Tutorials in Biostatistics

Volume 1: Statistical Methods in Clinical Studies

Edited by

R. B. D'Agostino,
Boston University, USA

John Wiley & Sons, Ltd

Tutorials in Biostatistics

Tutorials in Biostatistics

Volume 1: Statistical Methods in Clinical Studies

Edited by

R. B. D'Agostino,
Boston University, USA

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-02365-1

Typeset by Macmillan India Ltd
Printed and bound in Great Britain by Page Bros, Norwich
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

2.2 Prognostic/Clinical Prediction Models

Part III CLINICAL TRIALS

3.1 Design

3.2 Monitoring

3.3 Analysis

Preface

The development and use of statistical methods has grown exponentially over the last two decades. Nowhere is this more evident than in their application to biostatistics and, in particular, to clinical medical research. To keep abreast with the rapid pace of development, the journal *Statistics in Medicine* alone is published 24 times a year. Here and in other journals, books and professional meetings, new theory and methods are constantly presented. However, the transitions of the new methods to actual use are not always as rapid. There are problems and obstacles. In such an applied interdisciplinary field as biostatistics, in which even the simplest study often involves teams of researchers with varying backgrounds and which can generate massive complicated data sets, new methods, no matter how powerful and robust, are of limited use unless they are clearly understood by practitioners, both clinical and biostatistical, and are available with well-documented computer software.

In response to these needs *Statistics in Medicine* initiated in 1996 the inclusion of tutorials in biostatistics. The main objective of these tutorials is to generate, in a timely manner, brief well-written articles on biostatistical methods; these should be complete enough so that the methods presented are accessible to a broad audience, with sufficient information given to enable readers to understand when the methods are appropriate, to evaluate applications and, most importantly, to use the methods in their own research.

At first tutorials were solicited from major methodologists. Later, both solicited and unsolicited articles were, and are still, developed and published. In all cases major researchers, methodologists and practitioners wrote and continue to write the tutorials. Authors are guided by four goals. The first is to develop an introduction suitable for a well-defined audience (the broader the audience the better). The second is to supply sufficient references to the literature so that the readers can go beyond the tutorial to find out more about the methods. The referenced literature is, however, not expected to constitute a major literature review. The third goal is to supply sufficient computer examples, including code and output, so that the reader can see what is needed to implement the methods. The final goal is to make sure the reader can judge applications of the methods and apply the methods. The tutorials have become extremely popular and heavily referenced, attesting to their usefulness. To further enhance their availability and usefulness, we have gathered a number of these tutorials and present them in this two-volume set.

Each volume has a brief preface introducing the reader to the aims and contents of the tutorials. Here we present an even briefer summary. We have arranged the tutorials by subject matter, starting in Volume 1 with 18 tutorials on statistical methods applicable to clinical studies, both observational studies and controlled clinical trials. Two tutorials discussing the computation of epidemiological rates such as prevalence, incidence and lifetime rates for cohort studies and capture–recapture settings begin the volume. Propensity score adjustment methods and agreement statistics such as the kappa statistic are dealt with in the next two tutorials. A series of tutorials on survival analysis methods applicable to observational study data are next. We then present five tutorials on the development of prognostics or clinical prediction models. Finally, there are six tutorials on clinical trials. These range from designing

and analysing dose response studies and Bayesian data monitoring to analysis of longitudinal data and generating simple summary statistics from longitudinal data. All these are in the context of clinical trials. In all tutorials, the readers is given guidance on the proper use of methods.

The subject-matter headings of Volume 1 are, we believe, appropriate to the methods. The tutorials are, however, often broader. For example, the tutorials on the kappa statistics and survival analysis are useful not only for observational studies, but also for controlled clinical studies. The reader will, we believe, quickly see the breadth of the methods.

Volume 2 contains 16 tutorials devoted to the analysis of complex medical data. First, we present tutorials relevant to single data sets. Seven tutorials give extensive introductions to and discussions of generalized estimating equations, hierarchical modelling and mixed modelling. A tutorial on likelihood methods closes the discussion of single data sets. Next, two extensive tutorials cover the concepts of meta-analysis, ranging from the simplest conception of a fixed effects model to random effects models, Bayesian modelling and highly involved models involving multivariate regression and meta-regression. Genetic data methods are covered in the next three tutorials. Statisticians must become familiar with the issues and methods relevant to genetics. These tutorials offer a good starting point. The next two tutorials deal with the major task of data reduction for functional magnetic resonance imaging data and disease mapping data, covering the complex data methods required by multivariate data. Complex and thorough statistical analyses are of no use if researchers cannot present results in a meaningful and usable form to audiences beyond those who understand statistical methods and complexities. Reader should find the methods for presenting such results discussed in the final tutorial simple to understand.

Before closing this preface to the two volumes we must state a disclaimer. Not all the tutorials that are in these two volumes appeared as tutorials. Three were regular articles. These are in the spirit of tutorials and fit well within the theme of the volumes.

We hope that readers enjoy the tutorials and find them beneficial and useful.

RALPH B. D'AGOSTINO, SR. EDITOR
Boston University
Harvard Clinical Research Institute

Preface to Volume 1

This first volume of *Tutorials in Biostatistics* is devoted to statistical methods in clinical research. By this we mean statistical methods applied to medical problems involving human beings, either as members of populations or groups in observational and epidemiological research or as participants in clinical trials. The tutorials are divided into three parts. Here we briefly mention the general themes of each part and the articles within them.

Part I is on observational studies and epidemiology. These articles clarify the uniqueness and complications that arise from observational data and present methods to obtain meaningful and unbiased inferences. Section 1.1 is devoted to epidemiology and contains two tutorials. The first, by Beiser, D'Agostino, Seshadri, Sullivan and Wolf, presents a thorough discussion of epidemiological event rates such as incidence rates and lifetime risks, clarifies issues such as competing risks in the calculation of these rates and includes computer programs to carry out computations. The second tutorial, by Chao, Tsay, Lin, Shau and Chao, describes the computation of epidemiological rates for capture-recapture data such as would be obtained from multiple surveys attempting to estimate the disease prevalence rate for, say, hepatitis A virus or diabetes in a population. The issue of minimizing biases is discussed. Section 1.2, on adjustment methods, contains one article by Ralph D'Agostino, Jr., on the use of propensity scores for reducing bias in treatment comparisons from observational studies. This tutorial has become a standard reference for propensity scoring. The article from Section 1.3, on agreement statistics, by Kraemer, Periyakoil and Noda, covers in detail the use of the kappa statistic in medical research.

