therefore the model would be

$$\begin{aligned} k_t &= \alpha_0 + \alpha_1 * GDP + \alpha_2 * Smoking + e_t \\ e_t &\sim N(0, \sigma^2) \end{aligned} \qquad (5a)$$

However one would expect the existence of autocorrelation (of order n), so that we would modify the model with Cochrane-Ocurrt method.

In the case of order one, the modified model would be

$$\begin{aligned} k_t &= \alpha_0 + \alpha_1 * GDP + \alpha_2 * Smoking + e_t \\ e_t &= \rho * e_{t-1} + \nu_t \\ \nu_t &\sim N(0, \sigma^2) \end{aligned} \qquad (5b)$$

I call the model "COVARIATE model" or "model 3" in the following text.

In EViews the procedure is quite simple which is adding AR(1) in the space for explanatory variables.

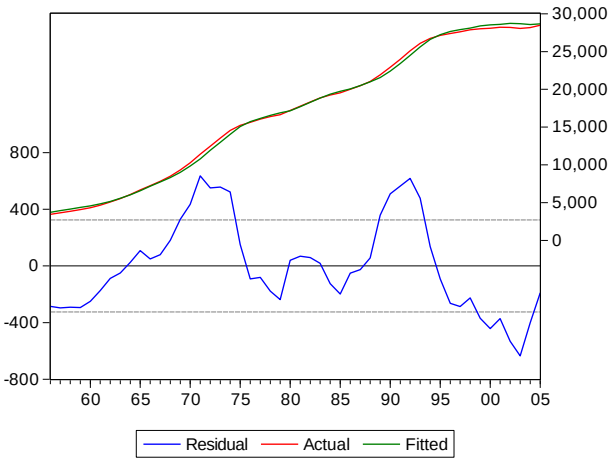
We have to watch out for ending up spurious regression when running model 3. If all variables are stationary it won't be spurious regression. If all variables are Random-walks (unit root processes), if there is a co-integration relationship which can express in model 3, the results from model 3 will not be spurious, either. However in other cases there is a risk to end up running spurious regression. GDP is known to be very close unit root process. It was the same story for smoking. The series seems to be random-walk but not conclusive. Usual procedure is to check the stationarity first and analyze the data after doing appropriate transformation. But our strategy is to run the model 3 and check whether or not the coefficients are small, the R squared is large, the t-value is large and DW (Durbin-Watson statistic) is small, all of which are observed when running spurious regressions. If none of these indications could be found I would assume it is not spurious.

The number of observations is 30, so that lower and upper bound for DW statistic are 1.28 and 1.57 in the case of two explanatory variables. These numbers are used roughly as a reference to see whether or not DW is small.

As a rule of thumb 5% significance level has been widely used for studies all over the world. However both GDP and smoking are known to be related to cancer and the number of observations is only 30 for the regression of k_t. My expectation about GDP is exogenous factor while for smoking I am not sure it is exogenous factor to the cancer incidence. So I do not specify the exact significance level but roughly 15% was used as my guidance to make sure all necessary variables are included in the final model.

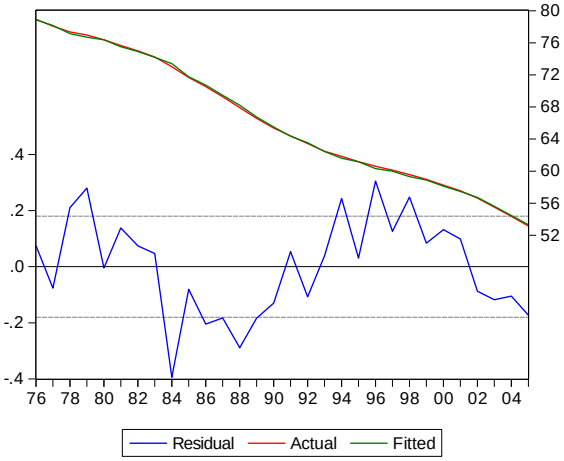
3.3. In COVARIATE model we need GDP_{2020} and $Smoking_{2020}$ to calculate k_{2020}

We need a plausible values for GDP and for smoking up to 2020 to forecast k_t up to k_{2020}. For simplicity I assume Random walk for GDP series. Actual fit was quite well enough (though residual curve showed that there seems to be some systematic movement left in the residual). Since the error size seems small compared with the actual GDP size and it would not be bad estimates.

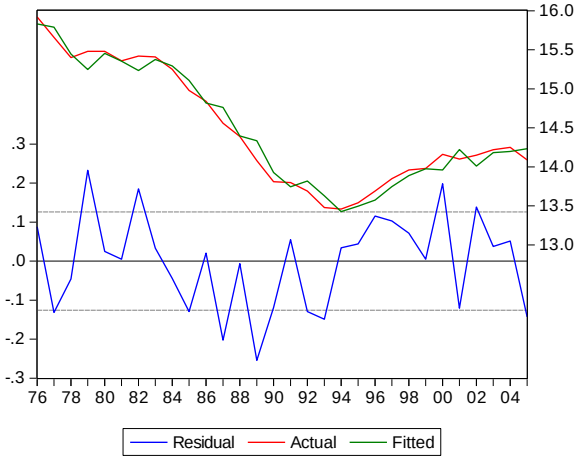


For smoking I assume ARMA(1,1) model and forecast smoking series up to 2020.

Smoking for men



Smoking for women



ARMA means stationary and residual must be white noise theoretically. Residual series for both sexes seem that there is a small trend left in the residual (actual values in the 1980s have less than estimated values). But overall fit seems quite reasonable.

RESULTS part

4. Cancer (over all; all sites together): ICD10(C00-C96 D05-D06)

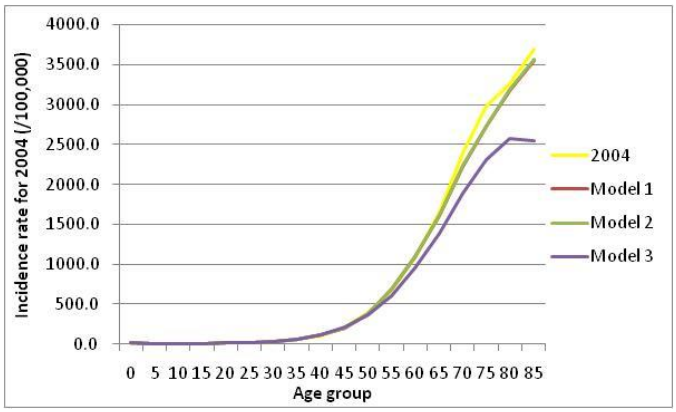
4.1. Male

Estimated formulas for each model are shown below.

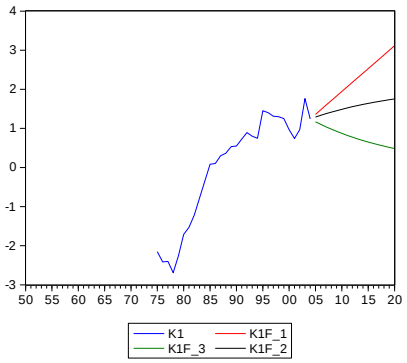
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.117 + k_{t-1}$	0.947
AR(1) model	$E(k_t) = 0.115 + 0.946k_{t-1}$	0.949
COVARIATE model	$E(k_t) = 0.943k_{t-1}$	0.943

Note : In COVARIATE model the GDP and smoking do not have significant explanatory power

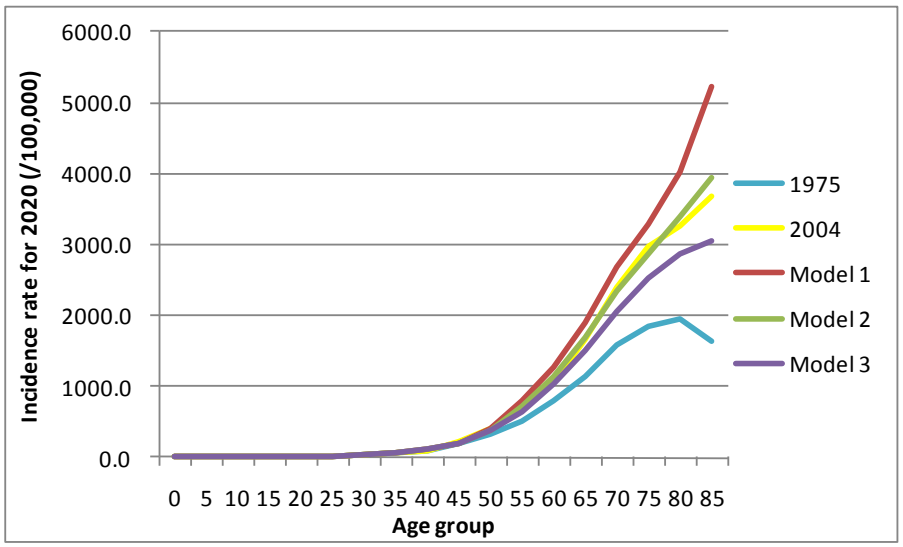
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



It doesn't matter which model we use the results are almost the same. The difference in the Adjusted R^2 is on the third decimal. Comparison with the actual data and estimated data for 2004 shows that both Random walk and AR(1) fit quite well while COVARIATE model (the model has no covariates this time) underestimates the actual data. COVARIATE model turned out to be the one with no exogenous factors included the model. The leading cause has changed from stomach to lung cancer during the period, so that it would be reasonable that exogenous factors were not chosen. Based on AR(1) the incidence for 2020 stays in the same level as 2004. Based on Random walk the incidence will increase. I would choose Random walk because it was chosen for female and it seems reasonable to assume increase in incidence.

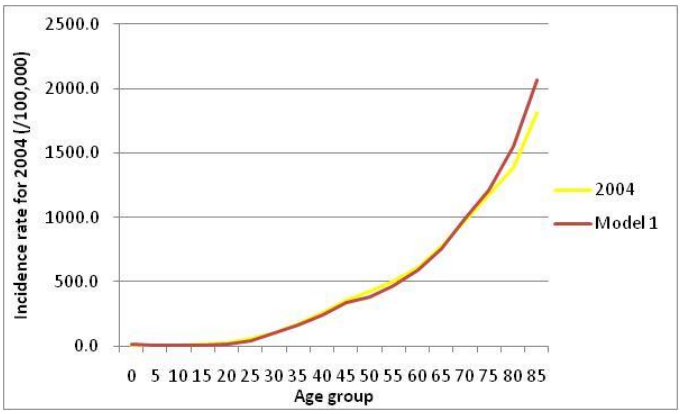
4.2. Female

Estimated formulas for each model are shown below.

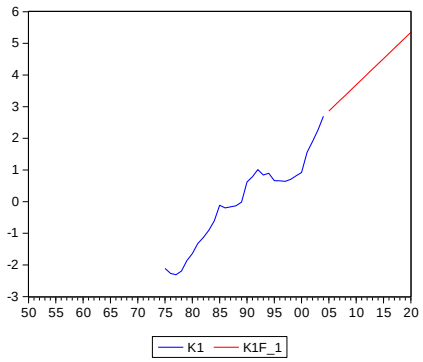
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.166 + k_{t-1}$	0.973
AR(1) model	$E(k_t) = 0.167 + 1.01 * k_{t-1}$	Identical to unit root
COVARIATE model	$E(k_t) = 1.01 k_{t-1}$	Identical to unit root

Note: Estimated AR(1) model or COVARIATE model are not autoregressive.

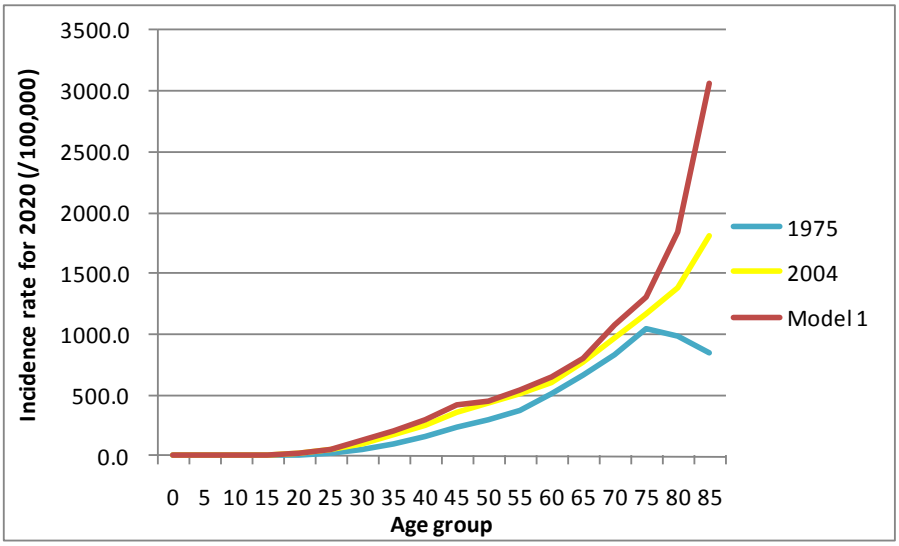
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Random walk model is the only model left. Actual data and estimated value is quite close for 2004. Time trend parameter k_t seems to be the natural extrapolation of the current dataset. For male it may be AR(1) which is very close to Random walk (coefficient=0.95). So it seems not contradicting that women's path is random walk (non-stationary). Cancer patients will become more common according to our forecast.

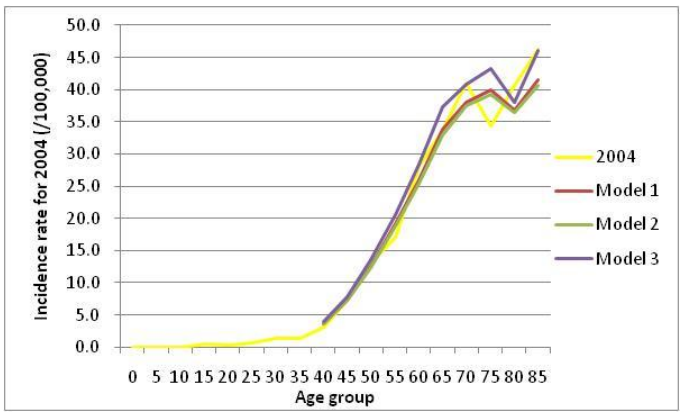
5. Oral cavity and pharynx: ICD10(C00-C14)

5.1. Male

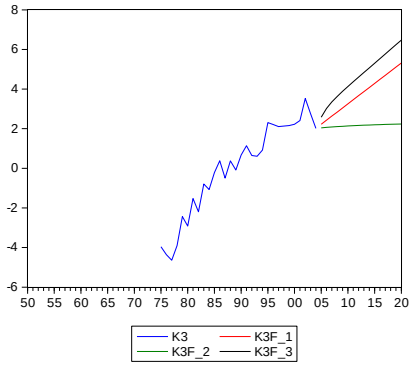
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.206 + k_{t-1}$	0.896
AR(1) model	$E(k_t) = 0.200 + 0.913k_{t-1}$	0.900
COVARIATE model	$E(k_t) = 0.000279 * GDP - 0.0938 * Smoking,$ $E(\varepsilon_t) = 0.523 * \varepsilon_{t-1}$	0.913

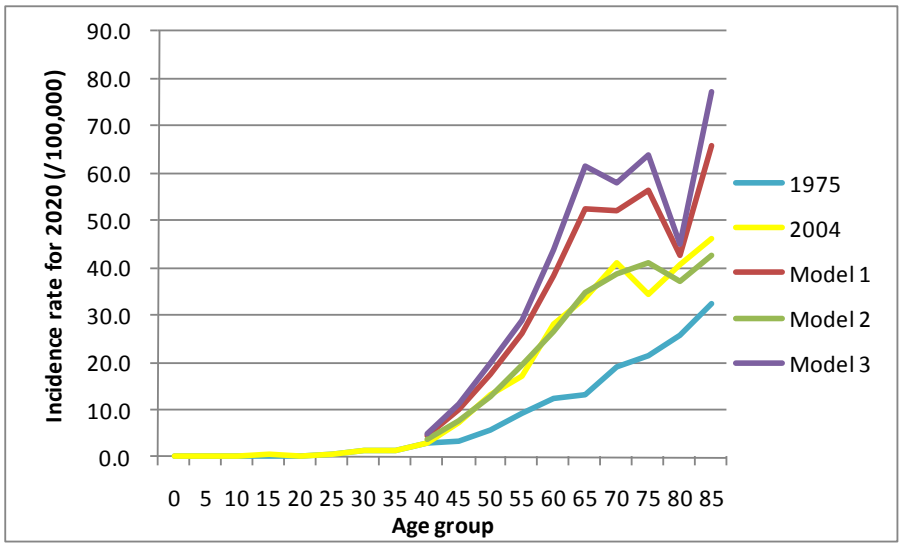
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Comparison with actual values for 2004 showed COVARITE fits better for 80 yrs or older while AR(1) and Random walk fit better for the age group less than 70 yrs old. R squared is around 90% for all three models. COVARIATE model has GDP and smoking left. What puzzles us is negative coefficient in smoking. But our smoking variable is not whether or not the individual is a smoker but the proportion of smoking population in a society. Our smoking parameter might be looking something different. For GDP as we expected the coefficient was positive. Because of its highest adjusted R squared and its capability to capture the change in GDP I would choose COVARIATE model. If exogenous information is not available it would not be easy to choose one model. For female case I chose AR(1) so I would choose AR(1) in order to be consistent between sexes.

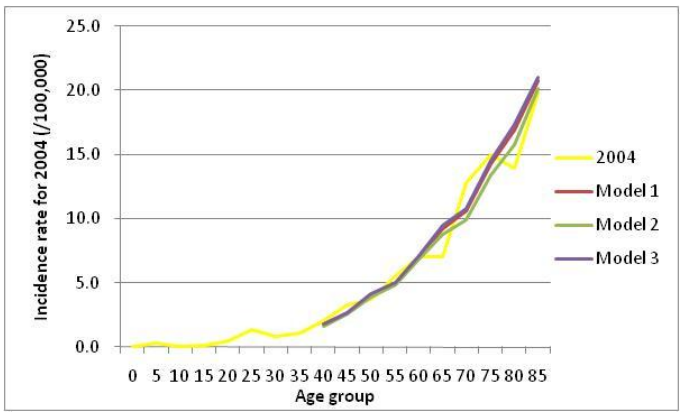
5.2. Female

Estimated formulas for each model are shown below.

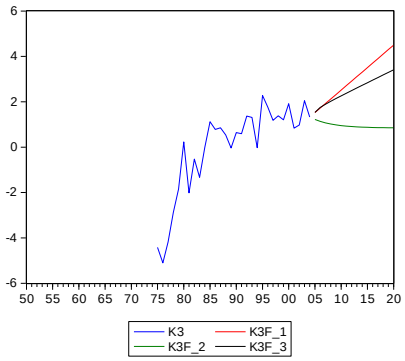
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.198 + k_{t-1}$	0.665
AR(1) model	$E(k_t) = 0.188 + 0.777k_{t-1}$	0.713
COVARIATE model	$E(k_t) = 0.000221 * GDP - 0.332365 * Smoking,$ $E(\varepsilon_t) = 0.536 * \varepsilon_{t-1}$	0.732

Note: Estimated AR(1) model is not autoregressive.

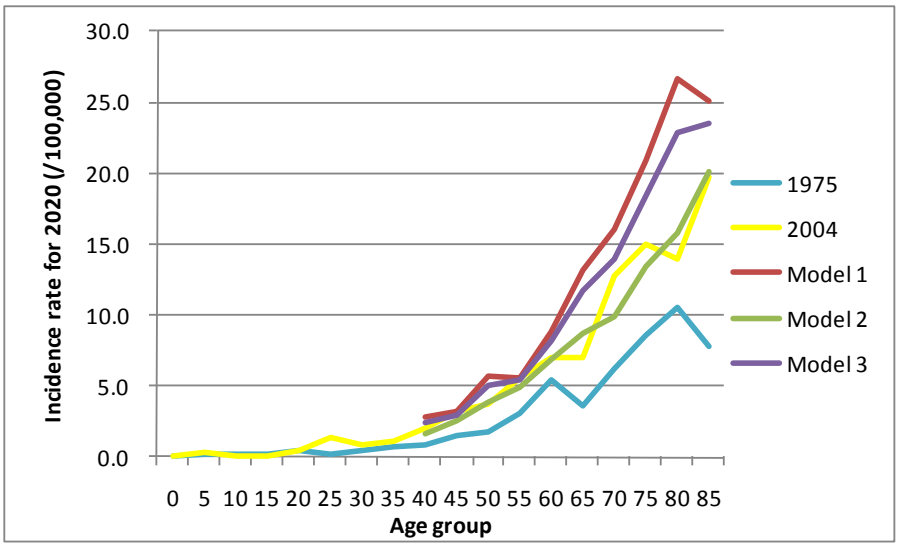
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Since the coefficient of AR(1) was 0.78 and Random walk model seems not good model. Whatever model we choose estimates for 2004 are quite close to the actual data of that year. COVARIATE model has the highest R squared (73%). The estimate for GDP is positive and smoking is negative. The sign is consistent as for male. We need some explanation for negative smoking, which I will write in Discussion part.

I would choose COVARIATE model because of its fit and its capability to capture the change in GDP and smoking. If exogenous information is not available I would choose AR(1) based on R squared and the coefficient of 0.78.

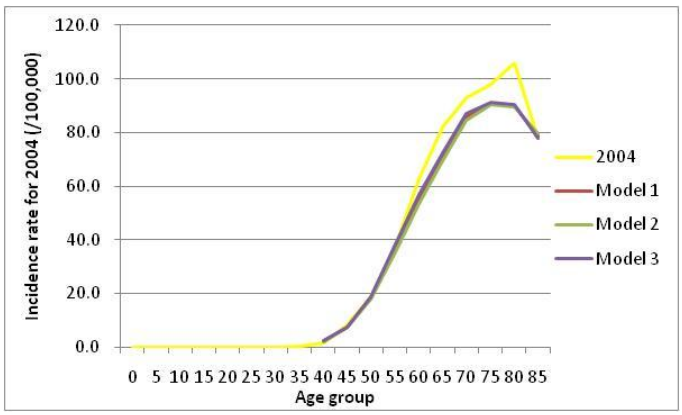
6. Esophagus: ICD10(C15)

6.1. Male

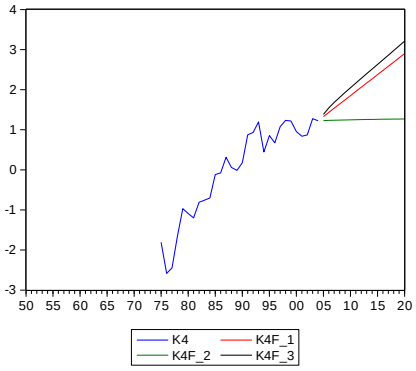
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.105 + k_{t-1}$	0.887
AR(1) model	$E(k_t) = 0.101 + 0.921k_{t-1}$	0.889
COVARIATE model	$E(k_t) = 0.000139 * GDP - 0.047 * Smoking,$ $E(\varepsilon_t) = 0.593 * \varepsilon_{t-1}$	0.903

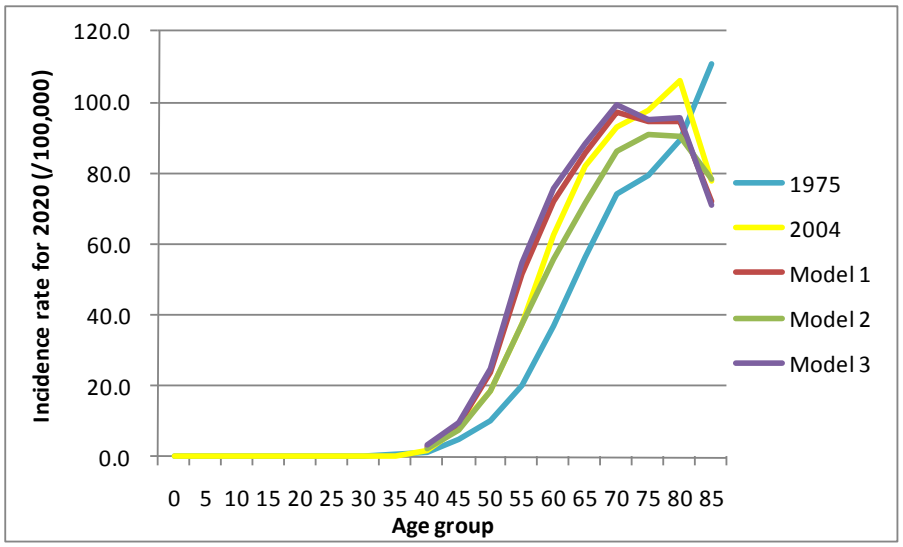
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



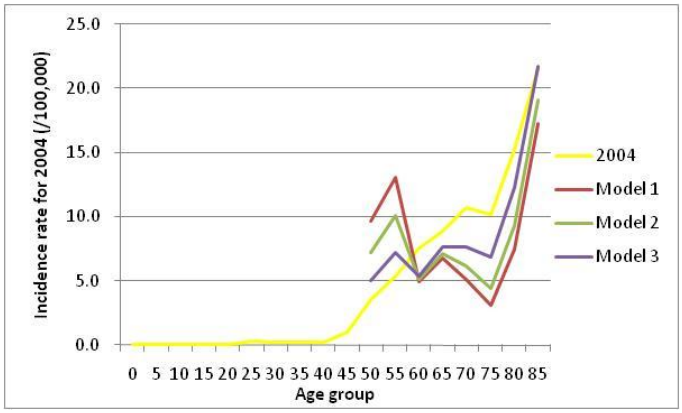
Fit is highest in COVARIATE model. As most of cancers with other sites of origin, the sign of estimate for GDP is positive and negative for smoking. Comparison with 2004 actual data shows all three models underestimate for the age groups 60yrs old or older. COVARIATE model can capture the exogenous factors of GDP and smoking, and the movement toward 2020 is between random walk and AR(1). I would choose COVARIATE model based on its highest R squared and its capability to capture the change in exogenous factors. If exogenous variable is not available I would not choose one model. Both could be plausible.

6.2. Female

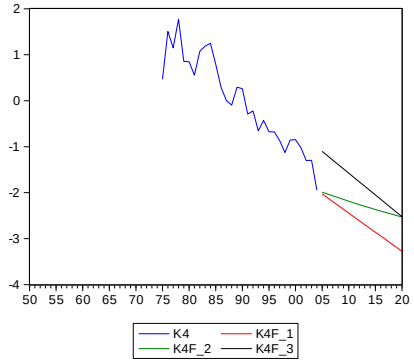
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.0832 + k_{t-1}$	0.827
AR(1) model	$E(k_t) = -0.0818 + 0.980k_{t-1}$	0.821
COVARIATE model	$E(k_t) = 4.234930 - 0.000187 * GDP$	0.848

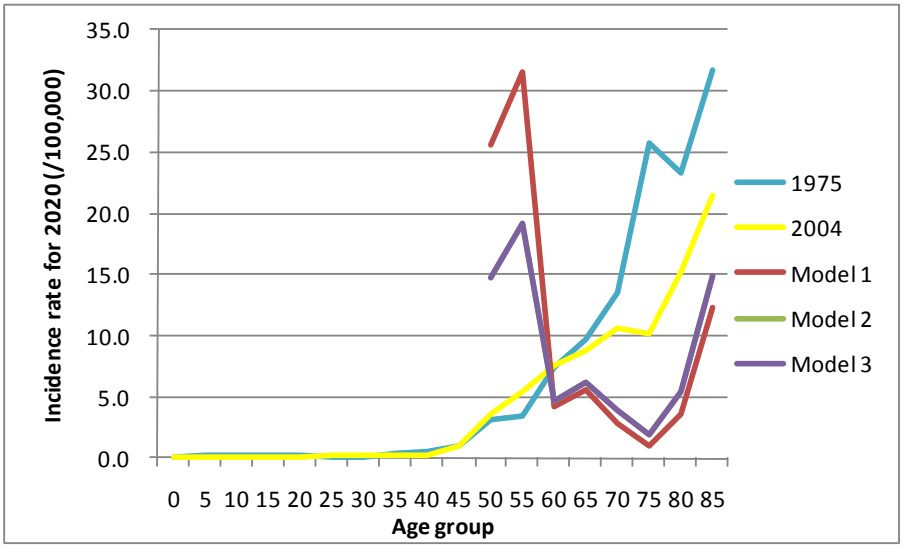
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



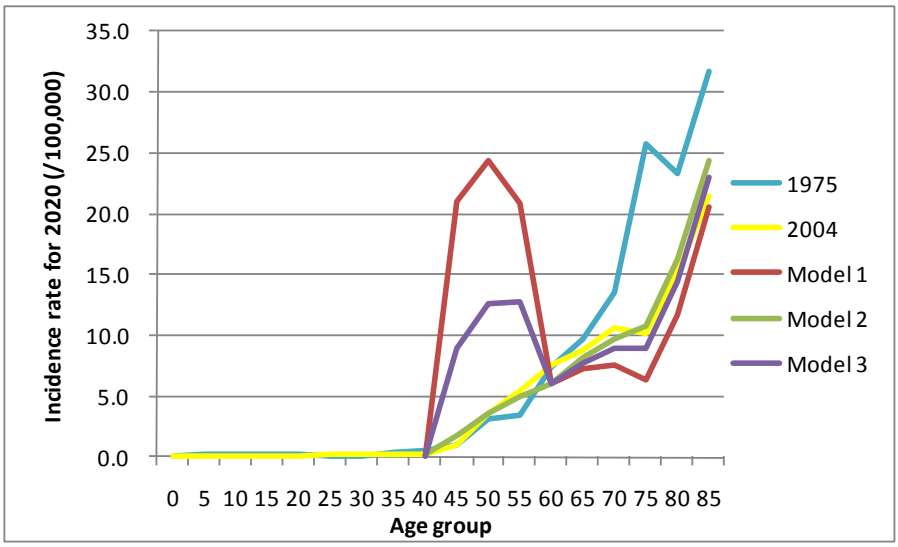
As can be seen in the figure which compares actual and estimated data for 2004, none of the model fits well. There is a big gap in shape between actual and estimated values. The best fit was COVARIATE model however residual plot and DW of 1.21 suggest that doubt would be casted on COVARIATE model. The fact that AR structure in the residual was turned out to be unnecessary also cast doubt on the correctness of COVARIATE model. Because of no AR structure in the error of COVARIATE model there is a gap between k_t in current dataset and forecastes k_t from 2005 to 2020. It looks funny for forecasted incidence rate for 2020. All models show peculiar shape of big jump in the middle age. When looking at the original data the incidence for people younger than 65 has a u-shape curve (both 1975 and 2004 are high. The trough was in 1980s). However the incidence for older people has a consistent decreasing trend. In other words Lee-Carter model tries to capture the u-shape trend for middle age while decreasing trend for old age. This is definitely

the case that time trend parameter cannot capture the whole scenario.

In Lee-Carter model we log the incidence rate and run the model. Logging means the distance between 0.1 and 0.4 is the same as the distance between 100 and 400. The incidence of 0.1 to 0.4 per 100,000 does not mean much in public health perspective (and it is easy to happen because of its huge variance) while 100 to 200 means a lot while the in log scale the latter has one half difference than the former. So in this specific situation the incidence in the middle age group during 1975 and 2004 changed little while it gets twice as much in ratio. Therefore it is expressed in huge increase in those age groups in 2020's forecast.

However if we look at the unit scale of incidence we will notice that the rate itself is one forth the level of male. Although the shape is very strange the mass number is still small. As long as incidence level is low the shape does not matter from the policy maker's perspective because the actual number is what matters.

If we run the Lee-Carter model separately to the age group 40-60 and the age group 65-85, we would get the following figure.



Lee-Carter model is non-parametric model and basically the shape does not differ much when doing separately. However it seemed to reduce the effect of trend in the middle age group to the older age group. The conclusion is the same and we could not believe the shape for the middle age group and the incidence itself is not so large that we don't have to pay much attention to the shape from the public health policy's perspective.

We were not able to choose any model based on data. But based on going back to the original Lee-Carter (Random walk). Random walk would be nice to give pessimistic scenario.

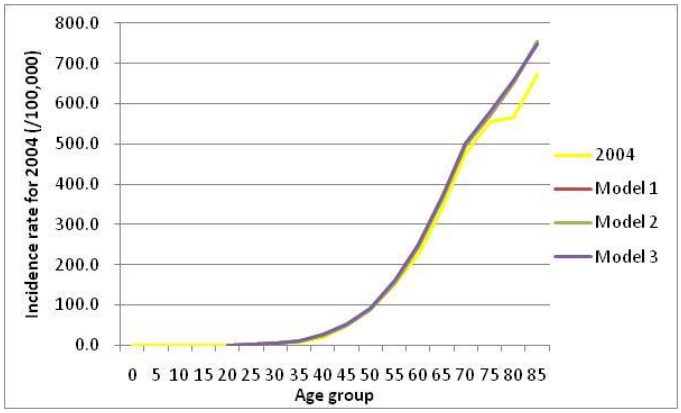
7. Stomach: ICD10(C16)

7.1. Male

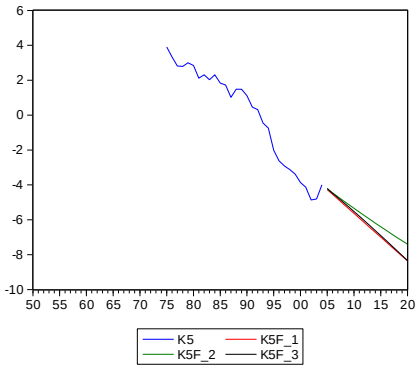
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.272 + k_{t-1}$	0.976
AR(1) model	$E(k_t) = -0.271 + 0.990 k_{t-1}$	0.975
COVARIATE model	$E(k_t) = -0.000357 * GDP + 0.119 * Smoking,$ $E(\varepsilon_t) = 0.881 * \varepsilon_{t-1}$	0.974

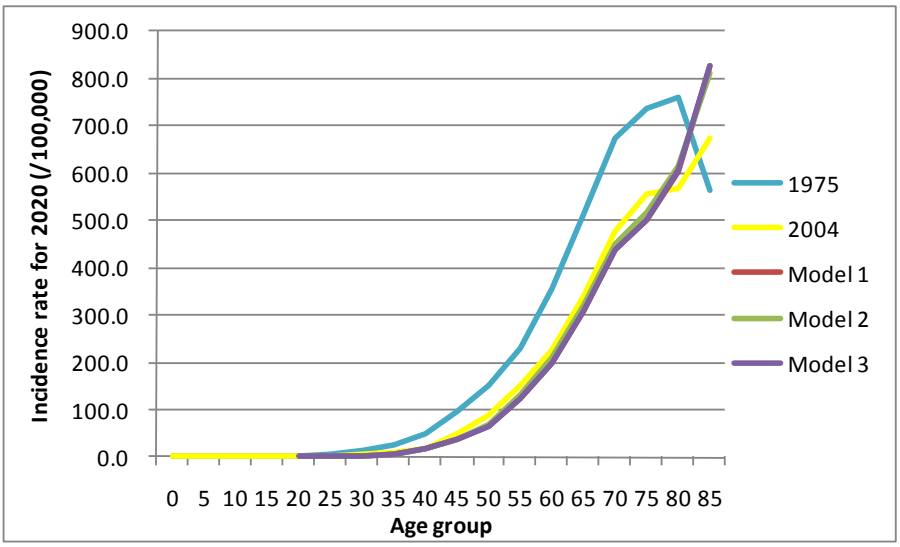
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



The difference in R squared is on the three decimal level and all have more than 97%. All models have almost the same value for 2004 estimated incidence rate, which is consistently bit overestimation. And all three are almost the same in the trajectory of k_t to 2020.Whatever model we use the forecast will be the same. Stomach cancer's incidence is expected to be the same level as for 2004. I would choose Random walk because of its fit in R squared and its simplicity and its consistency with Lee-Carter model in the case of mortality. It has decreased but be tapering off from now on according to our forecast.

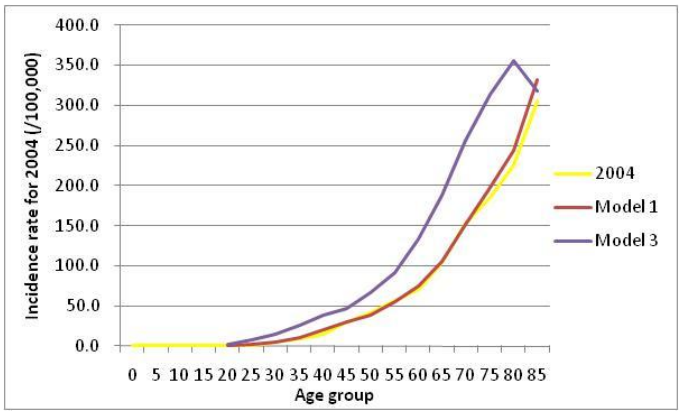
7.2. Female

Estimated formulas for each model are shown below.

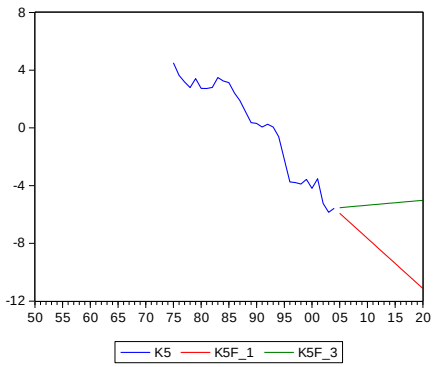
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.347 + k_{t-1}$	0.963
AR(1) model	$E(k_t) = -0.348 + 1.0004k_{t-1}$	Identical to unit root
COVARIATE model	$E(k_t) = 0.994k_{t-1}$	0.951

Note: Estimated AR(1) model is not autoregressive.

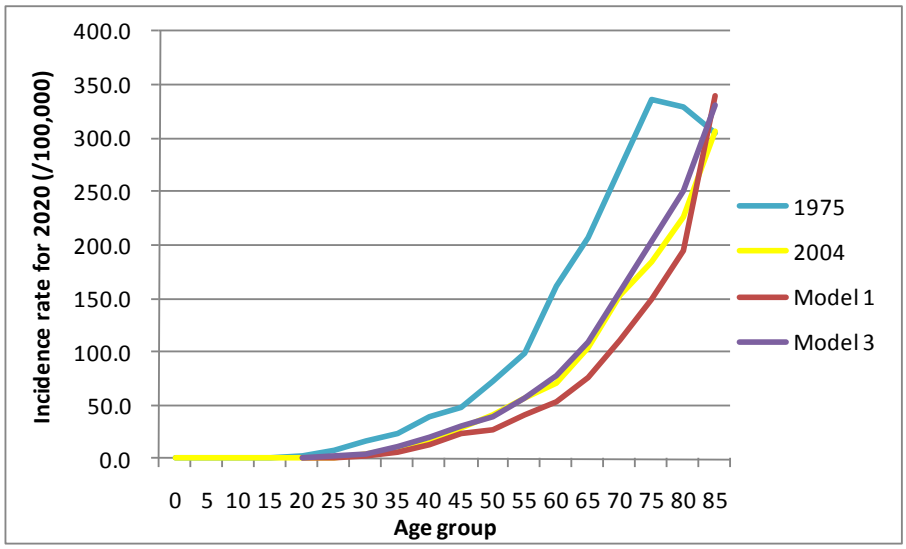
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



In COVARIATE model there is no exogenous factors left. Although R squared was as high as Random walk, COVARIATE model has huge divergence from actual data for 2004. During the period between 1975 and 2004 huge decrease was observed and if the current trend continues it should be at most the 2004 level (COVARIATE model) and hopefully went down more (Random walk). I would choose Random walk but it is an optimistic scenario (I hope it would happen).

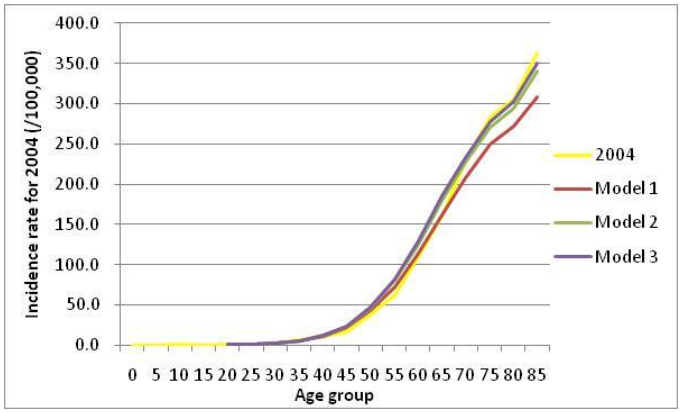
8. Colon: ICD10(C18)

8.1. Male

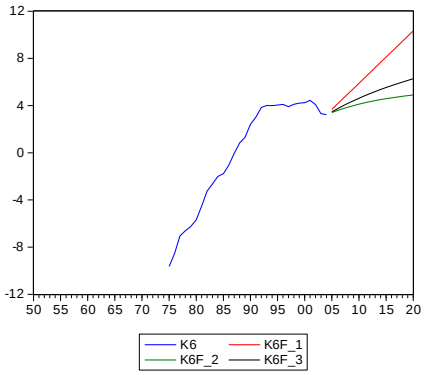
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.444 + k_{t-1}$	0.985
AR(1) model	$E(k_t) = 0.435 + 0.921k_{t-1}$	0.992
COVARIATE model	$E(k_t) = 0.000187 * GDP,$ $E(\varepsilon_t) = 0.913 * \varepsilon_{t-1}$	0.993

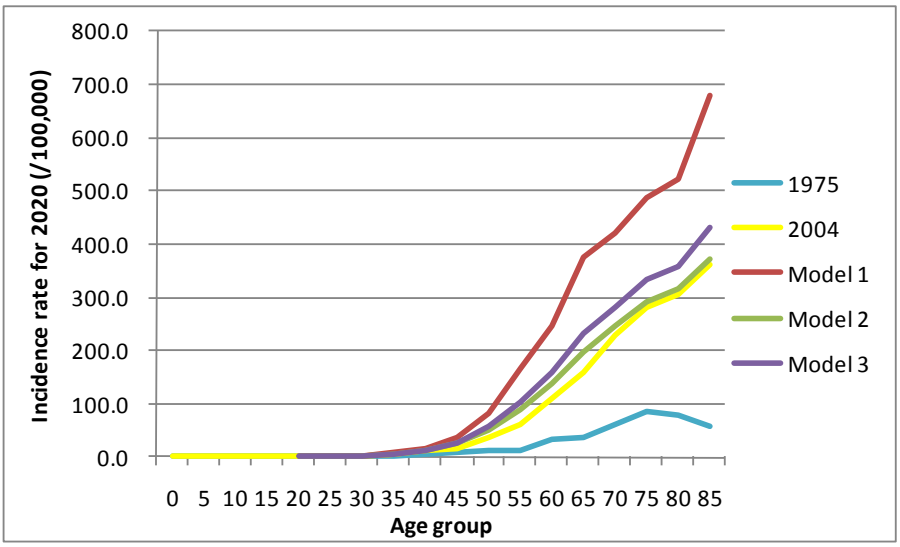
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



R squared for every model shows quite high. So basically k_t of the dataset explains itself (endogenously).

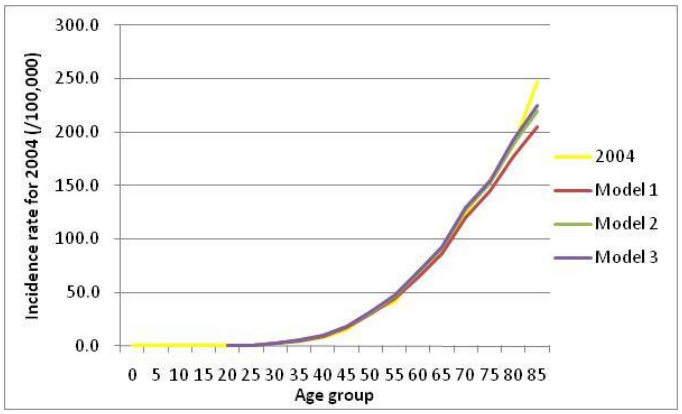
AR(1) and COVARIATE models have better fit to 2004 actual data. Different models show different trajectory for the future k_t. I would choose COVARIATE model because of its ability to control GDP. It has been said that the number of colon cancer is increasing as the lifestyle has been westernized. Westernization in food is related to and coincides the GDP growth. So it would be reasonable to think that GDP influence the incidence so that model can capture the GDP change in the future, which is the advantage for the model. However it would have been better if we had included lifestyle change parameter (if available).

8.2. Female

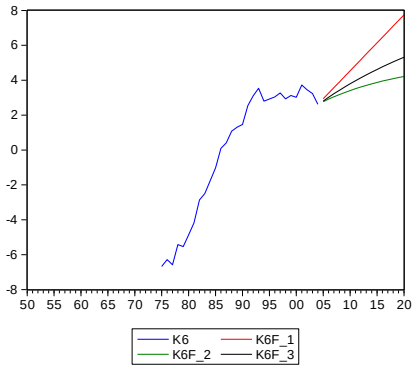
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.320 + k_{t-1}$	0.977
AR(1) model	$E(k_t) = 0.315 + 0.938k_{t-1}$	0.981
COVARIATE model	$E(k_t) = 0.000166 * GDP,$ $E(\varepsilon_t) = 0.933 * \varepsilon_{t-1}$	0.981

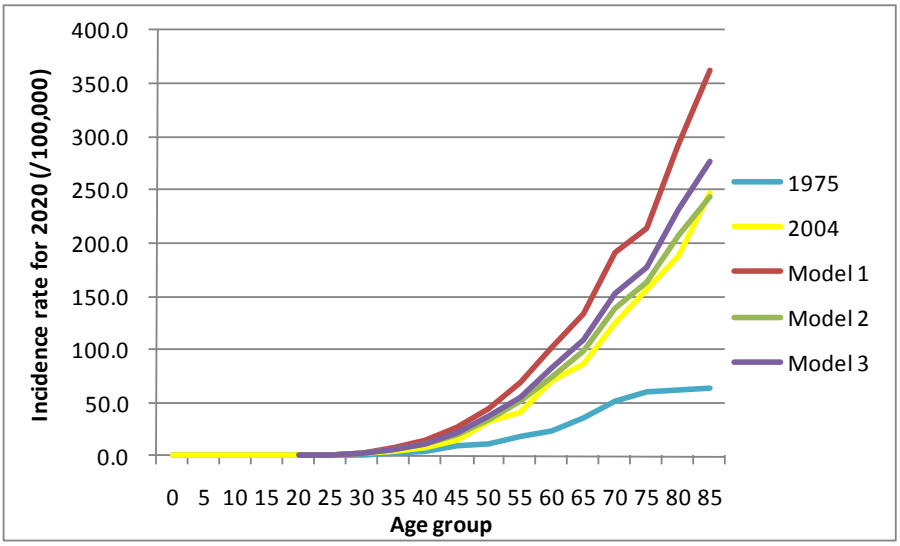
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



The result for female is quite similar with the result for male. Regardless of the choice of model almost 98% is explained. 2004 data would be estimated accurately and precisely by any model. As in the case of male I would choose COVARIATE model (model 3) because of its flexibility to the change in GDP. COVARIATE model might have been better if we had included lifestyle change parameter in the model however I doubt this available. If exogenous variable is not available I would not choose one but would conclude that it will be between Random walk and AR(1). AR(1) has better in fit but it would not be easy to buy the idea that the future colon cancer stays in the same level as for 2004.

