uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Chushu Gu

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Systems Science)

GRADE / DEGREE

Systems Science

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Estimating Life-expectancy Changes for Medical Decision Making: New Approximations

TITRE DE LA THÈSE / TITLE OF THESIS

Professor Kevin Brand

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

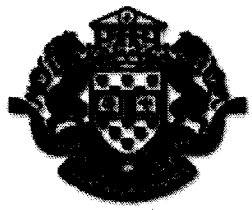
EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Professor Wojtek Michalowski

Professor Rama Chandran Nair

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies



Master's Program in Systems Science

Thesis

Estimating Life-expectancy changes for Medical Decision Making: New Approximations

Chushu Gu
#3123202

Thesis supervisor:
Professor Kevin Brand

March 20, 2006

Your file Votre référence
ISBN: 978-0-494-18420-2
Our file Notre référence
ISBN: 978-0-494-18420-2

Abstract

Life-expectancy and Life-years lost are frequently used and analyzed indices of survival. Life tables and Markov models are two exact approaches to calculate these indices; however cumbersome calculation limits their usage in real situations. Some simple approximation approaches have therefore been developed since a convenient and accurate approximation is critical both to develop a treatment plan of a patient by physicians and to assess health policies by health policy makers. These approximation approaches include the DEALE (Declining Exponential Approximation of Life Expectancy), new DEALEs, the IPH method (A method developed at Institute of Population Health, University of Ottawa) and the Keyfitz approach. A new approach has been developed to achieve better accuracy and maintain ease of application by extending the Keyfitz approach. To make the new approach less dependent on age–stratified tabulations, a convenient formula for the EME (Established market economics) region is developed. Its accuracy, robustness, and ease of application are demonstrated.

Keyword: the DEALE, new DEALEs, the IPH method, the Keyfitz approach, the EME region

Table of Contents

Table of Figures

Table of Tables

1. Introduction

Lay people, medical practitioners, researchers, and policy makers can relate to the concept of life-expectancy (LE). The life-expectancy of Canadians at birth is periodically reported in the media. For example, the life-expectancy of Canadians (at birth) was 79.3 years in year 2001. This index is sometimes compared to those from other countries. For example, Japanese have a LE of 81.4 years in year 2001, which is better than that of the Canadians. As another example, the people live in Southern Africa have shorter life-expectancy compared to the Canadians. The life-expectancy (i) is changed when the people are affected by some new health states; (ii) could be improved by some health interventions [1]; (iii) may worsen if exposed to some risk factors. The change of life-expectancy is a useful construct to express the impact of the new health state on a person's longevity prospect.

Medical researchers have applied the Life Table and the Markov model as two exact approaches to calculate the life-expectancy accurately. However, cumbersome calculation of these two methods limits their use in practice. Simple approximation approaches have therefore been proposed. The most popular approach is the DEALE [2, 3].

The objective of this thesis is to develop a new approximation approach, which fills a gap where other approximation approaches have bad performance. The existing approximation approaches depend on the availability of age-stratified tabulations, for example, the baseline life-expectancy tabulated by age. To make the new approach independent of these tabulations is another focus of this thesis. A convenient (not relying on the age-stratified tabulations) formula based on this new approach for all the populations in the EME(Established market economics) region will be developed. This new approximation approach is referred to as the "Extended Keyfitz Model".

In this thesis, first of all, the related health indices are introduced. Secondly, the exact approaches and the existing approximation approaches are visited. The new approximation approach "Extended Keyfitz Model" is derived thereafter. The accuracy of the Extended Keyfitz Model will then be demonstrated against the existing approximation approaches such as the DEALE, the new DEALEs and the Keyfitz approach [2, 3, 4, 5].

2. Background

2.1. Related Health Indices

Quantitative techniques have been used increasingly to help formulate health policy for whole populations and subpopulations and to aid in clinical decision making for individual patients. A health policy maker may be interested in the effect of a particular health intervention, so he can make the resource allocation policy accordingly. As another example, a physician may want to know the life-expectancy associated with the available therapeutic choices for a particular patient in order to make an optimized treatment plan. With many such applications of quantitative techniques, indices such as baseline life-expectancy, modified life-expectancy and life-years lost are commonly used for decision-making.

Brand examined a problem in the population health realm, which involved trying to quantify the population health impact from exposure to a particular environmental risk factor (Residential radon exposure) [6]. This analysis used a rigorous approach (the Life Table) to quantify impacts, but remarked upon some simple patterns in the data that suggested simpler approach might be used to obtain the same result. Brand, then derived a new approach to approximate the Life Table approaches by using the Taylor series expansion. He argued that his approach (Labeled the IPH approach, for the Institute of Population Health) would work well if the modification to mortality is not extreme [7]. He also pointed out other existing approximation approaches, including the aforementioned DEALE [2, 3] and an approach by Keyfitz [5]. He argued that the Keyfitz approach might be more elegant

than the IPH approach but suggested that these two approaches were likely to result in nearly identical answers (the Keyfitz approach was thought slightly more accurate). Brand argued that the IPH and Keyfitz approaches were likely to be superior to the DEALE approach, but he had not explored the issue in a comprehensive manner. This thesis was intended to fill this gap, by better establishing the performance of the IPH/Keyfitz approximations relative to the DEALE. As it progressed, however, we became interested in improving the performance of the IPH/Keyfitz approaches. Because of the superior elegance of the Keyfitz approach, the focus of these improvement efforts became the Keyfitz approach.

The IPH/Keyfitz approaches were originally motivated by a population health context wherein the modifications (to mortality) of interest are comparatively small compared to the modifications that are of interest in a clinical setting. For this reason, another purpose of this thesis was to fully evaluate the applicability of the IPH/Keyfitz approach in the clinical setting where the modifications of interest can get quite large! Indeed it was this interest in larger modifications that motivated efforts to try to improve the Keyfitz approach's ability to handle large modifications (the IPH/Keyfitz approaches' performance for more moderate or small modifications was argued to be quite good [5, 7]).

The Keyfitz/IPH approaches were designed to handle a particular form of modification (to mortality), which is called a multiplicative modification. The additive and multiplicative modification to mortality will be discussed later in Section 2.3.2.

In this thesis there are four types of modifications that might be explored:

1. "Tailoring prospects" to some extenuating health condition which will most often be used in the clinical contexts
2. "Exploring impact of intervention" where the counterfactual consists of the mortality rate schedule hypothesized to hold in the face of an intervention
3. "Exploring the side effects of an exposure to a risk factor"

4. Some combination of 1 through 3, where one is interested in the prognosis of those affected by an extenuating health condition both in the absence and the presence of an intervention (this too is more likely to be an issue in the clinical settings)

We start with introducing the concept of some health indices:

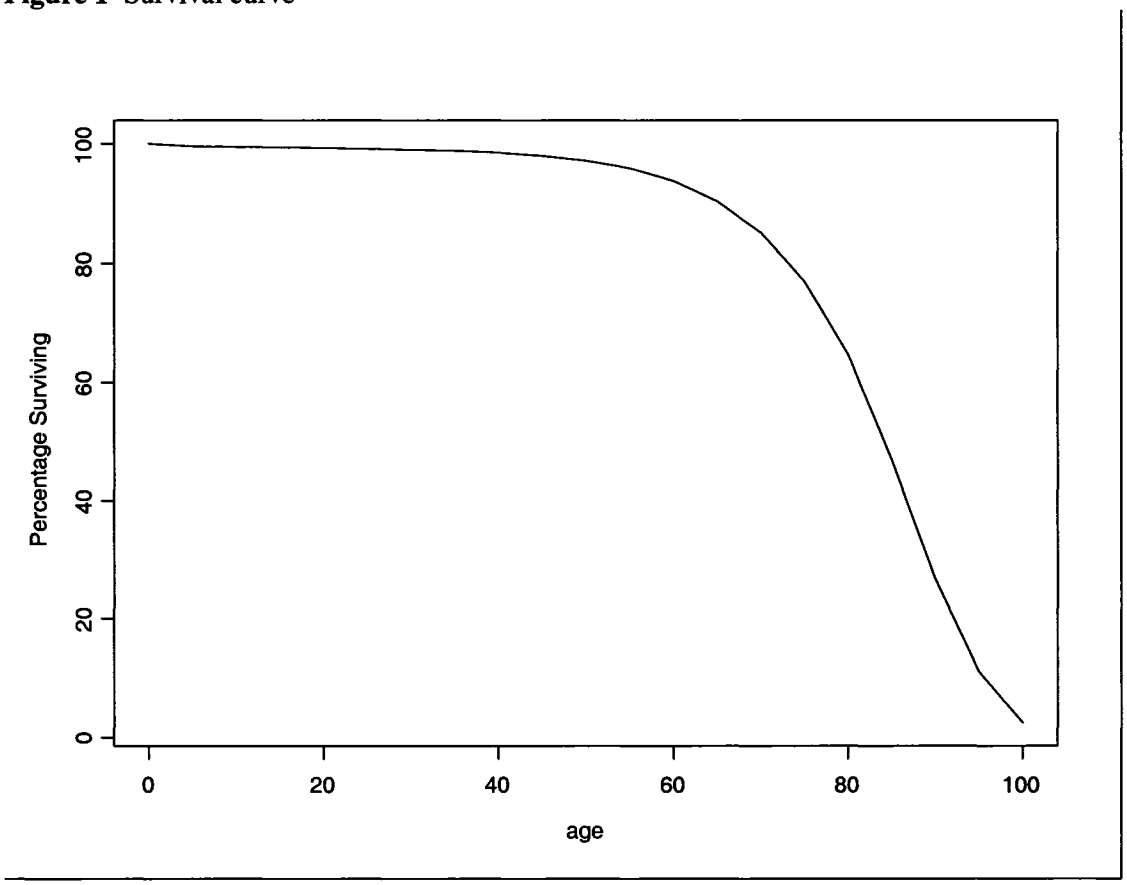
2.1.1. Baseline life-expectancy (LE_b)

When a life-expectancy is reported for a population, what does it mean? It is the average number of additional years a person in this population could expect to live if current mortality trends were to continue for the rest of that person's life. This reported life-expectancy is called baseline life-expectancy. There are three major demographic attributes associated with the baseline life-expectancy: age, sex and race.

Life-expectancy calculation can be illustrated by a survival curve [8, 9]. A survival curve expresses the survival prospects of a population as a function of age. Each point in this curve represents the fraction of individuals who are expected to survive to the given age. As an example, Figure 1 is a survival curve based on Canadian female population in year 2000. Figure 1 plots the percentage of the survival as a function of age. The fraction of population that will survive can be used as proxy for estimating probability of survival for an individual.

The area under the survival curve represents the life-expectancy for an individual at birth from the population of interest. The above focuses exclusively on the longevity prospects of those at birth. The more general case, namely examining the prospects of those who have already achieved some age of interest (such as age 40) is also of interest. How does one compute the (remaining) life-expectancy of those individuals that have already attained some age t? From the survival curve, it is actually the area under the curve from age t to the right, which is then divided by the probability of the original population just born will survive to age t (the height of survival curve at age t).

Figure 1 Survival curve



Note: A survival curve plots the percentage of survival as a function of age. The fraction of population that will survive can be used as proxy for estimating probability of survival for an individual. It is plotted for the Canadian female population in year 2000.

2.1.2. Modified life-expectancy (LE_M)

Based on the knowledge of baseline life-expectancy that was just introduced, suppose an individual has a health condition such as brain cancer, one would expect this condition to be associated with a higher mortality rate relative to similarly aged person who does not face this condition. One might be interested in his/her revised life-expectancy? A similar question could be asked about a population. Consider a hypothetical population that is subject to a smoking cessation program, the policy maker might be interested in the benefit in terms of a longer life-expectancy from this health intervention. In general, the revised life-expectancy is of special interest because it gauges

the adverse effect in the first example and the beneficial effect in the second example. We refer these new health conditions as modified cases.

What happens under these two kinds of new health conditions? In the first example, his/her hazard increases. The population with brain cancer has higher mortality than the general population. He/She is now subject to higher hazard of the population with brain cancer rather than the lower hazard of the general population. In the second example, a reduction in the mortality is caused by the smoke cessation program. Thus the health state change is also referred to as modification to mortality.

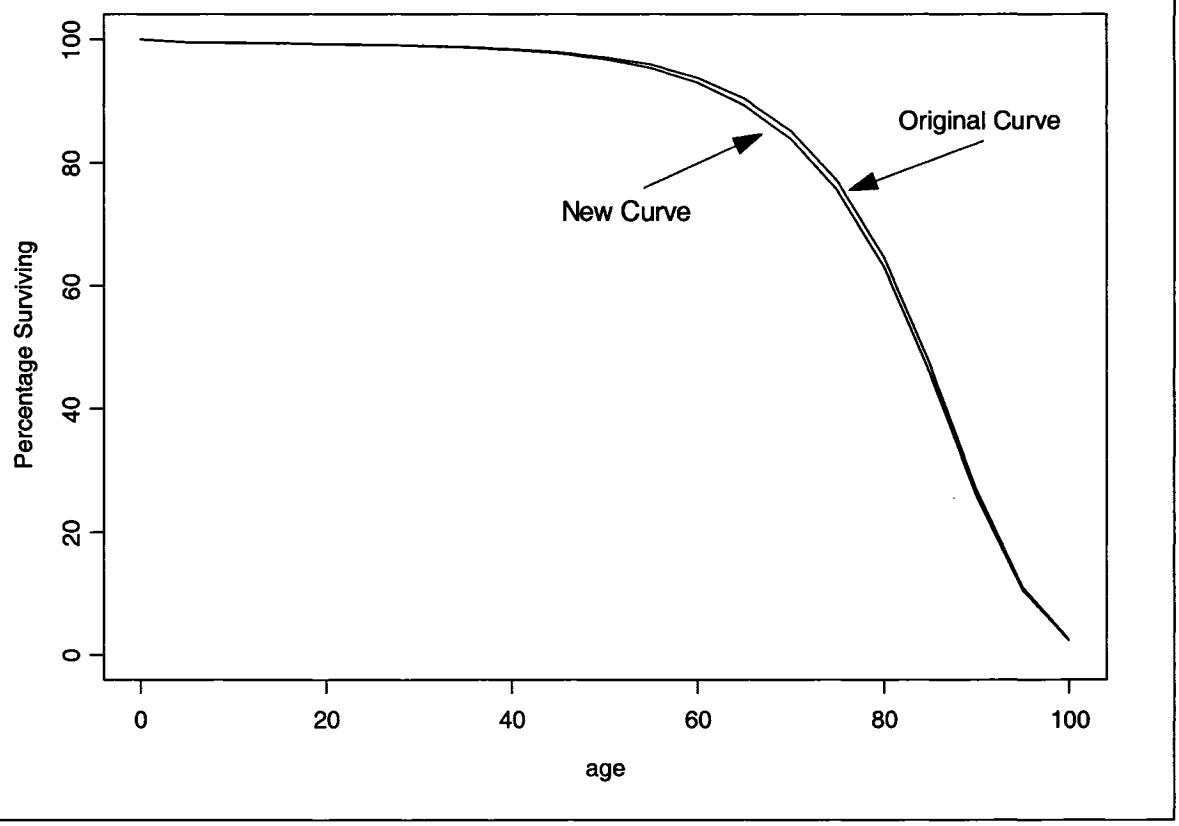
The next natural question is how to calculate the modified life-expectancy. It also can be computed by the survival curves. An example might help to illustrate this calculation.

Example A:

Consider that exposure to some risk factors of interest amplifies the risk of dying of brain cancer in Canadians. Suppose that evidence suggests as much as a 5 times hazard higher than the general population. Suppose the population under this exposure is Canadian female population in year 2000. In this case one might have a special interest to know the baseline and modified life-expectancy for an individual from this population at age 30?

The modified life-expectancy is not surprisingly calculated by the new survival curve. Figure 2 shows the new and original survival curve for Example A. I will use this example throughout this thesis to demonstrate different approaches on how to calculate the life-expectancy indices.

Figure 2 Survival curve under the modified case in Example A



Note: The original survival curve is presented in Figure 1. The new curve is plotted under the modified case, which is subject to 5 times higher brain cancer hazard. The baseline life-expectancy is calculated by the original survival curve. The modified life-expectancy is calculated by the new survival curve under the modified case. The two curves are very close indicates that brain cancer is not a major cause of death. All the calculations are based on the real mortality data of Canadian female population in year 2000.

2.1.3. Life-years lost (LYL)

It is more direct to use the change of life-expectancy to gauge the impact of the new health states, or exposure to some risk factors, or some health interventions. So rather than focusing on the absolute life-expectancy in the face of the modification (to mortality) of interest, one would focus on the change in life-expectancy associated with that modification. The index life-years lost (LYL) is defined to concentrate on this change. It is the difference between the baseline life-expectancy and the

modified life-expectancy ($LE_b - LE_M$). If the difference is negative, a different term is used, namely "Life-years gained".

The LE_M and LYL can be specific for a particular cause of death. When dealing with a specific cause of death, an added term (generally age-specific) is required in the calculations of LYL. Referred to as the intensity ratio [8] and denoted by ψ, this term represents the proportion of deaths in each age interval that can be anticipated to arise from the specific cause of death. For example, if the cause of interest is brain cancer, the added term ψ would consist of as many entries as there are in age intervals and would specify the fraction of deaths in each age interval attributable to brain cancer. The incorporation of this additional factor in the calculations of LYL by the Life Table is detailed in Section 2.2.1 and Appendix A.

The index LYL is very useful. Consider an individual with brain cancer who has LYL of 5 years and another individual with lung cancer who has LYL of 10 years. One would be clear that the impact of lung cancer is more serious than brain cancer.

Baseline life-expectancy, modified life-expectancy and LYL are my focus in this thesis. Especially LYL is most often used.

2.1.4. Baseline life time risk (LR_b)

The baseline life time risk is another important health index. The main focus of this thesis is life-expectancy indices, however life time risk is used in some parts of my thesis and the definition is introduced here.

The life time risk of dying of a specific cause of death is simply represented as the sum of the corresponding age-specific probability (across age):

$$LR_b = \sum_{i=1}^{N} S_i q_i^* \psi_i \qquad (1)$$

Where N denotes the number of age intervals,

S_i denotes the probability an individual just born will survive to age i,

q_i^* denotes the probability of death in the i-th age interval,

and ψ_i denotes fraction of deaths attributed to the specific cause of death in the i-th age interval, which is just discussed in the last section.

S_i can be calculated by:

$$S_i = \prod_{k=1}^{i-1}(1-q_k^*) \qquad (2)$$

Where k denotes the k-th age interval,

i denotes age,

q_k^* denotes the probability of death in the k-th age interval

and * denotes all cause of death.

2.2. Exact approaches

It is well known that there are two exact approaches to calculate the above life-expectancy indices accurately: the Life Table and the Markov model. The word "exact" may not be proper as these two approaches will not yield two exact values of Life-expectancy. Instead, these two approaches will yield two best estimates of Life-expectancy, which are accepted as the standards in literature. The Life Table is chosen in this thesis as the standard to evaluate the performance of different approximation approaches.

2.2.1. Life Table

A Life table is a tabular display of life expectancy and probability of dying at each age (or age group) for a given population, according to the age-specific death rates prevailing at that time. The life table gives an organized, complete picture of a population's mortality. It follows a hypothetical cohort of 100,000 persons born at the same time as they progress through successive

ages, with the cohort reduced from one age to the next according to the mortality rate schedule until all persons eventually die.

Life-tables may be complete or abridged, depending on the age intervals used in their compilation. Complete life-tables contain data broken by single year age intervals, while abridged life-tables contain data subdivided into longer age intervals: five-year intervals are most common while ten-year intervals are occasionally used.

From the standard expressions laying out the construction of the life-table, the baseline life-expectancy (LE_b) can be calculated by [8, 9]:

$$LE_b = \sum_{i=1}^{N-1} \left(S_i (1 - \alpha_i) + S_{i+1} \alpha_i \right) \omega_i \qquad (3)$$

Where N denotes the number of age intervals,

ω_i denotes the number of years in the i-th age interval,

S_i denotes the probability an individual just born will survive to age i,

and α_i denotes the average proportion of the i-th age interval lived by those who die in that age interval.

When calculating the life-expectancy under the modification to mortality, the constructed life-table is called cause-modified life-table (CMLT). The modified life-expectancy can be calculated using a similar equation [8]:

$$LE_M = \sum_{i=1}^{N-1} \left(S_{M,i} (1 - \alpha_i) + S_{M,i+1} \alpha_i \right) \omega_i \qquad (4)$$

Where $S_{M,i}$ denotes the probability an individual just born will survive to age i when he/she is subjected to a modified mortality rate schedule. Other terms have already been defined in Equation (3).

$S_{M,i}$ is calculated in the same way as S_i, namely,

$$S_{M,i} = \prod_{k=1}^{i-1}(1 - q^*_{M,k}) \quad (5)$$

Where $q^*_{M,k}$ denotes the probability of death in the k-th age interval subject to the modified mortality rate schedule.

The probability of death $q^*_{M,k}$ is calculated by the following Equation:

$$q^*_{M,k} = 1 - \left(1 - q^*_k\right)^{(1+\psi_k \varepsilon_c)} \quad (6)$$

Where * denotes all causes of death,

Subscript M denotes modified case,

ψ_k denotes fraction of deaths attributed to the specific cause of death in the k-th age interval,

ε_c is the excess rate ratio (ERR) on a specific cause of death (rate ratio of mortality by a specific cause of death after and before mortality modification minus 1). Subscript c denotes a specific cause of death. More of the details about ERR will be discussed in section 2.3.2.

and all other terms have already been defined before.

These equations are complicated and require a series of data inputs like the mortality rates ordered by age group; this requirement for the detailed age-specific data is one of the disadvantages of the exact approaches (one shared by the Life Table's counterpart, the Markov model). To get an idea of what the data inputs the Life Table must have, let us consider a life-table using 5-year age intervals, and suppose the age limit of this life-table is 110. In this case, the mortality rates of 22 age groups are needed to calculate the baseline life-expectancy. If you are going to calculate the LYL due to a specific cause of death, then additional 22 data inputs (ψ_i, fraction of deaths attributed to the specific cause of death in each age group) are needed. This is the best scenario. If this life-table is using 1-year age intervals, the mortality rates of 110 age

groups are needed to calculate the baseline life-expectancy. Again, if you are going to calculate the LYL due to a specific cause of death in this case, 220 data inputs are needed.

The calculation of Example A (in Section 2.1.2) by the Life Table is presented in Appendix A. Please refer to Appendix A for the details. From Appendix A, we know

$LE_b = 52.25$ years

$LE_M = 51.74$ years

$LYL = LE_b - LE_M = 0.51$ years

2.2.2. Markov Model

Some of the researchers use computer simulation programs based on the Markov model [10, 11] as their standard to calculate the life-expectancy. For example, the DEALE paper [2, 3] used computer simulation based on the Markov model to calculate the exact life-expectancies. These exact values were then used as the references for estimating the performance of their approximation approach.

The Markov model is introduced to show these researchers that the Life Table and the Markov model will get very close results for the life-expectancy calculation. Markov models [10, 11] succinctly represent situations in which there is an ongoing risk of a patient moving from one state of health to another. It is assumed that there are a set of possible health states, and specify the probability per unit of time that a patient in a given state will "transition" to each possible state. What we do assume, however, is that the process has no memory -- how we came to this state doesn't matter, only when we came. This is the prerequisite for applying Markov model. The detailed concepts about the Markov model are described in Appendix B.

Markov models require computer programs to conduct the simulation.

A simulation program in S-plus has been built to calculate LE_b and LE_M of Example A. This S-plus program is listed in Appendix I for your reference. The life-expectancy indices calculated for Example A by the two exact approaches are presented in Table 1:

Table 1 Life-expectancy indices by the exact approaches for Example A

INDEX	LIFE TABLE	MARKOV MODEL
LE_b	52.25	52.33
LE_M	51.74	51.84
LYL	0.51	0.49

Note: The LE_b, LE_M and LYL from the Life Table and the Markov model are very close. The LE_b and LE_M from the Markov model are the mean values of ten runs of simulation. The LE_b and LE_M by the Markov model are slightly larger than those from the Life Table.

In Table 1, the life-expectancy indices LE_b and LE_M calculated by the Markov model are slightly larger than those calculated by the Life Table.

2.3. Approximation approaches

2.3.1. Why seek an approximation?

The two exact approaches can be applied to calculate the health indices accurately. However, they both share similar limitations (some of which were alluded to already):

- The calculation is time consuming.
- The data may not be available. These two exact approaches need detailed information such as mortality rates stratified by age group (mortality rate schedule).

  For a complete life table, when the age limit is 110, 110 data inputs are needed. For an abridged life table (5-year age interval), in this case 22 data inputs are needed. The same data inputs are needed for the Markov models. Such a resolution of data is not always available.
- Special computation skill is needed. To apply Markov model, a simulation program has to be made. In practice, a physician most probably will have a calculator and such requirement to make a computer program to conduct a simulation is infeasible.

In practice, these disadvantages may limit the number of their applications, especially when the real case is urgent or the data is limited. As a result, simplified approximation approaches have been developed. The requirements for a useful approximation are

- It is simple to use. Suppose an individual just has a simple calculator, the good approach could rely on that calculator to get the results directly.
- The approach should work with limited data inputs.
- The end user doesn't need to have advanced skill on computation. For example, the end user does not need to make a computer program to use the approximation approach.

Currently available approximation approaches include the DEALE [2, 3], new DEALEs [4], the IPH method [7], the Keyfitz approach [5] and some other methods. These approaches approximate the life-expectancy indices under the modification to mortality. The underlying models of the modification to mortality are discussed in the next section.

2.3.2. Additive and Multiplicative modifications

For a long time, medical practitioners, policy-makers and researchers are interested in describing and projecting the mortality change (modification to mortality) in the future. In the real life, there is no evident way on how to model the mortality change in the future. Does this modification come about in an additive, or multiplicative way, or some other different ways? A modification to mortality can be modeled in a variety of ways, but two contrasting approaches stand out as representing common practice: The future mortality change is either treated as additive or multiplicative. "In the literature on the analysis of failure time data focused primarily on models that specify that the effect of covariates is to act multiplicatively on the baseline hazard rate"[12]. "On the contrary, the decision analysis literature has focused mainly on the additive ---- or excess-mortality---- concept ..."[12]. These two models are compared in terms of their effect on decision-analytic results in Kuntz's paper [12] for evaluating the decision between coronary artery bypass grafting and medical therapy in a 50-year-old male patient with three-vessel

coronary disease. The life-expectancies were found shorter and the differences in life-expectancies were found smaller in multiplicative model in this case.

To understand the difference between the two different models, it is helpful to re-express the multiplicative model in similar terms in an additive model. In multiplicative model, the relative increase/decrease is expressed by the excess relative rates (excess rate ratio). The implied additive change is obtained by forming a product with the age specific excess relative rate with age-specific baseline mortality rate. The implied absolute change will generally vary with age. In a similar way, an additive model could be re-expressed in terms of the multiplicative model. The resulting excess relative rate generally will vary with age. When the baseline mortality is constant across age, a constant multiplicative modification is interchangeable to a constant additive modification.

Which model is preferred depends on the biological insight. In reality, the mortality change of a specific cause of death is summarized either in additive or multiplicative terms by the epidemiology studies of the exposure response relationship.

What might the theoretic rationale in favor of one or the other model be? An additive mode of change is by definition independent of the baseline mortality. For some causes of death that somehow preyed upon some degree of susceptibility (pre-existing conditions), this independence is not expected. In this case, the mortality change and the baseline mortality rate share a common influence, namely the size of the pool of susceptible. The dependence between the mortality change and the baseline mortality is expected. The dependence upon susceptibility is likely to be a common trait among chronic diseases, which suggests that most chronic diseases would be more suitable to be modeled in multiplicative mode. The causes of death such as suicide or motor vehicle accidents might not have such dependence, that is, they may not depend upon a pool of susceptible. Generally the pathogenesis would dictate the mode of mortality change. It also depends on the data availability, namely the data collected by the studies of the exposure response relationship.

Analysts making projection of LYL involving mortality modification typically assume a simple mode of modification. A fixed additive or a fixed multiplicative change in mortality is typically assumed across all age. But what is the difference between the additive and the multiplicative constant mortality change? Under the constant additive change, baseline mortality rates are altered with an increment or decrement that is constant (in absolute terms) across age. In contrast, under a constant multiplicative change, a constant relative change in mortality is imposed on each age-specific mortality rate.

The DEALE, new DEALEs were originally devised for the additive modification. The Keyfitz and IPH are devised for the multiplicative modification. The special case is the DEALE. Although the DEALE is devised for the additive modification, it also can be used in the multiplicative modification. The reason is that the DEALE assumes a constant baseline mortality across age. As discussed in the previous text, a constant multiplicative modification can then be re-expressed as a constant additive modification. An example to illustrate this process will be discussed in the next section.

Excess mortality rate (ΔM) and Excess rate ratio (ERR ε_c)

In this thesis, there will be two symbols to denote age: i and t. In the previous sections, the symbol i is an integer and discrete variable. It represents certain age or age group. The symbol t will be used to denote the age too, but it is used as a continuous variable.