Section 1.4 presents three tutorials devoted to survival methods applicable to observational studies. First, Bull and Spiegelhalter clearly identify the complications and other issues involved in using survival methods in observational studies (in contrast to clinical trials). Interval estimation and binary outcomes in longitudinal studies are then developed in the next two tutorials by Lindsey and Ryan and by Carlin, Wolfe, Coffey and Patton. The latter two tutorials have uses beyond survival analysis in observational studies. We group them in this part of the volume mainly for convenience. The reader should quickly see the broader applicability of the methods and not be limited by our classification.

Part II is concerned with prognostic/clinical prediction models and contains two sections. Here the aim is to present methods for developing mathematical models that can be used to identify people at risk for an outcome such as the development of heart disease or for the prognosis of subjects with certain clinical characteristics such as cancer tumour size. Some of these tutorials have become major references for clinical prediction model development. Section 2.1 contains one article by Mazumdar and Glassman on categorizing prognostic variables. The question is often how best to dichotomize a diagnostic variable so that it can be used in a clinical setting. Issues such as multiple testing often render useless such 'obvious' methods as trying to find the best cut point. A careful review of the field is presented and helpful suggestions abound.

Section 2.2, 'Prognostic/Clinical Prediction Models', presents in four detailed tutorials methods for developing and evaluating multivariable clinical prediction models. The first, by D'Agostino, Belanger, Markson, Kelly-Hayes and Wolf, illustrates how to deal with a large set of potentially useful prediction variables. Methods such as principal components analysis and hierarchical variable selection methods for survival analysis are highlighted. The next two articles have Frank Harrell as the first author and deal in detail with developing prediction models for time to event, binary and ordinal outcomes. (The first is authored by Harrel, Lee and Mark, and the second by Harrell, Margolis, Gove, Mason, Mulholland, Lehmann, Muhe, Gatchalian and Eichenwald.) Questions of model development are explored completely, as are issues of making predictions and concerns about validation. The last tutorial in Section 2.2 deals with estimating prognosis based on observational data such as are obtainable in a registry. It is authored by DeLong, Nelson, Wong, Pryor, Peterson, Lee, Mark, Califf and Pauker. These four tutorials are among the best literature sources for the development and appropriate use of clinical prediction models.

Part III is on clinical trials and contains three sections. While given as tutorials, the articles of this section are innovative in understanding as well as in the presentation of the issues and methods. Section 3.1 contains a clever article by Wong and Lachenbruch on the optimal design of dose response studies. Section 3.2, on monitoring in clinical trails, contains an article on Bayesian data monitoring by Fayers, Ashby and Parmar pointing to the benefits of a Bayesian analysis even in this setting. Section 3.3 contains four articles on analysis. These bring together ideas and methods available for use, but not presented elsewhere with such completeness and clear focus. They fill a serious void and add wonderfully to the field. The first, by Albert, deals with longitudinal clinical trial data analysis. The second, by Senn, Stevens and Chaturvedi, deals with generating simple summary numbers from repeated measures studies so that the analysis and interpretations of the study are intuitive and meaningful. The next article, by Agresti and Hartzel, concerns binary data from multi-centre trials. Lastly, Chuang-Stein and Agresti discuss dose responses with ordinal data.

We hope these 18 tutorials will be of use to readers.

Part I
OBSERVATIONAL STUDIES/ EPIDEMIOLOGY

1.1 Epidemiology

Computing estimates of incidence, including lifetime risk: Alzheimer's disease in the Framingham Study. The Practical Incidence Estimators (PIE) macro

Alexa Beiser[1,*,†], Ralph B. D'Agostino, Sr[2], Sudha Seshadri[3], Lisa M. Sullivan[1] and Philip A. Wolf[3]

[1] *Department of Epidemiology and Biostatistics, Boston University School of Public Health, Boston, MA, U.S.A.*
[2] *Department of Mathematics, Boston University, Boston, MA, U.S.A.*
[3] *Department of Neurology, Boston University School of Medicine, Boston, MA, U.S.A.*

SUMMARY

The incidence of disease is estimated in medical and public health applications using various different techniques presented in the statistical and epidemiologic literature. Many of these methods have not yet made their way to popular statistical software packages and their application requires custom programming. We present a macro written in the SAS macro language that produces several estimates of disease incidence for use in the analysis of prospective cohort data. The development of the Practical Incidence Estimators (PIE) Macro was motivated by research in Alzheimer's Disease (AD) in the Framingham Study in which the development of AD has been prospectively assessed over an observation period of 24 years. The PIE Macro produces crude and age-specific incidence rates, overall and stratified by the levels of a grouping variable. In addition, it produces age-adjusted rates using direct standardization to the combined group. The user specifies the width of the age groups and the number of levels of the grouping variable. The PIE macro produces estimates of future risk for user-defined time periods and the remaining *lifetime risk* conditional on survival event-free to user-specified ages. This allows the user to investigate the impact of increasing age on the estimate of remaining lifetime risk of disease. In each case, the macro provides estimates based on traditional unadjusted cumulative incidence, and on cumulative incidence adjusted for the competing risk of death. These estimates and their respective standard errors, are provided in table form and in an output data set for graphing. The macro is designed for use with survival age as the time variable, and with age at entry into the study as the left-truncation variable; however, calendar time can be stubstituted for the survival time variable and the left-truncation variable can simply be set to zero. We illustrate the use of the PIE macro using Alzheimer's Disease incidence data collected in the Framingham Study.

* Correspondence to: Alexa Beiser, Department of Epidemiology and Biostatistics, Boston University School of Public Health, 715 Albany Street, Boston, MA 02118, U.S.A.
† E-mail: alexab@bu.edu

Contract/grant sponsor: Framingham Heart Study of the National Heart, Lung and Blood Institute
Contract/grant sponsor: National Institute of Aging; contract/grant number: 5-R01-A608122-11
Contract/grant sponsor: NIH/NHLBI; contract/grant number: N01-HC-38038

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
© 2004 John Wiley & Sons, Ltd. ISBN: 0-470-02365-1

1. INTRODUCTION

The Framingham Heart Study is a population-based cohort study wherein subjects have been evaluated biennially for cardiovascular risk factors and cardiovascular endpoints since 1948 [1]. Over the years, additional data have been accumulated on sociodemographic and life-style factors and the cohort members have been screened for a variety of novel risk factors and non-cardiac disease conditions such as dementia, osteoporosis, cancer, visual loss and hearing impairment. Of the original 5209 subjects, 920 subjects are alive; their current mean age is 86 years. Thus we have a well-characterized cohort of 'old-old' subjects, traditionally defined as subjects over the age of 80 or 85 years. This is the fastest growing segment of the U.S. population and the study of the incidence of dementia and risk factors associated with dementia in this population is of enormous public health importance.

