UNIVERSITÀ DEGLI STUDI DI MILANO

Dipartimento di Scienze Cliniche e di Comunità

Laboratorio di Statistica Medica, Biometria ed Epidemiologia "G. A. Maccacaro"

Corso di Dottorato di Ricerca in

Epidemiologia, Ambiente e Sanità Pubblica

XXXI Ciclo - Settore scientifico disciplinare MED/01

Cancer mortality data analysis and prediction

Dottoranda: **Greta CARIOLI** (n° matr R11439)

Tutor: **Chiar.mo Prof. Carlo LA VECCHIA**

Coordinatore del Dottorato: **Chiar.mo Prof. Carlo LA VECCHIA**

A.A. 2017/2018

Index

Abstract

Descriptive epidemiology has traditionally only been concerned with the definition of a research problem's scope. However, the greater availability and improvement of epidemiological data over the years has led to the development of new statistical techniques that have characterized modern epidemiology. These methods are not only explanatory, but also predictive. In public health, predictions of future morbidity and mortality trends are essential to evaluate strategies for disease prevention and management, and to plan the allocation of resources.

During my PhD at the school of "Epidemiology, Environment and Public Health" I worked on the analysis of cancer mortality trends, using data from the World Health Organization (WHO) database, available on electronic support (WHOSIS), and from other databases, including the Pan American Health Organization database, the Eurostat database, the United Nation Population Division database, the United States Census Bureau and the Japanese National Institute of Population database. Considering several cancer sites and several countries worldwide, I computed age-specific rates for each 5-year age-group (from 0–4 to 80+ or 85+ years) and calendar year or quinquennium. I then computed age-standardized mortality rates per 100,000 person-years using the direct method on the basis of the world standard population. I performed joinpoint models in order to identify the years when significant changes in trends occurred and I calculated the corresponding annual percent changes.

Moreover, I focused on projections. I fitted joinpoint models to the numbers of certified deaths in each 5-year age-group in order to identify the most recent trend slope. Then, I applied Generalized Liner Model (GLM) Poisson regressions, considering different link functions, to the data over the time period identified by the joinpoint model. In particular, I considered the identity link, the logarithmic link, the power five link and

the square root link. I also implemented an algorithm that generated a "hybrid" regression; this algorithm automatically selects the best fitting GLM Poisson model, among the identity, logarithmic, power five, and square root link functions, to apply for each age-group according to Akaike Information Criterion (AIC) values. The resulting regression is a combination of the considered models.

Thus, I computed the predicted age-specific numbers of deaths and rates, and the corresponding 95% prediction intervals (PIs) using the regression coefficients obtained previously from the four GLM Poisson regressions and from the hybrid GLM Poisson regression. Lastly, as a further comparison model, I implemented an average model, which just computes a mean of the estimates produced by the different considered GLM Poisson models.

In order to compare the six different prediction methods, I used data from 21 countries worldwide and for the European Union as a whole, I considered 25 major causes of death. I selected countries with over 5 million inhabitants and with good quality data (i.e. with at least 90% of coverage). I analysed data for the period between 1980 and 2011 and, in particular, I considered data from 1980 to 2001 as a training dataset, and from 2002 to 2011 as a validation set. To measure the predictive accuracy of the different models, I computed the average absolute relative deviations (AARDs). These indicate the average percent deviation from the true value. I calculated AARDs on 5-year prediction period (i.e. 2002-2006), as well as for 10-year period (i.e. 2002-2011). The results showed that the hybrid model did not give always the best predictions, and when it was the best, the corresponding AARD estimates were not very far from the other methods. However, the hybrid model projections, for any combination of cancer site and sex, were never the worst. It acted as a compromise between the four considered models. The average model is also ranked in an intermediate position: it never was the best predictive method, but its AARDs were competitive compared to the

other methods considered. Overall, the method that shows the best predictive performance is the Poisson GLM with an identity link function. Furthermore, this method, showed extremely low AARDs compared to other methods, particularly when I considered a 10-year projection period.

Finally, we must take into account that predicted trends and corresponding AARDs derived from 5-year projections are much more accurate than those done over a 10-year period. Projections beyond five years with these methods lack reliability and become of limited use in public health.

During the implementation of the algorithm and the analyses, several questions emerged: Are there other relevant models that can be added to the algorithm? How much does the Joinpoint regression influence projections? How to find an “a priori” rule that helps in choosing which predictive method apply according to various available covariates? All these questions are set aside for the future developments of the project.

Prediction of future trends is a complex procedure, the resulting estimates should be taken with caution and considered only as general indications for epidemiology and health planning.

Introduction

Descriptive epidemiology and its statistical techniques are fundamental instruments for exploratory studies in order to generate new hypotheses and/or verify them. Descriptive epidemiology is usually considered a first approach to define the purpose and scope of a research investigation.

The basic techniques of descriptive epidemiology were borrowed from demography and the key descriptive tools were morbidity and mortality rates. Their comparison and their standardization were and are still the main methods used; however, statistical variability was rarely taken into account, sometimes producing serious errors and misleading interpretations [1].

Over the years, especially cancer registries have improved the quantity and the quality of data, also working on the standardization of definitions/classifications and on registration procedures. Cancer incidence and mortality data are routinely recorded in cancer registries, and became the basic data for cancer surveillance. On the other hand, demographic data were also published on a more regular basis and became available for an increasing number of populations [2].

The improvement and the greater availability of epidemiological and time based mortality data over the years brought the development of techniques that characterized modern descriptive epidemiology. These new techniques, mainly based on mathematical modelling, were developed focusing on the analysis of time series. They aimed to identify different factors that underlie the changes in rates. Specifically, historical oncologic data recorded in cancer registries could provide us with rich information on the changes of cancer incidence and mortality over the years; the trends of changing rates reflect the changes of the underlying risks. Thus, the collection of increasingly detailed morbidity and mortality data, and the creation of data systems

which allow cases and deaths to be located in time and space, have provided a solid basis for the evaluation of time series trends, in turn requiring the development of appropriate statistical methods. These methods are not only explanatory, but also predictive [1].

Prediction of a future event is a complex process subject to large uncertainties and, for many aspects, questionable. However, in several human activities and working areas, it is useful to obtain information on future trends, even if uncertain or imprecise. In demography, for example, it is common practice to produce population structure projections for future decades, although it is known that the fertility, mortality and migratory patterns may vary considerably in relatively short periods, and thus substantially modify the subsequent population structure. Regarding oncologic data, which are at the basis of this thesis, the prediction of future cancer mortality rates is essential to plan the allocation of resources and to evaluate strategies for prevention and cancer management. Indeed, the actual available data are, usually, 2-3 or more years old [3].

During my PhD studies in the school of "Epidemiology, Environment and Public Health" I worked on the analysis of cancer mortality trends; my focus was on projections. Over the first year of my PhD I implemented a predictive method that I continued to investigate and improve during the second and third years. I started developing an algorithm to compute a "hybrid" regression, that was a mixture of linear, log-linear, power five, and square root regressions. This algorithm automatically chose the best model to apply for predictions according to R-squared values. I compared the hybrid regression results to those from linear and log-linear regressions. Then, during my second year, I replaced the simple linear regressions with more appropriate models: I considered Poisson Generalized Linear Model (GLM) regressions with different link functions (identity, logarithmic, power five and square root) and refined

the algorithm that automatically, now on the basis of Akaike information criterion (AIC) values instead of R-squared ones, selected the best fitting model to use, generating the hybrid regression.

All these models have been applied to the European Union (EU) data, including several cancer sites. Data referred to the period 1980-2011; I used data from 1980 to 2001 as a training dataset so that I could "predict" data for the following period 2002-2011 (validating dataset). Then, I compared the predicted data with the observed with the aim to identify the most performing model.

A further step that I have achieved during my third year of PhD was to extend the database to other 21 countries besides the EU, and to other several causes of death, for a total of 25, obtaining a more consistent database and, as a consequence, estimates. Then, in addition to reviewing the algorithm, I also added an average model, which just computes a mean of the estimates produced by the different considered GLM Poisson models.

Thus in this thesis, I will describe the methods and modelling techniques I used to study and project mortality rates in detail. I will also describe the main results obtained by applying the six different models previously described: the GLM Poisson models with the identity, logarithmic, power five, and square root link functions, the hybrid regression and the average regression.

Mortality rate

Introduction

In epidemiology, there are several indicators and measures to describe different aspects of a population's health. The rate is the most widely used. It measures the instantaneous change of a quantity (for example the switch from health to sickness or from living to dead) compared to the change in unit of another quantity (time).
If the event of interest is the number of deaths occurring in a population for a given cause in a certain period, the most suitable indicator is the mortality rate.

Instantaneous risk

To identify factors that cause an event (for example the disease onset or the death of an individual) it is necessary to calculate its risk, i.e. the probability of that event, which will depend on either individual characteristics, such as age, or environmental factors. The risk function determines how the risk of an event's occurrence changes over time or with age.
To give a definition of risk it is necessary to specify a time scale to measure it against an initial time point from which the risk will be measured. In epidemiology (as well as in demography), time can be measured in three ways: age, calendar period, and cohort of birth. The first two time indices correspond to the Lexis diagram axes [1].

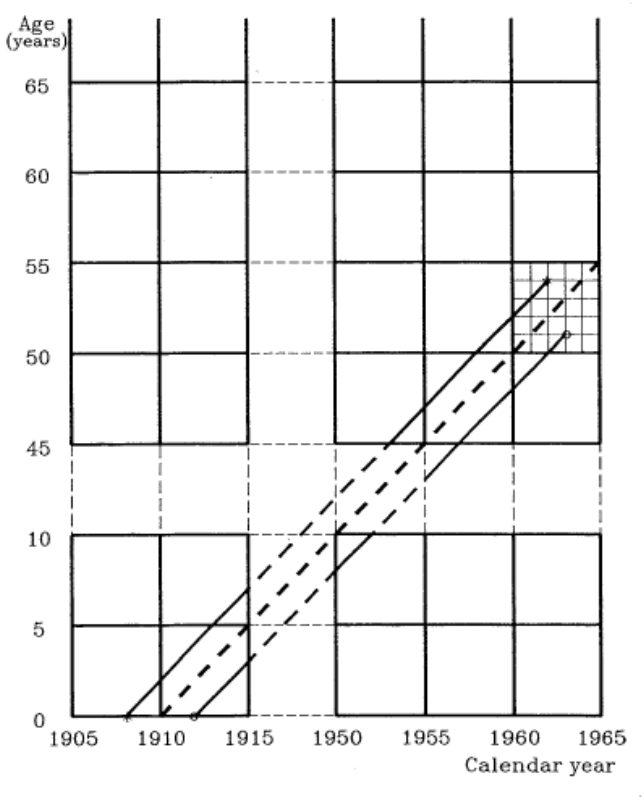

Figure 1. Lexis diagram.

Figure 1 illustrates the Lexis diagram structure. Every segment of oblique lines in this graph represent the observable fraction of an individual life, from the starting observation and the event realization (i.e. the interval of time and age during which an event of interest can occur). The left extremity of the segment represents the start of observation (for example the date of birth or the time of treatment initiation), the right one is the end of observation (the date and age at which either the event under study occurred or the subject stopped being observed). In order to obtain a measure of the time from initial observation since the occurrence of the event, you can project the oblique segment on one of the two axes of the diagram.

Once a time scale is established, the distribution of the time between the starting observation and the occurrence of the event is of fundamental importance. If for example the interest is in studying the risk of dying after a cancer diagnosis,

distribution knowledge of the time elapsing between diagnosis and death, would allow to calculate the risk of death within a year of diagnosis, or the probability of surviving up to a given age.

Thus, an appropriate mathematical model to calculate these risks is essential. The main assumption is that the event under study is a non-recurring event, i.e. once it has occurred, it cannot be repeated another time for the same individual. Death is obviously a non-recurring event, while the onset of a disease is not necessarily so.

Consider the following quantities:

- $\boldsymbol{T}$ - is a random variable representing the time before the occurrence of a specific event for a subject in a certain population. T must be positive ($T > 0$), i.e. at the beginning of observation the individual must not have experienced the event.
- $\boldsymbol{f(t)}$ - is the density function of T and $\boldsymbol{F(t)} = P(T \leq t)$ is the probability distribution ($F'(t) = f(t)$).
- $\boldsymbol{S(t)}$ - is the survival function, that is the probability that a subject experiences the event over a certain time t. $\boldsymbol{S(t)} = P(T > t) = 1 - F(t)$.

The instantaneous rate or risk function is defined as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T < t + \Delta t | T \geq t)$$

The previous formula is the ratio between the conditional probability that the event occurs at time t and the corresponding time interval Δt [4].

The instantaneous rate $\lambda(t)$ is not a probability, but a probability per unit time, also called probability rate.

The event of interest, for mortality data analysis, is death and it is called the instantaneous mortality rate. The higher $\lambda(t)$, more likely a death will occur between t and the next instant, therefore $\lambda(t)$ provides a measure of the force of mortality at time t.

It is possible to rewrite $\lambda(t)$ with the following formula:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{1 - F(t)}$$

Which, replacing $1 - F(t)$ with $S(t)$, becomes:

$$\lambda(t)S(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

$$\lambda(t)S(t) = F'(t)$$

Thus, the probability that an event occurs before a certain time t, $\pi(t)$, can be written as:

$$\pi(t) = \int_0^t F'(u)du$$

or also as:

$$\pi(t) = \int_0^t \lambda(u)S(u)du$$

The probability of death, π, between age t_0 and t_1, can be define as:

$$\pi = \int_{t_0}^{t_1} \lambda(u)S(u)du$$

Where $\lambda(u)$ is the age-specific rate and $S(u)$ the probability of survival without disease [1].

The main interest focuses on the calculation of the conditional probability of death, π_c, between the age t_0 and t_1 given that a subject is still at risk at age t_0. This probability is not influenced by overall survival until the age t_0 and, if the range between t_0 and t_1 is small, influenced very little by survival.
Then, it is possible to calculate π_c using the following formula:

$$\pi_c = \int_{t_0}^{t_1} \lambda(u)\frac{S(u)}{S(t_0)}du$$

Under the assumption that the interval $[t_0; t_1]$ is small enough, $\lambda(u)$ and $S(u)$ can be considered constant - $\lambda(t_0), S(t_0)$, then the equation can be rewritten in the following form:

$$\pi_c \approx \lambda(t_0)(t_1 - t_0)$$

If e is the number of observed events between t_0 and t_1, and n_{t_0} the number of subjects at risk at time t_0, then the following formula gives the π_c estimate:

$$\widehat{\pi_c} = \frac{e}{n_{t_0}}$$

And the estimate for λ is:

$$\hat{\lambda}(t_0) \approx \frac{e}{n_{t_0}(t_1 - t_0)}$$

Consequently, dividing the number of observed events by the number of person years (m) accumulated between t_0 and t_1, you obtain the more familiar estimate formula of the instantaneous rate at time t_0:

$$\hat{\lambda}(t_0) \approx \frac{e}{m}$$

If $\lambda(u)$ varies markedly between t_0 and t_1, or if the ratio $S(u)/S(t_0)$ is very different from unity, this approximation does not hold [1].