In additive term, excess mortality rate (ΔM_t) is used to describe the net change of mortality rate at age t. It is defined as:

$$\Delta M_t = M_{M,t} - M_{b,t} \qquad (7)$$

Where ΔM_t denotes the excess mortality rate at age t,

$M_{M,t}$ denotes the mortality rate in modified case at age t,

$M_{b,t}$ denotes the baseline mortality rate at age t.

Positive/Negative excess mortality rate (ΔM_t) implies a net increase/decrease of the mortality respectively at age t. In contrast, excess rate ratio (ERR $\varepsilon_{c,t}$) is used in a multiplicative term to describe the relative increase/decrease of the mortality. It is defined as:

$$\varepsilon_{c,t} = \Delta M_t / M_{c,t} \qquad (8)$$

Where $\varepsilon_{c,t}$ is the excess rate ratio of a specific cause of death at age t,

ΔM_t denotes the net change of mortality rate at age t,

$M_{c,t}$ denotes the mortality rate of a specific cause of death at age t,

And subscript c denotes a specific cause of death.

Similarly, positive/negative $\varepsilon_{c,t}$ indicates a relative increase/decrease of the mortality ($M_{c,t}$) for a specific cause of death. The lower limit of $\varepsilon_{c,t}$ is -1, corresponding to the elimination of a specific cause of death at age t. There is no theoretical upper limit for $\varepsilon_{c,t}$. However, $\varepsilon_{c,t}$ has the practical upper limit. In the approximation approach, a constant mortality change will be calculated based on the age specific ΔM_t and $\varepsilon_{c,t}$. In this case, ΔM and ε_c will be used (drop subscript t).

In the modified case, the mortality either increases or decreases. In Example A, the mortality increases multiplicatively with exposure to some risk factors. In contrast, the reduction in mortality is common, for example some health interventions could reduce mortality [1]. In DEALE paper, the reduction in mortality is not covered. The possible reason is that the authors of the DEALE focused on the clinical use of their approach and most of the modified cases in clinical setting are positive modified cases. To fully understand the performance of the DEALE, its performance under negative modifications is necessary to be explored.

- **Cause of death Projection**

When dealing with a specific cause of death, the approximations would be easy if ψ_i were fixed at a constant value across all the age intervals, where *i* denotes the i-th age interval. In that case the change in mortality rate would be adjusted in proportion to the fixed value, ψ. In the case of the IPH and Keyfitz approximations the characteristic numbers in these approaches would simply be scaled by this fixed value to reflect the role of the cause of death.

On the condition that $\varepsilon_{c,t}$ and ψ_i across age are fixed, ε_c and ψ are used to denote the excess rate ratio and fraction of death attributable to a specific cause of death (The symbol i and t are dropped to reflect these values are fixed across age). In this case, the LR_b is actually ψ (Through Equation 1). The net mortality change is $\varepsilon_c * \psi$, which can be treated either as ε_c fold change on a specific cause of death or as $\varepsilon_c * \psi$ fold change on all cause of death. This is because $LR_b=\psi=1$ for all cause of death.

On the condition that $\varepsilon_{c,t}$ and ψ_i are fixed across age, ε is used to denote the corresponding excess rate ratio on all cause of death. From the previous discussion, we could have the following equation:

$$\varepsilon = LR_b * \varepsilon_c \qquad (9)$$

Where ε denotes the constant excess rate ratio on all cause of death across age,

LR_b denotes the life time risk of a specific cause of death,

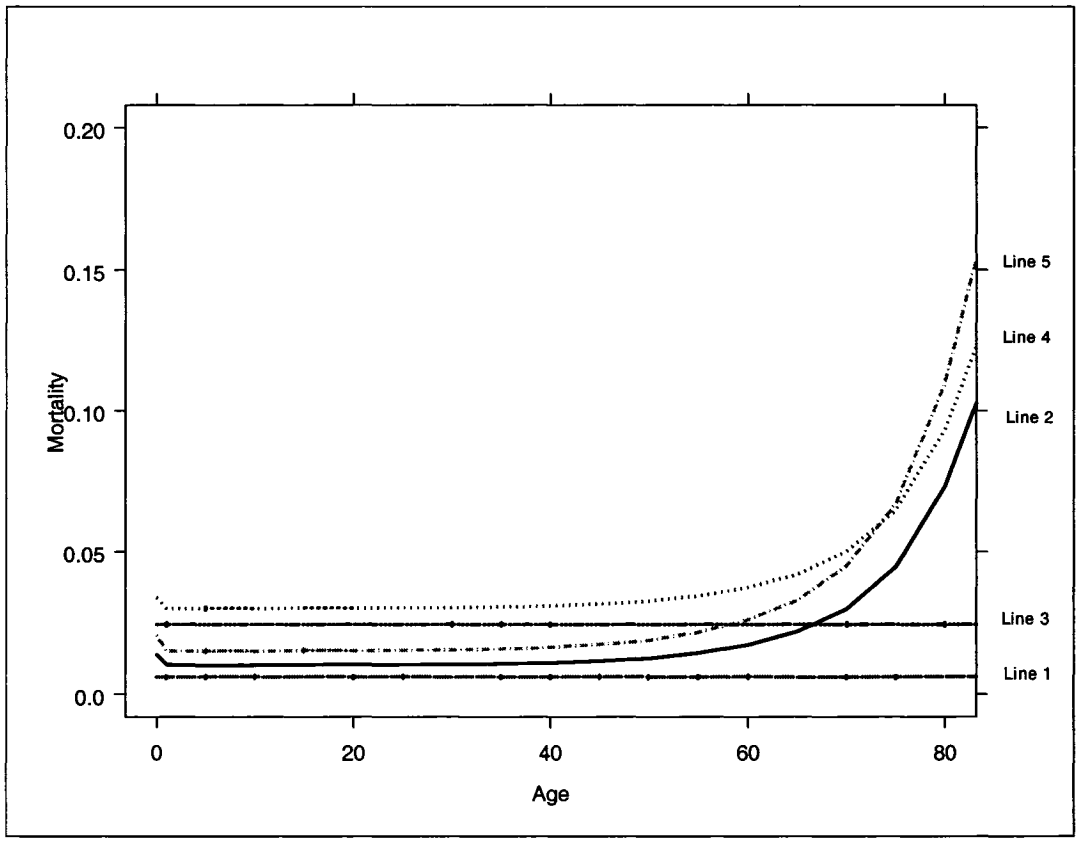
and ε_c denotes the constant excess rate ratio on a specific cause of death across age.

For the sake of easy presentation, ε_c is ERR which is more generally used on any specific cause of death and ε in this thesis represents the ERR on all cause of death.

In Figure 3, some examples are provided to show the constant additive modification and the constant multiplicative modification. The line denoted "Line 1" is an assumed case of constant baseline mortality rate. In reality, the baseline mortality can have any kind of shape.

Some causes of death that are almost independent of age such as car accidents, can be treated as approximately constant. The line denoted "Line 2" is the baseline mortality for a specific cause of death. Let's look at some modified examples applied on these two baseline mortalities.

Figure 3 Mortality modification examples



Note: This plot shows some examples of modification to mortality. Line 1 is the constant baseline mortality across age and Line 2 is the baseline mortality of a specific cause of death. Line 3 is the result of a constant additive/multiplicative modification applied on Line 1. Line 4 is the result of a constant additive modification applied on Line 2. Line 5 is the result of a constant multiplicative modification applied on Line 2.

The line denoted "Line 3" is the mortality plot from a constant mortality modification applied on "Line 1". The involved modification can be treated either as additive or multiplicative. When you treat it as an additive modification, the net change is $\Delta M = 3 * M_b$, where M_b is the baseline mortality represented by "Line 1"; when you treat it as a multiplicative modification, the constant relative change can be described by excess rate ratio ε_c =3. A modification can be treated

as either additive or multiplicative is a special situation. The reason is that the baseline mortality is constant across age.

The line denoted "Line 4" is the mortality plot from a mortality modification applied on "Line 2". The involved modification is additively constant over age. The line denoted "Line 5" is the mortality plot from another mortality modification applied on "Line 2". This modification is multiplicatively constant across age, which the excess rate ratio ε_c =0.5.

In the case of a multiplicative modification, some researchers applied the DEALE even though the DEALE was originally designed for the additive modification. The following steps show how the DEALE can be adapted for application involving multiplicative modifications. Firstly the constant baseline mortality (Average mortality, M_D) will be calculated, which uses the reciprocal of the life-expectancy. The constant excess mortality rate can be calculated by $\Delta M = \varepsilon * M_D$, ε is used because M_D is for all cause of death. In the case that a specific cause of death in the modified case, $\Delta M = \varepsilon_c * LR_b * M_D$ is used to calculate ΔM. The modified life-expectancy (LE_M) is obtained by $1/(M_D+\Delta M)$. The accuracy is the main concern of this process and will be explored.

2.3.3. Existing Approximation Approaches

2.3.3.1.DEALE

In 1982, Beck et al. [2, 3] presented a simple method, known as the "DEALE" method, to approximate the life-expectancy for an individual patient. The DEALE stands for Declining Exponential Approximation of Life Expectancy. It is based on the assumption that the survival curve follows a simple declining exponential function on time. This would apply if the annual of mortality rate μ is constant. In this case, the survival function can be expressed as:

$$S_t = e^{-\mu t} \qquad (10)$$

Where S_t denotes the probability to survive to age t,

μ denotes the constant mortality rate,

and t denotes the age.

In this case the baseline life-expectancy (LE) is the inverse of mortality rate μ:

$$LE=1/\mu \text{ year} \qquad (11)$$

The application of the DEALE is illustrated by solving Example A:

In Example A, the LE_b is 52.25 years for an individual at age 30 from the Life Table (in Section 2.2.1). This value is commonly tabulated by age, gender, and population.

Using Equation (11),

$$\mu=1/LE_b=1/52.25=0.0191 \text{ per year}$$

Example A is a multiplicative modified case and the relative change is constant (ε_c=5). To apply the DEALE, it is translated into the term used in an additive modification. Equation (8) and (9) are used to calculate ΔM. LR_b of the brain cancer at age 30 is 0.0056 from Equation (1).

$$\Delta M = \varepsilon^*\mu = LR_b^*\varepsilon_c^*\mu = (0.0056)^*(5)^*(0.0191)=0.00054$$

The modified overall mortality $\mu_M = \mu + \Delta M = 0.01964$, and hence the modified life expectancy $LE_M = 1/\mu_M = 1/0.01964 = 50.92$ years by using Equation (11) again.

In Table 2, the approximate health indices LE_M and LYL calculated by the DEALE are compared to those from the Life Table, which is one of the exact approaches.

Table 2 Life-expectancy indices approximated for Example A by the DEALE

INDEX	LIFE TABLE	DEALE
LE_b	52.25	52.25
LE_M	51.74	50.92
LYL	0.51	1.33

Note: The DEALE assumes the baseline life-expectancy is always available. The typical scenario is that the life-expectancy age-stratified tabulation is within user's arm. In this table, the baseline life-expectancy is calculated from the Life Table. N/A stands for "not applicable".

This approach is simple to use but the accuracy is not good when age is young and the excess mortality rate is between 0 and 0.05 (presented in the DEALE paper Figure 5).

2.3.3.2.IPH Method

This is a new method developed at Institute of Population Health, University of Ottawa [7]. The starting point of this approach is cause-modified life-table (CMLT). Brand used the Taylor series expansion to approximate the CMLT [7].

This method applies the Taylor series expansion to simplify Equation (4) and finally results in a very simple expression as below:

$$LE_M^t \approx LE_b^t\left(1-\Lambda^t\varepsilon\right) \qquad (12)$$

Where LE_M^t denotes the modified life-expectancy at age t,

LE_b^t denotes the baseline life-expectancy at age t,

ε denotes excess rate ratio on all cause of death,

And Λ^t is the core part of this approach, which can be calculated in advance. The Λ^t can be calculated by,

$$\Lambda^t = E\left[\sum_{k=t}^{i-1}\lambda_k\right] \qquad (13)$$

Where t denotes start age, k and i denotes the age groups,

E denotes a weighted average of the term inside [], where the weights are calculated as,

$$S_i\left(1-q_i^*\left(1-\alpha_i\right)\right)\omega_i \qquad (14)$$

Where ω_i denotes the number of years in the i-th age interval,

q_i^* denotes the probability of death in the i-th age interval,

α_i denotes the average proportion of the i-th age interval lived by those who die in that age interval,

and S_i denotes the cumulative survival.

λ_k can be calculated by

$$\lambda_k = \frac{q_k^* \psi_k}{1 - q_k^*} \qquad (15)$$

Where k denotes the k-th age interval,

ψ_k denotes the fraction of deaths attributed to a specific cause of death in the k-th age interval, which can be estimated by D_k^c / D_k (D_k^c denotes the deaths for a specific cause of death in the k-th age interval. D_k denotes the deaths for all cause of death in the k-th age interval),

and q_k^* denotes the probability of death in the k-th age interval.

In Example A, the Λ^t is calculated, which is 0.216, and the LE_b is 52.25 years for an individual age at 30 from the Life table. LR_b of the brain cancer at age 30 is 0.0056 from Equation (1). From Equation (9) and (12), LE_M = 52.25*(1-0.216*5*0.0056) = 51.93 years.

Table 3 lists the health indices calculated from the Life Table, the DEALE, the IPH and the Keyfitz approach for comparison purpose. LYL=0.51 is the true value from our discussion.

Table 3 Life-expectancy indices approximated for Example A by the different approaches

INDEX	LIFE TABLE	DEALE	IPH	KEYFITZ
LE_b	52.25	52.25	52.25	52.25
LE_M	51.74	50.92	51.93	51.91
LYL	0.51	1.33	0.32	0.34

Note: The DEALE, IPH and Keyfitz approach assumes the baseline life-expectancy is always available, which is calculated from the Life Table in this case. N/A stands for "not applicable".

The IPH method is easy to use. Based on Table 3, the accuracy of this approach is better than the DEALE in Example A.

The IPH and Keyfitz approaches are nearly identical but the Keyfitz approach is more elegant and can be shown to be at least as accurate as the IPH approach (although the difference

in accuracy between the two approaches is generally not that large [7]). The Keyfitz approach is chosen to compare with the Extended Keyfitz Model instead of the IPH approach.

2.3.3.3.Keyfitz method

In 1977, Nathan Keyfitz [5] developed an approximation approach. This approach also uses the Taylor series expansion. The following equation is the result of Keyfitz's efforts:

$$LE_M^t = LE_b^t(1 - H_c^{(1)}(t) * \varepsilon_c) \qquad (16)$$

Where LE_M^t denotes the modified life-expectancy at age t,

LE_b^t denotes the baseline life-expectancy at age t,

ε_c denotes excess rate ratio on a specific cause of death,

and $H_c^{(1)}(t)$ is a characteristic number derived by Keyfitz in his approach. It is cause-specific and age-specific. The superscript indicates it is the first order parameter from the Taylor series expansion. It is the core part of this approach, which can be calculated in advance.

$H_c^{(1)}(t)$ is defined as below:

$$H_c^{(1)}(t = a) = \frac{-\int_a^w S(t)Ln(S_c(t))dt}{\int_a^w S(t)dt} \qquad (17)$$

Where S(t) is the survival probability to age t,

$S_c(t)$ is the survival probability to age t under the modification, which can be calculated by Equation (5),

w denotes the upper limit of age,

a denotes the starting age.

From the definition, $H_c^{(1)}(t)$ is a function of age t and cause of death. It is the first order parameter during the derivation of this approach. The high order parameters are presented in Appendix C. When $H_c^{(1)}(t)$ is for all cause of death, ε will be used instead of ε_c, then Equation (16) has the same form as Equation (12) for the IPH except $H_c^{(1)}(t)$ is calculated differently from Λ^t.

In Example A, the $H_c^{(1)}(t)$ for all cause of death can be calculated from Equation (17), which is 0.2311, and the LE_b is 52.25 years for an individual age at 30 from the Life table. The LR_b of brain cancer is 0.0056 at age 30 from Equation (1). From Equation (9) and (16), LE_M = 52.25*(1-0.2311*5*0.0056) = 51.91 years. The results have already been listed in Table 3.

Suppose the LE_b^t and $H_c^{(1)}(t)$ are always available, then the Keyfitz approach is easy to use. The accuracy of this approach is good as well but will deteriorate when modification of mortality is high. It is a very similar approach as the IPH method, except that the characteristic parameter $H_c^{(1)}(t)$ is calculated differently from Λ^t.

2.3.3.4.New DEALEs

In 1992, Emmett Keeler and Robert Bell proposed five variants of the DEALE [4]. These five refinements have better results than the DEALE in most cases but sometimes their performance is worse. In addition, they are relatively complicated. Please refer to Appendix D for the details.

3. Initiative of new approximation approach

The DEALE is a popular approximation approach. More than 230 papers cited DEALE as reference from the website "Web in Knowledge". Many of them applied DEALE to their

application contexts. This is one of the evidences why we need the approximation approach. The performance of this approach was evaluated in the DEALE paper and was found good enough for the additive modifications when involving increases in mortality. However, no one has evaluated the accuracy of the DEALE when applying it in the multiplicative modifications. Despite this absence of evaluation, some papers have applied the DEALE in a multiplicative mode without first testing its performance [13, 14, 15, 16]. In addition, the accuracy of the DEALE in the negative modification to mortality is unknown. In this thesis, we show that the DEALE actually performs unreliably in the multiplicative cases. Especially, it has bad performance in the negative multiplicative modifications.

The performance of the Keyfitz approach deteriorates when the net mortality change is high. In his paper [5], Keyfitz only intended his approach to be applied for small incremental modifications to mortality.

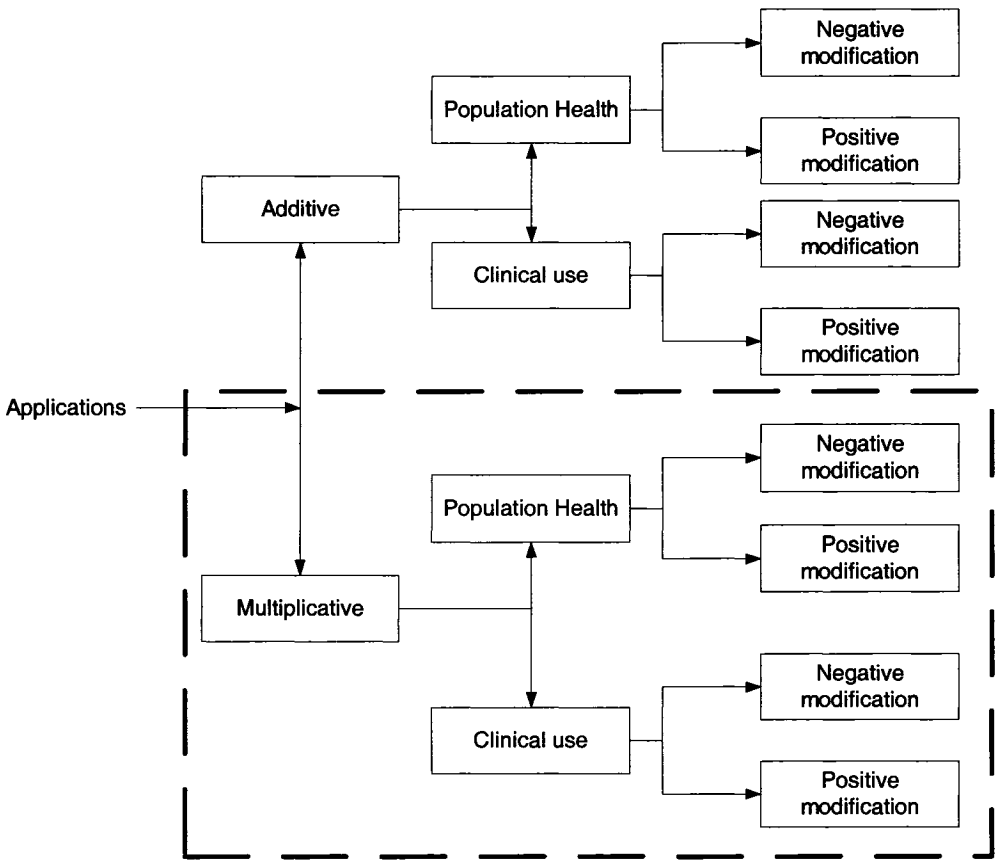
These aforementioned disadvantages of the DEALE and the Keyfitz approach motivated us to explore the possibility of a new better approach.

In reality, the impact of the modification to mortality usually will be studied in the clinical field or population health. Population health is currently the primary beneficiary. However, to study the impact of the modification to mortality in clinical field seems to have sufficient appeal. It has been used increasingly in this field to help clinical decision making.

To clarify the goal of the new approach, Figure 4 shows a map of the modification to mortality. In Figure 4, the DEALE can be applied everywhere. The DEALE actually works very well in the additive models with positive modification to mortality. It works well even when such modifications to mortality are large. Because it works so well it seems unwise to try to improve upon it in these application contexts. The purpose of this thesis is to fill a gap where the DEALE does not work well, namely multiplicative cases or when the multiplicative modification to

mortality is negative (e.g. the impact of health interventions). The new approach is aiming at performing better than the DEALE, the IPH and the Keyfitz approach for most applications lying within the dashed rectangle in Figure 4.

Figure 4 A map of the modification to mortality



Note: The DEALE can be applied everywhere in this map. It has good performance in the additive modification but performs uncertainly in the multiplicative modification. When the multiplicative modification is positively small or negative, it has bad performance. The new approach is trying to outperform the DEALE in the dashed rectangle area.

4. Derivation of the Extended Keyfitz Model

4.1.Data Source

For the purpose of the comparison and deriving a convenient formula for the EME (Established market economics) region based on the new approach, some public data sources are sought. Here is the data source used in this thesis: the life-tables for 191 Countries, which are available on a WHO website [17].

These life-tables are intended to represent the mortality rate schedules operating in year 2000 for each of these countries. For example, Table 4 lists one of the life-tables, which is for the Canadian female population. The WHO life-tables are organized by 191 countries and each country is separated into male and female populations.

Table 4 Life table for Canadian female population in year 2000

Age range	Actual Population	Actual Deaths	${}_nM_x$	${}_nq_x$
<1	156370	709	0.00453	0.00452
1-4	717870	154	0.00021	0.00086
5-9	992990	121	0.00012	0.00061
10-14	999870	130	0.00013	0.00065
15-19	1002200	299	0.00030	0.00149
20-24	1000580	297	0.00030	0.00148
25-29	1036960	328	0.00032	0.00158
30-34	1129870	515	0.00046	0.00228
35-39	1336990	947	0.00071	0.00354
40-44	1305780	1489	0.00114	0.00569
45-49	1165910	2085	0.00179	0.00890
50-54	1030530	2942	0.00285	0.01417
55-59	790650	3634	0.00460	0.02272
60-64	644440	4729	0.00734	0.03603
65-69	593930	6980	0.01175	0.05708
70-74	548100	10596	0.01933	0.09221
75-79	472690	15233	0.03223	0.14912
80-84	312070	17849	0.05720	0.25020
85-89	189710	18782	0.09900	0.39681
90-94	76200	12736	0.16714	0.55661
95-99	20230	5567	0.27521	0.70091
100+	2930	1295	0.44196	1.00000

Note: Column "Age range" denotes different age group;
Column "Actual population" denotes the actual population in each age group;
Column "Actual deaths" denotes the actual deaths in each age group;
Column "${}_nM_x$" denotes average mortality rate from age x to x+n;
Column "${}_nq_x$" denotes the probability that an individual alive at exact age x will die prior to age x+n.

4.2. ε range estimation

As mentioned before, ε has a practical range. Since the new approximation approach is devised for the multiplicative modifications, it is necessary to understand this practical range. To reiterate, in this thesis ERR on a specific cause of death is denoted as ε_c and ERR on all cause of death is denoted as ε. And ERR on a specific cause of death (ε_c) can be approximated as ERR on all cause of death (ε) by $\varepsilon = \varepsilon_c * LR_b$. Since it is more convenient to use ε to describe the extent of multiplicative modification to mortality, ε will be used in the new approximation approach and its performance evaluation.

The ε range is quite different in population health and clinical setting. In population health domain, I roughly define it from -0.5 to 2 [18] on all cause of death based on the literature. Here are some examples in Table 5:

Table 5 Examples of ERR in Population Health

Cause of death	**Risk Factor**	**ε_c**		**ε**	
		Women	**Men**	**Women**	**Men**
Stroke [18]	Body Height	0.73/5cm	0.84/5cm	0.088/5cm	0.101/5cm
Cardiovascular disease [18]	Body Height	0.93/5cm	0.82/5cm	0.246/5cm	0.216/5cm
Breast cancer [13]	Breast cancer screening	-0.3		-0.007	
Lung Cancer [6]	Radon	0.02~0.25		0.0003~0.0038	

Note: The ε_c on stroke increases 0.84 for men per 5 cm increase in height
The ε_c on stroke increases 0.73 for women per 5 cm increase in height
The ε_c on cardiovascular disease increases 0.82 for men per 5 cm increase in height
The ε_c on cardiovascular increases 0.93 for women per 5 cm increase in height
The ε_c on breast cancer is -0.3 by breast cancer screening on ESRD population
The LRs for Stroke, Cardiovascular disease, Breast Cancer and Lung Cancer are 0.12, 0.264, 0.022 and 0.015 for the Canadian population in year 2000.

In Table 5, there are several examples. The ε_c of stroke-introduced death is 0.84 for men and 0.73 for women per 5cm increase in height. The ε_c of cardiovascular diseases introduced death is 0.82 for man and 0.93 for woman per 5cm increase in height. Breast cancer screening reduces 30% mortality on breast cancer [13]. The ε_c of lung cancer introduced by radon is in a range [0.02, 0.25] and varies by age [6]. These ε_c values have their corresponding ε values, which can be approximated by $\varepsilon_c * LR_b$. And the implied ε values are also shown in Table 5. The LR_b used in Table 5 are for the Canadian population in year 2000.

The authors of the DEALE paper intended to focus on the clinical use of their approach. In the clinical setting, the ε range could be estimated from the DEALE paper. As the DEALE is devised for additive modifications, the excess mortality rate ΔM needs to be re-expressed as ε value.

The maximum excess mortality rate ΔM in the DEALE paper is 0.25. In this extreme case, consider an individual at birth, his life-expectancy is roughly around 80 years (For all the populations in the EME region), which implies the average mortality rate (u) is 1/80. From the definition of ε, $\varepsilon = \Delta M/u = 0.25 / (1/80) = 20$ in this case. From this simple calculation, it is shown that a rough range of ε for the clinical use is from 0 to 20.

In summary, the ε range is roughly from -0.5 to 2 in population health setting and from 0 to 20 in clinical setting. Put them all together, the range of ε of interest is from –0.5 to 20. This practical range will be used to evaluate different approximation approaches in this thesis.

To show the difference (in the performance of the approximation approaches) between population health and clinical use setting, I define the small and moderate modification of mortality as from –0.5 to 2 times modification of the overall mortality. The extreme modification of mortality is defined as greater than 2 and less than 20 times modification of the overall mortality. The small and moderate range has most cases in reality although some clinical cases

are in the extreme range. The convenient formula for the EME region will be developed in the ε range of [-0.5, 20] in this thesis.

4.3. Related issues in derivation

4.3.1. Target population

The populations belonging to the WHO are grouped into eight regions. The eight demographic regions are Established market economies (EME), Formerly socialist economies of Europe (FSE), India, China, Other Asia and islands (QAI), Sub-Saharan Africa (SSA), Latin America and the Caribbean (LAC), Middle Eastern crescent (MEC) [19].

A convenient formula for the EME region will be provided based on the new approach. The populations in the EME region are chosen to make the new approach less dependent on some age-stratified tabulations. The simple relationship between $LE_b(t)$ and $LE_b(t=0)$ in these populations could potentially lead to more convenient approximation formula, where t denotes age.

All the aforementioned approximation approaches rely on some baseline health indices tabulated by age, for example baseline LE tabulated by age. If it is baseline LE tabulation, the first step of these approaches is to check this tabulation to get the baseline LE at the age of interest. If the age of interest is somewhere in between the tabulation, interpolation has to be done to get baseline LE at that age. This scenario has some limitations:

1. The age-stratified tabulations are not always accessible.
2. Possible interpolation has to be done, which complicates the process, and influence the accuracy.

If the $LE_b(t)$ could be expressed in terms of $LE_b(t=0)$ in a simple way, then the new approach could depend on $LE_b(t=0)$ only and the age-stratified tabulation is not necessary any more.

The mortality data from the populations in the EME region in year 2000 are used to obtain the convenient formula. It is necessary to know what these populations are. Table 6 lists all the countries in the EME region that have their life-tables available in the WHO database.

Table 6 Countries of the target populations

Andorra	**France**	**Luxembourg**	**Spain**
Australia	**Germany**	**Monaco**	**Sweden**
Austria	**Greece**	**New Zealand**	**Switzerland**
Belgium	**Iceland**	**Norway**	**United Kingdom**
Canada	**Ireland Isle of Man**	**Netherlands**	**United States**
Denmark	**Italy**	**Portugal**	
Finland	**Japan**	**San Marino**	

Note: This table lists all the countries in the EME region that have their life-tables available in the WHO database. There are 26*2 (countries * sex) life-tables in total.