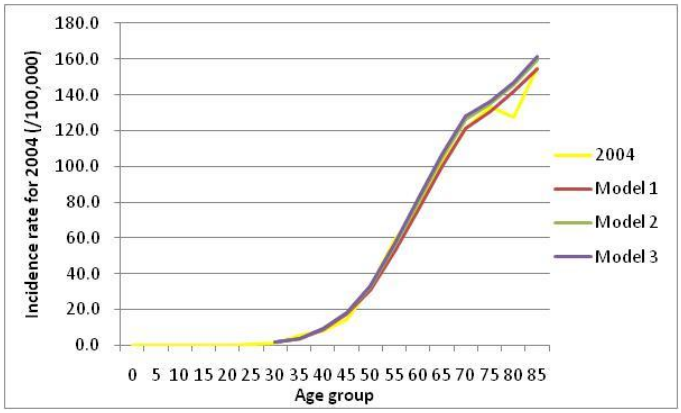
9. Rectum: ICD10(C19-C21)

9.1. Male

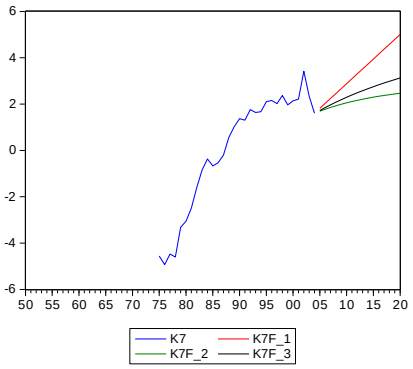
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.212 + k_{t-1}$	0.952
AR(1) model	$E(k_t) = 0.208 + 0.926k_{t-1}$	0.957
COVARIATE model	$E(k_t) = 0.000094 * GDP,$ $E(\varepsilon_t) = 0.917 * \varepsilon_{t-1}$	0.957

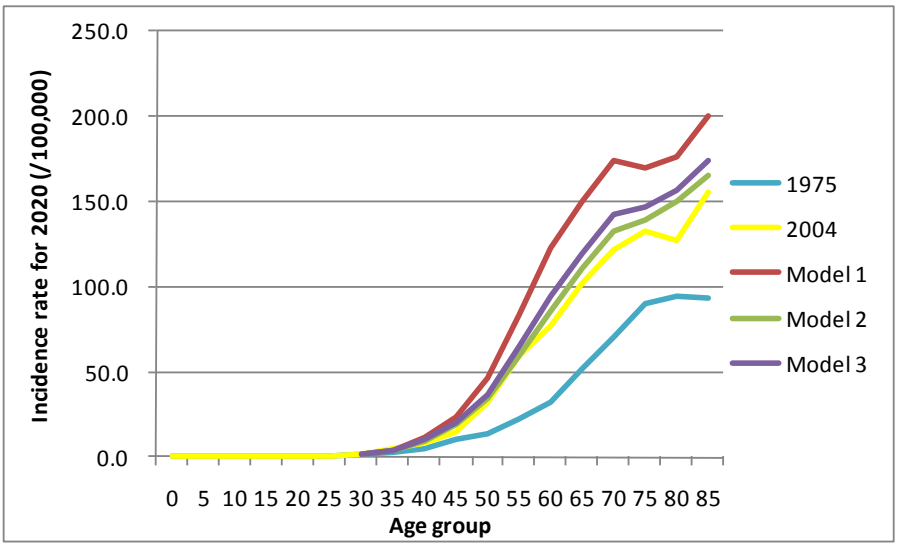
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



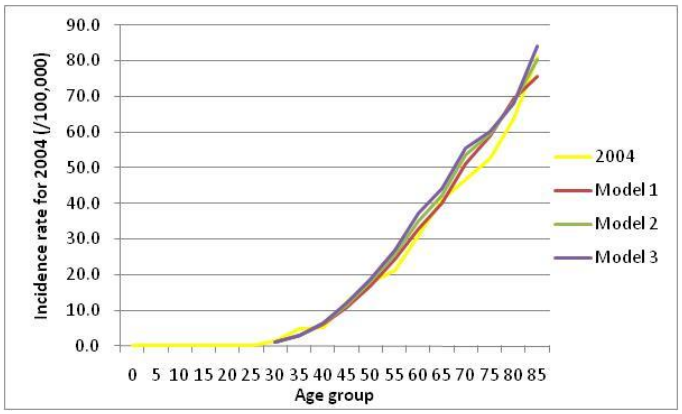
R squared is same for all models and the fit to the actual data for 2004 is same as well. The trajectory of k_t is plausible for all models. I could not choose based on data. The choice should be based on objective in this case. From policy making perspective model 1 (Random Walk) would be chosen because it may provide the pessimistic scenario. COVARIATE model which can capture the change in GDP might have some advantage over other two methods. I would choose COVARIATE model. As discussed in colon cancer rectum cancer is also related to westernized life styles, which attracts us to include GDP (exogenous factor) into the model. Only GDP is included in the model and it is consistent with colon cancer case. If exogenous information is not available I would not choose one because both are plausible.

9.2. Female

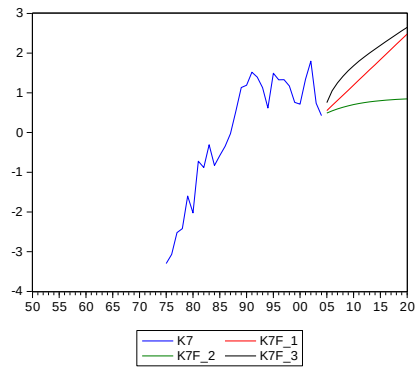
Estimated formulas for each model are shown below..

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.128 + k_{t-1}$	0.865
AR(1) model	$E(k_t) = 0.126 + 0.858k_{t-1}$	0.885
COVARIATE model	$E(k_t) = 0.000173 * GDP - 0.259 * Smoking,$ $E(\varepsilon_t) = 0.684 * \varepsilon_{t-1}$	0.890

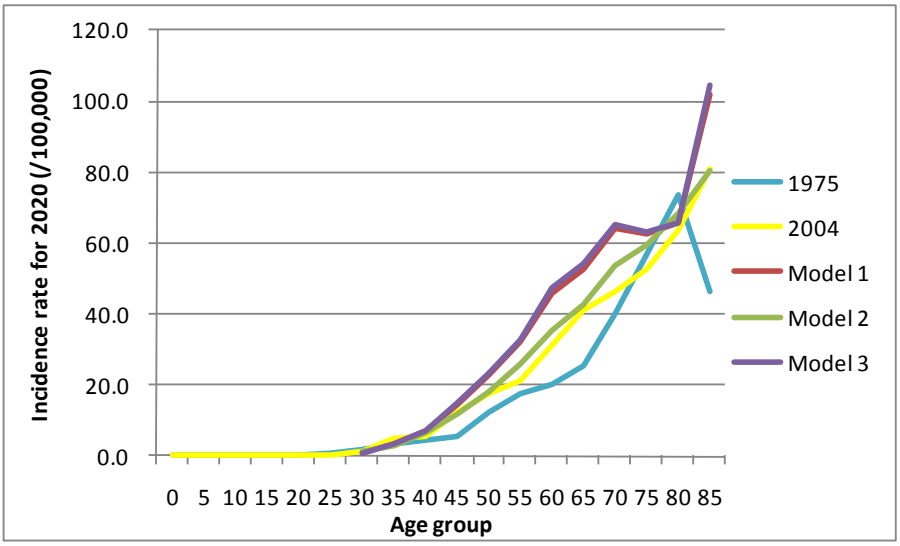
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



The coefficient of AR(1) tells implicitly that the process is a bit away from random walk. For 2004 data all models overestimated in older age groups a bit. All cases look plausible when we look at the trajectory of K to 2020. I would choose COVARIATE model because of its best fit and its capability to capture the change both in GDP and smoking. Smoking again is negative in sign. I will discuss about it later in the thesis (Discussion part). If exogenous variable is not available I would choose AR(1). Basically because it has better fit, the coefficient of 0.858 is bit far from one, and the future trajectory is at least on the increase path which is in line with the current empirical evidences.

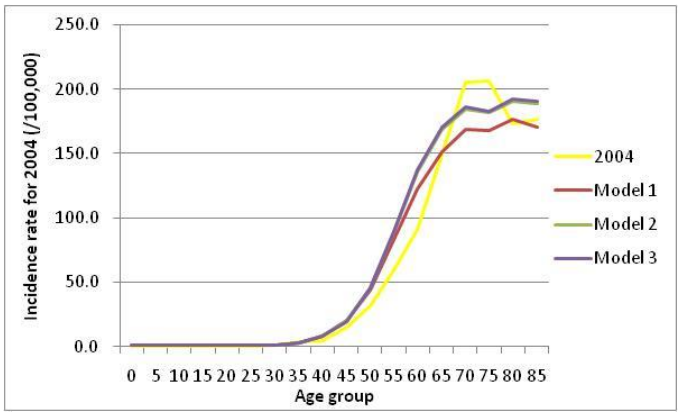
10. Liver: ICD10(C22)

10.1. Male

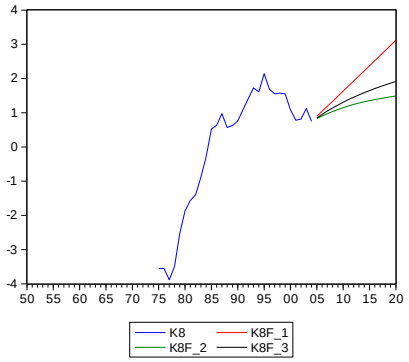
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.148 + k_{t-1}$	0.952
AR(1) model	$E(k_t) = 0.146 + 0.915k_{t-1}$	0.959
COVARIATE model	$E(k_t) = 0.000058 * GDP,$ $E(\varepsilon_t) = 0.908 * \varepsilon_{t-1}$	0.959

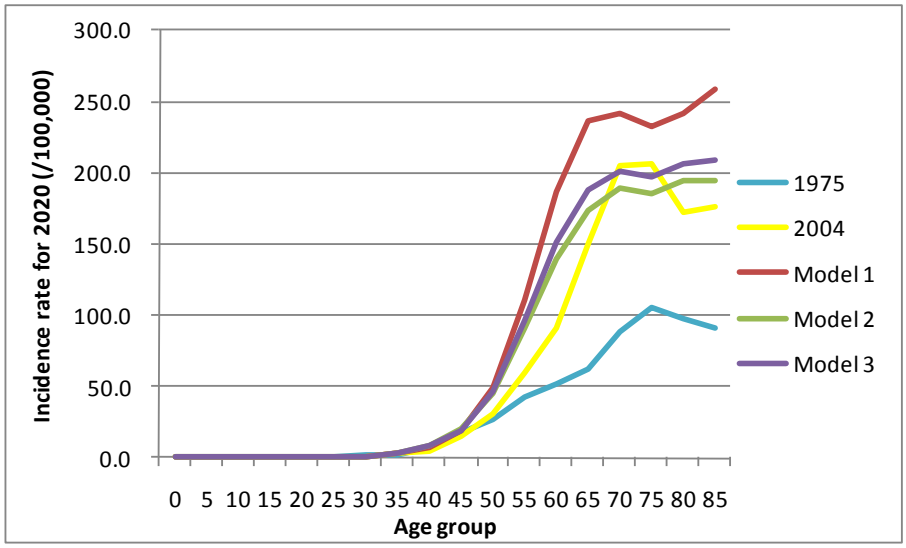
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



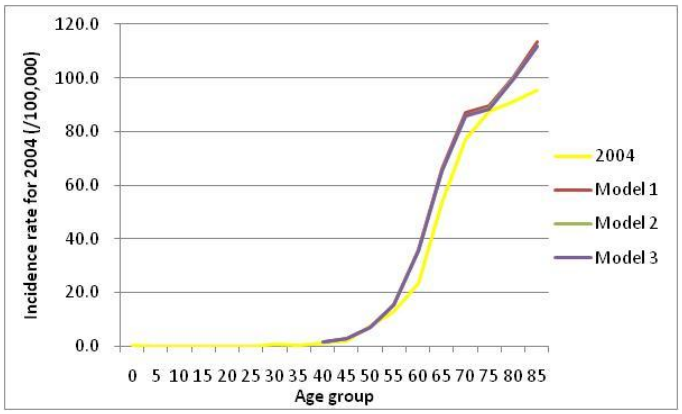
R squared for different models are similar and differ on three decimals. The trend in k_t seems to have changed in 1990s. This is when we were in economic recession. That implicitly tells that it might be better to put economic condition into the model. I would doubt Random walk. Random walk is predicting the simple upward trend. However hepatitis C is now the leading cause of liver cancer and precaution measure is currently taken so that I put some doubt on assuming simply upward trend. So AR(1) or COVARIATE model seem to be better than Random walk. I would choose COVARIATE model for its capability to capture GDP change.

10.2. Female

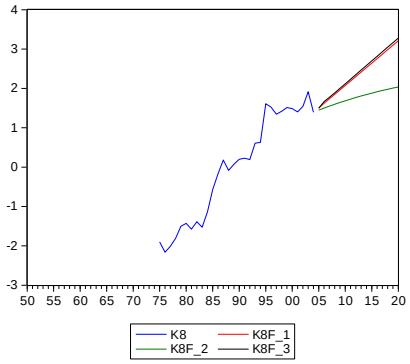
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.114 + k_{t-1}$	0.950
AR(1) model	$E(k_t) = 0.112 + 0.959k_{t-1}$	0.950
COVARIATE model	$E(k_t) = 0.000225 * GDP - 0.350 * Smoking,$ $E(\varepsilon_t) = 0.565 * \varepsilon_{t-1}$	0.956

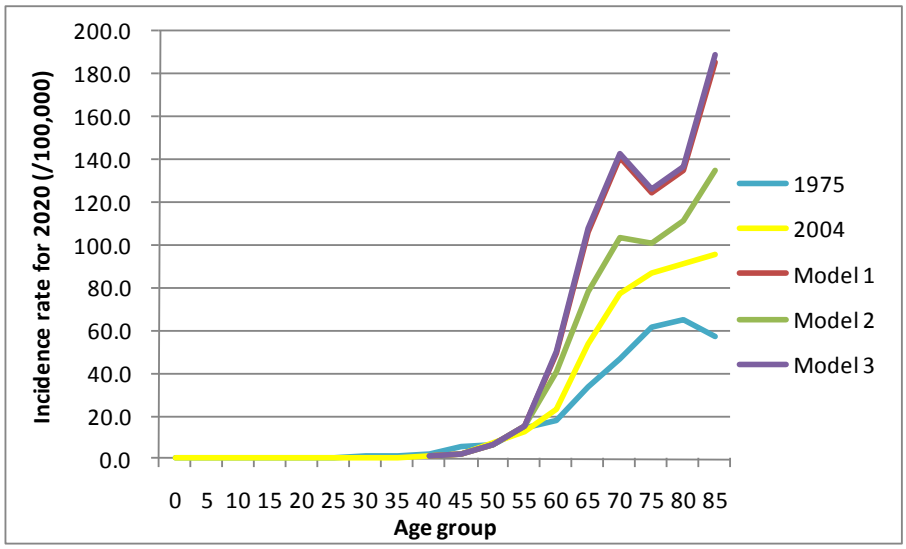
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Fit does not seem to be different between three models in terms of R squared. All models consistently overestimated 2004 data. In contrast with male, time trend k_t seems to have consistent increasing trend in the period between 1975 and 2004. In COVARIATE model GDP is positive effect while smoking is negative effect. Majority of liver cancer cases are progressed from Hepatitis C infections currently. Since infection precautions against this virus have been made over the last decade I don't think this current increasing trend will continue on in the future. So I would not choose random walk. I prefer COVARIATE model for its capability to capture GDP change (plus R squared is a bit higher for COVARIATE model). But both Random walk and COVARIATE models estimate huge increase in 2020, so that I cannot throw away AR(1) from the list of plausible options.

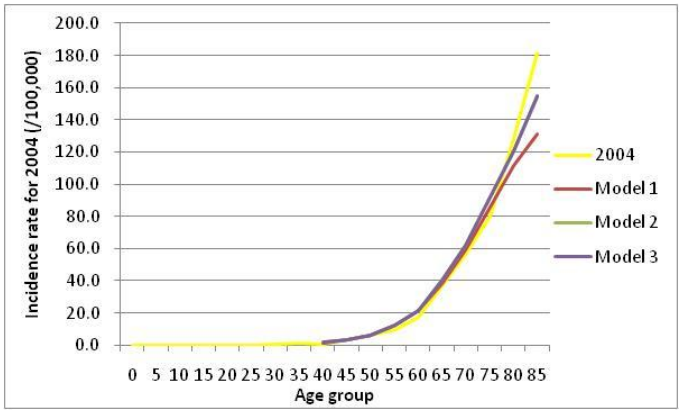
11. Gallbladder and bile ducts: ICD10(C23-C24)

11.1. Male

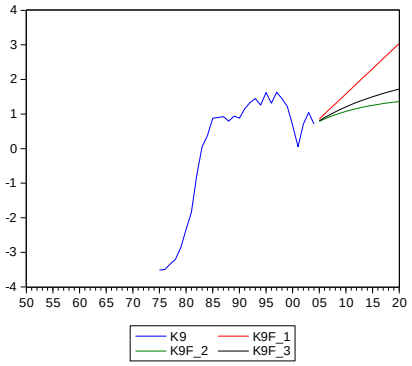
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.146 + k_{t-1}$	0.944
AR(1) model	$E(k_t) = 0.143 + 0.906k_{t-1}$	0.953
COVARIATE model	$E(k_t) = 0.0000515 * GDP,$ $E(\varepsilon_t) = 0.898 * \varepsilon_{t-1}$	0.953

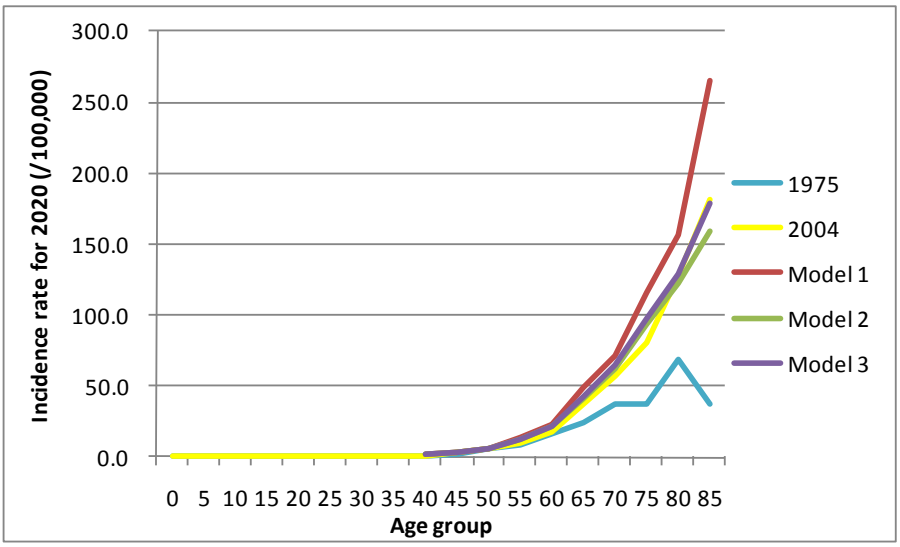
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



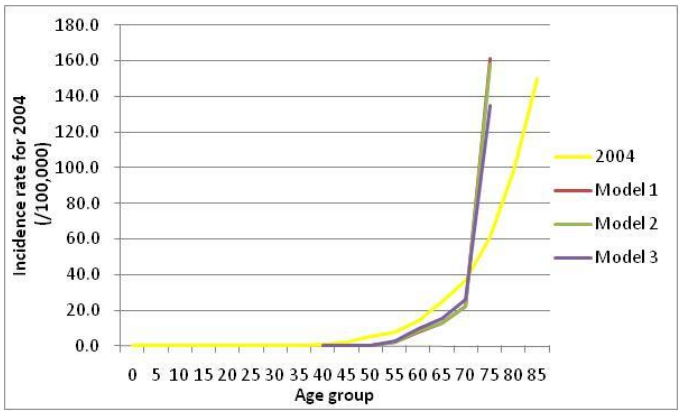
R square is bit lower for random walk. Comparing with 2004 actual data, COVARIATE model seems to fit best of all. The movement of k_t in the observational period is first going up dramatically and leveling off in the later period (the mid 1980s onward). Bile is essential for fat absorption (it has no enzyme but help the function of lipase) and westernized life style has had a lot to do so I think at least it has to increase in 2020 compared with the current level. I would use COVARIATE model because of including GDP. The sign of GDP is positive and is line with our expectation. If exogenous variable is not available I would use both Random walk and AR(1) to calculate the guidance boundaries for upper and lower.

11.2. Female

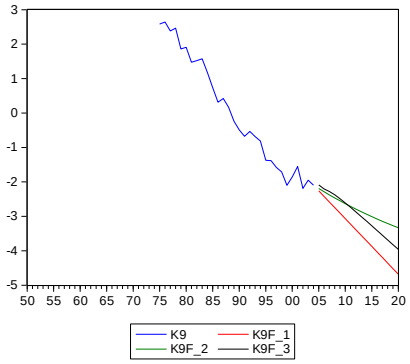
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.161 + k_{t-1}$	0.972
AR(1) model	$E(k_t) = -0.159 + 0.970k_{t-1}$	0.972
COVARIATE model	$E(k_t) = -0.000268 * GDP + 0.413 * Smoking,$ $E(\varepsilon_t) = 0.647 * \varepsilon_{t-1}$	0.970

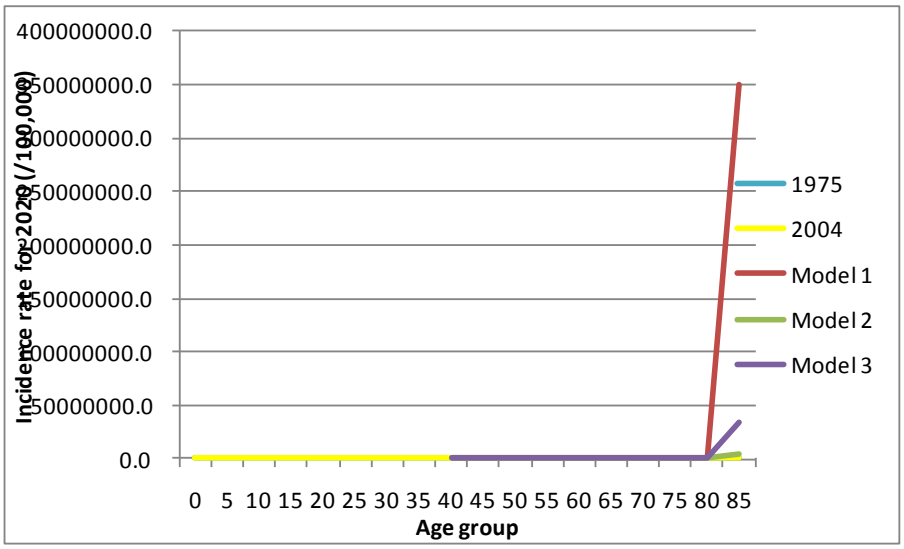
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)

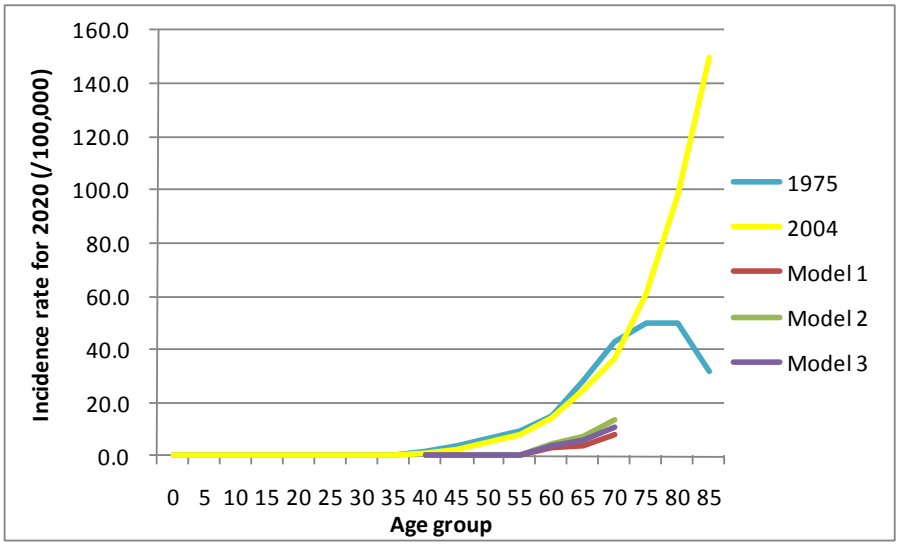


Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



R squared is similar and highest in random walk. However the fit for 2004 data shows that after 75 yrs old the expected values exploded. The forecasted figure for 2020 (above) using all data looks exploded. The incidence for example for age group 85 yrs old or older is estimated to be 300,000,000 per 100,000 populations.

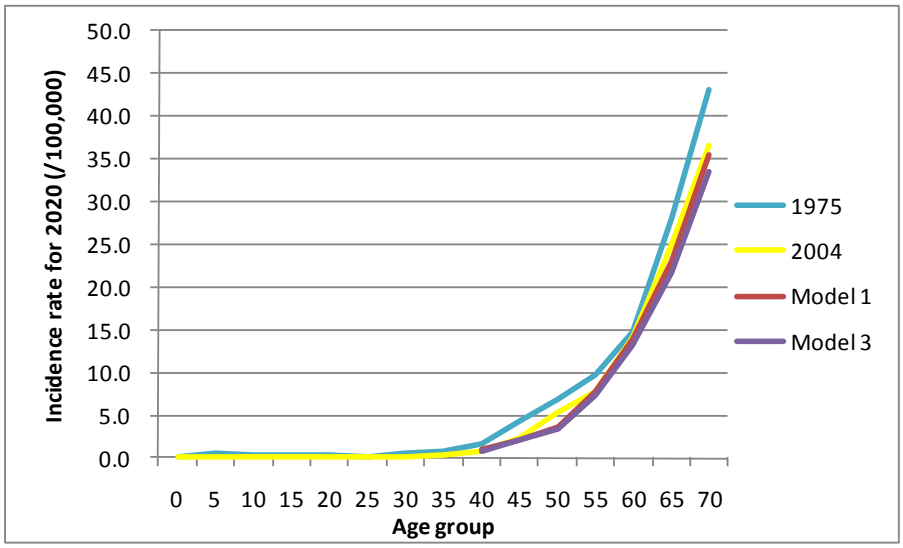
If we take away the age group 75 yrs old or older in order to see what is going on, the figure looks like follow.



Lee-Carter's trend parameter has captured the huge increase in incidence in the older age group. However this is the trend which data suggests (exponential increase since 1975) and in that case it is inappropriate to use any method of forecasting including Lee-Carter, which assumes the current trend is not changing in the future. Since inclusion of 75 yrs old or older makes k_t strange, I did the same thing but excluding the age group 75 yrs or older.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.635 + k_{t-1}$	0.896
AR(1) model	$E(k_t) = -0.0638 + 1.005k_{t-1}$	Identical with unit root
COVARIATE model	$E(k_t) = 6.698 - 0.00028 * GDP,$ $E(\varepsilon_t) = 0.834 * \varepsilon_{t-1}$	0.902

Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



2020 forecast based on both models are on the 2004 incidence rate. For the age groups under 75 the incidence did not change during 1975 and 2004, so that it would be no wonder.

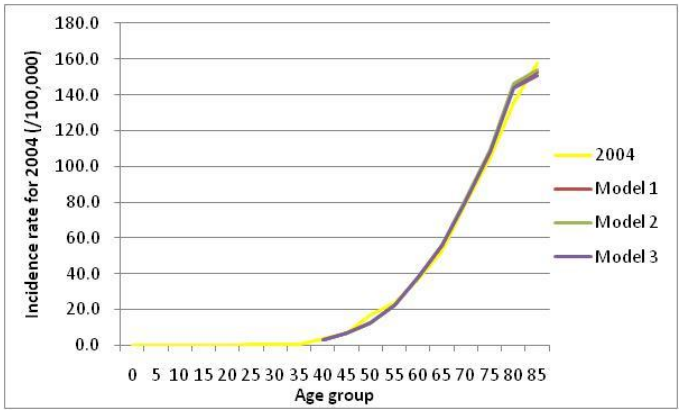
12. Pancreas: ICD10(C25)

12.1. Male

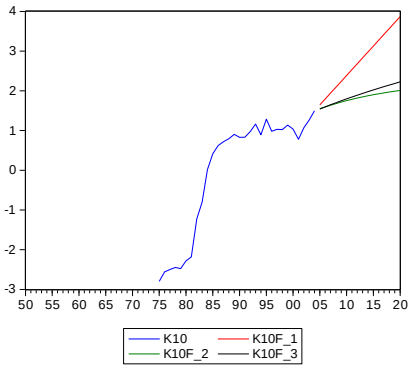
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.148 + k_{t-1}$	0.963
AR(1) model	$E(k_t) = 0.145 + 0.937k_{t-1}$	0.966
COVARIATE model	$E(k_t) = 0.000064 * GDP,$ $E(\varepsilon_t) = 0.922 * \varepsilon_{t-1}$	0.965

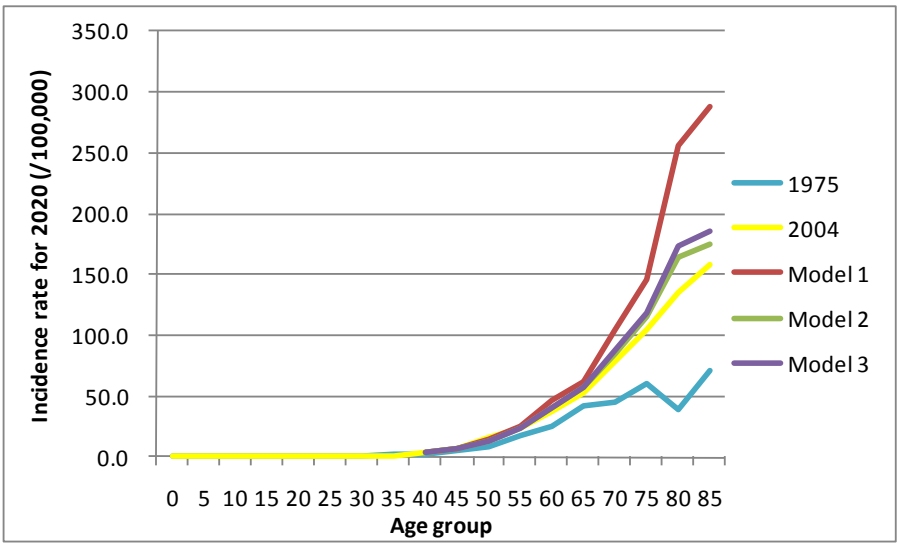
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



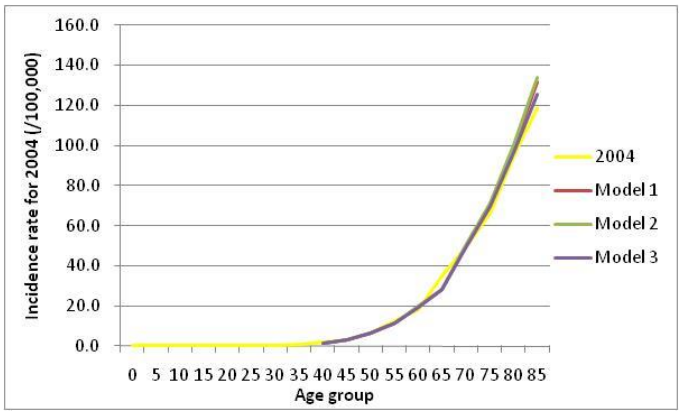
The trend (parameter k_t) changed in the mid 1980s and Random Walk does not capture this. Looking at the trajectory of K in the future might suggest that other two models seem to follow the trend in the period from the mid 1980s onward. Based on good fit and its capacity to capture GDP I would prefer COVARIATE model though AR(1) has slightly better fit. Pancreatic cancer is very hard to detect in the early stage because of its location in the abdominal cavity. This has been an issue for many years and the device or method for early detection has been researched and would be continue on. So I would imagine the positive relationship with GDP through technical advancement.

12.2. Female

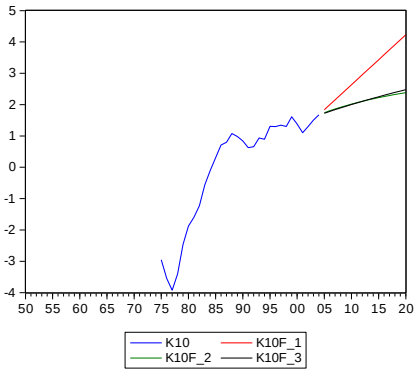
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.160 + k_{t-1}$	0.960
AR(1) model	$E(k_t) = 0.156 + 0.945k_{t-1}$	0.961
COVARIATE model	$E(k_t) = 0.0000716 * GDP,$ $E(\varepsilon_t) = 0.929 * \varepsilon_{t-1}$	0.960

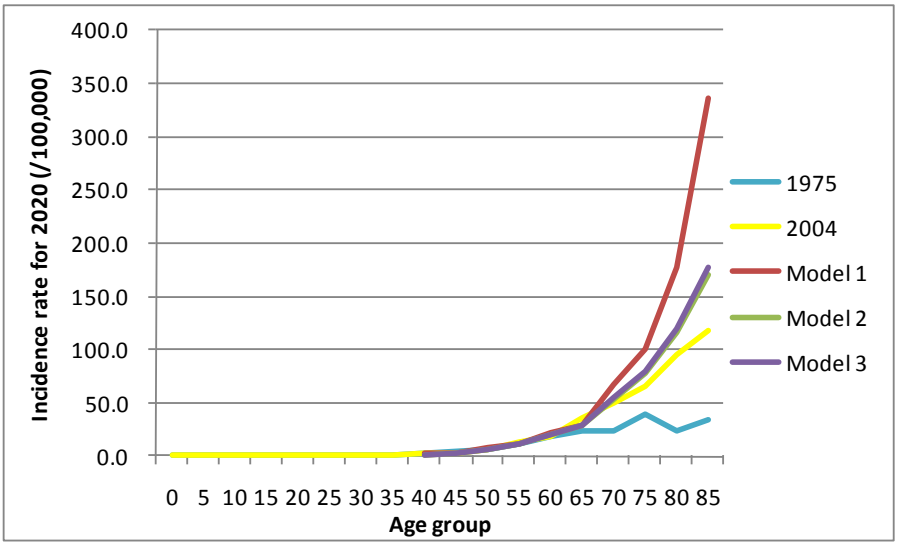
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



There is almost no difference in fit (96%). All three models accurately and precisely predicted 2004 data. The trend from the mid 1980s was captured by AR(1) and COVARIATE model. As is said in pancreas cancer for male, there is an advantage in including GDP. However DW for COVARIATE is 0.909. It might be the spurious regression. In this case it might be better to explain completely endogenously. Therefore I would pick AR(1) for female pancreatic cancer incidence.

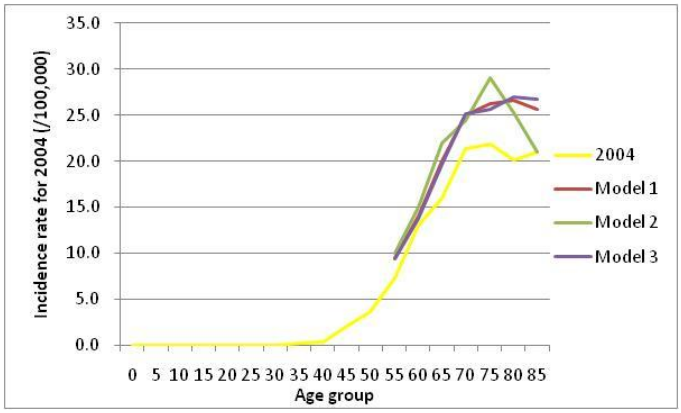
13. Larynx: ICD10(C32)

13.1. Male

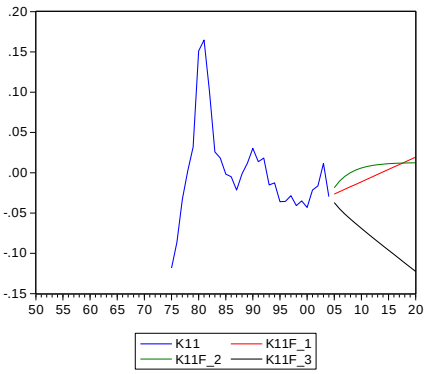
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.00306 + k_{t-1}$	0.555
AR(1) model	$E(k_t) = 0.00333 + 0.738k_{t-1}$	0.621
COVARIATE model	$E(k_t) = -0.00000589 * GDP + 0.00234 * Smoking,$ $E(\varepsilon_t) = 0.690\varepsilon_{t-1}$	0.673

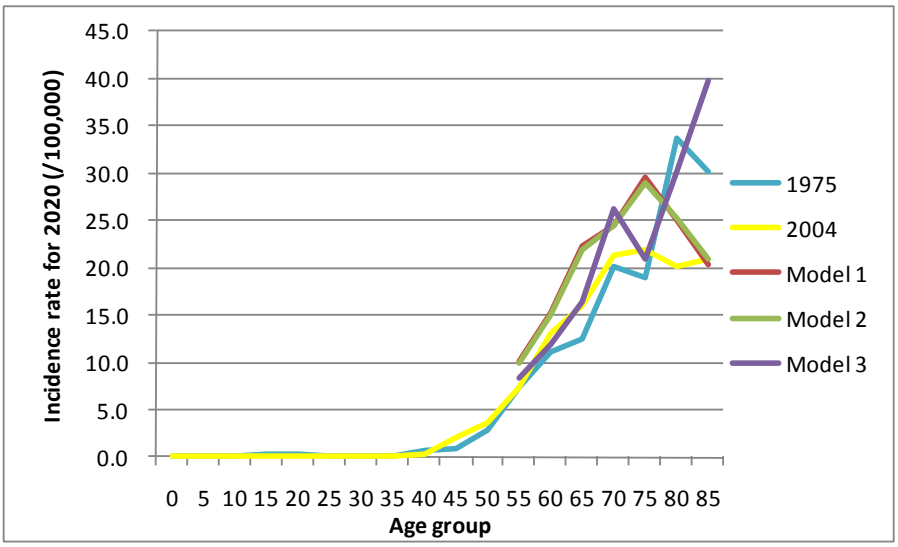
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



COVARIATE model was best in adjusted R squared. No matter what model we used it was always overestimation for 2004. Fit is only less than 70% level. I would choose COVARIATE model (GDP is negative but the size is 5 digits level). And poor fit (highest fit was observed in COVARIATE but only 67% in R squared) would suggest that we may need to include other exogenous factors into COVARIATE model.

If exogenous information is not available I would choose AR(1). R squared is bigger for AR(1) and the coefficient of 0.78 would make us feel the series is not close to Random walk.

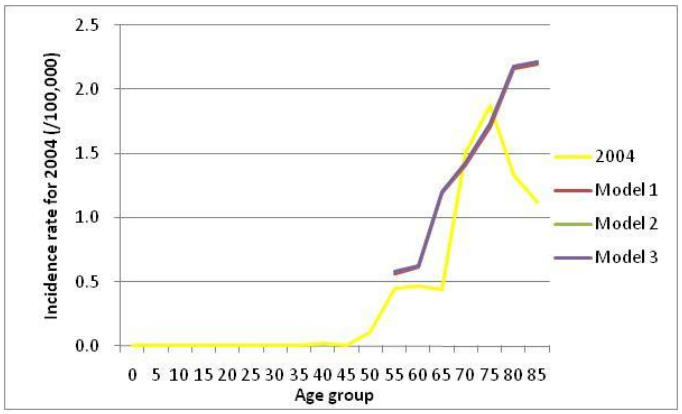
13.2. Female

Estimated formulas for each model are shown below.

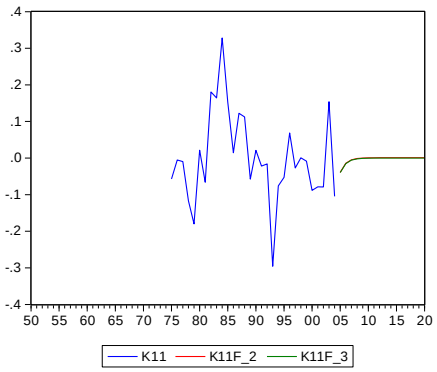
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.00163 + k_{t-1}$	--
AR(1) model	$E(k_t) = 0.000604 + 0.379k_{t-1}$	0.101
COVARIATE model	$E(k_t) = 0.379k_{t-1}$	0.141

Note : In COVARIATE model the GDP and smoking do not have significant explanatory power

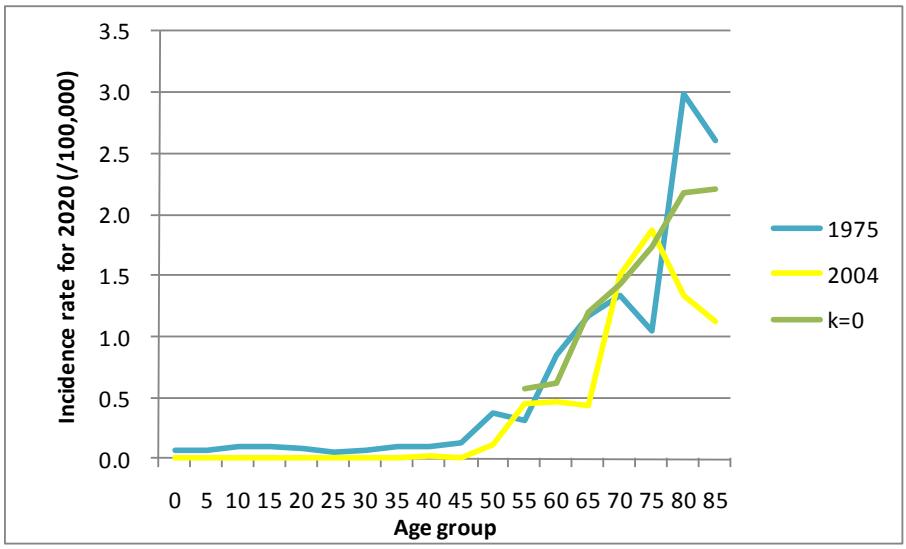
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



k_t does not change over time and there seems to be no trend either upward or downward. As shown above none of the model does not fit well to the current data (R squared is 10% level. Since this is time series data analysis setting 10% is quite low.). No model predicted 2004 data. The reason here is that incidence is almost zero. Highest incidence for 2004 was observed in the age group of 75 yrs old but that was less than 2.0. No model would be chosen but assuming $k_t = 0$ would be the reasonable choice. I would conclude none of the three models worked well.

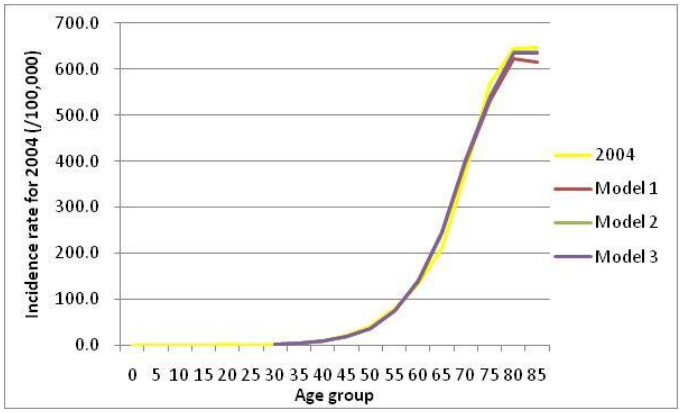
14. Lung Trachea: ICD10(C33-C34)

14.1. Male

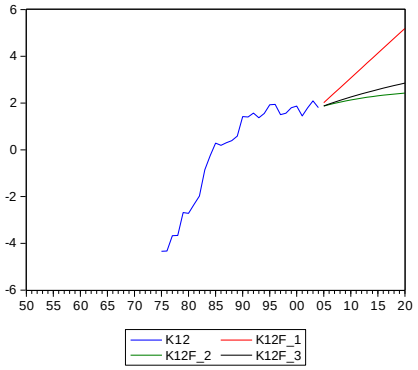
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.212 + k_{t-1}$	0.965
AR(1) model	$E(k_t) = 0.207 + 0.922k_{t-1}$	0.970
COVARIATE model	$E(k_t) = 0.000082 * GDP,$ $E(\varepsilon_t) = 0.908\varepsilon_{t-1}$	0.970

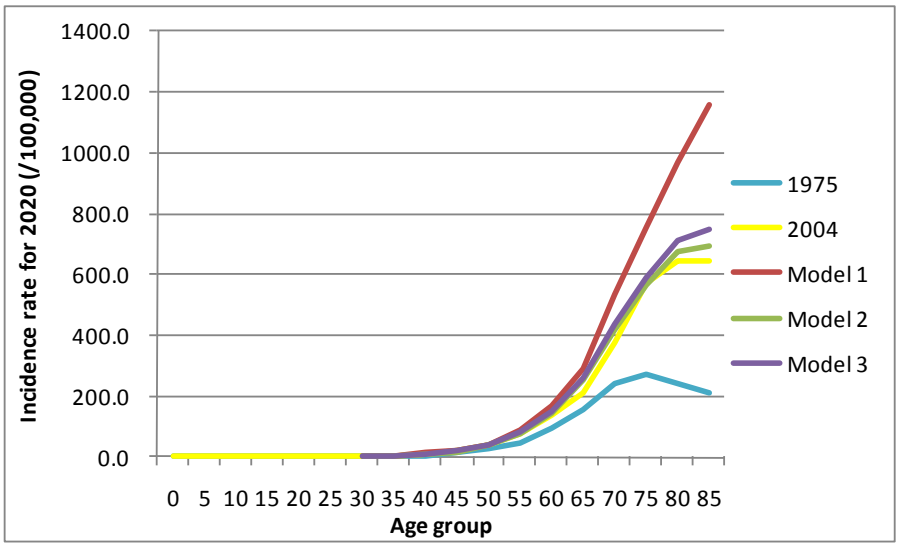
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



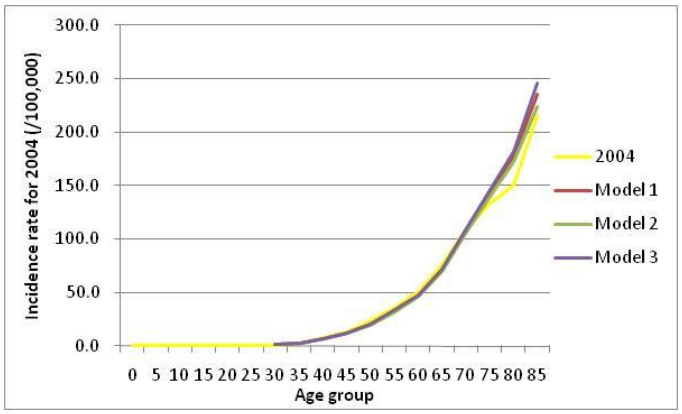
The fit seems to be in favor of both AR(1) and COVARIATE models. The incidence of Lung cancer has increased and this increase tapered off in 1990s when there was an economic recession in Japan. The movement is quite similar in GDP trend and that was shown in the result. Smoking is sure to have been a risk factor for lung cancer in individual level but when it comes to looking at the society as a whole proportion of smokers does not have significant effect. I would pick COVARIATE model because of its best fit and its capability to capture GDP. Lung cancer might be going up in number but stop smoking campaign is everywhere so it would be remained in the same level. If exogenous information is not available I would use both Random walk and AR(1) because both of them are plausible.