We can define the instantaneous rate as [4]:

$$\lambda = \frac{p}{t} = \frac{e/n}{t} = \frac{e}{n * t}$$

That is:

$$\lambda = \frac{event\ probability}{observation\ time} = \frac{number\ of events/subjects\ at\ risk}{observation\ time} =$$

$$= \frac{number\ of\ events}{subjects\ at\ risk * observation\ time}$$

Thus, the rate is an instantaneous quantity because the observation time is only used as an operational element, to make the calculations (average rate). However, since the follow-up time appears directly in the rate definition, the rate takes into account the events that occurred and the time during which they occurred [4].

Crude mortality rate

The crude mortality rate measures the frequency of deaths observed in a population in a given period of time (conventionally a calendar year) and is the easiest to calculate.

Formal definition of mortality rate:

The ratio between the number of deaths from a certain cause that occurred in the study population in a given period (= numerator) and the total population at risk in the same period considered (= denominator).

$$Mortality\ rate = \frac{\begin{array}{c} n°\ of\ deaths\ from\ a\ certain\ cause \\ that\ occurred\ in\ the\ study\ population \\ in\ a\ given\ period \end{array}}{\begin{array}{c} total\ person-time\ at\ risk \\ in\ the\ same\ given\ period \end{array}}$$

The mortality rate indicates the average speed with which a group of individuals switches from a state of risk to a state of death in the time unit.

In the specific case of annual mortality rates, the numerator of the rate is the number of deaths for the observed condition during the calendar year, while the denominator, consists of the population estimates derived from the census. In other words, the

person-years are expressed as the number of individuals present in mid-year (or the yearly average).

$$Annual\ mortality\ rate = \frac{\begin{array}{c} n°\ of\ deaths\ from\ a\ certain\ cause \\ that\ occurred\ in\ the\ study\ population \\ during\ the\ calendar\ year \end{array}}{\frac{(subjects\ at\ risk\ at\ the\ beginning + subjects\ at\ risk\ at\ the\ end)}{2}}$$

Since each population can be considered as a set of different homogeneous subgroups, the value of the generic rate can be seen as an average of the values measured in every subgroup.
These values are weighted by the size of the specific subgroup: the larger the subgroup, the greater the influence on the crude measurement.

Consider a population composed by age-subgroups, where:

- $\boldsymbol{N}$ - is the total size of the population at risk;
- $\boldsymbol{D}$ - is the total number of deaths observed in the population;
- $\boldsymbol{i}$ - indicates i-th stratum;
- $\boldsymbol{n_i}$ - is the size of the population in the i-th stratum;
- $\boldsymbol{d_i}$ - is the number of deaths observed in the i-th stratum.

The sum of all n_i gives the total size of the population (N), while the sum of all d_i gives the total number of deaths (D). The crude mortality rate is given by the ratio $^{D}/_{N}$, that is the weighted average of the stratum specific mortality rates ($^{d_i}/_{n_i}$). Each specific mortality rate contributes with a weight proportional to its stratum population ($^{n_i}/_{N}$):

$$\frac{D}{N} = \frac{\sum_{i=1}^{n} d_i}{N} = \frac{\sum_{i=1}^{n} n_i \, (d_i/n_i)}{N} = \sum_{i=1}^{n} \left(\frac{n_i}{N}\right)\left(\frac{d_i}{n_i}\right) = \sum_{i=1}^{n} w_i \left(\frac{d_i}{n_i}\right)$$

wi represent the weights and their sum is equal to unity:

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \left(\frac{n_i}{N}\right) = \frac{\sum_{i=1}^{n} n_i}{N} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} n_i} = 1$$

Since mortality is strongly associated with age, age-specific mortality rates vary strongly with age. The crude rate does not account for this heterogeneity. This is a significant limit if the aim is to compare rates between different populations or between different periods: part of the observed differences could be due to this heterogeneity and variability between strata.

The goal of the methods that will be presented in the next section is to obtain comparable measures between different populations.

Standardized rates

Since different intrinsic features characterize a population, the comparison of crude rates of different populations and different periods is inappropriate. In fact, each population differs from the other for socio-demographic, geographic, genetic, occupational, dietary, health and environmental aspects. There are significant differences regarding the distribution by age, sex, social class, occupation, etc.

In order to compare mortality rates between different geographical regions, groups or calendar periods it is necessary to consider those factors in the calculation of the rates. It is essential to take into account variables which are already recognized as possible

explanations of observed differences in rates. Among these factors, age plays a key role and its effects are large. As mentioned in the previous paragraph, particularly for mortality rates, the age distribution of the individuals may change between the compared populations. If, for example, older age-groups are dominant in a population, the crude mortality rate will be higher compared to populations where there is a higher proportion of young people and children. This is simply because the risk of death in older people is greater than in the young [1, 5].

Thus, "standardization" procedures are fundamental in order to make rates calculated on different populations comparable. Through standardization, it is possible to control certain known characteristics that can affect the value of the rates, obtaining estimates of weighted rates, based on a reference population, defined as a "standard population". Since the occurrence of many health conditions is related to age, the most common standardization for data concerning public health is standardizing by age.

There are two standardization techniques: direct standardization and indirect one. Below, the direct method is described [1, 6].

Direct standardization

This method aims to determine the annual rate that would be observed in a standard (or reference) population with a given age structure if it was subjected to the same mortality pressure of the studied population.

This procedure calculates the expected number of cases (deaths) in each age-group of the standard population, applying the person-years of the standard population to the corresponding specific estimated rates of the studied population. Then, the total number of expected cases is divided by the total number of person-year in the reference population [1, 5].

The resulting rate indicates the frequency of an event if study population had the same age structure as the reference population.

Consider:

- $\boldsymbol{g}$ – the number of age-groups considered;
- $\boldsymbol{L}$ – the size of **standard population**;
- $\boldsymbol{L_i}$ – the size of the i-th age-group of the **standard population**;
- $\boldsymbol{d_i}$ – the number of cases observed in the i-th age-group of the **population under study**;
- $\boldsymbol{m_i}$ – the number of person-year accumulated in the i-th age-group of the **population under study**;
- $\boldsymbol{\lambda_i} = {}^{d_i}/_{m_i}$ – the specific rate of the i-th age-group of the **population under study**.

The following formula gives the standardized rate with the direct method:

$$Standardized\ rate = \Lambda = \frac{1}{L}\sum_{i=1}^{g} L_i\lambda_i$$

$L_i\lambda_i$ represents the number of expected cases that might be observed in one year in the i-th age-group of the standard population if it was exposed to a level of risk defined by the rate λ_i.

Let:

$\boldsymbol{w_i}$ – be the weight (proportion of subjects) of the i-th group of the **standard population**, equal to L_i/L.

The previous formula may be also written as:

$$\Lambda = \sum_{i=1}^{g} w_i \lambda_i$$

$$\sum_{i=1}^{g} w_i = \sum_{i=1}^{g} \left(\frac{L_i}{L}\right) = \frac{\sum_{i=1}^{g} L_i}{L} = \frac{\sum_{i=1}^{g} L_i}{\sum_{i=1}^{g} L_i} = 1$$

The standardized rate Λ is a weighted average of age-specific rates (λ_i); the weights are the proportion of individuals in the various age-groups of the standard population [5].

Moreover, if two populations are characterized by the same age-specific rates, using the same standard population, the standardized rate will be the same regardless of their age structure.

Reference populations are not necessary real populations, but can also be theoretical ones.

The choice of the standard population depends on the study aim and influences the numerical results. When comparison among rates takes place in countries where age structures are similar to those of developed countries, the European population is suitable as a standard population, while the African population can be used as a reference for developing countries. It is also possible to restrict the standard population – truncated population – to certain age-groups, i.e. adult, when there is a specific interest. One of the most used reference populations is the world standard

population; the WHO provides it in order to make international comparisons easier. Its age structure and corresponding weights are reported below:

Age - group	Standard population 1960 (* 100,000)	Age - group	Standard population 1960 (* 100,000)
0-4	12	45-49	6
5-9	10	50-54	5
10-14	9	55-59	4
15-19	9	60-64	4
20-24	8	65-69	3
25-29	8	70-74	2
30-34	6	75-79	1
35-39	6	80-84	0.5
40-44	6	85 +	0.5
Total			100

Table 1. World standard population age structure (1960 version) [7].

Age - group	Standard population 2001 (* 100,000)	Age - group	Standard population 2001 (* 100,000)
0-4	8.86	45-49	6.04
5-9	8.69	50-54	5.37
10-14	8.60	55-59	4.55
15-19	8.47	60-64	3.72
20-24	8.22	65-69	2.96
25-29	7.93	70-74	2.21
30-34	7.61	75-79	1.52
35-39	7.15	80-84	0.91
40-44	6.59	85 +	0.63
Total			100

Table 2. World standard population age structure based on world average population between 2000-2025 [7].

Rate probability distribution and confidence intervals

The age-specific rates, and consequently the standardized rates, are estimated from observations which are subject to a certain amount of random variability. This variability affects the estimate of the standardized rates and can bring to biased conclusions if the observed differences between standardized rates are mainly due to random variation. In order to evaluate the importance of this kind of variation the standardized rate Λ should be presented with its standard error or its confidence interval [4, 8].

The exact probability distribution for the rate is complicated due to the presence of censored data. For each unit i observed over the time t_i, the events are distributed according to the Poisson probability distribution. Since events have a constant probability over time, the occurrence of an event does not influence the next one and the probability that two events occur at the same instant is zero.

If this reasoning is extended to the whole study, you can assume to have many small Poisson processes. Assuming that all the experimental units are independent (as for mortality data), the study can be thought of as a set of independent Poisson processes: the result will be a Poisson distribution relative to the total number of events [4, 9].

Let:

- $\boldsymbol{D}$ – be the total number of deaths;
- $\boldsymbol{\lambda}$ – be the mortality rate;
- $\boldsymbol{m}$ – be the total observed person time.

$$pr(d = x|D) = \frac{e^{-D} \cdot D^x}{x!} = \frac{e^{-\lambda m}(\lambda m)^x}{x!}$$

$$E(D) = \lambda m$$

$$Var(D) = \lambda m$$

If λ is sufficiently small, and D is much smaller than m, the Poisson distribution is a good approximation of the exact rate probability distribution. Moreover, the variability of the rate is considered to only be associated to its numerator (observed events), while the denominator (the population - time) is considered fixed and therefore not affected by random variability. Therefore, the accuracy of a rate only depends on the variability of the number of observed cases (D) [1].

When events are sufficiently numerous (≥20), the rate probability distribution is approximately Gaussian with mean:

$$\mu = \frac{D}{m}$$

and variance:

$$\sigma^2 = \frac{D}{m^2}$$

Thus, considering this approximation, the variance of the rate estimator $\hat{\lambda} = {}^{D}/_{m}$ is:

$$Var(\hat{\lambda}) = Var\left(\frac{D}{m}\right) = \frac{Var(D)}{m^2} = \frac{\lambda}{m}$$

Its estimate is obtained by replacing λ with ${}^{D}/_{m}$:

$$\widehat{Var}(\hat{\lambda}) = Var\left(\frac{D}{m}\right) = \frac{D}{m^2} = \frac{\hat{\lambda}^2}{D}$$

So, the variance of the specific rate $\hat{\lambda}_i$ is:

$$Var(\hat{\lambda}_i) = \frac{Var(D_i)}{m_i^2} = \frac{\lambda_i}{m_i}$$

For the standardized rate it is:

$$Var(\hat{\Lambda}) = \sum_{i=1}^{g} w_i^2 Var(\hat{\lambda}_i) = \sum_{i=1}^{g} w_i^2 \left(\frac{\lambda_i}{m_i}\right)$$

λ_i being unknown, this variance must be estimated by replacing λ_i by its estimate ${}^{d_i}/_{m_i}$ in the above expression [1]. Then:

$$\widehat{Var}(\hat{\Lambda}) = \sum_{i=1}^{g} \left(\frac{w_i^2}{m_i^2}\right) d_i$$

If the standardized rate is denoted by:

$$\mu = \hat{\Lambda} = \sum_{i} w_i \lambda_i$$

and if s is the estimate of its standard error, then $(\hat{\Lambda} - \lambda_{SD})/s$ approximates a standard normal variable. Therefore, it is possible to calculate the approximate confidence interval (IC) for μ:

$$IC_{1-\alpha}(\Lambda) = \left[\hat{\Lambda} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\Lambda})}\,, \hat{\Lambda} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\Lambda})}\right]$$

Where $1-\alpha$ is the chosen confidence level and $z_{1-\frac{\alpha}{2}}$ is the quantile of level $1-\frac{\alpha}{2}$ of a standard normal distribution - $N(0,1)$.

For practical purposes rates are usually given as $\hat{\Lambda}$ per 100,000 person years ($10^5 \times \hat{\Lambda}$), consequently the variance needs to be presented as $10^{10} \times \widehat{Var}(\hat{\Lambda})$.

Mortality trend analysis

Estimated Annual Percent Change

In order to describe mortality rate trends over time, it is useful to calculate the annual percent change (EAPC).

The assumption is that the rates change constantly from year to year. Rates that change with a constant percentage every year change linearly on a logarithmic scale. For this reason, to estimate the EAPC of a series of data, the following regression model is used:

$$\log(\lambda_x) = b_0 + b_1 x$$

Where $\log(\lambda_x)$ is the natural logarithm of the rate for the year x.

The EAPC between the year x and the year $x + 1$ is equal to:

$$APC_{x,x+1} = \left[\frac{(\lambda_{x+1} - \lambda_x)}{\lambda_x}\right] \times 100 = \left[\frac{\left(e^{b_0+b_1(x+1)} - e^{b_0+b_1 x}\right)}{e^{b_0+b_1 x}}\right] \times 100 = (e^{b_1} - 1) \times 100$$

However, it is unreasonable to describe the pattern of an entire data series in detail with a single EAPC, for this reason, it is useful apply a joinpoint regression model [10]. This model, through statistical criteria, determines when and how often the EAPC changes significantly over the considered period [11].

In mortality rates, the model is estimated through log-linear segments joined to each other. For example, rates can rise mildly for a certain period, have a strong upward trend during next years, to, then, decrease for the remaining time of the study.

Implementing the joinpoint model which best describes the data, is fundamental in order to determine how long the EAPC remains stable and when there is a change in trends.

Joinpoint regression model

The joinpoint regression model, proposed by Kim HJ et al [10], identifies the years characterized by a statistically significant change in mortality rates during the study period. This model is one of the most used and, moreover, is implemented by the Joinpoint software from the National Cancer Institute and freely available [12].
Briefly, the joinpoint regression model assumes that the trend of the logarithm of the rate is linear. A linear segment can approximate a curve quite well, provided that it has the appropriate length. The joinpoint model identifies linear segments that fit the observed rates best, minimizing the sum of the squares of the distances between the points and the segments themselves [13]. The points of statistically significant change in rates are called "joinpoints". The number of segments that make up the trend can't be more than the number of joinpoints arbitrarily set before the analysis. The year in which the joinpoint occurred is the year that identifies a change in trends.

