# 12

# *Analysis of Biased Survival Data: The Canadian Study of Health and Aging and beyond*

**Masoud Asgharian, Christina Wolfson, and David B. Wolfson**

*McGill University, Montréal, QC*

Often researchers are faced with analyzing data from a biased sample that is unrepresentative of the population of interest. Studies that lead to samples that are biased-by-design are only useful if statisticians have a way to compensate for biased procedures that would result from naïve use of such non-representative data. This chapter tells the story of how we encountered biased survival data in a major Canadian study of dementia in the elderly. We relate how we overcame the problem of bias and developed methods to answer questions about dementia. Although the story is woven around the Canadian Study of Health and Aging, there are many other areas in which the type of bias that we encountered also arises.

## 12.1 Introduction

Epidemiologists define the prevalence of a disease as the number of individuals with the disease per 100,000 in the general population at a given time. They define the incidence rate to be the number of new cases of the disease per 100,000 individuals per unit of time (usually per year). A shocking statistic is that roughly 40% of Canadians over the age of 85 have some form of dementia. Consequently, as life expectancies of Canadians rise there will, in the foreseeable future, be a corresponding increase in both the incidence and prevalence of dementia. Not only will our health care system be hard-pressed to cope, there will also be a growing burden on caregiver families; most of us will be affected either directly or indirectly. It is vital, therefore, that we understand the natural history of dementia; that is, how it evolves in those stricken with it, and the factors that may hasten or slow its progression. At the population

level, it is equally important to know how dementia incidence rates are changing and how this change, when combined with improving life expectancy, will affect the population burden of dementia. Further, we can only assess whether treatments or changes in lifestyle prevent or slow the course of a disease if we are able to measure changes in disease duration and incidence in populations.

Statisticians and epidemiologists collaborate to design studies and use the data collected from them to reveal patterns of changing incidence, and disease duration. With these goals in mind, in 1991 the Canadian Study of Health and Aging (CSHA) was launched (CSHA, 1994). A primary goal of the CSHA was to estimate the prevalence of dementia in elderly Canadians. Initially designed as a one time cross-sectional study, the investigator team obtained research funds to conduct two follow-ups on the participants over the subsequent decade.

Although there are many different types of dementia the two most common forms are those due to Alzheimer's disease and vascular dementia. Indeed, current research indicates that two thirds of all dementias that occur in the elderly are due to Alzheimer's disease. In this chapter, for simplicity, by the term "dementia" we shall mean either Alzheimer's disease or vascular dementia. Our story does not begin with incidence or prevalence but rather with the questions, "How long do people with dementia survive following onset of their disease?" and "What factors are associated with shorter (or longer) survival?"

These two questions were first posed by one of us, Christina Wolfson, Principal Investigator for the CSHA Progression of Dementia Study. Anticipating the use of data from the CSHA, she was concerned about the bias that invariably accompanies estimates of survival based on data collected from a study designed to follow-up a cohort of prevalent cases, i.e., cases that already have dementia at the time of their recruitment. Such studies are descriptively called prevalent cohort studies with follow-up. The CSHA was such a study.

In 1991 roughly 10,000 Canadians over the age of 65 were recruited and those living in the community were screened for dementia. All participants living in institutions, and those living in the community who screened positive at baseline were invited to undergo a thorough assessment for dementia that involved a battery of neurological and neuropsychological tests. Roughly 820 were diagnosed with prevalent (current) dementia in one of the two main categories above.

Although the full cohort of 10,000 was then followed forward, for the moment we focus on the 820 with prevalent dementia. By 1996, the end of the first phase of the CSHA, many of the 820 had died (and their dates of death recorded), and most of the rest were known to be still alive — said to have right censored survival times. (A survival time is said to be censored when only a lower bound of its value is known.) A small proportion of study subjects had been lost to follow-up in between 1991 and 1996 and the dates at which they were last known to be alive were recorded. Their survival times are also right censored.

In 1991 each caregiver provided an approximate date of dementia onset as well as covariate information, such as age at onset and number of years of education. That is, apart from their covariates, by 1996 each subject had contributed a survival time (possibly right censored) consisting of the time interval from their date of onset to the minimum of the date of death and their date of right censoring (see Figure 12.1).