The incidence of disease is estimated in medical and public health applications using a variety of different techniques. Most of these techniques are discussed in detail in books on survival analysis [2–7], epidemiologic methods [8, 9] or biostatistical methods [10, 11]. Other techniques have been presented in the statistical or epidemiologic literature [12–17], or have simply been applied in the medical or public health literature [18]. Many of these methods have not yet made their way to popular statistical software packages and their application requires custom programming.

We present a macro written in the SAS macro language that produces several estimates of disease incidence for use in the analysis of prospective cohort data. This work was motivated by research in Alzheimer's disease (AD) in the Framingham Study in which the development of AD has been prospectively assessed over an observation period of 24 years. Our goal is to use these data to estimate: (i) crude and age group-specific yearly incidence of AD; (ii) age-adjusted yearly incidence of AD within selected subgroups; and (iii) the future risk of developing AD conditional on survival dementia-free to selected ages. We estimate future risk for predefined periods and the remaining lifetime risk, using traditional unadjusted cumulative incidence (UCI), and cumulative incidence adjusted for the competing risk of death (ACI).

2. MOTIVATION

The estimation of yearly incidence is relatively straightforward; however, in prospective studies such as the Framingham Study, there are several issues that make the estimation of cumulative incidence difficult. In order to generate a valid estimate of future risk, including the lifetime risk of developing AD, we must address the following. First, we must consider that individuals are followed for different periods of time. Second, the time origin must be defined such that individuals are comparable at the time origin. Third, we must account for subjects entering the observation period at different ages. Finally, we must address the impact of the competing risk of death. We discuss each issue below.

2.1. Individuals are followed for different periods of time

If we could follow every subject in the sample until either the end of the study or until they developed AD, we could directly estimate the probability of developing AD as the proportion of subjects in our sample who developed AD during the observation period, or the *cumulative*

incidence [10]. As is generally the case in a long-term prospective study, there are many subjects who are not observed for the entire observation period and we must use survival analysis techniques to estimate the cumulative incidence. The primary technique we will rely on is a modified Kaplan–Meier method.

2.2. Time origin

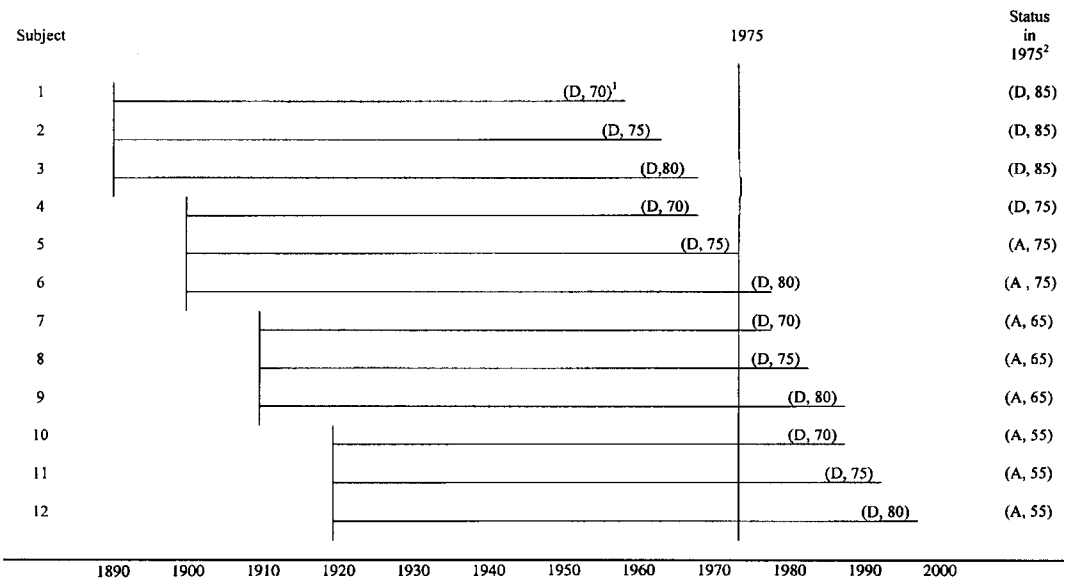
Prospective studies are often analysed using survival methods in which the dependent variable is survival time defined as the time from a designated origin to the event of interest (for example, diagnosis of AD). The time origin can be defined in several ways including the date of entry into a study or date of birth. If the time origin is defined as the date of entry into a study, an individual's time to event coincides with his/her time-on-study. If the time origin is defined as birth date, an individual's time to event is his/her survival age. However the time origin is defined, it is critical that individuals are comparable at the time origin [19]. In our case, subjects free of AD in 1975 enter the observation period at different ages. One subject may enter the observation period at age 50 years while another may enter at 70 years of age. Because the risk of AD is known to increase with age, it is important to take into account the different ages at which subjects enter the study. In this case, if the time origin is 1975, accounting for age may be done by adjusting for age as a covariate or by stratifying by age at entry (supposing we are able to categorize ages in a manageable way). This type of adjustment may not be sufficient.

Korn *et al.* [20] argue that survival age is more appropriate as a time scale than time-on-study for most outcomes. This approach assumes that the risk of development of the event of interest is more likely to change as a function of age than as a function of calendar time. This is certainly true for the development of dementia.

2.3. Subjects enter the observation period at different ages: selected risk set strategy

The survival age approach is ideal for population studies in which individuals are studied from their birth to event onset (for example, AD) or until they are right-censored with no event. Unfortunately, full survival information on an entire population is usually not available; rather, survival information is collected prospectively on a sample of event-free individuals selected from the population at the beginning of an investigation. When the time origin is defined to be birth, an individual's follow-up period is the sum of the observation period and the (event-free) period from birth to entry into the study. Individuals with incident events before the observation period are excluded from the entire study as is their follow-up time from birth to event. Thus follow-up time that occurred before the study is included, but *only if it is event-free.*

Very few, if any, subjects develop AD before the age of 65. In fact, in our data, the earliest AD case was diagnosed at age 69, and we only consider cases diagnosed at ages 70 or older. Consider a subject who enters the observation period at age 45 in 1975 and is observed for 24 years to 1998. In 1998 the subject reaches age 69 and is free of AD. In our estimate of the incidence rate, this subject contributes 24 person years (24 years free of AD) to the denominator. Including these person-years in the denominator decreases the estimate of incidence. A more useful estimator would be based on a restricted set of subjects, subjects who are truly 'at risk of developing AD' during the observation period (for example, subjects who are at least 70 years of age during the observation period).