The software implements two different methods to obtain the model estimates: Grid Search and Hudson. The first method considers the observed values as discrete numbers and allows joinpoints to fall precisely on an observation. A better estimate can be obtained refining the grid, changing the program settings on the number of points to be placed between the X values observed in the grid ("Grid Search") to a number greater than 0. In this way, the "Grid Search" method creates a grid of all possible positions in which the so-called joinpoints can fall, as specified in the settings,

and test the minimum sum of squared errors (SSE) for each model at k joinpoints, determining the best estimate. For low numbers of points between the observed values, this method is more efficient from a computational point of view.

Hudson search considers the observed data as continuous and is more computationally intensive.

The Joinpoint program also computes the EAPCs with the 95% confidence intervals [11].

Statistical model

The joinpoint regression for couple of observations (x_i, y_i), for i= 1, …, n, where y_i represents the observed mortality rates at the time x_i, can be written in the following way.

Let:

- $\boldsymbol{k}$ – be the number of unknown joinpoints;
- $\boldsymbol{\tau_k}$ – be the k-th unknown joinpoint.

$$E[y|x] = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_k (x - \tau_k)^+ = \beta_0 + \beta_1 x + \sum_{i=1}^{k} \delta_i \, (x - \tau_i)^+$$

$$(x - \tau_k)^+ = \begin{cases} x - \tau_k & per\ x > \tau_k \\ 0 & otherwise \end{cases}$$

In the literature, many authors studied this kind of non-linear model previously; it has been called in different ways, for example “piecewise regression”, “segmented regression”, “broken line regression” or “multi-phase regression” with a continuity constraint [14].

The parametrized log-linear model:

$$y = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_k (x - \tau_k)^+ + error$$

where $y = \log(rate)$.

In the case of a model without joinpoints the equation reduces to a simple linear model with intercept β_0 and slope β_1. While the terms $\delta_i (x - \tau_i)^+$ represent the change in slope for any subsequent segments and are equal to zero in the years prior to the joinpoints.

For example, to determine up to 2 joinpoints, you must test a null hypothesis of no change with the alternative of 2 joinpoints.

$$\begin{cases} H0: E[y|x] = \beta_0 + \beta_1 x \\ H1: exist\ \tau_1\ and\ \tau_2\ (\tau_1 < \tau_2)\ such\ that\ E[y|x] = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \delta_2 (x - \tau_2)^+ \end{cases}$$

- If the null hypothesis is rejected, the same procedure is applied to a hypothesis test with 1 joinpoint versus 2 joinpoints;

- If the null hypothesis is not rejected, it is tested against the 1 joinpoint alternative hypothesis.

The best model is identified by a permutation test algorithm, which involves comparisons among models with different numbers of joinpoints. The first comparison is between the model without joinpoints and the model with a number of joinpoints

equal to the fixed maximum number. The final model is the model with the fewest parameters, for which the addition of a further parameter (joinpoint) does not lead to significant improvements.

Once the best model is selected the EAPC estimate for each segment is:

$$APC_{(\tau_j,\tau_{j+1})} = \left(e^{b_1+\delta_1+\cdots+\delta_j} - 1\right) \times 100$$

The Joinpoint program provides the EAPCs and the corresponding confidence intervals.

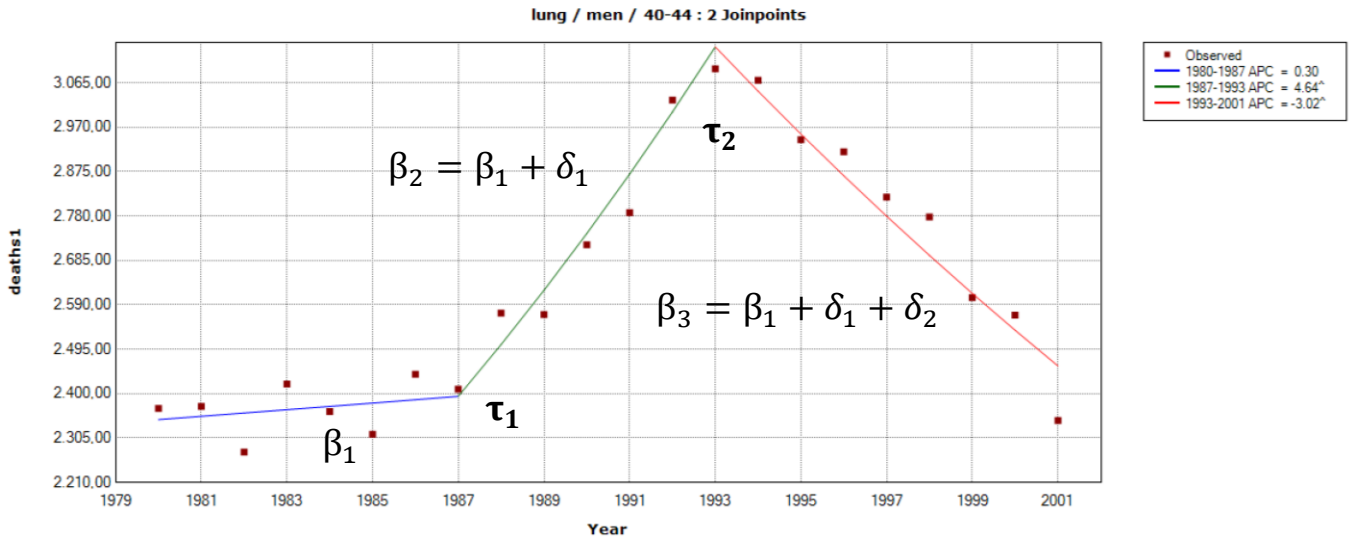

Figure 2. Example of joinpoint output for lung cancer in men aged 40-44 years old.

Predictive analysis

Introduction

The prediction of future trends in incidence and mortality is essential to generate the epidemiological information necessary for resource allocation in health planning.

This report focuses on cancer mortality rates. Cancer is a significant health issue with a huge burden on global population. Health specialists, planners and policy makers need information on the future cancer burden in order to prioritize prevention activities, allocate health services and resources, and evaluate the impact of interventions and treatments [3, 15].

In general, for health planning, which is an integral part of cancer control programs, having information on future trends is a necessity. This is why projection methods are so important and accurate projections of future burden of cancer are essential.

Statistical methods for cancer projections, which are commonly used when information on risk factors is not available, can be implemented in two steps:

1) using historical data to model trends of cancer risk;
2) extrapolating the trends into the future to project the numbers and rates.

Statistical modelling of past trends allows to project cancer incidence and mortality trends by extrapolating time trends from observed rates. The number of new cancer cases or deaths is calculated by applying the estimated rates to projected population numbers. Projections based on the extrapolation of trends in cancer incidence and mortality over time assume that trends in risk behaviour will remain stable, no intervention or screening program will be started, and there is no change in diagnostic techniques. However, this assumption of unchanged trends in rates is very strong and may not be realistic [15].

Trends in cancer incidence or mortality may be described as trends over age at diagnosis or at death, year of diagnosis or death (period), and/or year of birth (cohort). Age is the most important time scale that affects cancer risk; it characterizes the cumulative exposure of the body to carcinogens over time. Period effects correspond to events that change incidence risk regardless of the age-group and are usually due to an environmental change. Cohort effects involve risk factors that a specific generation shared [3, 16]. The trend of observed rates reflects the unobserved trend in cancer risk. Usually, the trend can be classified as (overall or age-specific) period trend and/or cohort trend, which lead to two classes of models: age-period models and age-period-cohort models. In general, the period effects can modify the risk of cancer in both the short- (usually less than 5 years ahead) and the long-term (around 25 years ahead), cohort effects on the risk of cancer are more important for a long period than for a short period. So in general, the short-term projections are based on age-period models, while the long-term projections take cohort effects into account and are based on age-cohort or age-period-cohort models.

Mathematically, trends can be described as linear or non-linear, and different statistical modelling techniques including parametric, semi-parametric and non-parametric models can be used. Because different statistical methods can result in different cancer projections, it may be difficult to determine which method is more appropriate. Thus, appropriate statistical modelling is fundamental to obtain valid cancer projections [15].

The literature proposes many statistical models for cancer projections, each focusing on different issues and aspects. Among these, short-term techniques include Poisson regression methods, those based on ARIMA models (Autoregressive Integrated Moving Average) for time series and Joinpoints. The most used long-term predictions methods rely on age-period-cohort models such as Nordpred [15, 17-21].

The following sections will present different models/projection methods, all based on an age-period model with joinpoints, that are used in an applicative scenario.

Joinpoint regression on the number of deaths

The trend analyses performed with joinpoint regression models can be used to predict mortality trends.

This prediction method proceeds as follows: a joinpoint regression model is fit to the logarithm of the number of age-specific deaths for each 5-year age-group to identify the most recent trend segments (a). Subsequently, a regression model is applied to the mortality data for each age-group over the period identified by the last segment of the joinpoint model, in order to estimate the regression coefficients (b). This model is then used to predict mortality for future years, to calculate the number of expected age-specific deaths and the corresponding 95% prediction intervals (IPs), that is, the confidence intervals for the prediction of each future value. These are calculated with a standard error that takes the variability of the new observation into account (c) [22, 23]. Age-standardized mortality rates, with corresponding 95% PIs, are calculated using the number of expected age-specific deaths and the projected population data for the period of interest (d).

During the statistical analysis for the paper titled “European cancer mortality predictions for the year 2016 with focus on leukaemia”, the authors noted that, for the 0-14 years age-group , the model that best fit the leukaemia data was logarithmic, while for the other age-groups (all ages, 15-44, 45-69) the linear one worked well. Also for

these situations, I decided to implement an algorithm that would combine the four aforementioned models and select the best for each specific age-group automatically.
In order to obtain projections of mortality data, I considered GLM Poisson regression models with four different link functions: the identity, logarithmic, power five and square root link. Furthermore, I implemented an algorithm that blends the above regression models, creating a new "hybrid" regression. This method calculates the number of expected age-specific deaths for each of the previous models; then for each age-group, sex and cancer site, the algorithm chooses the model with the best fit, based on the Akaike information criterion (AIC) statistic values. Thus, the algorithm chooses the best fitting model for each cause of death, sex and age-group, so the resulting total number of deaths and age-standardized rate could be calculated with different underlying link functions. The idea is that this hybrid model, automatically selects the best fitting model for each age-group, hopefully producing more accurate predictive estimates than the others.
AIC statistic was chosen to compare the different model performances. Another useful statistic to compare models is the Bayesian information criterion (BIC).
In addition to the previous models, I also considered a further simple model for projections: after obtaining the predicted estimates from the four GLM Poisson models, I computed the corresponding mean estimates, generating the average model.

The following section describes the main characteristics of the Generalized Linear Models, considering different distributions and link functions. GLM Poisson regressions with identity, logarithmic, power five and square root link functions were used for projections, in the algorithm and in order to obtain the average model, assuming that the number of deaths over the last segment identified by the joinpoint

model followed this distribution and hence were projected according to this class of models.

Exponential Family of Distribution and Generalized Linear Models

Introduction

The generalized linear model (GLM), as defined by Nelder and Wedderburn [24], could be considered as an extension of the classic linear regression model. GLMs aim to expand the classic linear regression to response variables (Y) which have distributions other than the Normal distribution. Since a non-Gaussian distribution is considered, the variance of Y is a function of its mean, thus the hypothesis of homoscedasticity at the basis of the linear regression collapses.

The Bernoulli, Binomial, Poisson and the Negative Binomial are typical examples of random variables on which GLMs perform well. The common characteristic of all previous random variables is that the dependent variable distribution belongs to a wider class of distributions called Exponential Family of Distributions (EF). This class of distributions shares many properties from the Normal distribution.

Exponential Family of Distribution

A single random variable Y has a probability function, if it is discrete, or a probability density function, if continuous, and belongs to the exponential family (EF) if follows the form:

$$f(y;\ \Theta, \phi) = exp\left\{\frac{y\Theta - b(\Theta)}{a(\phi)} + c(y, \phi)\right\}$$

Where:

- Θ is the canonical or natural parameter ($\Theta \in \Theta \subseteq \mathbb{R}$);
- ϕ is the dispersion parameter ($\phi \in \Phi \subseteq \mathbb{R}^{+}$);
- $a(\bullet), b(\bullet), c(\bullet)$ are specific functions that identify a specific distribution function belonging to EF. $a(\bullet)$ depends only on the parameter ϕ; $b(\bullet)$ depends only on Θ; $c(\bullet)$ depends on y and ϕ. Usually, $a(\phi)$ is defined as ϕ/ω, where ω is a known constant ($\omega > 0$).

If ϕ is known, the $f(y; \Theta, \phi)$ is called canonical of natural form of the exponential family.

Since $b(\Theta)$ is twice differentiable in Θ, its first derivative is an invertible function of Θ, and since Θ is a convex set, the expected value of Y is:

$$E(Y) = \mu = \frac{db(\Theta)}{d\Theta} = b'(\Theta)$$

And the variance of Y is:

$$Var(Y) = \frac{d^2 b(\Theta)}{d\Theta^2} a(\phi) = b''(\Theta) a(\phi) = \frac{d\mu}{d\Theta} a(\phi) = V(\mu) a(\phi)$$

where the quantity:

$$\frac{d\mu}{d\Theta} = V(\mu)$$

is called variance function and it expresses the dependence between the variance and the mean of Y. In particular, if $a(\phi) = 1$ the variance overlaps the variance function.

Moreover, under the same regularity conditions considered for $b(\Theta)$, the following relation is true:

$$\frac{d\Theta}{d\mu} = \frac{1}{V(\mu)}$$

Many well-known distributions belong to the exponential family. For example, considering the probability function of a Bernoulli random variable:

$$f(y;\ \pi) = \ \pi^{y}(1 - \ \pi)^{1-y} = (1 - \pi)\left(\frac{\pi}{1-\pi}\right)^{y}$$

Where π represents the success probability.

Applying the transformation:

$$f(y;\ \pi) = exp\{\ln(f(y;\ \pi)\}$$

We obtain:

$$f(y;\ \pi) = \ exp\left\{ln\left[(1 - \pi)\left(\frac{\pi}{1-\pi}\right)^{y}\right]\right\} = exp\left\{y\ ln\left(\frac{\pi}{1-\pi}\right) + \ln(1 - \pi)\right\}$$

In this case the correspondence between parameters is:

$\Theta = \Theta(\pi) = \ ln\left(\frac{\pi}{1-\pi}\right)$ with $\Theta \in (-inf, +inf)$

$a(\phi) = \phi = 1$

$b(\Theta) = ln(1 + e^{\Theta})$

$c(y, \phi) = 1$

The expected value is:

$$E(Y) = \mu = \frac{db(\Theta)}{d\Theta} = \frac{d}{d\Theta}[\ln(1 + e^{\Theta})] = \frac{e^{\Theta}}{1 + e^{\Theta}}$$

The variance (equal to the variance function) is:

$$Var(Y) = V(\mu) = \frac{d\mu}{d\Theta} = \frac{d}{d\Theta}\left[\frac{e^{\Theta}}{1 + e^{\Theta}}\right] = \pi(1 - \pi)$$

Similarly, for the Binomial random variable:

$$f(y; n, \pi) = \binom{n}{y} \pi^{y}(1 - \pi)^{n-y}$$

$$y = 0,1,2, \dots, n \quad \text{and} \quad 0 < \pi < 1$$

Where n is the number of independent sets.