To estimate how long people with dementia live from onset of their disease such data cannot be used without adjustment, because the sample is biased. The bias stems from the fact that in order to be among the 820 subjects identified with dementia in 1991 one would have to survive long enough to have a chance of being recruited into the study. For example, consider two subjects who had the same date of onset, say in 1988. If one of these subjects had died before 1991 they could not have become part of the CSHA, while the second subject would have been recruited into the CSHA if they had survived longer than three years. That is, the longer survivor would have been recruited.

This phenomenon occurs whenever prevalent cases are identified through a cross-sectional survey and then followed forward with a view to estimating survival from a meaningful origin such as date of onset, the origin that we shall use in this chapter. The observed survival times are said to be left-

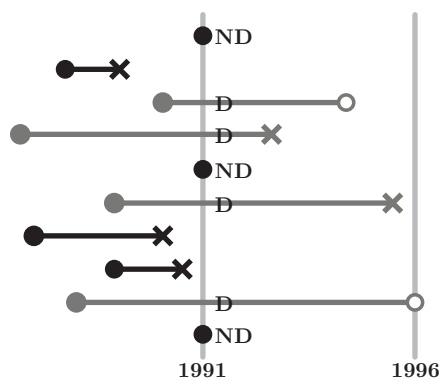



FIGURE 12.1: A schematic demonstrating the sampling mechanism and follow-up for the first phase of the Canadian Study of Health and Aging. The letters "D" and "ND" by the vertical line together depict some of the roughly 10,000 who were screened for dementia in 1991. Letters "D" depict those with prevalent dementia. Their ascertained onset times prior to 1991 are denoted by gray dots and their times of death/censoring, obtained from follow-up after 1991, are denoted by gray crosses/gray circles. The letters "ND" depict those without dementia in 1991. Also pictured are subjects with dementia who died prior to 1991; their unobserved onset times are denoted by black dots and their unobserved failure times are denoted by black crosses.

truncated, and if the incidence rate is constant (called the stationary case) the survival times are termed length-biased. It should be noted, however, that left-truncated survival times are, in general, length-biased and the two terminologies just provide a convenient way of differentiating between when the incidence process is stationary and when it is non-stationary (that is, the incidence rate changes over time).

Our first goal was to estimate the survivor function of those who develop dementia; we had carried out a literature search and found that based exclusively on prevalent cohort studies with follow-up, published estimated median survival times with dementia (for the two types combined) ranged from around 5 to 9 years; see Mölsä et al. (1986), Walsh et al. (1990) and Stern et al. (1997). Although the most recent of these studies mentioned survivor bias, none had included analyses that adjusted for it and we felt that proper adjustment would reduce the currently accepted range of median survival times.

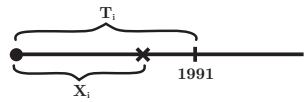
## 12.2 Nonparametric Estimation of the Survivor Function

We begin with some basic terminology and notation. Let $X_i$ be the survival time of subject $i$, measured from onset of dementia, had they been observed as an incident case (that is, a subject in an initially disease-free cohort, whose onset of disease is observed). Let $S_U(t) = \Pr(X_i > t)$ be the survivor function of $X_i$, the function that we wish to estimate. In a prevalent cohort study with follow-up there is an underlying incidence process that must be taken into account. Each point of incidence defines a random truncation time $T_i$, the time from the date of incidence to the date of recruitment. However, their onsets are not all observed and hence, neither are their truncation times (see Figure 12.2a).

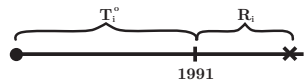
Let $T_i^0$ be the observed (left) truncation time of a subject that is recruited. Other commonly used terminology for the observed time interval between disease onset and recruitment into the prevalent cohort (the observed truncation time), is the backward recurrence time. The forward recurrence time, $R_i$, is the time interval from recruitment until death (see Figure 12.2b). This might be censored by its associated censoring time $C_i$ (see Figure 12.2c).

Let $S_{LB}$ denote the survivor function of the observed length-biased survival times. It is not difficult to show (Asgharian et al., 2006) that there is a simple relationship between $S_U$ and $S_{LB}$, viz.