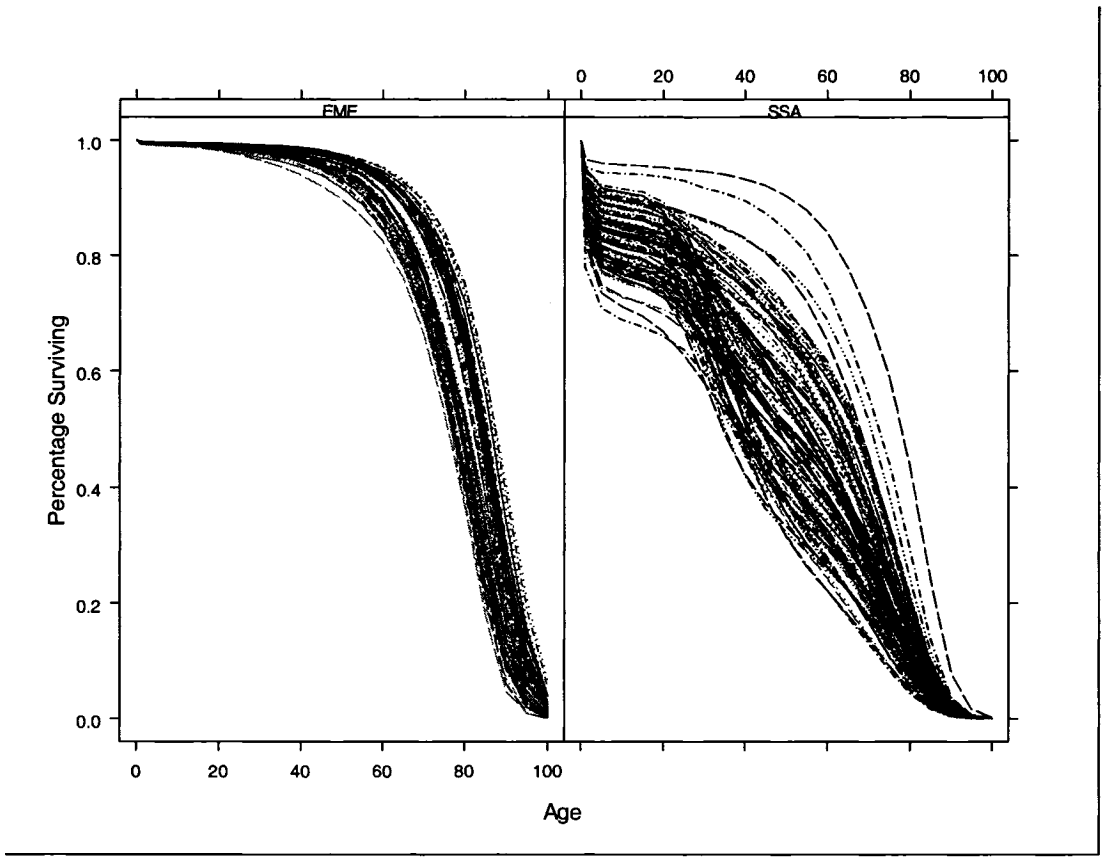
To show the difference between the EME region and the other region, SSA region is chosen as an example to contrast with the EME region. Table 7 lists the range of LE_b and $H^{(1)}$ at birth in the EME and SSA region for the population in year 2000. It shows that the range of the LE_b and $H^{(1)}$ is narrower in the EME region than that in the SSA region.

Table 7 LE_b and $H^{(1)}$ at birth in the EME region

	LE_b	$H^{(1)}$
Region	**At Birth** - *Mean (range)*	
EME	78.5 (71.7-84.7)	0.137 (0.111-0.171)
SSA	48.1 (37.0-72.3)	0.487 (0.183-0.725)

To understand the spectrum of mortality rate schedules for the EME populations, we can look at the survival curves of these populations. For the purpose to contrast this with the remaining populations in the WHO database, SSA region is chosen as an example to contrast with the EME region. Figure 5 shows the survival curves for the populations in the EME region in the left panel and the survival curves for the populations in the SSA region in the right panel.

Figure 5 Survival Curves in the EME and SSA region



From Figure 5, we know that the mortality schedules are quite different from region to region. In the EME region, the mortality schedules are quite close compare to those in the SSA region. As the LE_b is calculated from the survival curves, it is understandable that the LE_b in the EME region has a narrower range than that in the SSA region.

4.3.2. ε values in derivation

It is known that the practical ε range is within [-0.5, 20] from previous discussion. In this range, 20 values are chosen to calculate the LYL in derivation.

4.3.3. Age values in derivation

The age range is within [0, 80]. In this range, 9 age values are chosen to calculate the LYL. These values are 0, 10, 20, 30, 40, 50, 60, 70, and 80.

4.3.4. Other issues

- Source data are grouped by 5-year age intervals. Conventionally researchers use one-year or 5-year/10-year age intervals to construct the life-table. The life-table constructed with 1-year age intervals is called Complete Life Table and the life-table constructed with 5-year/10-year age intervals is called Abridged Life Table. The Abridged Life Table with 10-year age intervals is occasionally used. The Abridged Life Table with 5-year age intervals is chosen to be used in this thesis since (i) the life tables with one-year and 5-year age intervals come to very close value of life-expectancy and (ii) the life-tables from the WHO database are constructed with 5-year age intervals due to the data availability (Mortality data).
- The LYL relative error is defined as

  LYL Relative error = (approximate-truth)/truth

  Where approximate denotes the LYL calculated by the approximation approach, truth denotes the exact value of LYL as obtained by the exact approach, namely the cause modified life-table (CMLT).
- Variable t represents age throughout this thesis unless otherwise stated.

4.4. Preliminary analysis

As the first step, the special case the LYL at birth is studied.

To refine the Keyfitz approach (in Section 2.3.3.3), the rest of terms of the Taylor's expansion from the Keyfitz approach derivation are studied. The LYL can be expressed as below after applying the Taylor series expansion using all cause of death (drop subscript 'c'):

$$LYL = LE_b\left(H^{(1)}\varepsilon - H^{(2)}\frac{\varepsilon^2}{2!} + H^{(3)}\frac{\varepsilon^3}{3!} - H^{(4)}\frac{\varepsilon^4}{4!} + ...\right) \qquad (18)$$

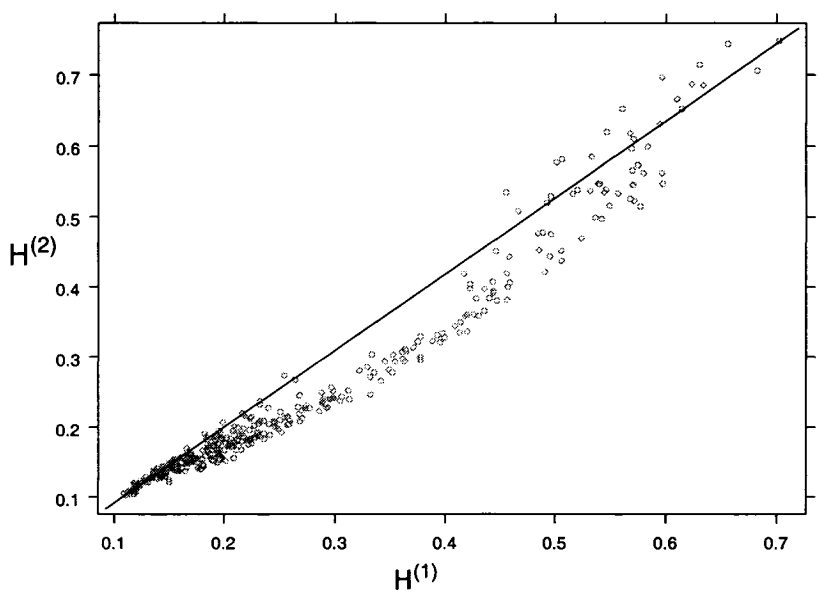
The Keyfitz approach only takes the first term. To improve the accuracy of the Keyfitz approximation, one strategy involves considering the Taylor series expansion beyond the first term. Theoretically the more terms included the better it would approximate the true value of LYL. Unfortunately increasing the number of terms undermines the objective of identifying a simple expression: each added term entails the need for an additional characteristic number ($H^{(n)}$). While such characteristic number could presumably be pre-calculated and tabulated for each population of interest, applications of the approximation would prove unwieldy, requiring the selection of multiple characteristic numbers for each country --- note that some preliminary investigations indicated that the expansion would have to be expanded significantly beyond five terms in order to attain adequate improvement in performance. In addition, each successive term in Equation (18) has ε and factorial term. It is complicated for an end user.

To avoid this unwieldiness while at the same time taking advantage of an extended Taylor series expansion, another strategy is explored: simplify the Equation (18). This is done in the hopes that Equation (18) might resemble the Taylor series expansion for a familiar function. If successful, our plan would identify a function, readily available on a calculator. The hope is that this familiar function would improve the approximation of LYL. The strategy to simplify Equation (18) relies one the existence of a simple relationship between the first and each of the high order coefficients as a scalar multiple of the first order coefficient. In this way, the first order coefficient can be pulled out of the series expansion (enclosed in the parentheses) entirely, leaving behind an expansion having no dependence upon the population of interest.

In Equation (18), $H^{(1)}$ and $H^{(2)}$ are first and second order parameters at birth for all cause of death. The $H^{(1)}$ is defined by Equation (17). The Keyfitz approach only takes the first term in Equation (18), which is $LE_b * H^{(1)} * \varepsilon$, to approximate the LYL. The general formula including all higher order terms ($H^{(n)}$) is listed and explained in Appendix C.

From Equation (18), if $H^{(1)}$ and $H^{(n)}$ have simple relation as $H^{(n)} = \vartheta_n * H^{(1)}$ for all n, the right expression in Equation (18) can change to $LE_b * H^{(1)} *(\ldots)$. Then there may be a chance to simplify the expression inside parenthesis. To explore the relation between $H^{(1)}$ and $H^{(2)}$, these two terms are calculated based on "all cause of death" and at birth for the 382 populations in the WHO database. Figure 6 shows the relation between $H^{(1)}$ and $H^{(2)}$ from these populations. In Figure 6, the straight solid line is the one to one line. Every point represents one population. It shows $H^{(1)}$ and $H^{(2)}$ are highly correlated and suggests an approximately linear relation. Although a non-linear model might fit these data better, in the interest of a simple relationship, the non-linear option will not be pursued.

Figure 6 The relation between $H^{(1)}$ and $H^{(2)}$



Note: All the points are close to the one to one line. It is plotted for all the populations in the WHO database.

A simple linear regression analysis has been conducted. $H^{(2)}$ acts as the response variable and $H^{(1)}$ acts as the predict variable. This analysis is conducted by forcing the regression line through the origin. The $H^{(3)}, H^{(4)}, H^{(5)}$, and $H^{(6)}$ are also calculated at birth based on "all cause of death" for all the populations in the WHO database. The study of the relations between these parameters shows they are also highly correlated. In all these analyses, simple linear regression models have been fit between these parameters and $H^{(1)}$ by forcing the line through the origin. The coefficient estimates are listed in Table 8:

Table 8 The Linear regression parameter estimates

	Slope (SE*)	**R-Square**	**Residual SE***
$H^{(2)} \sim H^{(1)}$	0.94 (0.005)	0.95	0.033
$H^{(3)} \sim H^{(1)}$	1.67 (0.011)	0.93	0.067
$H^{(4)} \sim H^{(1)}$	4.57 (0.031)	0.91	0.187
$H^{(5)} \sim H^{(1)}$	16.97 (0.124)	0.90	0.738
$H^{(6)} \sim H^{(1)}$	78.91 (0.634)	0.88	3.818

Note: $H^{(2)}$ - $H^{(6)}$ are the response variables and $H^{(1)}$ is the predict variable. The simple linear regression analyses have been conducted by forcing the regression line through the origin to obtain a simpler form. These analyses are based on the data of all the populations in the WHO database. SE* stands for Standard Error.

Based on the above proportional relationships, more generally, it shows a strong relation which can be expressed by:

$$H^{(n)} \approx \theta_n H^{(1)} \qquad (19)$$

Where θ_n s are constants.

Thus after substituting Equation (19) into Equation (18), the following equation is obtained:

$$LYL \approx LE_b H_1^{(1)} (\theta_1 \varepsilon - \theta_2 \frac{\varepsilon^2}{2!} + \theta_3 \frac{\varepsilon^3}{3!} + ... + (-1)^{(k+1)} \theta_k \frac{\varepsilon^k}{k!} + ...) \qquad (20)$$

Where LYL denotes Life-years lost,

LE_b denotes baseline life-expectancy,

ε denotes ERR on all cause of death,

θ_n is described in Equation (19) and partly listed in Table 7.

Equation (20) is based on all the populations in the WHO database and all cause of death.

4.5. New approach at birth

4.5.1. Candidate models

The expression form of the $LE_b * H^{(1)} *(\ldots)$ was obtained in the previous section (Equation 20). The next step is to explore the potential for simplifying the terms inside the parenthesis. Here are two candidate models:

4.5.1.1. Model 1

As we already know that $\theta_1 = 1$ by definition and $\theta_2 \approx 1$ by Table 7. When ε is reasonably small, the contribution of the higher order terms, relative to these first two terms, may be negligible. With this in mind it may be reasonable to assume $\theta_n \approx 1$ for all higher order terms. Equation (20) then changes to the following expression:

$$LYL \approx LE_b H^{(1)} (\varepsilon - \frac{\varepsilon^2}{2!} + \frac{\varepsilon^3}{3!} + \ldots + (-1)^{(k+1)} \frac{\varepsilon^k}{k!} + \ldots)$$

It is obvious that the same expression applying to Equation (21) according to Taylor's expansion,

$$LYL \approx LE_b H^{(1)} (1 - \exp(-\varepsilon)) \qquad (21)$$

This equation is simple and easy to use. It is taken as candidate model 1.

4.5.1.2.Model 2

As an alternative approach, we account for the best-fit values of θ_n. From the Table 7, the following equation can be obtained:

$$LYL \approx LE_b H^{(1)} (\varepsilon - \frac{\varepsilon^2}{2.13} + \frac{\varepsilon^3}{3.59} - \frac{\varepsilon^4}{5.25} + \frac{\varepsilon^5}{7.07} - \frac{\varepsilon^6}{9.12} + \ldots)$$

Inspection of the Taylor's expansion for Ln(1+ε) shows some resemblance to the above.

$$Ln(1+\varepsilon) = \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \frac{\varepsilon^4}{4} + \frac{\varepsilon^5}{5} + \ldots$$

Given this resemblance, we try the following equation:

$$LYL \approx LE_b H^{(1)} Ln(1+\varepsilon) \qquad (22)$$

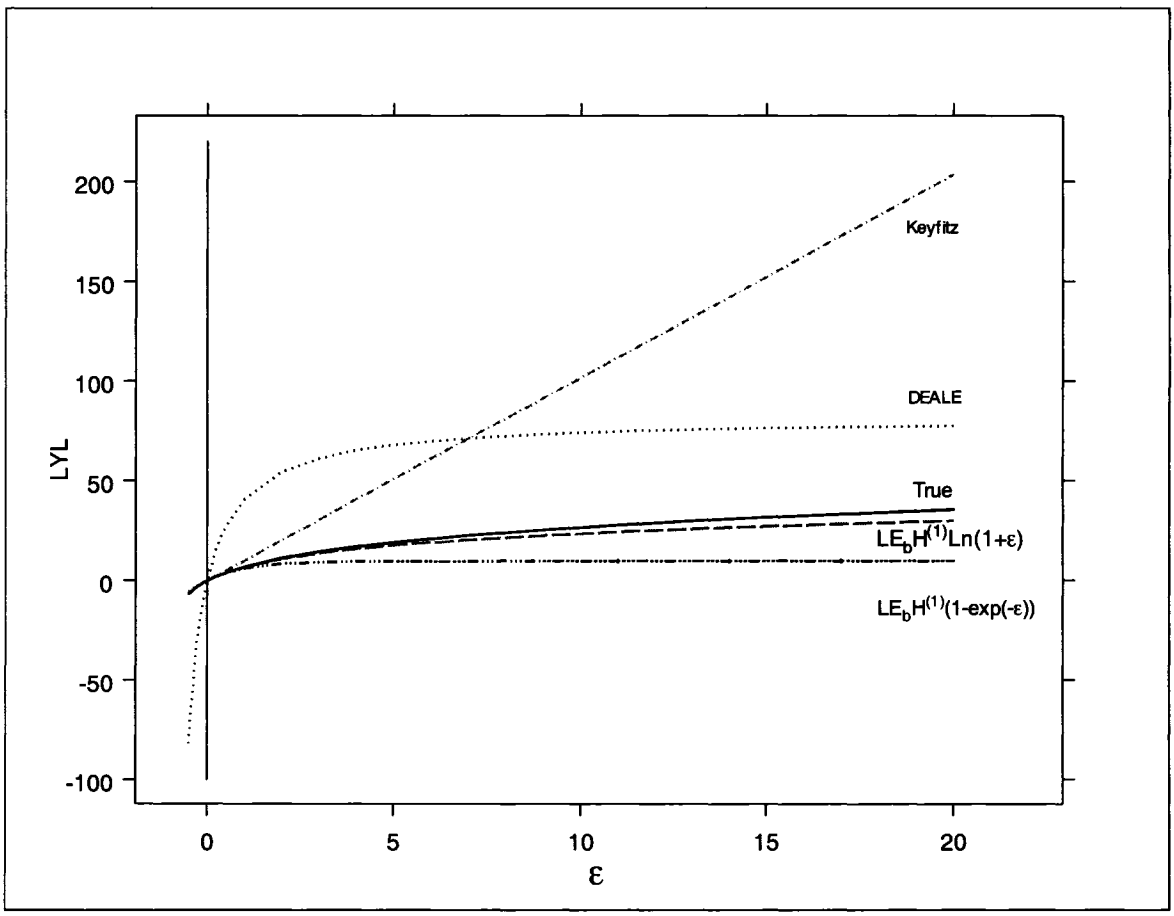
This equation is simple and easy to use as well. Equation (22) is taken as candidate model 2.

4.5.2. Evaluation of these two models

To compare the performance of these two models, the LYL are calculated by the Life Table (True value), the DEALE, the Keyfitz approach, candidate model 1 and candidate model 2. Figure 7 plots LYL as a function of ε for the particular case of Canadian female population in year 2000 at birth. A similar pattern to that seen in Figure 7 was found more generally across all the populations in the WHO database (Results not shown).

In Figure 7: The line denoted by "True" is the calculation result from the CMLT, which is the standard. The other lines denoted by "Keyfitz", "DEALE", "$LE_bH^{(1)}(1-\exp(-\varepsilon))$", "$LE_bH^{(1)}Ln(1+\varepsilon)$" represent LYL approximations based on the Keyfitz approach, the DEALE approach, candidate model 1 and candidate model 2 respectively.

Figure 7 Performance comparison between candidate model 1 and 2 at birth

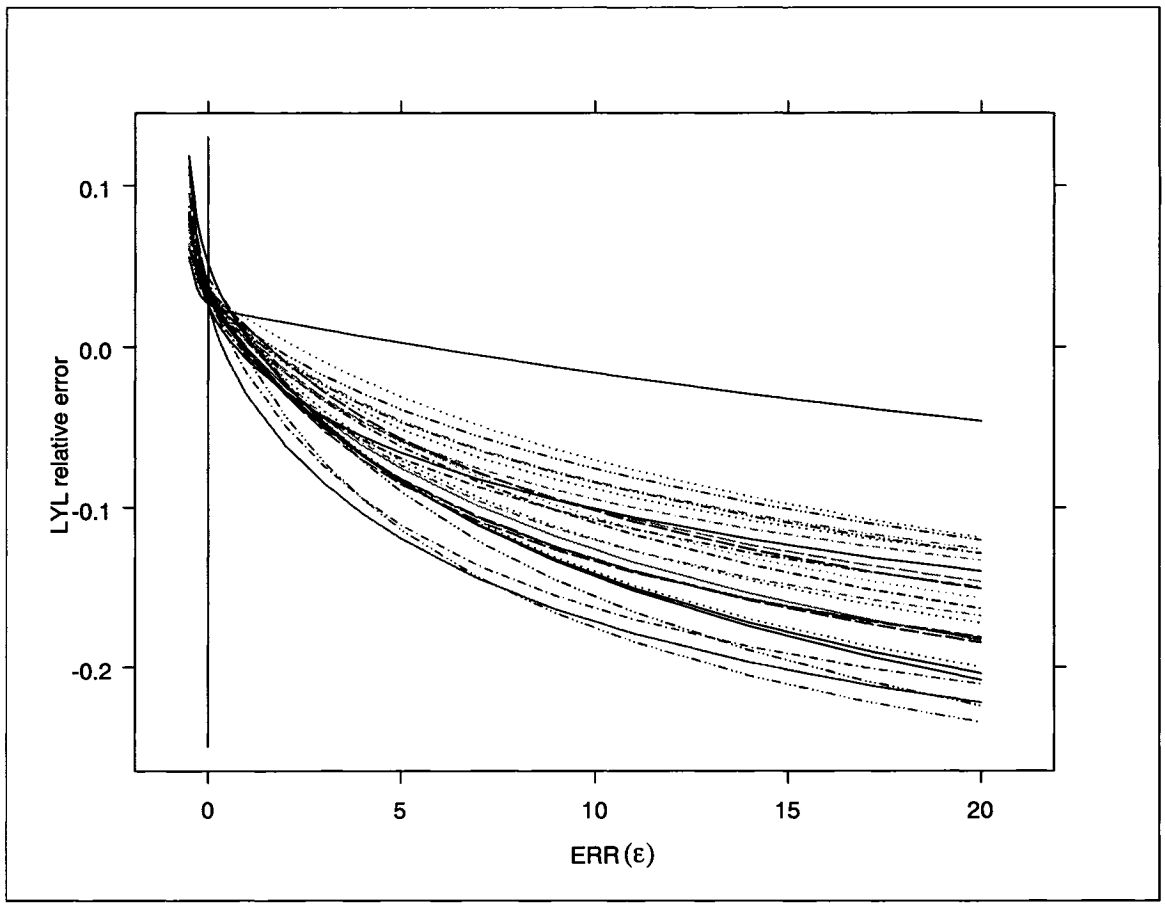


Note: This plot shows the LYL as a function of ε by the different approaches as well as the candidate model 1 and 2. It is plotted under the condition that age=0. The associated population is Canadian female population in year 2000.

Figure 7 shows that candidate model 2 (Dashed line) is systematically better than candidate model 1 at birth. More comparisons by other populations show it is a general result across all the populations in the WHO database. The limitation of the Keyfitz approach is also evident. In Figure 7, it is considered accurate only when $\varepsilon <= 2$.

Candidate model 2 has the best performance at birth. If this superior performance holds at any age, Equation (22) would be what we hoped for. Figure 8 shows its LYL relative error for all female populations in the EME region when age=0. The LYL relative error is defined in Section 4.3.4.

Figure 8 Performance evaluation of candidate model 2



Note: This plot shows the performance of model 2 (Equation 22) at birth. It describes the LYL relative error as a function of ε. This plot is specific for the female populations in the EME region at birth. The vertical line at $\varepsilon = 0$ separates the graph into two parts: The left part is for the negative modified cases and the right is for positive modified cases. This style will be used throughout this thesis to show the difference between the positive and negative multiplicative modifications.

Figure 8 shows the LYL relative error as a function of ε at birth for all the female populations in the EME region. Each line represents one population. For the case of Denmark, the LYL relative error is smallest, which stands out in Figure 8. The LYL relative errors for the male populations in the EME region show a similar pattern (Results not shown).

Table 9 lists the upper and lower bound values on the LYL relative errors in Figure 8 as a function of a couple of sample values for the ε.

Table 9 Upper and lower bound values in Figure 8

Age=0		
ε	Upper limit	Lower limit
-0.5	0.118	0.053
2	0.015	-0.061
20	-0.046	-0.234

Note: The absolute largest relative error is 23% in ε range [-0.5, 20] and 12% in ε range [-0.5, 2].

The largest error is about 23% when ε=20, which is a rare case. The typical error is within the range from –6% to 12% since most cases will involve ε in the range [-0.5, 2].

4.6. The Extended Keyfitz Model beyond Birth

4.6.1. Method 1

Equation (22) is only useful when estimating the LYL at birth. Thus its applicability is limited.

Most of the applications will have the factor age involved. To extend it to any age, a natural way of thinking is to think about the makeup of Equation (22) to check which term is age dependent. Upon replacing all the age dependent terms with the terms modified to depict the age dependency, Equation (22) becomes:

$$LYL_t \approx LE_b(t)H^{(1)}(t)Ln(1+\varepsilon) \qquad (23)$$

Where LYL_t denotes Life-years lost for those at age t,

$LE_b(t)$ denotes the baseline life-expectancy of an individual at age t,

$H^{(1)}(t)$ denotes the $H^{(1)}$ of an individual at age t and for all cause of death. It can be calculated by Equation (17).

In Equation (23), ε is not displayed as being dependent on age t because ε is presumed to be fixed across age in the approximation approach.

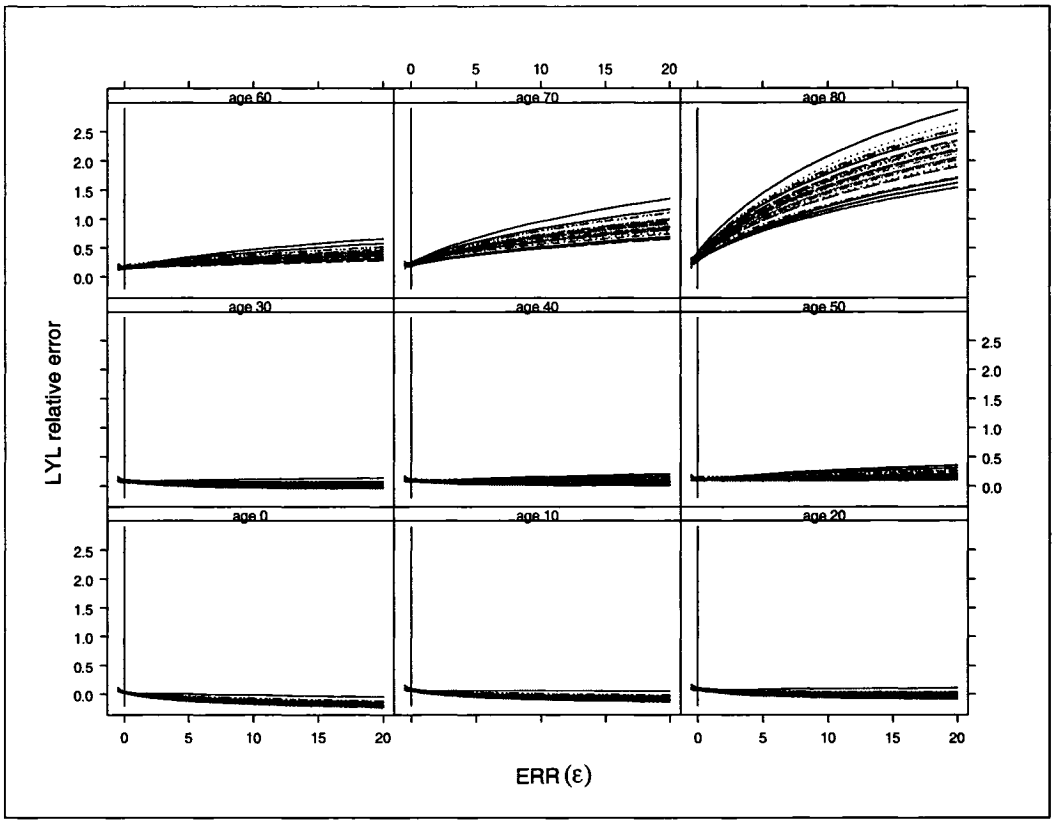
If $LE_b(t)$ and $H^{(1)}(t)$ are known in advance in Equation (23), then LYL at any age can be approximated. When you use other existing approximation approaches, they all assume some pre-existing tabulations. The tabulations of age-specific baseline life-expectancy $LE_b(t)$ are typically available for most populations. Because $H^{(1)}(t)$ is a less standard term, such pre-existing tabulations are less likely. Indeed even the $H^{(1)}(t=0)$ are not typically available. However, this Keyfitz characteristic number seems to have sufficient appeal that in the future, it may be more widely available, possibly even tabulated by age.

The evaluation of Equation (23) is conducted by presuming that tabulations of both LE and H exist as stratified by age. To evaluate how Equation (23) performs in this idealized circumstance the analysis that was done to prepare Figure 8 can be repeated, only now we need to examine the performance of our model for several cases (allowing the age for which LYL is computed to take on different values rather than that just exploring the LYL at birth, which is what Figure 8 did).