The correspondence among parameters is:

$\Theta = \Theta(\pi) = ln\left(\frac{\pi}{1-\pi}\right)$ con $\Theta \in (-inf, +inf)$

$a(\phi) = \phi = 1$

$b(\Theta) = n\, ln(1 + e^{\Theta})$

$c(y, \phi) = \ln\binom{n}{y}$

The expected value is:

$$E(Y) = \mu = \frac{db(\Theta)}{d\Theta} = \frac{d}{d\Theta}[n \ln(1 + e^{\Theta})] = n\frac{e^{\Theta}}{1 + e^{\Theta}} = n\pi$$

The variance (equal to the variance function) is:

$$Var(Y) = V(\mu) = \frac{d\mu}{d\Theta} = \frac{d}{d\Theta}\left[n\frac{e^{\Theta}}{1 + e^{\Theta}}\right] = n\pi(1 - \pi)$$

Considering the Poisson random variable:

$$f(y; \lambda) = \frac{\lambda^{y} e^{-\lambda}}{y!}$$

$$\lambda > 0$$

Applying the transformation:

$$f(y;\ \lambda) = exp\{\ln(f(y;\ \lambda)\} = exp\{y\ ln\lambda - \lambda - ln(y!)\}$$

We obtained the following correspondence:

$\Theta = \Theta(\lambda) =\ ln\,\lambda$ with $\Theta \in (-inf, +inf)$

$a(\phi) = \phi = 1$

$b(\Theta) = e^{\Theta}$

$c(y, \phi) = -\ln(y!)$

The expected value is:

$$E(Y) = \mu = \frac{db(\Theta)}{d\Theta} = \frac{d}{d\Theta}[e^{\Theta}] = e^{\Theta} = \lambda$$

The variance (equal to the variance function) is:

$$Var(Y) = V(\mu) = \frac{d\mu}{d\Theta} = \frac{d}{d\Theta}[e^{\Theta}] = e^{\Theta} = \lambda$$

Generalized Linear Models

Consider Y a dependent random variable belonging to the Exponential Family of Distributions. The GLM models are specified by three components:

1) the random component, which regards the dependent variable and its distribution;
2) the systematic component, which regards the linear predictor (η=$x^t\beta$) that is the set of covariates (qualitative, quantitative or both). Usually the maximum likelihood method is used for parameter estimation;
3) the link function (g), which specifies the relation between the expected value of Y and the systematic component. More precisely, the link is the function of the *E(Y)* which, in the model, will be equal to the linear predictor. It establishes a relation between the random component and the systematic one:

$$g(\mu) = \eta = x^t\beta$$

The link function must be monotonous and differentiable, thus an inverse link function (g^{-1}) exists such that:

$$\mu = E(Y) = g^{-1}(\eta) = g^{-1}(x^t\beta)$$

The link function choice and the Y distribution choice are independent. Given a specific Y distribution and a set of covariates, it is possible to define different link functions, and consequently different GLMs.

However, some matches between the link function and the distribution of Y have particular properties. One of the most frequent link choice is the canonical link ($\Theta = \eta$), which connect the canonical parameter and the linear predictor linearly:

$$g(\mu) = \Theta(\mu) = \Theta = \eta = x^t\beta$$

In the previous formula, the canonical parameter depends from the mean μ and:

$$\mu = E(Y) = \Theta^{-1}(\eta) = \Theta^{-1}(x^t\beta)$$

The identity link ($\mu = \eta = x^t\beta$) is a particular link function that specifies a linear model.

In the case of a dependent variable with a Poisson distribution ($Y \sim Poi(\lambda)$), the canonical parameter is:

$$\Theta = \Theta(\lambda) = \ln(\lambda)$$

$$\text{and } \lambda = E(Y) = \mu$$

Then, the GLM model using the canonical link is the following:

$$g(\lambda) = \Theta(\lambda) = \ln\lambda = \eta = x^t\beta$$

The previous is called log-linear Poisson model and it is usually used for count data response variables.

Another possible link function is the probit link:

$$\eta = \phi^{-1}[E(Y)]$$

where ϕ is the standardized Normal distribution function.

In addition to the identity and logarithmic link, for the present report I also considered the power five link:

$$g(\mu) = \Theta(\mu) = \Theta^{1/5} = \eta = x^t\beta$$

and the square root link:

$$g(\mu) = \Theta(\mu) = \Theta^{2} = \eta = x^t\beta$$

The link function results in a linear transformation on the population averages and not on the values of the dependent variable. Unlike methods of transformation, the link function takes advantage from the source distribution of the response variable, allowing the results to be expressed in the source scale.

It is crucial to identify the link function that can better interpret the response variable and its relation to the set of explanatory variables. Indeed, if the GLM model is not correctly specified, the results distribution and any inference drawn from them are not valid.

Comparison tests

To measure the accuracy of the predicted figures and to compare the performance of the different projection methods, the prediction error is estimated by computing the predicted minus the observed values in absolute terms for every projected year. Then the ratio between the prediction error and the observed count is calculated (when the observed count is zero, it is necessary to add 0.5 to the denominator to avoid numerical errors). The ratio is the percentage error of the prediction, compared to the observed count [25, 26]:

$$error\ ratio = \frac{|estimated\ count - observed\ count|}{observed\ count\ (+0.5 \text{ when denominator is } 0)}$$

Finally, the average absolute relative deviation (AARD) is computed as the average of these error ratios:

$$AARD = \frac{1}{N}\sum_{i=1}^{N}\frac{|estimated\ count_i - observed\ count_i|}{observed\ count_i + 0.5}$$

where i =1 … N indicates a specific scenario.

The AARD indicates the average percent deviation from the true value (number of observed deaths) relative to the true value. This measure attempts to take the relative differences in observed mortality counts and assess the extent to which the estimates deviate from the observed.

Smaller values of AARD indicate the predicted estimates are close to the observed value. In general a prediction is considered reliable when the AARD value is less than 5% [18].

Similar considerations of the number of deaths apply to rates.

In particular in this thesis AARDs are computed for the six projection methods considered (Poisson GLM model with identity, logarithmic, power five and square root link, the hybrid regression and the average one); more specifically for each combination among cause of death, sex and projection methods, and also for categories of the mortality counts, from the rarest causes of deaths to the most common [25, 26].

Application to real data

In this section the previously described projection methods are applied to real mortality data.

Data and methods

I obtained official cancer death certification data from 1980 to 2011 from the World Health Organization (WHO) database, available on electronic support (WHOSIS) [27]. Figures were derived for 22 countries worldwide, including the European Union (EU) as a whole (28 countries as defined in July 2013, minus Cyprus due to data unavailability), and for 25 major causes of death (23 cancer sites and 2 cardiovascular diseases). I only selected countries with over 5 million inhabitants and with data coverage above 90% [28].

Mortality data was coded according to the ICD - International Classification of Diseases, developed by the WHO. During the calendar period considered, three different Revisions of the International Classification of Diseases were used, the eighth (ICD-8), the ninth (ICD-9) and the tenth revision (ICD-10) [29, 30, 31]. Since coding differences between various revisions were generally minor, all cancer deaths were recoded according to the tenth Revision of the ICD.

From the WHO database, I obtained estimates of the resident population for the corresponding calendar periods, based on official censuses. When population data were missing for some European countries, they were derived from Eurostat [32]. For the USA I retrieved population estimates from the Pan American Health Organization database (PAHO) [33].

I split the data: I used observed data from 1980 to 2001 as a training dataset, to which I applied the projection methods in order to predict data for the 2002-2011 period, and I used observed data from 2002 to 2011 as a validation dataset, to compare the different methods results.

From the matrices of certified deaths and resident population, I computed age-specific observed rates for each 5-year age-group (from 0–4 to 80+ or 85+ years) and calendar year or quinquennium. I then computed age-standardized mortality rates per 100,000 using the direct method on the basis of the world standard population. For the calculation of the EU rates, if data was missing for one or more calendar years within a country, I performed extrapolations using the nearest available data.

Projections were derived by fitting a joinpoint model to the number of certified deaths in each 5-year age-group in order to identify the most recent trend period. Subsequently, Poisson GLM regressions with identity, logarithmic, power five and square root link functions and the hybrid regression (a combination of the previous four models), were applied to the mortality data in each age-group over the time period identified by the joinpoint model. I thus computed the predicted age-specific number of deaths, and the 95% prediction intervals (PIs) using the previously obtained regression coefficients and simulated standard error. Predicted standardized mortality rates, and their 95% PIs, were computed using the 2002-2011 populations. In this specific case, I applied the joinpoint regression model to the certified numbers of deaths over the 1980-2001 period, with the following constraints:

- the number of available years following the last estimated joinpoint must be at least equal to 5;
- the number of years between two subsequent joinpoints must be at least equal to 4;
- the maximum number of joinpoints, decided before the analysis, is 5.

These constraints were set in order to consider more stable periods; however, it is possible that if there were significant changes in slope in more recent periods, the joinpoint couldn't detect them. The last constraint, in particular, was set because I was working on the number of deaths, which has more fluctuation than rates.
After obtaining predicted rates and the PIs for the models under study, I computed the average model, obtaining the corresponding average estimates.
I then compared the performance of the different projection methods using the AARD score.

The datasets submitted to the Joinpoint program were created with SAS 9.4 software, while for the projections, including the implementation of the hybrid model algorithm, I used R 3.2.3 software.

Hybrid model

In this paragraph, I explain the inner workings of the predictive algorithm.

```
###Poisson - link log
modPlog<-function(dati,predyears,n=100){
  n_anni<-nrow(dati)
  media_anni<-mean(dati$Year)
  dev_anni<-sum((dati$Year-media_anni)^2)
  new<-predyears
  data$deaths_S<-ifelse(data$deaths1<=0,0.5,as.integer(round(data$deaths1,0)))

  Poislog<-glm(deaths_S~Year, data=dati,  family=poisson(link=log))

  prednew<-predict(Poislog,new,int="pred",se=T, type="response")
  predetti<-predict(Poislog,int="pred",se=T, type="response")
  pred<-ifelse(prednew$fit<0,0,prednew$fit)
  fittati<-predetti$fit

  ##### Bootstrap methods to estimate Prediction Inverval for Poisson Count Regression (requires plyr library)

  bootSimFun<-function(preddata,fit,data){
    bdat<-data[sample(seq(nrow(data)),size=nrow(data),replace=TRUE),]
    bfit<-update(fit,data=bdat)
    bpred<-predict(bfit,type="response",newdata=preddata)
    rpois(length(bpred),lambda=bpred)
  }
  #simulated bootstrap distribution of projections
  simvals<-raply(n,bootSimFun(preddata=new,fit=Poislog,data=dati))
  #selection of percentiles for IP
  int_pred<-t(apply(simvals,2,quantile,c(0.025,0.975),na.rm=TRUE))

  #IC number of the deaths
  deaths_pred_poi<-ifelse(prednew$fit<0,0,prednew$fit)
  low_conf_deaths_poi<-ifelse((prednew$fit-1.96*prednew$se.fit)<0,0,(prednew$fit-1.96*prednew$se.fit))
  upp_conf_deaths_poi<-ifelse((prednew$fit+1.96*prednew$se.fit)<0,0,(prednew$fit+1.96*prednew$se.fit))
  se_c_poi_deaths<-(upp_conf_deaths_poi-prednew$fit)/1.96

  #IP number of the deaths
  low_prev_deaths_poi<-ifelse(int_pred[,1]<0,0,int_pred[,1])
  upp_prev_deaths_poi<-ifelse(int_pred[,2]<0,0,int_pred[,2])
  se_p_poi_deaths<-apply(simvals,2,sd,na.rm=TRUE)
```

Figure 3. A piece of the R program - “modPlog” function for the Poisson GLM logarithmic link function.

In R, I constructed a function for each of the four Poisson GLM regressions considered. This function requires as arguments: a dataset containing the population estimates, the certified deaths, the age-standardized observed rates by cause of death, sex, year, age-group (Figure 4) and a dataset only containing the years for the predictions.

	canctab	Sex	Year	X_LABEL_	POP1	age	deaths1	rate	yearmax	sgstart
1	1	1	1998	deaths at age 0-4 years	13315860	1	7.0	0.05256889	2001	1998
2	1	1	1999	deaths at age 0-4 years	13162899	1	3.0	0.02279133	2001	1998
3	1	1	2000	deaths at age 0-4 years	13086879	1	4.0	0.03056497	2001	1998
4	1	1	2001	deaths at age 0-4 years	12989647	1	2.0	0.01539688	2001	1998
5	1	1	1980	deaths at age 5-9 years	16556390	2	8.2	0.04952770	2001	1980
6	1	1	1981	deaths at age 5-9 years	16219993	2	7.2	0.04438966	2001	1980
7	1	1	1982	deaths at age 5-9 years	15909589	2	8.2	0.05154124	2001	1980
8	1	1	1983	deaths at age 5-9 years	15693259	2	8.2	0.05225173	2001	1980
9	1	1	1984	deaths at age 5-9 years	15541581	2	7.2	0.04632733	2001	1980
10	1	1	1985	deaths at age 5-9 years	15419952	2	6.0	0.03891063	2001	1980
11	1	1	1986	deaths at age 5-9 years	15539653	2	5.0	0.03217575	2001	1980
12	1	1	1987	deaths at age 5-9 years	15273731	2	6.0	0.03928313	2001	1980
13	1	1	1988	deaths at age 5-9 years	15192943	2	2.0	0.01316401	2001	1980

Showing 1 to 13 of 16,074 entries

Figure 4. Example of dataset required in the Poisson GLM function (EU data).

I used the "glm" function in R with a Poisson family, with the corresponding link functions.

Then, I constructed a loop that, for each combination of cause of death, sex and age, submits the four GLM Poisson functions and compares the AIC values (Figure 5).

```
for (c in c(1,2,3,4,5,6,7,8,9,10,12,13,14,15,22,23,24,25,26,27,28,29,30,31,100)){
  for (s in (1:2)){
    for (a in c(1:18)){

      if (c==1 & s==1 & a==1){

        lin_mod<-plin[[as.character(c),as.character(s),as.character(a)]]$modello
        log_mod<-plog[[as.character(c),as.character(s),as.character(a)]]$modello
        sqrt_mod<-psqrt[[as.character(c),as.character(s),as.character(a)]]$modello
        power5_mod<-ppower5[[as.character(c),as.character(s),as.character(a)]]$modello

        aic<-AIC(lin_mod, log_mod, sqrt_mod, power5_mod)
        aic$df<-NULL
        aic<-data.frame(t(aic))

        best<-which.min(c(aic$lin_mod, aic$log_mod, aic$sqrt_mod, aic$power5_mod))

        { if (best==1) {aic$modello_fin<-"lineare"}
          else if(best==2) {aic$modello_fin<-"logaritmico"}
          else if(best==3) {aic$modello_fin<-"sqrt"}
          else if(best==4) {aic$modello_fin<-"power5"}
```

Figure 5. A piece of the R program, the algorithm for predictions.