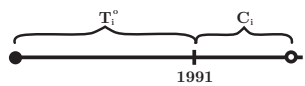
$$S_{LB}(t) = \frac{\int_t^\infty \Pr(T \le x) \mathrm{d}S_U(x)}{\int_0^\infty \Pr(T \le x) \mathrm{d}S_U(x)}. \tag{12.1}$$




(a) An unobserved survival time $X_i$ with its longer associated truncation time $T_i$, which is also not observed.




(b) An observed survival time $X_i^0 = T_i^0 + R_i$, which is the sum of the observed truncation time from onset to 1991 (backward recurrence time) and the time from 1991 to death, for a prevalent case (forward recurrence time).




(c) An observed truncation time $T_i^0$ from onset to 1991 with its associated forward censoring time $C_i$. By definition, censoring can only occur during follow-up of prevalent cases.

FIGURE 12.2: Illustration of truncation, failure, and censoring times.

Let $X_i^0 = T_i^0 + R_i$ be the left-truncated survival time, with survivor function $S_{\text{LB}}$. Since $X_i^0$ is only observed if $X_i \geq T_i$, we have that

$$S_{LB}(t) = \Pr(X_i > t | X_i \geq T_i).$$

Under stationarity, Expression (12.1) becomes

$$S_{LB}(t) = \int_t^\infty x \, \mathrm{d}S_U(x) \Big/ \int_0^\infty x \, \mathrm{d}S_U(x), \tag{12.2}$$

a relationship that is of central importance in this account.

Since the forward recurrence times are possibly right censored by their corresponding forward censoring times, the full data observed for $n$ subjects are $\{T_i^0, \min(R_i, C_i), \delta_i\}$ for $i \in \{1, \ldots, n\}$, where $\delta_i = \mathbf{1}(R_i \leq C_i)$ is the censoring indicator. The censoring indicator takes the value 1 if $R_i \leq C_i$ and is 0 otherwise. Ignoring covariates for the moment, the CSHA survival data from the prevalent dementia cases were of this form, where $T_i^0$ corresponded to the interval between the date of onset of dementia and the date of recruitment into the CSHA, and the end point was death from any cause/censoring. In our initial analysis we exploited Expression (12.2) to find the nonparametric maximum likelihood estimator (MLE) of $S_U$ by substituting the MLE of $S_{LB}$, obtained directly from the observed length-biased data. That is, we obtained an estimator of $S_U$ that does not assume any specific parametric form for $S_U$.

We chose this route rather than use the more robust product limit estimator of Tsai et al. (1987) that does not depend on stationarity, because there was no reason to believe that the incidence rate of dementia had increased much, if at all, in the approximately twenty years prior to 1991. If the incidence rate of dementia could be regarded as roughly unchanged, we felt that with these stronger model assumptions we would obtain a more efficient estimator of the survivor function than the product limit estimator (that is, an estimator with a smaller variance). To our surprise, although other researchers such as Vardi (1982) had found the Nonparametric Maximum Likelihood Estimator (NPMLE) of $S_U$ under stationarity, none had allowed for censoring. Alas, our first attempts ended in failure. For, we fell into the trap of using the Kaplan–Meier estimator as the NPMLE of the length-biased survivor function, $S_{LB}$; after all, did we not have randomly right censored survival data from $S_{LB}$?

Our mistake, pointed out by a referee, was that censoring for length-biased data obtained from a prevalent cohort study with follow-up is informative. Censoring is informative when censoring and survival times are not independent, as required by the methodology. Fortunately, we were able to repair the damage by using the correct NPMLE for $S_{LB}$. This had previously been obtained by Vardi (1989), although in a context entirely different from that of a prevalent cohort study with follow-up. Therefore, when we came to work out the asymptotic properties (that is, the properties in the ideal situation where samples are very large) of our NPMLE we ran into another problem. Clearly, from Expression (12.2) these properties would follow from the asymptotic properties of the NPMLE of $S_{LB}$ and, even though Vardi's likelihood and our likelihood were proportional, the sampling properties of an MLE depend on how the data are obtained. We could not use the results of Vardi and Zhang (1992), who later derived the asymptotic properties of the NPMLE under multiplicative censoring. However, we were able to learn from their methods. The paper by Asgharian et al. (2002) in which we present our research also demonstrates the markedly smaller variance of the NPMLE that exploits the assumed stationarity of the incidence process over the product limit estimator, which does not require the assumption of stationarity. Our first article was followed by a detailed investigation (Asgharian and Wolfson, 2005) of various censoring mechanisms and a rigorous presentation of the arguments underpinning our paper.