14.2. Female

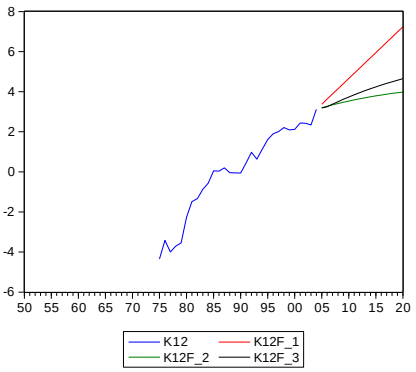
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.258 + k_{t-1}$	0.962
AR(1) model	$E(k_t) = 0.252 + 0.945k_{t-1}$	0.964
COVARIATE model	$E(k_t) = 0.460 * Smoking,$ $E(\varepsilon_t) = 0.956\varepsilon_{t-1}$	0.967

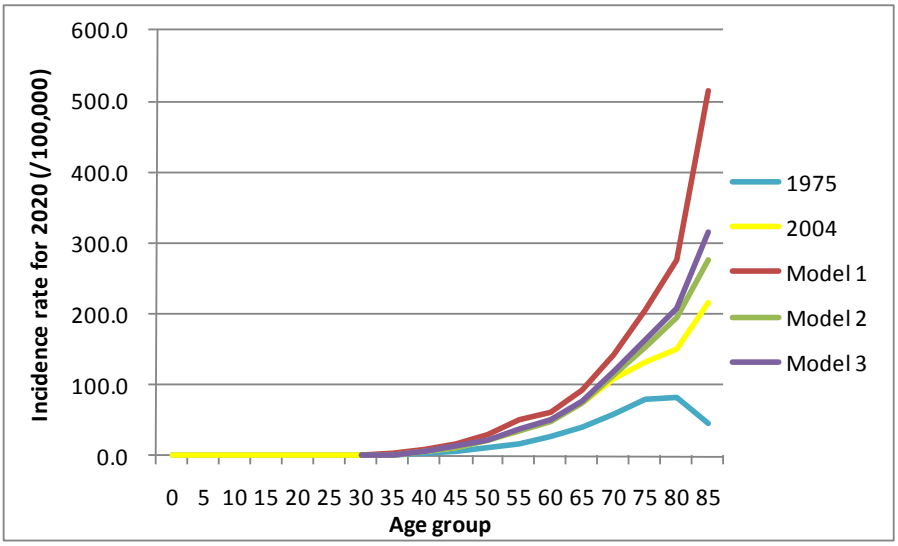
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



All three models fit quite well in terms of R squared and the comparison with 2004 actual data. R squared suggests that COVARIATE model fits best of all and smoking has left in the explanatory variable. The result was positive smoking effect and no GDP effect. In women's case proportion of smoker was already around 15% level and changed little in the 30 yrs period. Time trend k_t looks consistent upward movement. It is lung cancer and positive smoke effect would be easy to accept though most of other cases smoking is either insignificant or negative in sign. I would choose COVARIATE model. If exogenous variable (smoker's proportion) is not available, both Random walk and AR(1) should be calculated.

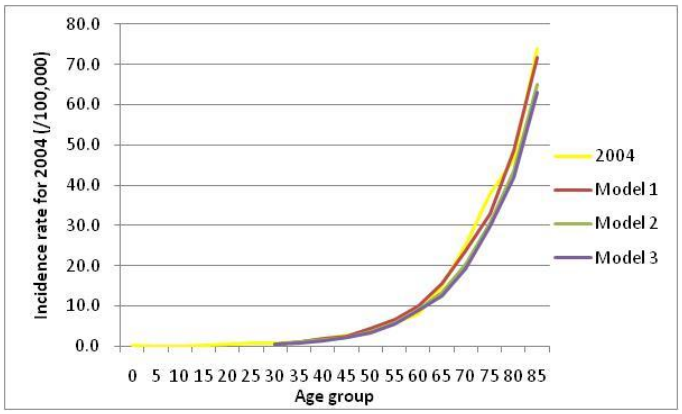
15. Skin: ICD10(C43-C44)

15.1. Male

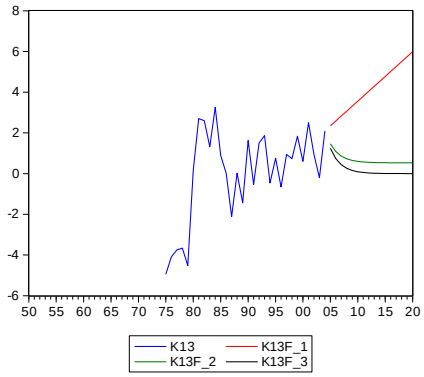
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.243 + k_{t-1}$	0.223
AR(1) model	$E(k_t) = 0.214 + 0.595k_{t-1}$	0.392
COVARIATE model	$E(k_t) = 0.592k_{t-1}$	0.403

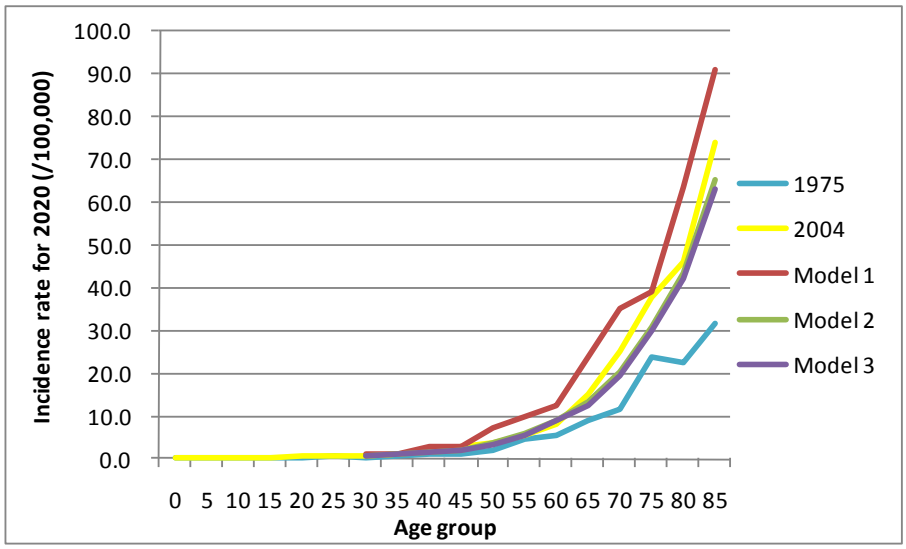
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



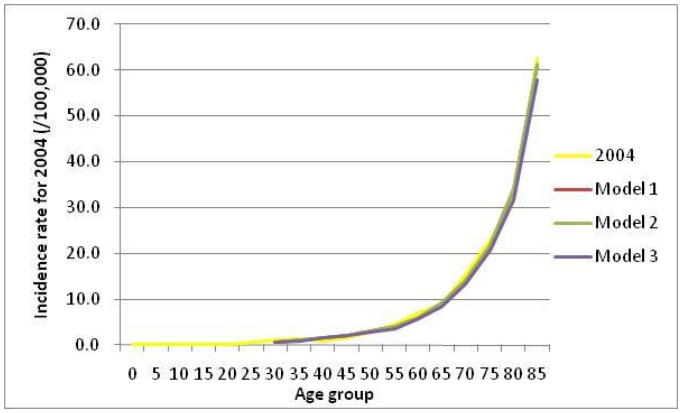
R squared is less than 50%. The current trend (K) is hard to capture. However in spite of poor fit in R squared, expected values for 2004 quite match the actual data. Exogenous factors are turned out to be not included in COVARIATE model. Since the coefficient of AR(1) is 0.6 I would not choose random walk. Both Model 2 and 3 are close to assuming $k_t = 0$ in the trajectory to 2020. Which one would I choose? For that moment assuming $k_t = 0$ seems to be reasonable because I don't see dramatic difference in incidence rate for skin cancer in the last 30 years when looking at the data.

15.2. Female

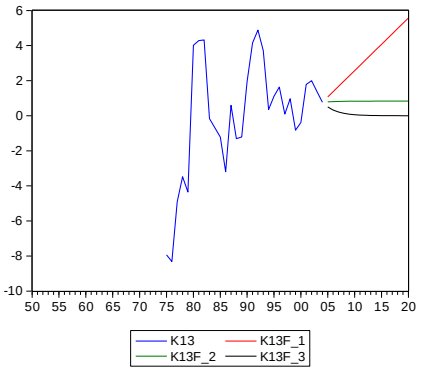
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.300 + k_{t-1}$	0.375
AR(1) model	$E(k_t) = 0.291 + 0.651k_{t-1}$	0.510
COVARIATE model	$E(k_t) = 0.650 * k_t$	0.517

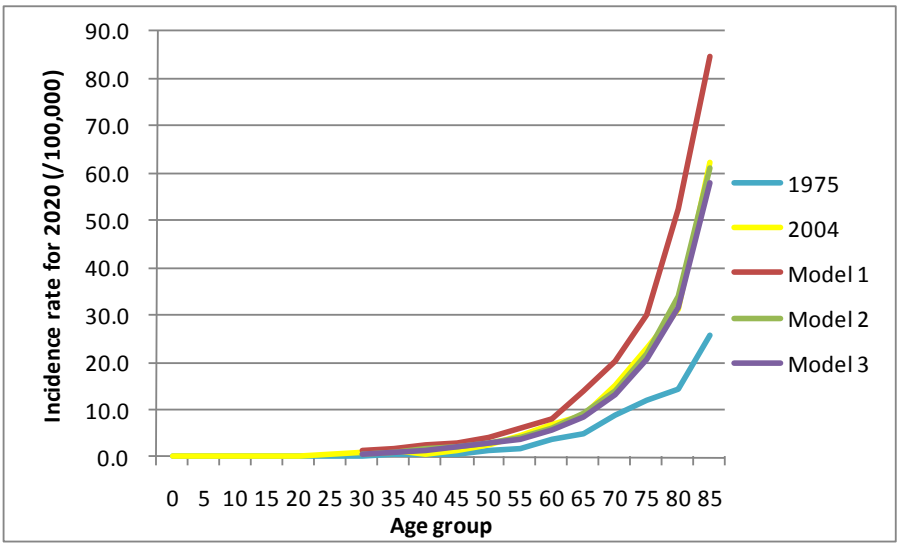
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Female case is quite similar with the male case for skin cancer. R squared is not high. But for 2004 actual data and predicted data fit very well. COVARIATE model in this case does not have any exogenous factors. The coefficient of AR(1) is far from unit, it is no surprise that Random-walk model fits worst (R squared is 38%). My preference is AR(1). However I guess assuming $k_t = 0$ after 2004 is probably the best and it is almost the same as picking AR(1) or COVARIATE model.

16. Breast: ICD10(C50 D05)

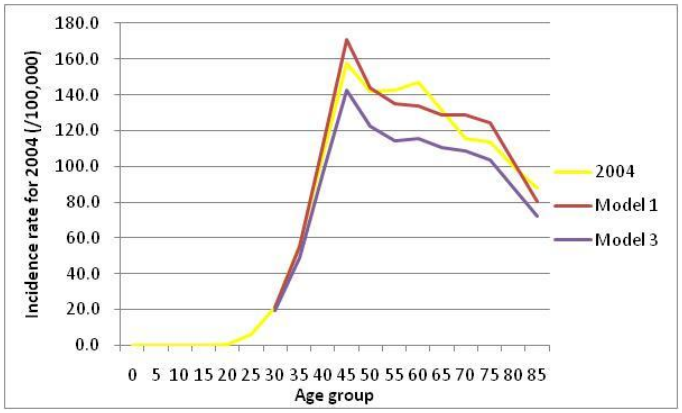
16.1. Female

Estimated formulas for each model are shown below.

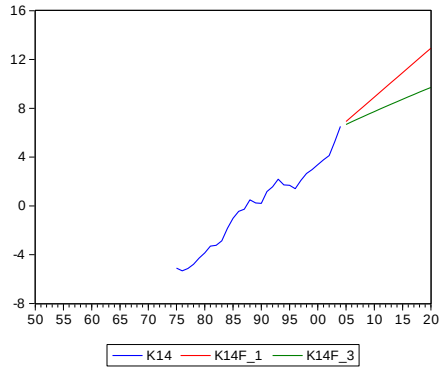
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.401 + k_{t-1}$	0.982
AR(1) model	$E(k_t) = 0.405 + 1.02k_{t-1}$	Identical to unit root
COVARIATE model	$E(k_t) = 0.000302 * GDP,$ $E(\varepsilon_t) = 0.967\varepsilon_{t-1}$	0.977

Note: Estimated AR(1) model is not autoregressive.

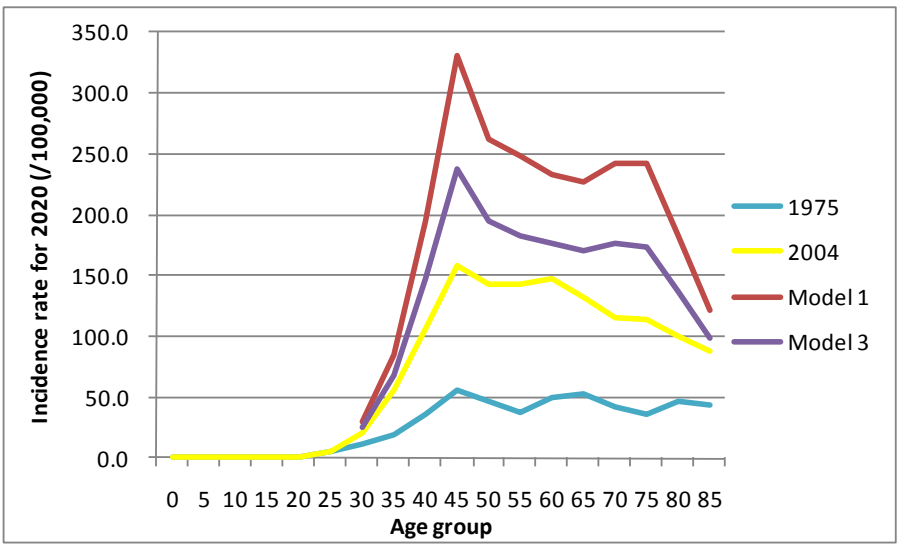
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



The fit for Random walk is more than 98% and the coefficient of AR(1) is more than unit. So k_t is probably unit root. DW for COVARIATE model is 1.12 and it is on the border line of spurious regression or cointegration (suppose GDP is unit root). I would choose Random walk based on highest R squared. Random walk seems to have better fit to actual data for 2004. The cancer is sex hormone related and the fecund period (from menarche till menopause) matters. I guess that is why inclusion of GDP did not reach as high as Random-walk model in terms of R squared and why I chose Random walk over COVARIATE model.

17. Uterus (incl. epithelial carcinoma) : ICD10(C53-C55 D06)

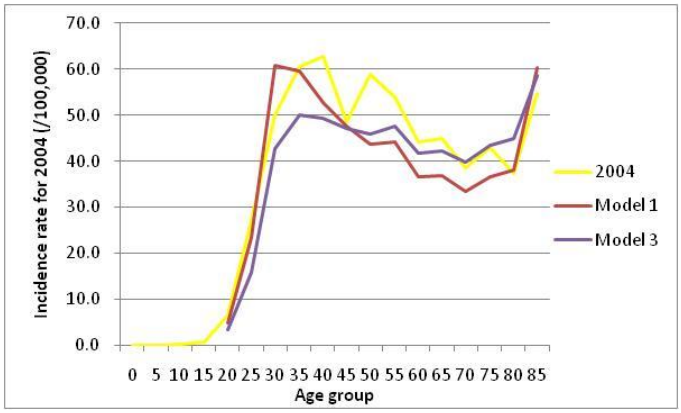
17.1. Female

Estimated formulas for each model are shown below.

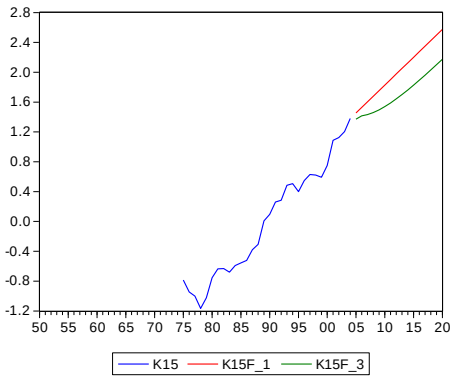
Models	Estimates	Adjusted R^2
Random Walk model	$k_t = 0.0746 + k_{t-1}$	0.973
AR(1) model	$E(k_t) = 0.076 + 1.03k_{t-1}$	Identical to unit root
COVARIATE model	$E(k_t) = 0.000149 * GDP - 0.234 * Smoking,$ $E(\varepsilon_t) = 0.822\varepsilon_{t-1}$	0.968

Note: Estimated AR(1) model is not autoregressive.

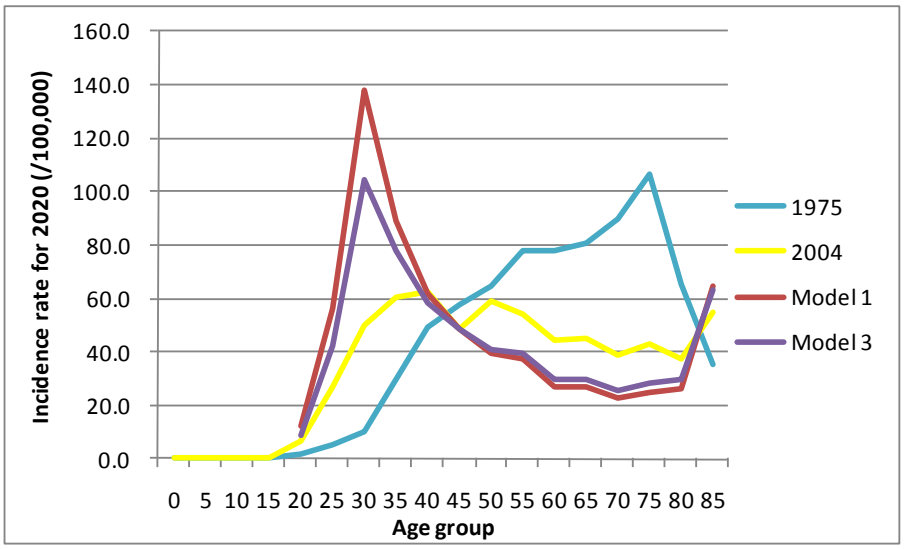
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



As in the case of breast cancer Random walk has best fit. It is closer to the actual data for 2004 as shown in the figure. COVARIATE model underestimated the 2004 data for most of the age groups. The shape for 2020 seems funny because its huge peak in the early age group and U shaped in the middle to older age group. It may be the true picture for 2020 or time trend parameter did not catch the trend. However this is the sum of uterine and cervix cancer and we may want to see the results separately which will be shown in the next two chapters.

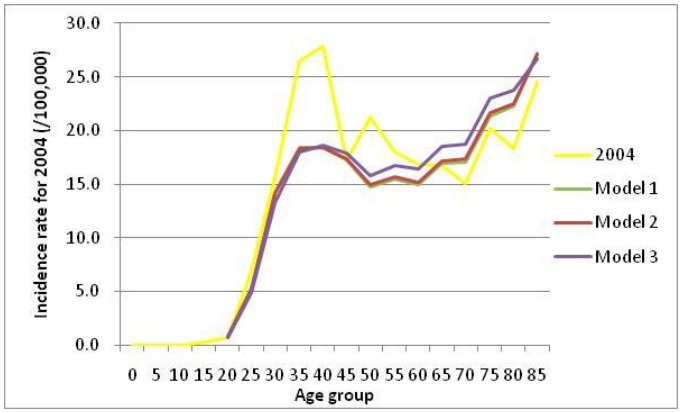
18. Cervix uteri: ICD10(C53)

18.1. Female

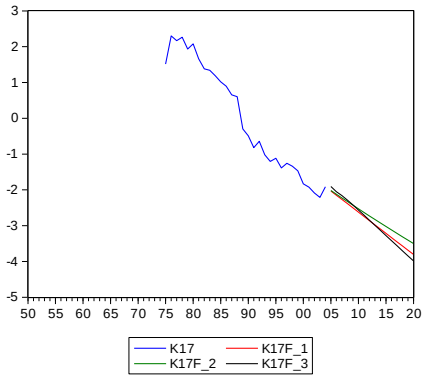
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = -0.118 + k_{t-1}$	0.965
AR(1) model	$E(k_t) = -0.118 + 0.993k_{t-1}$	0.964
COVARIATE model	$E(k_t) = -0.000276 * GDP + 0.433 * Smoking,$ $E(\varepsilon_t) = 0.485\varepsilon_{t-1}$	0.978

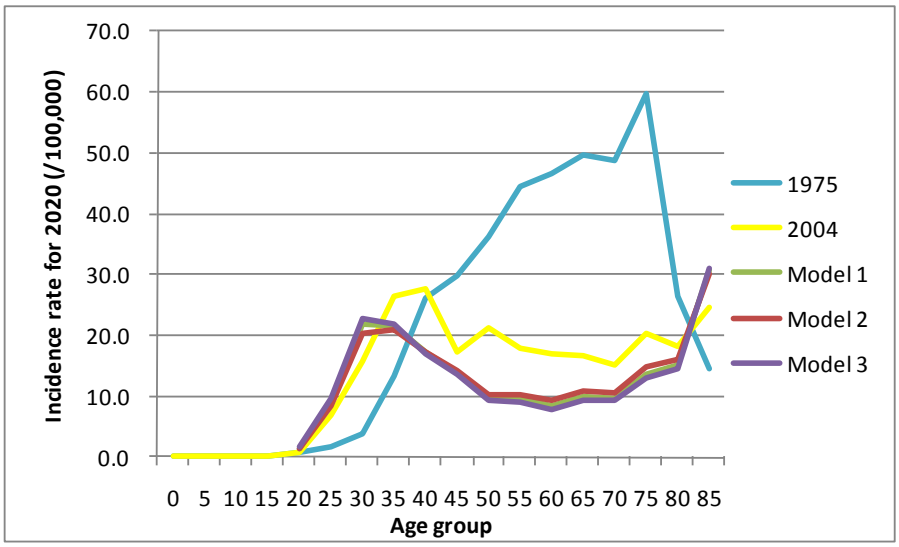
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



All of the models fit more than 96%. However none of the model could predict the jump in incidence for 2004. Regardless of the models the forecasted incidence for 2020 are the same. The risk of cancer is said to be multiple sex or Human Papillomavirus infection. So the shape of taking off in 20s seems reasonable. COVARIATE model has negative GDP and positive smoking. Infectious disease related profile might suggest that the disease is related to health education or prevention measures which are easier to implement when the country gets richer. Smoking has correlated with high risk sex behavior as well. So the positive relationship between proportion of smokers and number of cancer is also understandable. I would choose COVARIATE model based on best fit and the inclusion of exogenous factors.

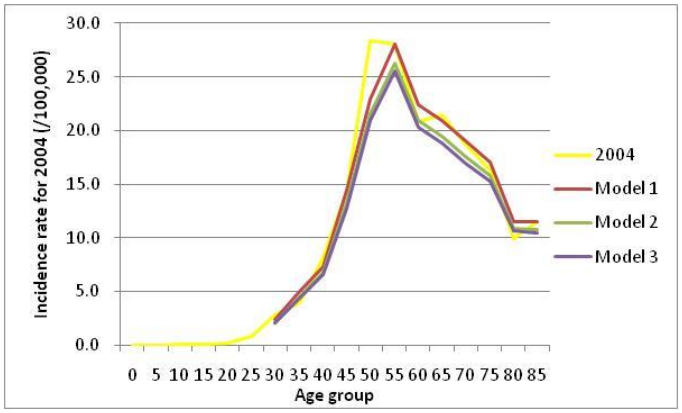
19. Corpus uteri: ICD10(C54)

19.1. Female

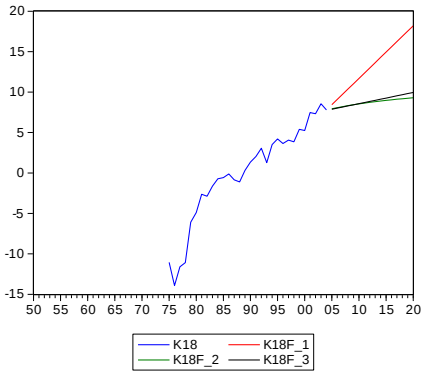
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.650 + k_{t-1}$	0.935
AR(1) model	$E(k_t) = 0.633 + 0.938k_{t-1}$	0.937
COVARIATE model	$E(k_t) = 0.000276 * GDP,$ $E(\varepsilon_t) = 0.918\varepsilon_{t-1}$	0.937

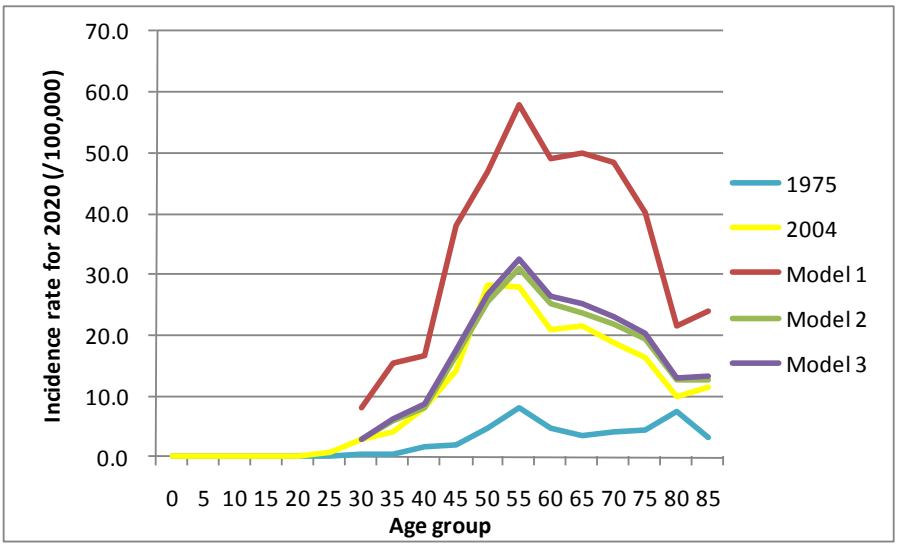
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



There is almost no difference in R squared (94%). All models fit quite well for 2004 and well capture the jump in the age group 40s to 60s. This is in line with the epidemiological evidence that it takes off in 40s and has peak in 50s or 60s. Only GDP is left and positive in COVARIATE model. It might be related to the advances in detection which is related to GDP. I would take COVARIATE model because of its fit and the inclusion of GDP. According to COVARIATE model a little increase will be expected in the incidence rate in 2020. However the coefficient of AR(1) is more than 0.93 and one could not throw away Random walk when exogenous information is not available. Random walk provides us with pessimistic scenario.

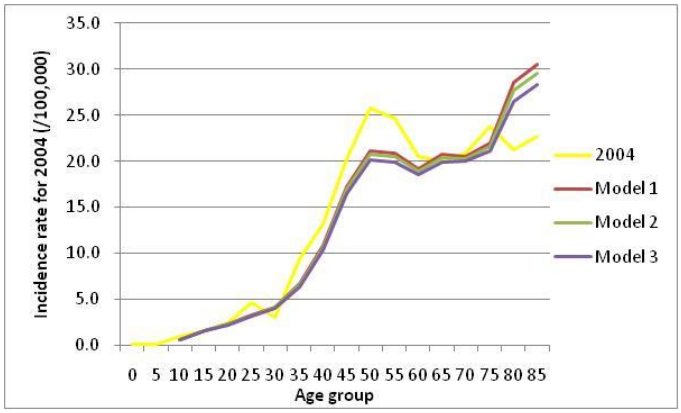
20. Ovary: ICD10(C56)

20.1. Female

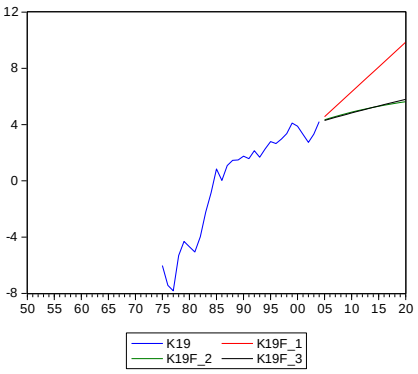
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.353 + k_{t-1}$	0.943
AR(1) model	$E(k_t) = 0.345 + 0.948k_{t-1}$	0.944
COVARIATE model	$E(k_t) = 0.000164 * GDP,$ $E(\varepsilon_t) = 0.930\varepsilon_{t-1}$	0.943

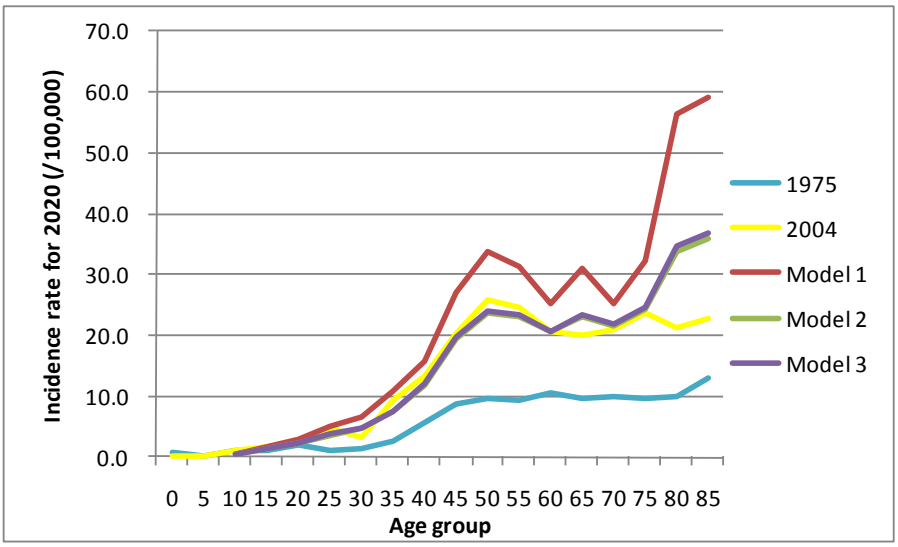
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



R squared is similar for all three models. The comparison with actual data for 2004 shows every model is similar. AR(1) and COVARIATE model are similar both in fit and in forecast. I would choose COVARIATE model because of its ability to capture GDP change. Early stage of ovary cancer has no symptom at all and most women go to the hospital when they get in advanced stage. Early detection is an issue for the cancer and that makes me attracted with the model which includes GDP. However if exogenous information is not available I would choose Random walk. It is ovum related and I think it should have one peak before menopause. AR(1) does not seem to have a significant peak before menopause and the coefficient of 0.95 cannot throw away the possibility of Random walk.

21. Prostate: ICD10(C61)

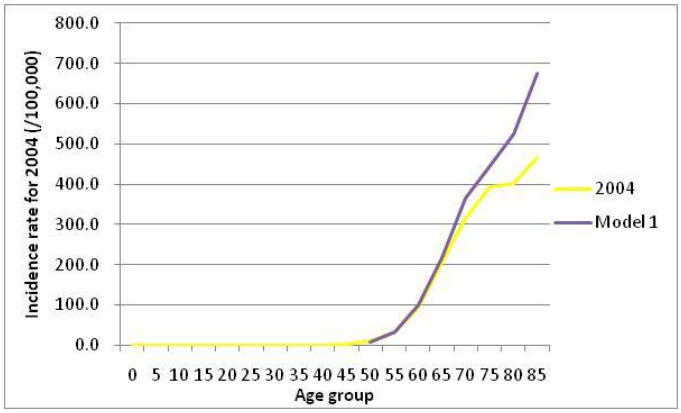
21.1. Male

Estimated formulas for each model are shown below.

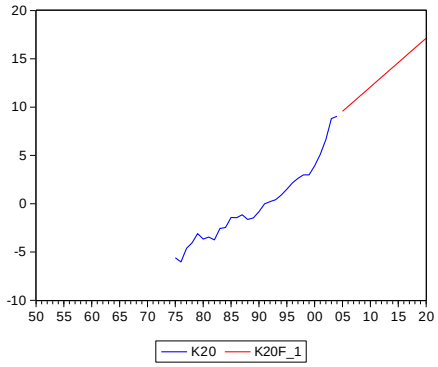
Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.505 + k_{t-1}$	0.973
AR(1) model	$E(k_t) = 0.523 + 1.055k_{t-1}$	Identical to unit root
COVARIATE model	--	--

Note: Estimated AR(1) model is not autoregressive. The estimation for COVARIATE model could not find any appropriate model. (It is either coefficient for AR is more than one or is ended up spurious regression).

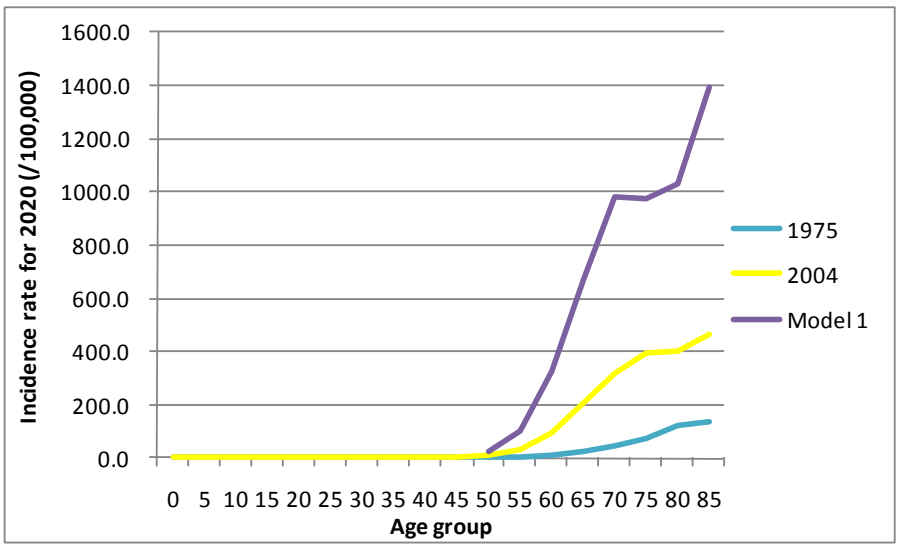
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Only Random walk was left. But the expected and actual data for 2004 has discrepancy for older age group. The trajectory of K based on random walk is simple extrapolation of the current dataset. However the estimated incidence for 2020 is very high. It is related to western life style because in Europe or North America prostate cancer accounts for almost 20% of death for male. Therefore it is no wonder to have been estimated very high level. It is true we have been westernized in lifestyle. But I don't believe Japan will become closer in lifestyle to Europe or North America significantly than the current level. I would guess the true incidence is less than the estimates based on assuming time trend follows random walk.

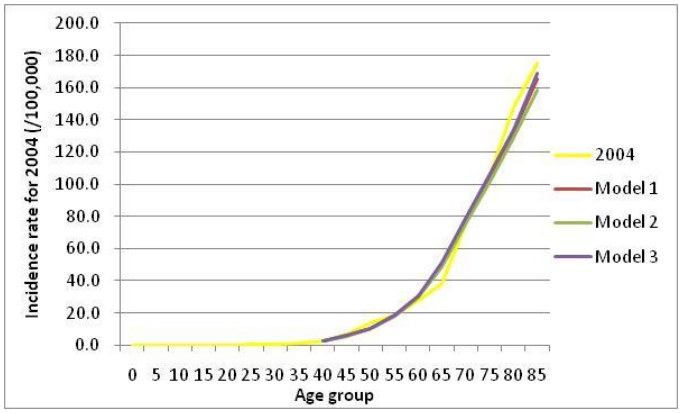
22. Bladder: ICD10(C67)

22.1. Male

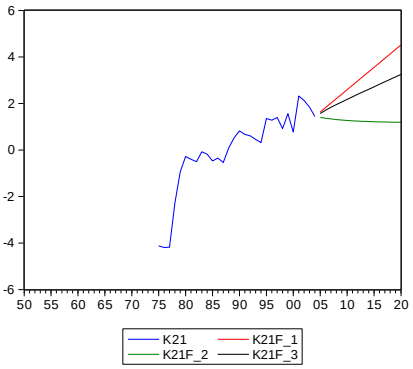
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.192 + k_{t-1}$	0.841
AR(1) model	$E(k_t) = 0.184 + 0.843k_{t-1}$	0.867
COVARIATE model	$E(k_t) = 0.000132 * GDP - 0.0397 Smoking,$ $E(\varepsilon_t) = 0.708\varepsilon_{t-1}$	0.868

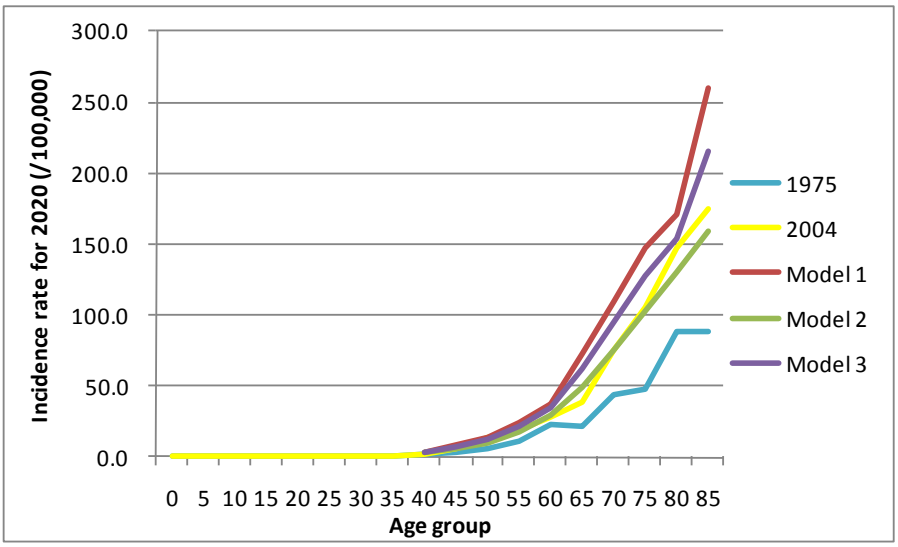
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



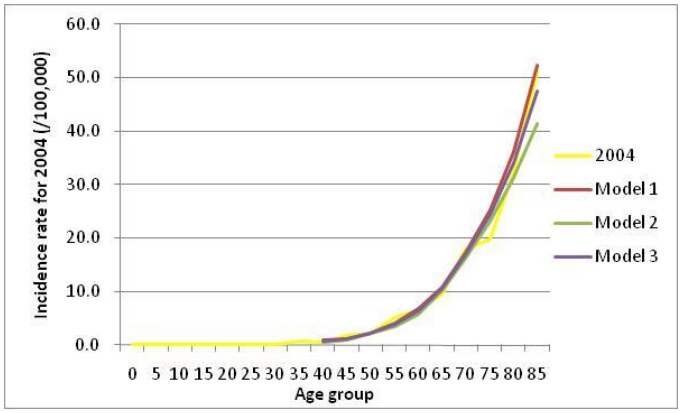
The fit is best for COVARIATE model (87%). 2004 data is well estimated by all three models. COVARIATE model has positive GDP and negative smoking. The incidence for bladder cancer is said to be taking off in 40s or 50s and gets higher when people get older. And the trajectory of k_t to 2020 is most reasonable for COVARIATE model. Therefore I would choose COVARIATE model. AR(1) predicts the 2020 incidence as same level as 2004. But since the series is a bit away from Random walk process it might be the case.

22.2. Female

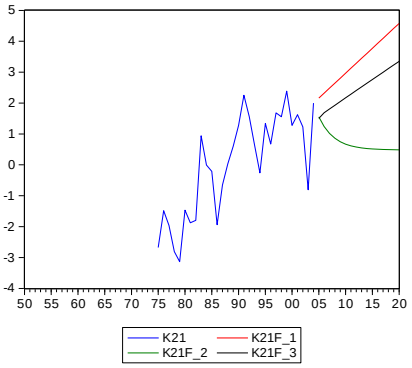
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.161 + k_{t-1}$	0.432
AR(1) model	$E(k_t) = 0.141 + 0.707k_{t-1}$	0.505
COVARIATE model	$E(k_t) = 0.000228 * GDP - 0.355 * Smoking$	0.658

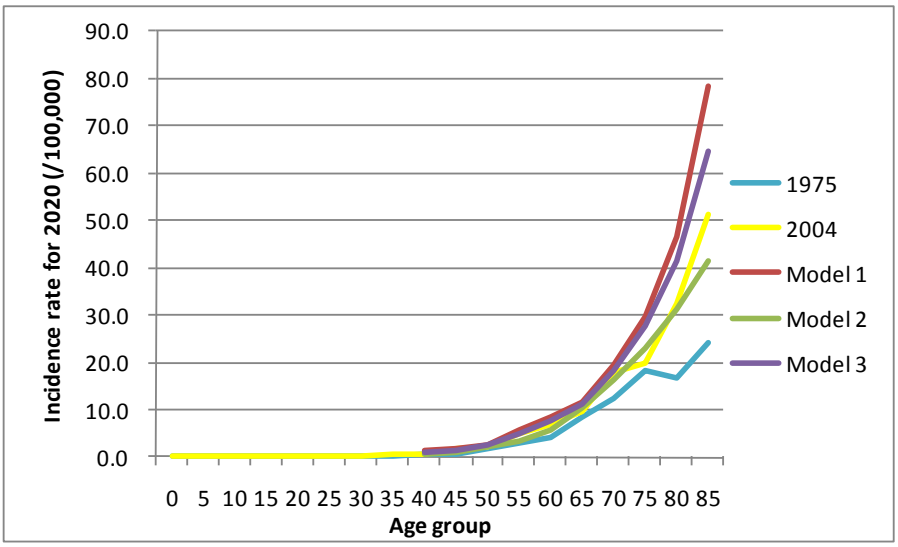
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



It is difficult. R squared is highest for COVARIATE model. Comparison with actual data for 2004 shows there is no difference between three models. They are well at predicting. No autoregressive error term is included in COVARIATE model. This means it is possible to have huge change in incidence if either GDP or smoking changes dramatically. That is where I would put some doubt. However the coefficient of AR(1) shows it is far from random walk and AR(1) model is estimating the decrease in time trend parameter k_t to 2020. Although two models are different in R squared, forecasted values are similar between Random Walk and COVARIATE model. I would choose COVARIATE model because of highest fit, capacity to capture exogenous factors and consistency with male bladder cancer.

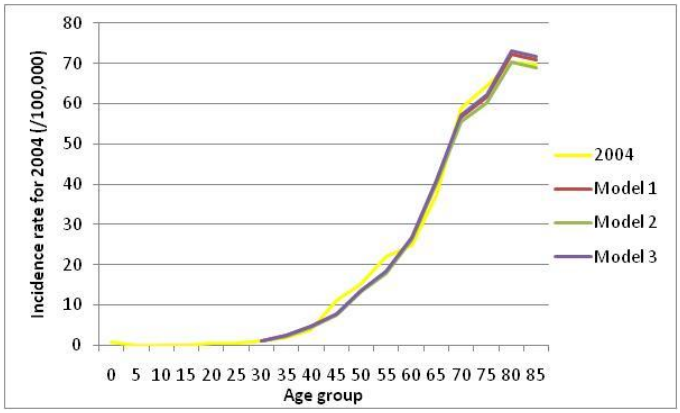
23. Kidney and other urinary organs: ICD10(C64-C66 C68)

23.1. Male

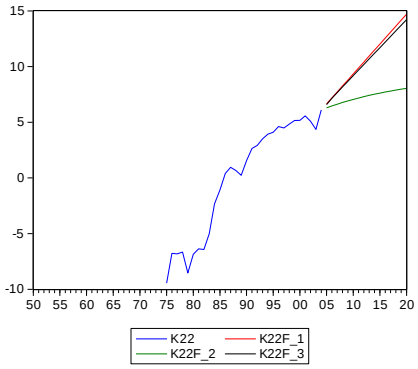
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.537 + k_{t-1}$	0.959
AR(1) model	$E(k_t) = 0.525 + 0.944k_{t-1}$	0.961
COVARIATE model	$E(k_t) = 0.000607 * GDP - 0.200 * Smoking,$ $E(\varepsilon_t) = 0.713\varepsilon_{t-1}$	0.966

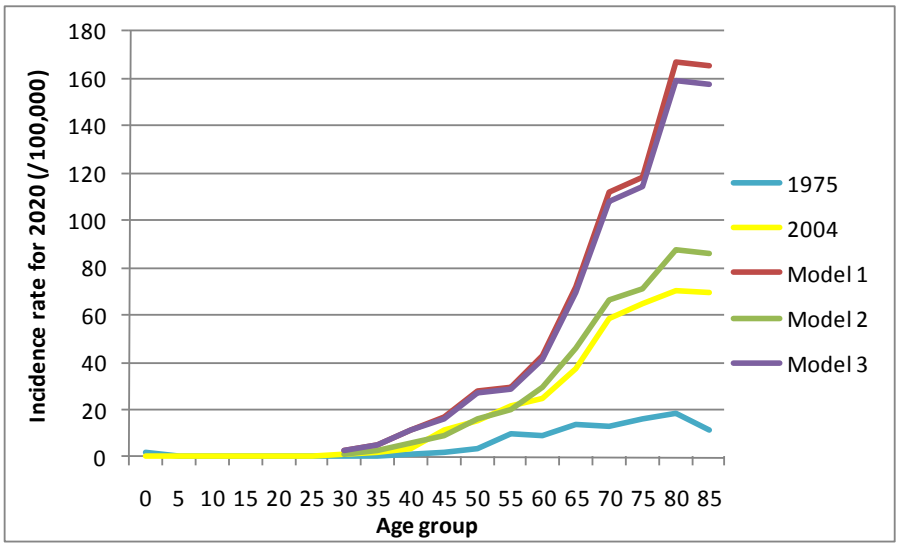
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



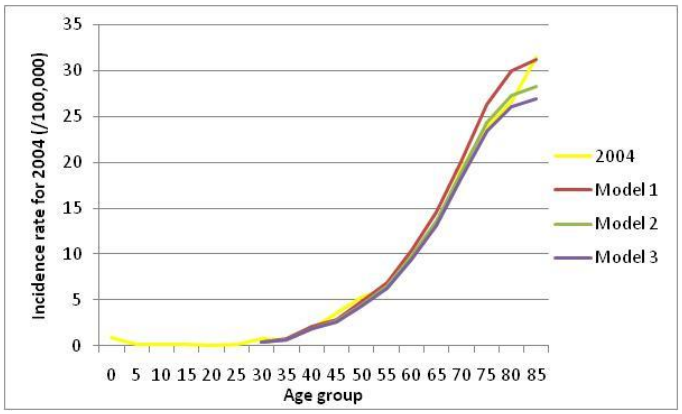
R squared is around 96% level for all models. The difference is very small. The fit to the actual data for 2004 by every model seems very good level. The COVARIATE model has positive GDP and negative smoking. I would choose COVARIATE model based on fit and its ability to take into account exogenous factors. If exogenous information is not available the choice between Random walk and AR(1) cannot be assessed. Random walk provides us with pessimistic scenario (huge increase is expected), which might be helpful from policy making perspective.