On the basis of the AIC values, the program choses which of the previous models to apply to each specific time series defined by cause of death, sex and age (i.e. 5 year age-group) variable combinations. For each combination, I obtained parameter estimates with their accuracy measures (Figure 6).

Year	fit	lwr	upr	se_fit	SE_pred	Model	se_pred	se_conf	age	canctab	Sex
2002	1.51185789	1.506852e-01	15.168808	1.5145619	3.514971	log	3.174077	1.5145619	1	1	1
2003	1.06852437	6.463675e-02	17.664012	1.7453415	4.276147	log	3.857036	1.7453415	1	1	1
2004	0.75519289	2.629896e-02	21.685889	2.0195201	5.117875	log	4.864603	2.0195201	1	1	1
2005	0.53374197	1.037694e-02	27.453239	2.3414153	6.006385	log	6.256488	2.3414153	1	1	1
2006	0.37722877	4.017870e-03	35.417164	2.7176195	6.923690	log	8.143798	2.7176195	1	1	1
2007	0.26661113	1.536701e-03	46.255913	3.1564010	7.859714	log	10.688593	3.1564010	1	1	1
2008	0.18843073	5.828396e-04	60.919227	3.6676481	8.808492	log	14.114733	3.6676481	1	1	1
2009	0.13317576	2.197526e-04	80.707956	4.2630017	9.766308	log	18.726768	4.2630017	1	1	1
2010	0.09412363	8.249613e-05	107.389977	4.9560809	10.730741	log	24.937140	4.9560809	1	1	1
2011	0.06652304	3.086838e-05	143.360788	5.7627767	11.700155	log	33.303702	5.7627767	1	1	1
2002	3.57294927	-3.289957e-01	7.474894	0.7548254	1.870807	Square Root	1.870572	0.7548254	2	1	1
2003	3.40571513	-5.400702e-01	7.351501	0.8054999	1.891907	Square Root	1.891589	0.8054999	2	1	1
2004	3.23852274	-7.541942e-01	7.231240	0.8570024	1.914502	Square Root	1.914088	0.8570024	2	1	1
2005	3.07137206	-9.712574e-01	7.114002	0.9091895	1.938542	Square Root	1.938015	0.9091895	2	1	1
2006	2.90426306	-1.191148e+00	6.999674	0.9619471	1.963972	Square Root	1.963319	0.9619471	2	1	1
2007	2.73719570	-1.413755e+00	6.888146	1.0151839	1.990740	Square Root	1.989944	1.0151839	2	1	1
2008	2.57016996	-1.638965e+00	6.779305	1.0688261	2.018791	Square Root	2.017837	1.0688261	2	1	1
2009	2.40318581	-1.866669e+00	6.673041	1.1228134	2.048074	Square Root	2.046946	1.1228134	2	1	1

Figure 6. Dataset obtained after running the algorithm.

Moreover, in the resulting dataset there is “model”, a variable that indicates which model, among the four considered, fits the data better in a specific cause of death, sex, and age-group combination. The previous figures show that for the predicted year 2002, for the cause of death 1 (corresponding to the “Oral cavity and Pharynx” cause), for men aged 0-4 years the logarithmic link function model fit better according to the AIC statistic. Instead, in the 5-9 age-group, in the same cause of death and sex, the square root link function regression had a lower AIC.

This new dataset was used to calculate the rates with their accuracy measures; in the end, we sum the projections by cancer site, and sex over all the age-groups.

The procedure for the Poisson GLM regressions with the four different link functions is similar, with the difference that each combination of age, cause of death, and sex, had the same model.
I repeated the whole process for every country, sex and cause of death considered.

Predictive analysis results

Below, I will show and comment the most significant results obtained from the comparison of the six different projection methods (other specific comparisons are available in the supplementary material). Firstly, I compare results from specific cancer site and country in order to make the comprehension of more general tables easier. Then, I will analyse the more general results in detail, without country distinction.

When I considered specific causes of death and specific countries, the resulting best prediction models are different. In order to make comparisons, I selected results from lung cancer, one of the major cause of deaths, for the EU, the USA and Japan. Moreover, I distinguished between 10-years projection (Table 3a) and 5-years projection, i.e. up to 2006 (Table 3b). From these specific tables, 5-years prediction are more precise and produced lower AARDs. Nevertheless, AARDs from 10- years projection are, most of the times, lower or around 5%, indicating good predicted estimates.

Table 3a. AARDs on rate by projection method, and sex in the EU, in the USA and in Japan (projections up to 2011).

	AARD											
	Men						Women					
	Hybrid	Identity	Log	Average	Power5	Square Root	Hybrid	Identity	Log	Average	Power5	Square Root
EU	0.00320	0.01113	0.01241	0.00305	0.00725	**0.00276**	0.04161	0.07846	**0.03301**	0.05394	0.04485	0.05944
USA	0.05465	0.06605	**0.05091**	0.05721	0.05373	0.05814	0.05024	0.05492	**0.04007**	0.04394	0.04129	0.04514
Japan	0.02231	0.05043	0.03335	**0.01575**	0.01810	0.01773	0.06243	**0.02182**	0.10163	0.06360	0.08147	0.05447

Table 3b. AARDs on rate by projection method, and sex in the EU, in the USA and in Japan (projections up to 2006).

	AARD											
	Men						Women					
	Hybrid	Identity	Log	Average	Power5	Square Root	Hybrid	Identity	Log	Average	Power5	Square Root
EU	**0.00249**	0.00339	0.00733	0.00326	0.00537	0.00281	0.02600	0.04384	**0.02158**	0.03147	0.02673	0.03371
USA	0.06116	0.06649	0.**05956**	0.06246	0.06088	0.06290	0.06367	0.06578	**0.05626**	0.06042	0.05834	0.06128
Japan	0.00853	0.03739	**0.00800**	0.01623	0.01001	0.01956	0.04575	**0.02606**	0.07281	0.05108	0.06239	0.04735

The EU

Considering the whole projection period, in the EU, the lowest AARD for men was that of square root link function model with an AARD (computed on rate) of 0.00276. However, also the hybrid and the average model worked very well (AARDs around 0.003). At a graphical level (Figure 7a, d, f), observing these three models, it is possible notice that the continuous black line (the predicted trend) overlaps the points of the observed rates perfectly. Moreover, it seems that, in the final hybrid model, the square root and power five link functions are predominant. The square root and power five link figures (Figure 7e, f) in men are very similar to the hybrid one, even if the PIs, in particular for the square root, are slightly closer. Instead, the Poisson GLM identity and logarithmic link function model produced worse predictions with very similar AARD of about 0.012.

In EU women, over the 10-years period, the logarithmic link function predicted trend produced the lowest AARD, with a value of 0.03301. The predicted trend remained closer to the real one, but they did not overlap, however, the real trend is at least included in the PIs (Figure 7c). The hybrid model (AARD of 0.04161) was mainly influenced by the logarithmic link and by the power five link, this latter showed an AARD of 0.04485. Also in women, the identity link function produced the worst AARD (0.07846) and in Figure 7b it is possible to see that the predicted estimates underestimated the real trend. The square root link function model and the average model had AARDs of about 0.059 and 0.054 respectively, showing non-satisfactory predicted trends and PIs (Figure 7d, f).

For this tumour and this country, the prediction intervals in men tended to contain the true rates, while in women the PIs are often at the limit. The AARDs for this cancer site are equal or lower than 5% (with the exception of identity link in women), thus the projections are quite good. Overall, in women the AARDs are higher than those of men.

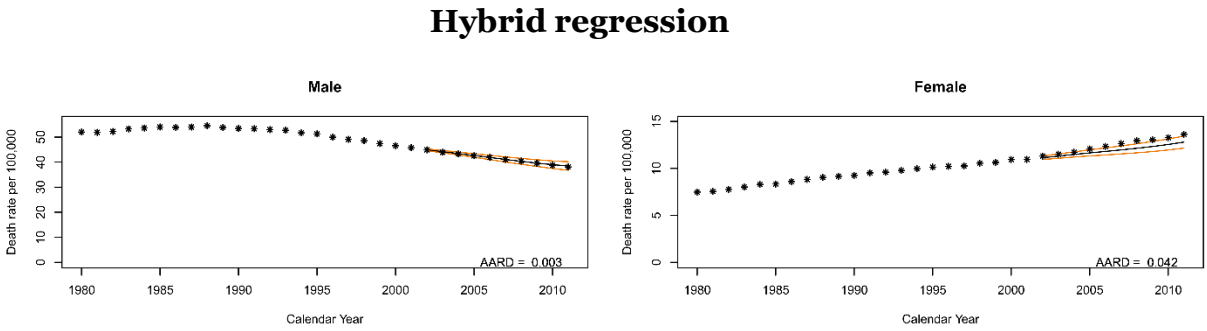

Figure 7a. Projected lung cancer trends with the hybrid regression in men and women from the EU.

Poisson GLM – identity link

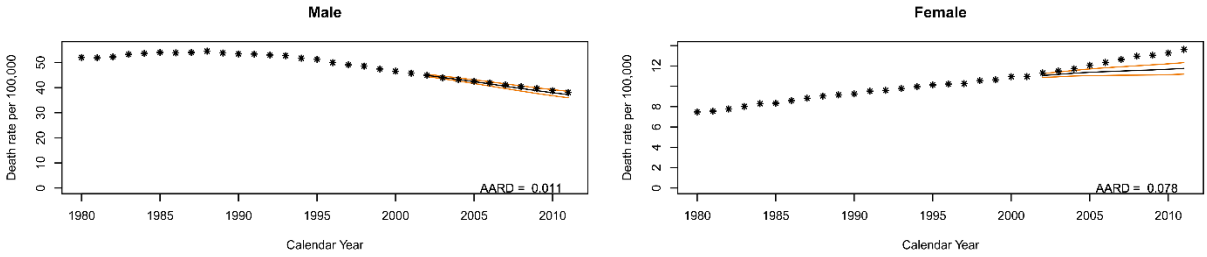

Figure 7b. Projected lung cancer trends with the Poisson GLM identity link function in men and women from the EU.

Poisson GLM – logarithmic link

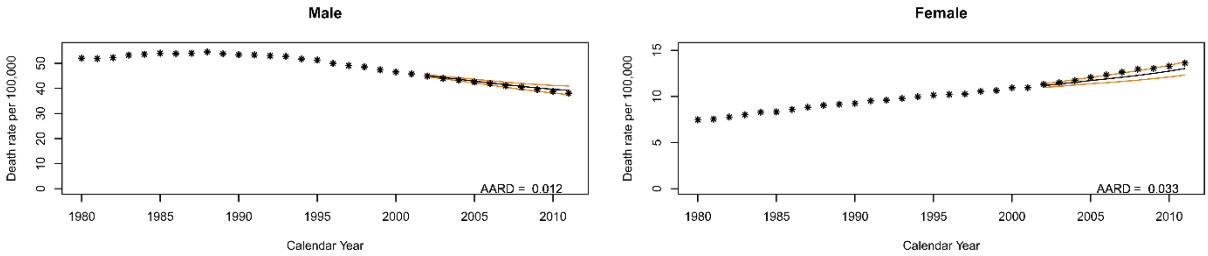

Figure 7c. Projected lung cancer trends with the Poisson GLM logarithmic link function in men and women from the EU.

Average

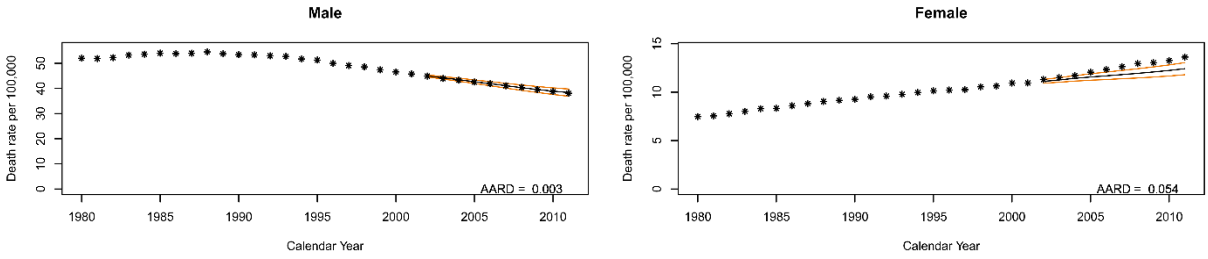

Figure 7d. Projected lung cancer trends with the Average model in men and women from the EU.

Poisson GLM – power five link

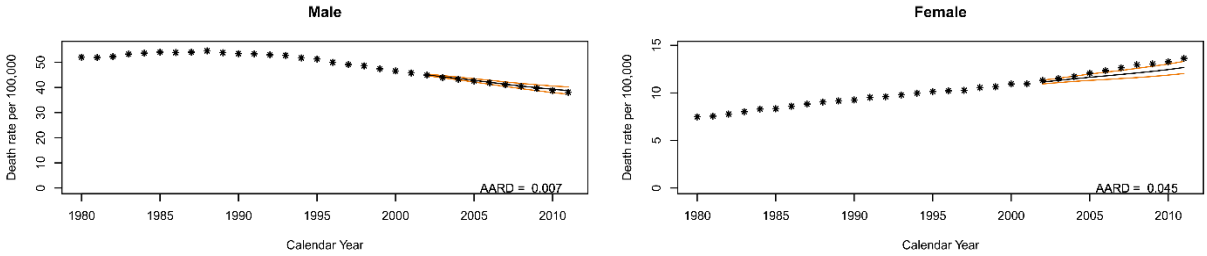

Figure 7e. Projected lung cancer trends with the Poisson GLM power five link function in men and women from the EU.

Poisson GLM – square root link

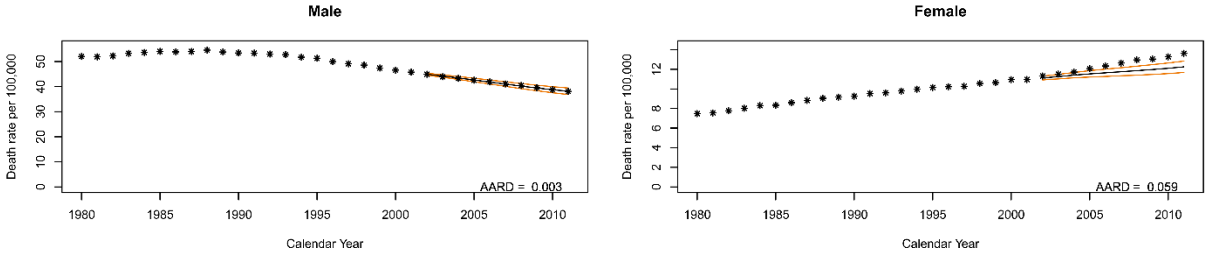

Figure 7f. Projected lung cancer trends with the Poisson GLM square root link function in men and women from the EU.