While we had embarked on our methodological research in order to estimate the survival distribution from the onset of dementia until death, when we submitted our CSHA paper to the *New England Journal of Medicine* (NEJM) our methodology paper had not yet been published. The Editor of the NEJM therefore, required us to conduct our analyses using the product limit estimator, which was the standard method for analyzing right censored left-truncated survival data. We estimated median survival from onset of dementia to be 3.3 years. This was considerably less than had previously been thought and it caused quite a stir in the media.

Figure 12.3 is taken with permission from the NEJM (Wolfson et al., 2001), and demonstrates the difference between the unadjusted estimator of survival and the correct product limit estimator. A possible explanation, put forward by some, for the difficult-to-believe 3.3 years was that in the CSHA, those at least 85 years of age had been oversampled. However, without adjustment the estimated median survival of 6.6 years was consistent with the estimated median survival in other studies, obtained without adjustment. Also, we re-estimated the survivor function assuming a constant incidence rate for dementia, using our new approach, and obtained a similar estimated median of 3.75 years, although with a much narrower confidence interval (Asgharian et al., 2002). This was not surprising since the method we employed for the NEJM article and our newly developed approach yield consistent estimators, which have the property that they will be close to the target with high probability for large sample sizes. Following its publication, with an accompanying editorial, our NEJM paper has been cited over 400 times.

An incident cohort study, although logistically difficult and more expensive, is the gold standard for estimating disease incidence and the survival distribution. Under this design, a cohort of disease-free subjects is followed for the development of dementia. The dates of onset are recorded, and the cohort is followed over time until some end-of-study date. The dates of death or censoring are recorded for all who become incident cases while under follow-up. Of relevance are three recent studies, based on incident cohorts (Helzner et al., 2008; Xie et al., 2008; Matsui et al., 2009), which have confirmed our findings that survival with dementia is much shorter than previously thought. Further, stratification on type of dementia did not produce any anomalous results; the effect of length bias was strong, and survival from onset, short.

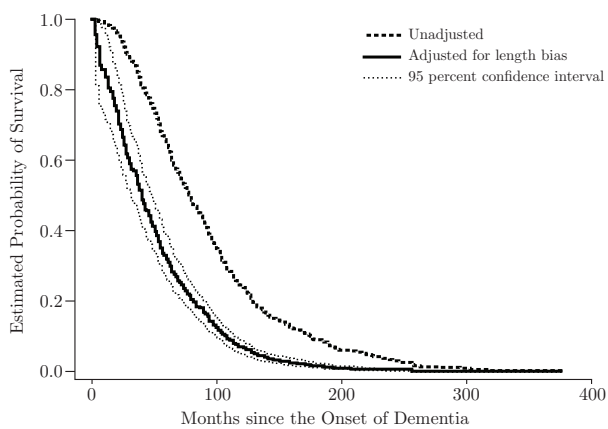



FIGURE 12.3: Estimated probability of survival, without adjustment (dotted line) and with adjustment (solid line, including 95% confidence bands).

Since in the NEJM article we had used the product limit estimator for left-truncated data we compared the point estimates given there, with our estimates based on the stationarity assumption; see Asgharian et al. (2002).

## 12.3 Checking for Stationarity of the Incidence Process

Carrying out research is akin to trying to slay the mythical Hydra; a problem solved creates two new problems. So it was with us. Our NPMLE and its asymptotic sampling properties had been developed under the assumption of a stationary incidence process. In order to trust our methods for the CSHA survival data we had to check our assumption that the incidence rate of dementia had remained fairly constant over the time period leading up to 1991. More generally, we wondered if one could test for stationarity using only data collected from a prevalent cohort study with follow-up. Wang (1991) had suggested how one might check for stationarity graphically using her estimator of the truncation time distribution; under stationarity this should be uniform. However, she proposed no formal test.