23.2. Female

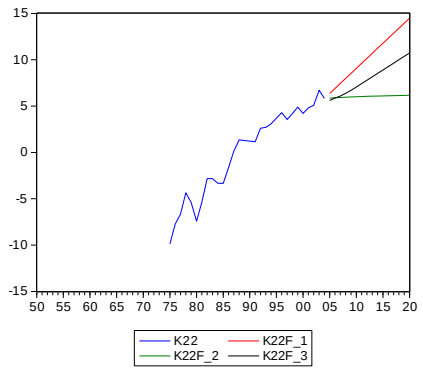
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.541 + k_{t-1}$	0.935
AR(1) model	$E(k_t) = 0.524 + 0.916k_{t-1}$	0.941
COVARIATE model	$E(k_t) = 0.000714 * GDP - 1.09Smoking,$ $E(\varepsilon_t) = 0.508\varepsilon_{t-1}$	0.951

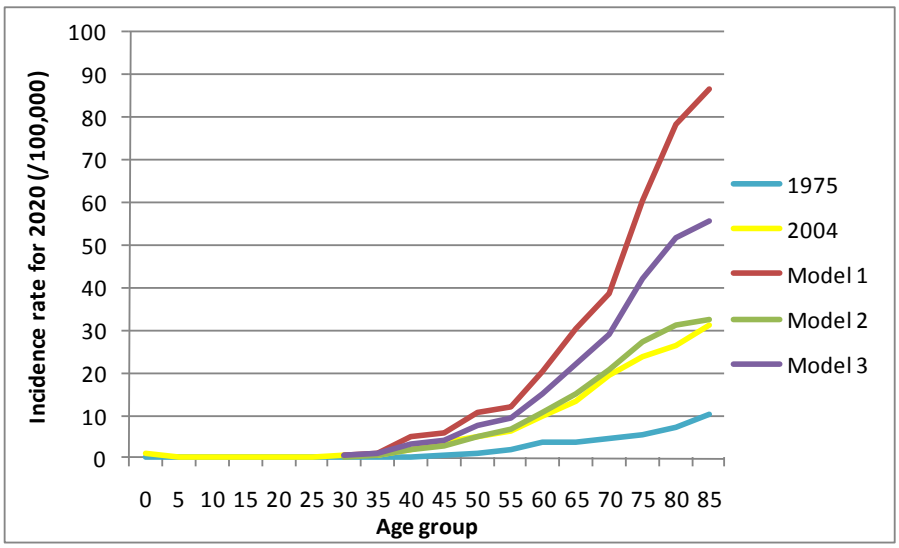
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



COVARIATE model is highest in R squared (95%) though other twos have more than 93% and comparing with 2004 data shows no difference in prediction is observed. Kidney's function is excretion of waist and it is quite natural "the older, the higher the incidence will be". Plus accumulated exposure to no-good substances might trigger the cancer so it makes sense that COVARIATE model includes both GDP and smoking. I would choose COVARIATE model based on best fit, its ability to capture exogenous factors and to be consistent with male case.

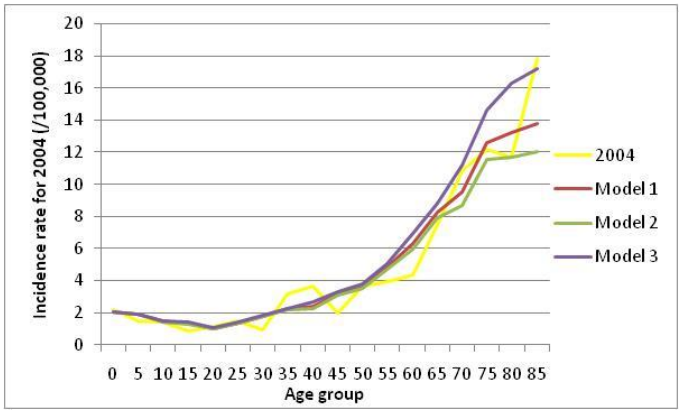
24. Brain, nervous system: ICD10(C70-C72)

24.1. Male

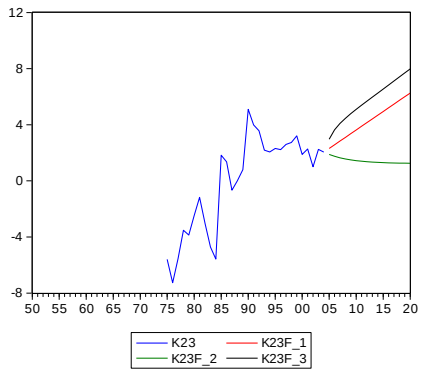
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.263 + k_{t-1}$	0.650
AR(1) model	$E(k_t) = 0.249 + 798k_{t-1}$	0.683
COVARIATE model	$E(k_t) = 0.000344 * GDP - 0.115 * Smoking,$ $E(\varepsilon_t) = 0.512\varepsilon_{t-1}$	0.718

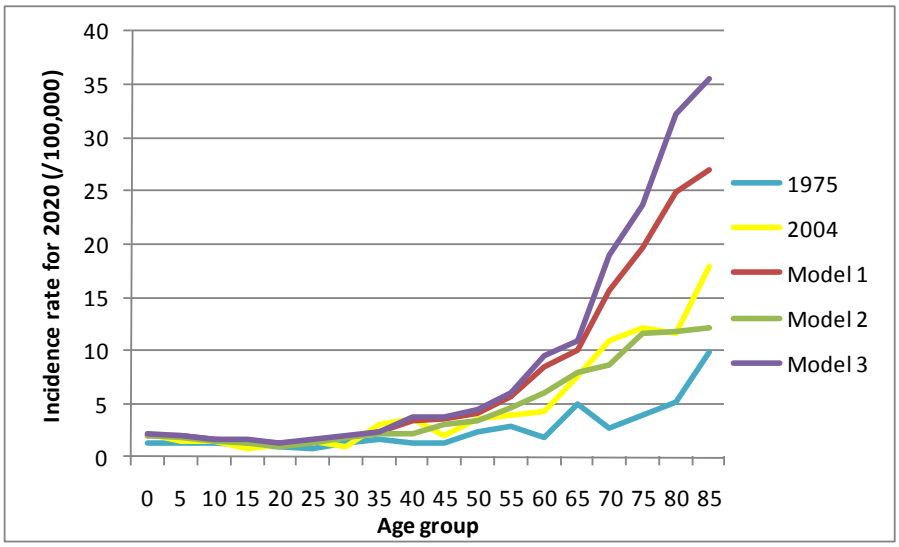
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



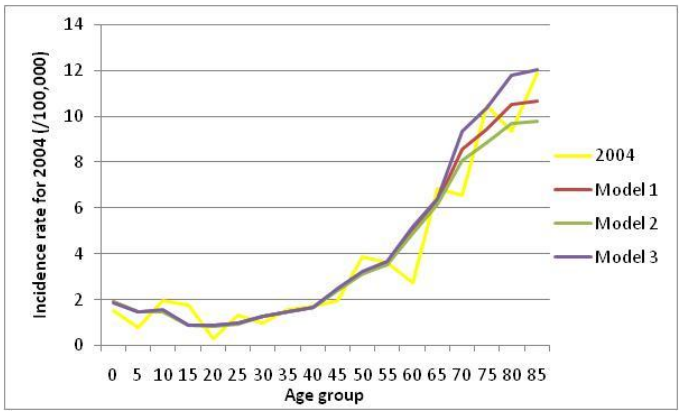
R squared is highest for COVARIATE model. Comparison between actual and expected for 2004 shows Random walk fits more for age 70s while COVARITE fits better for age 80s. COVARIATE model has positive GDP and negative smoking, and it is consistent with many of other cancers. Trajectory of k_t in the observation period seems a slow upward trend, and the forecast is assuming the trend is continuing on for Random walk and COVARIATE, while AR(1) is staying in the same level. I would choose COVARIATE model based on highest fit and its ability to capture exogenous factors. COVARIATE model provides us with most pessimistic picture for 2020.

24.2. Female

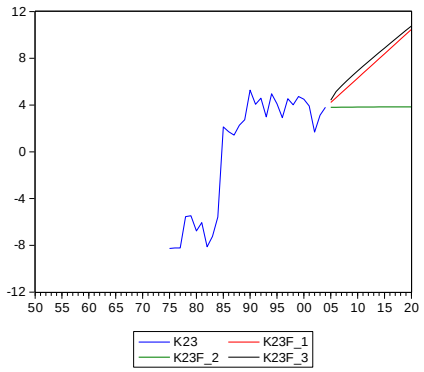
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.416 + k_{t-1}$	0.855
AR(1) model	$E(k_t) = 0.403 + 0.896k_{t-1}$	0.862
COVARIATE model	$E(k_t) = 0.000732 * GDP - 1.14 * Smoking,$ $E(\varepsilon_t) = 0.676\varepsilon_{t-1}$	0.872

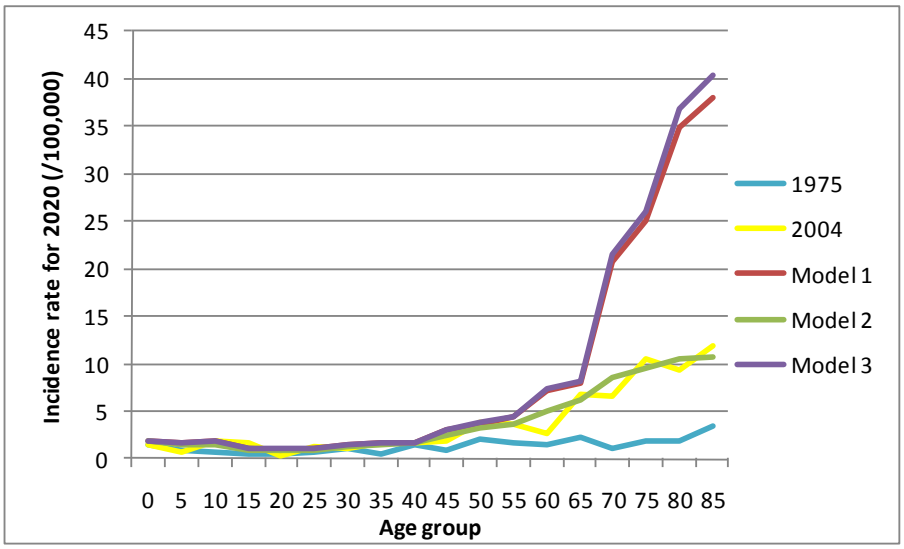
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



R squared is highest for COVARIATE model. But in contrast with male, all three models have relatively high fit (more than 85%). Random walk seems to predict most precisely the 2004 data. k_t seems to have a structural break in the 1980s or it is just a trend with high variation. According to the adjusted R squared I would pick COVARIATE model. COVARIATE model has positive GDP and negative smoking, which is also consistent with male case. It is related aging and it may be true to have more cases in the future because current level is only 10 to 100 000 population and the detection is related to diagnostic devices such as MRI, CT all of which are related to GDP (technical advances).

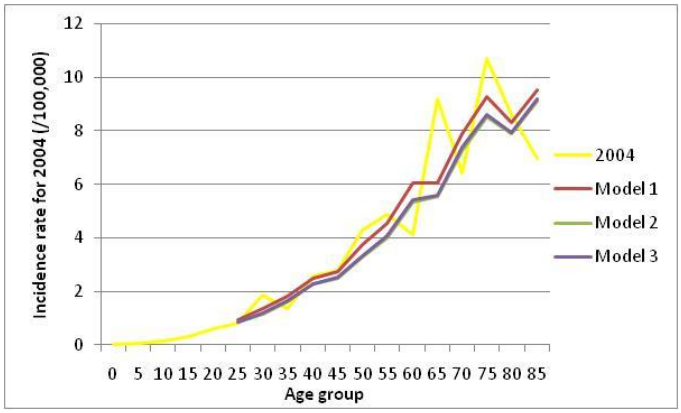
25. Thyroid: ICD10(C73)

25.1. Male

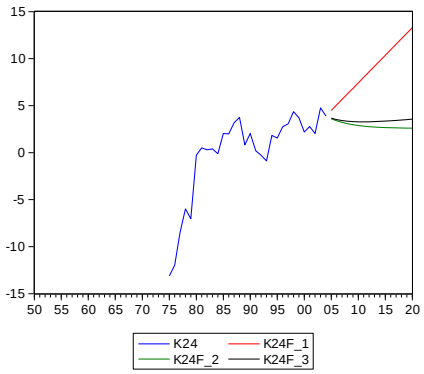
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.587 + k_{t-1}$	0.774
AR(1) model	$E(k_t) = 0.558 + 0.783k_{t-1}$	0.832
COVARIATE model	$E(k_t) = 0.0000983 * GDP,$ $E(\varepsilon_t) = 0.765\varepsilon_{t-1}$	0.833

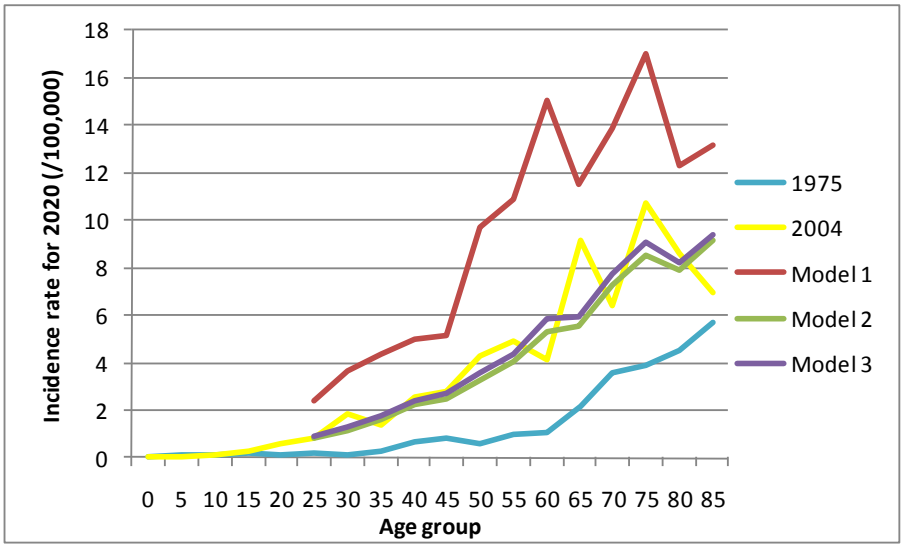
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



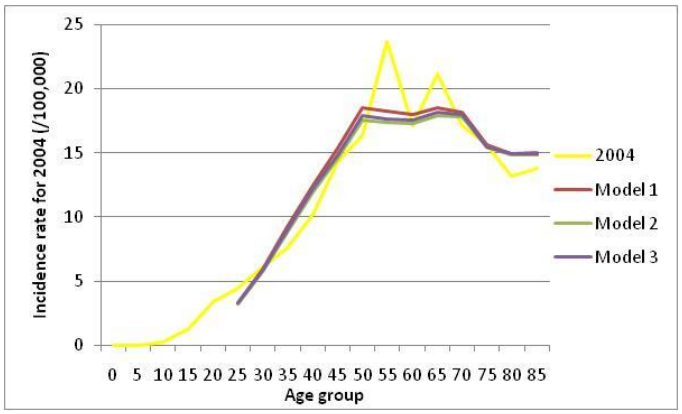
AR(1) and COVARIATE models are high in fit compared with Random walk. Bumps observed in 2004 actual data couldn't be captured by any model completely. The coefficient of AR(1) shows that it would deviate from unit root process. COVARIATE model has positive GDP as an explanatory variable and the positive sign is in line with our expectation. I would choose COVARIATE model based on fit and its ability to capture exogenous factors. However the number of cases is small (2~10 cases per 100 000) and the variation could be huge so one should be open to any other options.

25.2. Female

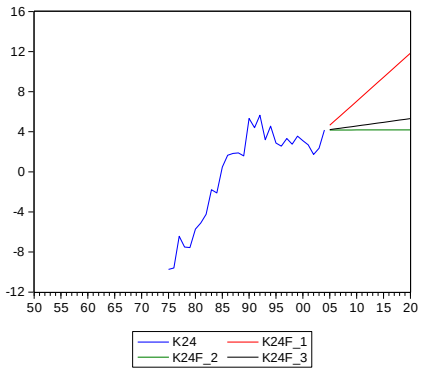
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.479 + k_{t-1}$	0.894
AR(1) model	$E(k_t) = 0.464 + 0.889k_{t-1}$	0.905
COVARIATE model	$E(k_t) = 0.000147 * GDP,$ $E(\varepsilon_t) = 0.875\varepsilon_{t-1}$	0.906

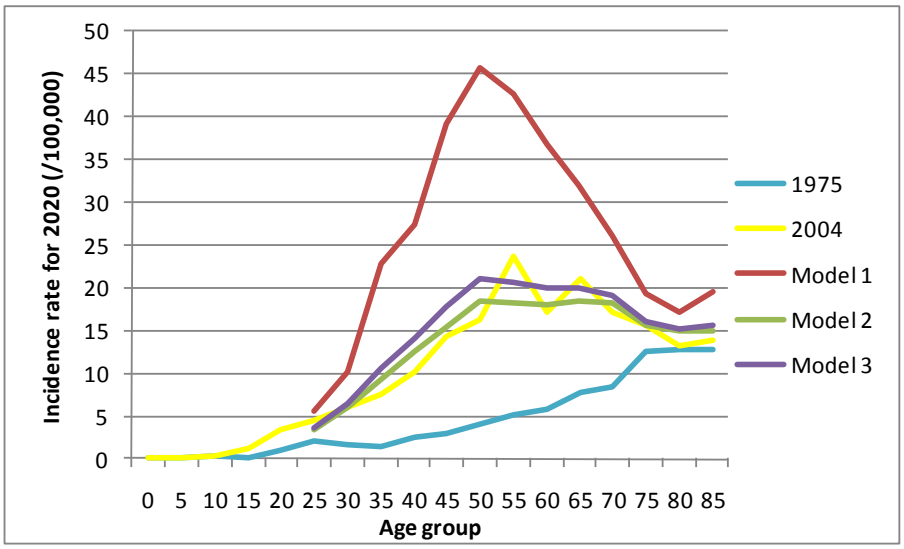
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Random walk does not seem to capture the trend. The coefficient of AR(1) is 0.88 suggesting it is a bit far from Random Walk. R squared for AR(1) and COVARIATE are similar (90%). No big difference in fit to 2004 was observed. COVARIATE model has positive GDP and the positive sign is in line with our expectation. I would choose COVARIATE model based on fit and its ability to capture exogenous factors.

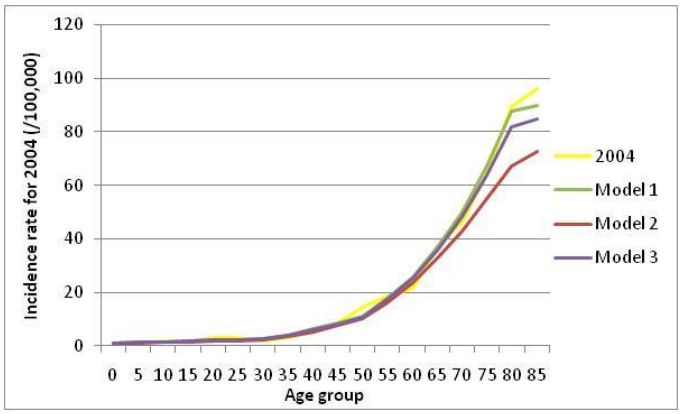
26. Malignant lymphoma: ICD10(C81-C85 C96)

26.1. Male

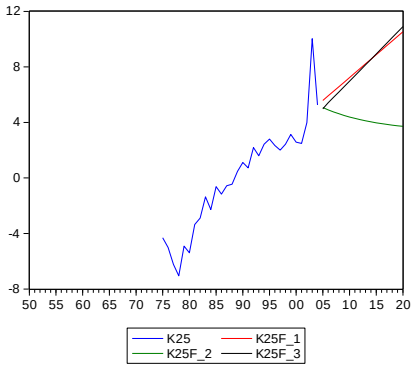
Estimated formulas for each model are shown below..

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.330 + k_{t-1}$	0.793
AR(1) model	$E(k_t) = 0.313 + 0.901k_{t-1}$	0.794
COVARIATE model	$E(k_t) = 0.000472 * GDP - 0.159Smoking$	0.847

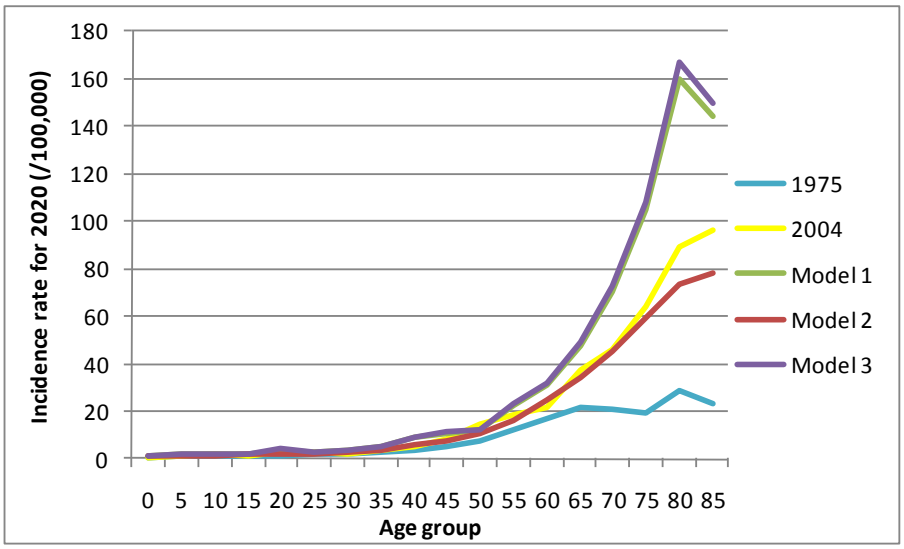
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



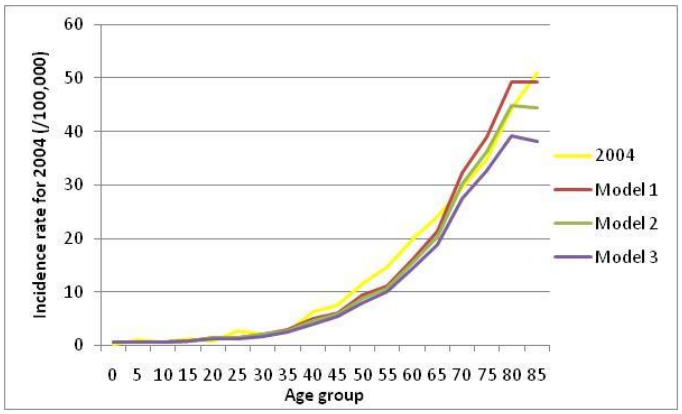
COVARIATE model with both GDP and smoking has highest fit of three models. However the structure of COVARIATE model does not include autoregressive error. It might be the case. But with this structure dramatic change in GDP or smoking would increase the expected incidence and a little doubt should be put in COVARIATE model. As we can see from the figure Random walk fits best in 2004 actual data. COVARIATE model has positive GDP and negative smoking which is consistent with many of other cancer cases. I would conclude COVARIATE model is best because of its fit and its ability to take into account exogenous factors. If exogenous information is not available, I would choose Random walk because AR(1) will have incidence for 2020 which is less than 2004 level. I think it is unlikely.

26.2. Female

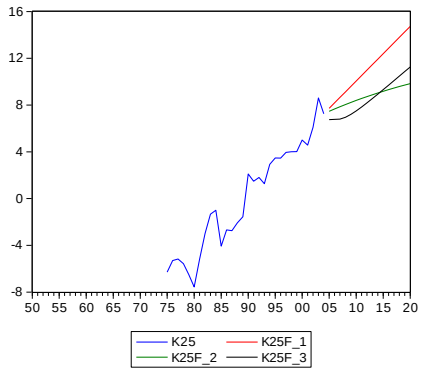
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.467 + k_{t-1}$	0.909
AR(1) model	$E(k_t) = 0.459 + 0.965k_{t-1}$	0.907
COVARIATE model	$E(k_t) = 0.000761 * GDP - 1.17 * Smoking,$ $E(\varepsilon_t) = 0.639\varepsilon_{t-1}$	0.913

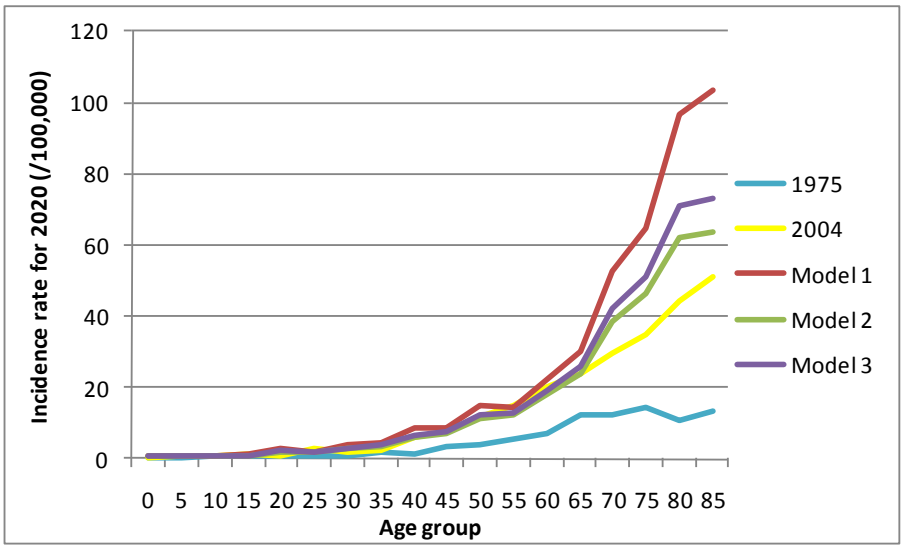
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



It is quite similar case as male. COVARIATE model has highest in fit (91%). COVARIATE model has positive GDP and negative smoking. What is good for female is that the model has autoregressive error structure left in the model. I would choose COVARIATE model based on R squared and its inclusion of exogenous factors. If exogenous information is not available it might be AR(1) or it might be Random walk.

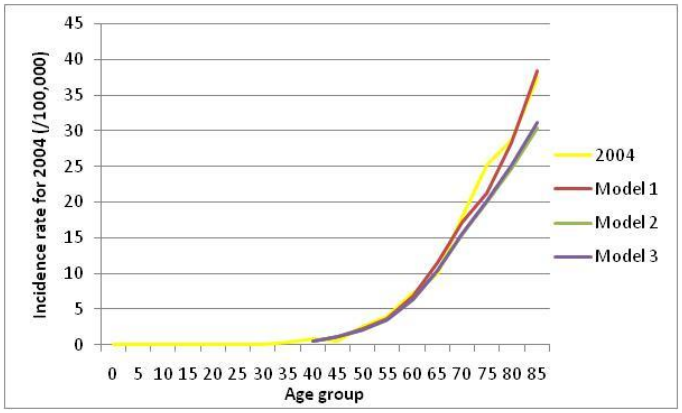
27. Multiple myeloma: ICD10(C88-C90)

27.1. Male

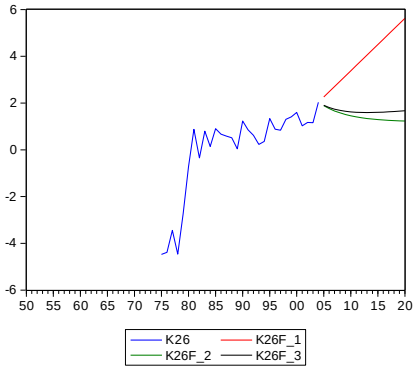
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.224 + k_{t-1}$	0.774
AR(1) model	$E(k_t) = 0.212 + 0.823k_{t-1}$	0.805
COVARIATE model	$E(k_t) = 0.0000457 * GDP,$ $E(\varepsilon_t) = 0.805\varepsilon_{t-1}$	0.807

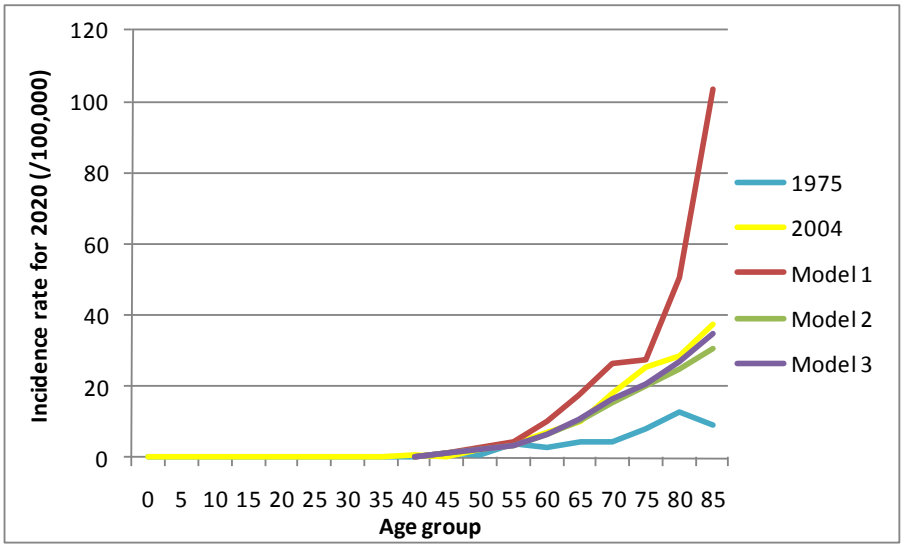
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



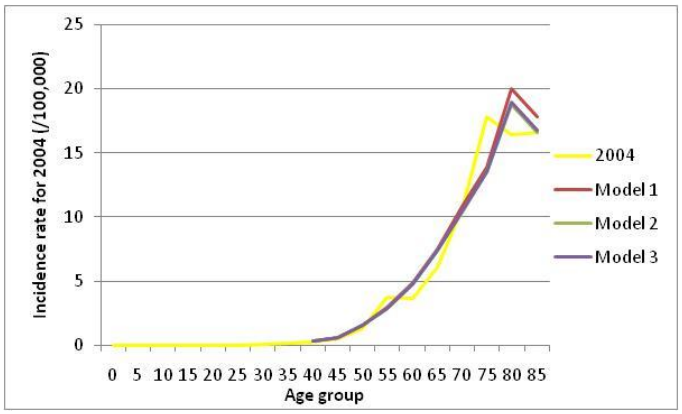
Looking at the coefficient of AR(1) the time trend parameter k_t does not seem to be unit root. The best fit was observed for COVARIATE model (80.7%). However one has to mention that Random walk seems to fit best to 2004 actual data. COVARIATE model has positive GDP which is in line with the expectation. The shape of forecasted values for 2020 does not look funny. I would choose COVARIATE model based on best fit and its inclusion of exogenous factors. The number of cases is small for that cancer and it should have huge variation. So if exogenous variable is not available one cannot throw away Random walk but I would choose AR(1) because R squared was higher and because since 1980 k_t has remained the same level and would be reasonable to stay in the same level in the future.

27.2. Female

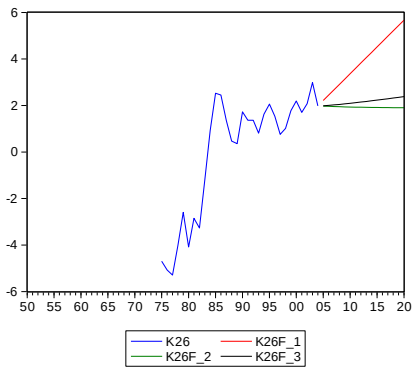
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.230 + k_{t-1}$	0.847
AR(1) model	$E(k_t) = 0.222 + 0.882k_{t-1}$	0.857
COVARIATE model	$E(k_t) = 0.000657 * GDP,$ $E(\varepsilon_t) = 0.868\varepsilon_{t-1}$	0.857

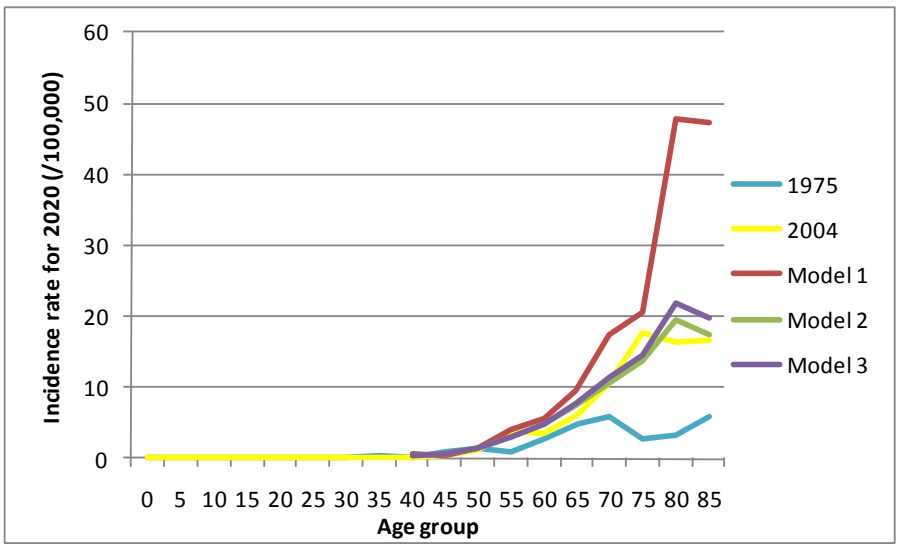
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



The situation and the structure of COVARIATE model are quite similar for female with the male's case. Although the difference is only a little COVARIATE model has best fit. Although the p-value was 0.25 I included GDP in COVARIATE model for the purpose of consistency with male. P is relatively large but the coefficient itself is not small (standard error is large). The shape of forecasted values for 2020 does not look funny. I would choose COVARIATE model based on best fit and its inclusion of exogenous factors. If exogenous variable is not available I would choose AR(1) based on R squared and the trajectory of k_t. Considering the rareness of the disease expected incidence of more than 40 seems too high.

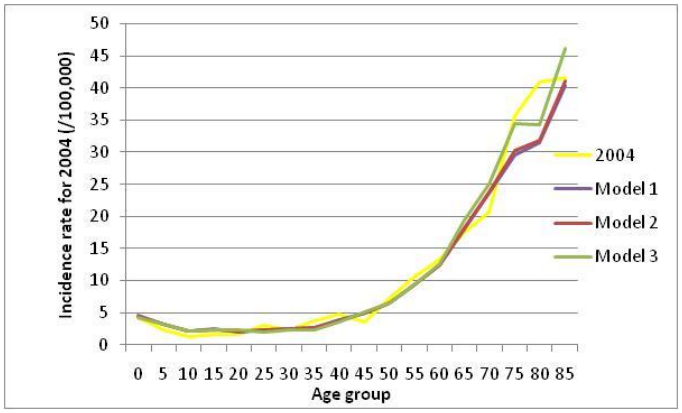
28. Leukemia: ICD10(C91-C95)

28.1. Male

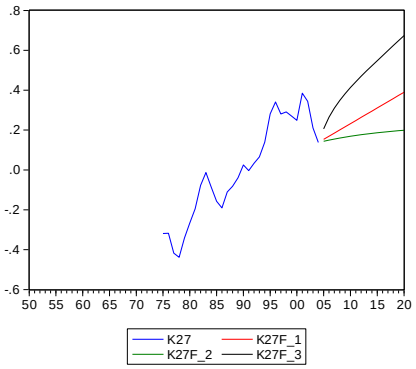
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.157 + k_{t-1}$	0.910
AR(1) model	$k_t = 0.0154 + 0.933k_{t-1}$	0.912
COVARIATE model	$E(k_t) = 0.0000294 * GDP - 0.0101 * Smoking,$ $E(\varepsilon_t) = 0.674\varepsilon_{t-1}$	0.920

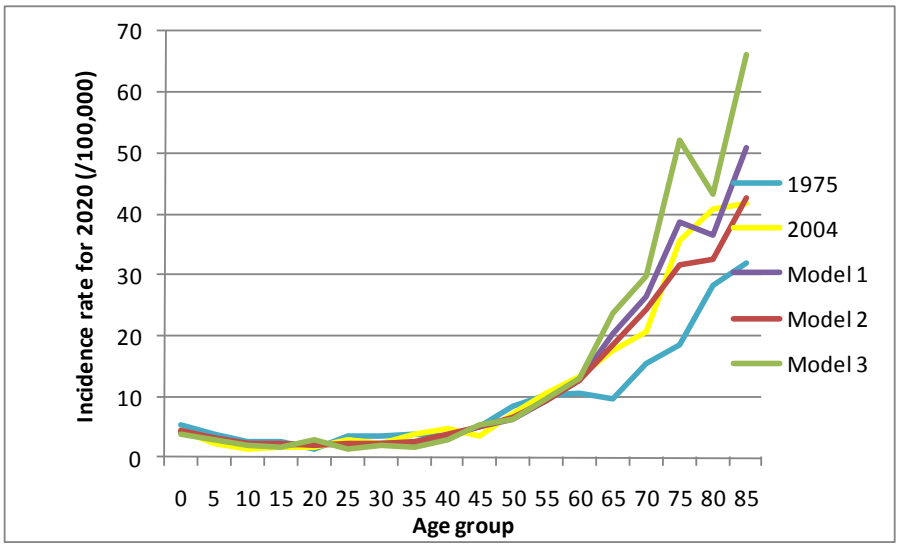
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



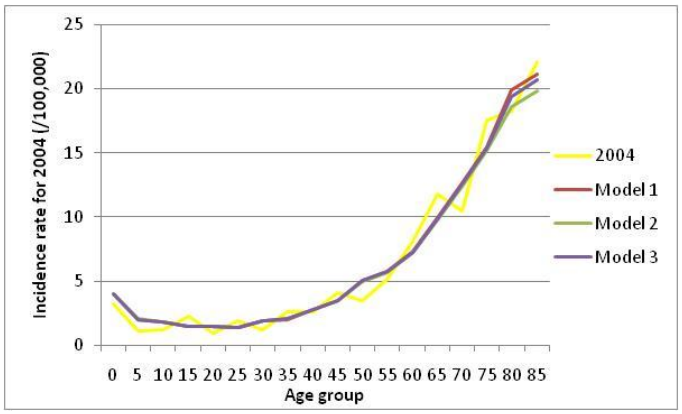
It might not be a good idea to include smoking because leukemia is common for small children as well as elderly. Whatever model we use R squared was more than 90% for all models. 2004's data is most close to the estimates from COVARIATE model. Looking at the forecasted incidence for 2020 by three models does not show any strangeness in shape. So I would choose COVARIATE model based on fit and its ability to control exogenous factors. Positive GDP and negative smoking in COVARIATE model is consistent with most of other cancer cases. If GDP data or smoking data is not available, I would choose Random walk because the coefficient of AR(1) is 0.93 and it would be reasonable to guess that the incidence will increase in 2020 (Random walk) than decrease (AR(1)).

28.2. Female

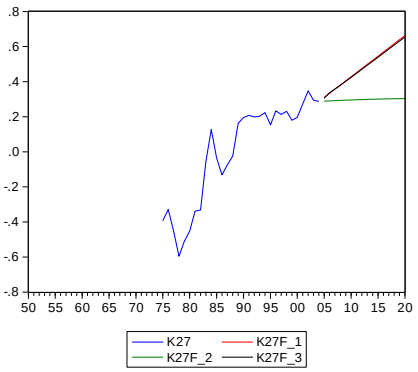
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.0234 + k_{t-1}$	0.882
AR(1) model	$E(k_t) = 0.0228 + 0.927k_{t-1}$	0.883
COVARIATE model	$E(k_t) = 0.0000447 * GDP - 0.0695 * Smoking,$ $E(\varepsilon_t) = 0.659\varepsilon_{t-1}$	0.892

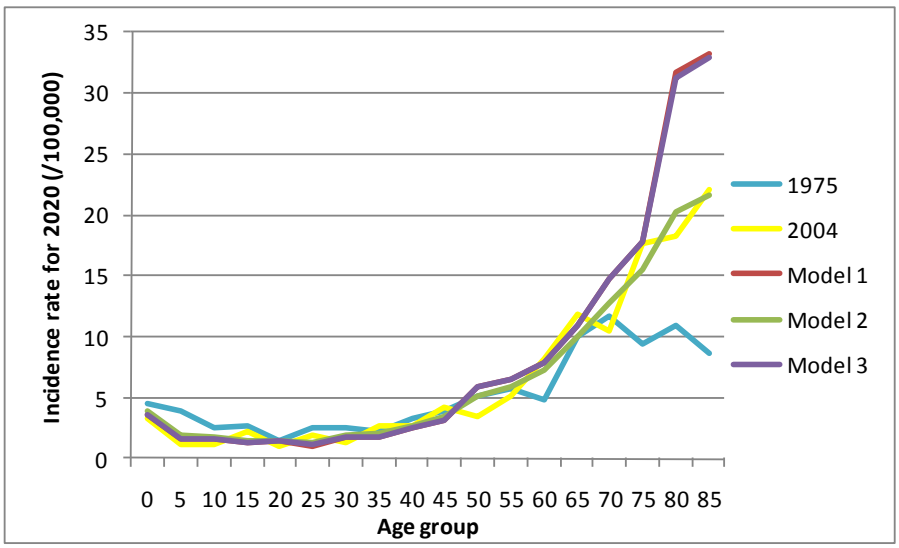
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Whatever model we use R squared was bit less than 90% for all models. All models predicted quite well 2004 data. Looking at the forecasted incidence for 2020 by three models does not show any strangeness in shape. So I would choose COVARIATE model based on fit and its ability to control exogenous factors. Positive GDP and negative smoking in COVARIATE model is also consistent with most of other cancer cases. If exogenous variable is not available I cannot choose one model. However according to the forecast based on Random walk the incidence for 2020 will be at most less than 35 per 100 000 population for the age group 80 or 85, and this gives the useful information to policy making.

29. Lower digestive organ (Colon and Rectum together) : ICD10(C18-C21)

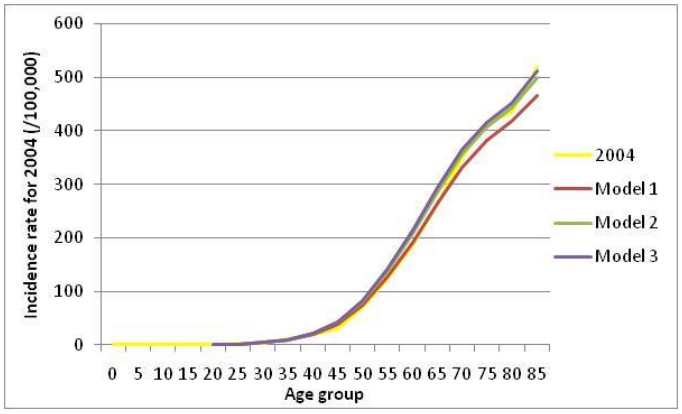
In clinical practice colon and rectum are treated in the same department (ex. Chemotherapy does not differentiate between colon and rectum). In this chapter I will focus lower digestive organ, colon and rectum together.

29.1. Male

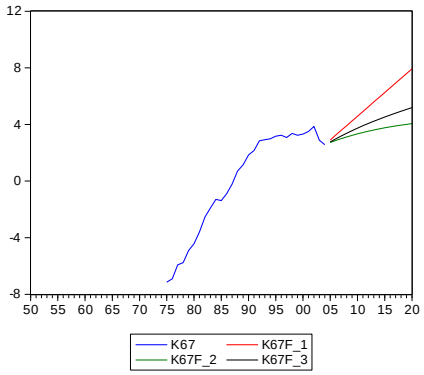
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.335 + k_{t-1}$	0.983
AR(1) model	$E(k_t) = 0.329 + 0.931k_{t-1}$	0.988
COVARIATE model	$E(k_t) = 0.000159 * GDP,$ $E(\varepsilon_t) = 0.924\varepsilon_{t-1}$	0.989

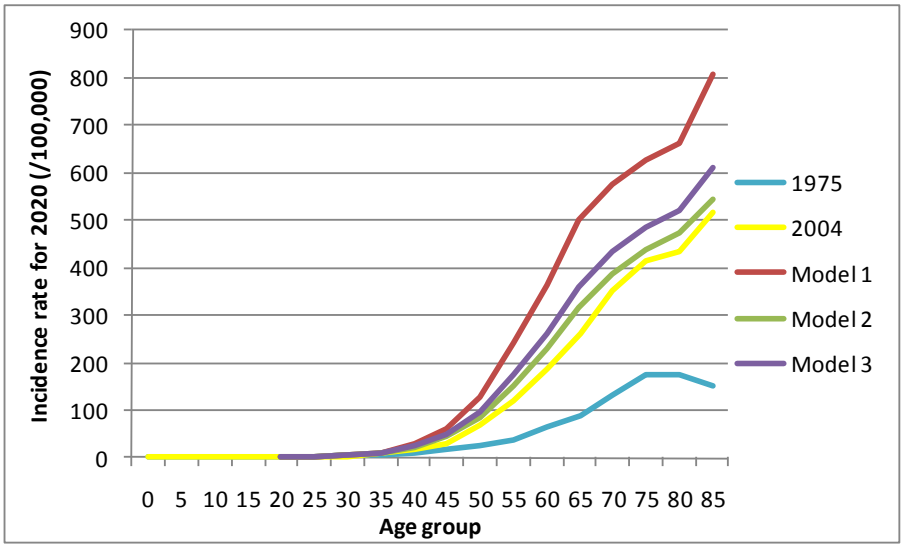
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are show simultaneously)



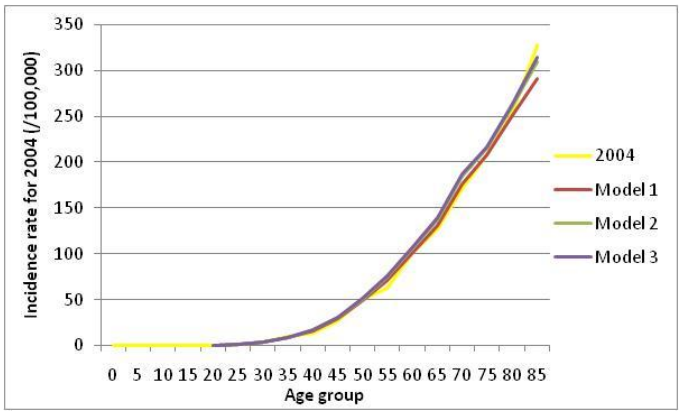
All models have 98% in fit. COVARIATE model has GDP left in explanatory variable. All models fit quite well to the actual data for 2004. As explained in colon and rectum cancer, life style change has been said to be related to this change life style change coincides with GDP. I would choose COVARIATE model based on its ability to capture exogenous factors. No matter what model we chose the incidence will increase in 2020. The result is consistent with the results of analyses for colon and for rectum.

29.2. Female

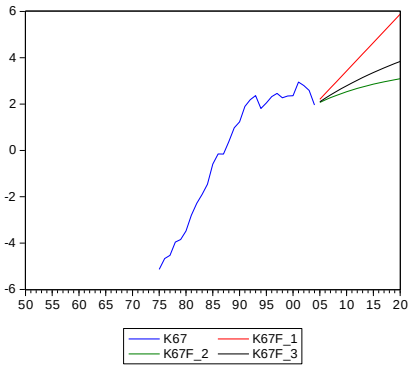
Estimated formulas for each model are shown below.

Models	Estimates	Adjusted R^2
Random Walk model	$E(k_t) = 0.245 + k_{t-1}$	0.980
AR(1) model	$E(k_t) = 0.24 + 0.935k_{t-1}$	0.985
COVARIATE model	$E(k_t) = 0.000118 * GDP,$ $E(\varepsilon_t) = 0.927\varepsilon_{t-1}$	0.985

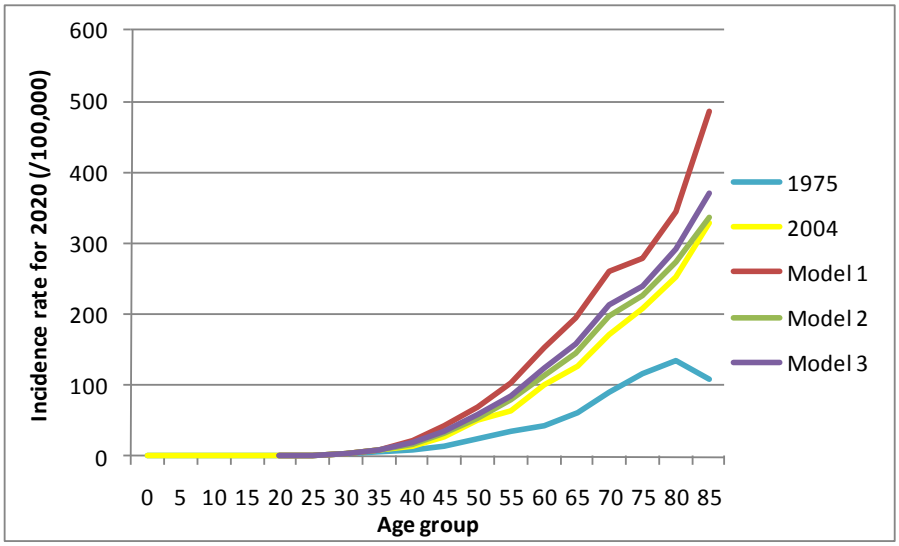
Comparison between data and estimates(2004)



Trajectory of K (time trend parameter)



Forecasting 2020 (for the purpose of comparison actual data for 1975 and 2004 are shown simultaneously)



Female case is quite similar as male case. Every model has more than 98% in fit and almost perfectly predicts 2004 data. Regardless of the choice of model the forecasted trend seems to be probable for all models. COVARIATE model has positive GDP effect and it is in line with our expectation. I would choose COVARIATE model. COVARIATE model was consistently chosen for colon cancer and rectum cancer. It is natural that the accumulation of these two would be explained by the same model, regardless of sex. Number of cases is large for cancer in lower digestive organ and it will be continuing on being an issue for public health, medicine (clinical practice) and health care expenditures.