Just as in the analysis that produced Figure 8, the true LYL based on the CMLT as well as the approximate LYL by Equation (23) are calculated. This is done for all 52 populations in our EME database (The life-tables of the populations in the EME region from the WHO database), and the LYL relative errors are calculated. Figure 9 plots these relative errors for 9 different ages. The first represents the LYL relative errors for those at birth, the second represents the LYL relative errors for those of age 10, the third for those of age 20, and so on for the remaining ages (30, 40, 50, 60, 70 and 80).

Figure 9 shows the accuracy of this extension also holds for other ages. But the performance of the Equation (23) deteriorates on age 80. Equation (23) is called the Extended Keyfitz Model.

Figure 9 Performance of Equation (23)



Note: This plot shows the performance of Equation (23), in which $LE_b(t)$ and $H^{(1)}(t)$ are assumed to be known in advance. Each panel shows the performance under a specific age. In each panel, the LYL relative errors are plotted as a function of ε. The different lines in each panel represent the populations in the EME region.

The Upper and lower bound values in Figure 9 are listed by age in Table 10 as a function of a couple of sample ε values.

Table 10 The upper and lower bound values in Figure 9

	Upper limit (ε = -0.5)	Lower limit (ε = -0.5)	Upper Limit (ε = 2.0)	Lower Limit (ε = 2.0)	Upper limit (ε = 20)	Lower limit (ε = 20)
Age 0	0.118	0.053	0.015	-0.061	-0.046	-0.234
Age 10	0.159	0.077	0.058	-0.008	0.056	-0.139
Age 20	0.169	0.082	0.075	0.013	0.108	-0.086
Age 30	0.171	0.084	0.087	0.031	0.147	-0.042
Age 40	0.088	0.181	0.111	0.055	0.208	0.016
Age 50	0.199	0.101	0.167	0.093	0.354	0.111
Age 60	0.233	0.122	0.257	0.166	0.662	0.287
Age 70	0.274	0.143	0.446	0.284	1.354	0.654
Age 80	0.321	0.151	0.909	0.524	2.875	1.545

The largest error is about 288% when ε=20 and age=80. In that situation, an individual at age 80 would face 20 times modification of the overall mortality, which is a rare case. The typical error is within the range from –6% to 91% since most cases will involve ε in the range [-0.5, 2].

4.6.2. Method 2

4.6.2.1. Initiative

Equation (23) looks promising. However, to use Equation (23), first the $LE_b(t)$ and $H^{(1)}(t)$ must be obtained from the pre-existing tabulations. As mentioned in the previous section, while such tabulations exist for $LE_b(t)$, they do not (as far as we are aware) exist for $H^{(1)}(t)$. Furthermore, such tabulations even if in existence would be unlikely to be within end users' arm reach. It is more likely that the approximations would be applied in contexts where LE and $H^{(1)}$ are not known for an arbitrary age of interest, but only, most likely, just for the population/individual of interest at birth. For this reason, Equation (23) might only be helpful for examining the special case of age zero. Let's take a look at what kind of age-stratified tabulations the other approximation approaches need. Table 11 lists the age dependent indices required for different approximation approaches.

Table 11 Age dependent indices required for the approximation approaches

Approximation Approach	Age dependent indices required in advance
DEALE	LE, LR
New DEALEs	LE, LR
IPH	LE, LR, Λ_2
Keyfitz	LE, LR, $H^{(1)}$
Extended Keyfitz Model	LE, LR, $H^{(1)}$

Note: LR is used for projection from a specific cause of death to "all cause of death".

All of these approximation approaches seem to presume that the tabulations of the related indices stratified by age would be available.

In summary, none of the existing approximation approaches has a good way to settle this issue (as far as I know). Three models are proposed in the following sections as an attempt to reduce the data inputs. These models only apply for the populations in the EME region. The best model is selected and acts as an implementation of the Extended Keyfitz Model. This yields a convenient (less data dependent) and accurate approach, though it loses some generality (The result is only valid for the populations in the EME region in year 2000).

4.6.2.2. Candidate models

The Extended Keyfitz Model relies on age dependent terms for $LE_b(t)$ and $H^{(1)}(t)$. The purpose is to modify its expression to allow it to depend on LE_b and $H^{(1)}$ at birth only. In this section, three candidate models will be proposed. The first candidate model is to model $LE_b(t)$ by $LE_b(0)$, and model $H^{(1)}(t)$ by $H^{(1)}(0)$ separately. The second candidate model is trying to model the $LE_b(t)*H^{(1)}(t)$ as a single term by $LE_b(0)*H^{(1)}(0)$. The third candidate model is to use the "true" values from the Life Table to adjust the model, more details will be explained later.

4.6.2.2.1. Model 1

Look at Equation (23), the first intention is to study $LE_b(t)$ and $H^{(1)}(t)$ respectively. Because Equation (23) is based on an approximation that only requires $H^{(1)}(t)$, there will be no need to retain the superscript in the subsequent description --- all of the subsequent derivations and discussions shall refer only to the first order Keyfitz characteristic number, $H^{(1)}(t)$, which shall henceforth be denoted as $H(t)$ (dropping the superscript).

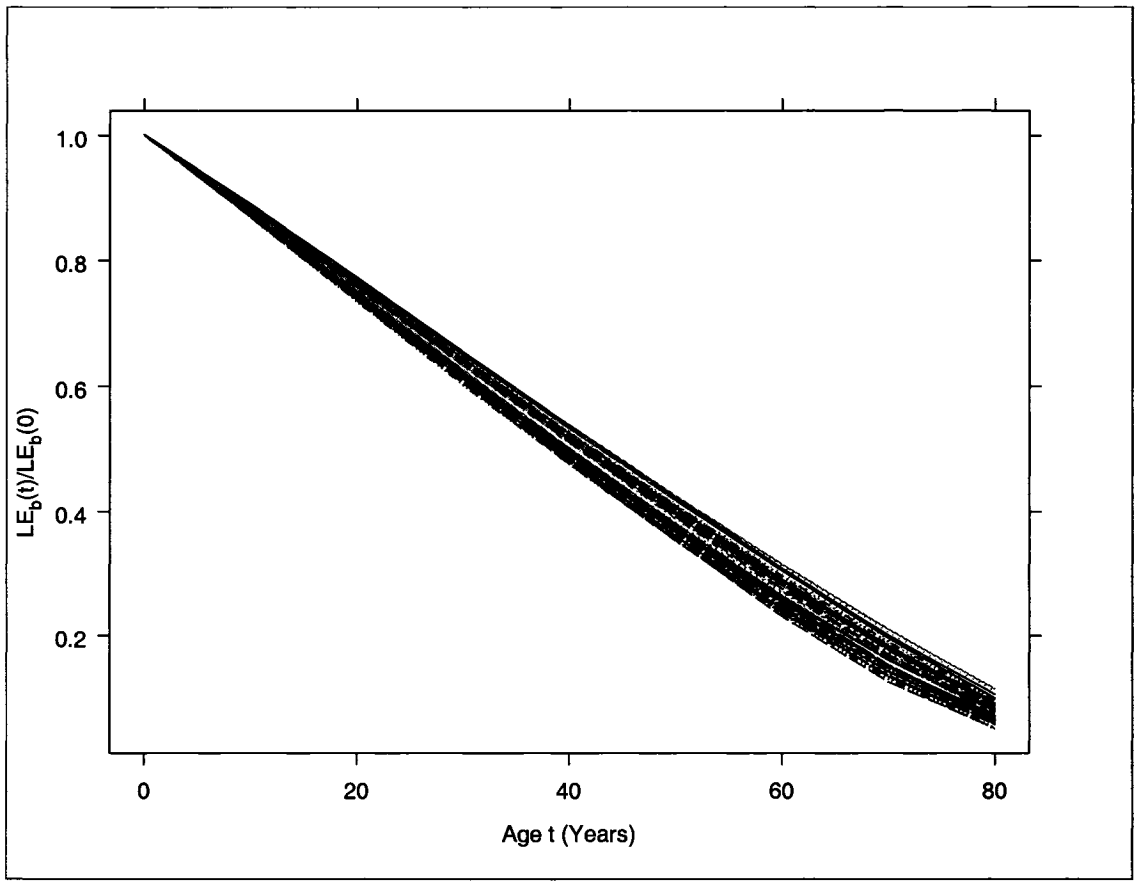
The goal is to express $LE_b(t)$ in terms of $LE_b(0)$ and $H(t)$ in terms of $H(0)$. In an effort to standardize, I examine $LE_b(t)/LE_b(0)$ and $H(t)/H(0)$ as a function of age to study if any pattern exists.

Figure 10 plots $LE_b(t)/LE_b(0)$ as a function of age t for all populations in the EME region. Every line represents one population in the EME region. The lines show a striking consistency; all show a similar linear decline. The average baseline life-expectancy at birth for the populations in the EME region is about 80. When age increases one unit, the remaining life-expectancy drops one year approximately.

To model this decline, there are two alternative ways.

1. The data points in every line are independent. A simple linear regression analysis is conducted for every population. A final linear function is obtained by averaging the coefficient estimates from all 52 regression lines.

Figure 10 The relation between $LE_b(t)/LE_b(0)$ and age t



Note: This plot shows the relation between $LE_b(t)/LE_b(0)$ and age t for all the populations in the EME region. Each line represents one population. It shows strong linear relation.

2. The mean life-expectancy on every age could be calculated first, and then fit a simple linear regression line.

The analysis results by these two methods are listed in Appendix F for reader's reference. It shows that all these two ways lead to similar regression lines. It is unclear what is the better way to conduct this kind of analysis. Because the results are close, I choose the analysis result from the first approach.

For this analysis, the $LE_b(t)/LE_b(0)$ is calculated by the Life Table and is organized in three dimensions: countries, sex and age. It yields 26*2*9 pairs of $LE_b(t)/LE_b(0)$ and age t for the analysis.

Each simple linear regression analysis has been conducted as follows: For each population, $Y=LE_b(t)/LE_b(0)-1$ is the dependent variable and age t is the independent variable. The simple linear regression has been conducted by forcing the line through the original since $Y=LE_b(t)/LE_b(0)-1=0$ when t=0 by definition. The final average coefficient estimates are listed in Table 12:

Table 12 Average coefficient estimates in modeling $LE_b(t)$

Slope (SE^*)	R-Square	Residual SE^*
-0.012 (0.0001)	0.999	0.0163

Note: SE^* stands for standard error. This table lists the results from the first approach. The coefficient estimates from both two methods are listed in Appendix F.

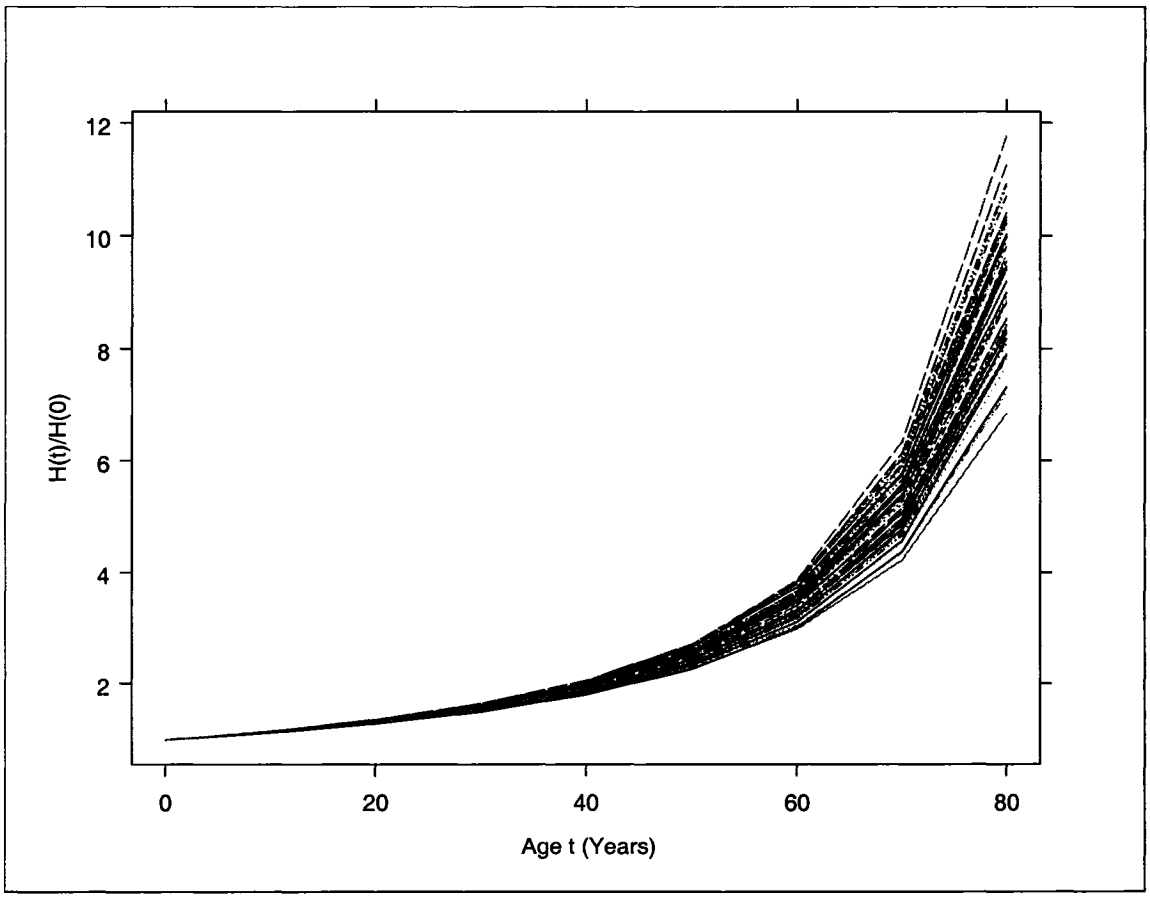
The result of the analysis with these coefficient estimates suggests that:

$$LE_b(t) = LE_b(0) * (1-0.012*t) \qquad (24)$$

It shows that 99.9% of the data variation is explained by Equation (24). This Equation allows a user to obtain the estimate of the remaining life-expectancy for an arbitrary age of interest, provided the user has a single input, namely $LE_b(0)$, available. We would like to be able to do the same for H(t). Specifically we would expect to express H(t) as a function of H(0).

Figure 11 plots H(t)/H(0) as a function of age t for all populations in the EME region. The 52 curves (one for each population in our EME database) show a similar pattern. In comparison to Figure 10, this pattern is somewhat less consistent (showing some inter-country spread as age increases).

Figure 11 The relation between H(t)/H(0) and age t

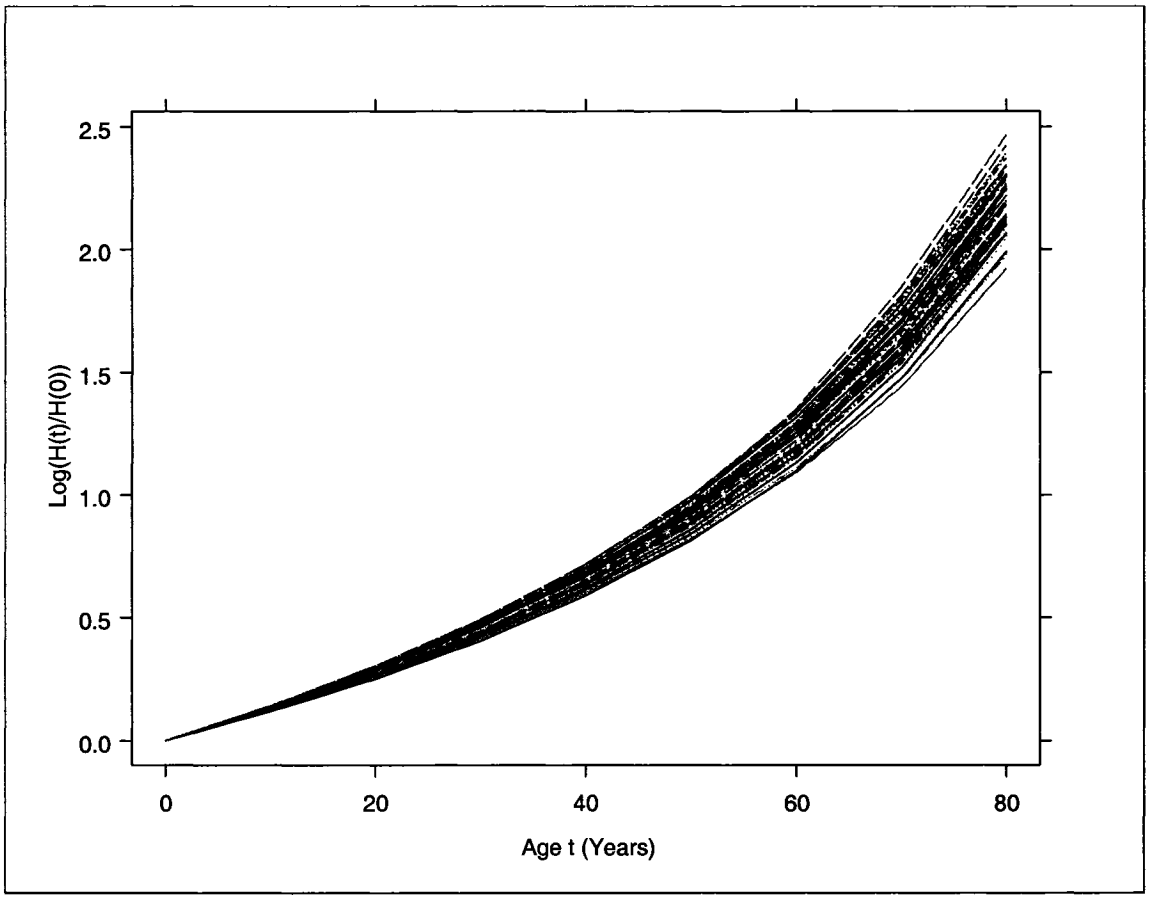


Note: The plot shows H(t)/H(0) as a function of age t. Every line represents a population in the EME region. It shows an exponential relation.

The relations are nonlinear, but have a clear pattern. A transformation is sought to "linearize" the relationship. Because the curves are reminiscent of an exponential relation, the following transformation is performed:

Let Y = Log(H(t)/H(0))

Figure 12 The relation between Log(H(t)/H(0)) and age t



Note: This plot shows Log(H(t)/H(0)) as a function of age t. Each line represents one population in the EME region. It shows an approximately linear relation.

Figure 12 plots composite variable Y as a function of age t. It suggests a linear relationship. (A nonlinear model appears to fit better. But the goodness of fit only changes a little comparing to the linear model. As simplicity is another goal of ours, we are not going to find the best-fit model in this case). There are two alternative ways to conduct this regression analysis:

1. Fit a simple linear regression line for every population and get final linear function by averaging the coefficients for 52 linear regression lines (52 populations in the EME region).
2. Calculate the mean Y on every age and then fit a simple linear regression line.

These two ways come to similar regression lines (Please refer to Appendix G). It is arguable what is the better way to conduct this kind of analysis. As they both come to the same result, the first way is chosen to conduct this analysis.

The Log(H(t)/H(0)) is calculated and organized in two dimensions: population and age. It yields 52*9 pairs of Log(H(t)/H(0)) and age t for the simple linear regression analysis.

For each population, a simple linear regression line has been fit: Y is the response variable and age t is the explanatory variable. The simple linear regression has been conducted by forcing the line through the origin since Log(H(0)/H(0))=0 by definition.

The average coefficient estimates are listed in Table 13:

Table 13 Average coefficient estimates in modeling H(t)

Slope (SE*)	R-Square	RESIDUAL SE*
0.0224 (0.0015)	0.965	0.2155

Note: SE* stands for standard error. The linear regression coefficient estimates between Log(H(t)/H(0)) and age t are obtained by forcing the regression line through the origin.

Using these estimates, the relation can be described by:

$$H(t) = H(0) * e^{0.0224*t} \qquad (25)$$

We now have an expression relating $LE_b(t)$ as a function of $LE_b(0)$ (Equation 24) and an expression modeling H(t) as a function of H(0) (Equation 25). Substituting these expressions into Equation (23) yields:

$$LYL(\varepsilon,t) = LE_b * H * Ln(1+\varepsilon) * e^{0.0224*t} * (1-0.012*t) \qquad (26)$$

Where ε denotes ERR on all cause of death,

$LYL(\varepsilon,t)$ denotes Life-years lost for a specific ε on age t,

LE_b denotes baseline life expectancy at birth,

H denotes the first order parameter in the Keyfitz approach on all cause of death at birth,

and t denotes age.

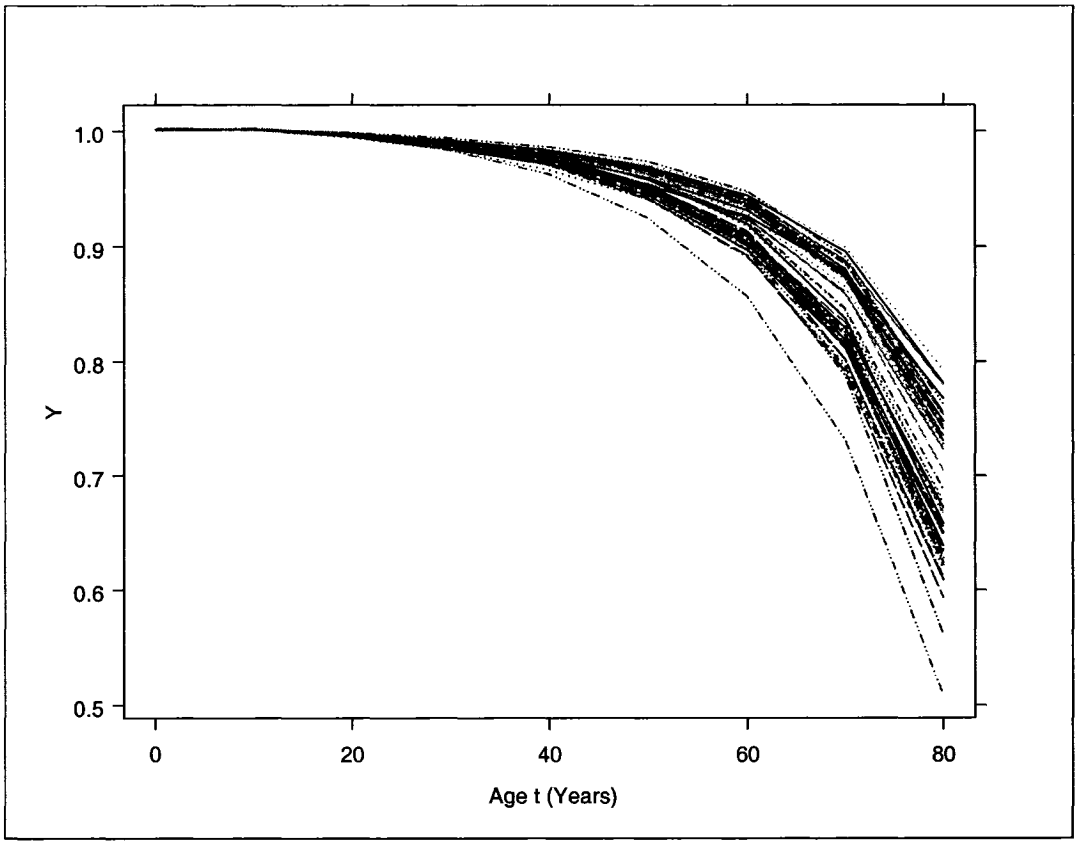
Equation (26) looks complicated, but it only requires two data inputs: $LE_b(0)$ and H(0). $LE_b(0)$ and H(0) are population specific. The remaining parameters (t and ε) are part of the problem specification. This equation can be applied to any age, and requires the same data inputs ($LE_b(0)$ and H(0)) regardless of age. The equation does however require the quick recall of two constants (0.0224 and 0.012), which could impede its usage.

4.6.2.2.2. Model 2

Equation (26) is somewhat complicated. This complexity leads us to consider the possibility of a simpler form.

Based on the understanding of the relation between $LE_b(t)$ and H(t) with age t, it may be better to model these two terms together as a composite variable. Defining this composite variable as Y(t) (where $Y(t)=LE_b(t)*H(t)$), it is expected to reduce the error and be more accurate by analyzing this composite variable. To explore the potential of this strategy, Figure 13 plots our new composite variable Y(t)/Y(0) as a function of age t.

Figure 13 The relation between Y (t)/Y(0) and age t



Note: This plot shows Y (t)/Y(0) as a function of age t, where $Y(t)=LE_b(t)*H(t)$. Each line represents one population in the EME region. It explores the possibility of modeling $LE_b(t)$ and H(t) together as a single term.

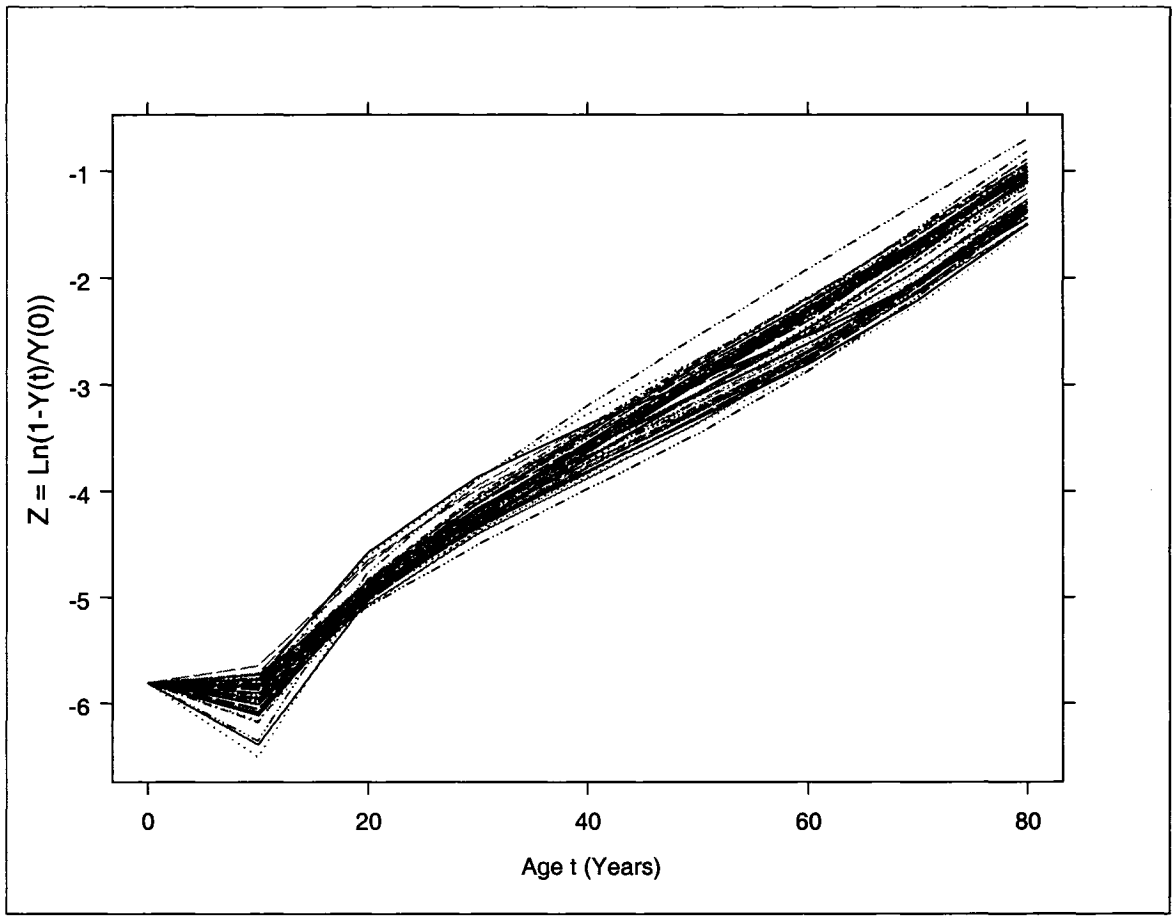
Figure 13 shows a clear nonlinear relation.

Just as in the case of the H(t) / H(0) plot (Figure 12), A transformation is sought to "linearize" the relationship in Figure12. Based on the shape of the curves in Figure13, the following transformation is performed:

$$\text{Let } Z = Ln(1-\frac{Y(t)}{Y(0)})$$

For the 52 populations, Z (as defined above) is calculated for all nine age values. Figure 14 plots these values as a function of age t, using different lines for the 52 populations.

Figure 14 The relation between Z and age t



Note: This plot shows Z as a function of age t, where Z=Ln(1-Y(t)/Y(0)). The strong linear relation suggests a simple linear regression. The deviations between age 0 and 20 are not considered to be of special concern.

The lines in Figure 14 display basically linear pattern, suggesting the linearization was effective. Although there are some distinct deviations to the linear pattern between age 0 and 20s, these deviations are not considered to be of special concern. There are two different ways to conduct this simple linear regression analysis which are same as just mentioned in Section 4.6.2.2.1. Since these two methods give the similar regression lines, the first method is chosen to conduct this analysis (Please refer to Appendix H).

For each population, a simple linear regression analysis is conducted: Z is the dependent variable and age t is the independent variable.