The problem is this: We must make inference about a stochastic process (the incidence process) based on incomplete information. Data from some individuals are missing because we only observe the onsets of those who survive to be recruited into the prevalent cohort. Equivalently, we need to make inference about the distributions of the $T_i$'s based on an observed subset of them, the $T_i^0$'s (see Figure 12.4). Our idea was this: From renewal theory (Karlin and Taylor, 1975) it is known that under stationarity of a renewal process the forward and backward recurrence times should have the same distribution. Even though our setting was not exactly that of renewal processes some of the results from renewal theory also hold in the setting of prevalent cohort studies with follow-up, albeit with different proofs. In particular we were able to show that the forward and backward recurrence time distributions are equal if and only if the full incidence process is stationary. Conversely, if the incidence rate is increasing one would expect an excess of short backward recurrence times and long forward recurrence times. This suggested a simple graphical check for stationarity: On the same axes plot the Kaplan–Meier estimates of the forward and backward recurrence times, respectively, and assess their similarity; see Asgharian et al. (2006). An application of our simple procedure to the CSHA data supported our speculation that the incidence rate of dementia did not appear to have changed over roughly a twenty year period prior to 1991.

Now, although like Wang, we had proposed a simple graphical check for stationarity, this would have limited value unless we could also put forward a formal test for stationarity. One of our PhD students, Vittorio Addona, addressed this problem as part of his doctoral dissertation. The difficulties that had to be overcome were that forward and backward recurrence times

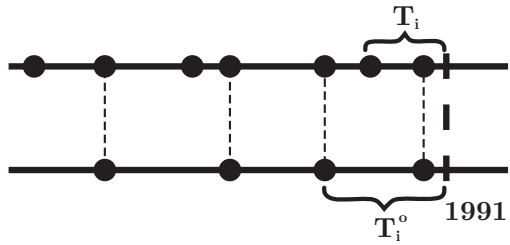



FIGURE 12.4: All onset times (top line) and observed onset times (bottom line).

are dependent. Further, forward recurrence times are often right censored. The dependence and censoring precluded a conventional test of equality of two distributions. A search of the literature, however, yielded a nonparametric Wilcoxon rank sum type test due to Wei (1980), that could be adapted to our situation. We were able to confirm what we had seen graphically: the CSHA data were not inconsistent with a roughly constant incidence rate of dementia (Addona and Wolfson, 2006).

## 12.4 Estimating the Incidence Rate

A new problem now emerged. One of the goals of the CSHA had been to estimate the incidence rate of dementia by following, for five years, the subjects that had been declared dementia-free in 1991. Ideally this cohort would have been monitored closely so that dates of dementia onset could be recorded near the time that they occurred. However, this cohort was only revisited in 1996, and information about those who had died between 1991 and 1996 was scant; in many cases there was considerable uncertainty whether they had died with dementia, and even if it was decided that they had (by means of an algorithm based on logistic regression), it was impossible to ascertain their dates of onset. By relying on subjects with prevalent disease only, we were able to avoid the ambiguity of whether or not subjects had dementia at the time of death. Although the date of onset for dementia is difficult to pin down, we at least had access to caregivers at the time that the subjects were diagnosed. By using a simple two-stage procedure that relied on the recollections of caregivers we obtained "best guess" dates of onset for each subject (Rouah and Wolfson, 2001).

Armed with this information we sought to exploit the well known epidemiological relationship between the prevalence odds, $P/(1-P)$, the (constant)

incidence rate, $\lambda_U$, and the mean disease duration, $\mu_U$ (of the time from onset of dementia to death), to estimate the incidence rate, viz.

$$\frac{P}{1-P} = \lambda_U \times \mu_U. \tag{12.3}$$

The proportion of the 10,000 sampled CSHA participants diagnosed with dementia in 1991 gave us an estimate of $P$, the prevalence of dementia. Then, using Expression (12.3) above, since we had estimated the survivor function, we were able to estimate $\mu_U$, using $\hat{\mu}_U = \int_0^\infty \hat{S}_U(s)\mathrm{d}s$, and hence the incidence rate as

$$\hat{\lambda}_U = \frac{1}{\hat{\mu}_U} \times \frac{\hat{P}}{1-\hat{P}}\,. \tag{12.4}$$

Of course, overall incidence rates are of less interest than age-specific incidence rates and we next set about estimating these (Addona et al., 2009). This required a generalization of Expression (12.4) to allow for the changing age distribution in the general population, information that was obtained through Statistics Canada. The asymptotic properties of our estimators built on those of $\hat{S}_U$, which we had previously obtained in Asgharian and Wolfson (2005) and Asgharian et al. (2002). This was not the end of the incidence rate story, however, because Carone et al. (2012) have derived an estimator of the incidence rate without the restriction that it be constant.