1: Event status (D=dead, A=alive) and survival age
2: Event status (D=dead, A=alive) and age in 1975 (at start of observation period)

Figure 1. Lifetimes of 12 hypothetical patients.

Analytic approaches that rely on person-years at risk can underestimate hazards and incidence rates. For example, the hazard at age 70 years is the conditional probability of event onset at age 70, or the number of events at age 70 divided by the number of individuals at risk at age 70. Individuals who contribute time that occurred before the observation period (in other words, individuals who were older than 70 when the study began) appear in the denominator, but not in the numerator.

The bias in estimating hazards and incidences can be removed by excluding follow-up time that occurs before the observation period, by *left-truncation.* The risk set, at any age, must include only those individuals who were at risk at that age during the observation period.

Consider, for example, a hypothetical population of 12 individuals whose lifetimes are displayed in Figure 1. Suppose the event of interest is death. The population hazard at age 70 is defined as the ratio of the number of deaths at age 70 to the number of subjects in the risk set (defined as the number of subjects at risk entering age 70) and is equal to $4/12 = 0.33$. If an investigation is initiated in 1975 (see Figure 1), then 8 subjects are included in the investigation (subjects 5–12 who are alive at the beginning of the observation period in 1975). In a standard survival analysis of this 1975 cohort using survival age as the dependent variable, each of the 8 subjects contributes the number of years from his/her birth to survival age. The risk set entering age 70 includes 8 subjects, of whom 2 died at age 70 (Subjects 7 and 10). The population hazard is underestimated as $2/8 = 0.25$. Subjects 5 and 6 are older than 70 years of age at the beginning of the observation period (1975); they were both at risk at age 70, however, they were not at risk at

age 70 during the observation period. These subjects should be excluded from the risk set at age 70 (in 1975). If we restrict the risk set at age 70 to only those who are at risk at age 70 during the observation period, the risk set includes 6 subjects, and we correctly estimate the population hazard as $2/6 = 0.33$. This approach reflects a selected risk set strategy and we use this strategy in our computations.

2.4. Competing risks

The estimation of cumulative incidence of AD is complicated by a fairly common situation: the development of AD is subject to the competing risk of death. Subjects who die during the observation period are treated as censored observations in traditional survival analytic techniques such as the Kaplan–Meier method [21]. This method is inappropriate as it assumes that failure from the event of interest is still possible beyond the time at which the censoring occurred. For example, a person who dies of cardiovascular disease cannot develop AD and should not contribute to the estimate of development of AD. Gooley [13] shows that the potential contribution of censored observations to the probability of failure from the event of interest is distributed among those subjects remaining at risk. However, the potential contribution of a subject who has died should be zero. Treating such subjects as censored inflates the estimate of cumulative incidence. Various analytic solutions to the problem of competing risks have been proposed and implemented [12–17], but there is still no software available that addresses this issue.

We provide estimates of both the unadjusted cumulative incidence (UCI) and the cumulative incidence adjusted for the competing risk of death (ACI). (Note that these are generally referred to in the literature as 1-KM – the complement of the Kaplan–Meier estimate of survival – versus CI [12, 13]). The ACI is useful as an estimate of the probability of actually developing AD, while the UCI estimates the probability of developing AD assuming no competing risk (that is, all subjects living for the entire lifespan). The former estimator is particularly useful from a public health standpoint as it allows the estimation of the numbers of cases of AD one can expect in a given population. Further, by adjusting the observed cumulative incidence using the mortality experience of a 'standard population' one can estimate a standardized lifetime risk. In some diseases which appear to be associated with aging *per se*, the exponential rise in annual incidence with increasing age is balanced by the exponential decrease in life expectancy seen with age, resulting in a relatively invariant estimate of the lifetime risk in elderly individuals. Thus for instance, in the Framingham Study, the lifetime risks of Alzheimer's disease [18] and congestive heart failure [22] were found to remain relatively constant with increasing age beyond 65 years. The ability to generate a single sex-specific estimate of lifetime risk is useful in educating the public regarding the true risk of the disease. The unadjusted cumulative incidence may be useful in the pathophysiological investigation of potential risk factors for AD. As an example, cigarette smoking may appear to provide protection from AD when the ACI is used to estimate cumulative incidence. This could be a simple consequence of the fact that cigarette smoking is associated with increased mortality, thus decreasing the observed incidence of AD. Smoking may, however, increase the physiological risk of AD; this would be seen only if the UCI is used to estimate cumulative incidence (Seshadri *et al.*, submitted to the American Academy of Neurology, 2000).

2.5. Statistical software

Many standard statistical computing packages do not handle these issues in a straightforward manner, if at all. Even the calculation of one-year incidence rates and age-adjusted rates using

direct standardization requires a certain amount of programming. Many standard statistical software packages do not exclude follow-up that occurred before the observation period and thus underestimate hazards and cumulative incidence. For example, SAS *Proc Lifetest* provides estimates of cumulative incidence using the Kaplan–Meier method, but has no mechanism for left-truncation. SAS *Proc Phreg* performs proportional hazards modelling and allows left-truncation but does not consider the adjustment for competing risk that is often necessary. We developed an SAS macro that produces: one-year incidence rates by age group; age-adjusted rates to compare rates among the levels of a grouping variable; estimates of traditional Kaplan–Meier cumulative incidence; and estimates of cumulative incidence adjusted for the competing risk of another event. Confidence intervals for each of the cumulative incidence estimates can also be provided.

3. THE FRAMINGHAM STUDY

The Framingham Study is a longitudinal study of 5209 participants (2336 men and 2873 women) which began in 1948 [1]. Participants have been examined in biennial exam cycles from 1948 to the present. At study onset, the initial ages ranged from 28 to 62 years.

3.1. Neuropsychological assessment in the Framingham Study

Several standardized neuropsychological batteries have been administered to participants at biennial exam cycles beginning with exam 14 in 1975/1976 to prospectively assess dementia. The Kaplan Albert (KA) battery [23] was introduced in exam 14 and includes sub-tests taken or derived from: (i) the original Weschler Memory Scale (including the logical memory, logical memory-delayed, logical-memory retained tests); (ii) sub-tests of the Weschler Adult Intelligence Scale (including similarities, digit span forward and digit span backward tests); and (iii) a measure of word fluency taken from the Aphasia Examination. Starting with exam 17 in 1982/1983, the Folstein Mini Mental State Examination (MMSE) [24] has been administered to participants on a biennial basis.