The average coefficient estimates are listed in Table 14:

Table 14 Average coefficient estimates in modeling composite variable Z

Slope (SE*)	Intercept (SE*)	R-Square	Residual SE*
0.0733 (0.001)	-6.94 (0.04)	0.962	0.29

Note: These estimates are for the simple linear regression between composite variable Z and age t

Applying an inverse transformation by these estimates yields:

$$LYL(\varepsilon,t) = LE_b * H * Ln(1+\varepsilon) * (1-e^{-6.94+0.0733*t}) \qquad (27)$$

Equation (27) is slightly more elegant than Equation (26). Equation (26) has two terms age related and Equation (27) only has one term age related. Thus the influence from age t is more straightforward by Equation (27) than Equation (26). When age goes up, you may feel easier to understand the trend by Equation (27) than Equation (26).

4.6.2.2.3. Model 3

The candidate model 1 and 2 are all from Equation (23). Equation (23) is an approximation itself that assumes the age dependent indices are available. In the case where all the age dependent indices are available, Using Equation (23) to approximate the LYL has some errors. In addition, there are some errors introduced by candidate model 1 or 2. In the Equation (26) and Equation (27), the term $LE_b(0) * H(0) * Ln(1+\varepsilon)$ is common and is not age related. The rest of the equations depends only on age. To reduce these two errors, also we want to take out the age irrelevant part(Equation 22), a composite variable Y is devised to meet this need.

Let $Y = LYL(\varepsilon,t)/\{LE_b(0) * H(0) * Ln(1+\varepsilon)\}$

Where ε denotes ERR on all cause of death,

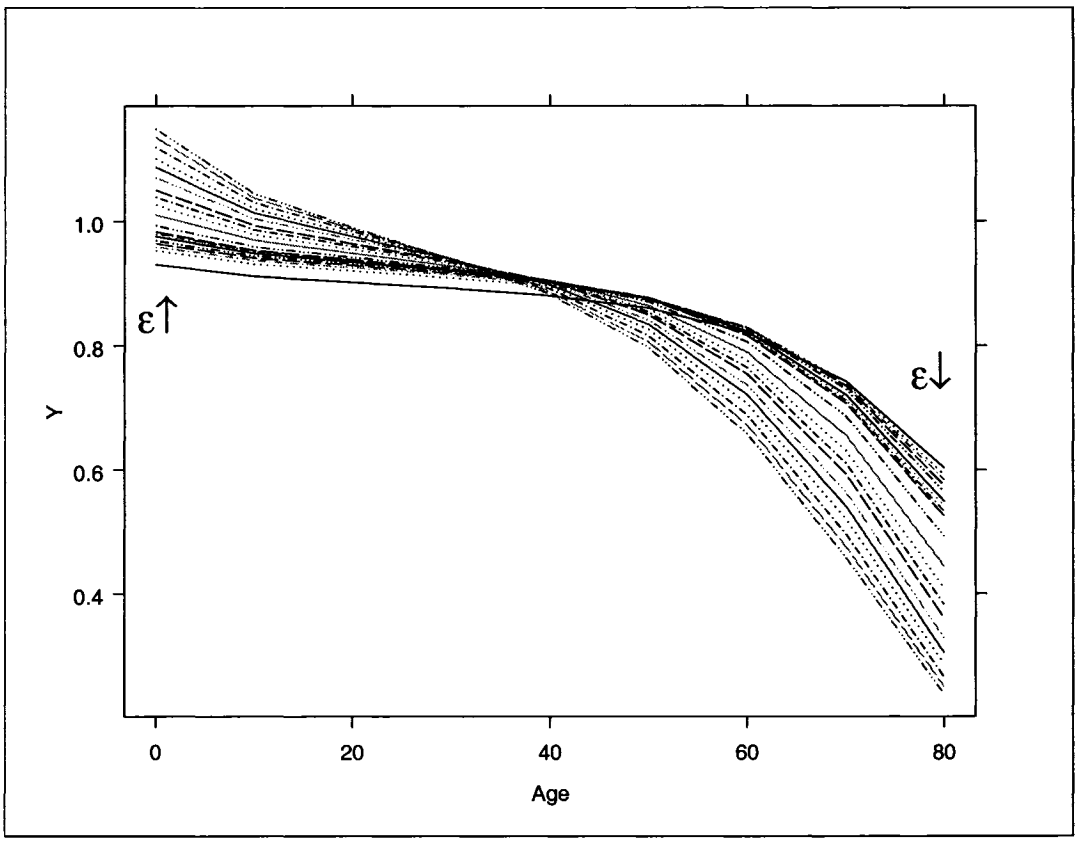
$LYL(\varepsilon,t)$ denotes Life-years lost on a specific ε and age t, which is calculated from the Life Table, and all the other terms as defined before.

To explore the potential of this strategy, the composite variable Y is plotted as a function of age t. Here is the data preparation process, which is different from model 1 and 2:

One more dimension, ε, is needed to calculate the Y (Defined above). The dimensions are: populations, age and ε. The sizes of these three dimensions are: 56 populations in our EME database, 9 age values and 20 ε values. The true value of LYL is calculated by the CMLT. The composite variable Y is calculated accordingly using the population specific H and LE.

Figure 15 plots the composite variable Y as a function of age t. It is specific for the Canadian female population in year 2000. Every line represents one specific ε value. The ε increases clockwise from the most flat line. All the other populations in the EME region have similar pattern as shown in Figure 15 (Results not shown).

Figure 15 The relation between composite variable Y and age t



Note: This plot shows the relation between $Y = LYL(\varepsilon, t) / \{ LE_b(0) * H(0) * Ln(1+\varepsilon) \}$ and age t. It is specific for Canadian female population in year 2000. Each line represents a ε value in the range [-0.5, 20].

In Figure 15, all the lines appear nonlinear and a transformation needs to be done. If a constant subtracts all the values of Y, the new pattern is very close to an exponential relation. The new relation could be described as

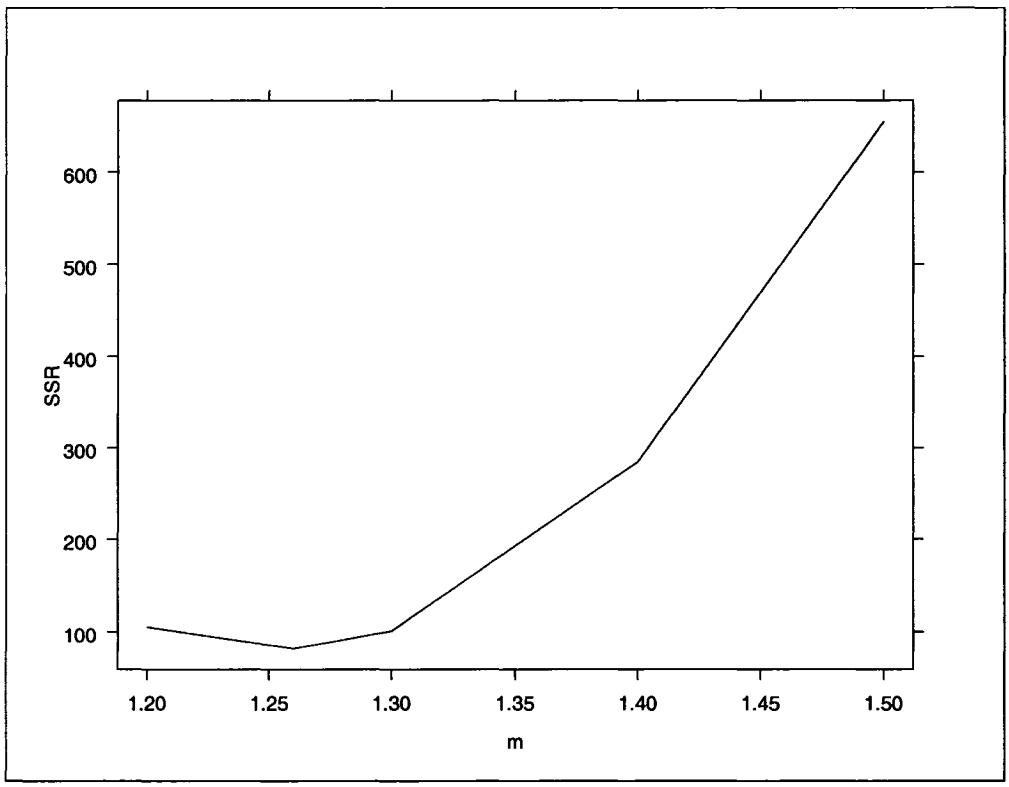
$$Y = m - \exp(a * t + b)$$

Where m, a, and b are constants.

There are three parameters that need estimation. It can be started from any value of m. Each m value will enable us to do an analysis: Let r = m - Y ; The next step is to fit an exponential model to get a pair of estimates (a and b) for each population. The value of a and b will be averaged across populations. An optimized value of m was found by achieving smallest model error (Sum of squared residuals, SSR).

Figure 16 describes the relation between SSR and m.

Figure 16 The relation between SSR and m



Note: This plot shows the relation between sum of squared residuals (SSR) and m. An optimized m value will achieve lowest SSR.

The optimized solution has the following average estimates across populations:

Table 15 Estimates in Model 3

	Slope	Intercept	m
Estimate	0.0214454	-1.8099500	1.2643700
SE*	0.000391231	0.038280600	0.007949570
T value	54.8152	-47.2812	159.0490

Note: The regression analysis in model 3 achieves lowest SSR when m=1.26437. The estimates in the table are obtained from Splus function: nls. SE* stands for the average standard error.

Applying an inverse transformation by these estimates yields:

$$LYL(\varepsilon,t) = LE_b * H * Ln(1+\varepsilon) * (1.264 - e^{-1.81+0.0214*t}) \quad (28)$$

Equation (28) looks similar as Equation (27) (Model 2). It has the same level of simplicity as Model 2. The structure of the Equation (27) and Equation (28) are same: There are three constants in Equation (27): 1, -6.94 and 0.0733; Comparing to the three constants in Equation (28): 1.264, -1.81 and 0.0214.

Before evaluating their performance, it is useful to review the difference between these three models:

In Table 16, these three candidate models are summarized without the mathematical analysis details.

Equation (27) and (28) are quite similar, and they are different from Equation (26). Let's do a simple comparison when age=0.

Table 16 A brief summary for three candidate models

Model	Equation	Description
Model 1	(26)	Model $LE_b(t)$ and H(t) separately
Model 2	(27)	Model $LE_b(t)$ and H(t) together
Model 3	(28)	Using Equation (23) at birth and the true LYL by the CMLT to find a model

When age=0, Equation (26) reduces to

$$LYL(\varepsilon) = LE_b * H_1 * Ln(1+\varepsilon)$$

Equation (27) reduces to

$$LYL(\varepsilon) = 0.999 * LE_b * H_1 * Ln(1+\varepsilon)$$

Equation (28) reduces to

$$LYL(\varepsilon) = 1.1 * LE_b * H_1 * Ln(1+\varepsilon)$$

Equation (26) reduces to Equation (22) (Equation 23 at birth) when age=0. And the reduced form of Equation (27) is next closest to Equation (22). The reduced form of Equation (28) is a little far from Equation (22). The reason for this difference is because Equation (26) and (27) start from Equation (23). If these models are proper, they should reduce to Equation (22) when age=0. Equation (28) is using the true LYL by the Life Table to find a model based on Equation (22). The reduced form of Equation (28) might be different from Equation (22).

4.6.2.3. Comparison between these candidate models

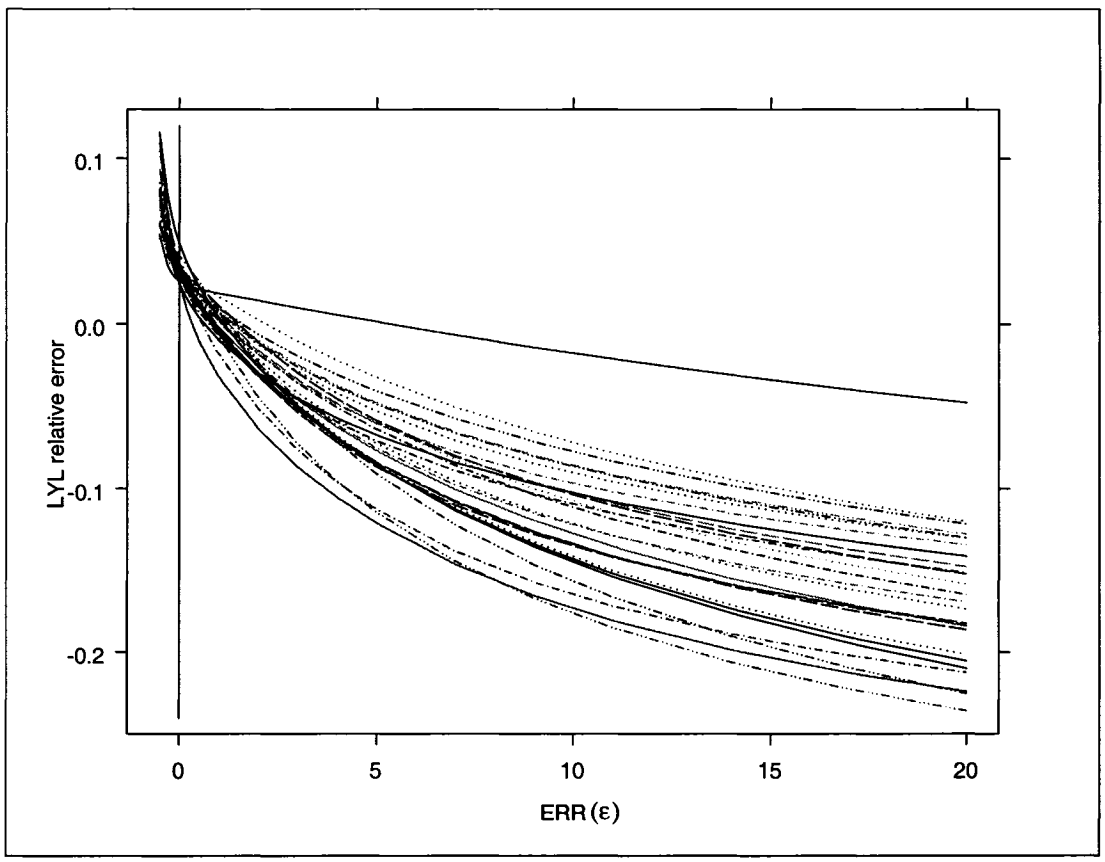
This section evaluates the three candidate models by comparing their associated LYL relative errors. The female populations in our EME database are chosen to present these relative errors. The male populations in our EME database have the same pattern (Results not shown).

Let's look at these relative errors in model 1 first.

- **Performance of model 1**

The LYL relative error at birth for the female populations in the EME region by model 1 is presented in Figure 17. Figure 17 is identical to Figure 8.

Figure 17 The performance of Model 1 at birth



Note: This plot evaluates the performance of model 1 when applying the model to the female populations in the EME region at birth. The LYL relative errors are plotted as a function of ε.

Data preparation process for Figure 17:

There are two dimensions: population (26 female populations in the EME database) and ERR (20 ε values). The exact LYLs at birth are calculated by the CMLT and the approximate LYLs at birth are calculated by Equation (26). The relative errors can then be calculated and organized in the same dimensions. These data are then grouped by populations for plotting purpose. Each population has 20 pairs of LYL relative error and their corresponding ε values.

The upper and lower bound values of the LYL relative error (in Figure 17) are listed in Table 17 as a function of a couple of sample ε values. Table 17 is identical to Table 8.

Table 17 The upper and lower bound values in Figure 16

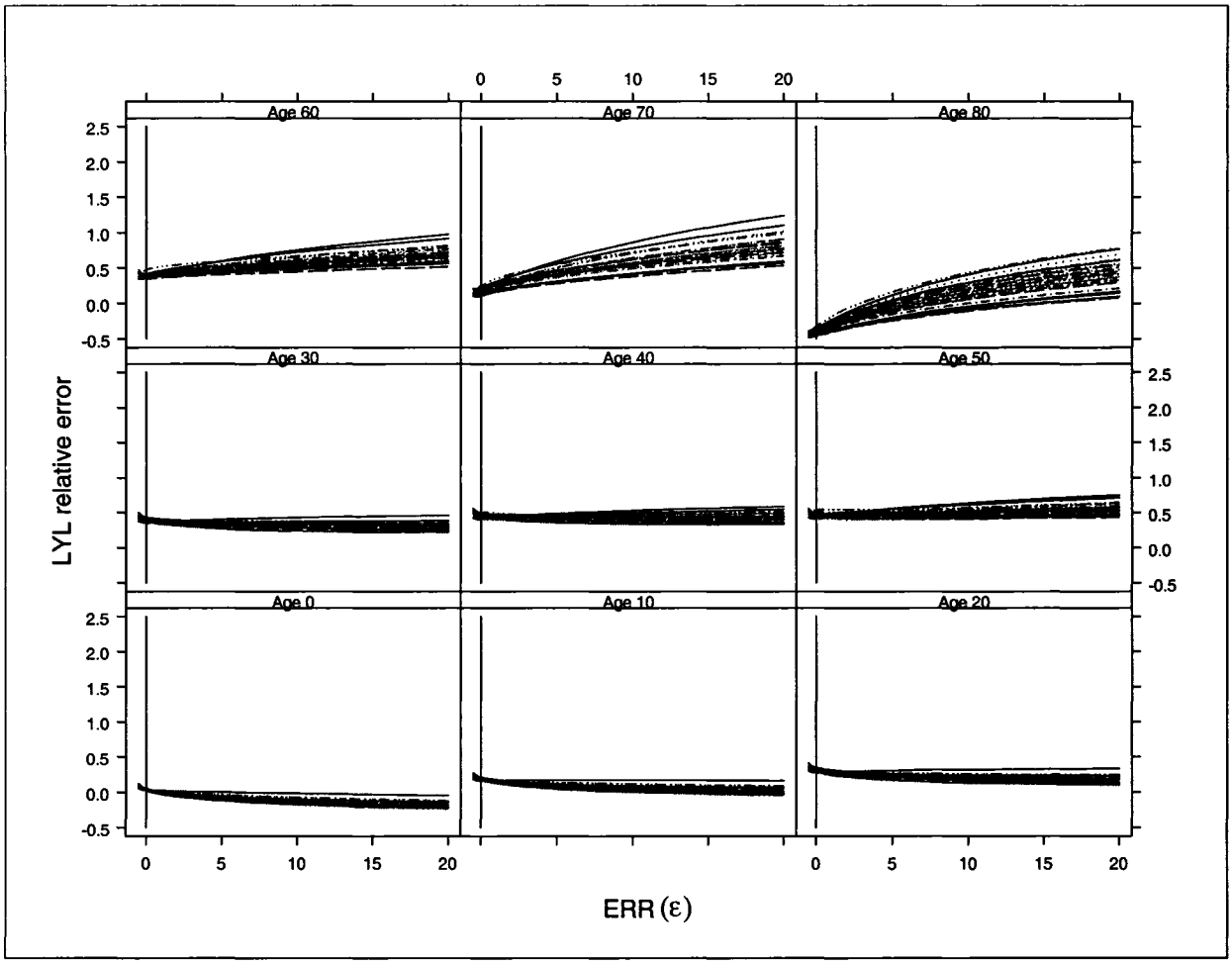
	Upper limit (ε = -0.5)	Lower limit (ε = -0.5)	Upper Limit (ε = 2.0)	Lower Limit (ε = 2.0)	Upper limit (ε = 20)	Lower limit (ε = 20)
Age 0	0.118	0.053	0.015	-0.061	-0.046	-0.234

Note: The upper and lower bound values when ε=2 is presented to show the performance in Population Health

The maximum relative error is about 23% and appears when ε=20. When ε is in the range [-0.5, 2], the maximum relative error is 12%. The results for the male populations have a similar pattern (Results not shown).

Figure 17 shows LYL relative errors for the special case of age=0 (ie. for populations at birth). We are interested in the performance of model 1 cross the range of possible ages. Figure 18 addresses this by plotting 9 panels, where each panel shows the performance at different age (0, 10, 20, ..., 80). The panel corresponding to age=0, is actually identical to Figure 18.

Figure 18 The performance of Model 1 at different ages



Note: This plot shows the performance of model 1 as a function of ε at 9 different ages (0, 10, ...). The performance is getting worse when age is 80.

The upper and lower bound values of LYL relative error are listed in Table 18 as a function of a couple of sample ε values.

The largest relative error is 125% when ε=20 and age=70. When we confine our attention to the ε in the range [-0.5, 2], the largest relative error is 56%.

Table 18 The upper and lower bound values in Figure 18

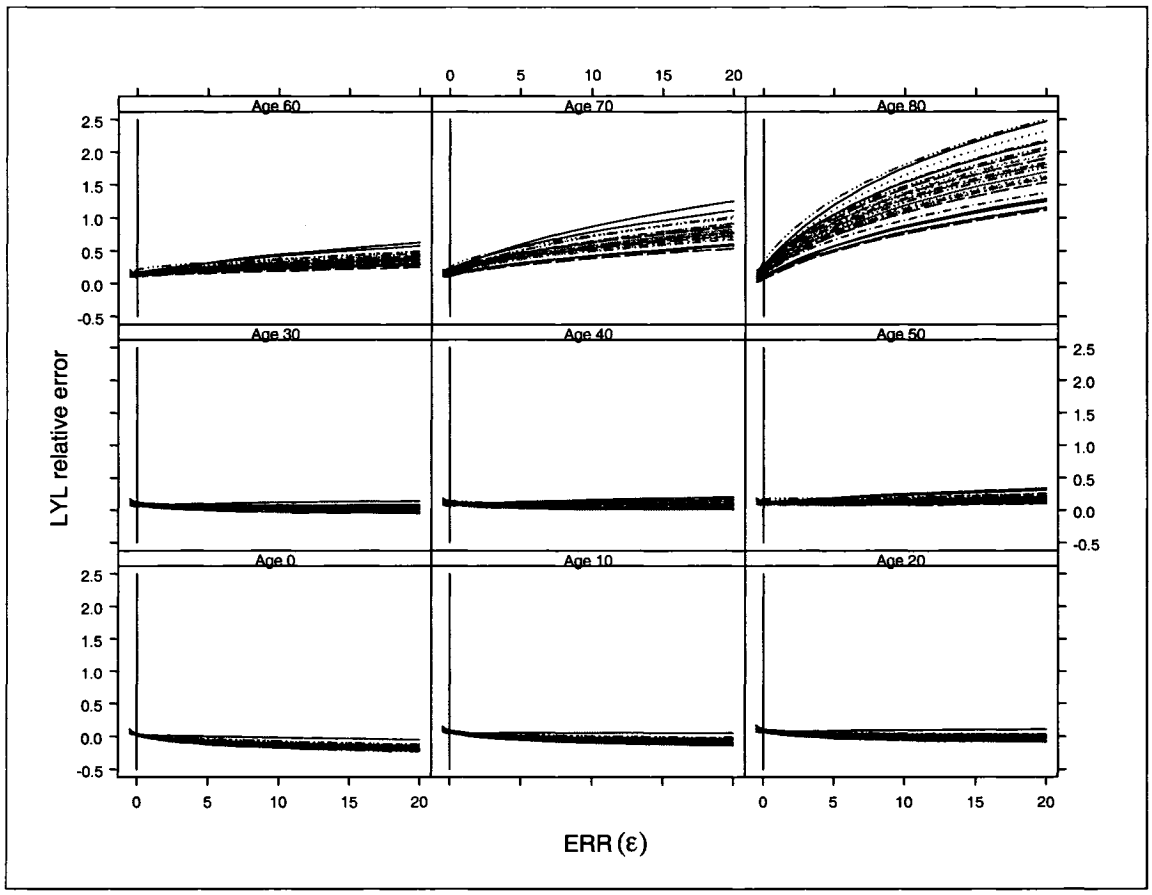
	Upper limit (ε = -0.5)	Lower limit (ε = -0.5)	Upper Limit (ε = 2.0)	Lower Limit (ε = 2.0)	Upper limit (ε = 20)	Lower limit (ε = 20)
Age 0	0.116	0.052	0.014	-0.063	-0.047	-0.236
Age 10	0.276	0.186	0.166	-0.092	0.164	-0.051
Age 20	0.406	0.297	0.291	0.214	0.328	0.096
Age 30	0.503	0.383	0.396	0.315	0.463	0.222
Age 40	0.562	0.431	0.482	0.389	0.589	0.338
Age 50	0.559	0.430	0.548	0.412	0.754	0.439
Age 60	0.469	0.341	0.549	0.378	0.977	0.522
Age 70	0.207	0.106	0.415	0.197	1.246	0.534
Age 80	-0.386	-0.476	-0.076	-0.353	0.787	0.079

Note: The LYL relative errors are calculated for the female populations in the EME region in year 2000.

- **Performance of model 2**

Let's study the performance of model 2, the LYL relative errors are presented in Figure 19. The panels are arranged in the same style as Figure 18.

Figure 19 The performance of Model 2 at different ages



Note: This plot shows the performance of model 2 at different ages (0, 10, ...). The performance is getting worse when age is 80.

Data preparation for Figure 19 is similar as Figure 18 except that the approximation equation is Equation (27). The upper and lower bound values of LYL relative errors at different ages are listed in Table 19 as a function of a couple of sample ε values.

Table 19 The upper and lower bound values in Figure 19

	Upper limit (ε = -0.5)	Lower limit (ε = -0.5)	Upper Limit (ε = 2.0)	Lower Limit (ε = 2.0)	Upper limit (ε = 20)	Lower limit (ε = 20)
Age 0	0.115	0.051	0.013	-0.064	-0.048	-0.236
Age 10	0.154	0.073	0.054	-0.012	0.052	-0.141
Age 20	0.170	0.080	0.075	0.011	0.106	-0.087
Age 30	0.178	0.084	0.094	0.030	0.146	-0.042
Age 40	0.186	0.086	0.125	0.054	0.206	0.015
Age 50	0.194	0.095	0.185	0.081	0.343	0.102
Age 60	0.209	0.103	0.275	0.134	0.627	0.252
Age 70	0.210	0.108	0.418	0.199	1.250	0.537
Age 80	0.204	0.027	0.811	0.268	2.504	1.117

The largest relative error is 250% when ε=20 and age=80. When narrow ε in the range [-0.5, 2], the largest relative error is 81%.

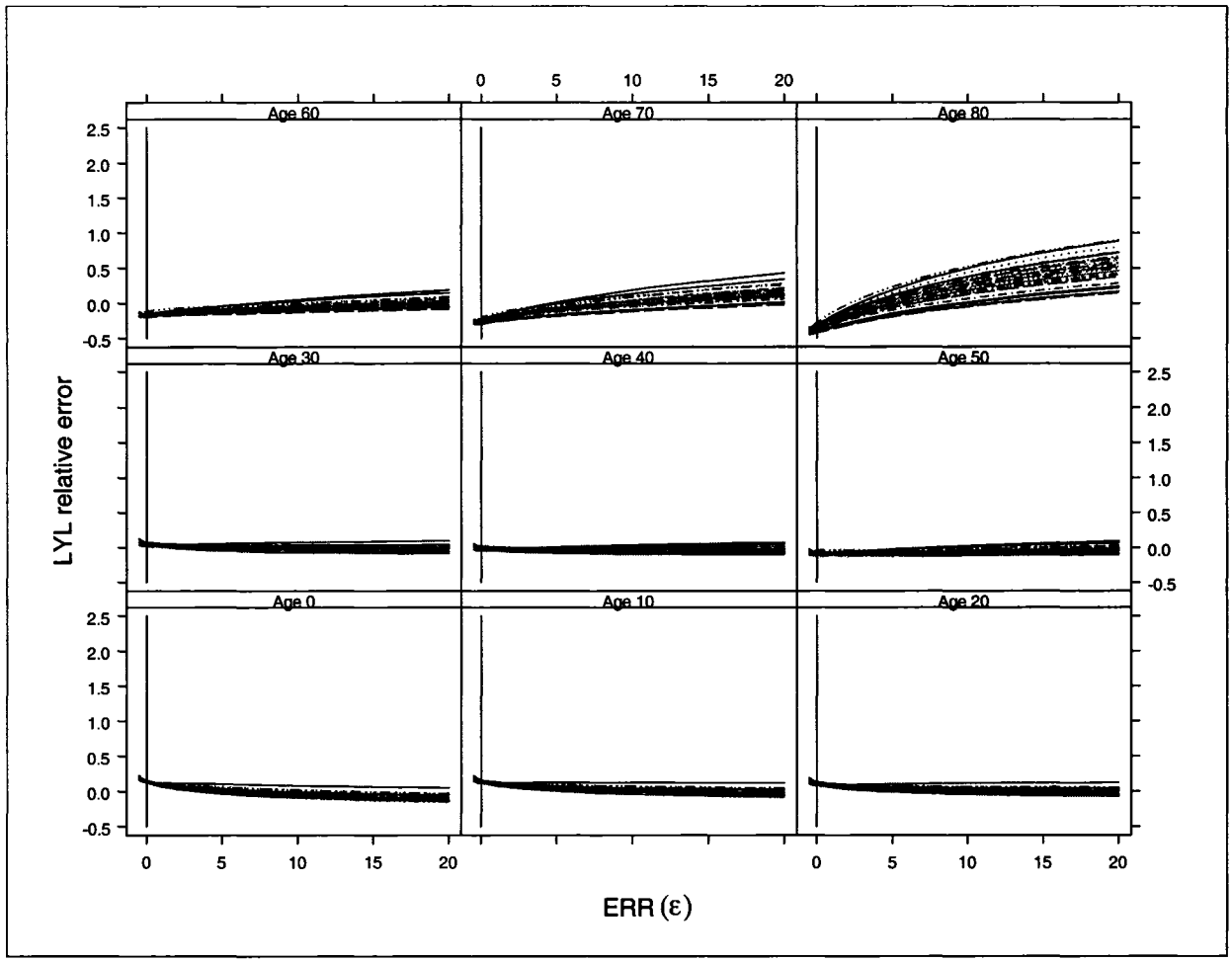
- **Performance in Model 3**

Let's look at the performance of model 3, the LYL relative errors are presented in Figure 20, the data preparation is same as model 1 and 2. In Figure 20, the panels are arranged in the same style as Figure 18 and 19.