The USA

In the USA, the 10-years predicted trends derived from Poisson GLM logarithmic link regression models are better than the other projection methods in both sexes, with AARDs of 0.05091 and 0.04007, in men and women respectively. In Figure 8c, it is possible to notice that the predicted estimates from this model followed the real rates quite well.

The highest AARDs, in both sexes were for the identity link function model with values of 0.06605 in men and 0.05492 in women.

In men, the hybrid model, had an AARD value of about 0.055, very similar to the power five AARD and lower than the square root one (Figure 8a, e, f).

In women, the hybrid regression was strongly influenced by the identity one, indeed the corresponding AARD is the second highest (0.05024). Instead AARDs for the average model, the power five and the square root are lower than 5%, the predicted trends are quite close to the real rate and the PIs are at the limits (Figure 8d, e, f).

Overall, in women the AARDs are lower than in men.

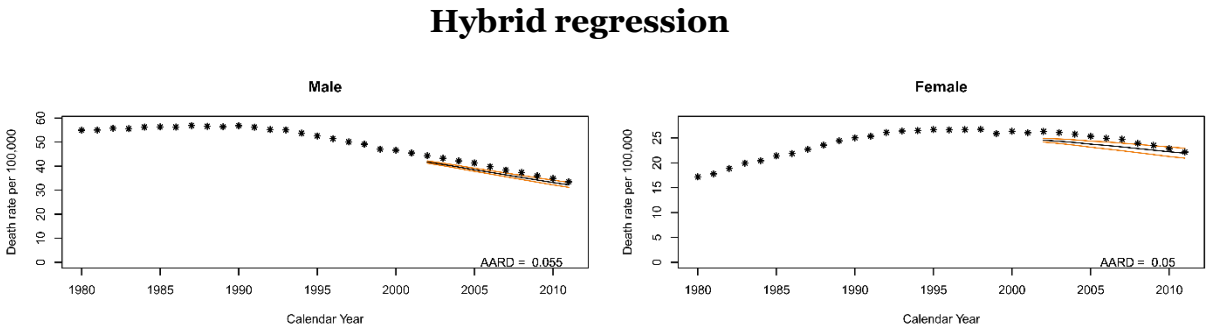

Figure 8a. Projected lung cancer trends with the hybrid regression in men and women from the USA.

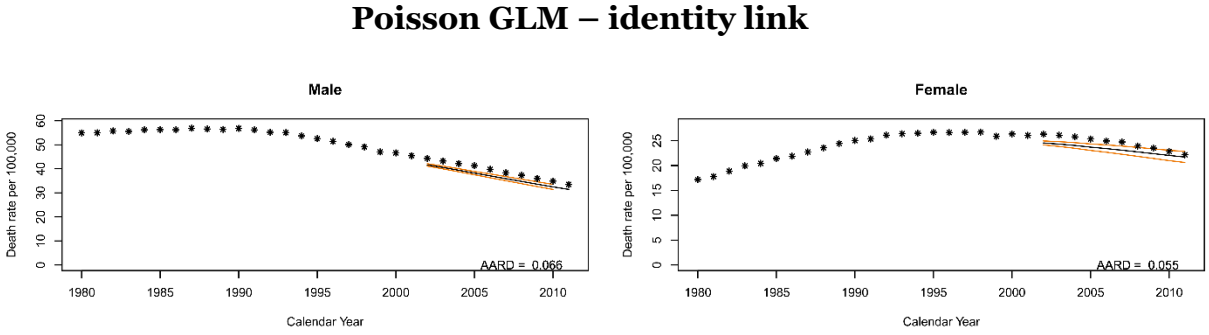

Figure 8b. Projected lung cancer trends with the Poisson GLM identity link function in men and women from the USA.

Poisson GLM – logarithmic link

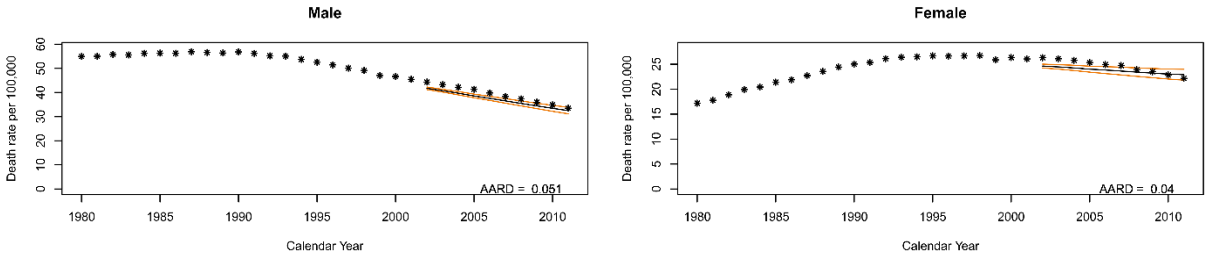

Figure 8c. Projected lung cancer trends with the Poisson GLM logarithmic link function in men and women from the USA.

Average

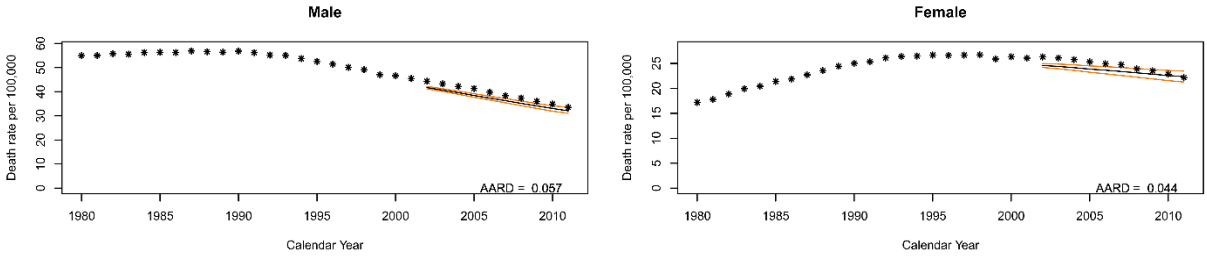

Figure 8d. Projected lung cancer trends with the Average model in men and women from the USA.

Poisson GLM – power five link

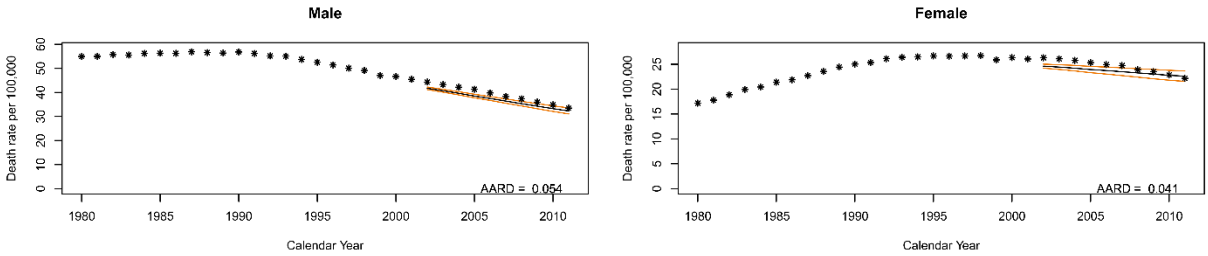

Figure 8e. Projected lung cancer trends with the Poisson GLM power five link function in men and women from the USA.

Poisson GLM – square root link

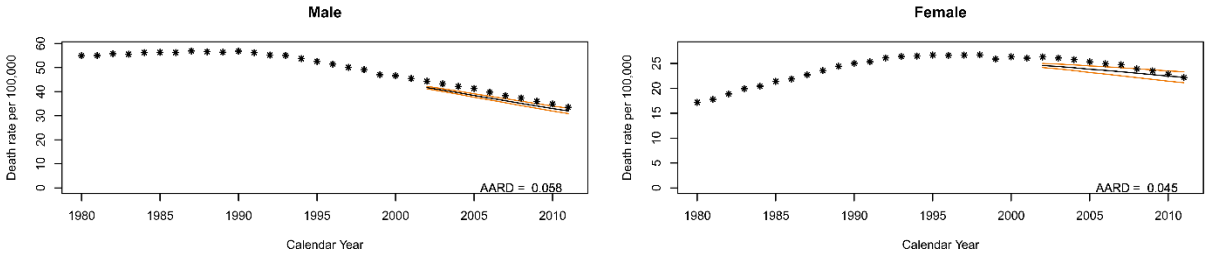

Figure 8f. Projected lung cancer trends with the Poisson GLM square root link function in men and women from the USA.

Japan

In males from Japan, the best predictive method over the period 2002-2011 is the average model, with an AARD of 0.01575. The predicted trend overlaps the observed rates very well (Figure 9d); similarly, the Poisson GLM power five and square root link function regression methods work well with AARD values of around 0.018. The hybrid model was influenced negatively by the logarithmic link function, AARDs of 0.02231 and 0.03335 respectively; the predicted trends from these models tended to overestimate the real data (Figure 9a, c), however, the data is within the PI limits. The identity link function regression produced the worst AARD, indeed the predicted trend substantially did not overlaps the real data, underestimating them (Figure 9b).

In women, the more performant projection method is the Poisson GLM identity link function regression (AARD of 0.02182). As it is possible to see from Figure 9f, also the square root link function produced quite good predicted estimates (AARD of 0.05447). The hybrid model predicted slightly better than the average one with AARDs around 0.06. The worst projections were from the logarithmic and power five link function; for these latter models the PIs did not include the real data (Figure 9c, e).

Hybrid regression

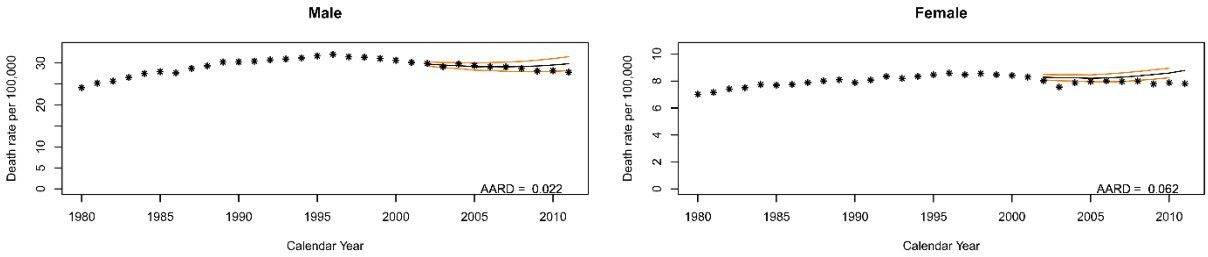

Figure 9a. Projected lung cancer trends with the hybrid regression in men and women from Japan.

Poisson GLM – identity link

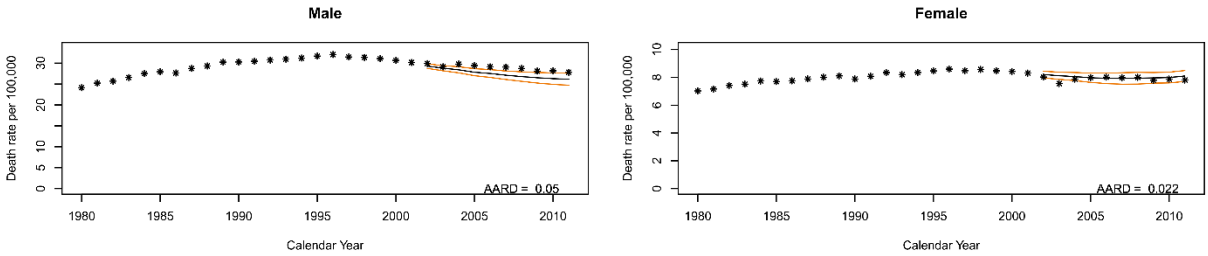

Figure 9b. Projected lung cancer trends with the Poisson GLM identity link function in men and women from Japan.

Poisson GLM – logarithmic link

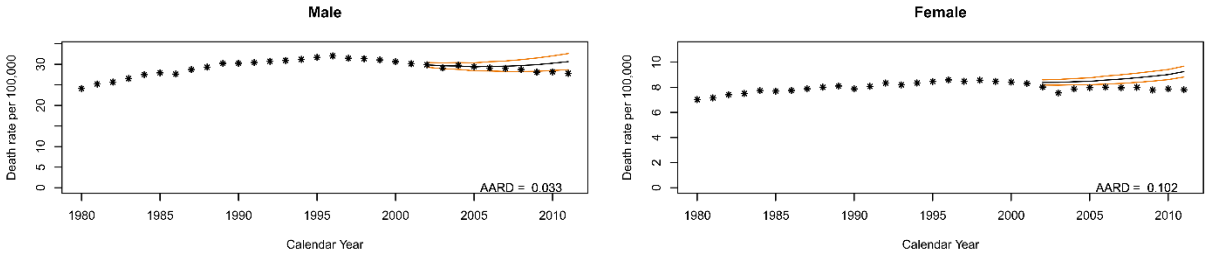

Figure 9c. Projected lung cancer trends with the Poisson GLM logarithmic link function in men and women from Japan.

Average

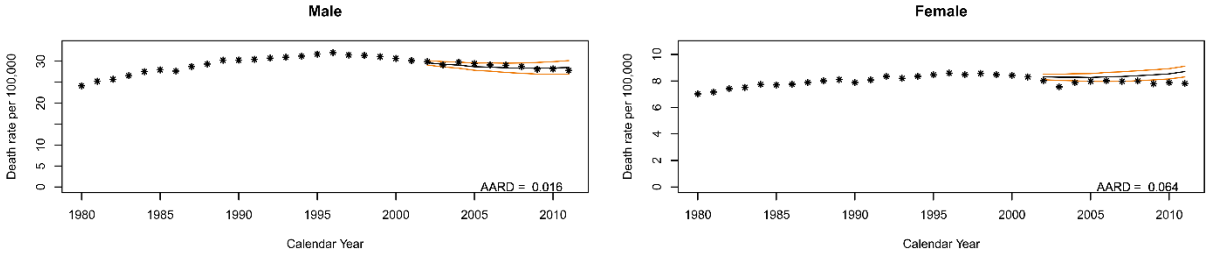

Figure 9d. Projected lung cancer trends with the Average model in men and women from Japan.

Poisson GLM – power five link

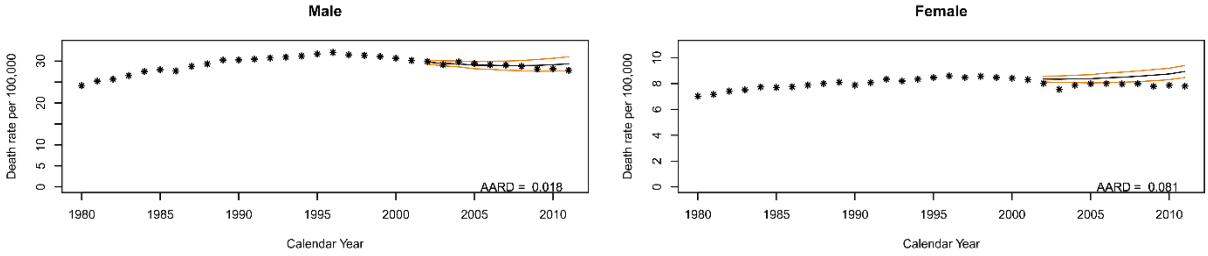

Figure 9e. Projected lung cancer trends with the Poisson GLM power five link function in men and women from Japan.