3.2. Generating the Framingham Dementia Cohort

The Framingham Dementia Cohort includes $n = 2611$ participants who were dementia-free in 1975. The criteria outlined below used to determine inclusion in the dementia cohort are strict enough to ensure that members were indeed dementia-free. Participants had to pass the Kaplan–Albert battery or the MMSE (score at least 24 of a possible 30 points) to be included. The dementia cohort contains people who satisfied the following (see Figure 2):

(i) Passed the Kaplan–Albert battery in 1975 at exam 14 ($n = 2083$), or did not take the Kaplan–Albert Battery at exam 14 (as it was introduced part way through the cycle) and either:
(ii) passed the MMSE at exam 17 ($n = 474$), or
(iii) did not take the MMSE at exam 17 but passed it at exam 18 ($n = 38$), or
(iv) did not take the MMSE at exams 17 or 18 but passed it at exam 19 ($n = 17$).

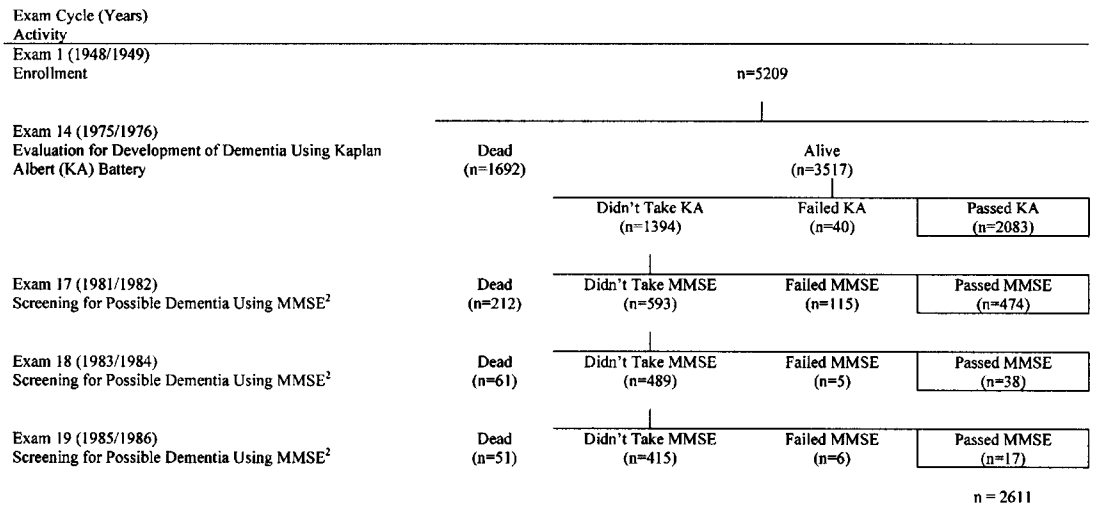


Figure 2. The Framingham dementia cohort: participants dementia free in 1975 ($n = 2611$). (Cases in the Framingham dementia cohort satisfied the following: passed the KA exam ($n = 2083$), or did not take the KA exam and either passed the MMSE at exam 17 ($n = 474$), or did not take the MMSE at exam 17 but passed the MMSE at exam 18 ($n = 38$), or did not take the MMSE at exams 17 or 18 but passed at exam 19 ($n = 17$). MMSE is Mini Mental State Examination.)

3.3. Dementia surveillance protocol; Identification of cases

Framingham Study participants are examined for cognitive decline according to a standardized protocol. At each biennial examination, subjects can be flagged for further evaluation based on: (i) self-report or family report of memory loss; (ii) referral by the physician conducting the biennial examination; or (iii) their performance on the MMSE. For this purpose, poor performance on the MMSE is defined as an absolute score below an education-based cut-off, a score more than 3 points lower than the score at the previous assessment or a score more than 5 points less than any previous score. Subjects may also be referred for evaluation by their primary care physician or by another source, such as the ongoing study 'Precursors of Stroke Incidence and Prevalence' in the same cohort.

A neuropsychologist reviews participants flagged by the initial screen, performs additional neuropsychological assessments and enters the participant in the dementia tracking protocol. Simultaneously but independently, a neurologist evaluates the flagged participants and, based on the neurology examination, classifies patients as not demented, or as mildly, moderately or severely demented. Patients classified as not demented or as mildly demented continue to be monitored with the dementia tracking protocol that includes annual neuropsychological re-examination and neurological re-evaluation as indicated, usually at least once every two years. Those who are classified as moderately or severely demented are evaluated at an in-depth dementia review.

A panel of at least two neurologists and one neuropsychologist conducts the dementia review. The review is based on neurology examination findings, neuropsychological assessments,

Framingham Study records, hospital and nursing home records, brain imaging, information from primary care physicians, and data gathered by telephone interview of family or next of kin. The panel is responsible for determining the presence of definite dementia (based on DSM-III [25] and, later, DSM-IV [26] criteria). At Framingham, a diagnosis of definite dementia also requires the presence of symptoms for at least 6 months and presence of 'moderate' dementia (severity not less than 1 on the Clinical Dementia Rating Scale). Participants who do not satisfy the criteria for definite dementia continue to be monitored according to the neuropsychology tracking protocol and are rereviewed by the dementia review panel as necessary. Participants identified as definitely demented are assigned a subtype of dementia, a year of dementia onset, and a year of diagnosis, that is, the year in which the criteria for a diagnosis of dementia were first fully satisfied. Criteria established by the National Institute of Neurological and Communicative Disorders – Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA [27]) are used in the diagnosis of probable Alzheimer's disease.

Participants who have suffered a stroke (usually after the onset of dementia) but do not meet the DSM-IV criteria for vascular dementia are classified as 'dementia complicated by stroke, relationship unknown'. These participants are combined with those classified as probable AD to estimate the upper limit of the risk of AD in the Framingham population.

Members of the Framingham Study cohort who have not participated in recent biennial examinations may be identified for dementia review after their death when a separate panel reviews their medical history as part of the Framingham Study protocol.

3.4. Dementia cohort follow-up status

Participants included in the dementia cohort ($n = 2611$) were followed from 1975 to 1998. Each participant is classified as a case (that is, he/she satisfied the criteria for dementia as outlined in Section 3.3 during the follow-up period), a death (that is, he/she died dementia-free during the follow-up period) or as a censored observation. Each classification is described in detail below. In all cases, follow-up times (that is, until development of dementia, death or censoring) are based on the last year of follow-up as opposed to an exact date.

The last year of follow-up for participants classified as cases is the year of diagnosis of dementia. This is necessarily later than the year of dementia onset, but may be determined using objective criteria as compared to the year of dementia onset.