Our research had been preceded by the work of Keiding (1991), who took a multi-state approach to the problem. He defined three states: disease-free, diseased, and death, and modeled the transitions between these using data that included no follow-up of the prevalent cases. In particular he showed how to estimate a constant incidence rate from such data. However, the avoidance of follow-up requires stronger model assumptions that may sometimes be difficult to justify.

## 12.5 Covariates

From the beginning, CSHA researchers had been interested in factors (covariates) associated with survival with dementia, such as sex, age at onset, years of education, and the presence of extrapyramidal signs and/or psychiatric symptoms. To investigate the association between selected factors and survival, in our NEJM article we chose to use a proportional hazards model with a Weibull baseline hazard (Lawless, 2002). That is, we assumed a model relating covariates to survival, as if our data had arisen from an incident cohort. Next, we corrected this miss-specified model by exploiting a well known relationship between a length-biased and unbiased model, to obtain a likelihood for our observed data. Finally, we maximized this likelihood with respect to the regression parameters inherited from the proportional hazards model. Having

originated from a proportional hazards model, the estimated parameters were easily interpreted. Had we begun with a proportional hazards model for the length-biased data, we would not have been able to interpret the estimated regression coefficients as easily in the model of ultimate interest, the model that describes survival in incident cohorts.

We found, not surprisingly, that older age at onset is associated with shorter survival, that females with dementia tend to live longer than males and that there was a trend toward those with vascular dementia having shorter survival than those with Alzheimer's disease. We found no effect of level of education on survival.

Our PhD student Pierre-Jérôme Bergeron investigated the effect of length-bias on the sampling properties of the estimated regression parameters (Bergeron et al., 2008). At first we thought that the estimators would be asymptotically biased for the true parameters because covariates that accompany length-biased survival times would, themselves, not be representative of covariates in the general population. For example, one observes longer survivors in a prevalent cohort and it is known that females in general tend to live longer than males. Therefore, in a prevalent cohort one is more likely to find an excess of females among the prevalent cases, and might conclude that females with dementia survive longer than do males. However, our suspicions were unfounded, as it was found that bias is not the issue. Rather, length-biased sampling could lead to a loss in efficiency in the parameter estimators.

## 12.6 Concluding Remarks

An assumption almost always made when analyzing length-biased survival data is that survival is independent of the date of onset. If one assumes stationarity then it is possible to test the assumption that the survival distribution did not change over time. This problem can be regarded as a dual to the problem of testing for stationarity assuming unchanging survival, and also depends on a comparison between the forward and backward recurrence time distributions; see Addona et al. (2012).

The history of left-truncated survival data is long and the range of applications broad. We have listed many of the important historical articles in the references below which are by no means restricted to medical settings. Since we began working in the field of length-biased data there has been a surge of methodological papers on the subject; some of these are also listed.

Our saga began with a scientific research question to be addressed through data collected as part of the CSHA: "What can be said about survival with dementia?" Much substantive and methodological research has resulted from this modest beginning. The Canadian Longitudinal Study on Aging (CLSA) (Raina et al., 2009), the largest aging study undertaken to date in Canada

and broader in scope than the CSHA, is now underway (`www.clsa-elcv.ca`). The CLSA will follow 50,000 Canadians aged 45 to 85 years for 20 years, to ascertain determinants not only of disease but also of healthy aging. Since the CLSA is a prospective study, and a large number of initially disease-free subjects will be followed over a long time period, it will permit the identification of incident cases of a number of late life diseases. This will facilitate the study of survival of these diseases through the gold standard of an incident cohort study. Indeed, the study design requires the implementation of several disease ascertainment algorithms, not only for dementia, but also for Parkinson's disease, epilepsy, diabetes, chronic airflow disruption and ischemic heart disease, among others. However, when recruited, not all participants will be disease-free, as some of them will have prevalent disease. Consequently, the CLSA is also a prevalent cohort with follow-up. For example, our first venture into the use of the CLSA is to investigate survival with Parkinson's disease from the follow-up of participants who have Parkinson's disease when they are recruited.