The upper and lower bound values at different ages in Figure 20 are listed in Table 20 as a function of sample ε values.

The largest relative error is 90% when ε=20 and age=80. When we confine our interest to the ε in the range [-0.5, 2], the largest relative error is 44%.

Figure 20 The performance of Model 3 at different ages



Note: This plot shows the performance of model 3 at different ages (0, 10, 20, 30, 40, 50, 60, 70, 80). The performance is getting worse when age is 80.

Table 20 The upper and lower bound values in Figure 20

	Upper limit (ε = -0.5)	Lower limit (ε = -0.5)	Upper Limit (ε = 2.0)	Lower Limit (ε = 2.0)	Upper limit (ε = 20)	Lower limit (ε = 20)
Age 0	0.228	0.157	0.116	0.031	0.048	-0.159
Age 10	0.227	0.141	0.121	0.051	0.119	-0.087
Age 20	0.191	0.099	0.093	0.029	0.125	-0.071
Age 30	0.132	0.042	0.051	-0.010	0.102	-0.079
Age 40	0.061	-0.028	0.007	-0.056	0.080	-0.091
Age 50	-0.024	-0.104	-0.030	-0.116	0.099	-0.099
Age 60	-0.117	-0.194	-0.069	-0.172	0.189	-0.085
Age 70	-0.230	-0.295	-0.098	-0.237	0.432	-0.022
Age 80	-0.347	-0.443	-0.018	-0.312	0.900	0.148

For ease of comparison, the same scale and range are used in Figure 17-19. The performance of male populations in the EME region is similar (Results not shown).

4.6.2.4. Choice of these three models

Model 1 has larger error than Model 2 when age is young but Model 1 outperforms Model 2 when age is above 80. Model 2 has similar performance as Model 3 when age is less than or equal to 50, but the performance deteriorates faster than Model 3 when age is greater than 50. Model 3 has better performance than Model 1. From the comparisons, Model 3 is the best of all the three models. It is not surprising as it uses the exact LYL values to adjust the Equation (22).

In model 3, the maximum relative error is 90%. This worst case appears when ε=20 and age=80. A modification involving 20-fold change of all-cause mortality is an extreme change. Most applications of practical interest will involve much lower ε. When we restrict our attention to the moderate to small ε, this relative error is 44%. This worst case appears when $\varepsilon = -0.5$ and age = 80.

The maximum LYL relative error appears at age 80. When age=80, the LYL is a smaller number relative to the younger age. From the definition of the relative error (T-A)/T, where T represents true value and A represents approximation, in this case the denominator is a small number and prone to get a larger relative error. However, the absolute error may not be large in this case.

Our final choice is Model 3. Model 3 is an implementation of the Extended Keyfitz Model.

As the next step, this convenient formula (Model 3) is compared with other existing approximation approaches to evaluate its performance.

5. Performance of the Extended Keyfitz Model

This section will address the performance issue of the Extended Keyfitz Model (Equation 28) by comparing it with the existing approximation approaches. Here is a list of the existing approximation approaches: the DEALE, New DEALEs, the IPH, and the original (First order) Keyfitz approach.

Before conducting the performance evaluation of the Extended Keyfitz Model, it is helpful to compare the data requirements of the different approaches. The data requirement is one of the considerations in weighing the relative favorability of the various approaches.

Table 21 summarizes the data requirement of the various approaches. From Table 21, we understand the approximation approaches all require the age-stratified LE tabulation except the Extended Keyfitz Model.

Table 21 Data requirement for the approximation approaches

Approximation Approach	LR	LE	H/Λ	Constants
DEALE	$LR_b(t)$	$LE_b(t)$		0
New DEALEs	$LR_b(t)$	$LE_b(t)$		0
IPH	$LR_b(t)$	$LE_b(t)$	$\Lambda_2(t)$	0
Keyfitz	$LR_b(t)$	$LE_b(t)$	H(t)	0
Extended Keyfitz Model	$LR_b(t)$	$LE_b(0)$	H(0)	3

Note: The age dependent item in this table represents the requirement of the age-stratified tabulation. In contrast, $LE_b(0)$ and H(0) represent the requirement of a single number.

It is assumed that the age-stratified tabulations are always available for all these approaches to make the performance comparison. For this assumption, the comparisons are

actually conducted in favor of the approximation approaches that require the age-stratified tabulations, which are the DEALE, the new DEALEs, the IPH and the Keyfitz approach.

The comparisons are conducted across all the populations in our EME database. The results are consistent across these populations. Specifically, the Canadian female population in year 2000 is chosen to present the comparison results. Those results of other populations in the EME region have similar patterns (Results not shown).

The next section presents the performance comparison between the Extended Keyfitz Model and the DEALE for the Canadian female population in year 2000.

5.1. The Extended Keyfitz Model VS the DEALE

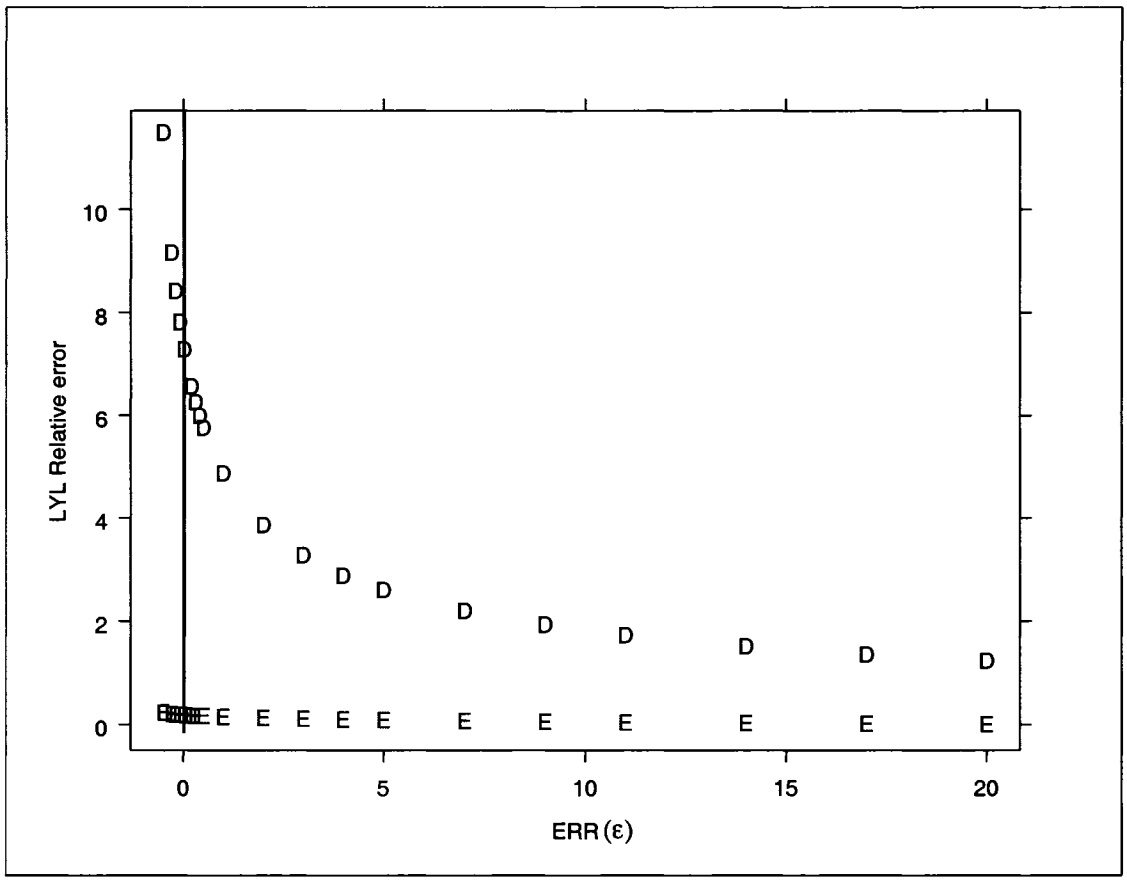
We start out by examining the special case of age=0 before examining the general cases (for other ages). In this section, the Extended Keyfitz Model (Equation 28) is compared to the DEALE.

At first, the LYLs are calculated using the DEALE, the Extended Keyfitz Model and the CMLT. The LYLs calculated by the CMLT serve as "true" values. As the population and age is fixed (age=0), the only dimension is ε. So the LYLs on different ε values by these three approaches are calculated. The LYL relative errors by the DEALE and the Extended Keyfitz Model are then calculated accordingly.

The comparison results are presented in Figure 21.

In Figure 21, the character "D" denotes the LYL relative errors associated with the DEALE approach and "E" denotes the LYL relative errors associated with the Extended Keyfitz Model.

Figure 21 Comparison between the DEALE and the Extended Keyfitz Model at birth



Note: This plot shows the performance of the DEALE and the Extended Keyfitz Model when age=0. "D" denotes the DEALE and "E" denotes the Extended Keyfitz Model. The performance of the Extended Keyfitz Model is better than the DEALE in this case. Another fact is the bad performance of the DEALE in negative multiplicative modification (reduction to mortality). This graph is specific for the Canadian female population in year 2000. Those graphs for other populations in the EME region have similar patterns.

Figure 21 shows that the relative errors associated with the Extended Keyfitz Model are much less than those associated with the DEALE. The bad performance of the multiplicative negative modification (reduction in mortality) of the DEALE is obvious under this situation. It shows progressively larger error as ε decreases.

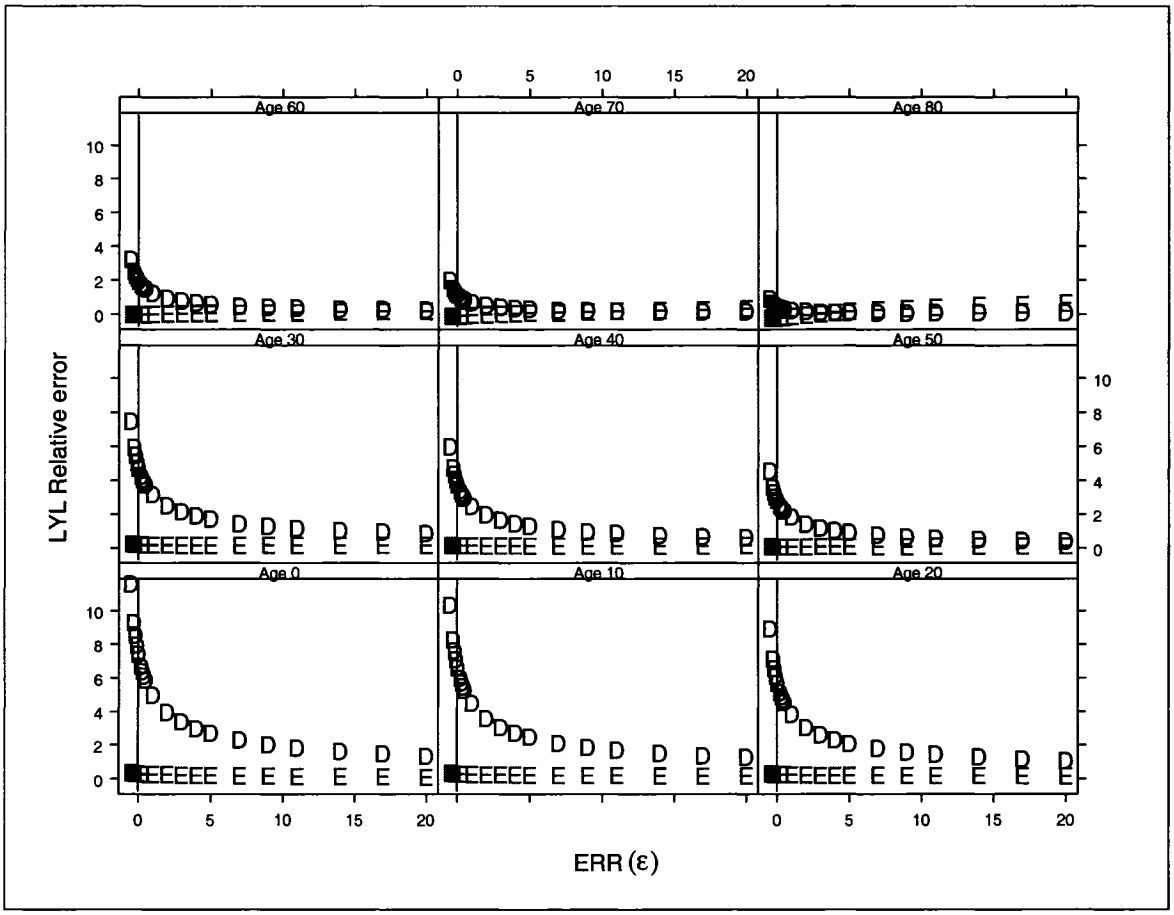
The Extended Keyfitz Model looks stable and performs well under this situation.

Let's look at their performance in a more general sense.

Figure 21 only examines the special case of age=0. We want to show the LYL relative errors associated with the DEALE approach versus those of the Extended Keyfitz Model at different ages. There is one more dimension: age. The LYL relative errors are calculated in the same way as preparing for Figure 21 but allowing the age to take different values.

Figure 22 plots the LYL relative errors for the Extended Keyfitz Model and the DEALE as a function of ε at different ages. The "D" and "E" in each panel represents relative errors associated with the DEALE approach and the Extended Keyfitz Model respectively.

Figure 22 Comparison between the DEALE and the Extended Keyfitz Model



Note: This plot presents the performance comparison between the DEALE and the Extended Keyfitz Model at different ages. It is obvious that the Extended Keyfitz Model outperforms the DEALE except at age 80.

The following is a summary of the comparison results in Figure 22:

1. When the multiplicative modification is positively small or negative, the DEALE has bad performance.
2. The LYL relative error associated with the DEALE decreases when ε increases.
3. The Extended Keyfitz Model is generally better than the DEALE approach. When age=80, these two approaches have similar performance. When age>=80, the performance of the DEALE is getting better because the future mortality rate is close to be modeled as a constant, which is the assumption in the DEALE. When age is greater than or equal to 80, it is arguable which approach is better and it depends on the ε value. Roughly in this case when ε <=5, the DEALE approach has bigger relative error. When the condition is ε >5, the DEALE could possibly outperform the Extended Keyfitz Model.
4. Though this is the pattern for Canadian female population in year 2000, it is a general pattern for all the populations in the EME region in year 2000 (Results not shown).

In Figure 22, it is quite clear that the Extended Keyfitz Model outperforms the DEALE in most cases.

5.2. The Extended Keyfitz Model VS the new DEALEs

Please refer to Appendix E for the detailed comparison results. The results are similarly arranged as those in Section 5.1.

The following is a summary of the comparison results:

1. The new DEALEs have lower relative error than the DEALE when the multiplicative modification is negative.
2. The new DEALEs generally have better performance than the DEALE.
3. The Extended Keyfitz Model has better performance than the new DEALEs in general.

4. When age is approximately 80 years or more, the performance of the new DEALEs gets close to the Extended Keyfitz Model. Please refer to Appendix E for some details on the performance comparison under this particular condition.
5. Though these above results are for Canadian female population in year 2000, it is a general pattern for all the populations in the EME region.

From these comparisons, in general the Extended Keyfitz Model is superior to the new DEALEs. In addition, the Extended Keyfitz Model does not require age-specific inputs. Another disadvantage of the new DEALEs is that they are generally more complicated than the DEALE. The new DEALEs have similar level of complexity as the Extended Keyfitz Model.

5.3. The Extended Keyfitz Model VS the Keyfitz approach

The Extended Keyfitz Model is derived from the Keyfitz approach. It is interesting to make a comparison between these two approaches.

There are two dimensions: age and ε, the LYLs by the CMLT and the approximate LYLs by the Keyfitz approach and the Extended Keyfitz Model on different age and ε values are calculated. The relative errors can then be calculated and organized in these two dimensions. The relative errors calculated by the Keyfitz approach and the Extended Keyfitz Model are presented in Figure 23.

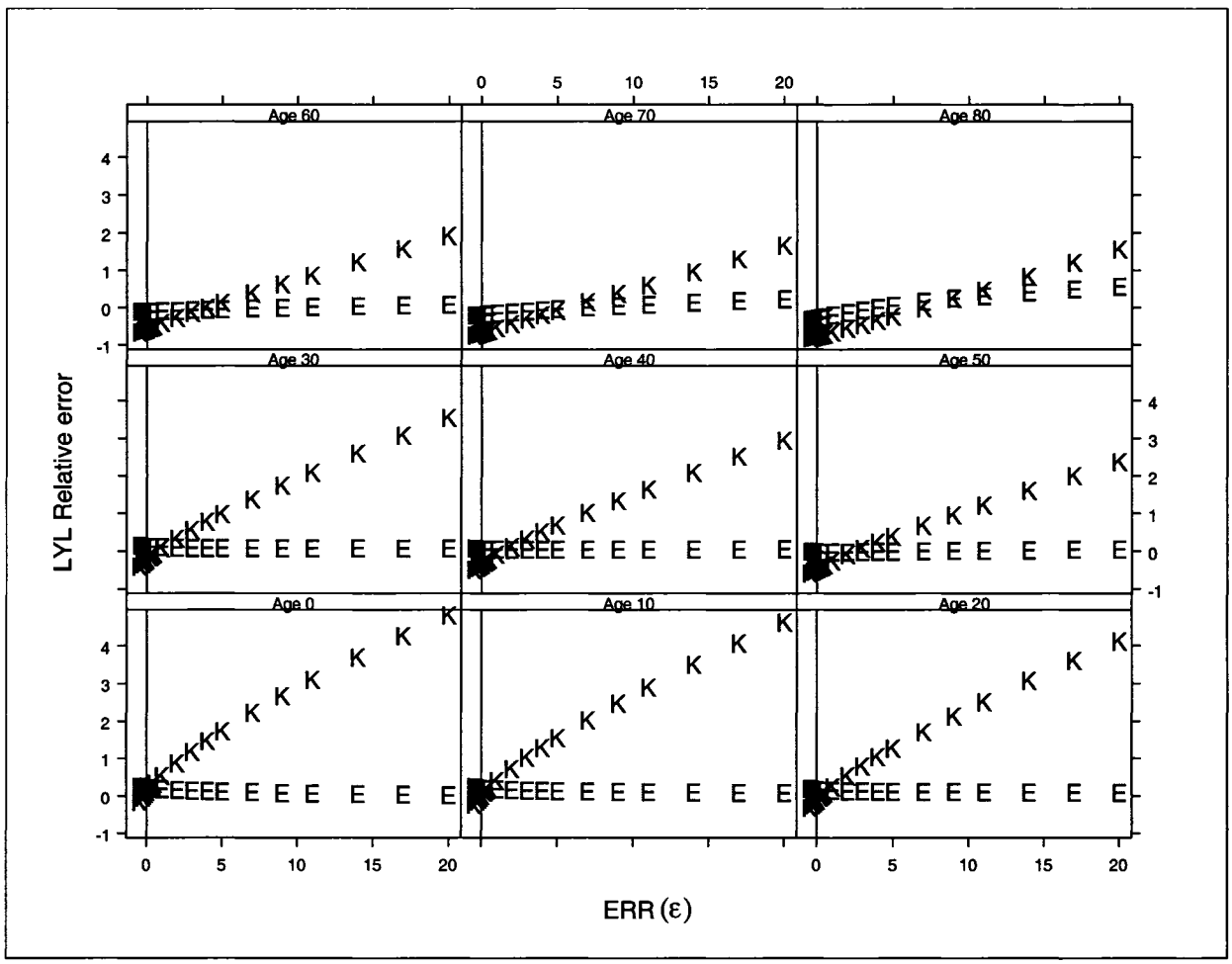
Figure 23 has the same configuration as Figure 22 except that the "K" in each panel represents the LYL relative errors associated with the Keyfitz approach and "E" represents the LYL relative errors associated with the Extended Keyfitz Model.

Here is a summary of the comparison results:

1. The LYL relative errors associated with the Keyfitz approach increases when age increases.
2. The Extended Keyfitz Model is systematically better than the Keyfitz approach at all ages and under any value of ε.

3. When ε is around 0, Keyfitz has similar accuracy as the Extended Keyfitz Model.
4. Though this is the pattern for Canadian female population in year 2000, it is a general pattern for all populations in the EME region.

Figure 23 Comparison between the Keyfitz approach and the Extended Keyfitz Model



Note: This plot presents the performance comparison between the Keyfitz approach and the Extended Keyfitz Model (Model 3, Equation 28). The Keyfitz approach assumes $LE_b(t)$ and $H(t)$ are always available, if the same assumption applies on the Extended Keyfitz Model, Equation (23) will be used.

The next section will have a real case study to show the advantage of the Extended Keyfitz Model. The decision or conclusion made by a more accurate approximation approach is more reliable.

6. Case Study

- **Initiative**

This section is going to apply the Extended Keyfitz Model in a real case. In this real case, the DEALE has been used in practice [13, 14, 15, 16]. This section will show how to apply the Extended Keyfitz Model in this case. The performance will be compared between the DEALE and the Extended Keyfitz Model.

- **Background**

Cancer screening is an important service provided by primary health care providers. Some researchers have tried to determine the cost effectiveness of cancer screening in the ESRD (End-Stage Renal Disease) population compared to the general population. In this real case, the cancer screening is the secondary health prevention. In their papers [13, 14, 15, 16], they use the DEALE to approximate the Life-years gained for cancer screening.

Here is one of their research topics:

The health policy maker is interested to know if it is better to conduct a breast cancer screening in ESRD population than in general population for women age from 50 to 69.

The answer is related to two aspects:

1. The life-years gained associated with the breast cancer screening in these two populations
2. The economic cost associated with the breast cancer screening in these two populations

Since the second issue is related to some health economics knowledge, it is beyond our scope. We narrow our interest to the life-years gained of the breast cancer screening in the general population and the ESRD population.

- **Study population**

In this real case, the LYL of the breast cancer screening is going to be estimated in the general population and the ESRD population by the Life Table, the DEALE and the Extended Keyfitz Model.

Here are the populations used in this real case:

1. General population refers to the US female population.
2. ESRD population refers to the US female population with ESRD disease.

There are some difficulties to replicate their results:

1. Different paper may study different year of these populations. It is not clear in some of these papers which year of these populations is studied.
2. The data associated with the specific populations they studied may not be available publicly.

These populations in year 2000 are chosen as the associated mortality data is available from some public data sources: 1. United States Renal Data System [20]; 2. Centres for disease control and prevention [21]. These populations in year 2000 are believed to be similar as those in all these four papers [13, 14, 15, 16].

- **Effect of the breast cancer screening**

In their paper, they mentioned "The breast cancer screening decreases mortality approximately 30% in women" [13]. It indicates this effect is independent of the populations. In another word, the breast cancer screening has the same effect on the general population and the ESRD population. This effect is a multiplicative modification (to mortality) and the ERR (ε_c) = - 0.3.

- **Replication process**

The mortality data of breast cancer for the general population and ESRD population are collected. These data are from United States female population in year 2000. The Life-years gained are calculated in two cases by the CMLT, the DEALE and the Extended Keyfitz Model:

1. The general population receives breast cancer screening.
2. The ESRD population receives breast cancer screening.

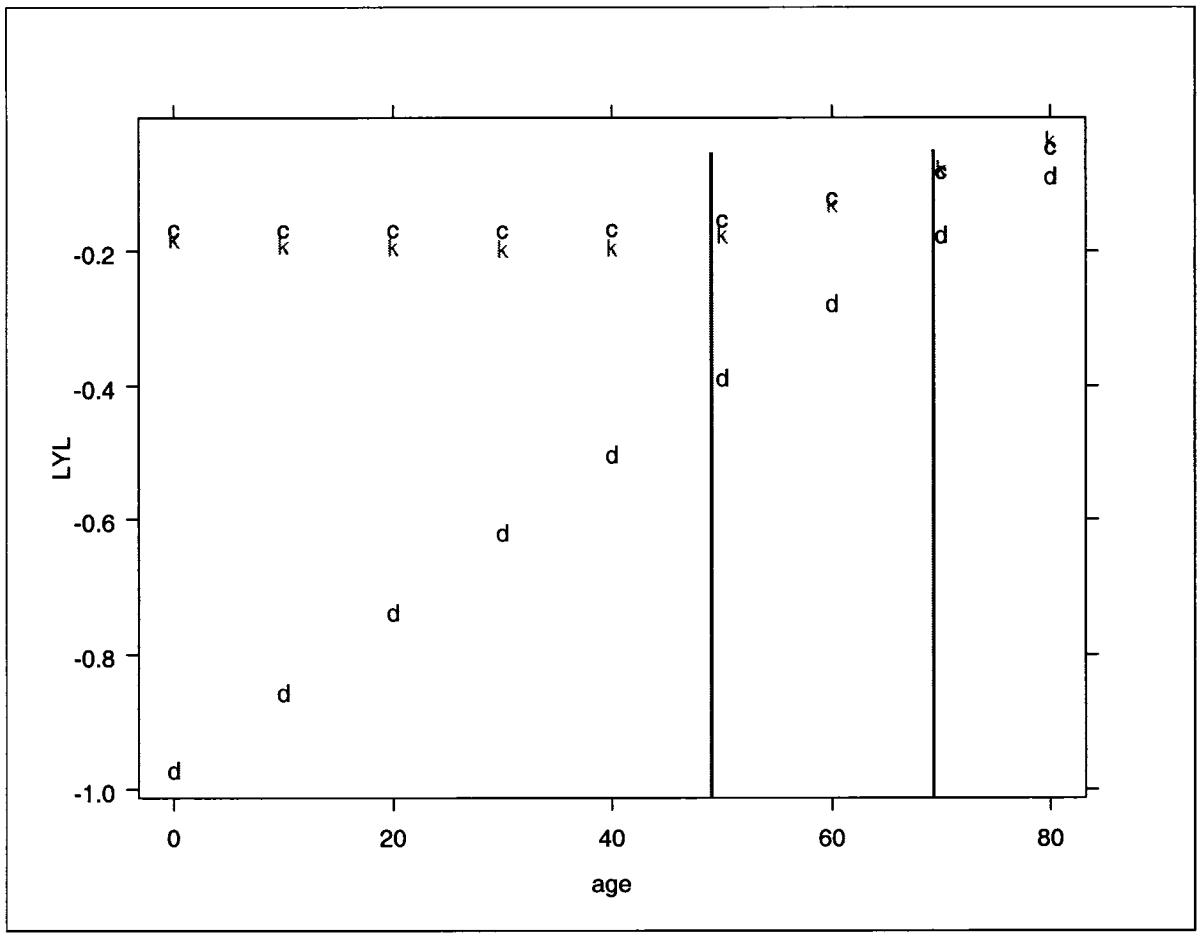
- **Comparisons**

Figure 24 shows the LYL calculated by the Life Table, the DEALE and the Extended Keyfitz Model for the general population. In Figure 24, character "c" denotes the LYL calculated by the Life Table, "d" denotes the LYL calculated by the DEALE and "k" denotes the LYL calculated by the Extended Keyfitz Model.

Some observations about the comparison in Figure 24 are listed below:

1. The error of the DEALE decreases when age increases. When age is young, the DEALE tends to have bigger error.
2. The Extended Keyfitz Model has better performance and appears more reliable in this case.