Participants classified as deaths died during the follow-up period and were determined to be dementia-free by (i) post-mortem dementia review or (ii) post-mortem medical record review confirming the absence of cognitive decline. The last year of follow-up for this group is their year of death.

All other members of the dementia cohort are classified as censored. This group contains participants who are dead and (i) have not yet had a post-mortem medical record review, (ii) were referred for dementia review after the post-mortem medical record review but have not yet been reviewed by the dementia review panel, or (iii) were not identified as demented at initial review by the dementia review panel but need a post-mortem re-review. In addition, this group contains participants who are alive and (i) have never have been flagged for cognitive evaluation, or (ii) were flagged for cognitive evaluation and are being monitored according to the neuropsychology tracking protocol. The participants in this group can be in any stage of the dementia surveillance protocol but are all censored in the last year they could be classified as not demented. This is the year of their last normal neurology exam (in which they were classified as not demented or as

mildly demented) or the year of their last normal MMSE (as defined using the education-based cut-off), whichever is later.

4. METHODS

In this section we establish notation and provide the formulae that are operationalized in the macro. The description of the SAS macro follows in Section 5.

4.1. Data structure

Alzheimer's disease develops insidiously over time as compared to other events, such as stroke, that occur suddenly. For this reason, its development is reported using a year (rather than a specific date) of diagnosis. To be consistent, we use years as the measure of time for all analyses. Consider the following three subjects:

Subject	Date of birth	Year of AD diagnosis	Date of death
A	07/03/1903	1980	XXX
B	01/23/1905		11/29/1978
C	03/03/1905	1975	06/22/1989

The traditional data structure for a life-table analysis of these data would consider each subject's contribution in terms of years on study (starting in 1975). The risk set would initially contain all subjects and its size would be a non-decreasing function of time. For example, subject A is at risk of developing AD at entry in 1975 and entering each subsequent year to 1980. Subject A thus contributes six years of follow-up, with an event in the sixth year. Subject B contributes four years and is censored in the fourth year in 1978. Subject C contributes one year (1975) with an event in that year. The traditional data structure for these three subjects is as follows:

Years on study	Number of subjects entering risk set
1	3
2	2
3	2
4	2
5	1
6	1
7	0

We can reorganize the data with age as the time scale and also allow for the left-truncation necessary to account for subjects entering at different ages. Each member of the dementia cohort provides one observation with the following variables:

(i) *entry age*, age, in years, at entry in the observation period (that is, age in 1975);
(ii) *survival age*, age, in years, in their last year of follow-up;

(iii) *event status*, an indicator variable coded 1 for subjects who developed AD during the observation period and 0 for those who did not;
(iv) *any event status*, an indicator variable coded 1 for subjects who either developed AD or died during the observation period and 0 for those who did neither.

For example, subject A has *entry age* = 72 (1975–1903); *survival age* = 77 (1980–1903); *event status* = 1; and *any event status* = 1. This subject contributes 6 years of data at ages 72, 73, 74, 75, 76 and 77. Notice that subject A is considered to contribute an entire year of age 72 follow-up (in 1975) although this subject will not actually be 72 until 3 July 1975. We also consider that subject A becomes demented at age 77 (in 1980) when the age at diagnosis could really be as young as $76\frac{1}{2}$. It is not possible to more accurately assign a time of diagnosis and we choose calendar year to index subjects' ages.

The second subject, subject B, has *entry age* = 70 (1975–1905); *survival age* = 73 (1978–1905); *event status* = 0; *any event status* = 1, and contributes 4 years, at ages 70, 71, 72 and 73. Subject C has *entry age* = 70 (1975–1905); *survival age* = 70 (1975–1905); *event status* = 1; and *any event status* = 1. This subject contributes one year at age 70.

We define the following notation to summarize the data structure we will use to perform analyses with age as the time scale:

r_A = the number of persons at risk entering age A;
e_A = the number of incident events of interest (for example cases of AD) at age A;
w_A = the weighted number of persons at risk entering age A, computed by assigning a weight of 0.5 observations censored free of the event of interest during age A and a weight of 1 to all other observations (the actuarial method);
c_A = the number of persons who fail due to either the event of interest, or due to the competing risk (here, death).

In this case the size of the risk set is not a non-decreasing function of time; rather it decreases as subjects fail or are censored, but increases as subjects age into the study period. A summary of data on the three subjects in the example above is as follows:

Age	r_A	e_A	w_A	c_A
70	2	1	2	1
71	1	0	1	0
72	2	0	2	0
73	2	0	1.5	1
74	1	0	1	0
75	1	0	1	0
76	1	0	1	0
77	1	1	1	1

4.2. One-year incidence rates by age group

We summarize the age-specific data described in Section 4.1 by collapsing the age-specific data into age groups. The one-year incidence rate (per 1000 person-years) in each age group G is

calculated as the total number of incident events divided by the total number of (weighted) person-years at risk, times 1000:

$$\mathrm{IR}_G = 1000 \times \{(\Sigma_{A \in G}(e_A))/(\Sigma_{A \in G}(w_A))\} \tag{1}$$

where the summations are over the ages, A, included in the age group. For example, the 1-year incidence rate per 1000 person-years for the age group 70–74 is

$$\mathrm{IR}_{70\text{–}74} = 1000 \times \{(1)/(2 + 1 + 2 + 1.5 + 1)\} = 1000(1/7.5) = 133.33.$$

The crude 1-year incidence rate (over all ages) is estimated as the total number of incident events divided by the total number of person-years at risk times 1000:

$$\mathrm{IR}_C = 1000 \times \{(\Sigma_{\mathrm{all}\, A}(e_A))/(\Sigma_{\mathrm{all}\, A}(w_A))\} \tag{2}$$

where the summations are over all ages.

4.3. Age-adjusted rates

We use direct standardization to calculate age-adjusted rates for comparison of rates among levels of a grouping variable. The combined group (over all levels of the grouping variable) is used as the standard population. For example, if the grouping variable has two levels and we define, in age group G, the 1-year incidence rates in levels one and two, IR_{1G} and IR_{2G}, respectively, then the age-adjusted rates are

$$\mathrm{IR}_{1A} = \Sigma_{\mathrm{all}\, G}(\mathrm{IR}_{1G})(p_G)$$

and

$$\mathrm{IR}_{2A} = \Sigma_{\mathrm{all}\, G}(\mathrm{IR}_{2G})\,(p_G) \tag{3}$$

where $p_G = \{w_G/(\Sigma_{\mathrm{all}\, G}\,(w_A)\}$ is the proportion of (weighted) person-years in age group G.