Poisson GLM – square root link

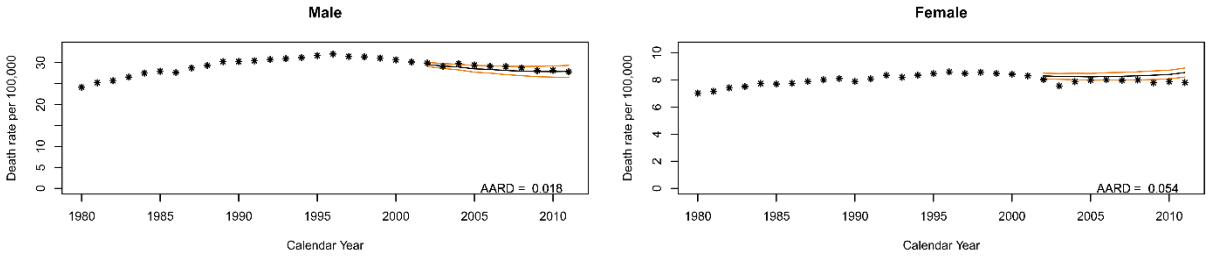

Figure 9f. Projected lung cancer trends with the Poisson GLM square root link function in men and women from Japan.

In general, as can be seen in the various figures, the predicted estimates are more accurate when considering a 5-year period. The predicted line overlaps the observed rates more precisely and the PIs completely include the real trends more often.

Comprehensive analysis

Considering the whole database (all countries, sex and causes of death), the AARD estimates are worse, probably due to the high data variability. Despite this, I obtained more consistent results than those for single country. Analysing the total AARDs computed on number of deaths over the entire period 2002-2011, the Poisson GLM regression with identity link function shows the lowest AARD value (0.16674), while the logarithmic link function has the highest AARD value (0.99326). The average regression has intermediate value (Table 4a).

Table 4a. AARDs on number of deaths by projection method.

AARD					
Hybrid	Identity	Log	Average	Power5	Square Root
0.92669	**0.16674**	0.99326	0.38447	0.21846	0.18069

AARDs from 5-years projection are clearly lower (Table 4b), indicating definitely better predicted estimates. However, the best performance remains that from the identity link function method. The greater improvements passing from the AARDs computed on 10-years projection to those computer on 5-year projection are for the hybrid model and for the logarithmic link function method, that are also the worse ones.

Table 4b. AARDs on number of deaths by projection method (projections up to 2006).

AARD					
Hybrid	Identity	Log	Average	Power5	Square Root
0.15823	**0.12932**	0.17311	0.14416	0.14652	0.13479

Compared to the total AARDs computed on number of deaths, the ranking among AARDs computed on rates is the same, for both projections period (Table 5a and 5b). Over the whole period 2002-2011, the Poisson GLM regression with identity link function shows the lowest AARD value (0.22997), while the logarithmic link function has the highest AARD value (1.73099).

Table 5a. AARDs on rate by projection method.

AARD					
Hybrid	Identity	Log	Average	Power5	Square Root
1.64702	**0.22997**	1.73099	0.62624	0.30994	0.25376

Also for 5-years projections, the identity link function method shows the best predicted estimates with an AARD value of 0.19064. The hybrid and the logarithmic AARDs strongly decrease considering a short projection period.

Table 5b. AARDs on rate by projection method (projections up to 2006).

AARD					
Hybrid	Identity	Log	Average	Power5	Square Root
0.23828	**0.19064**	0.25712	0.21487	0.21881	0.20229

The following tables show the AARDs by projection method, sex and cause of death computed on numbers of deaths (Table 6) and on rates (Table 7) focusing on five-year projections; there is no by country distinction. In both tables and in both sexes, the identity link function most frequently presented lower AARDs as compared to the other link function models. For AARDs computed on numbers of deaths (Table 3), the square root link function model follows the identity link as the second best model for prediction for men, while in women the second best model was the hybrid. In men, the

power five model, and in women, the logarithmic and the average one were never the best model for any causes of deaths considered. Regarding AARDs calculated on rates (Table 7), in women, the logarithmic link function model, the average prediction method and the power five link were never the best.

From these results it would seem that none of these methods are appropriate, since the best result shows AARDs over 10% for deaths and around 20% for rates.

Table 6. AARDs on number of deaths by projection method, sex and cause of death (projections up to 2006).

	AARD											
	Men						Women					
	Hybrid	Identity	Log	Average	Power5	Square root	Hybrid	Identity	Log	Average	Power5	Square root
ORAL CAVITY, PHARYNX	0.07262	0.06155	0.08511	0.07212	0.07700	0.06896	0.15308	0.12232	0.16828	0.14393	0.15369	0.13823
OESOPHAGUS	0.07133	0.07512	0.07368	0.06856	0.06930	0.06839	0.14795	0.12162	0.17686	0.15135	0.16165	0.14399
STOMACH	0.07223	0.09342	0.06420	0.07319	0.06644	0.07415	0.06437	0.08217	0.06489	0.06540	0.06409	0.06612
INTESTINE (COLON AND RECTUM)	0.03912	0.03621	0.04201	0.03710	0.03892	0.03608	0.05858	0.05600	0.07177	0.06216	0.06584	0.05994
GALLBLADDER AND BILE DUCTS	0.14349	0.14056	0.15491	0.14180	0.14649	0.13904	0.13800	0.14606	0.15459	0.13229	0.13601	0.13183
PANCREAS	0.05757	0.06728	0.05260	0.05599	0.05357	0.05754	0.05130	0.05142	0.06169	0.05115	0.05440	0.04986
OTHER DIGESTIVE ORGANS	1.38842	0.81078	1.46002	1.06020	1.02685	0.90199	0.58909	0.49725	0.62150	0.54329	0.54489	0.52369
LARYNX	0.10403	0.09904	0.12830	0.10965	0.11418	0.10387	0.49016	0.42971	0.51526	0.48485	0.49302	0.46360
LUNG	0.03284	0.03474	0.03660	0.03365	0.03453	0.03328	0.04851	0.06969	0.05821	0.05508	0.05348	0.05695
BONE & ARTICULAR CARTILAGE	0.22357	0.22299	0.23907	0.23094	0.23167	0.22588	0.31135	0.26414	0.34306	0.29235	0.29004	0.27036
SKIN INCLUDING MELANOMA	0.13714	0.10573	0.15014	0.11730	0.12201	0.10840	0.08844	0.08431	0.09946	0.08890	0.09307	0.08659
BREAST	.	.	.	.	.	.	0.04232	0.03512	0.05327	0.04143	0.04682	0.03945
UTERUS (CERVIX AND CORPUS)	.	.	.	.	.	.	0.06198	0.06281	0.06685	0.06109	0.06320	0.06011
PROSTATE	0.07282	0.06390	0.08249	0.07099	0.07503	0.06815	.	.	.	.	.	.
BLADDER	0.11233	0.07651	0.12230	0.08503	0.08071	0.07513	0.09444	0.09540	0.11690	0.10080	0.10583	0.09824
KIDNEY AND OTHER URINARY SITES	0.12198	0.07792	0.13996	0.09885	0.09859	0.08532	0.11427	0.08205	0.12966	0.10121	0.10922	0.09327
BRAIN AND NERVES, BENIGN OR MALIGNANT	0.07920	0.06313	0.08949	0.07049	0.07744	0.06570	0.08079	0.07778	0.09621	0.07630	0.08260	0.07345
THYROID	0.27272	0.19900	0.28689	0.24014	0.24684	0.22153	0.21924	0.17783	0.23825	0.19805	0.20117	0.18736
HODGKIN'S DISEASE	0.26370	0.28292	0.28640	0.26527	0.26317	0.25691	0.57478	0.36638	0.59919	0.44173	0.43186	0.38608

	AARD											
	Men						Women					
	Hybrid	Identity	Log	Average	Power5	Square root	Hybrid	Identity	Log	Average	Power5	Square root
MULTIPLE MYELOMA	0.12009	0.09531	0.18267	0.13341	0.14672	0.11850	0.13512	0.10900	0.16121	0.13423	0.14504	0.12763
LEUKEMIAS	0.10111	0.07495	0.12112	0.09628	0.10630	0.08958	0.09162	0.06242	0.11609	0.08885	0.09939	0.08166
ALL CANCERS (malignant and benign)	0.03104	0.02730	0.03364	0.02918	0.03102	0.02833	0.03674	0.02951	0.03894	0.03417	0.03642	0.03319
ALL CAUSES	0.03035	0.03602	0.02864	0.03058	0.02922	0.03083	0.04413	0.04429	0.04628	0.04452	0.04530	0.04427
CHD (CORONARY HEART DISEASES)	0.06321	0.07187	0.06490	0.06225	0.06306	0.06237	0.06278	0.06585	0.07141	0.06487	0.06771	0.06411
CVD (CEREBROVASCULAR DISEASES)	0.09642	0.08349	0.10787	0.09579	0.10114	0.09304	0.07860	0.07900	0.08606	0.07999	0.08305	0.07941

Table 7. AARDs on rate by projection method, sex and cause of death (projections up to 2006).

	AARD											
	Men						Women					
	Hybrid	Identity	Log	Average	Power5	Square root	Hybrid	Identity	Log	Average	Power5	Square root
ORAL CAVITY, PHARYNX	0.08685	0.07052	0.10284	0.08615	0.09385	0.08256	0.21148	0.16990	0.23104	0.20067	0.21403	0.19373
OESOPHAGUS	0.07450	0.07583	0.08200	0.07354	0.07622	0.07231	0.26010	0.21376	0.29004	0.25281	0.26807	0.24383
STOMACH	0.06707	0.09041	0.06342	0.06963	0.06442	0.07043	0.06539	0.07187	0.07580	0.06785	0.07057	0.06743
INTESTINE (COLON AND RECTUM)	0.04331	0.03643	0.04842	0.04195	0.04472	0.04052	0.06612	0.05889	0.07757	0.06740	0.07183	0.06533
GALLBLADDER AND BILE DUCTS	0.18946	0.17490	0.21373	0.19332	0.20293	0.19074	0.19530	0.15859	0.22735	0.18158	0.19218	0.16922
PANCREAS	0.05080	0.05955	0.05027	0.05058	0.04982	0.05149	0.06200	0.05672	0.06899	0.05851	0.06076	0.05678
OTHER DIGESTIVE ORGANS	2.07846	1.36193	2.15328	1.64826	1.61948	1.46753	1.03143	0.82767	1.08501	0.93199	0.94094	0.88373
LARYNX	0.15079	0.13833	0.17478	0.15283	0.16214	0.14828	1.01596	0.95679	1.04293	1.00179	1.01758	0.99147

	AARD											
	Men						Women					
	Hybrid	Identity	Log	Average	Power5	Square root	Hybrid	Identity	Log	Average	Power5	Square root
LUNG	0.03573	0.03904	0.03820	0.03692	0.03699	0.03680	0.05882	0.06312	0.07398	0.05756	0.05960	0.05490
BONE & ARTICULAR CARTILAGE	0.26898	0.26509	0.28674	0.27286	0.27804	0.26981	0.50321	0.42584	0.56015	0.47836	0.48385	0.44901
SKIN INCLUDING MELANOMA	0.15280	0.10337	0.17464	0.12906	0.13621	0.11638	0.12384	0.11533	0.13534	0.12470	0.12940	0.12292
BREAST	.	.	.	.	.	.	0.04116	0.03911	0.04680	0.03961	0.04308	0.03921
UTERUS (CERVIX AND CORPUS)	.	.	.	.	.	.	0.07315	0.06902	0.08053	0.07188	0.07498	0.07032
PROSTATE	0.08702	0.07091	0.10015	0.08427	0.08951	0.08075	.	.	.	.	.	.
BLADDER	0.14718	0.07789	0.15881	0.09710	0.08867	0.07856	0.17785	0.15203	0.20240	0.17588	0.18513	0.16955
KIDNEY AND OTHER URINARY SITES	0.13912	0.08243	0.15562	0.10635	0.10353	0.09019	0.14026	0.11729	0.15498	0.13486	0.14201	0.13032
BRAIN AND NERVES, BENIGN OR MALIGNANT	0.08826	0.07132	0.09641	0.07992	0.08514	0.07607	0.10102	0.07960	0.10699	0.08796	0.09145	0.08243
THYROID	0.49748	0.43418	0.51711	0.47629	0.48819	0.46597	0.51653	0.39189	0.54625	0.45269	0.45688	0.42257
HODGKIN'S DISEASE	0.33115	0.33212	0.39608	0.34502	0.35679	0.33964	1.15592	0.65710	1.20671	0.84497	0.82199	0.71912
MULTIPLE MYELOMA	0.16202	0.13072	0.21126	0.16728	0.18135	0.15479	0.21172	0.17681	0.23699	0.20895	0.22110	0.20267
LEUKEMIAS	0.10453	0.07941	0.12279	0.09867	0.10895	0.09338	0.09370	0.08592	0.11023	0.09265	0.10039	0.08964
ALL CANCERS (malignant and benign)	0.03313	0.03048	0.03546	0.03228	0.03367	0.03164	0.02785	0.02534	0.03086	0.02715	0.02878	0.02681
ALL CAUSES	0.03630	0.04679	0.03503	0.03867	0.03608	0.03916	0.03780	0.04431	0.03899	0.03846	0.03815	0.03865
CHD (CORONARY HEART DISEASES)	0.05349	0.06672	0.05634	0.05303	0.05395	0.05320	0.06354	0.07371	0.07697	0.06839	0.07190	0.06766
CVD (CEREBROVASCULAR DISEASES)	0.09670	0.08037	0.10850	0.09460	0.10101	0.09176	0.07486	0.07353	0.08655	0.07730	0.08216	0.07682

Conclusions

Cancer mortality trend analyses are important for public health, but current year and future rate trend predictions are essential in order to allocate resources and health services wisely and to prioritize specific prevention activities.

This report aims to describe statistical techniques used in mortality trend analyses, both for descriptive (the EAPC) and inferential (the joinpoint model and the projection) studies. Moreover, it describes and compares projections obtained through six different models: Poisson GLM regressions with identity, logarithmic, square root, power five link functions, a "hybrid" model and an "average" model. The hybrid model is the results of an algorithm implemented in the R software that combines, in the final standardized rate, estimates from the four previous Poisson GLM model (identity, logarithmic, square root, power five link function regression models), choosing the more performing model for each age-group, sex, and cancer site according to the AIC statistic. The average model, simply computes a mean of the predicted estimates obtained from the same four models.

The overall results show that, differently from what I expected, the hybrid model does not give the best predictions, and when it does, the corresponding AARD estimate is not very far from the AARDs of the other methods. However, the hybrid model projections, for any combination of cancer site and sex, are never the worst. Rather, it appears as a compromise of the four models considered, though heavily influenced by the logarithmic model. In the examples from this thesis, its predictive trend does not perfectly overlap the observed trend, but it is not very far off. The average model

predicted estimates are in general better than the hybrid ones, even if they were never the best.