DISCUSSION part

I applied Lee-Carter model to the incidence data for cancer in Japan. There is for example Gompertz model (1825) or Heligmand and Pollard model (1980) which could be applicable to the incidence data. However both models are parametric models and assuming the specific shape of mortality age profile. For example the shapes (graphs with age specific incidence rate on y-axis and age on x-axis) are quite different between uterus cancer and lung cancer. What is advantageous for Lee Carter model is the capability to express all types of age specific rates because it is a non-parametric model. If the shape changed over time there is no way to forecast (ex. female esophagus cancer).

Suppose the shape of age specific incidence rate does not change over time, Lee-Carter would be useful tool except the case the change during the observation period is drastic (ex. female gallbladder cancer. But this is not the problem only for Lee-Carter but any forecasting methods). In Lee-Carter model there is one parameter k_t to express the time trend. According to our analyses our time trend parameter k_t. is close to unit root. However it should not be restricted only to random walk but open to the option of other models such as AR(1) or COVARIATE model.

Our analysis shows each series would be better seen as individual case. There is a tendency for example that sex hormone related cancer such as breast cancer, uterine cancer and prostate cancer are all inclined to have unit-root. But as a rule of thumb seeing individually is recommended. This seems reasonable because the disease called cancer is the group of diseases which show uncontrollable cells growth and the characters of cancers (cause, development speed, proliferation speed, metastases and so on) differ significantly between different sites of origin.

According to my forecast, taking no exogenous factors into account is well capturing the dynamics of cancer incidence series. I would apply AR(1) as well as Random walk because even if it seems unit root sometimes it is just very close to unit root but not the same as unit root. Two models have different when it comes to extrapolation (forecasting the future), so it would be better to calculate both and compare.

It would be nice to include exogenous variables into the model to improve the precision if the exogenous information is available. I called it COVARIATE model and use it as one alternative to random walk model (= original Lee Carter model). I included only GDP and smoking effect. If one could think of other indicators which might affect the incidence rate, it might be interesting to include those as well. Smoking has known as a risk factor for cancer and it would be easy if we had positive coefficient. However what we got was negative for many cases. I have two reasons for this. The proportion of people who are smoking is affected by the political campaign and the tax rate imposed on tobacco. In other words it may be the case that GDP and smoking proportion may not be independent (Even if smoking affects the incidence rate in micro level, it is a different story when it comes to macro level). Other could-be explanation is the following. Smoker's proportion was high with 80% for male and 16% for female in 1965 when people died of diseases other than cancer. But in 2009 only 40% of male and 12% of female are smokers when people die of cancer most often. I guess there were many people who died of non-cancer disease in the past who would have died of cancer if they had lived in the 21st century. If this is the case smoking was looking at the era's profile. When smoker's proportion was high, there were many people who died of infectious disease or stroke but

would have had cancer if they had survived. Negative sign for smoking means that the data is from the era in which there are many who can die from diseases other than cancer. Smoking's harm for cancer incidence must be looked by micro level analysis or by using different measure other than smoker's proportion.

What about GDP? Should it be included as it is in our COVARIATE model? Or is it unnecessary? I would like to touch GDP's role in health care. Newhouse (1977) drew two implications about health care expenditures.

1. Rich country (high GDP) tends to pay more on health care.
2. Institutional factors are endogenous factors to health care cost.

Kenjo (2005)[7] showed there is no evidence to show a difference in mortality and incidence between countries with high health care expenditures and low health care expenditures. In other words there is not a difference in incidence and mortality between high GDP countries and low GDP countries (among OECD countries). However cancer incidence is related to screening tests (most likely provided by the government and taking screening tests or not is up to individual so that public awareness is also important) and longevity in life and from this point of view GDP seems to be important. My conclusion about exogenous factors is we should include GDP while it is questionable for smoker's proportion when it comes to macro level analysis because smoking proportion is affected by the institutional factor (tax and campaign) or might be just looking at the difference in era. It has a possibility to correlate with GDP (multi-collinearity problem) too.

GDP is most of the cases positive coefficient because I guess GDP is partly the proxy of longevity (Getting old simply the risk factor of many cancers. But there are cancers not mainly affected by aging such as cervix cancer which is related mainly by sex behavior and its peak is relatively in the younger age groups). Some cancer such as stomach cancer gets negative GDP. GDP can increase the health care expenditure. It can increase the early detection (to increase the incidence) or it creates better prevention effect for the whole society (to decrease the number of cancer patients). In addition to that, life style change which has not been taken into account in our analyses may affect this phenomenon. As our lifestyle has been westernized cancer in colon and rectum has been more common than before while stomach cancer has been on decreasing trend. I think GDP should be included whether it gets positive or negative coefficient.

Therefore I would conclude that COVARIATE model with explanatory variable of GDP plus taking into account the correlation between errors should be the first choice. Smoking (the proportion of smoker in the society) could be included however it has to be chosen based on individual cancer case because sometimes it might be endogenous variable or it might be simply not related. If COVARIATE model does not work or exogenous variables are not available, then both AR(1) and Random-walk (original Lee-Carter) would be considered as our second choice and see their similarity and divergence.

[7] Kenjo (2005), p.191-192, Figure 3. He used the OECD data of 1970 and 1998, and calculated correlation coefficient between health care expenditures and the index such as infant mortality, perinart mortality, maternal mortality, low weight birth rate, life expectancy for women and life expectancy for men. Sample sizes are around 10 to 20 because of the availability of those health related index and the coefficients are in the range of ± 0.5.

ACKNOWLEGEMENT

I appreciate my supervisor Panagiotis Mantalos for the helpful comment and suggestion for the thesis. Also I thank every staff and teacher of courses in Economic demography program for their supports and instructions in the last two years.

DATA AND REFERENCE

Cancer incidence data

Matsuda T, Marugame T, Kamo K, Katanoda K, Ajiki W, Sobue T, and The Japan Cancer Surveillance Research Group. (2008) "Cancer incidence and incidence rates in Japan in 2002: based on data from 11 population-based cancer registries". Japanese Journal of Clinical Oncology, 38: 641-8.

http://ganjoho.jp/professional/statistics/statistics.html

GDP data

Data from Penn World Tables and from Mr. Michael Bordro (University of Pennsylvania) are by country from 1950

http://pwt.econ.upenn.edu/php_site/pwt_index.php

Tobacco data (sales and smoker's rate)

Sales data:

Japan Monopoly Corporation (Predecessor Company to JT). Sales data found between 1950 and 1974 (not deflated, this is the document for annual shareholders meeting at that time). At JT inc. website I found sales data between 1990 and 2008.

Smoker's rate:

http://www.health-net.or.jp/tobacco/product/pd090000.html

Reference

Gompertz, B. (1825). "On the nature of the function expressive of the law of human mortality, COVARIATEn a new mode of determining the value of life contingencies". In a letter to Francis Baily, Esc. F.R.S. Phil. Trans. R. Soc. London 115, 513-585.

Heligman L, Pollard JH. (1980)."The Age Pattern of Mortality". Journal of the Institute of Actuaries, 107: Part 1: 49-80.

Hirose M, Imanak Y, Ishizaki T, Evans E. (2003) “How can we improve the quality of health care in Japan? Learning from JCQHC Hospital Accreditation”. Health Policy no.66 29-49

Kenjo (2005) “Redistribution political economics: social security and health care in Japan” Keio university publication. (This book is in Japanese; title is 再分配政策の政治経済学)

Kogure , Hasegawa. (2005) “Statistical modeling of future life table: Lee-Carter model and its extension”. Keio University Shonan Fujisawa Campus, Faculty of policy management, working paper: 71(in Japanese).
http://coe21-policy.sfc.keio.ac.jp/ja/wp/WP71.pdf

Lee RD, Carter LR. (1992). “Modeling and Forecasting U.S. Mortality”. Journal of the American Statistical Association, 87: 419: 659-675

Lee RD. (2000). “The Lee-Carter Method for Forecasting Mortality, with Various Extensions and Applications”. North American Actuarial Journal, 4: 1: 80-93

Murray CJL, Lopez AD. (1997a). “Mortality by cause for eight regions of the world: Global Burden of Disease Study”. Lancet, 349: 9061: 1269-1276

Murray CJL, Lopez AD. (1997b). “Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: Global Burden of Disease Study”. Lancet, 349: 9062: 1347-1352

Murray CJL, Lopez AD. (1997c). “Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study”. Lancet, 349: 9063: 1436-1442

Murray CJL, Lopez AD. (1997d). “Alternative projections of mortality and disability by cause 1990-2020: Global Burden of Disease Study”. Lancet, 349: 9064: 1498-1504

Newhouse (1977). "Medical-Care Expenditure: A Cross-National Survey". The Journal of Human Resources, 12: 1: 115-125

OECD (2006) “Projecting OECD health and long-term care expenditures: What are the main drivers?” Economic department working paper No.477