4.4. Kaplan–Meier estimate of unadjusted cumulative incidence (UCI)

Suppose that $t_1 < t_2 < \ldots < t_j < \ldots < t_J$ are the ordered failure times among N subjects ($J < N$), e_j is the number of patients who fail from the event of interest (AD) at time t_j, and r_j is the number of patients at risk at time t_j (their failure or censoring times are greater than or equal to t_j). The Kaplan-Meier estimate of survival beyond time t (for example, probability of not developing AD) is given by

$$\hat{S}(t) = \Pr\{T > t\} = \prod_{j=1}^{k}\left(1 - \frac{e_j}{r_j}\right)$$

where k is the largest j such that $t_j < t$, $r_1 = N$ and $h_j = e_j/r_j$ is the estimate of the hazard, or conditional probability, of developing AD at time t_j given survival beyond time t_{j-1}. A perhaps more intuitive relationship between the hazard and survival functions involves $\hat{f}(t_j)$, the unconditional probability of failure at time t_j:

$$\hat{f}(t_j) = h_j \hat{S}(t_{j-1})$$

Thus the unconditional probability of failing at time t_j is the product of the conditional probability of failing at time t_j given survival beyond t_{j-1}, and the probability of surviving

beyond time t_{j-1}. The cumulative incidence function is

$$\hat{F}(t) = \sum_{j=1}^{k} \hat{f}(t_j)$$

where k is the largest j such that $t_j < t$, and the survival function is $\hat{S}(t) = 1 - \hat{F}(t)$. This recursive method for calculating the survival function and the cumulative incidence function begins at time t_1 and we define $\hat{S}(t_0) = \hat{S}(0) = 1$.

The traditional Kaplan–Meier method assumes that the time scale for failure times is time on study or some other function of calendar time. It can, however, be modified for use with a survival age time scale:

$$\hat{f}_A = h_A \hat{S}_{A-1} \tag{4}$$

where $\hat{S}_A$is the probability of surviving beyond age A. Notice that the hazard of developing the event at age A, h_A, is zero for ages less than $A_{\min}$, the youngest age at which the event occurs. In our case, the earliest diagnosis of AD was for a subject who was 70 and so $A_{\min} = 70$. Then the cumulative incidence at age A is

$$\hat{F}_A = \sum_{j=A_{\min}}^{A} \hat{f}_j = \sum_{j=A_{\min}}^{A} h_j \hat{S}_{j-1} \tag{5}$$

and the survival function is $\hat{S}_A = 1 - \hat{F}_A$. As was noted in Section 4.1, the size of the risk set is not necessarily non-decreasing; however, this does not impact the calculation of $\hat{S}_A$.

The *remaining lifetime risk* of failure from the event of interest is simply the cumulative incidence, $\hat{F}_{A_{\max}}$, where $A_{\max}$ is the maximum age. This method can easily be modified to condition on survival to a particular age, A_S, by setting $\hat{S}_A = 1$ for $A < A_S$, and by summing from $j = A_S$ instead of from $j = A_{\min}$. In this way, we can calculate the remaining lifetime risk of failure conditional on survival to age A_S. For example, we can calculate the remaining lifetime risk of developing AD for a cognitively intact 70-year-old or 75-year-old.

The variance of the estimated cumulative incidence at age A is given by Greenwood's formula [2, 28]:

$$\text{var}(\hat{F}_A) = \hat{S}_A^2 * \sum_{j=A_{\min}}^{A} \left(\frac{e_j}{r_j * (r_j - e_j)} \right)$$

95 per cent confidence limits can be constructed in the usual way as

$$\hat{F}_A \pm 1.96 \sqrt{\{\text{var}(\hat{F}_A)\}} \tag{6}$$

Note that in the calculation of $\hat{F}_A$ and its variance, deaths are treated as censored events. The estimate is not adjusted for the competing risk of death. For this reason, we term $\hat{F}_A$ the unadjusted cumulative incidence or UCI.

4.5. Cumulative incidence adjusted for competing risk (ACI)

We now return to the issue of competing risks. As discussed in Section 2.4, the unadjusted cumulative incidence, or UCI, overestimates the risk of actually developing AD. The method we

use to adjust the cumulative incidence of AD adjusted for the competing risk of death is described in detail by Gaynor *et al.* The adjustment is to the estimate of unconditional probability of failure, $\hat{f}_A$. In equation (4), the unconditional probability of failure at age A is the product of the hazard of failure at age A given survival to age (A-1). The estimate of the probability of survival to age (A-1) is based on an analysis in which deaths are censored and thus is an estimate of the probability of survival AD-free but not necessarily alive. A more appropriate condition is survival free of AD *and alive.* This probability may be obtained by performing a survival analysis (as described in Section 4.4) in which deaths are not censored but are counted as events of interest along with AD events. Using the notation in Section 4.1, the hazard of failing from either AD or death at age a is (c_a/r_a). The estimated survival probability from such an analysis is:

$$\hat{U}_A = 1 - \sum_{j=A_{\min}}^{A} (c_j/r_j)\,\hat{U}_{j-1}$$

Then we modify equation (4) as follows:

$$\hat{f}_A^* = h_A \hat{U}_{A-1} \tag{7}$$

The ACI, or cumulative incidence adjusted for the competing risk of death, is

$$\hat{F}_A^* = \sum_{j=A_{\min}}^{A} h_A \hat{U}_{A-1}. \tag{8}$$

The variance of the ACI can be estimated using a Taylor series linear expansion [12]:

$$\mathrm{SE}(\hat{F}_A^*) = \sqrt{\left\{ \sum_{j=A_{\min}}^{A} h_j \hat{U}_A \times \frac{(r_j - e_j)}{(e_j \times r_j)} + \sum_{1=A_{\min}}^{i-1} \frac{c_1}{r_1(r_1 - c_1)} + 2 \sum_{j=A_{\min}}^{A-1} \sum_{k=j+1}^{A} h_j \hat{U}_A \times h_k \hat{U}_A \left[\frac{-1}{r_j} + \sum_{1=A_{\min}}^{j-1} \frac{c_1}{r_1(r_1 - c_1)} \right] \right\}}$$

5. MACRO DESCRIPTION

In this section we describe our SAS macro *practical incidence estimators* (PIE) which provides estimates of age-specific incidence rates, crude and age-adjusted incidence rates, estimates of the unadjusted cumulative incidence and cumulative incidence rates adjusted for competing risk of death. The macro also provides the remaining lifetime risk of developing the event of interest conditional on survival to selected ages. We describe the macro parameters and the modules (sub-macros) it calls.