Overall, the AARDs from the six methods are quite similar, there was a strong difference only in a few cases. Moreover, it is possible to notice that, often, the hybrid regression shows AARDs closer to those of Poisson GLM logarithmic link function regression, compared to the other methods. Checking the data, I saw that the algorithm that generated the hybrid model, selected the logarithmic function more frequently as this function fits the data better more often. In any case, the Poisson GLM logarithmic link function method turns out to be a bad predictive function, in spite of fitting existing data better.

Paradoxically, the method that shows the best predictive performance is the Poisson GLM with the identity link function. This method showed much lower AARDs compared to other methods, even when I considered a 10-years projection period. Annually, my research group produced projection estimates for major cancer site in Europe and worldwide using a simple identity model. The results from this thesis encourage continuing to produce predicted estimates through an identity method; furthermore my research group predicts only for very short periods.

Some more general considerations. Projection methods which apply joinpoint regression models to the number of deaths, produce better predicted estimates on number of deaths compared to rates. Considering small countries and minor causes of deaths, i.e. low numbers, causes unreliable projections regardless of the methods used. Finally, we must take into account that predicted trends and corresponding AARDs estimates derived from 5-year projections are definitely better than those on long

periods. Projections over more than five years lack accuracy and, become less relevant to discussions.

One of my aims, through the application to greater more varied data, was to be able to classify problems in order to select the most accurate predictive model according to geographic area, cause of death, sex, age structure and other available covariates. Instead, the results from specific analyses, single country and single cancer sites, are quite discordant. There is no model that emerges as the best in predictive performances. This suggest that there is still a lot to do in order to find an "a priori" or mechanic rule that helps in choosing which predictive method to apply according to various covariates.

During the implementation of the algorithm and the analyses, several interesting angles for further analysis emerged.

To compare and choose the best model for each age-group, sex, and cancer site, the algorithm uses the AIC statistic, but there may be better statistics for comparing the different models?

We retrieved the four transformations used in the algorithm from the literature, are there other relevant and more performant models?

How much does the Joinpoint program influence the projections? It would be interesting to use the Joinpoint program for all the data available through 2011 and then break the dataset to 2001. In this way, the last segment identified by the joinpoint could lead to different coefficient estimates. Furthermore, are the 5 years following the last estimated joinpoint too many?

All these questions are set aside for future development of the project.

In conclusion, prediction of future trends is a complex procedure; hence the resulting estimates should always be taken with caution and considered only as a general indication for epidemiology and health planning.

References

1. Esteve J, Benhamou E, Raymond L. Descriptive Epidemiology *Statistical Methods in Cancer Research* ed., vol. IV: International Agency for Research on Cancer, 1994.

2. Frova L, Marchetti S, Pace M. La codifica automatica delle cause di morte in Italia. *Istituto Nazionale di Statistica* 2005.

3. Qiu Z, Hatcher J, Wang M. Review ancer projection methods for canadian partnership against cancer analytic network. *Alberta Health Services for the Canadian Partnership Against Cancer*.

4. Zocchetti C, Consonni D. Mortality rate and its statistical properties. *Med Lav* 1994;**85**: 327-43.

5. Doll R, G. SP. Comparison between registries: age-standardized rates ed., vol. IV: IARC 1982.

6. Isabel dS. Cancer epidemiology: principles and methods ed.: IARC, 1999.

7. Ahmad OB, Boschi-Pinto C, D. LA, Murray CJL, Lozano R, Inoue M. Age standardization of rates: a new who standard. *GPE Discussion Paper Series: No31* 2001.

8. Petrie A, Sabin C. Rates and Poisson regression. In: Wiley-Blackwell. *Medical statistics at a glance* ed., 2013.

9. Brillinger DR. The natural variability of vital rates and associated statistics. *Biometrics* 1986;**42**: 693-734.

10. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. (Erratum in: Stat Med 2001;20: 655). *Stat Med* 2000;**19**: 335-51.

11. Clegg LX, Hankey BF, Tiwari R, Feuer EJ, Edwards BK. Estimating average annual per cent change in trend analysis. *Stat Med* 2009;**28**: 3670-82.

12. National Cancer Institute https://surveillance.cancer.gov/joinpoint/download.

13. Lerman P. Fitting segmented regression models by grid search. *Appl Statist* 1980;**29(1)**: 77-84.

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

15. Qiu Z, Wang M, J. H, Jiang Z, Hatcher J, Cancer Projection Analytical Network Working Team. Long-Term Projection Methods: Comparison of Age-Period-Cohort Model-Based Approaches. *Alberta Health Services for the Canadian Partnership Against Cancer* 2010.

16. Case RA. Cohort analysis of mortality rates as an historical or narrative technique. *Br J Prev Soc Med* 1956;**10**: 159-71.

17. Moller B, Fekjaer H, Hakulinen T, Sigvaldason H, Storm HH, Talback M, Haldorsen T. Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. *Stat Med* 2003;**22**: 2751-66.

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

19. Clements MS, Armstrong BK, Moolgavkar SH. Lung cancer rate predictions using generalized additive models. *Biostatistics* 2005;**6**: 576-89.

20. Moller H, Fairley L, Coupland V, Okello C, Green M, Forman D, Moller B, Bray F. The future burden of cancer in England: incidence and numbers of new patients in 2020. *Br J Cancer* 2007;**96**: 1484-8.

21. Moller B, Fekjaer H, Hakulinen T, Tryggvadottir L, Storm HH, Talback M, Haldorsen T. Prediction of cancer incidence in the Nordic countries up to the year 2020. *Eur J Cancer Prev* 2002;**11 Suppl 1**: S1-96.

22. Faraway JJ. Linear Model with R ed.: Chapman & Hall/CRC, 2005.

23. Verzani J. *Using R for Introductory Statistics*ed.: Chapman & Hall, 2005.

24. Nelder JA, Wedderburn RWM. Generalized Linear Models. *Journal of the Royal Statistical Society* 1972;**135**: 370-84.

25. Chen HS, Portier K, Ghosh K, Naishadham D, Kim HJ, Zhu L, Pickle LW, Krapcho M, Scoppa S, Jemal A, Feuer EJ. Predicting US- and state-level cancer counts for the current calendar year: Part I: evaluation of temporal projection methods for mortality. *Cancer* 2012;**118**: 1091-9.

26. Zhu L, Pickle LW, Ghosh K, Naishadham D, Portier K, Chen HS, Kim HJ, Zou Z, Cucinelli J, Kohler B, Edwards BK, King J, et al. Predicting US- and state-level cancer counts for the current calendar year: Part II: evaluation of spatiotemporal projection methods for incidence. *Cancer* 2012;**118**: 1100-9.

27. World Health Organization Statistical Information System. WHO mortality database. Geneva: World Health Organization *Available at: http://wwwwhoint/healthinfo/statistics/mortality_rawdata/en/indexhtml* Last accessed November, 2015.

28. Mathers CD, Fat DM, Inoue M, Rao C, Lopez AD. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bull World Health Organ* 2005;**83**: 171-7.

29. World Health Organization. *International Classification of Disease: 8th revision*ed. Geneva: World Health Organization, 1965.

30. World Health Organization. *International Classification of Disease: 9th revision*ed. Geneva: World Health Organization, 1977.

31. World Health Organization. *International Classification of Disease and related Health Problems: 10th revision*.ed. Geneva: World Health Organization, 1992.

32. European Commission. Eurostat population database, vol. 2014 (Last access July), 2014 (Last access July).

33. Pan American Health Organization (PAHO). Health Statistics from the Americas, 2006 Edition, Chapter VI, 2-20. *Available at: http://wwwpahoorg/English/DD/AIS/HSA2006htm* (Last accessed April 2018).

Supplementary material

Tables

Table 1S. AARDs on rate by cancer site and projection method (projections up to 2006).

	AARD					
	Hybrid	Identity	Log	Average	Power5	Square Root
ORAL CAVITY, PHARYNX	0.14916	0.12021	0.16694	0.14341	0.15394	0.13815
OESOPHAGUS	0.16730	0.14480	0.18602	0.16318	0.17215	0.15807
STOMACH	0.06623	0.08114	0.06961	0.06874	0.06749	0.06893
INTESTINE (COLON AND RECTUM)	0.05472	0.04766	0.06299	0.05468	0.05827	0.05292
GALLBLADDER AND BILE DUCTS	0.19238	0.16674	0.22054	0.18745	0.19755	0.17998
PANCREAS	0.05640	0.05813	0.05963	0.05455	0.05529	0.05414
OTHER DIGESTIVE ORGANS	1.55495	1.09480	1.61915	1.29012	1.28021	1.17563
LARYNX	0.58337	0.54756	0.60885	0.57731	0.58986	0.56988
LUNG	0.04728	0.05108	0.05609	0.04724	0.04830	0.04585
BONE & ARTICULAR CARTILAGE	0.38609	0.34547	0.42344	0.37561	0.38094	0.35941
SKIN INCLUDING MELANOMA	0.13832	0.10935	0.15499	0.12688	0.13280	0.11965
BREAST	0.04116	0.03911	0.04680	0.03961	0.04308	0.03921
UTERUS (CERVIX AND CORPUS)	0.07315	0.06902	0.08053	0.07188	0.07498	0.07032
PROSTATE	0.08702	0.07091	0.10015	0.08427	0.08951	0.08075
BLADDER	0.16251	0.11496	0.18061	0.13649	0.13690	0.12406
KIDNEY AND OTHER URINARY SITES	0.13969	0.09986	0.15530	0.12060	0.12277	0.11026
BRAIN AND NERVES, BENIGN OR MALIGNANT	0.09464	0.07546	0.10170	0.08394	0.08830	0.07925
THYROID	0.50700	0.41303	0.53168	0.46449	0.47254	0.44427
HODGKIN'S DISEASE	0.74354	0.49461	0.80139	0.59499	0.58939	0.52938
MULTIPLE MYELOMA	0.18687	0.15376	0.22413	0.18812	0.20123	0.17873
LEUKEMIAS	0.09911	0.08267	0.11651	0.09566	0.10467	0.09151
ALL CANCERS (malignant and benign)	0.03049	0.02791	0.03316	0.02972	0.03122	0.02923
ALL CAUSES	0.03705	0.04555	0.03701	0.03856	0.03712	0.03890
CHD (CORONARY HEART DISEASES)	0.05852	0.07022	0.06665	0.06071	0.06292	0.06043
CVD (CEREBROVASCULAR DISEASES)	0.08578	0.07695	0.09752	0.08595	0.09159	0.08429

Table 2S. AARDs on number of deaths by cancer site and projection method (projections up to 2006).

	AARD					
	Hybrid	Identity	Log	Average	Power5	Square Root
ORAL CAVITY, PHARYNX	0.11285	0.09193	0.12669	0.10803	0.11535	0.10359
OESOPHAGUS	0.10964	0.09837	0.12527	0.10995	0.11548	0.10619
STOMACH	0.06830	0.08779	0.06454	0.06929	0.06526	0.07014
INTESTINE (COLON AND RECTUM)	0.04885	0.04611	0.05689	0.04963	0.05238	0.04801
GALLBLADDER AND BILE DUCTS	0.14075	0.14331	0.15475	0.13705	0.14125	0.13543
PANCREAS	0.05444	0.05935	0.05715	0.05357	0.05398	0.05370
OTHER DIGESTIVE ORGANS	0.98876	0.65402	1.04076	0.80174	0.78587	0.71284
LARYNX	0.29709	0.26437	0.32178	0.29725	0.30360	0.28373
LUNG	0.04067	0.05221	0.04741	0.04437	0.04401	0.04512
BONE & ARTICULAR CARTILAGE	0.26746	0.24357	0.29106	0.26165	0.26085	0.24812
SKIN INCLUDING MELANOMA	0.11279	0.09502	0.12480	0.10310	0.10754	0.09750
BREAST	0.04232	0.03512	0.05327	0.04143	0.04682	0.03945
UTERUS (CERVIX AND CORPUS)	0.06198	0.06281	0.06685	0.06109	0.06320	0.06011
PROSTATE	0.07282	0.06390	0.08249	0.07099	0.07503	0.06815
BLADDER	0.10339	0.08595	0.11960	0.09291	0.09327	0.08669
KIDNEY AND OTHER URINARY SITES	0.11812	0.07998	0.13481	0.10003	0.10390	0.08930
BRAIN AND NERVES, BENIGN OR MALIGNANT	0.07999	0.07046	0.09285	0.07340	0.08002	0.06958
THYROID	0.24598	0.18842	0.26257	0.21910	0.22401	0.20444
HODGKIN'S DISEASE	0.41924	0.32465	0.44280	0.35350	0.34752	0.32149
MULTIPLE MYELOMA	0.12760	0.10215	0.17194	0.13382	0.14588	0.12307
LEUKEMIAS	0.09637	0.06869	0.11860	0.09256	0.10284	0.08562
ALL CANCERS (malignant and benign)	0.03389	0.02840	0.03629	0.03168	0.03372	0.03076
ALL CAUSES	0.03724	0.04016	0.03746	0.03755	0.03726	0.03755
CHD (CORONARY HEART DISEASES)	0.06299	0.06886	0.06815	0.06356	0.06539	0.06324
CVD (CEREBROVASCULAR DISEASES)	0.08751	0.08125	0.09696	0.08789	0.09209	0.08622

Table 3S. AARDs on rate by sex and projection method (projections up to 2006).

	AARD					
	Hybrid	Identity	Log	Average	Power5	Square Root
Men	0.21457	0.16912	0.23211	0.19103	0.19375	0.17869
Women	0.26100	0.21127	0.28110	0.23772	0.24284	0.22493

Table 4S. AARDs on number of deaths by sex and projection method (projections up to 2006).

	AARD					
	Hybrid	Identity	Log	Average	Power5	Square Root
Men	0.16006	0.12534	0.17409	0.14160	0.14253	0.13017
Women	0.15647	0.13313	0.17217	0.14661	0.15035	0.13923

Table 5S. AARDs on rate by categories of mortality counts and projection method (projections up to 2006).

	AARD					
	Hybrid	Identity	Log	Average	Power5	Square Root
≤50000	1.14924	0.79471	1.21027	0.94256	0.93480	0.85250
50000-100000	0.44655	0.37925	0.47756	0.42005	0.42674	0.40184
100000-500000	0.19836	0.17060	0.21839	0.19114	0.19930	0.18436
>500000	0.06931	0.06469	0.07756	0.06676	0.06909	0.06458

Table 6S. AARDs on number of deaths by categories of mortality counts and projection method (projections up to 2006).

	AARD					
	Hybrid	Identity	Log	Average	Power5	Square Root
≤50000	0.70400	0.48933	0.74178	0.57762	0.56669	0.51717
50000-100000	0.25672	0.21599	0.27682	0.24037	0.24243	0.22628
100000-500000	0.13278	0.11481	0.15116	0.12848	0.13453	0.12206
>500000	0.06285	0.06079	0.07005	0.06174	0.06372	0.06008