APPENDIX FILES

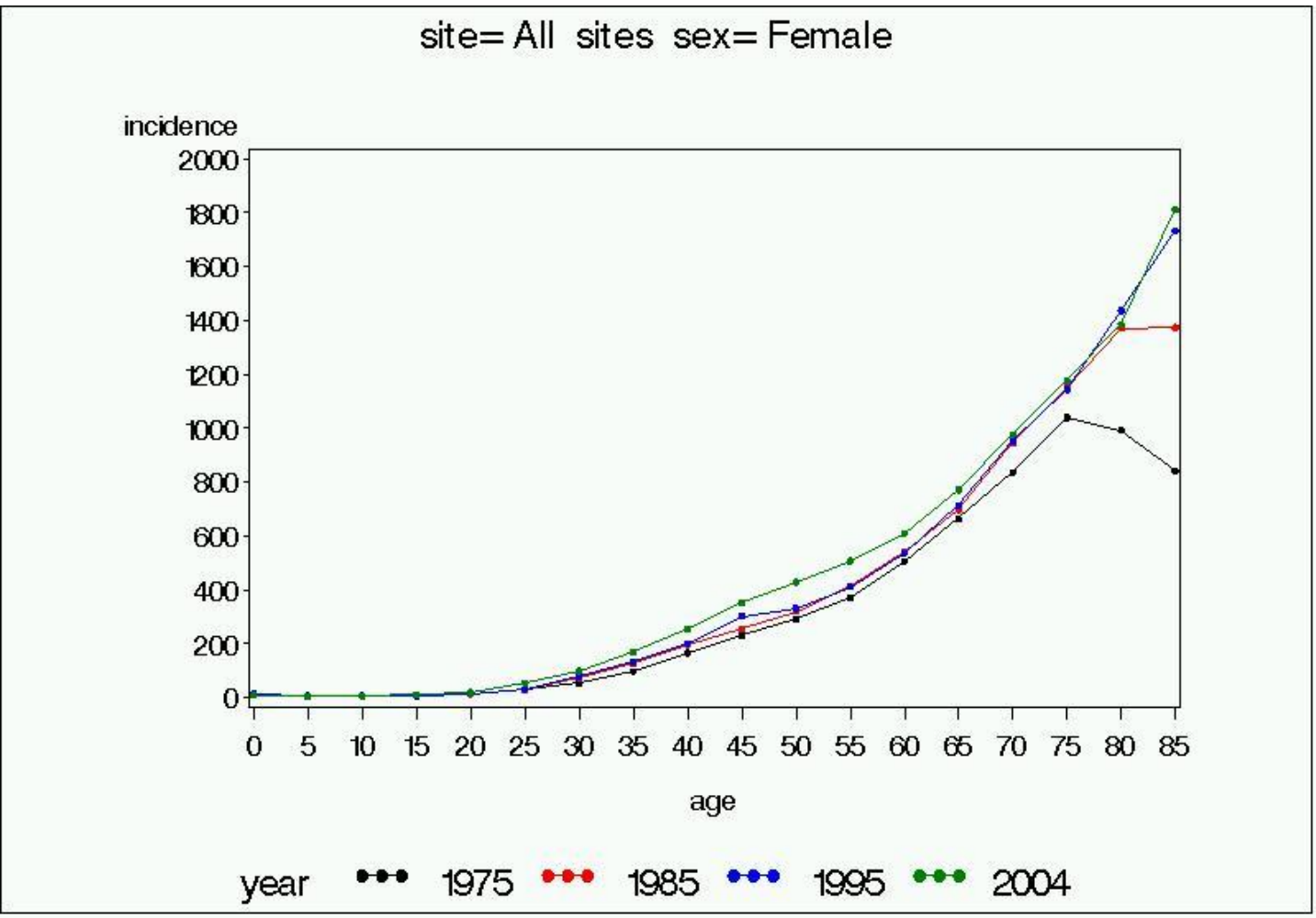


Figure. 1

APPENDIX FILES

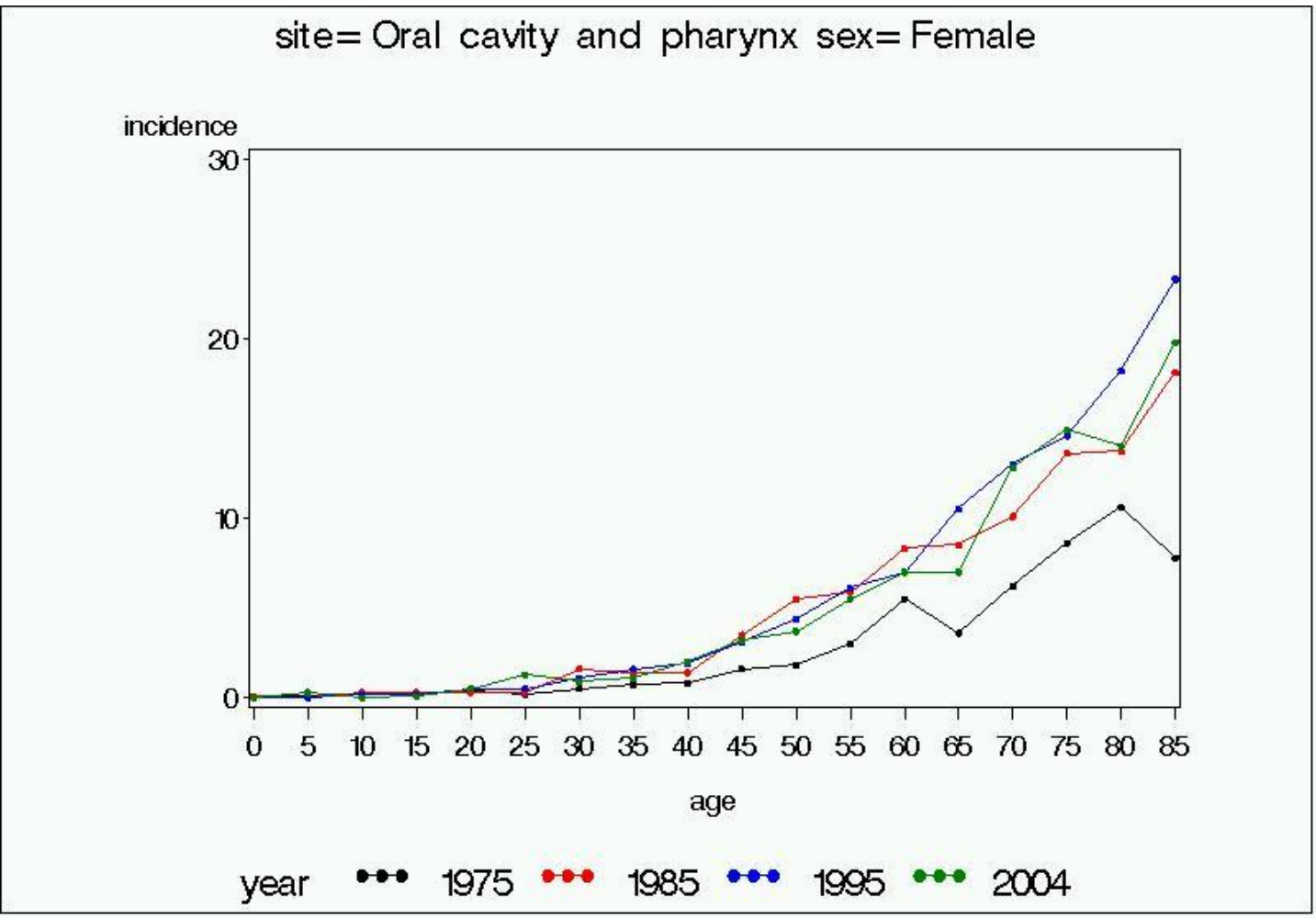


Figure. 2

APPENDIX FILES

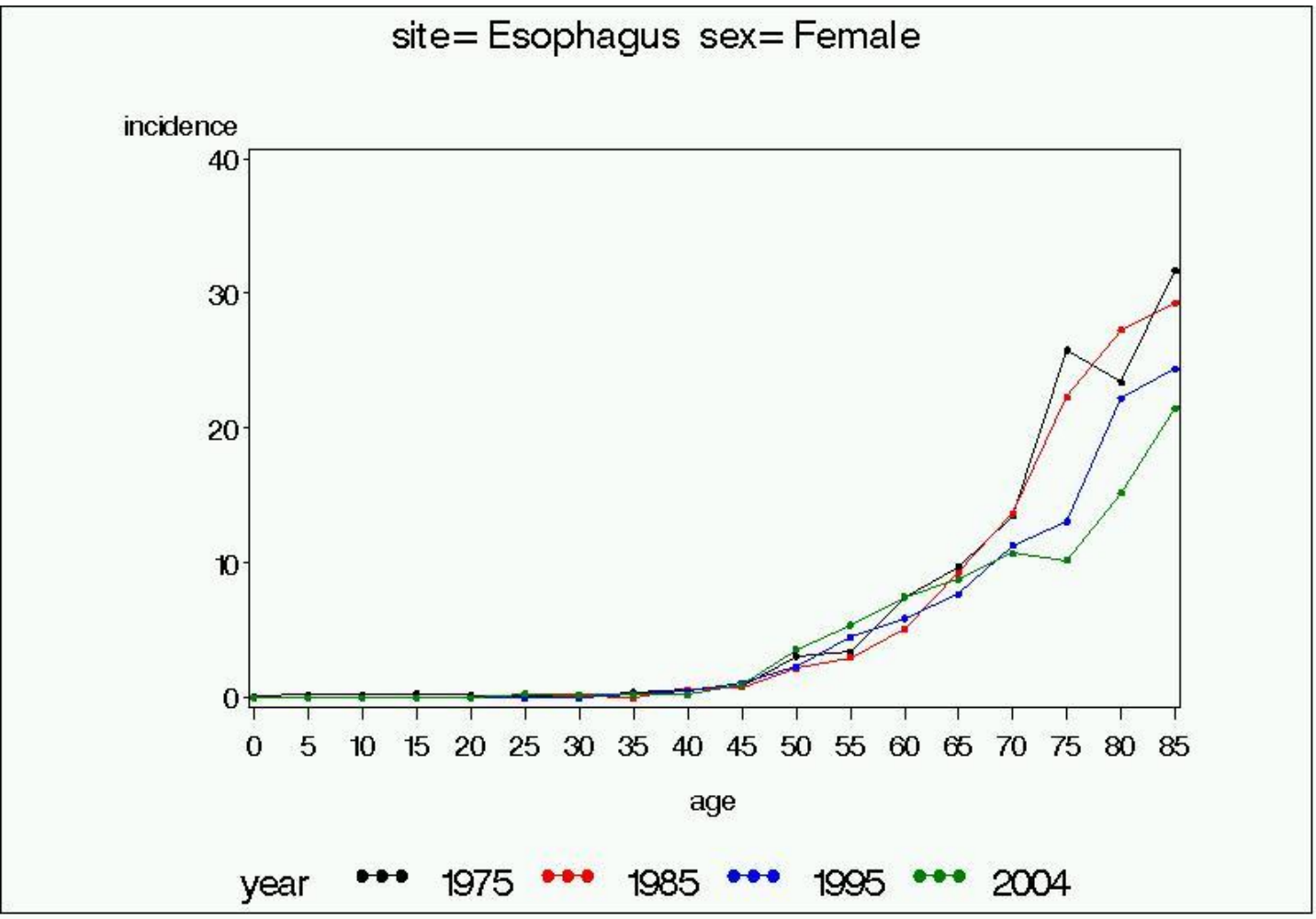


Figure. 3

APPENDIX FILES

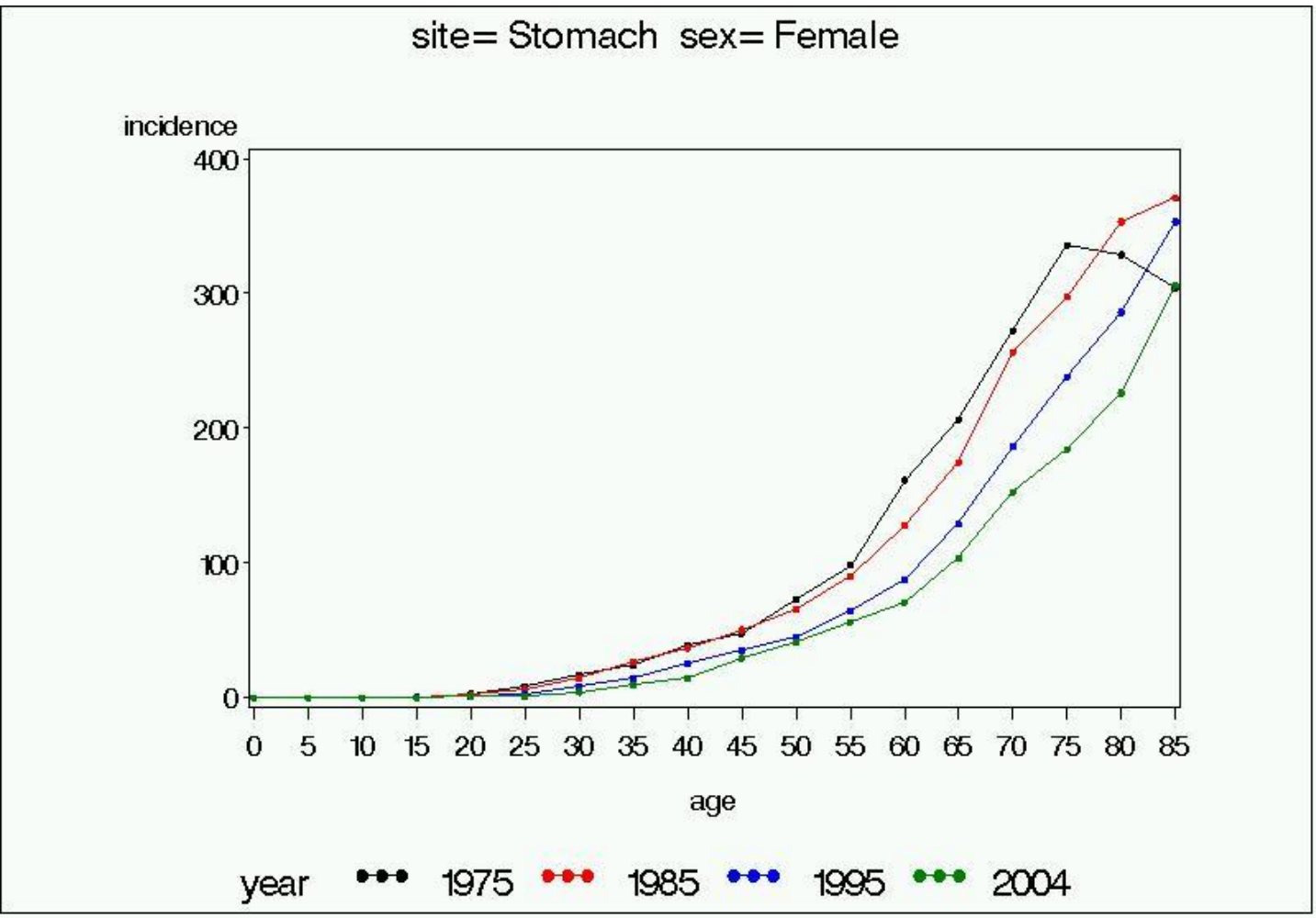


Figure. 4

APPENDIX FILES

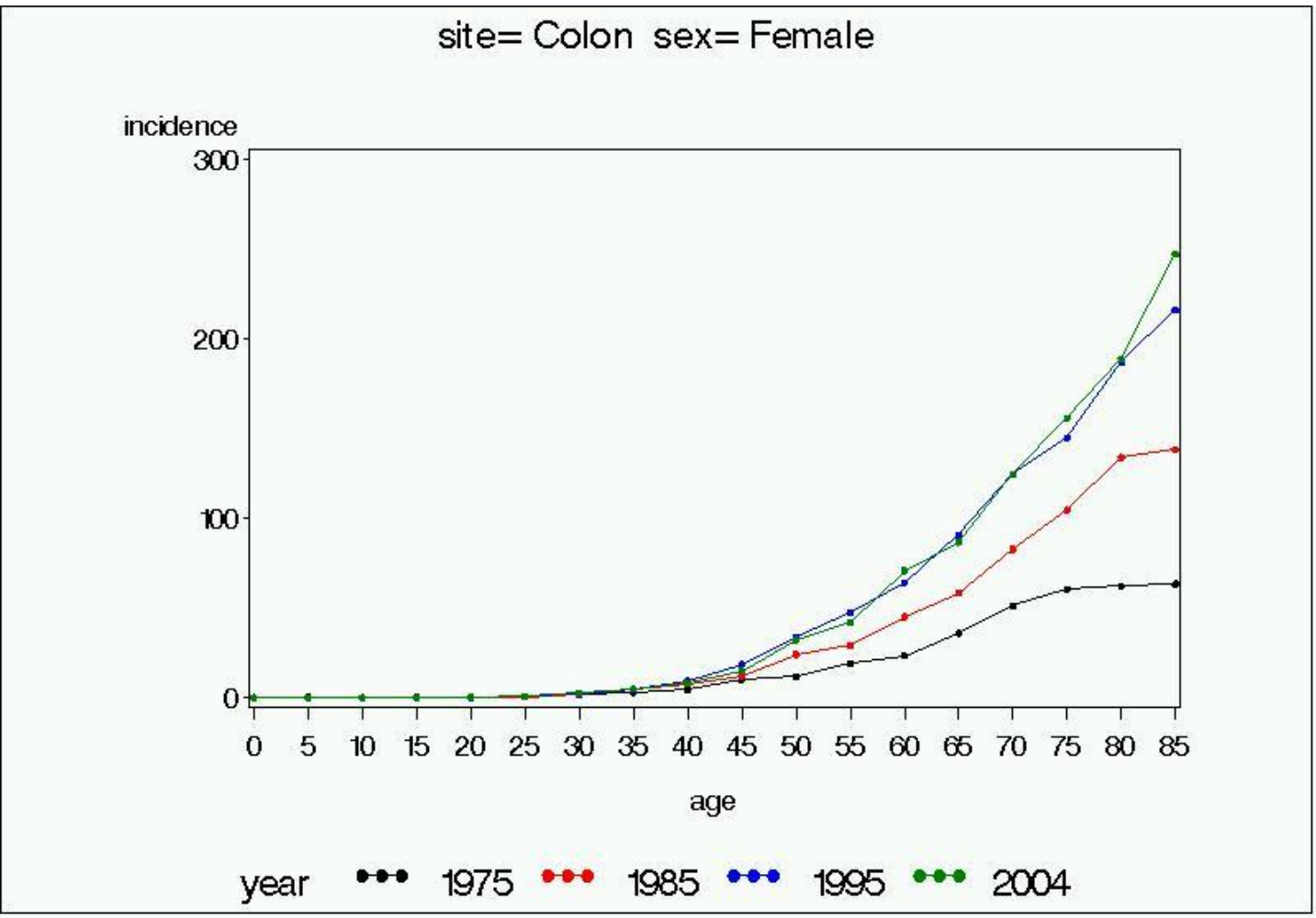


Figure. 5

APPENDIX FILES

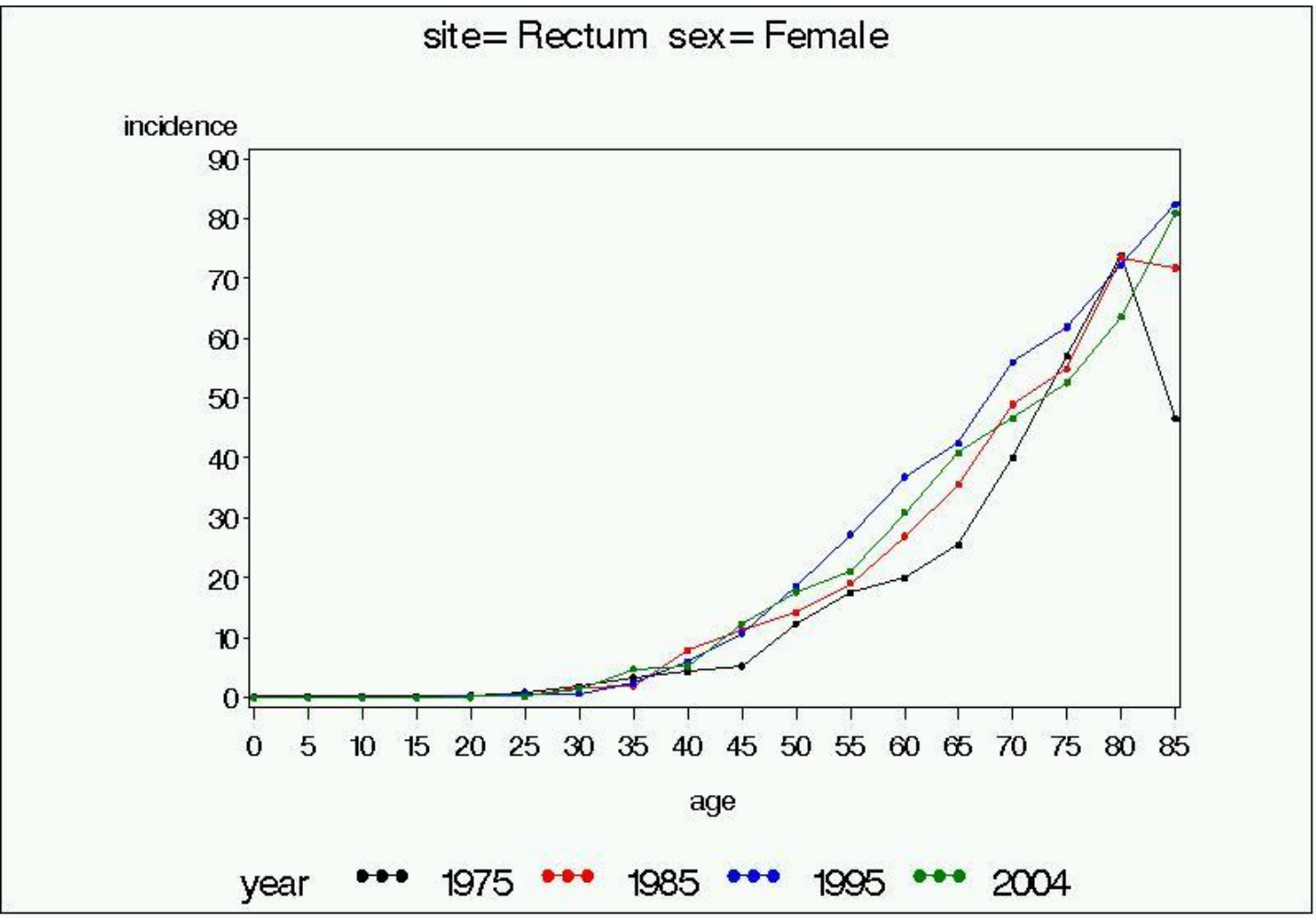


Figure. 6

APPENDIX FILES

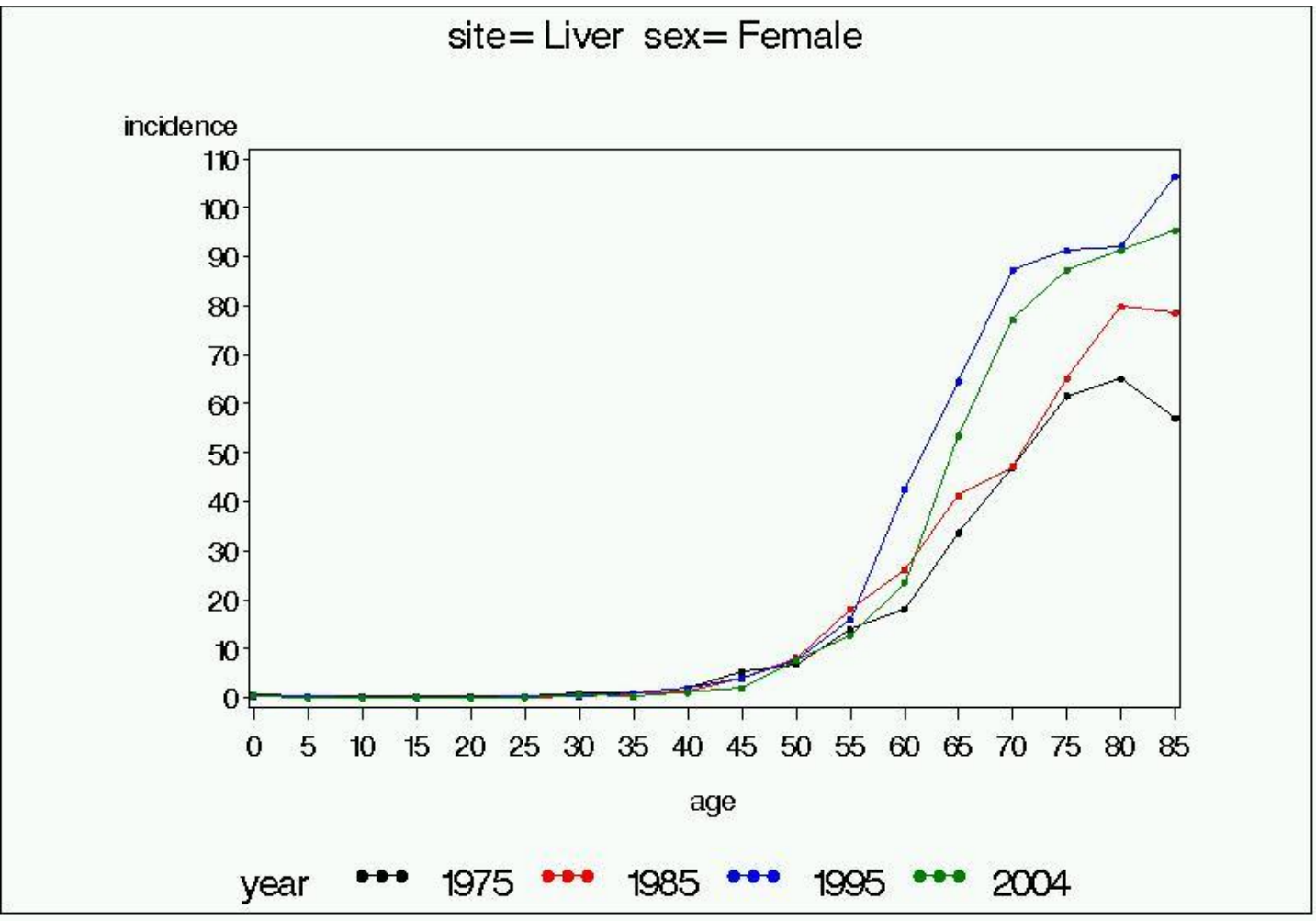


Figure. 7

APPENDIX FILES

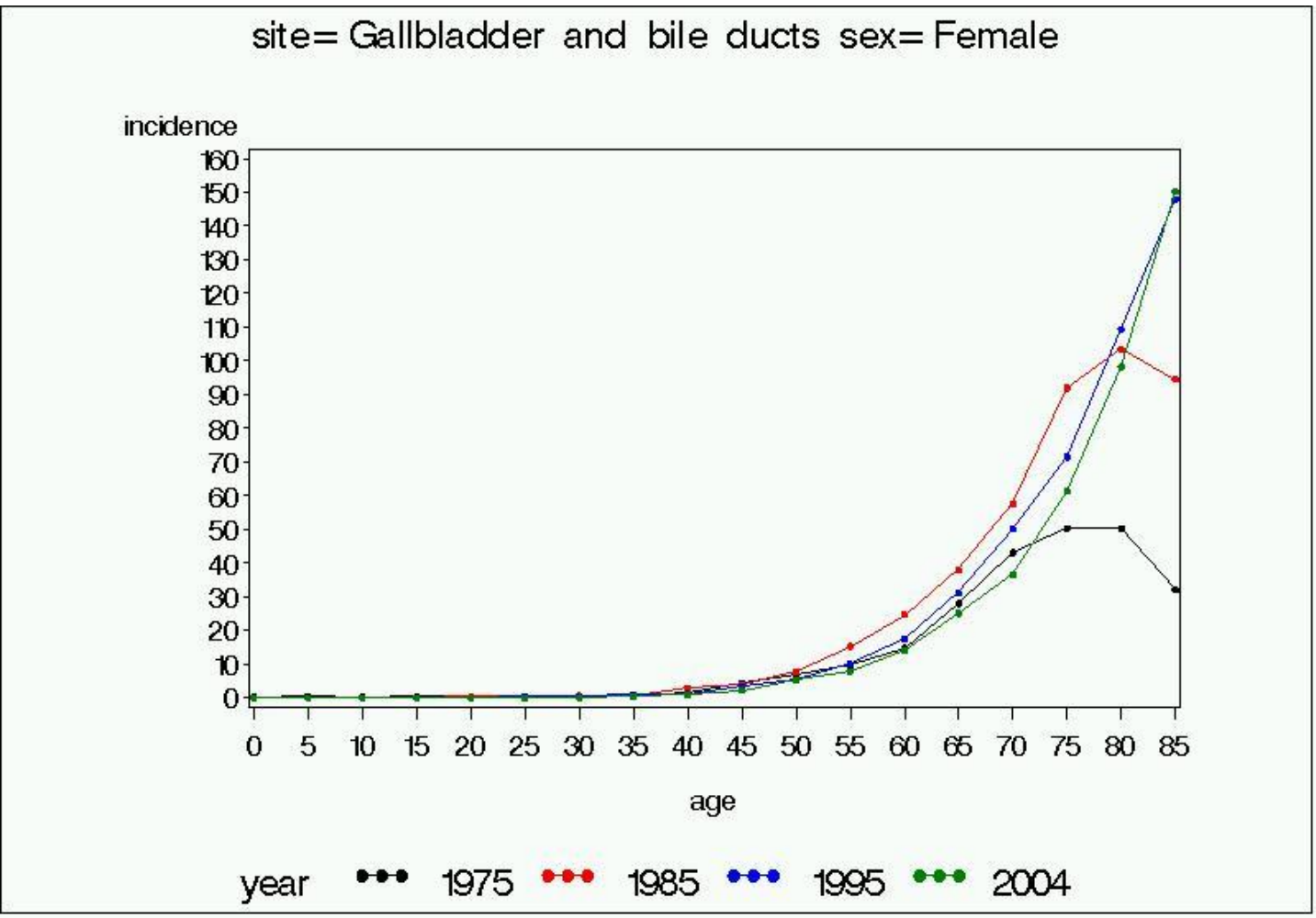


Figure. 8

APPENDIX FILES

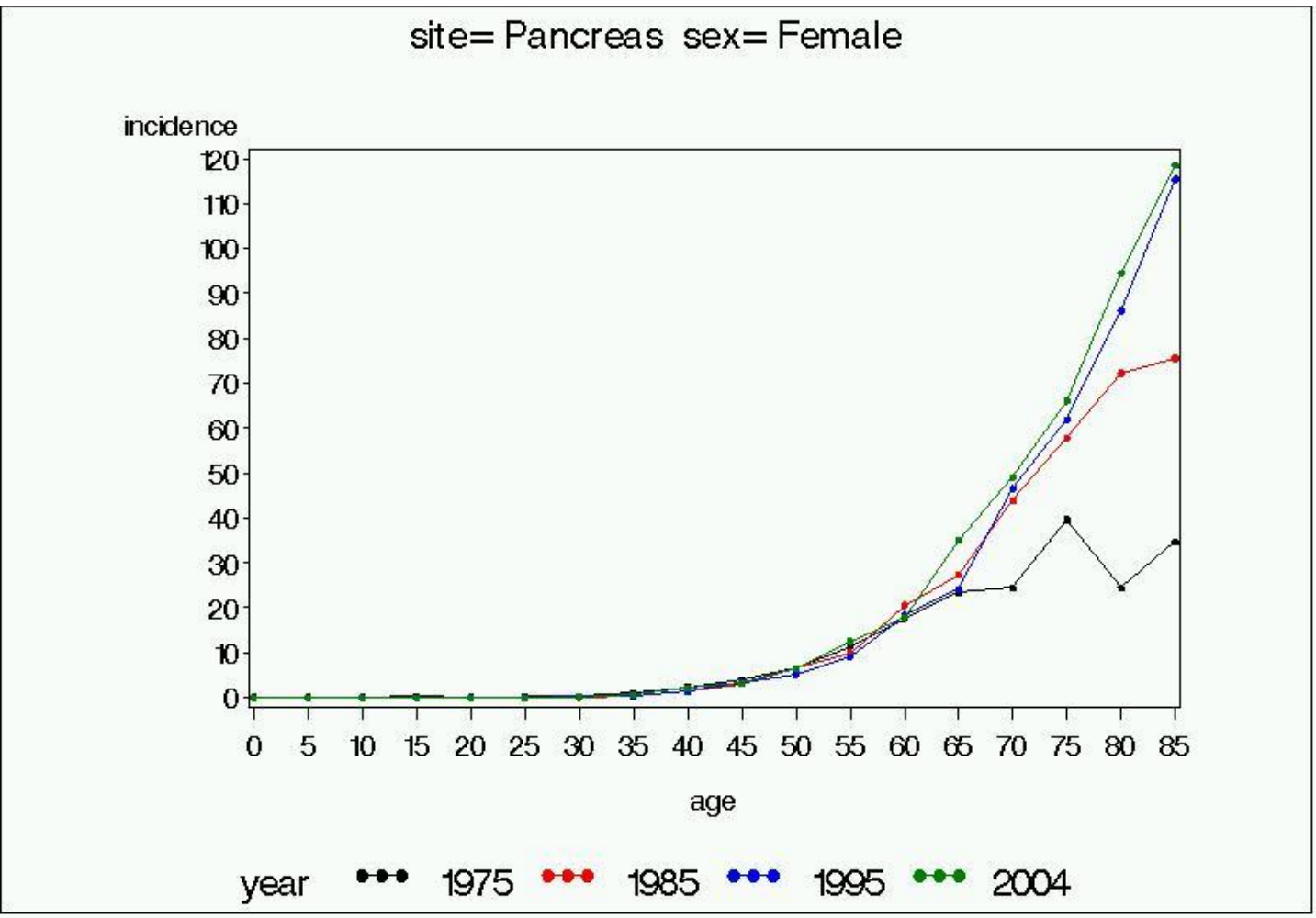


Figure. 9

APPENDIX FILES

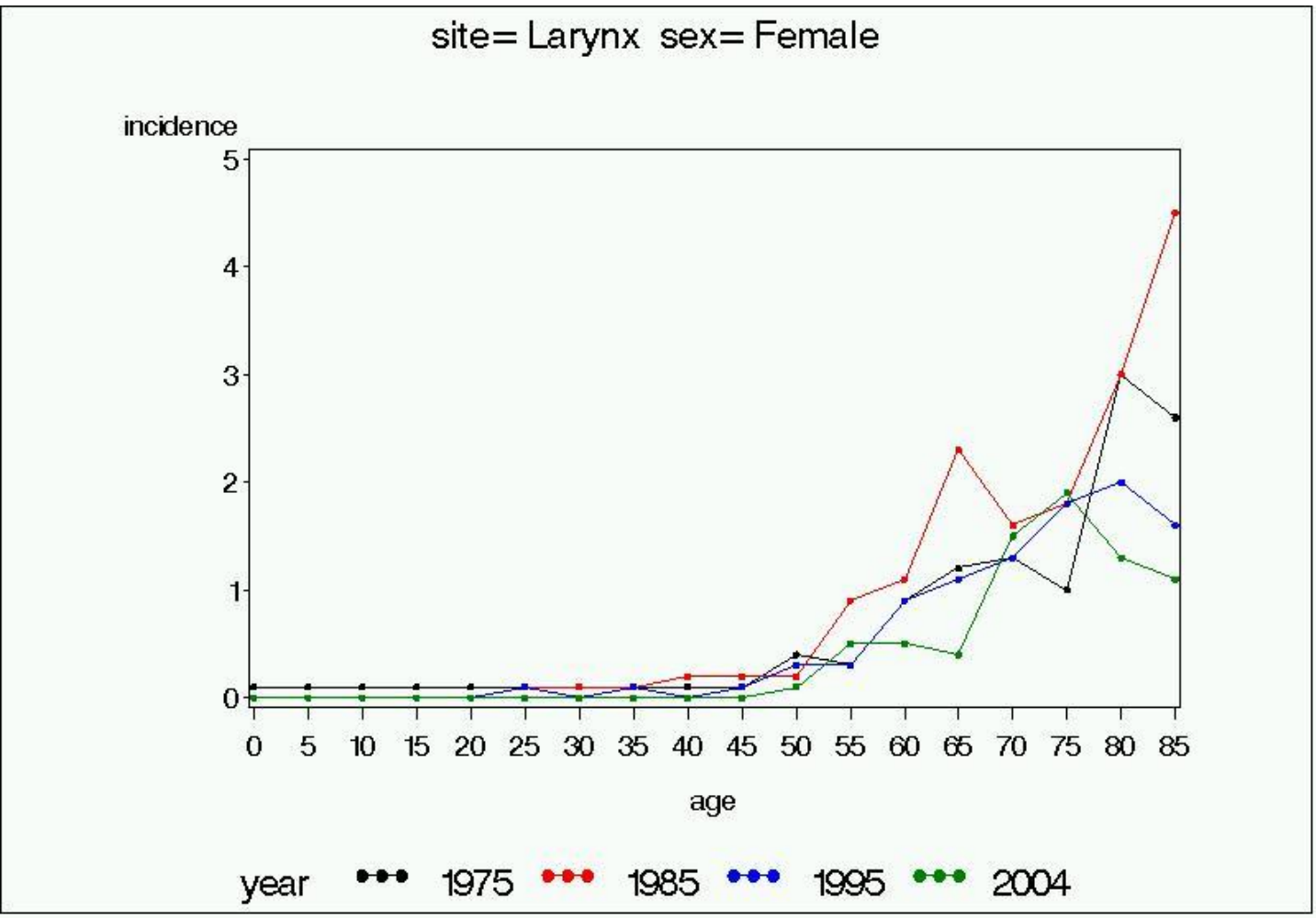


Figure. 10

APPENDIX FILES

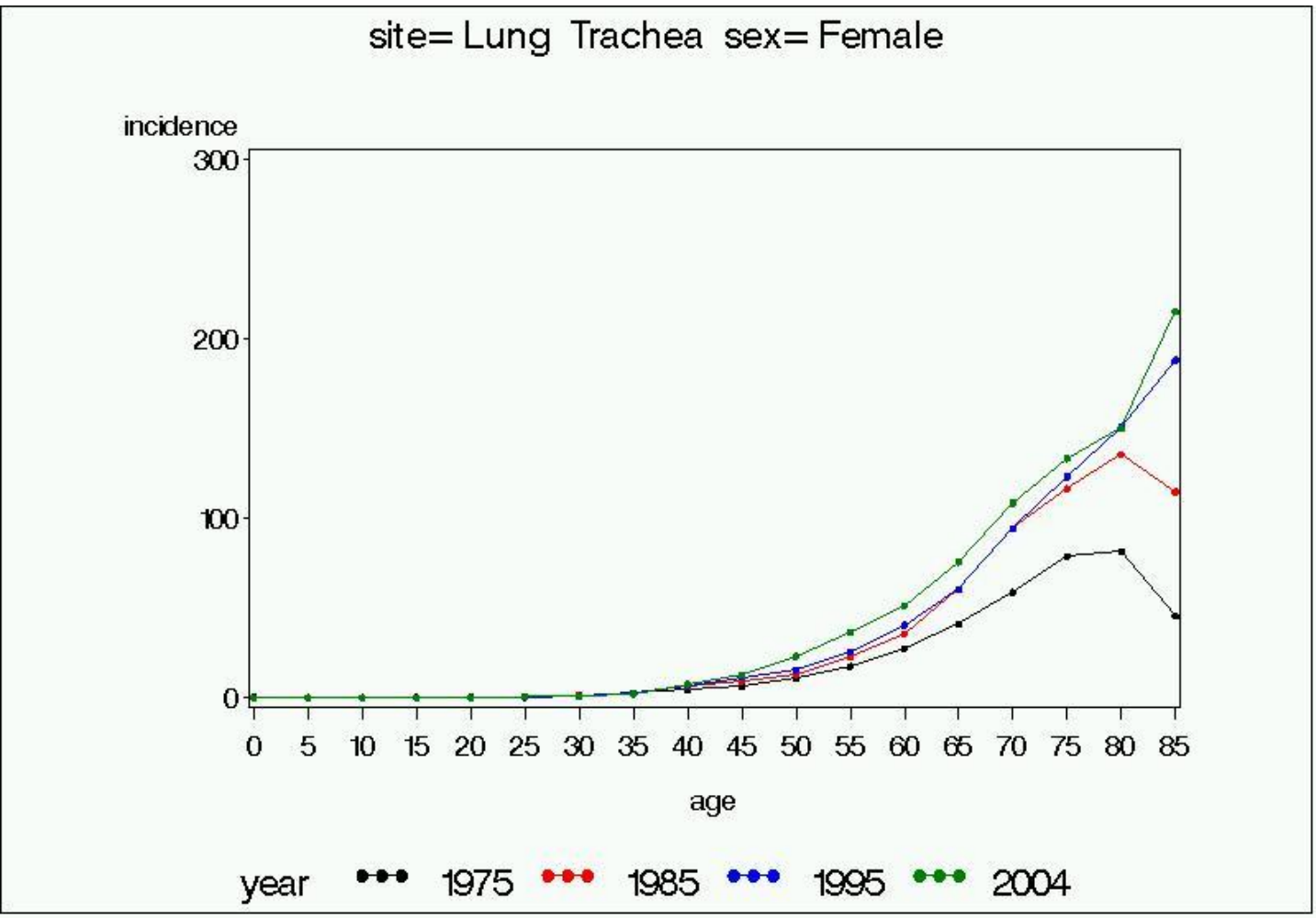


Figure. 11

APPENDIX FILES

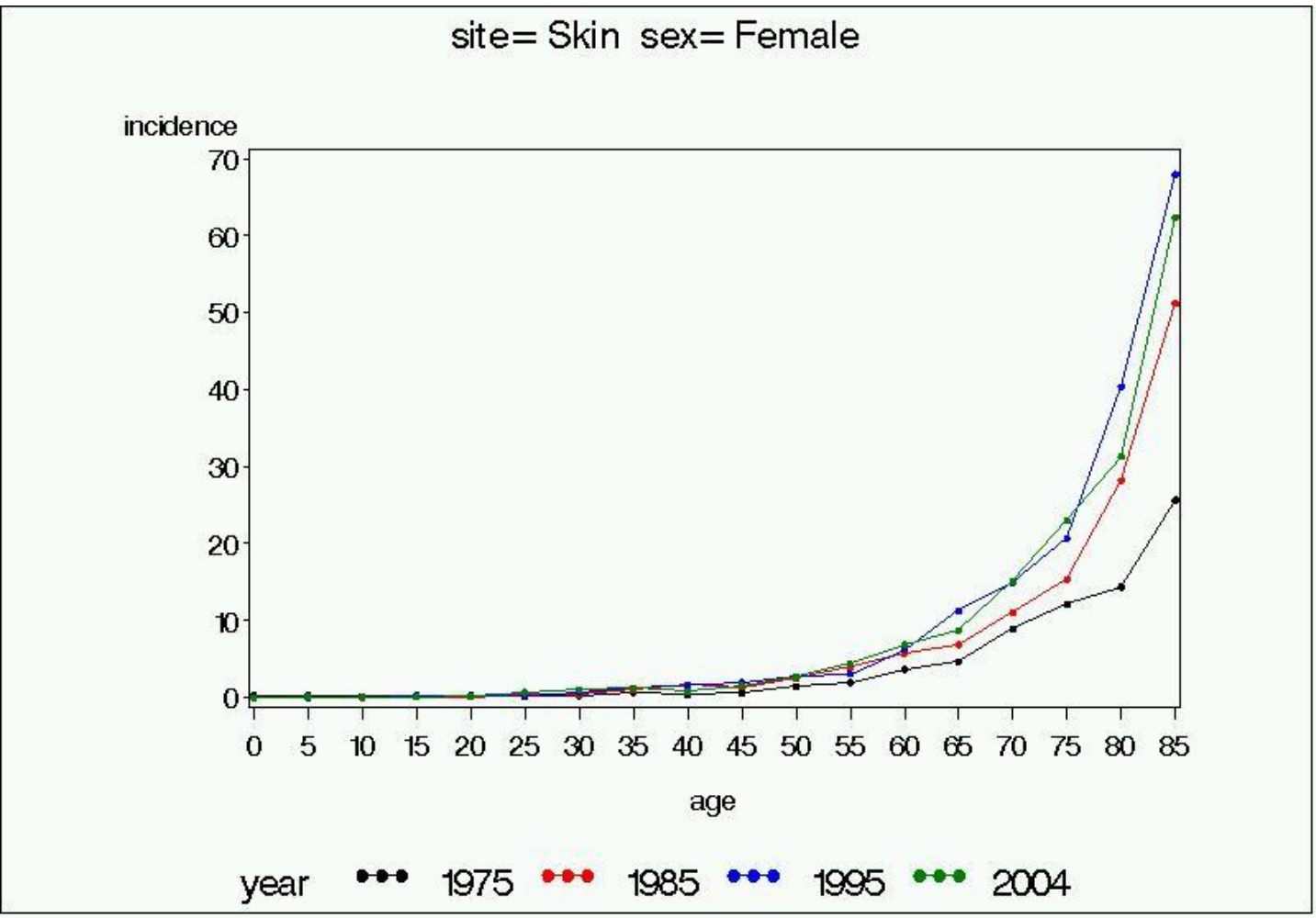


Figure. 12

APPENDIX FILES

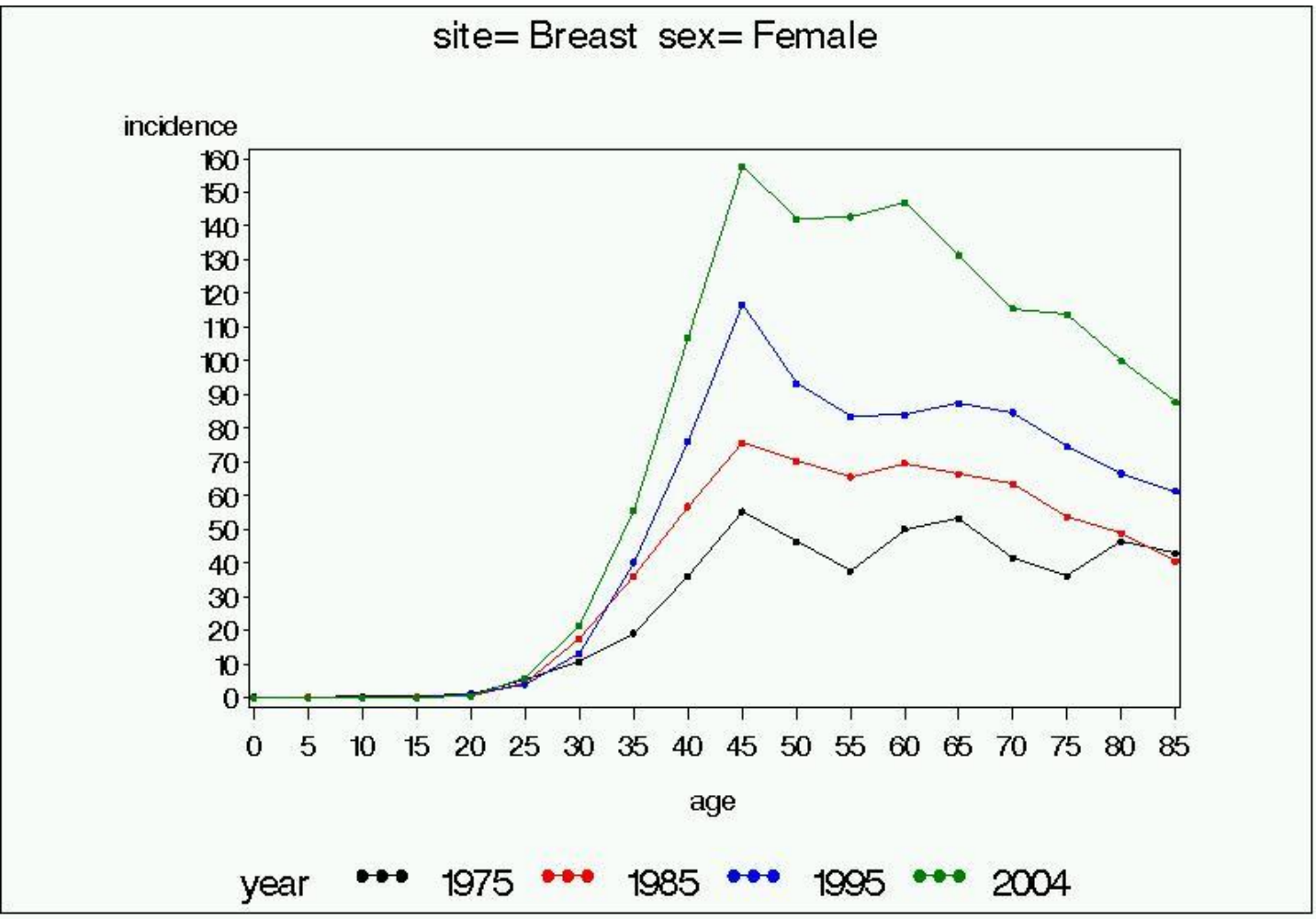


Figure. 13

APPENDIX FILES

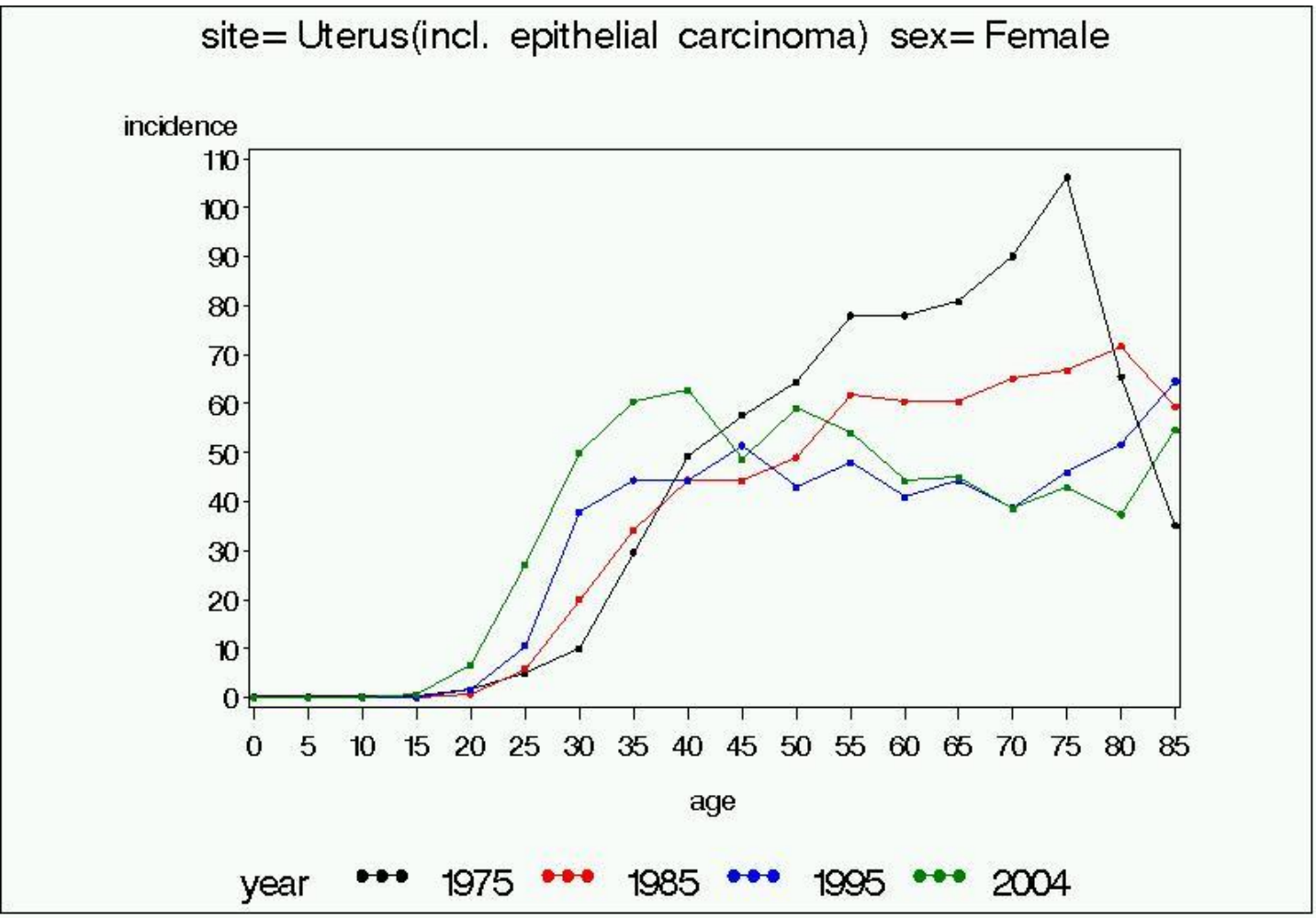


Figure. 14

APPENDIX FILES

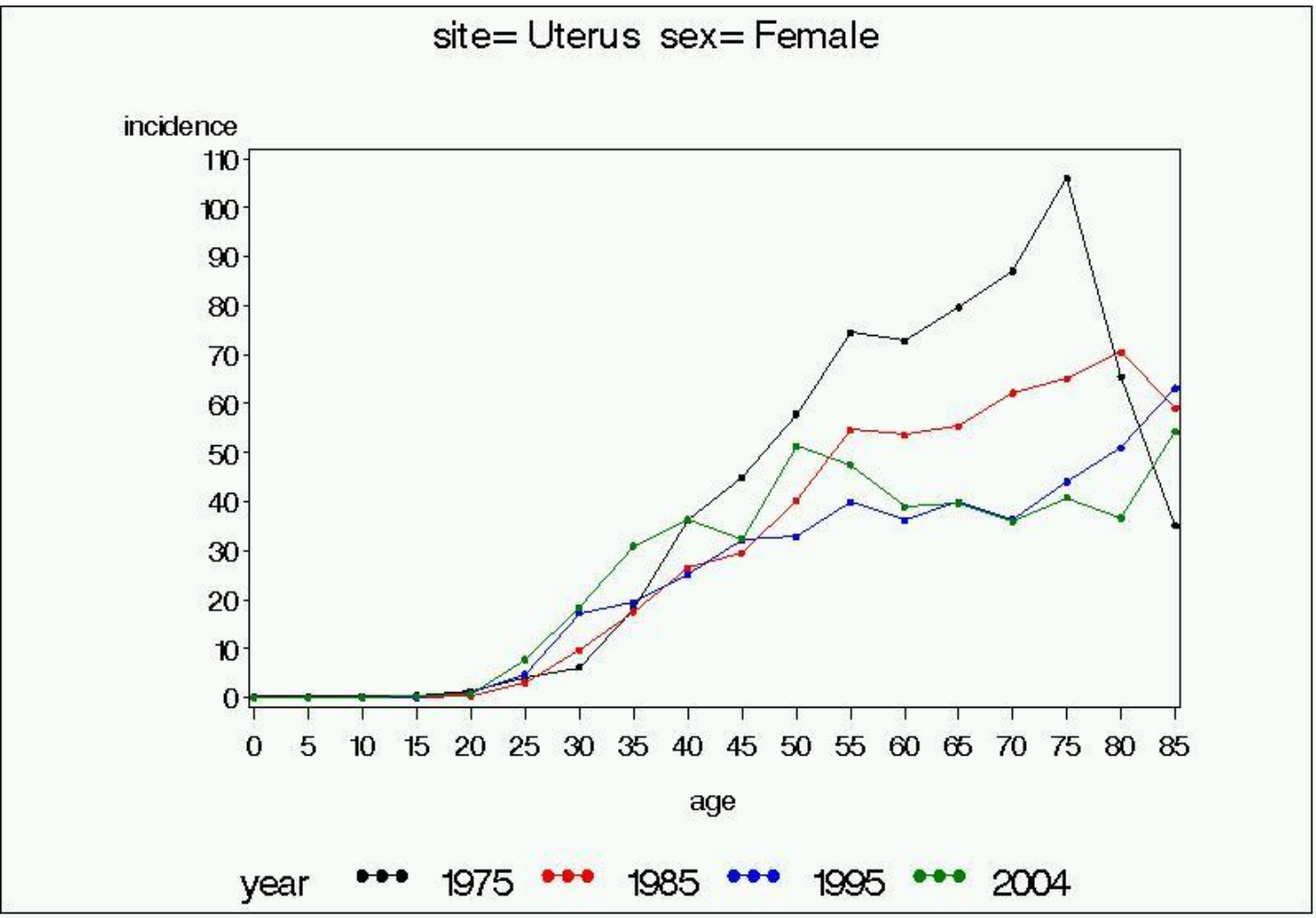


Figure. 15

APPENDIX FILES

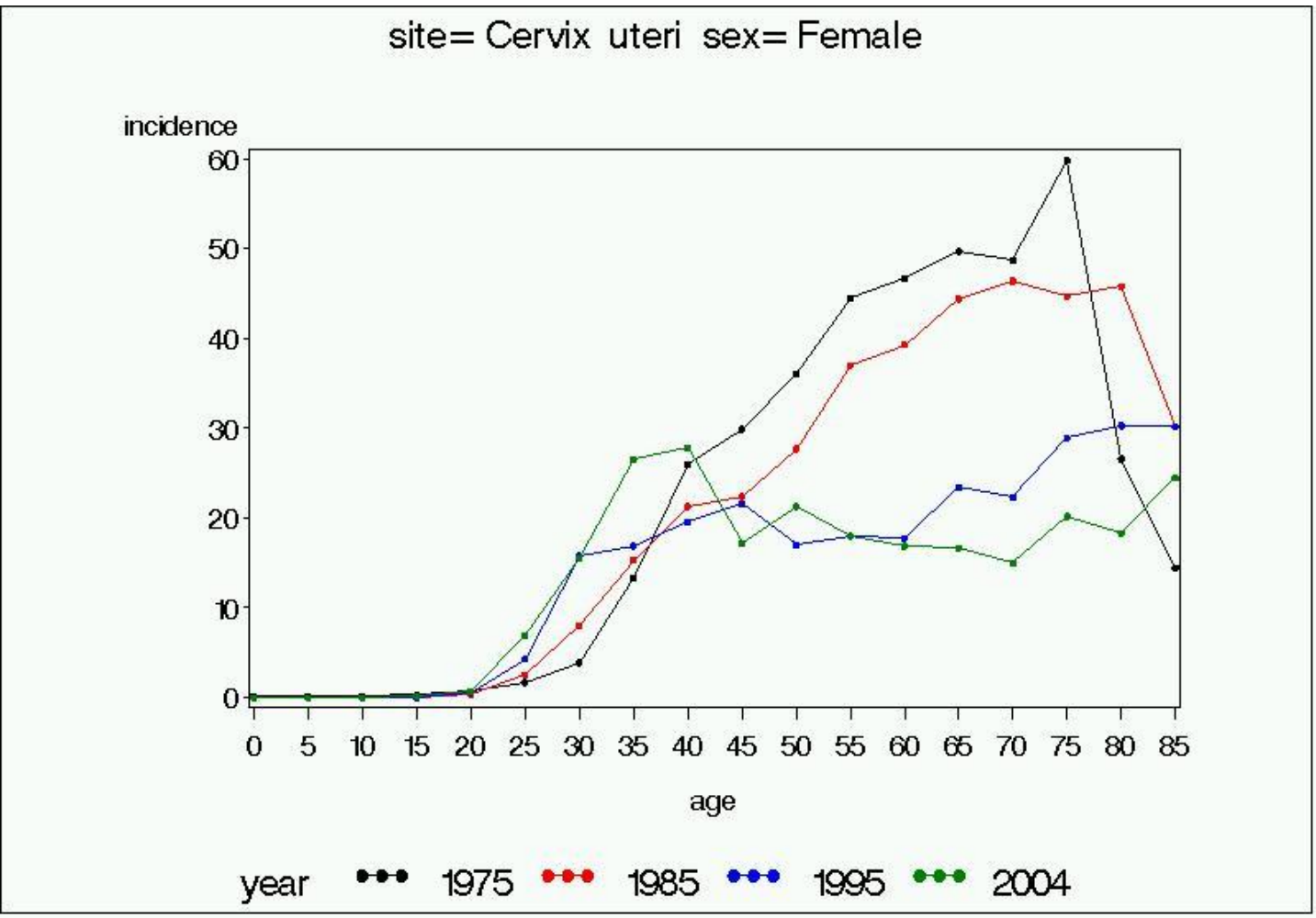


Figure. 16

APPENDIX FILES

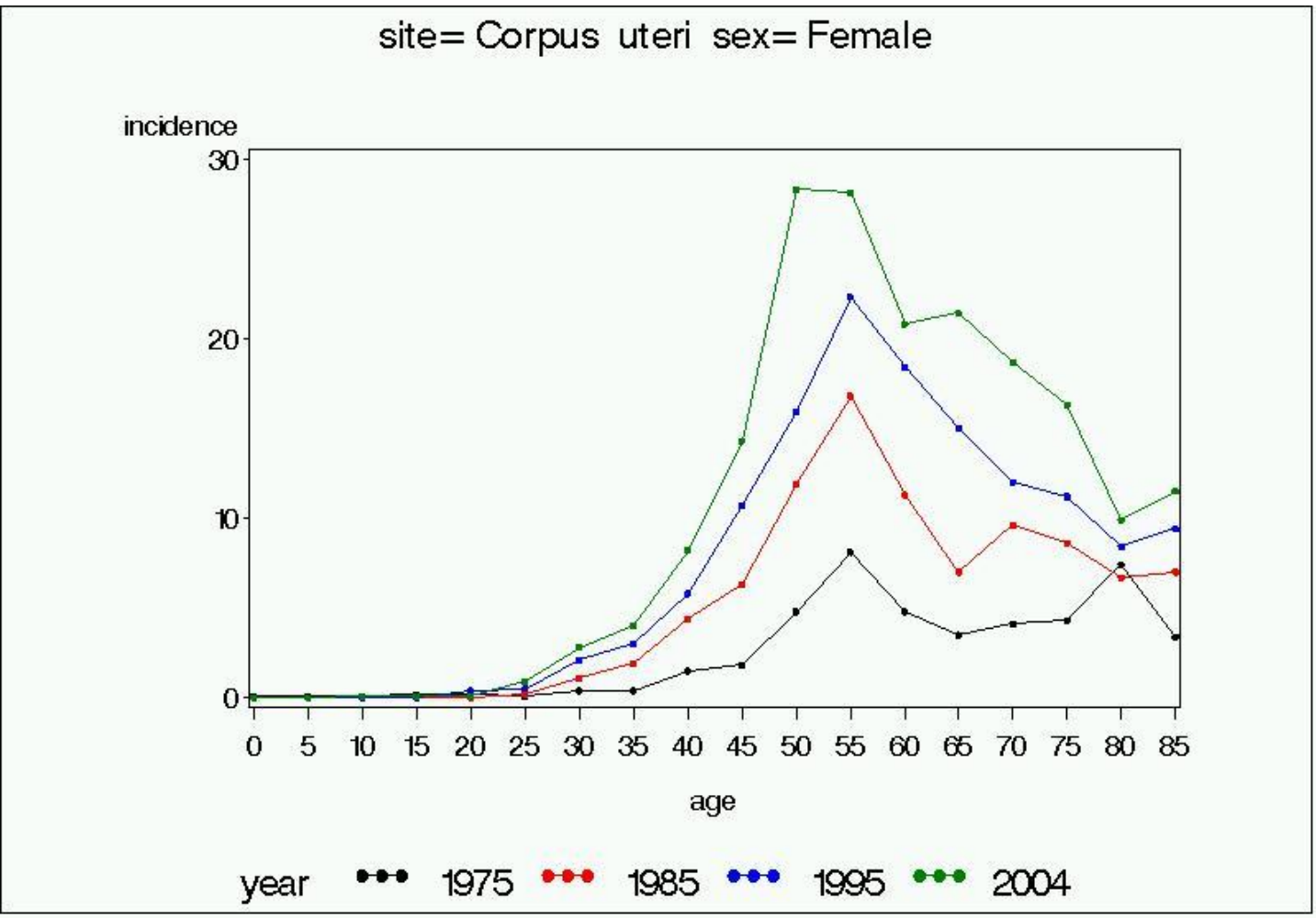


Figure. 17

APPENDIX FILES

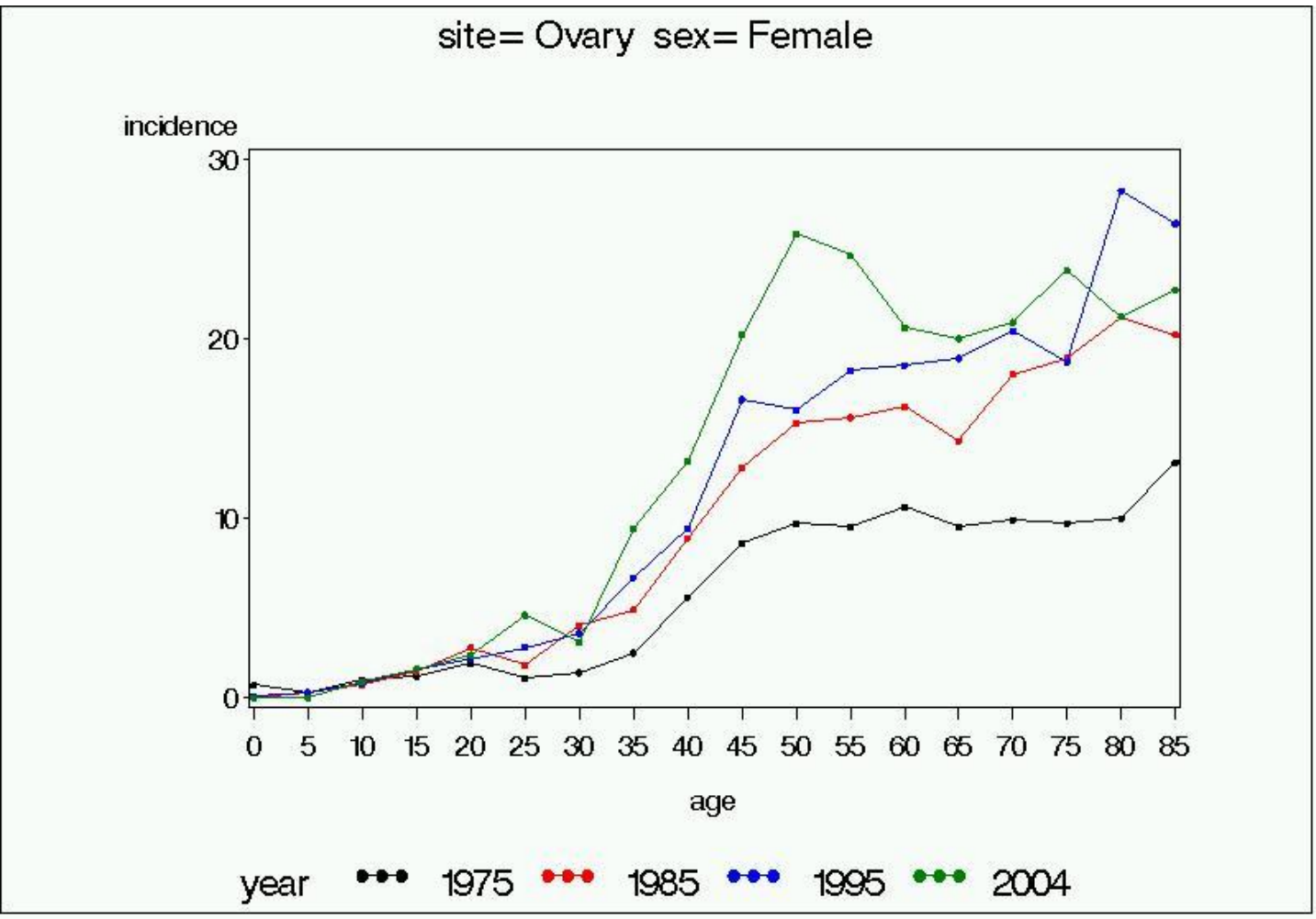


Figure. 18

APPENDIX FILES

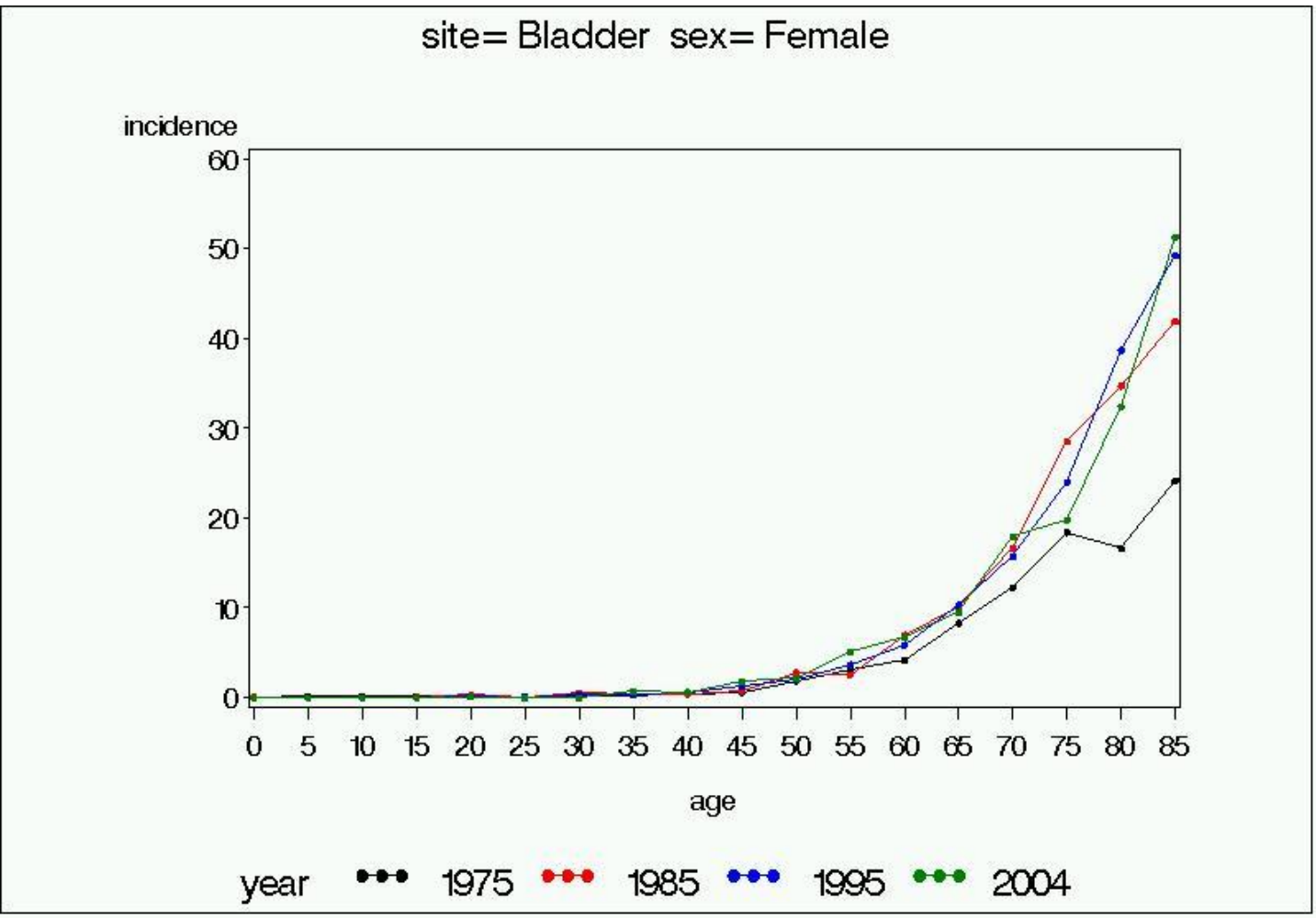


Figure. 19

APPENDIX FILES

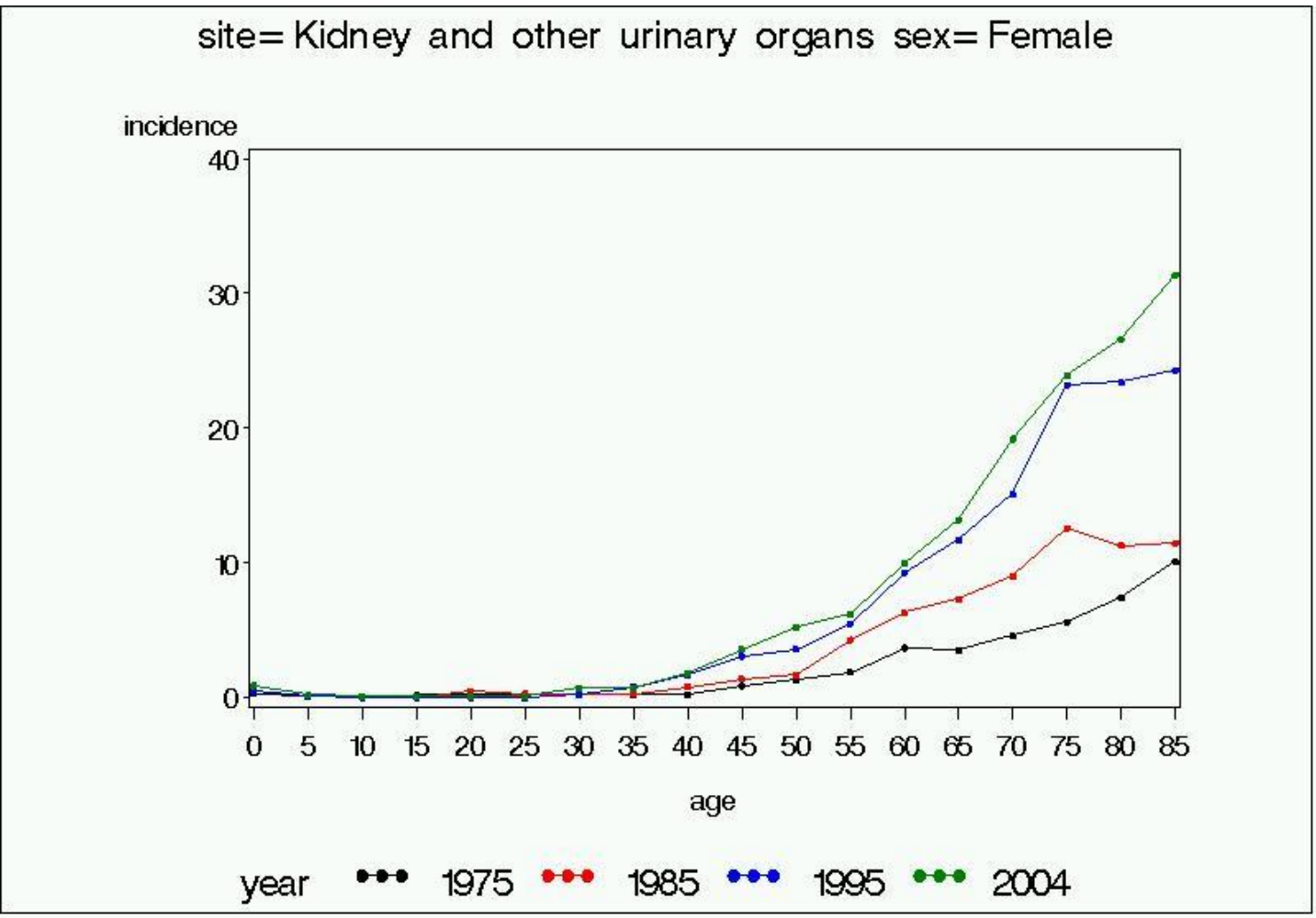


Figure. 20

APPENDIX FILES

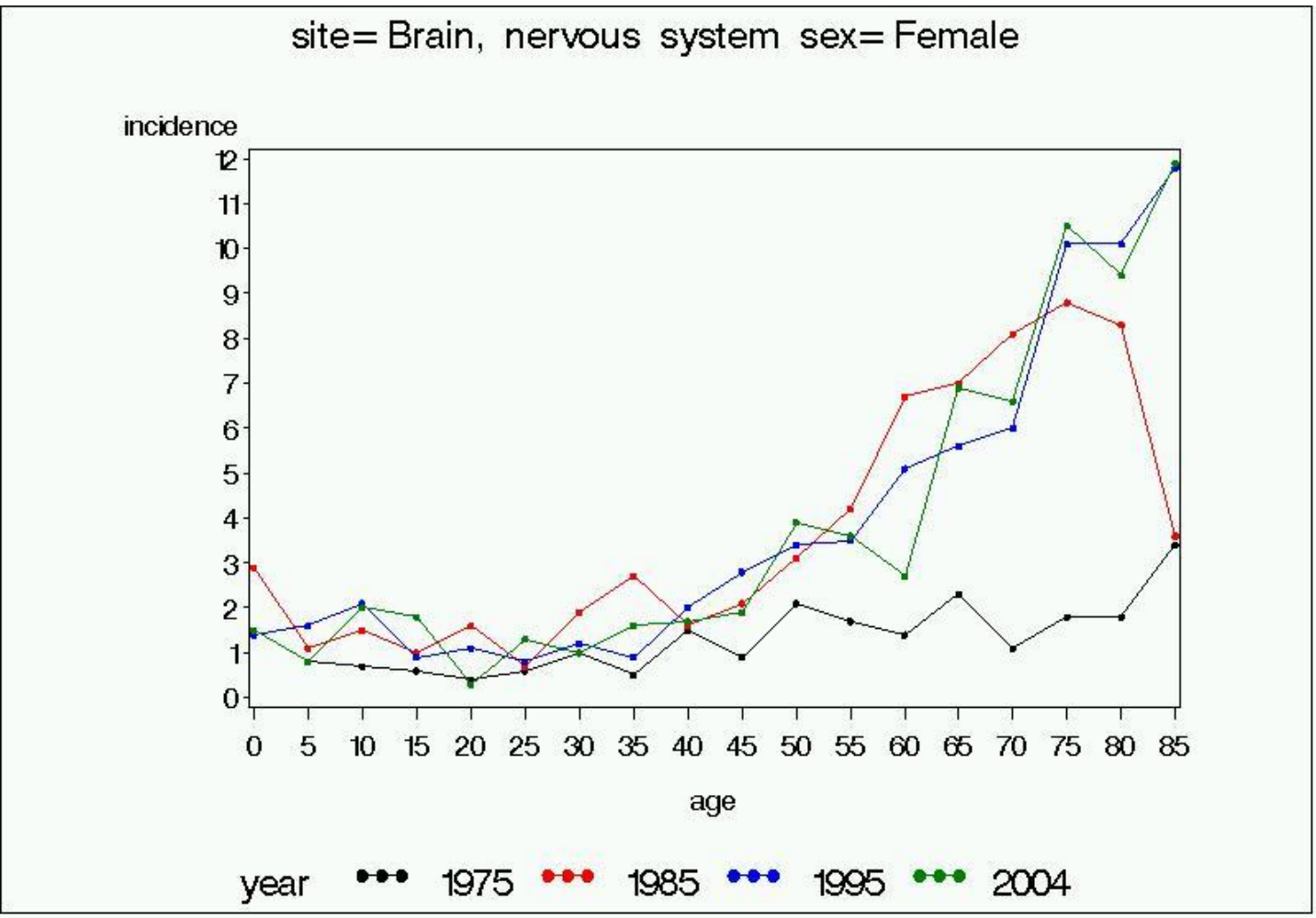


Figure. 21

APPENDIX FILES

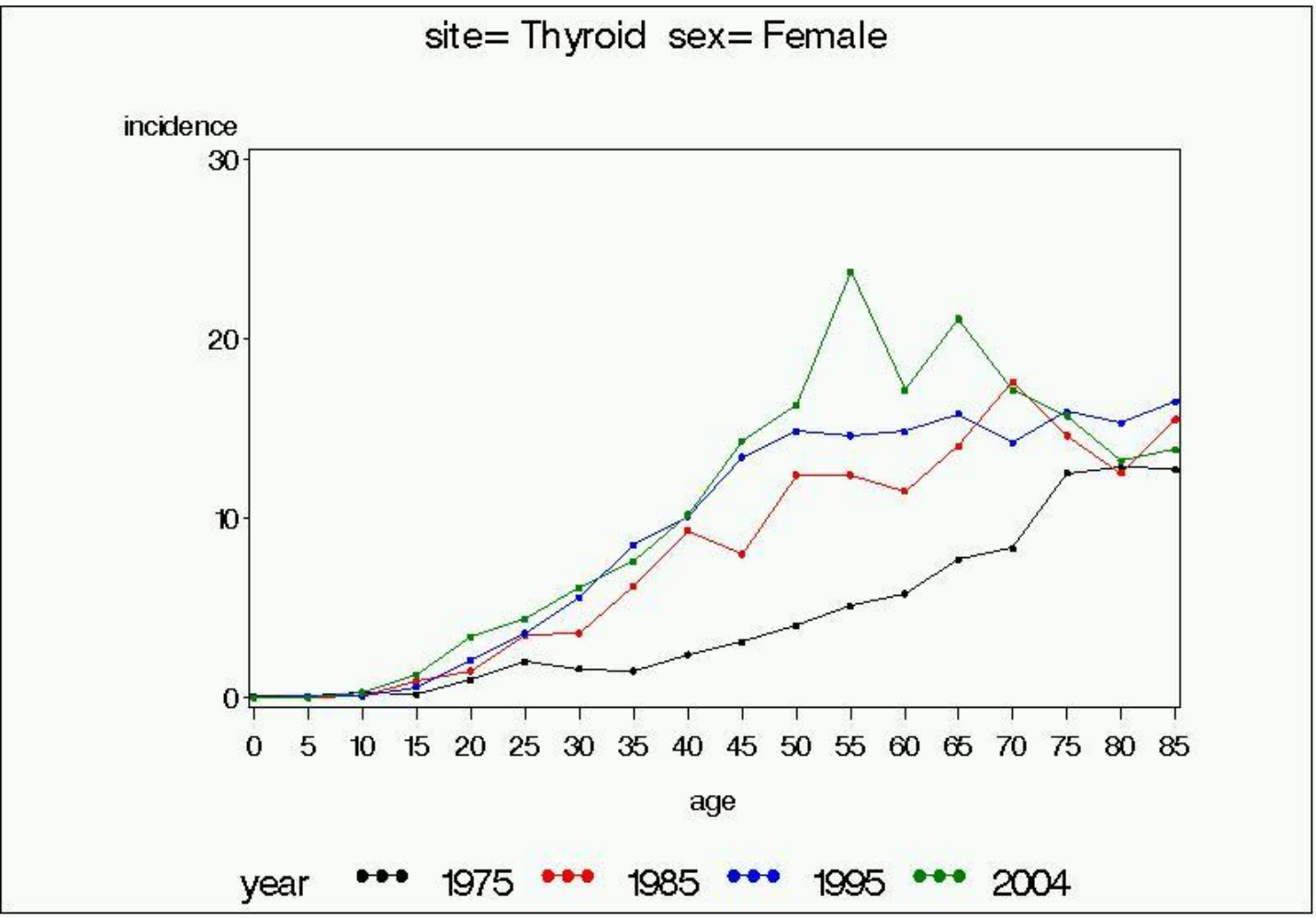


Figure. 22

APPENDIX FILES

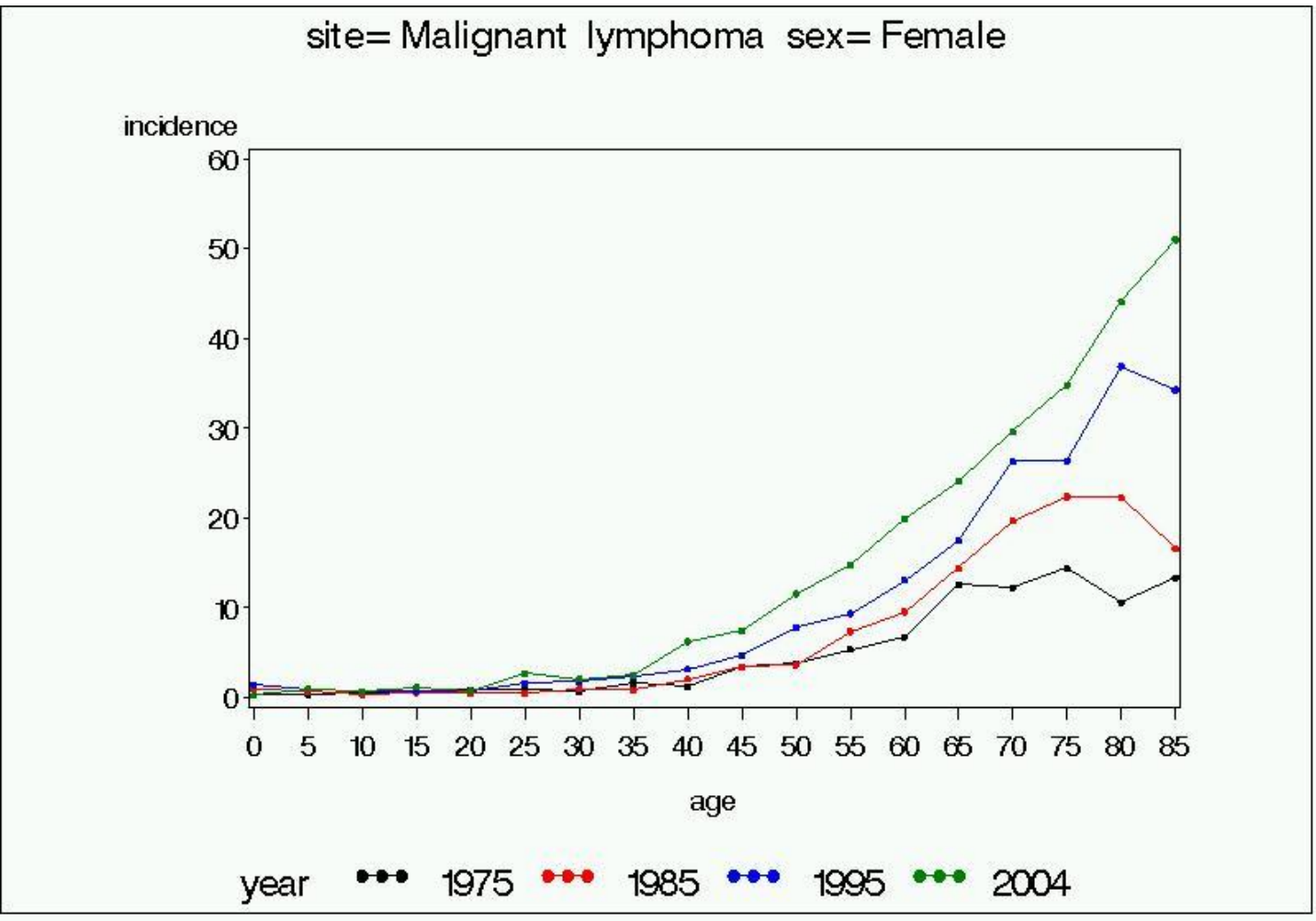


Figure. 23

APPENDIX FILES

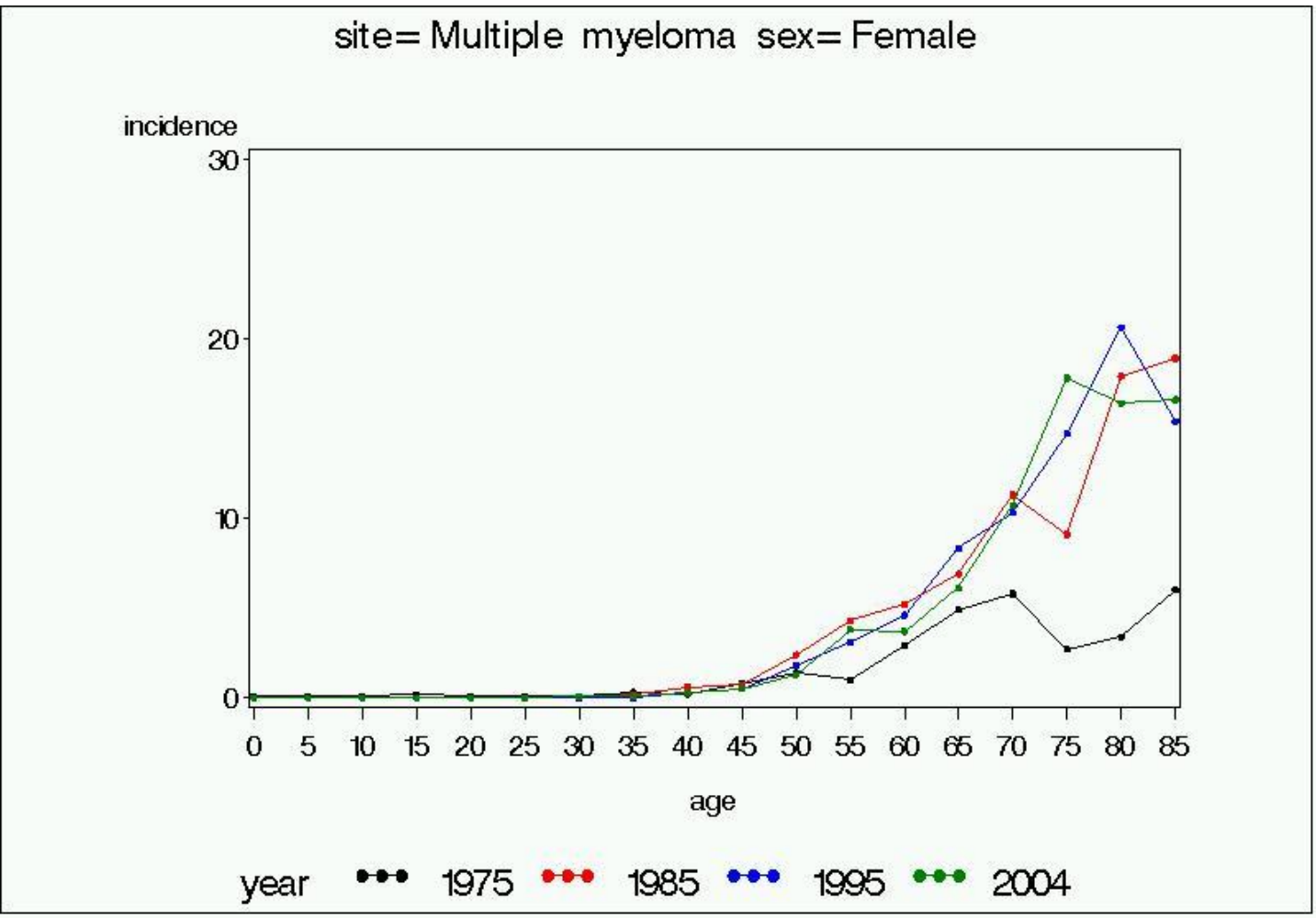


Figure. 24

APPENDIX FILES

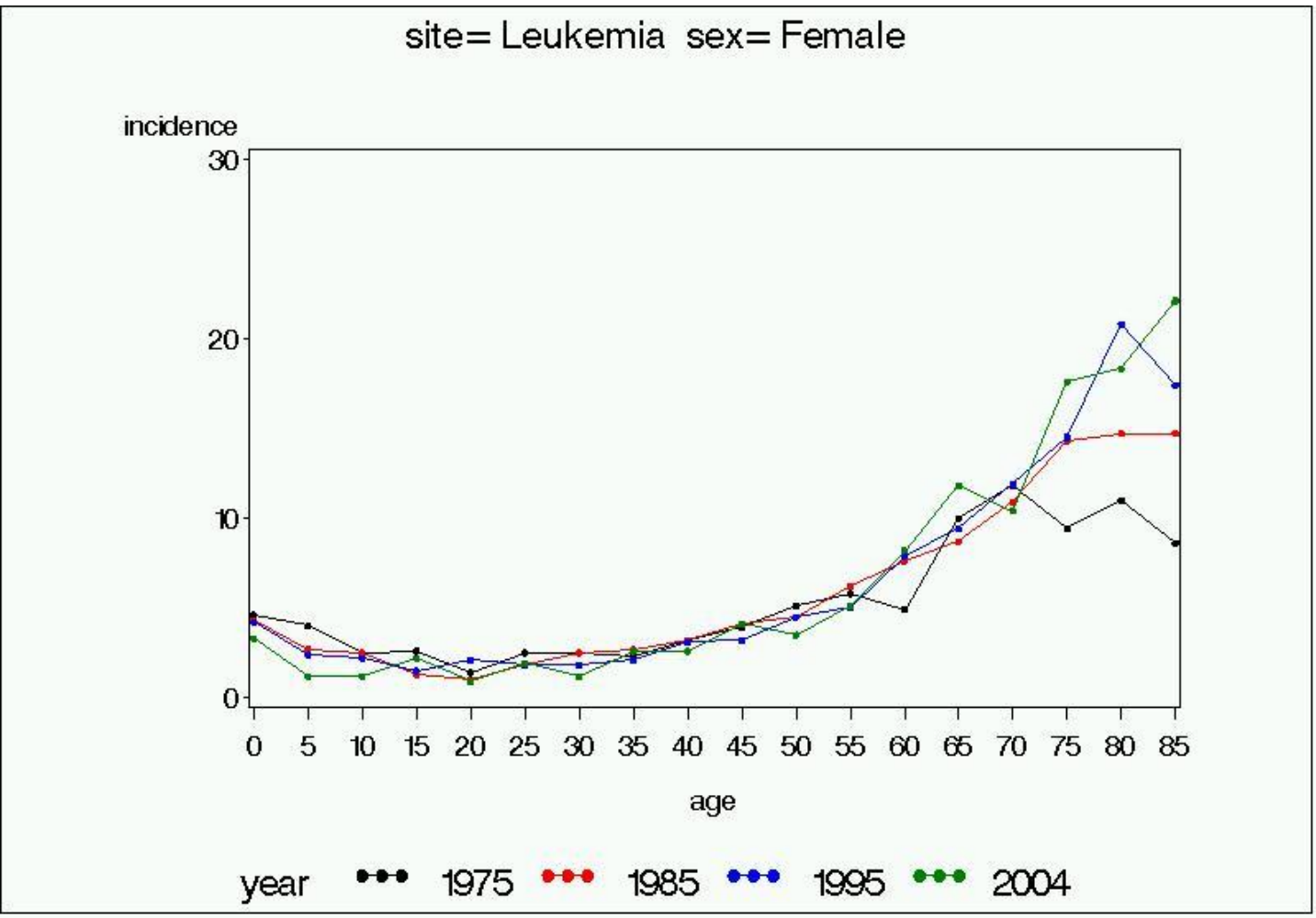


Figure. 25

APPENDIX FILES