Figure 24 LYL approximation for the general population

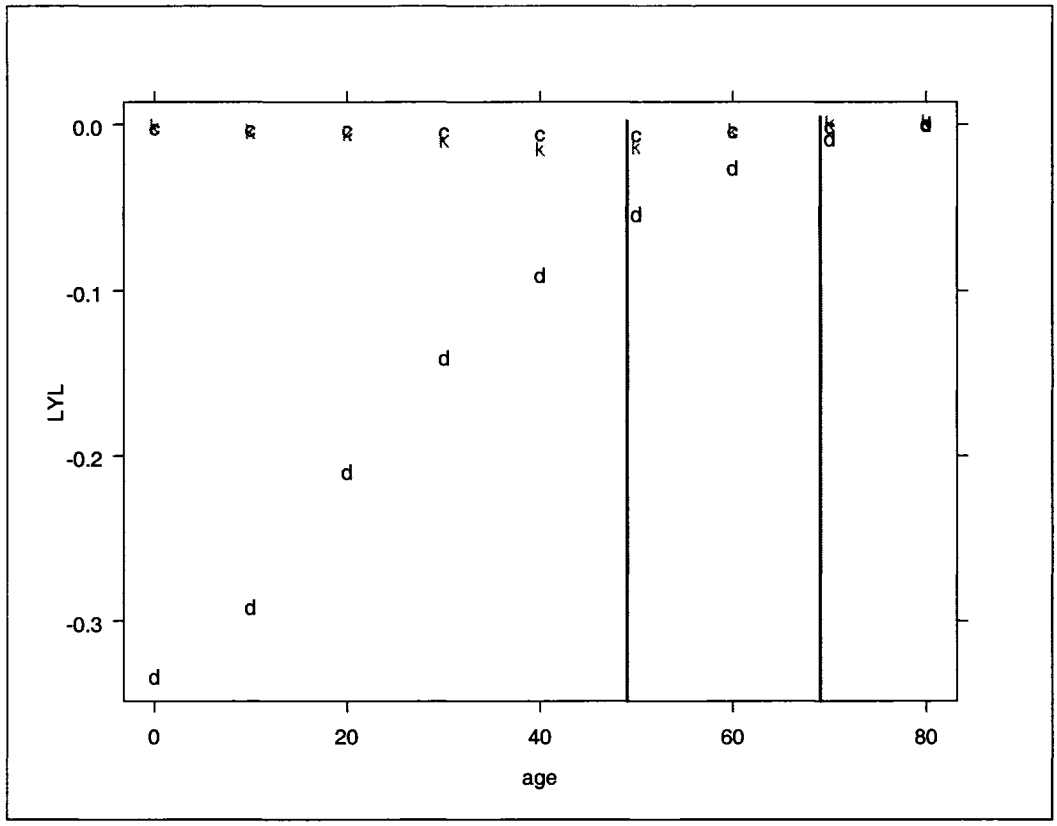


Note: LYL is approximated for the general population to evaluate the effect of the breast cancer screening by the Extended Keyfitz Model versus the DEALE. "c" denotes the exact value, "k" denotes the Extended Keyfitz Model and "d" denotes the DEALE. It is obvious the LYL approximated by the Extended Keyfitz Model is far more accurate than those approximated by the DEALE. The age range of interest is between those two solid vertical lines.

A similar graph for the ESRD population is presented in Figure 25.

In Figure 25, the pattern is similar as in Figure 24. However, the DEALE performs better in comparison with its performance in Figure 24.

Figure 25 LYL approximation for the ESRD population



Note: LYL is approximated for the ESRD population to evaluate the effect of the breast cancer screening by the Extended Keyfitz Model versus the DEALE. "c" denotes the exact value, "k" denotes the Extended Keyfitz Model and "d" denotes the DEALE. It is obvious the LYL approximated by the Extended Keyfitz Model is far more accurate than those approximated by the DEALE. The age range of interest is between those two solid vertical lines.

In these two graphs, the Extended Keyfitz Model is more stable and reliable.

The policy makers evaluated if it is wise to conduct the breast cancer screening in ESRD population instead of the general population. The large LYL relative errors associated with the DEALE may potentially lead to wrong decision. If one has to use an approximation, the Extended Keyfitz Model is a more wise choice when conducting such analysis.

7. Discussion

In this thesis, when modeling LE_b at any age by LE_b at birth for the populations in the EME region, it is assumed that the mortality rate schedules of these populations are independent. Actually, the mortality rate schedules for male and female populations in the same country are not independent. Even the mortality rate schedules for the populations in different countries may not be independent. For example, Canada and America are quite similar in many economic aspects and their mortality rate schedules are somewhat dependent. As independence is the prerequisite for the simple linear regression analysis, this could be a concern. In Section 4.6.2.2.1, two alternative analysis approaches are used and found little difference. The way to conduct this analysis is arguable.

The Health indices are used for the decision making. The individual variation will be a concern when a medical decision is made on the individual. In the DEALE paper, the individual variation is not discussed probably it is assumed that such effect are identical across different treatments for a specific individual. In an application context to compare multiple treatments on multiple individuals with same disease, the individual variation has to be considered. The exact approaches and the approximation approaches are not designed for that purpose.

The specific formula for the EME region is derived from the WHO database in year 2000. And any extrapolation to other time period is probably questionable. If the study population has similar mortality rate schedule as the populations in the EME region in year 2000 do, the Extended Keyfitz Model can be applied in this situation without lose its accuracy. Otherwise, applying the Extended Keyfitz Model beyond the EME region and year 2000 is questionable and need further research to explore its behavior and performance.

The performance of the Extended Keyfitz Model has been evaluated for a wide range of circumstances. For example, the performance was compared on the excess ratio for all cause

of death varies from -0.5 to 20, which covers almost all of the real health state change. The Extended Keyfitz Model allows you to apply it to an individual at any age. The error is relatively smaller than the DEALE in general. As the level of error in DEALE has been accepted in most application contexts. We can conclude this applies to the Extended Keyfitz Model as well.

The life expectancy calculated from the Life Table is used as a standard to evaluate the performance of different approximation approaches in this thesis. Some researchers may know some advanced sophisticated techniques to obtain more precise estimate of life expectancy, the techniques they use won't affect the performance evaluation in this thesis. When constructing the life-table, special techniques may be used for interpolating mortality rate and finessing the formulation of the life-tables. If one is trying to provide precise estimates for particular populations, then these techniques have possible importance. However, these techniques correct minor errors. When evaluating the performance of the new approximation approach, this minor error is less likely a concern to the comparison results.

The Extended Keyfitz Model is the first approximation approach that avoids reliance upon the age-stratified tabulations. The formula developed in this thesis is for the EME region only because the $LE_b(t)/LE_b(0)$ and $H(t)/H(0)$ in the EME region show some clear patterns. In other regions, $LE_b(t)/LE_b(0)$ might have such pattern but probably not for $H(t)/H(0)$, especially the populations in South Africa. If the clear patterns are identified in a region as those shown in the EME region, the formula to calculate the LYL for this region can be obtained similarly. If no pattern can be identified, then Equation (23) might be used to approximate the LYL.

In the derivation of the Extended Keyfitz Model, the relation between $H^{(1)}$ and $H^{(n)}$ is studied. The final result is Equation (19). This result is based on the life tables in year 2000 in

our WHO database. It indicates that this relation probably holds for the populations in some other years as well. A research on a more broad scope may be needed to address this issue.

In the derivation of the Extended Keyfitz Model, we only show the performance of Equation (23) for all the populations in the EME region, which is the region of our interest. Actually this equation is fit for all the populations in year 2000 in our WHO database. The performance of the populations other than in the EME region is similar as the populations in the EME region (Results not shown). If a general formula is our interest, we could possibly stop at Equation (23). However, Equation (23) requires age stratified tabulations, which could be necessarily a limitation.

The practical ε range [-0.5, 20] was used to develop the Extended Keyfitz Model. A narrower ε range would probably lead to an approximation more accurate within that range. The ε range to use for the derivation depends on the practical needs.

The bad performance in the positively small or negative multiplicative modification by the DEALE was shown in this thesis. This indicates the bad performance probably exists in the negative additive modification by the DEALE as well. Further research needs to be done on this topic.

There would be two error sources associated with the Extended Keyfitz Model:

- The convenient formula (Equation 26, 27, 28) tries to use $LE_b(0)$ and $H(0)$ to model Equation (23). The resulting models will introduce some errors. These errors can possibly be reduced by choosing some nonlinear models instead of the linear model in this thesis. However, the complexity of these nonlinear models introduced may not be worthwhile.
- The above error is introduced by modeling Equation (23). Equation (23) itself has some errors, that is, when $LE_b(t)$ and $H(t)$ are always available, the

approximation by Equation (23) will have some error to predict LYL. This error is possibly reduced by devising some new approximation approaches.

The next step to continue my work would be to adapt the Extended Keyfitz Model to the other regions (e.g. SSA) and study its behavior and performance.

In practice, the Extended Keyfitz Model has a number of strengths, especially the accuracy and ease of application.

8. Conclusion and Recommendations

Here are several guidelines when you select the proper approach to approximate the life-expectancy indices:

1. If the modified case is multiplicative, the best approximation approach is undoubtedly the Extended Keyfitz Model. The Extended Keyfitz Model is simple to use and has superior performance.
2. If the modified case is additive, the DEALE may be preferable. It is simple to use, and it performs well in the additive case. However, in the additive model, the DEALE works worse when excess mortality rate has similar magnitude as the baseline mortality rate and when the age is young. Under this condition caution is advised when using the DEALE.
3. When the modification is negative (reduction in mortality), it is wise to choose an approximation approach other than the DEALE: In multiplicative model, choose the Extended Keyfitz Model; in additive model, choose one of the new DEALEs.
4. If accuracy is important when you have an additive model, you may choose the exact approaches or one of the new DEALEs.

References

[1] Wright JC, Weinstein MC. Gains in Life expectancy from medical interventions – standardizing data on outcomes. N Engl J Med 1998;339:380-386

[2] Beck JR, Kassirer JP, Pauker SG. A convenient approximation of life expectancy (the 'DEALE'). I Validation of the model. AM J Med. 1982;73: 883-8

[3] Beck JR, Pauker SG, Gottlieb JE, Klein K, Kassirer JP. A convenient approximation of life expectancy (the 'DEALE'). II.Use in medical decision making. AM J Med. 1982;73: 889-97

[4] E. Keeler and R. Bell. New DEALEs: other approximations of life expectancy. Med-Decis-Making. 12(4):307-11, Oct-Dec 1992

[5] Nathan Keyfitz. What difference would it make if cancer were eradicated? An examination of the Taeuber Paradox. Demography. 14.4:411-418, Nov 1977

[6] Kevin P. Brand, Jan M. Zielinski, and Daniel Krewski. Residential Radon in Canada: An Uncertainty analysis of Population and Individual Lung Cancer Risk. Risk Analysis. 2005; 25: 253-269

[7] Kevin P. Brand. Approximations and Heuristics for the Cause Modified Life-Table. Risk Analysis. 2005; 25: 1-15

[8] CL. Chiang. The life-table and its applications. Robert E. Krieger, Publishing, Co, Malabar, FL, USA, 1984

[9] K. G. Manton and E. Stallard. Recent trends in mortality analysis. Academica Press, Inc., Toronto, 1984

[10] Sonnenberg FA, Beck JR. Markov Models in Medical decision making - A Practical Guide. Med-Decis-Making. 13 (4): 322-338

[11] Stephen C. Newman A Markov Process Interpretation of Sullivan's Index of Morbidity and Mortality Statistics in Medicine 7: 787-794

[12] Kuntz KM, Weinstein MC. Life expectancy biases in clinical decision making. Med-Decis-Making. 1995; 15: 158-69.

[13] Christopher J. LeBrun, Louis F. Diehl, Kevin C. Abbot, Paul G. Welch, Christina M. Yuan

Life Expectancy Benefits of Cancer Screening in the End-Stage Renal Disease Population

Journal of Kidney Diseases 35: 237-243

[14] Sahar Kajbaf, Graham Nichol, Deborah Zimmerman Nephrol Dial Transplant 17:1786-1789

[15] Chris J. LeBrun, Christina M. Yuan, Paul G. Welch A Reconsideration of the Benefits of Cancer Screening in Dialysis Patients Seminar in Dialysis 12:140-143

[16] Glenn M. Chertow, A. David Paltiel, William F. Owen, J. Michael Lazarus Cost-effectiveness of Cancer Screening in End-Stage Renal Disease ARCH INTERN MED 156:1345-1350

[17] http://www3.who.int/whosis/life_tables/life_tables_process.cfm?country=can&language=en

[18] http://gosset.wharton.upenn.edu/~foster/mortality/References/References

[19] http://www.cdc.gov/ncipc/pub-res/epi_of_violence.htm

[20] http://www.usrds.org/

[21] http://www.cdc.gov/cancer/natlcancerdata.htm

[22] Gompertz B. On the nature of the function expressive of the law of human mortality. Philos Trans R Soc London. 1982;115: 513-85

Appendix A Life Table calculation for Example A

In Example A, ERR ε_c=5.

Given the values from the following table, LE_b and LE_M can be calculated by Equation (3) and (4). Please refer to Appendix I for the detailed S-plus program.

The results are:

$LE_b = 52.25$ $LE_M = 51.74$

Table 22 Life table calculation for Example A

	q^*	S	Ψ	q^*_M	S_M
1-5	0.004479	1.000000	0.001317	0.004508	1.000000
6-10	0.000580	0.995521	0.050748	0.000728	0.995491
11-15	0.000652	0.994943	0.048466	0.000810	0.994767
16-20	0.001405	0.994294	0.010184	0.001477	0.993961
21-25	0.001339	0.992896	0.023453	0.001496	0.992493
26-30	0.001591	0.991567	0.024226	0.001783	0.991008
31-35	0.002043	0.989989	0.041204	0.002463	0.989241
36-40	0.003168	0.987966	0.031147	0.003661	0.986804
41-45	0.005188	0.984836	0.035564	0.006108	0.983191
46-50	0.008377	0.979726	0.024493	0.009398	0.977185
51-55	0.012923	0.971519	0.028590	0.014757	0.968001
56-60	0.022258	0.958964	0.022795	0.024764	0.953717
61-65	0.036252	0.937619	0.020451	0.039884	0.930099
66-70	0.058131	0.903628	0.011826	0.061460	0.893003
71-75	0.095064	0.851100	0.009000	0.099122	0.838119
76-80	0.159743	0.770191	0.005952	0.164084	0.755043
81-85	0.272077	0.647159	0.003998	0.276683	0.631153
86-90	0.429685	0.471081	0.001812	0.432580	0.456523
91-95	0.589071	0.268664	0.001237	0.591325	0.259040
96-100	0.776288	0.110402	0.000939	0.777856	0.105863
101-105	1.000000	0.024698	0.000729	1.000000	0.023517

Where q* denotes the probability of death,

S denotes the percentage of survival,

Ψ denotes fraction of death attributed to brain cancer,

$q^*_{M,i}$ denotes the probability of death in the modified case in Example A,

S_M denotes the percentage of survival in the modified case in Example A.

Appendix B Markov model

Markov models [9, 10] succinctly represent situations in which there is an ongoing risk of a patient moving from one state of health to another. We assume that there are a set of possible health states, and specify the probability per unit of time that a patient in a given state will "transition" to each possible state. These transition probabilities may depend on the current time (for example, the chance of death increases with time due to aging, independent of health). We also need to know the utilities of the states. Utilities may also be a function of the time at which they're entered (if, for example, utilities for health states are discounted, bad health sooner is worse than bad health later.) What we do assume, however, is that the process has no memory -- how we came to this state doesn't matter, only when we came. This is the prerequisite for applying Markov model.

Markov models are often represented using two figures:

A state diagram, which shows the possible states as nodes and arrows indicating possible transitions between states. For example, it must be one of the health states.

A transition-probability matrix that shows the probability of transitioning from one state to another. This matrix is mathematically called transition matrix.

Actually, you can show all of the information using either a diagram or a matrix, but both are often employed because they each make certain kinds of questions and operations easier.

To evaluate a Markov model, we can imagine a hypothetical patient who begins at some state, and then we can follow that patient until death. Each year of life, the patient gains the utility associated with the state s/he's in. Each year of life, the patient has a given probability of transitioning to a new state. When the patient's dead, we examine his/her accumulated utility. If we repeat this simulation for a few thousand patients, we can get a pretty good idea of the total expected utility associated with a life beginning at the initial state. (We can also measure

variance, confidence intervals, etc.) This approach to evaluating Markov models is called "Monte Carlo simulation".

Another way to think about this is to imagine a few thousand patients, and use the transition probabilities to apportion them into groups that transition into different states, adding up the total utility for each patient in each group. This is called "cohort simulation".

If you have two different cohorts of patients with different utilities, you run the simulation separately for each group. If utilities depend on past history, you can also create separate states associated with each past history.

If the transition probabilities don't change with time, you can get an exact solution, without simulation, using matrix algebra.

Another important representation is Markov-cycle trees, a way to represent the information that's more typically available clinically (e.g., the chance of death following surgery due to infection, rather than the overall chance of death following surgery.) These are like recursive decision trees, with only chance nodes.

Appendix C H calculation in Keyfitz approach

From the Keyfitz approach, we have Equation (17) to calculate $H^{(1)}$. The superscript in $H^{(1)}$ denotes the H value is related to the first term in Taylor expansion expression of LYL.

More generally, the $H_c^{(n)}$ can be calculated by the following formula:

$$H_c^{(n)} = \frac{-\int_a^{\omega} S(t)(Ln(S_c(t)))^n dt}{\int_a^{\omega} S(t)dt}$$

Where S(t) is the survival probability of age t for all cause of death,

$S_c(t)$ is the survival probability of age t under modification,

ω denotes the upper limit of age,

a denotes the starting age,

and n denotes the H is for n-th order parameter.

$S_c(t)$ is related to cause of death and age, which can be calculated by Equation (5). The H is therefore also varying with cause of death and age. Please refer to Appendix I for detailed S-plus program to calculate these two values.

Appendix D New DEALEs

In Section 2.3.3.4, five variants of the DEALE are mentioned. Here are some details:

- **The ERFALE**

This method approximates the hazard of death for an individual by $h_0(t)=bt$. The hazard is linearly increased by age. We can express the survival function s in terms of b.

$$\frac{ds}{dt} = -sh_0(t)$$

So

$$Ln[S_0(t)] = -\int_0^t h_0(u)du = -\frac{bt^2}{2}$$

And life expectancy (L) is the area under the survival curve,

$$L = \int_0^\infty e^{-\frac{bt^2}{2}} dt$$

$$\Rightarrow \quad L = \sqrt{\pi / 2b}$$

$$\Rightarrow \quad b = \frac{\pi}{2L^2}$$

Let the hazard from disease be d, Assume hazards are combined by linearly adding.

$$\text{Let } x = \frac{d}{\sqrt{b}} = dL\sqrt{\frac{2}{\pi}} = 0.798dL$$

Denote the cumulative normal to x by $\Phi(x)$, and the modified life expectancy for a diseased person is proposed to be calculated by:

$$LE_M = \frac{x\Phi(-x)}{d\Phi(x)}$$

- **The Delayed DEALE**

In the delayed DEALE, the assumption is no death for an initial part of remaining life and there is a constant hazard in the later part. Suppose the life expectancy for a healthy person is L, we can deduce that no death in the initial period kL, that is $h_0(t)=0$ when $t<kL$ and $h_0(t)=1/(1-k)L$ when $t>kL$. Let the hazard from disease be d.

From the survival function

$$\frac{ds}{dt} = -sh_0(t)$$

We can calculate the modified life expectancy (LE_M)

$$LE_M = \int_0^{kL} e^{-dt} dt + \int_{kL}^{x} e^{-dkL} e^{-\left[d+\frac{1}{(1-k)L}\right](t-kL)} dt$$

Simplify above formula, we finally get

$$LE_M = \frac{1}{d}\left[1 - \frac{\exp(-dkL)}{1+d(1-k)L}\right]$$

- **The mixed DEALE**

In mixed DEALE, it supposes that p of the population live exactly L more years and the remaining (1-p) follow the original DEALE. Let the hazard from disease be d. Then the modified life expectancy for a disease person using mixed DEALE is the weighted sum of life expectancy for the two types:

$$LE_M = p\frac{1-\exp(-dL)}{d} + \frac{1-p}{d+\frac{1}{L}}$$

- **Delayed Adaptive DEALE**

Delayed adaptive DEALE use well-chosen value of k for delayed DEALE to improve the accuracy of the estimate. The best constant value for k is about 0.5.

- **Mixed Adaptive DEALE**

Mixed adaptive DEALE uses well-chosen value of p for mixed DEALE to improve the accuracy of the estimate. The best constant value for p is about 0.75.

Appendix E Performance comparison ---New DEALEs vs the Extended Keyfitz Model

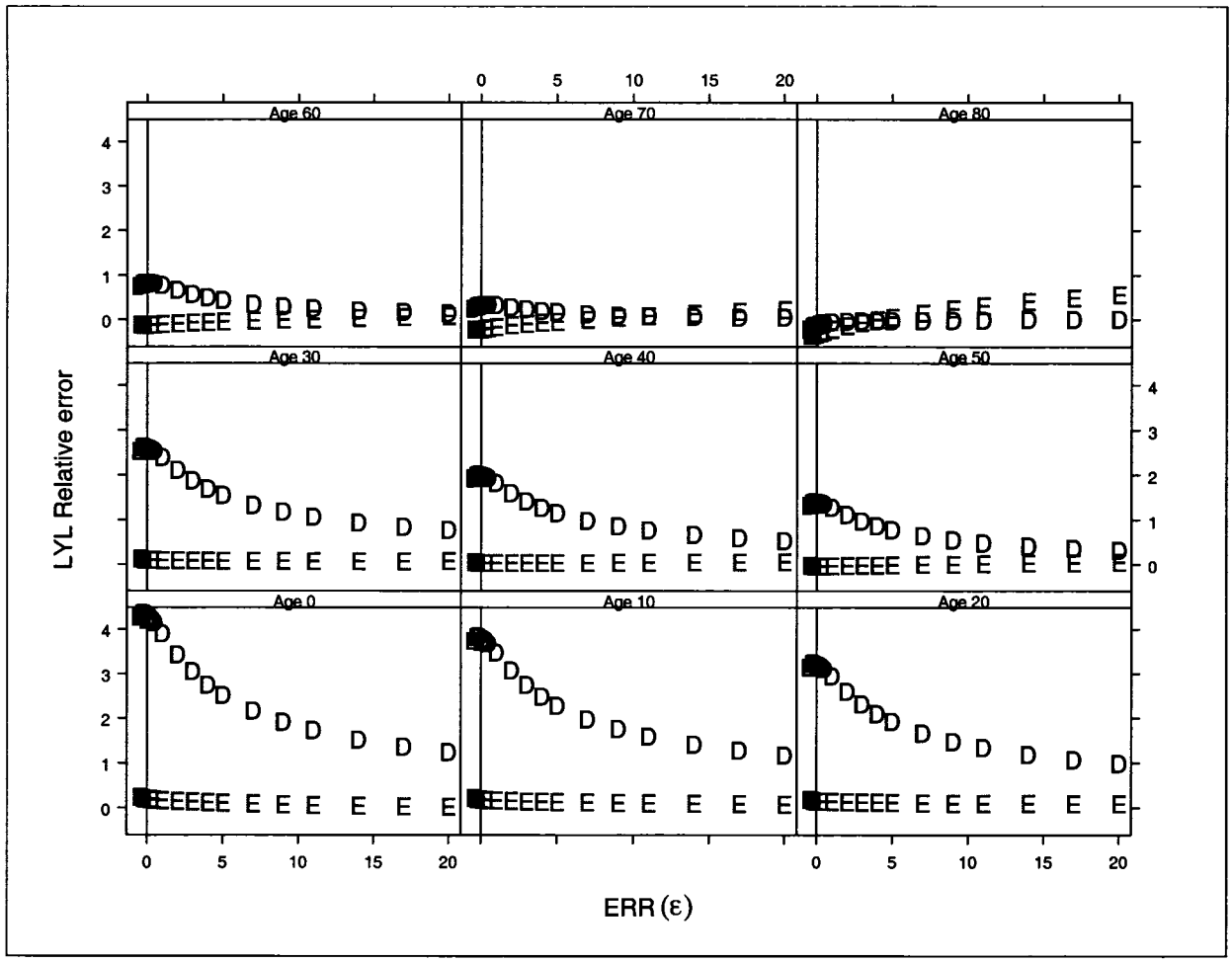
In this Appendix, we will compare the performance of the Extended Keyfitz Model with new DEALEs. Figure 26-30 show the LYL relative error with five new DEALEs approaches versus the Extended Keyfitz Model. The last two approaches use the optimized parameter as depicted in the New DEALEs paper. LYL relative error is defined same as before and LYL is approximated by different new DEALEs approaches in different graphs. The "D" in these graphs represents the new DEALEs approach, each graph has one of them, for example "D" represents ERFALE in Figure 26 and represents Delayed DEALE approach in Figure 27 etc. All the "E"s in 5 graphs represents the Extended Keyfitz Model. The panels denote different age in an ascending order from lower-bottom to upper-right starting from age 0 to age 80.

All five graphs are in same scale and range.

From these graphs, we can see:

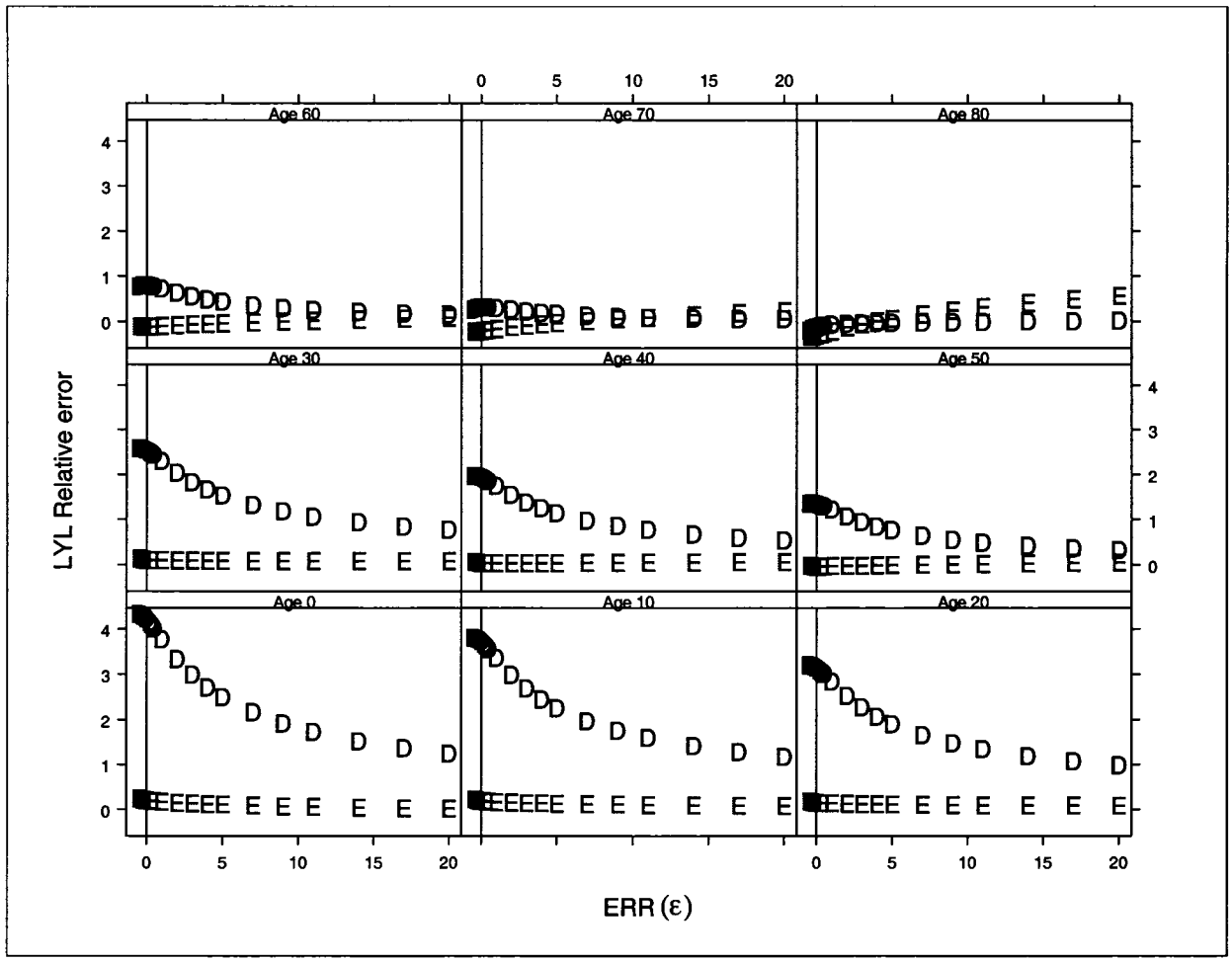
1. The Extended Keyfitz Model outperforms all the five new DEALEs.
2. When age is 80, some of the new DEALEs are competitive to the Extended Keyfitz Model.
3. They may not be worth to be used since they are complicated approaches and have worse performance than the Extended Keyfitz Model.
4. Though these graphs show the pattern for the Canadian female population in year 2000, we found that it is a general pattern for all the populations in the EME region.