5.1. Preparing the data

The syntax used to call the SAS macro *practical incidence estimators* is as follows:

%macro PIE (IDS, minage, maxage, agegrpw, group, level1, level2, agefree, o1,o2);

The ten parameters are:

1. The input data set (IDS). This data set must contain the following variables and one observation per subject
   (i) Study identification number (*id*): unique identification numbers to distinguish one subject in the sample from another.
   (ii) Entry age (*entryage*): age, in years, at the beginning of the observation period.
   (iii) Survival age (*survage*): age, in years, in the last year of follow-up (for example, the year of failure from the event of interest or the year of censoring).
   (iv) Event status (*status*): an indicator variable coded 1 for subjects who develop the event of interest during the observation period (for example, develop AD during 1975–1998) and 0 for subjects who do not. (Subjects who die during the observation period are coded 0 with respect to event status.)
   (v) Any event status (*astatus*): an indicator variable coded 1 for subjects who fail due to the event of interest *or* the competing risk during the observation period (for example, develop AD or die during 1975–1998) and 0 for subjects who do not.
   (vi) Grouping variable (*group*): a variable that defines comparison groups of interest (for example, subject gender).
2. *minage*: minimum age at event onset.
3. *maxage*: maximum age at event onset.
4. *agegrpw*: the width of (or number of years in) each age interval used in the incidence tables.
5. *group*: the variable which defines comparison groups of interest with levels as in points 6 and 7.
6. *level* 1.
7. *level* 2.
8. *agefree*: age to which subjects are assumed to be free of the event (used in calculating future risk conditional on survival to age *agefree*).
9. *o*1: the name of the SAS data set with estimates of UCI and ACI derived from data in *level* 1 of the variable *group*.
10. *o*2: the name of the SAS data set with estimates of UCI and ACI derived from data in *level* 2 of the variable *group*.

The Framingham dementia cohort data set ('addata') is the IDS used in the PIE macro. Recall that each subject contributes one observation with the five requisite variables given above. The grouping variable we will consider here is *male* which is coded 1 for men and 0 for women. The constant parameter values are as follows:

minage = 70, the youngest age at diagnosis in our data;
maxage = 99, the oldest age at diagnosis in our data;
agegrpw = 5 requests 5-year age groups in the tables of one-year incidence;
group = male;
level 1 = 1;
level 2 = 0;
agefree = 70;
*o*1 = out1;
*o*2 = out0.

Thus our macro call is

%macro PIE (addata, 70, 99, 5, male, 1, 0, 70, out1, out0);

The PIE macro contains the following modules, each of which is described below:

%macro PIE (IDS, minage, maxage, agegrpw, group, level1, level2, agefree, o1, o2);

```
  %data1 (&IDS, PDS)
  %sdsmac (PDS, (&level1, &level2), SDS)
  title "&group = &level1";
  %sdsmac (PDS, (&level1), SDS1)
  title "&group = &level2";
  %sdsmac (PDS, (&level2), SDS2)
  title;
  %incid (SDS, I, 1)
  %incid (SDS1, i1, 0)
  %incid (SDS2, i2, 0)
  %aa2g (i1, i2)
  title "&group = &level1";
  %lr (SDS1, &e o1)
  title "&group = &level2";
  %lr (SDS2, &o2)
%mend;
```

5.2. Module one: creating a pooled data set

In the first module, a data set is created with one line per subject for each year at risk during the observation period. Subjects are considered at risk at a given age during the observation period as long as they are free of the event of interest at least until that age. Only ages that are between the specified minimum and maximum ages are included. For example, suppose a person enters the study at age 50 in 1975 and is still alive and free of AD in 1998. If we specify the minimum and maximum ages to be 70 and 99, respectively, this person contributes 4 years of risk, starting in 1995 at age 70, to 1998 at age 73. The pooled data set excludes those years when subjects were free of the event of interest but were observed at ages less than the specified minimum age at event onset. With respect to our example subject, the first 20 years of observation, when the subject was aged 50–69, are excluded from the pooled data set. The pooled data set also excludes ages at risk that occurred before the observation period.

The calculation of one-year incidence rates uses a weighting scheme similar to the strategy employed in the construction of actuarial estimates (Section 4.1). In most cases, the last year of follow-up for subjects who are censored free of AD is assigned a weight of 0.5. However, subjects who are censored at the maximum age to be considered but who are known to have survived free of AD to an older age are assigned weights of 1.0 in their last year of follow-up. All other observations are assigned weights of 1.0.

The module *data* 1 is called as follows, where *in* and *out* are the input and pooled data sets, respectively. The PIE macro invokes this module to produce the pooled data set *PDS* based on the initial data set *IDS*.

```
%macro data1 (in, out);
  data &out;set &in;
    keep id age status astatus weight &group;
    if survage gt &maxage then do;
      survage = &maxage; status = 0; astatus = 0; full = 1;
    end;
    start = max(entryage, &minage);
    stop = survage-1;
    age = survage;
    if status eq 0 and full ne 1 then weight = 0.5;else weight = 1;
    output;
    do age = start to stop;
      status = 0; astatus = 0; weight = 1; output;
    end;
    run;
%mend;
```

5.3. Module two: creating a summary data set for each age

In the second module, the pooled data set (created in module one) is summarized for each year of age. Specifically, for each year of age, A, between the minimum and maximum ages, the module outputs r_A, the number of subjects in the risk set, e_A, the number of events of interest, c_A, the total number of events (including the competing risk) and w_A, the weighted number of person years at risk (see Section 4.1). In addition, the module outputs the sum over all ages of each of r, e, c and w.

The PIE macro invokes the module three times, each time inputting the pooled data set *PDS* created in module one. The first call produces a summary data set, *SDS*, for all subjects combined. The next two calls of the module produce summary data sets *SDS*1 and *SDS*2, for the two levels of the comparison group of interest. In our example, SDS is a summary data set for men and women combined, and *SDS*1 and *SDS*2 are summary data sets for men and women, respectively.

```
%macro sdsmac (in, level, out);
  proc means noprint n sum data = &in;
    where (&group in &level);
    class age;
    var status astatus weight;
    output out = t n = r sum = e c w;
  run;
  data &out; set t;
    if age eq . then age = 999;
    proc sort; by age; run;
%mend;
```

5.4. Module three: computing age-specific incidence rates

In the third module, each summary data set output in module two (*SDS*, *SDS*1 and *SDS*2) is used to calculate one-year incidence rates. First, ages are collapsed into intervals of specified width (for example, if *minage* = 65, *maxage* = 94, and *agegrpw* = 5, ages are collapsed into 5-year