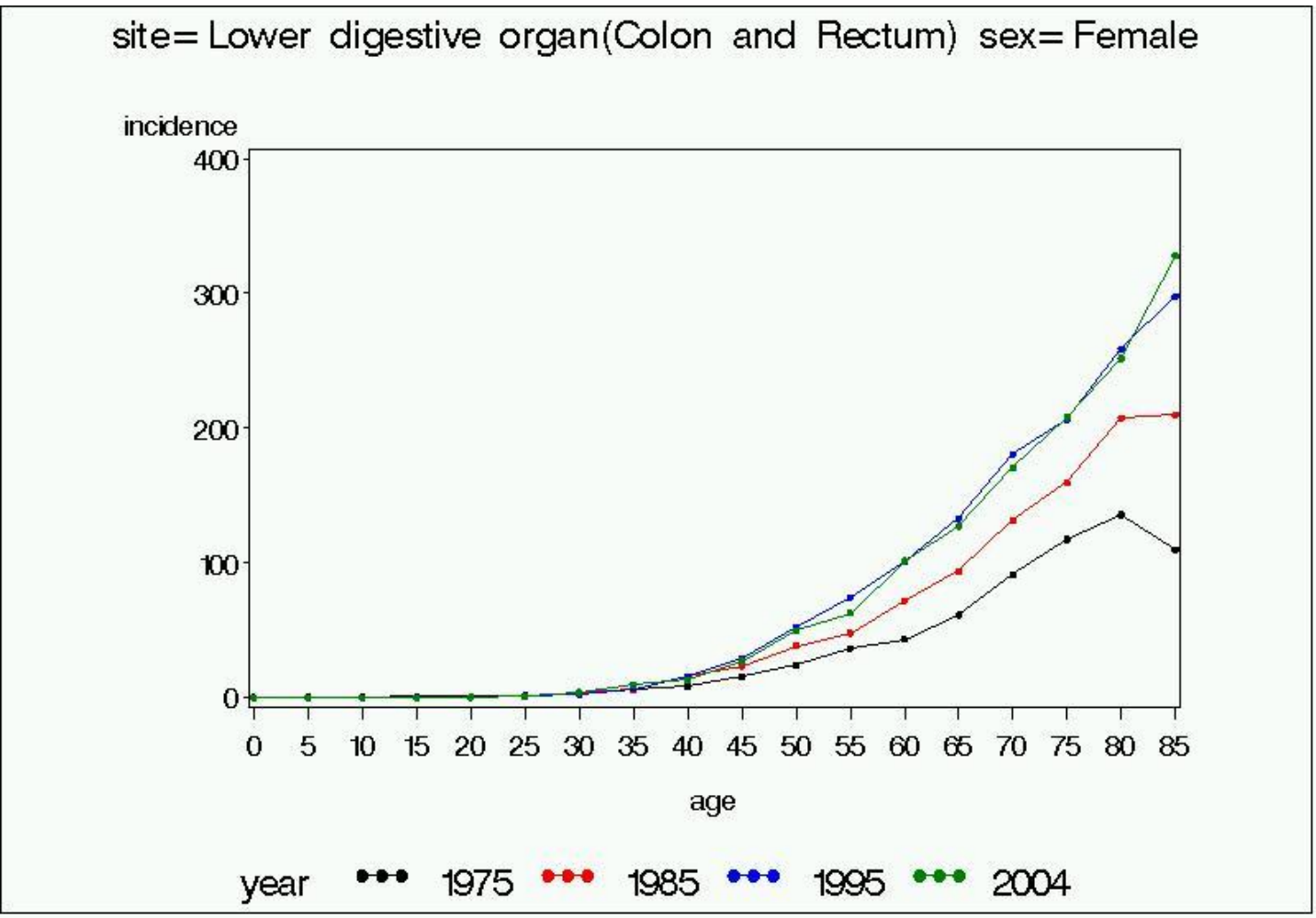


Figure. 26

APPENDIX FILES

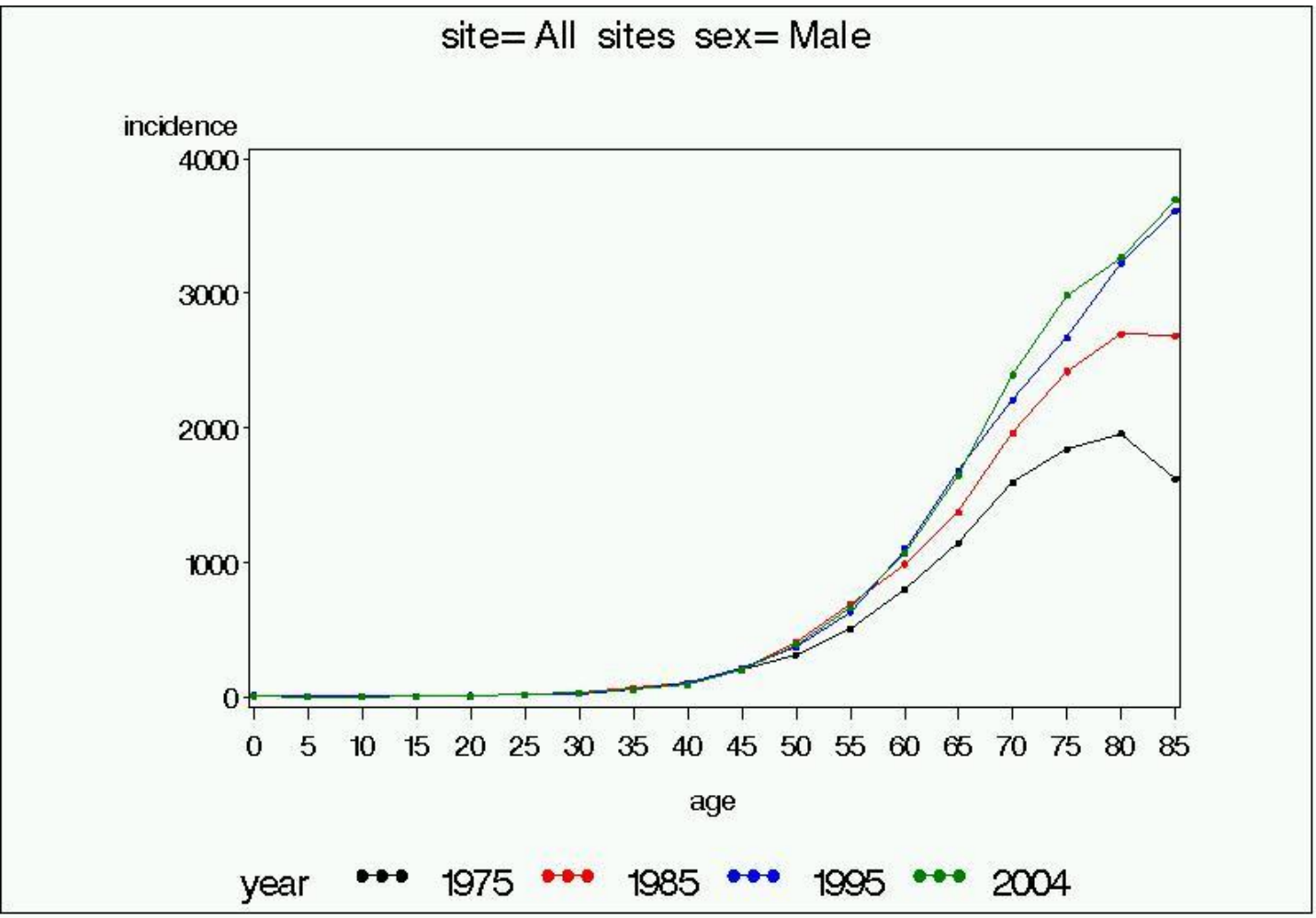


Figure. 27

APPENDIX FILES

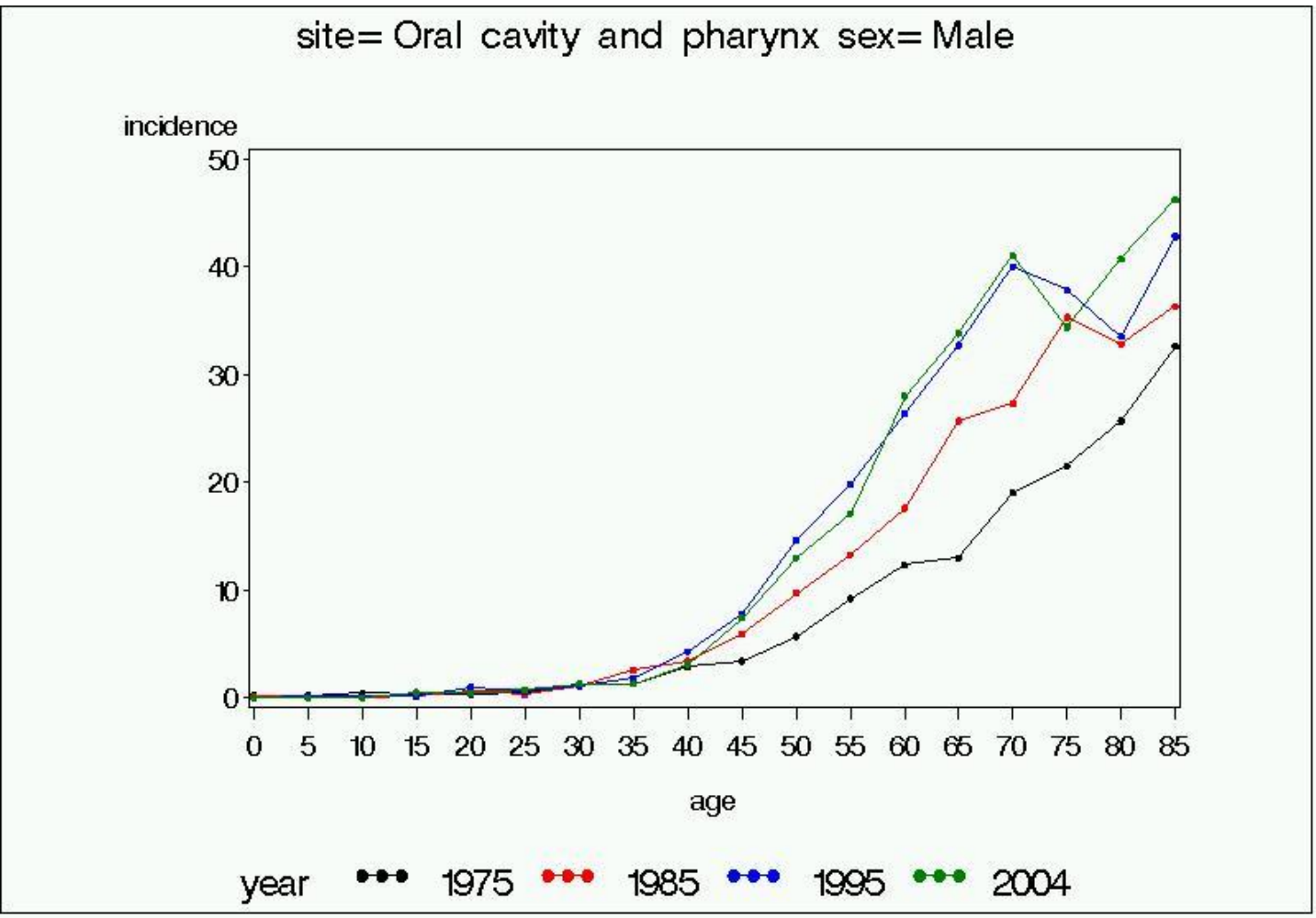


Figure. 28

APPENDIX FILES

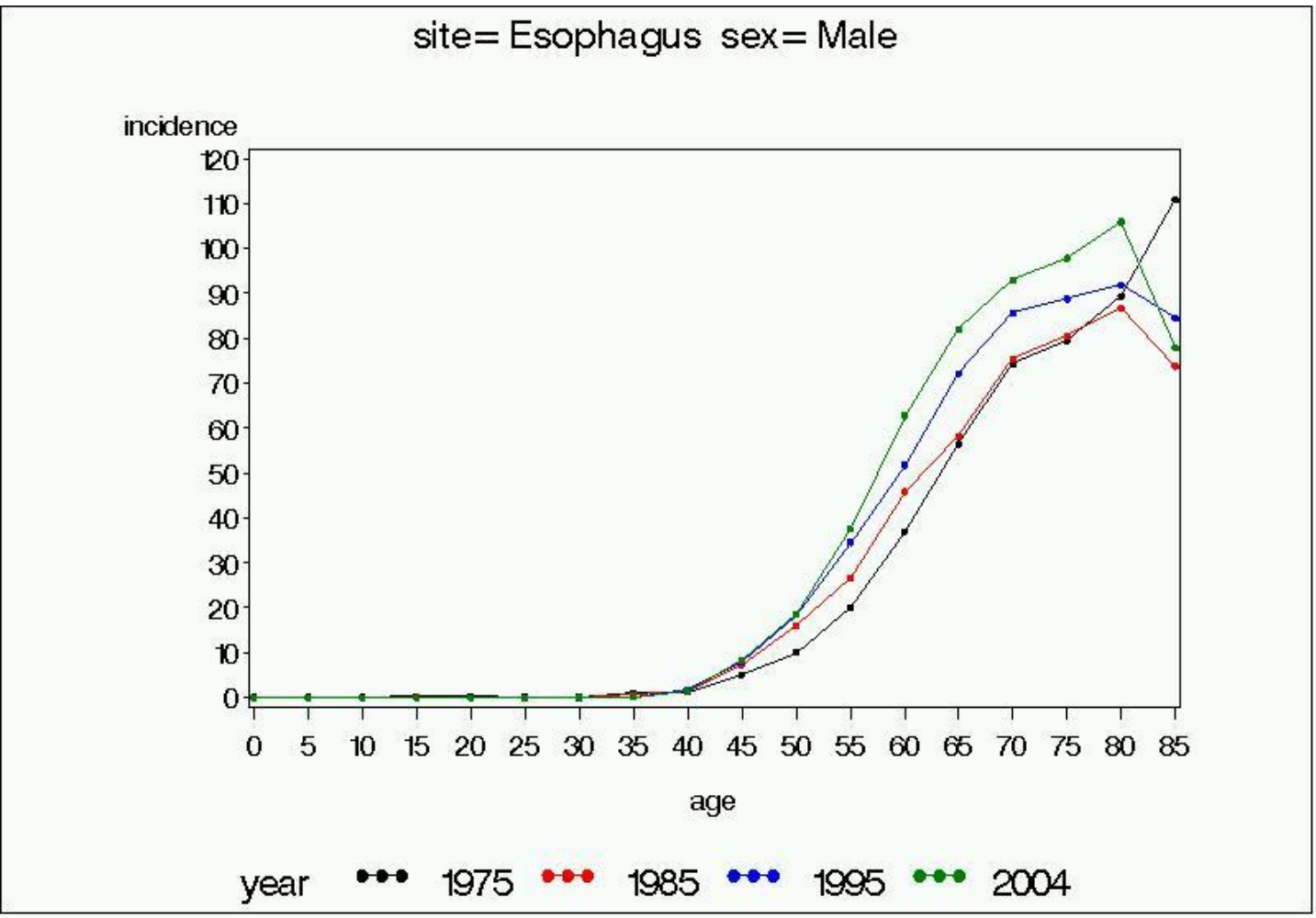


Figure. 29

APPENDIX FILES

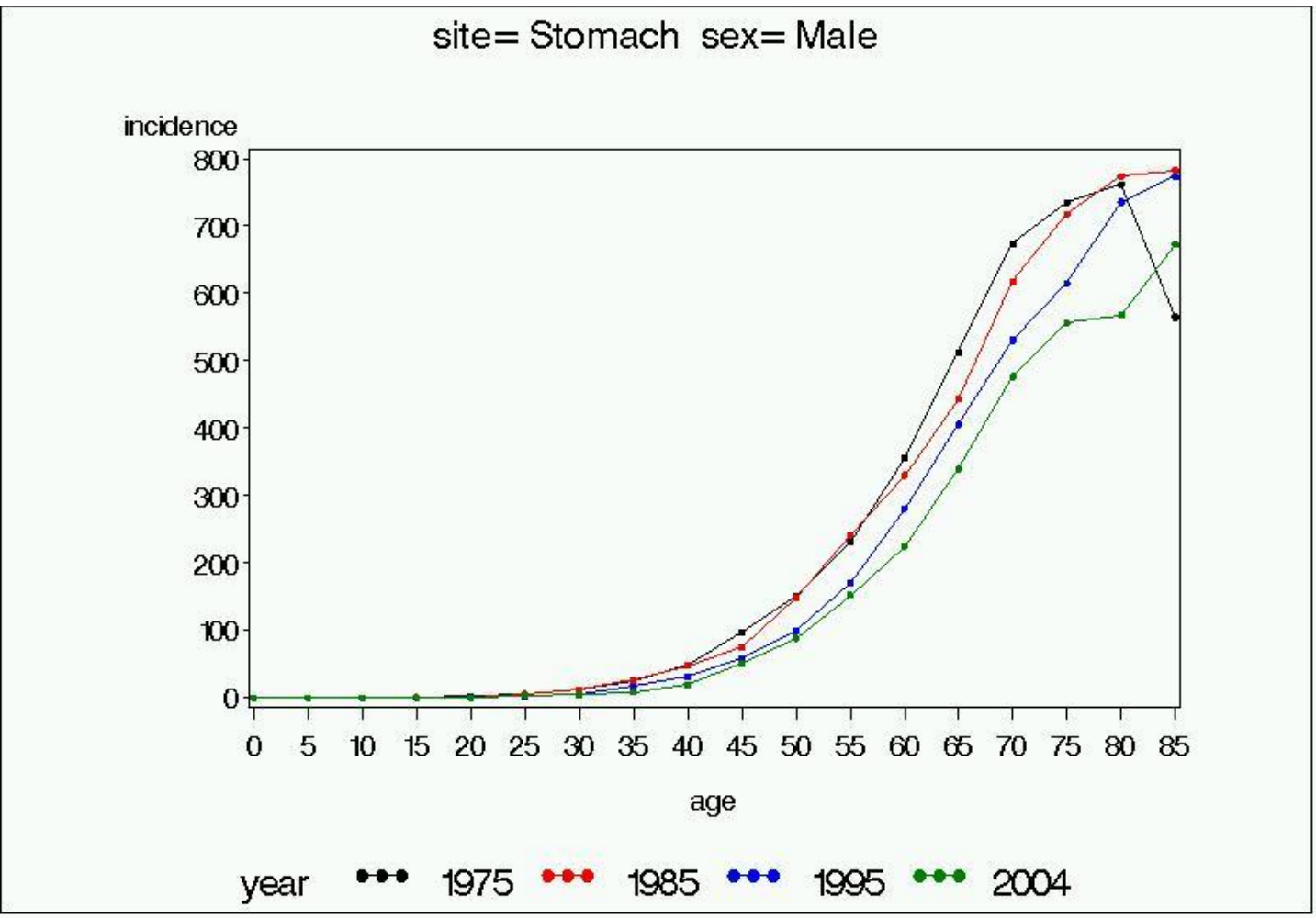


Figure. 30

APPENDIX FILES

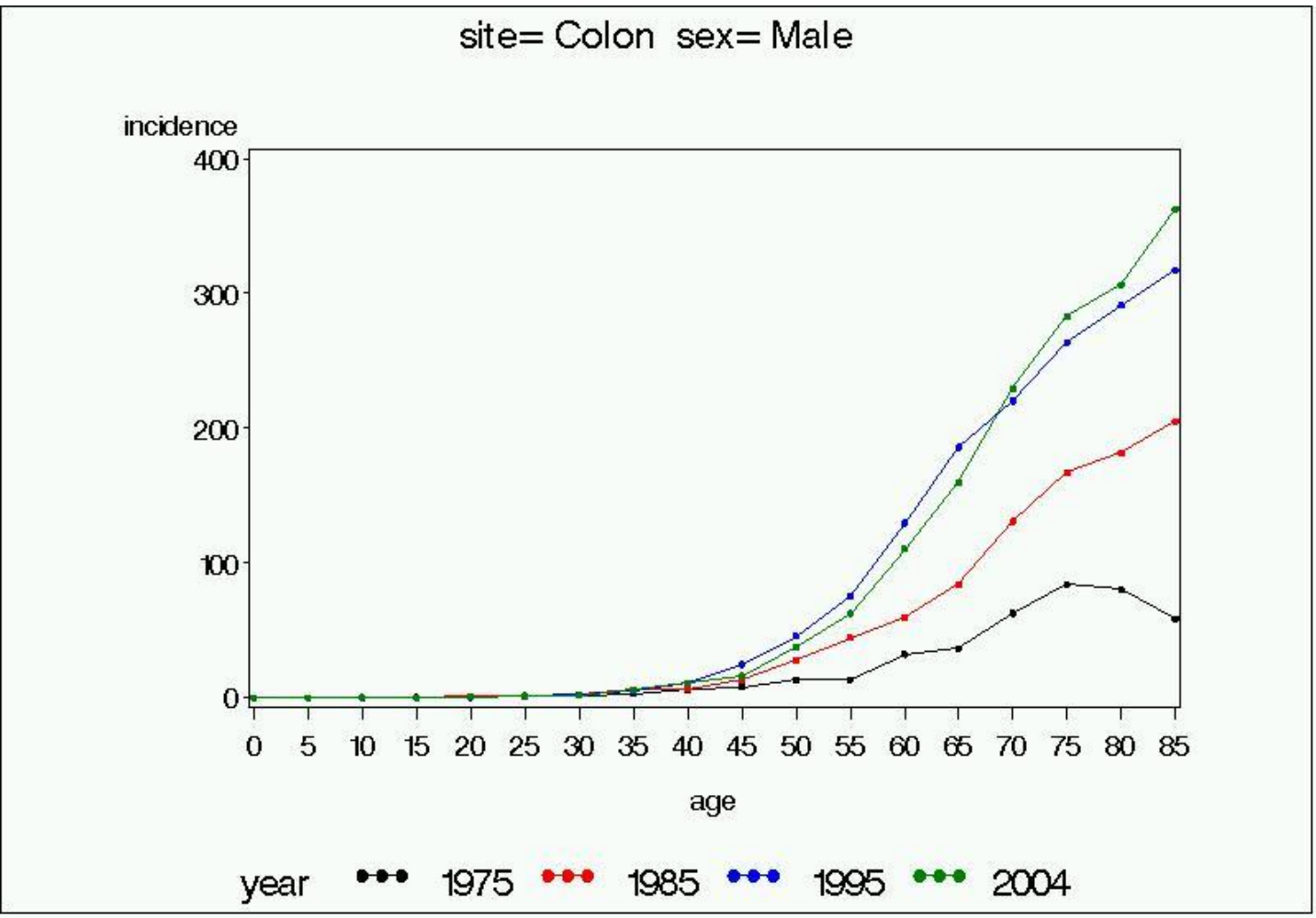


Figure. 31

APPENDIX FILES

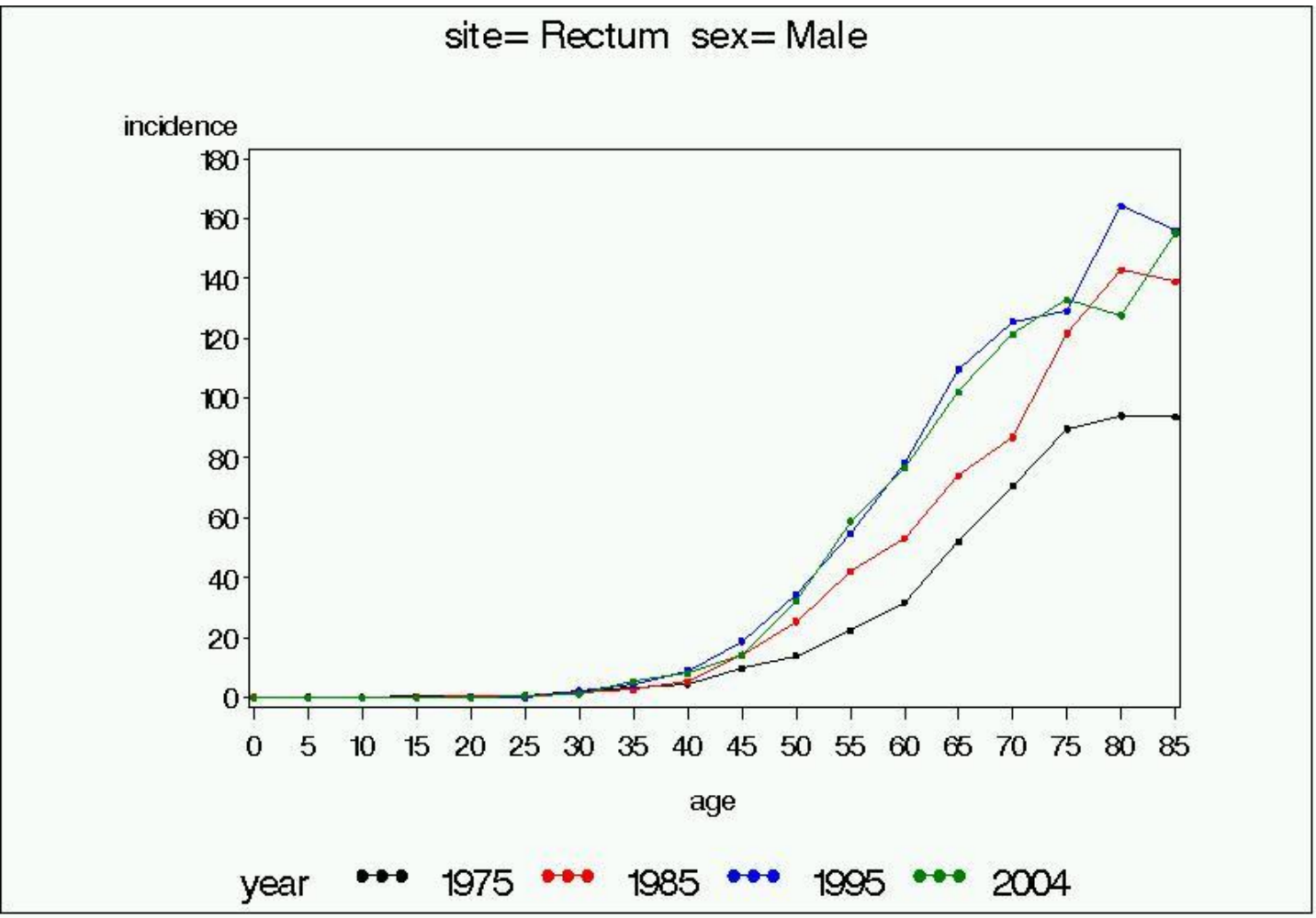


Figure. 32

APPENDIX FILES

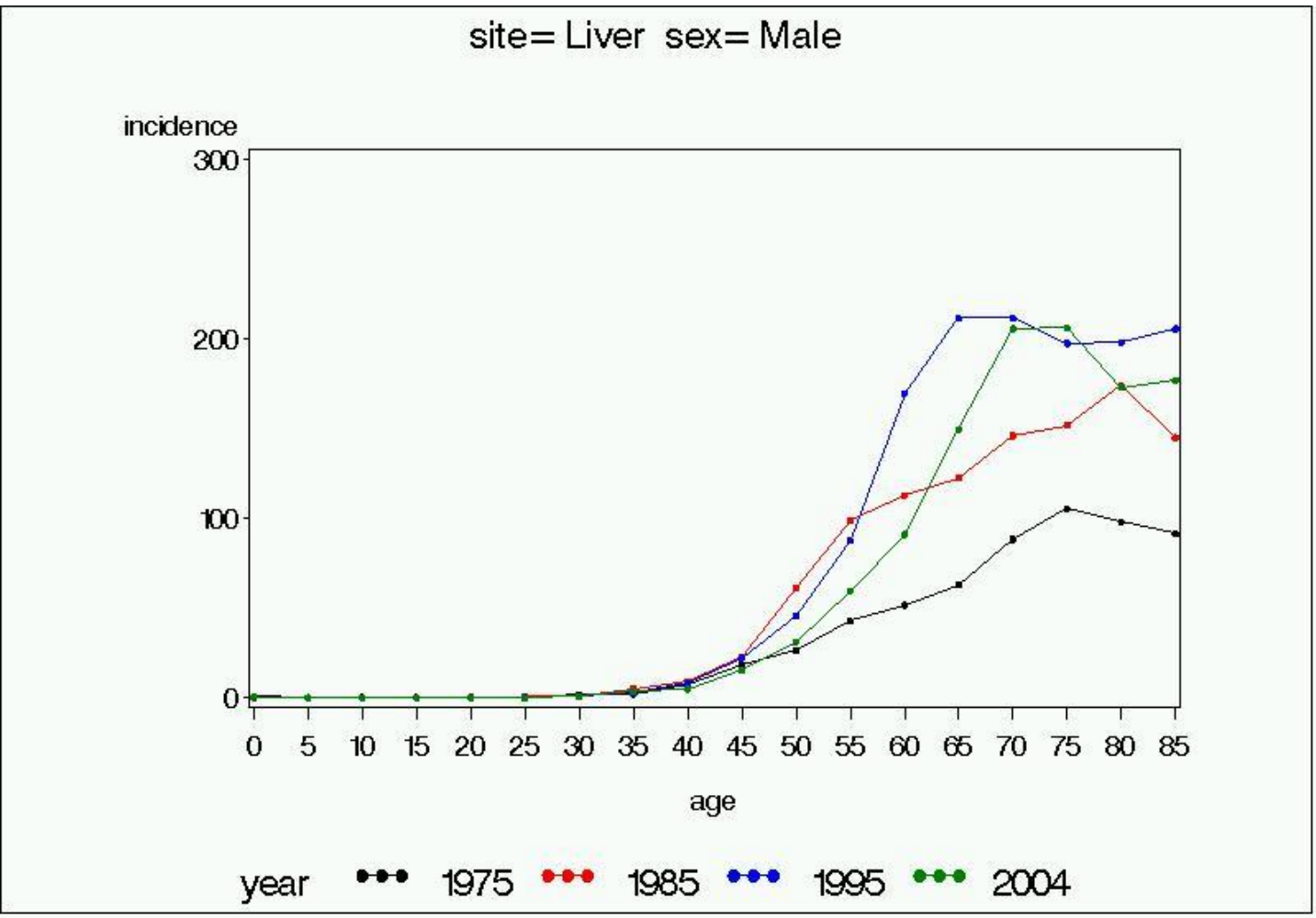


Figure. 33

APPENDIX FILES

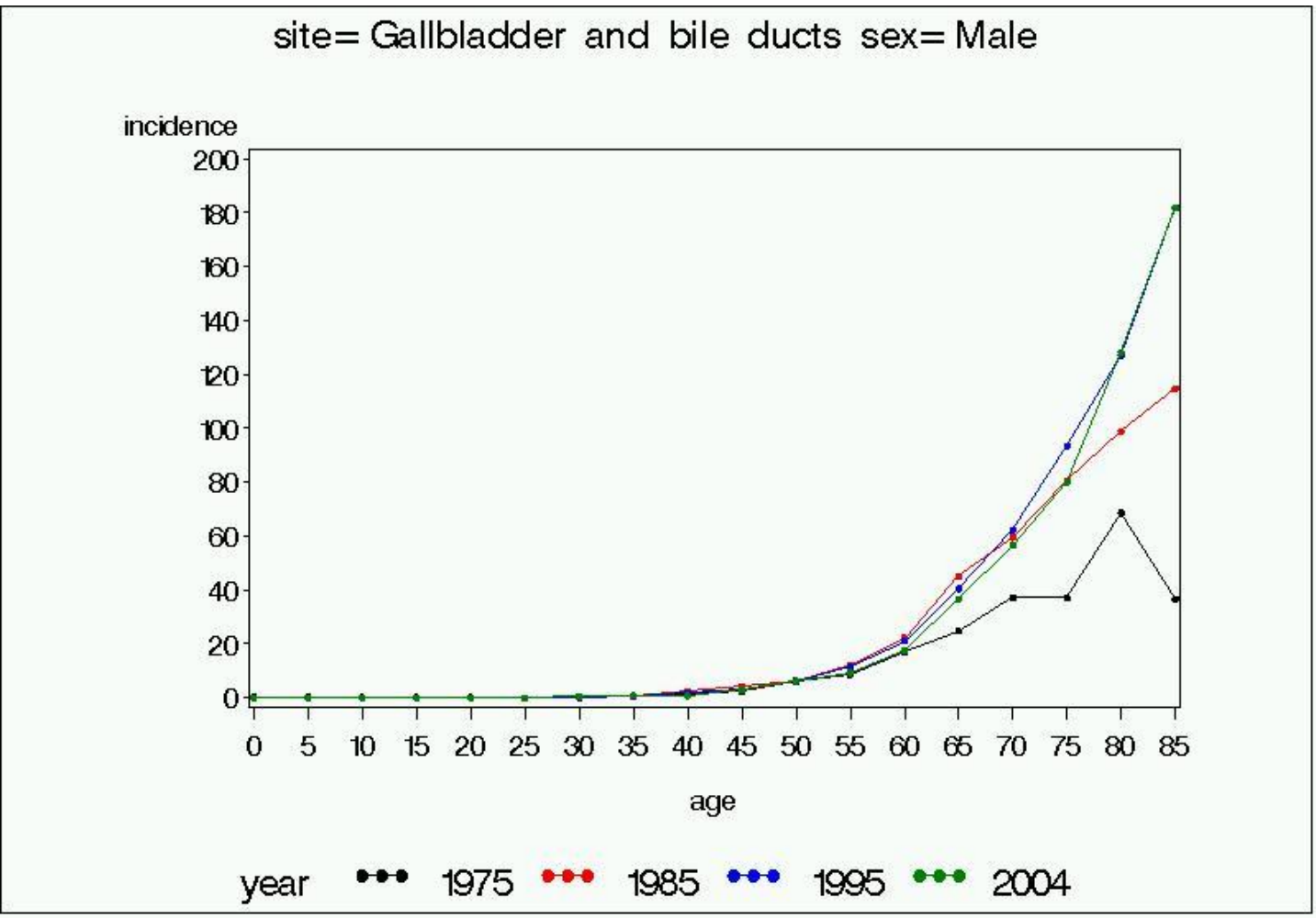


Figure. 34

APPENDIX FILES

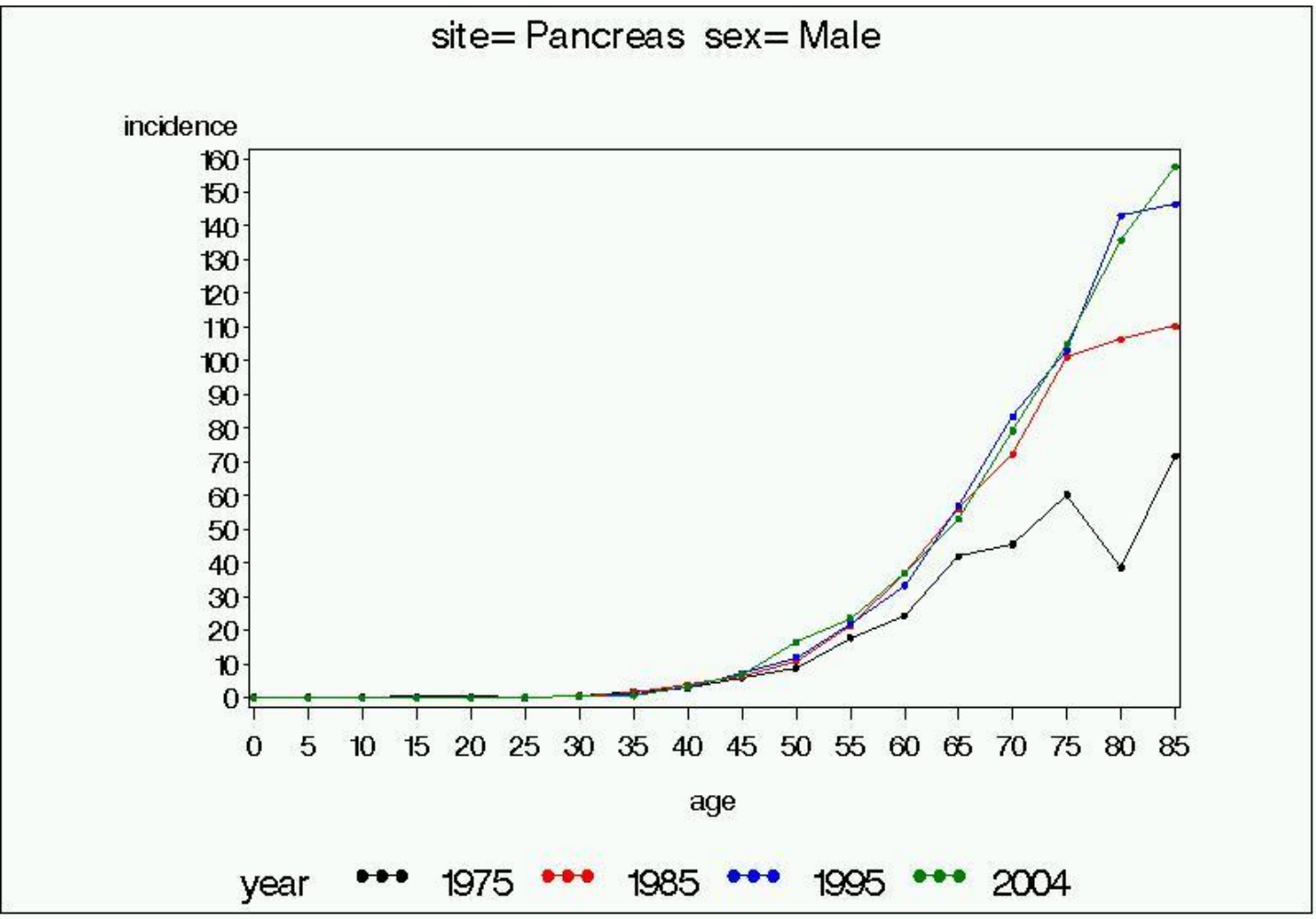


Figure. 35

APPENDIX FILES

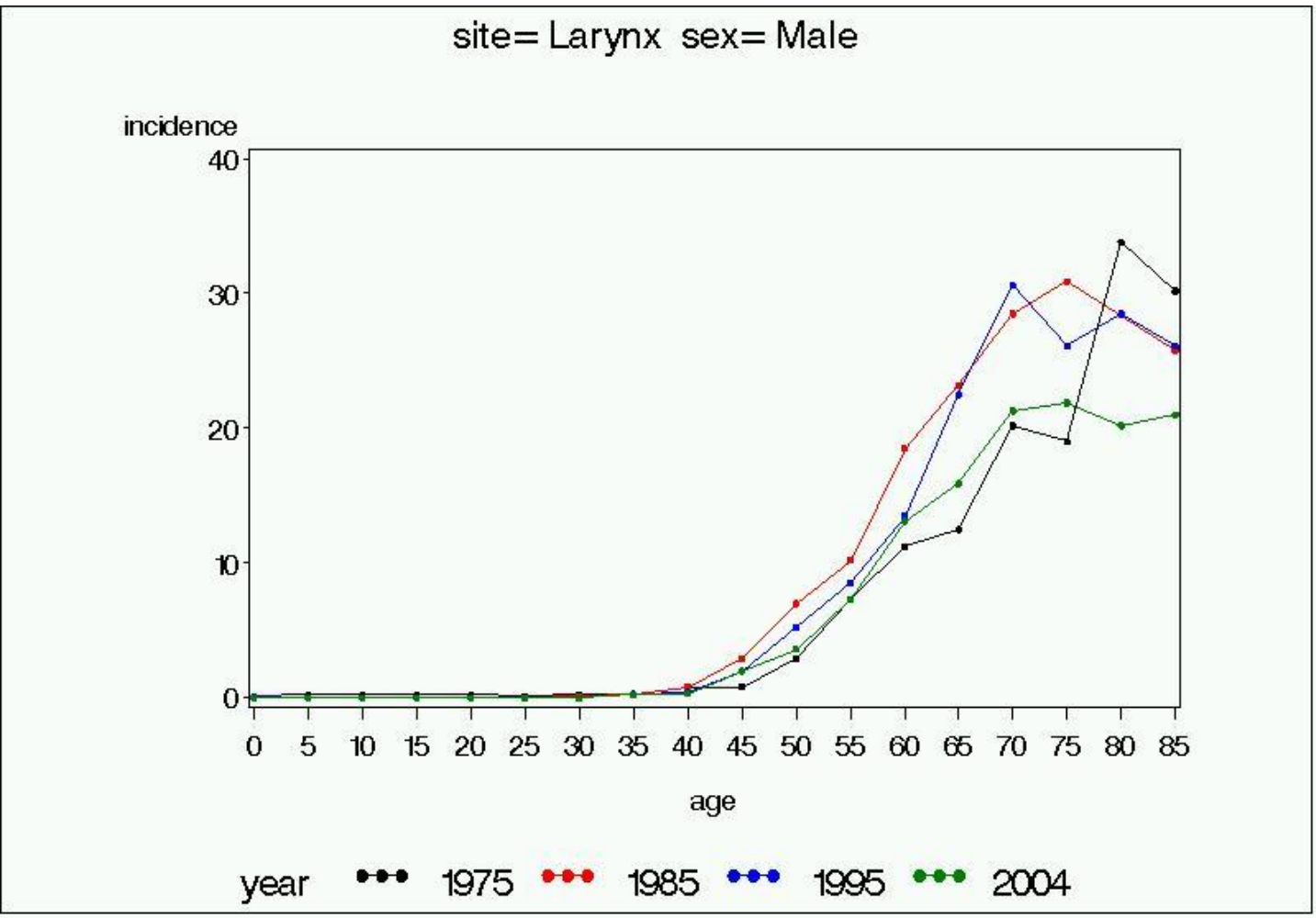


Figure. 36

APPENDIX FILES

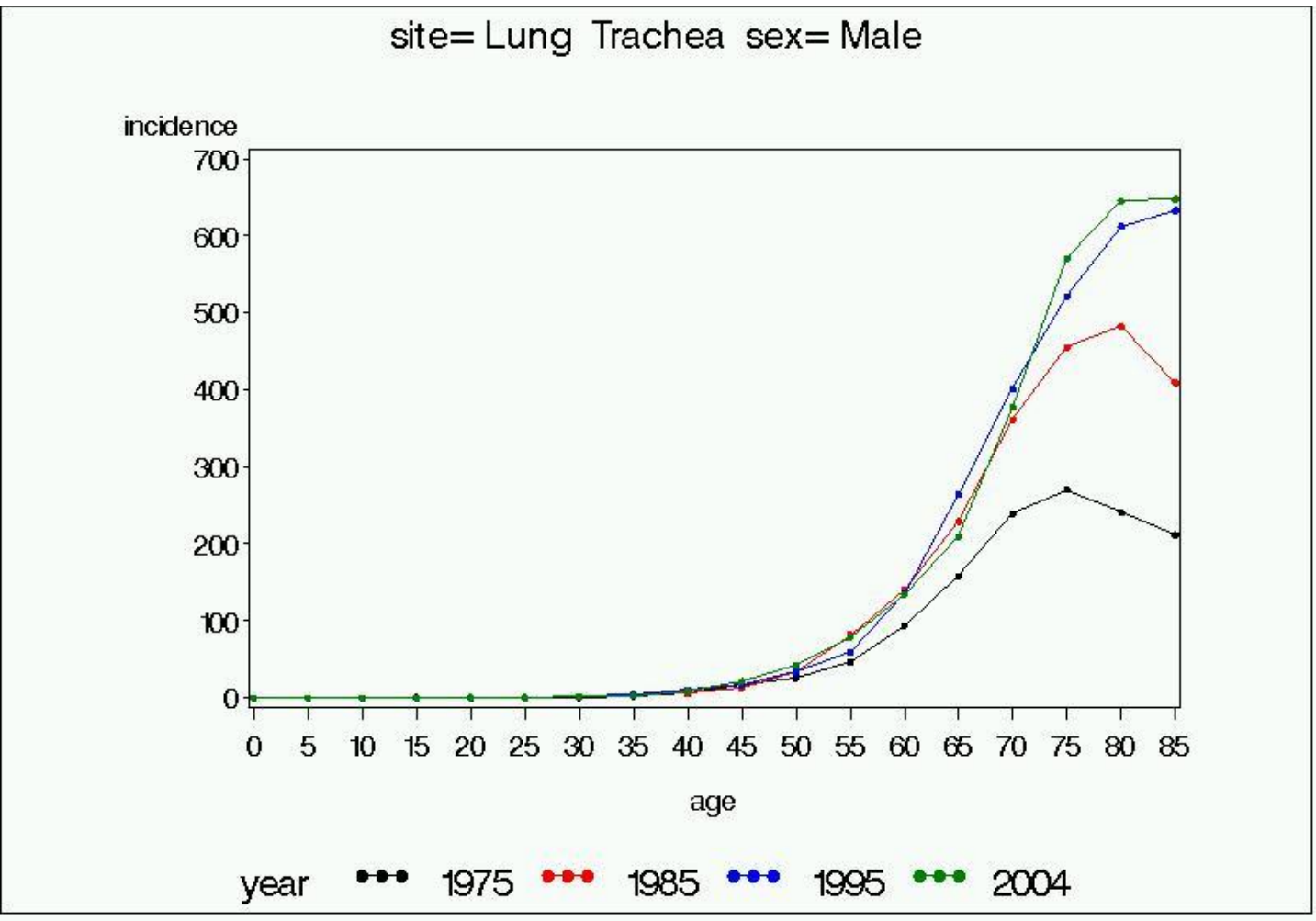


Figure. 37

APPENDIX FILES

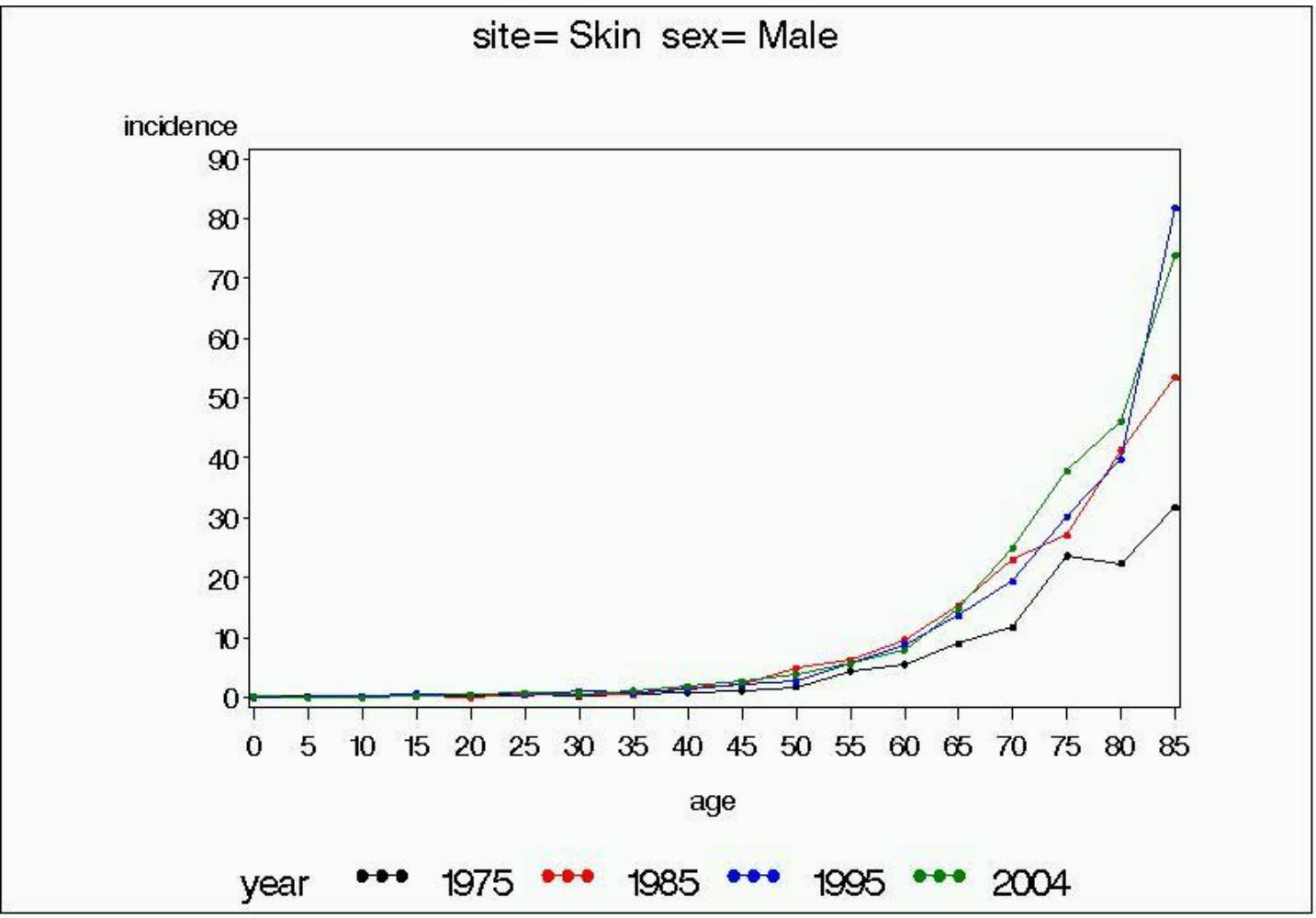


Figure. 38

APPENDIX FILES

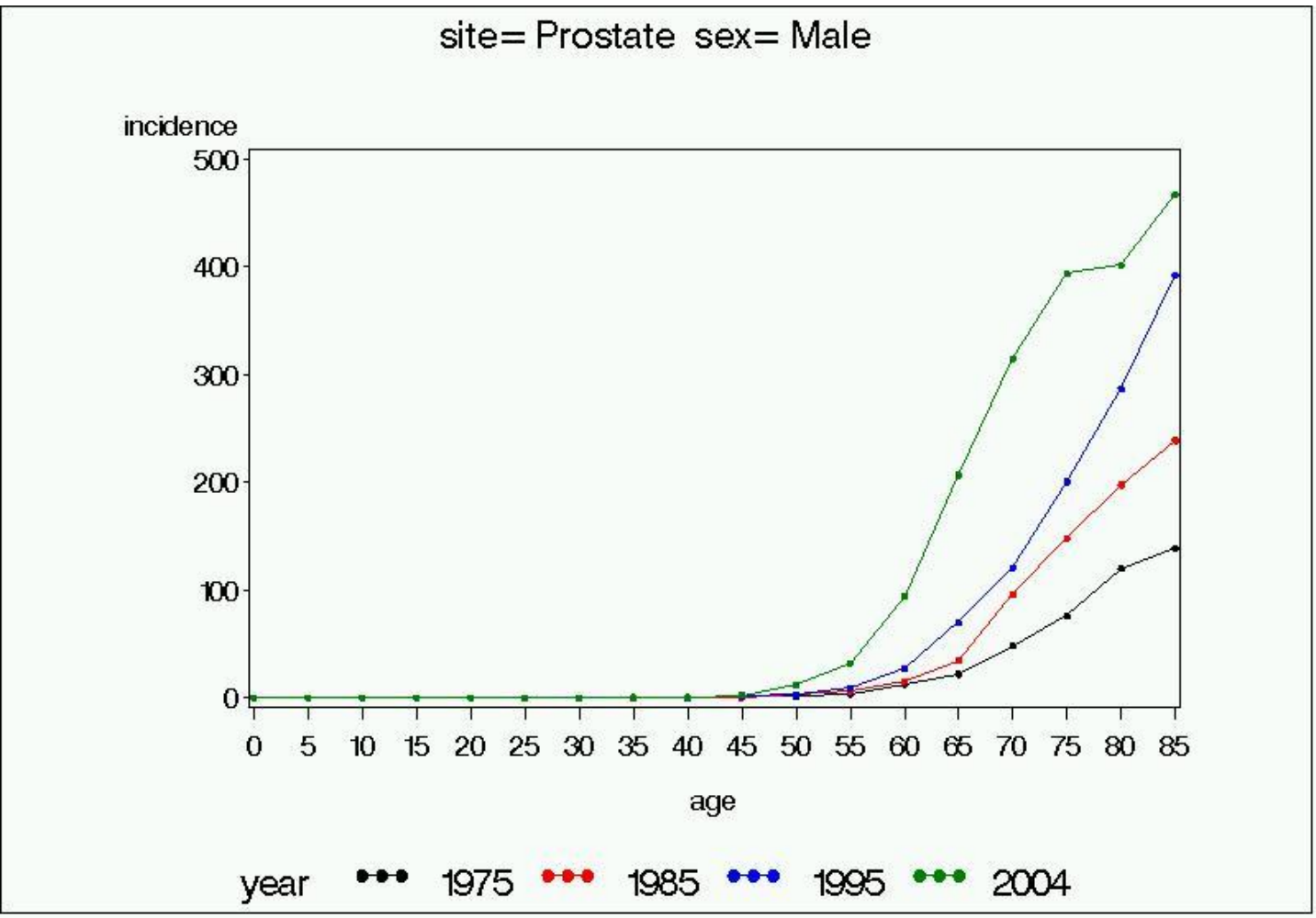


Figure. 39

APPENDIX FILES

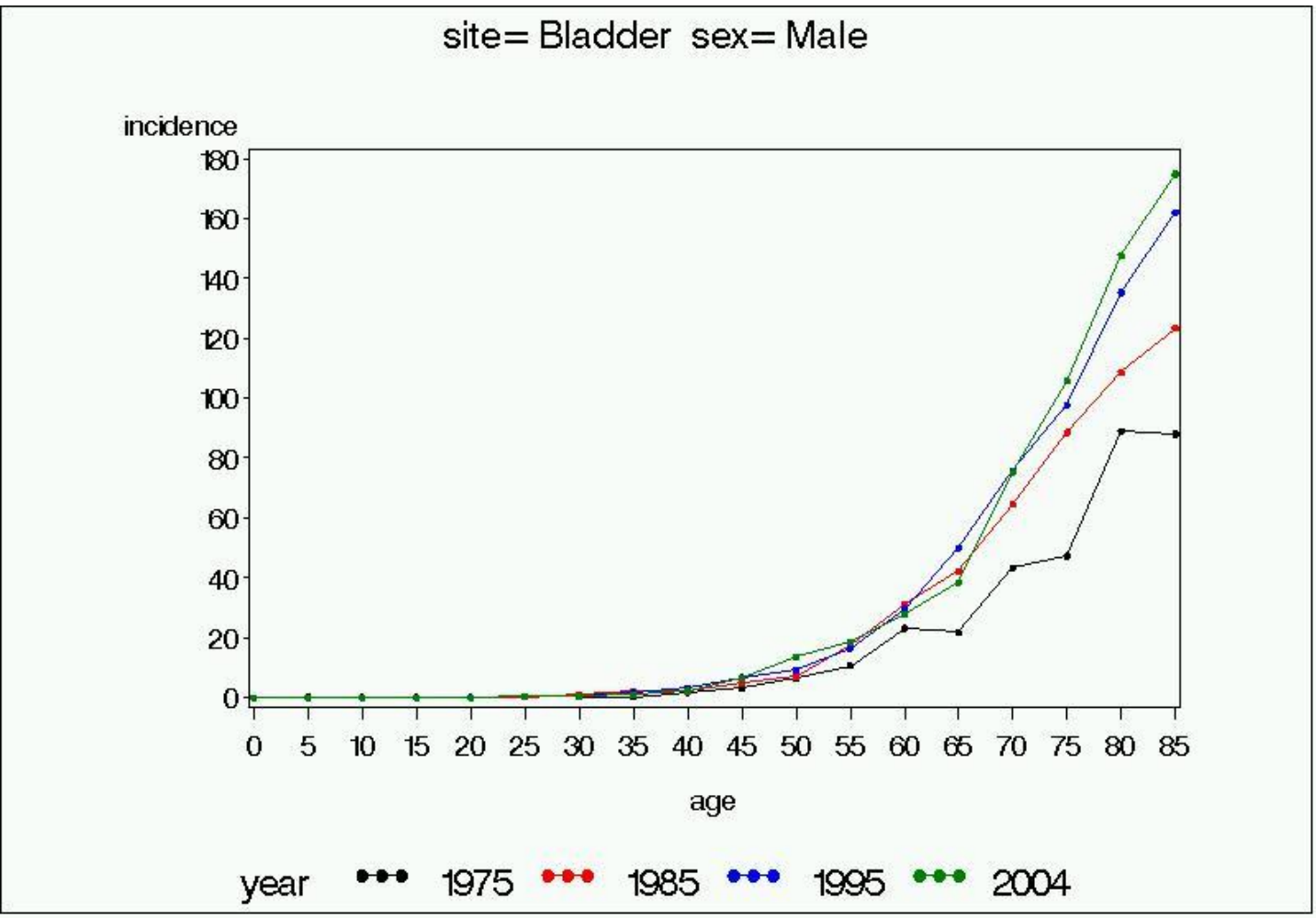


Figure. 40

APPENDIX FILES

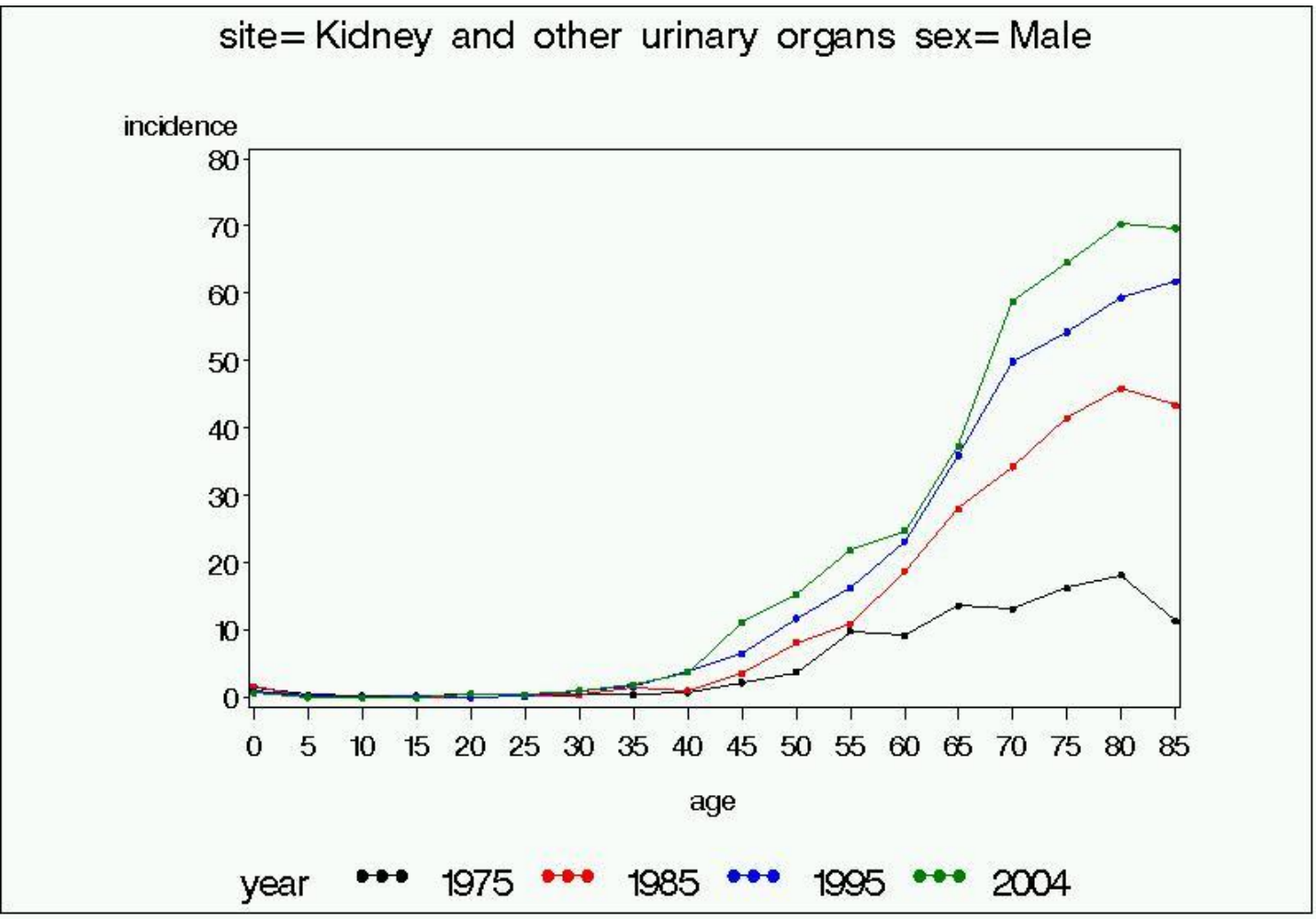


Figure. 41

APPENDIX FILES

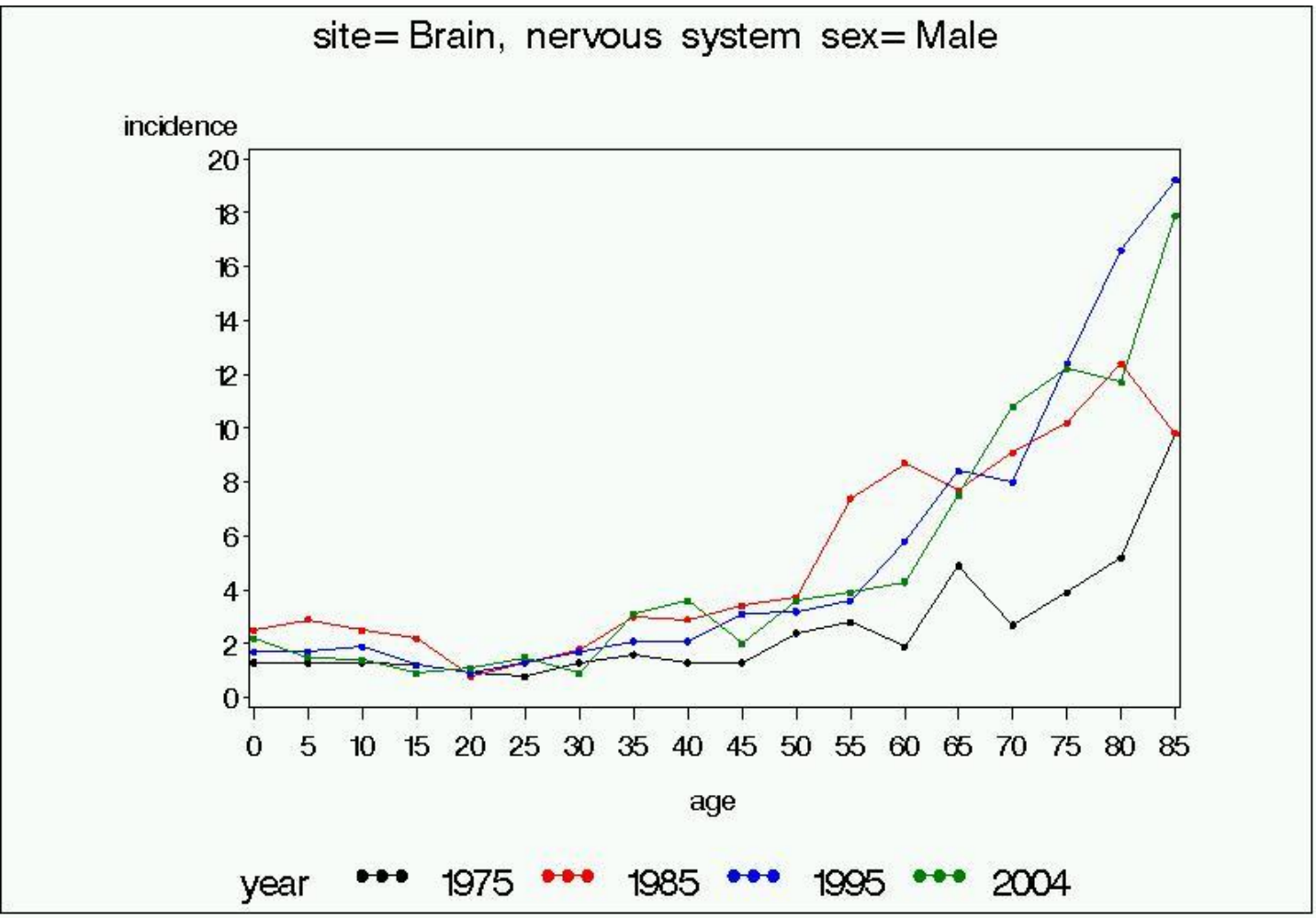


Figure. 42

APPENDIX FILES

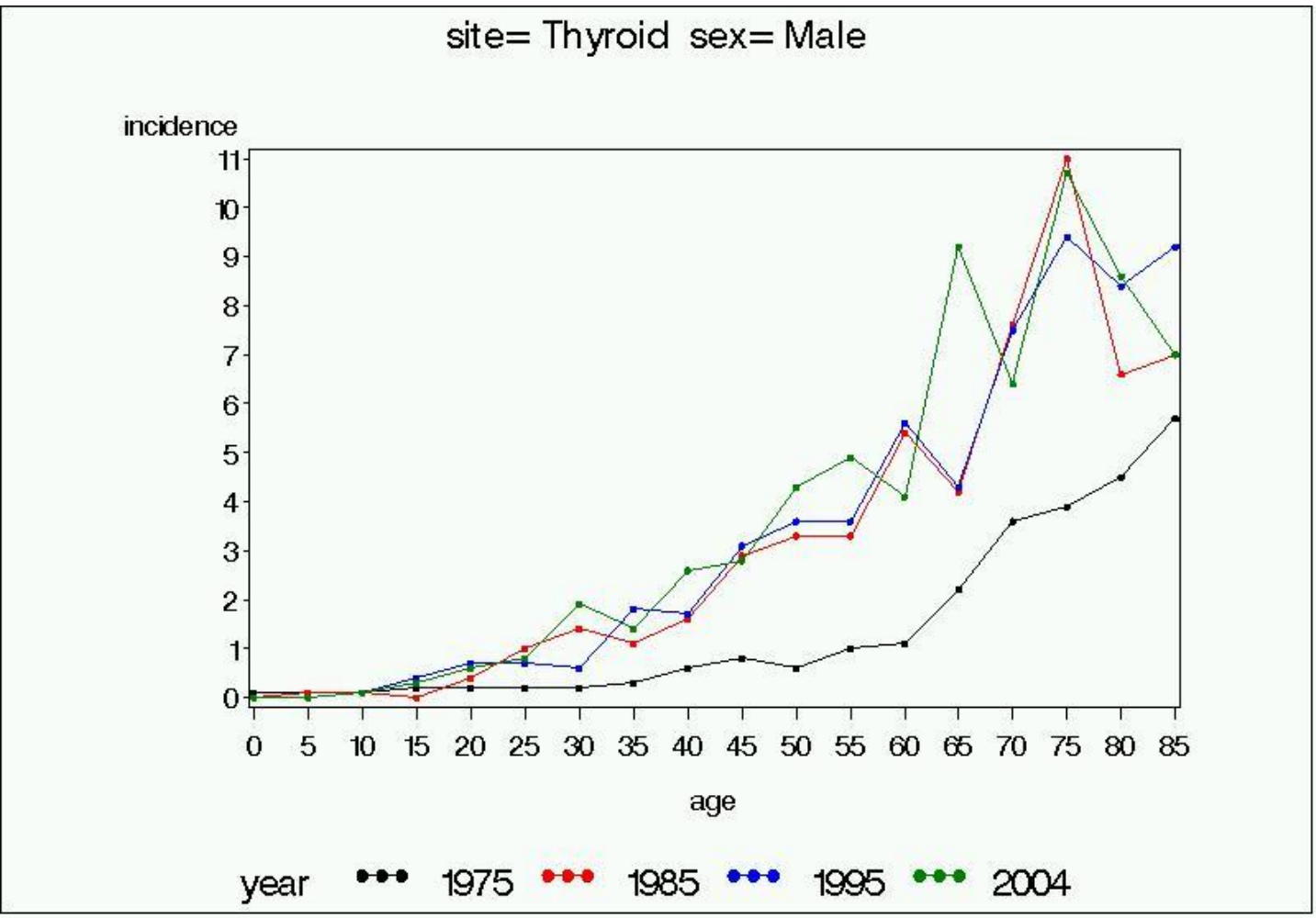


Figure. 43

APPENDIX FILES

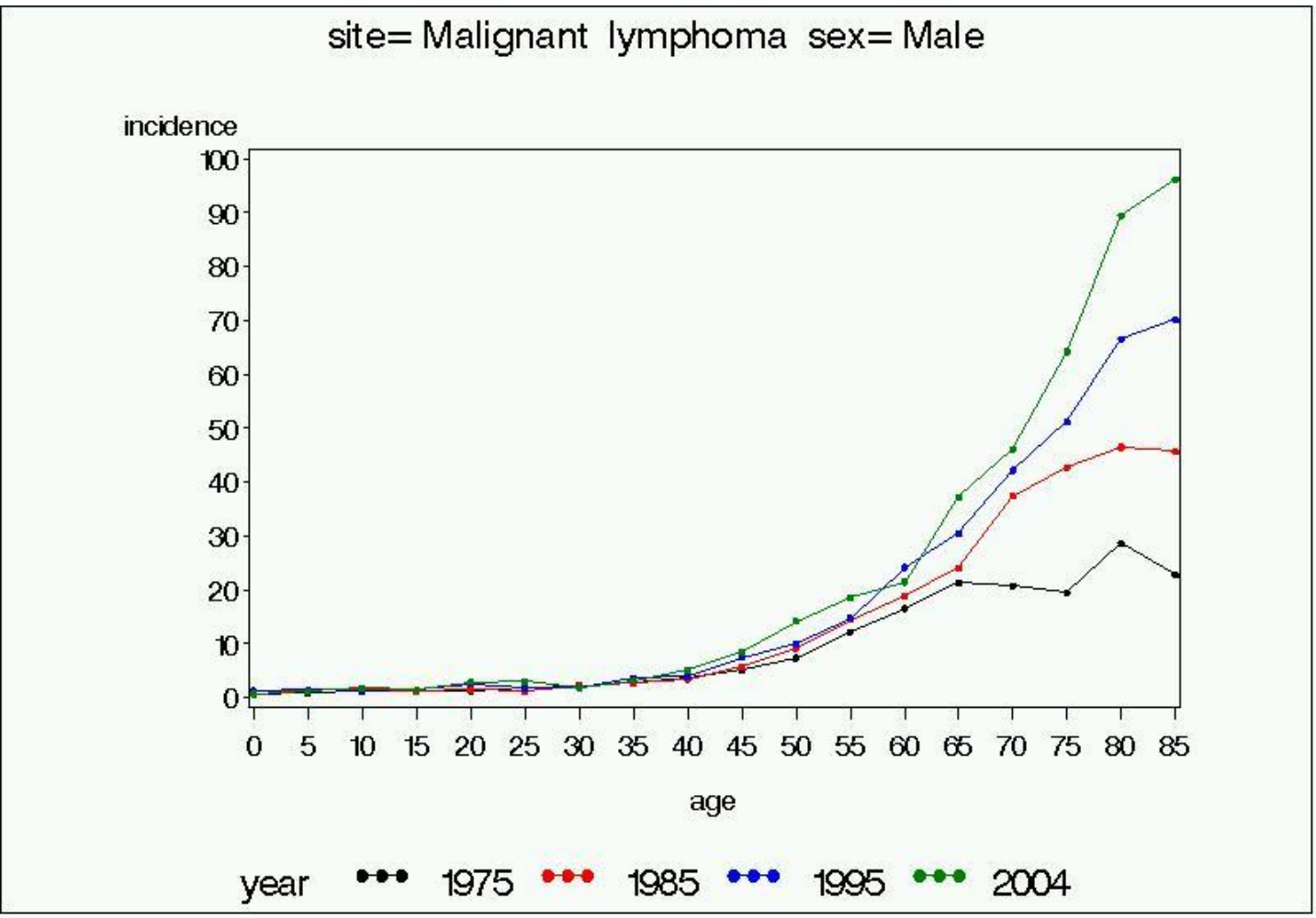


Figure. 44

APPENDIX FILES

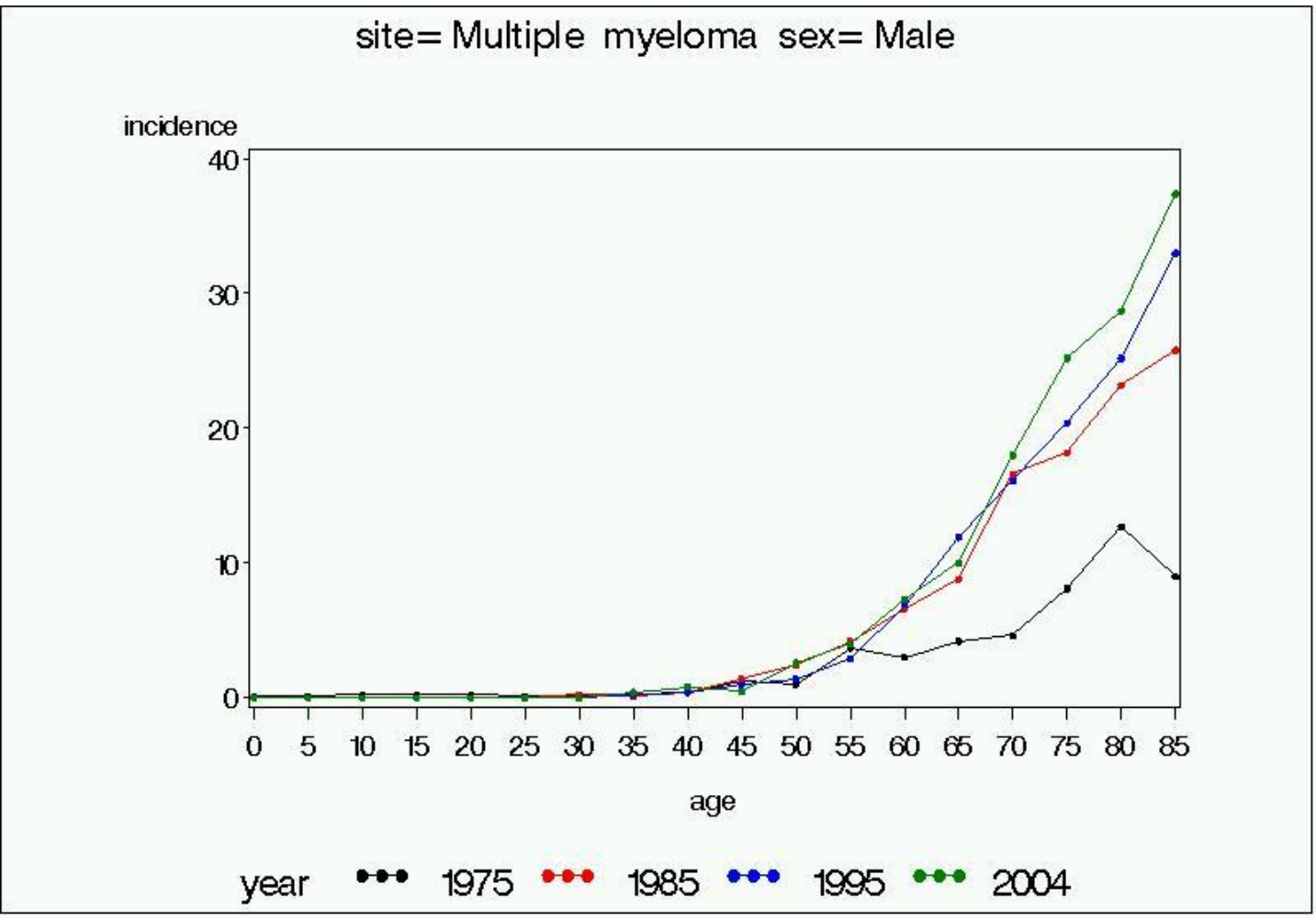


Figure. 45

APPENDIX FILES

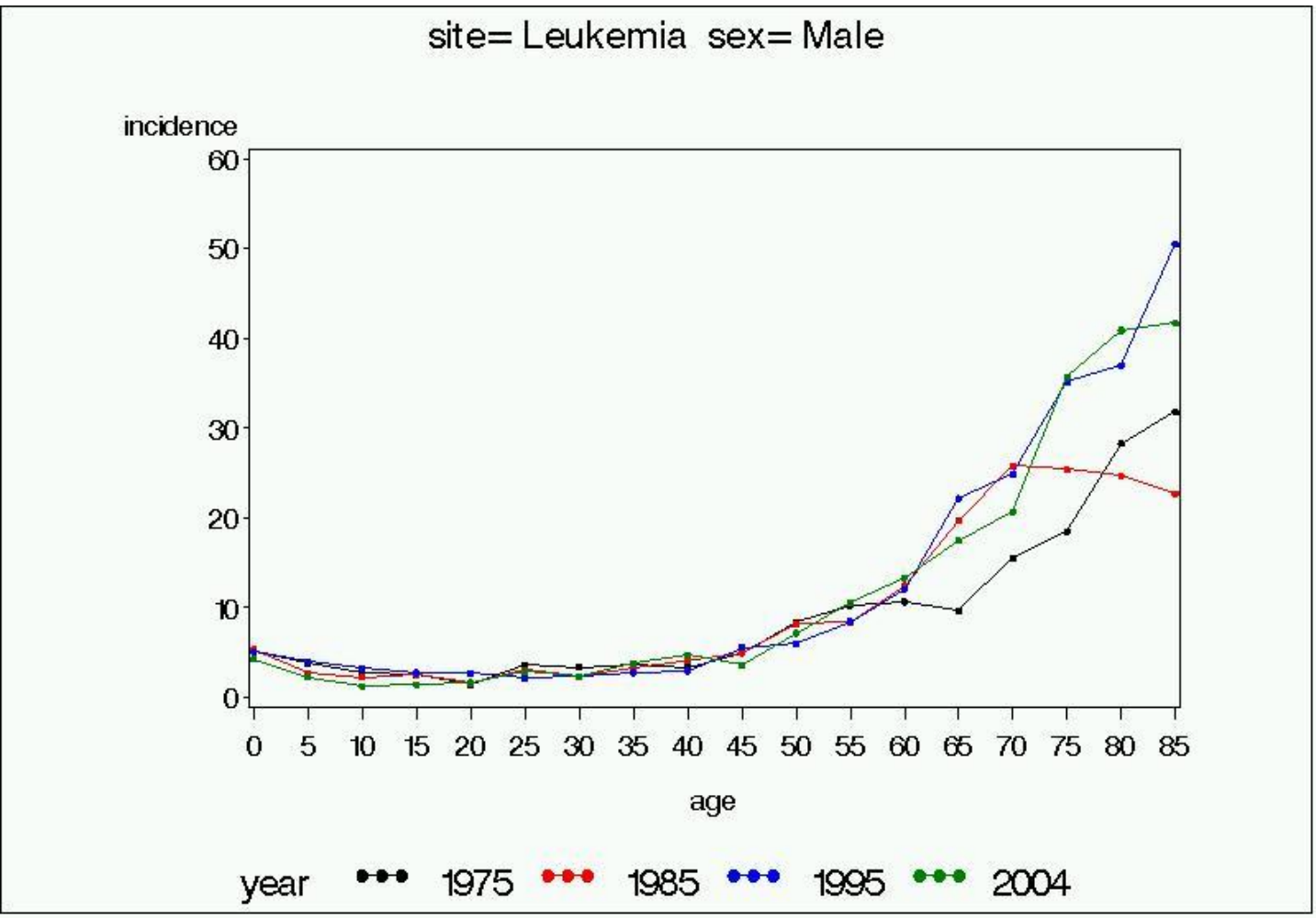


Figure. 46

APPENDIX FILES

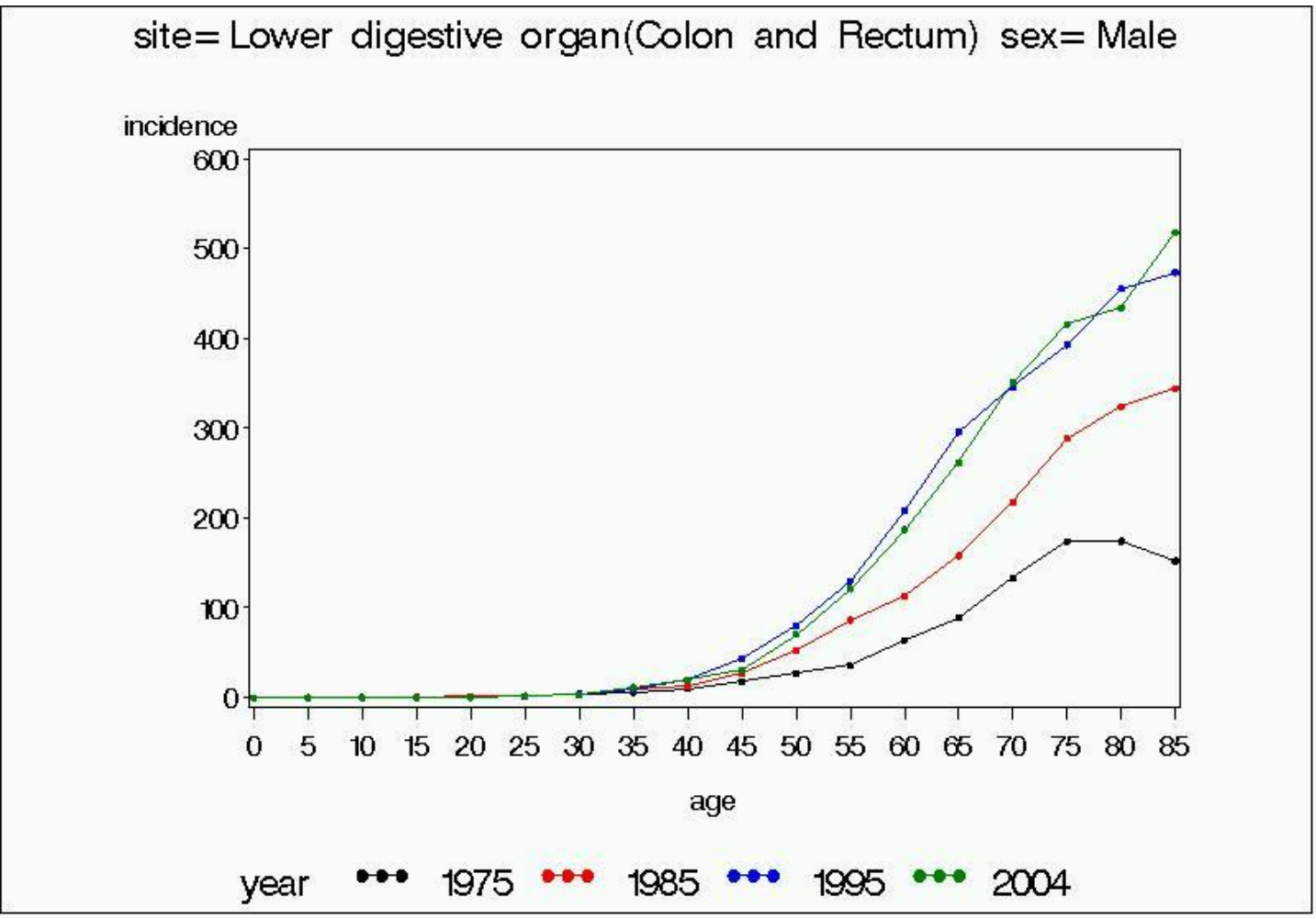


Figure. 47

APPENDIX FILES

code	site of origin	Age Range	Principal Component 1	Principal Component 2
1	All sites	0~	0.6353	0.1054
3	Oral cavity and pharynx	40~	0.8194	0.0685
4	Esophagus	40~	0.6526	0.1447
5	Stomach	20~	0.7652	0.1436