Figure 26 Comparison between the ERFALE and the Extended Keyfitz Model



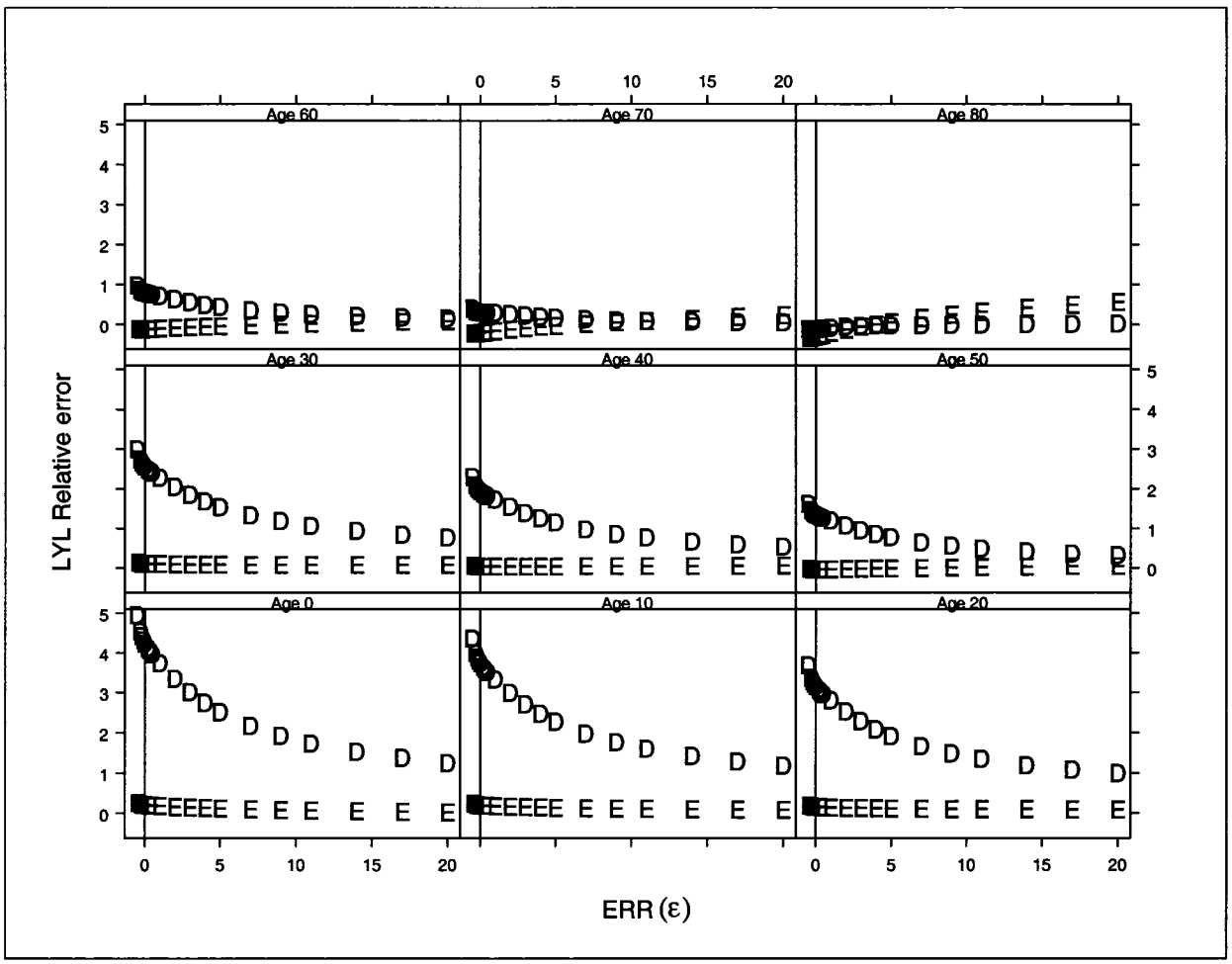
Note: This plot shows the performance of the ERFALE and the Extended Keyfitz Model. In general, the Extended Keyfitz Model outperforms the ERFALE.

Figure 27 Comparison between the Delayed DEALE and the Extended Keyfitz Model



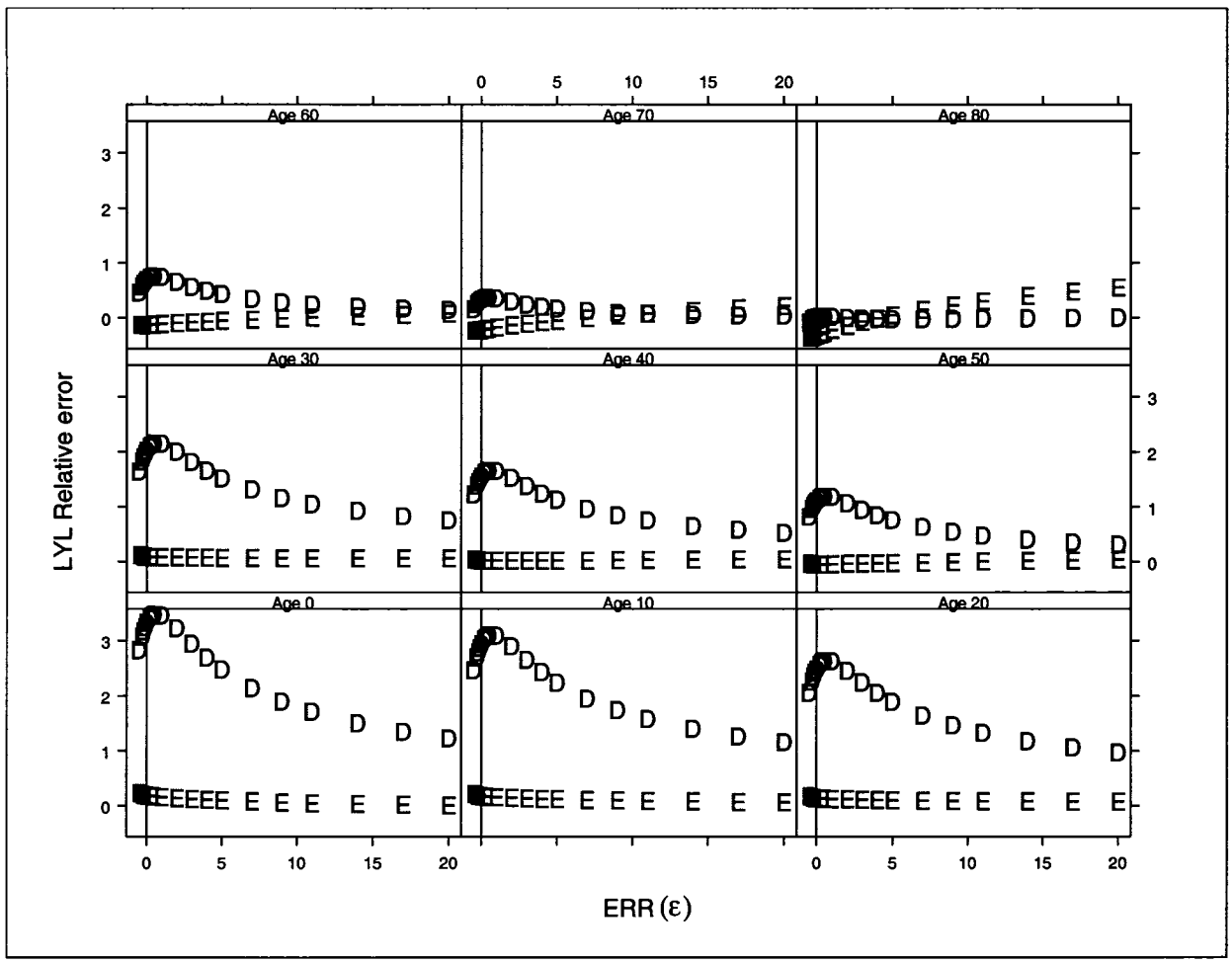
Note: This plot shows the performance of the Delayed DEALE and the Extended Keyfitz Model. In general, the Extended Keyfitz Model outperforms the Delayed DEALE.

Figure 28 Comparison between the Mixed DEALE and the Extended Keyfitz Model



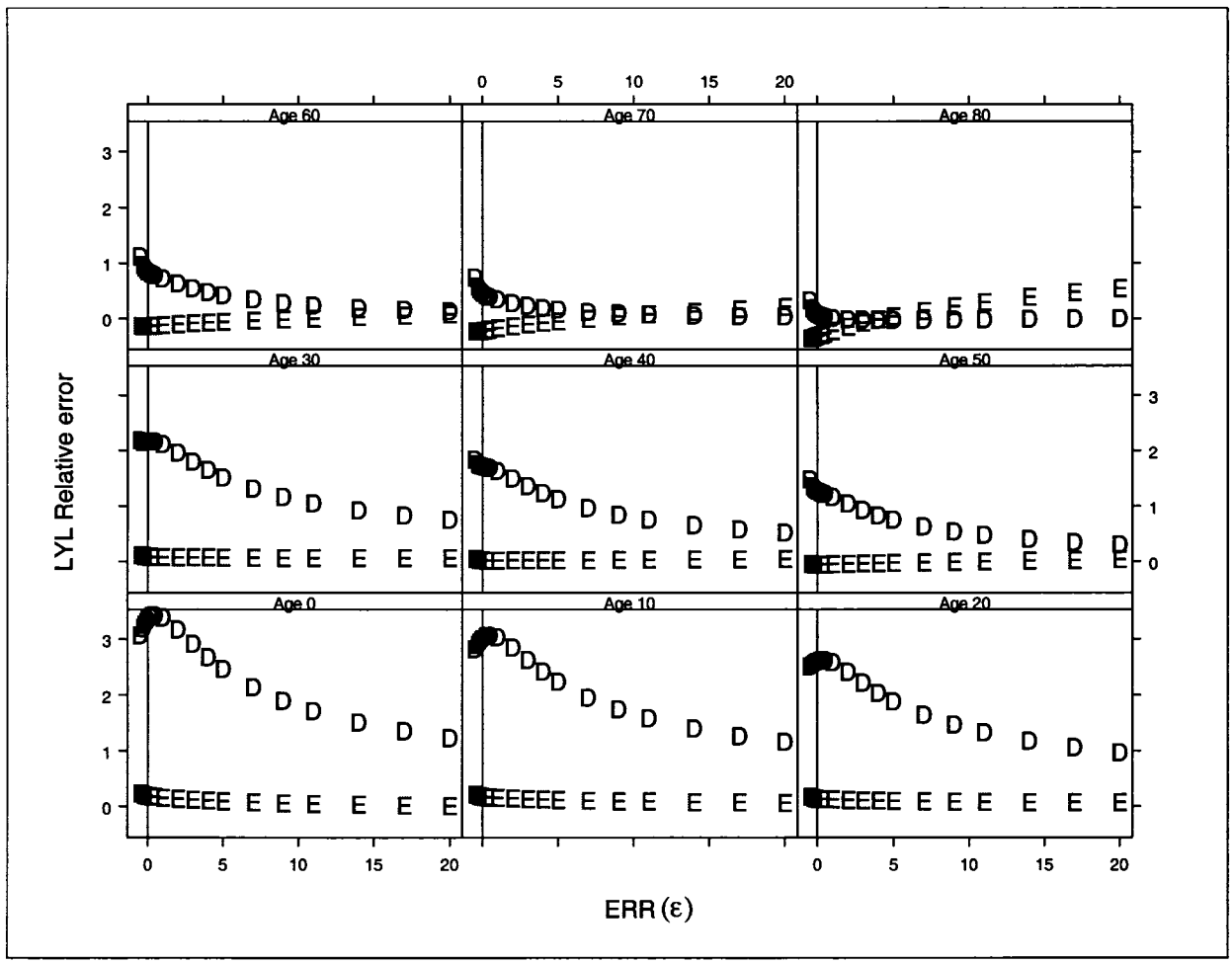
Note: This plot shows the performance of the Mixed DEALE and the Extended Keyfitz Model. In general, the Extended Keyfitz Model outperforms the Mixed DEALE.

Figure 29 Comparison between the Delayed Adaptive DEALE and the Extended Keyfitz Model



Note: This plot shows the performance of the Delayed Adaptive DEALE and the Extended Keyfitz Model. In general, the Extended Keyfitz Model outperforms the Delayed Adaptive DEALE.

Figure 30 Comparison between the Mixed Adaptive DEALE and the Extended Keyfitz Model



Note: This plot shows the performance of the Mixed Adaptive DEALE and the Extended Keyfitz Model. In general, the Extended Keyfitz Model outperforms the Mixed Adaptive DEALE.

Appendix F Modeling between $LE_b(t)/LE_b(0)$ and age t

- **Method 1**

The first way is to conduct the linear regression analysis by population, and then average all the coefficient estimates. The following S-plus analysis output is for the Canadian female population in year 2000:

Residuals:

Min	1Q	Median	3Q	Max
-0.02661	-2.463e-16	0.004339	0.011	0.01342

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
age	0.0117	0.0001	134.7506	0.0000

Residual standard error: 0.0124 on 8 degrees of freedom

Multiple R-Squared: 0.9996

F-statistic: 18160 on 1 and 8 degrees of freedom, the p-value is 1.033e-14

The detailed S-plus analysis results from other populations are not shown, but the slope and the standard error range are provided below:

Slope	0.0113 ~ 0.0125
Slope standard error	0.00005 ~ 0.00017

The final result of method 1 is the following equation:

$$\mathbf{LE_b(t) = LE_b(0) * (1-0.01196*t)}$$

- **Method 2**

The second way is to get the mean $LE_b(t)/LE_b(0)$ on each age, and then conduct the simple linear regression analysis by forcing the regression line through the origin.

Here is the S-plus analysis output:

Residuals:

Min	1Q	Median	3Q	Max
-0.03425	-0.001097	0.007903	0.01307	0.01702

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
age	0.0120	0.0001	105.3254	0.0000

Residual standard error: 0.01622 on 8 degrees of freedom

Multiple R-Squared: 0.9993

F-statistic: 11090 on 1 and 8 degrees of freedom, the p-value is 7.372e-14

The final result of method 2 is the following equation:

$$\mathbf{LE_b(t) = LE_b(0) * (1-0.012*t)}$$

There is little difference between the regression lines by these two approaches. No matter which approach is better, the final result is very close. It is arguable which kind of analysis is best. As the results are very close, a choice has to be made. The first way was finally adopted.

Appendix G Modeling between H(t)/H(0) and age t

- **Method 1**

The first way is to conduct the linear regression by population, and then average all the coefficient estimates. The following S-plus analysis output is for the Canadian female population in year 2000:

Residuals:

Min	1Q	Median	3Q	Max
-0.2293	-0.2011	-0.1088	-3.331e-16	0.3888

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
age	0.0216	0.0014	14.9621	0.0000

Residual standard error: 0.2064 on 8 degrees of freedom

Multiple R-Squared: 0.9655

F-statistic: 223.9 on 1 and 8 degrees of freedom, the p-value is 3.93e-07

The detailed S-plus analysis results from other populations are not shown, but the slope and its standard error range are provided below:

Slope	0.0198 ~ 0.0248
Slope standard error	0.0012 ~ 0.0018

The final result of method 1 is the following equation:

H(t) = H(0) * exp(0.02244948*t)

- **Method 2**

The second way is to get the mean $LE_b(t)/LE_b(0)$ on each age, and then conduct the linear regression analysis. The following is the analysis output from S-plus:

Residuals:

Min	1Q	Median	3Q	Max
-0.2389	-0.2108	-0.1154	-4.718e-16	0.4066

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
age	0.0224	0.0015	14.8834	0.0000

Residual standard error: 0.2154 on 8 degrees of freedom

Multiple R-Squared: 0.9651

F-statistic: 221.5 on 1 and 8 degrees of freedom, the p-value is 4.094e-07

The final result of method 2 is the following equation:

H(t) = H(0) * exp(0.0224*t)

There is little difference between the regression lines by these two approaches. It is arguable which kind of analysis is best. As the results are close, a choice has to be made. The first way was finally adopted.

Appendix H Modeling between $(LE_bH)(t)/(LE_bH)(0)$ and age t

- **Method 1**

The first way is to conduct the linear regression by population, and then average all the coefficient estimates. The following S-plus analysis output is for the Canadian female population in year 2000:

Residuals:

1	2	3	4	5	6	7
-0.3684	0.2317	0.2322	0.09683	-0.03778	-0.09178	-0.06276

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.2315	0.2350	-30.7786	0.0000
age	0.0736	0.0044	16.8713	0.0000

Residual standard error: 0.2309 on 5 degrees of freedom

Multiple R-Squared: 0.9827

F-statistic: 284.6 on 1 and 5 degrees of freedom, the p-value is 1.338e-05

The detailed S-plus analysis results from other populations are not shown, but the intercept, slope and their standard error ranges are provided below:

Intercept	-7.7234 ~ -6.1214
Intercept standard error	0.1420 ~ 0.3092
Slope	0.0590 ~ 0.0819
Slope standard error	0.0026 ~ 0.0057

The final result of method 1 is the following equation:

$$(LE_b*H)(t)=(LE_b*H)(0)(1-\exp(-6.94+0.0733*t))$$

- **Method 2**

The second way is to get the mean $LE_b(t)/LE_b(0)$ on each age, and then conduct the linear regression analysis. The S-plus result is listed below:

Residuals:

1	2	3	4	5	6	7
-0.3449	0.188	0.2087	0.1213	0.002417	-0.07358	-0.1018

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.9418	0.2175	-31.9168	0.0000
age	0.0733	0.0040	18.1375	0.0000

Residual standard error: 0.2137 on 5 degrees of freedom

Multiple R-Squared: 0.985

F-statistic: 329 on 1 and 5 degrees of freedom, the p-value is 9.362e-06

The final result of method 2 is the following equation:

$$(LE_b*H)(t)=(LE_b*H)(0)(1-\exp(-6.94+0.0733*t))$$

There is little difference between the regression lines by these two approaches. It is arguable which kind of analysis is best. As the results are same, a choice has to be made. The first way was finally adopted.

Appendix I S-plus Code

- **Life-years lost and life-expectancy calculation for different approaches**

```
LYL <- function(qstar, qcause, alpha = c(0.09, 0.4, 0.5), err = 0, method = "base", age = 0, LE = F, k = 0.5,
p = 0.75)
{
# Make sure qstar and qcause are vectors and no missing value
# qstar represents probability of dying on all causes of death
# qcause represents probability of dying on specific cause of death
# err represents excess rate ratio
# LE =T returns Life expectancy; LE=F returns Life years lost
# k,p are parameters in new DEALEs
# alpha is the fraction of last age interval
        qstar <- as.vector(qstar)
        qcause <- as.vector(qcause)
        qstar <- qstar[!is.na(qstar)]
        qcause <- qcause[!is.na(qcause)]# Match method with legal choices, exit if no match
        which.method <- pmatch(method, c("base", "chiang", "deale", "kevin", "erfale", "delayed",
"mixed", "delayed.adaptive",
                "mixed.adaptive", "lambda"))
        if(age < 0)
                stop("age can not be a negative value")
        if(is.na(which.method))
                stop("method must be one of the following values: \n
base,chiang,deale,crude,erfale,delayed,mixed,delayed.adaptive, mixed.adapt
ive,lambda"
                        )       # LE base
        nincrement <- length(qstar)
        if(length(qstar) != length(qcause))
                stop("Length in parameters qstar and qcause does not match!")
        if(nincrement >= 100) {
                repeats <- c(1, rep(2, 4), rep(3, nincrement - 5))
                a <- alpha[repeats]
                delta.yr <- rep(1, nincrement)
                delta.temp <- c(rep(1, trunc(age)), age - trunc(age))
                if(nincrement != 110)
                        warning("The age upperbound is not 110, here is the current age upperbound:",
nincrement)
        }
        else {
```

```
            if(nincrement <= 21)
                stop("Error in parameters!")
            a <- c(mean(alpha[1:2]), rep(alpha[3], nincrement - 1))   #alternative
(alpha[1]+4*alpha[2])/5
            delta.yr <- c(rep(5, nincrement))
            delta.temp <- c(rep(5, trunc(age/5)), age - 5 * trunc(age/5))
        }
        delta.temp <- delta.temp[delta.temp > 0]
        temp.length <- length(delta.temp)
        delta.real <- delta.yr - c(delta.temp, rep(0, length(delta.yr) - temp.length))
        S.base <- c(1, cumprod(1 - qstar))[ - (nincrement + 1)]
        coef <- S.base * (1 - qstar * (1 - a))
        if(temp.length <= 1)
            LE.base <- sum(coef * delta.real)
        else LE.base <- sum(coef * delta.real)/coef[temp.length - 1]
        sci <- qcause/qstar
        q.modified <- 1 - (1 - qstar)^(1 + sci * err)
        S.modified <- c(1, cumprod(1 - q.modified))[ - (nincrement + 1)]
        coef <- S.modified * (1 - q.modified * (1 - a))
        if(temp.length <= 1)
            LE.chiang <- sum(coef * delta.real)
        else LE.chiang <- sum(coef * delta.real)/coef[temp.length - 1]
        LYL.chiang <- LE.base - LE.chiang       # LYL based on Chiang method
        u.asr <- 1/LE.base
        u.d <- u.asr * err
        LE.deale <- 1/(u.asr + u.d)
        LYL.deale <- LE.base - LE.deale         # LYL based on Deale approximation
        x <- 0.798 * u.d * LE.base
        LE.erfale <- (x/u.d) * (pnorm( - x)/dnorm(x))
        LYL.erfale <- LE.base - LE.erfale       # LYL based on ERFALE approximation
        LE.delayed <- (1/u.d) * (1 - exp( - u.d * k * LE.base)/(1 + u.d * (1 - k) * LE.base))
        LYL.delayed <- LE.base - LE.delayed     # LYL based on delayed DEALE approximation
        LE.mixed <- (p * (1 - exp( - u.d * LE.base)))/u.d + (1 - p)/(u.d + 1/LE.base)
        LYL.mixed <- LE.base - LE.mixed         # LYL based on mixed DEALE approximation
        k <- LE.base/(17 + LE.base * (1 + 11 * u.d))
        LE.delayed.adaptive <- (1/u.d) * (1 - exp( - u.d * k * LE.base)/(1 + u.d * (1 - k) * LE.base))
        LYL.delayed.adaptive <- LE.base - LE.delayed.adaptive # LYL based on adaptive delayed DEALE
approximation
        p <- (1.14 * LE.base)/(LE.base + 15)
        LE.mixed.adaptive <- (p * (1 - exp( - u.d * LE.base)))/u.d + (1 - p)/(u.d + 1/LE.base)
        LYL.mixed.adaptive <- LE.base - LE.mixed.adaptive     # LYL based on adaptive mixed DEALE
approximation
```

```
        lam.LE <- lambda(qstar, qcause, a, delta.real, delta.yr)
        LYL.kevin <- LE.base * lam.LE * err
        LE.kevin <- LE.base - LYL.kevin
        return(switch(which.method,
                LE.base,
                ifelse(LE, LE.chiang, LYL.chiang),
                ifelse(LE, LE.deale, LYL.deale),
                ifelse(LE, LE.kevin, LYL.kevin),
                ifelse(LE, LE.erfale, LYL.erfale),
                ifelse(LE, LE.delayed, LYL.delayed),
                ifelse(LE, LE.mixed, LYL.mixed),
                ifelse(LE, LE.delayed.adaptive, LYL.delayed.adaptive),
                ifelse(LE, LE.mixed.adaptive, LYL.mixed.adaptive),
                lam.LE))
}
```

- **Calculate Λ using in the IPH method**

```
lamda <- function(qstar, qcause, a, delta.real, delta.yr)
{
        sci <- qcause/qstar
        len <- length(delta.yr)
        lam.k <- (sci * qstar)/(1 - qstar)
        lam <- c(0, cumsum(lam.k))
        lam <- lam[ - length(lam)]
        s.base <- c(1, cumprod(1 - qstar))[ - (len + 1)]
        w <- s.base * (1 - (1 - a) * qstar) * delta.real
        w <- w/sum(w)
        lamb <- sum(w * lam)
        return(lamb)
}
```

- **Calculate life-expectancy using Monte-Carlo simulation (Markov model)**

```
LE.monte<-function(pop, ageupbound, method = "Gompertz", param = c(7.59e-005, 0.0875, 0.0005), age =
0)
{
        qc <- qcompare(ageupbound, method = method, para = param, method = "Gompertz")
        cohort <- matrix(ifelse(runif(ageupbound * pop) - qc < 0., F, T), nrow = ageupbound)
        cohort <- apply(cohort, 2., function(x)
        {
```

```
            !cumsum(!x)
        }
        )
        cohort <- cohort[(age + 1):ageupbound, ]
        LE <- sum(cohort)/sum(cohort[age + 1, ]) + 0.5
        return(LE)
}
```

- **Fit the data into Gompertz function** [22] (Used in simulation)

```
Gfit<-function(qstar, step = 1)
{
# Make sure qstar and qcause are vectors and no missing value
# qstar represents probability of dying on all causes of death
        qstar <- as.vector(qstar)
        qstar <- qstar[!is.na(qstar)]
        nincrement <- length(qstar)
        q <- 1 - (1 - qstar)
        S.modified <- cumprod(1 - q)[ - (nincrement + 1)]
        x <- seq(step, nincrement * step, step)
        y <- log( - log(S.modified))
        s <- lm(y ~ x)
        b <- s$coef[2]
        a <- b * exp(s$coef[1])
        return(a, b)
}
```

- **Return mortality rate using Gompertz function** [22] (Used in simulation)

```
qcompare <- function(ageupbound, method = "Gompertz", para = c(7.59e-005, 0.0875, 0.0005))
{
        which.method <- pmatch(method, c("Gompertz", "Markham", "Logistic"))
        if(which.method == 1) {
                if(length(para) < 2)
                        stop("Error in para, it must at least have two values")
                temp <- (exp(para[2] * (0:ageupbound)) * para[1])/para[2]
                hx <- diff(temp)
                qx <- 1 - exp( - hx)
                return(qx)
        }
        if(which.method == 2) {
```

```
        if(length(para) < 3)
                stop("Error in para, it must at least have three values")
        temp <- (exp(para[2] * (0:ageupbound)) * para[1])/para[2] + para[3] * (0:ageupbound)
        hx <- diff(temp)
        qx <- 1 - exp( - hx)
        return(qx)
    }
    if(which.method == 3) {
        if(length(para) < 3)
                stop("Error in para, it must at least have three values")
        temp <- exp(para[2] * (0:ageupbound)) * para[1]
        temp <- temp/(1 + temp) + para[3]
        delta <- diff(temp)/2
        hx <- temp[ - ageupbound] + delta
        qx <- 1 - exp( - hx)
        return(qx)
    }
}
```

- **H calculation**

```
Keyfitz.age<- function(qstar, qcause, age = 0., alpha = c(0.09, 0.4, 0.5), method = "H1", err =
0.01)
{
    q.star <- as.vector(qstar)
    q.cause <- as.vector(qcause)
    q.star <- q.star[!is.na(q.star)]
    q.cause <- q.cause[!is.na(q.cause)]
    #len <- length(q.cause) - (sum(q.cause == 1) - 1)
    #q.star <- q.star[1:len]
    #q.cause <- q.cause[1:len]
    which.method <- pmatch(method, c("H1", "H2", "H3", "H4", "H5", "H6"))
    if(is.na(which.method))
        stop("method must be one of the following values: \n H1 H2 H3 H4 H5 H6")
    nincrement <- length(q.star)
    if(length(q.star) != length(q.cause))
        stop("Length in parameters qstar and qcause does not match!")
    if(nincrement >= 100.) {
        repeats <- c(1., rep(2., 4.), rep(3., nincrement - 5.))
        a <- alpha[repeats]
        delta.yr <- rep(1., nincrement)
        delta.temp <- c(rep(1., trunc(age)), age - trunc(age))
        delta <- 1.
        if(nincrement >= 110.)
```

```
                warning("The age upperbound is larger than 110, here is the current age
upperbound:", nincrement)
        }
        else {
                if(nincrement <= 15.)
                        stop("Error in parameters!")
                a <- c(0.4, rep(alpha[3.], nincrement - 1.))
                delta.yr <- c(rep(5., nincrement))
                delta.temp <- c(rep(5., trunc(age/5.)), age - 5. * trunc(age/5.))
                delta <- 5.
        }
        delta.temp <- delta.temp[delta.temp > 0.]
        temp.length <- length(delta.temp)
        delta.real <- delta.yr - c(delta.temp, rep(0., length(delta.yr) - temp.length))
        S.star <- c(1., cumprod(1. - q.star))
        S.star.m <- S.star[ - length(S.star)] + diff(S.star)/2.
        S.star <- S.star[ - length(S.star)]
        S.cause <- c(1., cumprod(1. - q.cause))
        S.cause.m <- S.cause[ - length(S.cause)] + diff(S.cause)/2.
        S.cause <- S.cause[ - length(S.cause)]
        h1.m <- sum( - delta.real * S.star.m * logb(S.cause.m))/sum(delta.real * S.star.m)
        h2.m <- sum( - delta.real * S.star.m * logb(S.cause.m) * logb(S.cause.m))/sum(S.star.m *
delta.real)
        h3.m <- sum( - delta.real * S.star.m * (logb(S.cause.m))^3.)/sum(S.star.m * delta.real)
        h4.m <- sum( - delta.real * S.star.m * (logb(S.cause.m))^4.)/sum(S.star.m * delta.real)
        h5.m <- sum( - delta.real * S.star.m * (logb(S.cause.m))^5.)/sum(S.star.m * delta.real)
        h6.m <- sum( - delta.real * S.star.m * (logb(S.cause.m))^6.)/sum(S.star.m * delta.real)
        return(switch(which.method,
                h1.m,
                h2.m,
                h3.m,
                h4.m,
                h5.m,
                h6.m))
}
```