6	Colon	20~	0.8798	0.0492
7	Rectum	30~	0.8021	0.0811
8	Liver	40~	0.7055	0.2102
9	Gallbladder and bile ducts	40~	0.6486	0.2424
10	Pancreas	40~	0.7612	0.1125
11	Larynx	55~	0.4859	0.2789
12	Lung Trachea	30~	0.7276	0.1015
13	Skin	30~	0.4949	0.2791
20	Prostate	50~	0.9272	0.0335
21	Bladder	40~	0.7191	0.0946
22	Kidney and other urinary organs	30~	0.8445	0.0551
23	Brain, nervous system	0~	0.4666	0.2083
24	Thyroid	25~	0.7205	0.0763
25	Malignant lymphoma	0~	0.6015	0.1410
26	Multiple myeloma	40~	0.5188	0.1817
27	Leukemia	0~	0.3633	0.1695
67	Colon and Rectum = lower digestive organ	20~	0.8637	0.0506

Table. 1
The age range used in Lee-Carter method and the proportion of eigen value to the total (MALE)

APPENDIX FILES

code	site of origin	Age Range	Principal Component 1	Principal Component 2
1	All sites	0~	0.6353	0.1054
3	Oral cavity and pharynx	40~	0.8194	0.0685
4	Esophagus	50~	0.5464	0.1909
5	Stomach	20~	0.7652	0.1436
6	Colon	20~	0.8798	0.0492
7	Rectum	30~	0.8021	0.0811
8	Liver	40~	0.5486	0.2945
9	Gallbladder and bile ducts	40~	0.6486	0.2424
10	Pancreas	40~	0.7612	0.1125
11	Larynx	55~	0.4859	0.2789
12	Lung Trachea	30~	0.7276	0.1015
13	Skin	30~	0.4949	0.2791
20	Prostate	50~	0.9272	0.0335
21	Bladder	40~	0.7191	0.0946
22	Kidney and other urinary organs	30~	0.8445	0.0551
23	Brain, nervous system	0~	0.4666	0.2083
24	Thyroid	25~	0.7205	0.0763
25	Malignant lymphoma	0~	0.6015	0.1410
26	Multiple myeloma	40~	0.5188	0.1817
27	Leukemia	0~	0.3633	0.1695
67	Colon and Rectum = lower digestive organ	20~	0.8637	0.0506
25	Malignant lymphoma	0~	0.5069	0.1354
26	Multiple myeloma	40~	0.6226	0.1487
27	Leukemia	0~	0.4187	0.1565
67	Colon and Rectum = lower digestive organ	20~	0.7223	0.1295

Table. 2
The age range used in Lee-Carter method and the proportion of eigen value to the total (FEMALE)