Will the CLSA provide ground as fertile as the CSHA for statistical research, and in particular for survival analysis? We are sure it will, through the necessity to extend survival models and methods for length-biased data to recurrent events, multivariate failure times, and joint inference for longitudinal and survival data.

## Postscript

Studies on biased sampling can be traced as far back as Wicksell (1925) and his corpuscle problem which is now a classical example in stereology. The next important contribution can perhaps be attributed to Fisher (1934) on bias induced by the method of ascertainment. Neyman (1955) identified a type of bias that epidemiologists often encounter and coined the term incidence-prevalence bias. This was followed by the work of Cox (1969) in the quality control of fabrics. Zelen and Feinleib (1969) identified biased sampling in screening for chronic diseases. For more recent examples of biased sampling, see Morgenthaler and Vardi (1986) in econometrics, Drummer and McDonald (1987) in botany, Nowell et al. (1988) in land valuation, Nowell and Stanley (1991) in marketing, Terwilliger et al. (1997) in genetics and linkage mapping, Gilbert et al. (1999) in causal inference, Feuerverger and Hall (2000) in applied physics, De Uña Álvarez (2004) in labor force studies, Kvam (2008) in nano-physics, and Leiva et al. (2009) in water quality.

In recent years, there has been considerable interest in statistical methods for length-biased survival data, resulting in several important papers. We cite five of these: Andersen and Keiding (2002), Mandel and Fluss (2009), Qin and Shen (2010), Huang and Qin (2012), and Shen and Cook (2013).

## Acknowledgments

The authors thank Ana Best for her invaluable technical assistance in preparing the manuscript. This work was supported in part by Discovery Grants from the Natural Sciences and Engineering Research Council to David Wolfson and to Masoud Asgharian, and by a grant from the Canadian Institutes of Health Research to Christina Wolfson and David Wolfson.

## About the Authors

**Masoud Asgharian** is an associate professor of statistics at McGill University. He earned BSc and MSc degrees from Shahid Beheshti University, Tehran, and a PhD from McGill in 1998. His main areas of interest are survival analysis, causal inference, variable selection, clustering and classification, change-point problems, and longitudinal data analysis. He is an associate editor for *The Canadian Journal of Statistics*.

**Christina Wolfson** is a professor in the Departments of Epidemiology, Biostatistics and Occupational Health and of Medicine at McGill University. She holds degrees in mathematics, statistics, and epidemiology and biostatistics from McGill. Her research interests include the design and analysis of observational studies, as well as the epidemiology of neurodegenerative disorders. She is a fellow of the American College of Epidemiology.

**David B. Wolfson** is a professor in the Department of Mathematics and Statistics at McGill University. He completed undergraduate studies and a master's degree at the University of Natal in Durban, South Africa. He obtained his PhD in statistics from Purdue University in 1974 and has been a faculty member at McGill since then. His current research is focused on survival analysis, change-point problems and optimal design, along with statistical applications to medicine, in particular to dementia and multiple sclerosis.

## Bibliography

Addona, V., Asgharian, M., and Wolfson, D. B. (2009). On the incidence–prevalence relation and length-biased sampling. *The Canadian Journal of Statistics*, 37:206–218.

Addona, V., Atherton, J., and Wolfson, D. B. (2012). Testing the assumptions for the analysis of survival data arising from a prevalent cohort study with follow-up. *International Journal of Biostatistics*, 8:Online Publication.

Addona, V. and Wolfson, D. B. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Analysis*, 12:267–284.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11:91–115.

Asgharian, M., M'Lan, C.-É., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association*, 97:201–209.

Asgharian, M. and Wolfson, D. B. (2005). Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. *The Annals of Statistics*, 33:2109–2131.

Asgharian, M., Wolfson, D. B., and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in Medicine*, 25:1751–1767.

Bergeron, P.-J., Asgharian, M., and Wolfson, D. B. (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association*, 103:737–742.

Canadian Study of Health and Aging Working Group (CSHA) (1994). Canadian study of health and aging: Study methods and prevalence of dementia. *Canadian Medical Association Journal*, 150:899–913.

Carone, M., Asgharian, M., and Wang, M.-C. (2012). Nonparametric incidence estimation from prevalent cohort survival data. *Biometrika*, 99:599–613.

Cox, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, pp. 506–527. Wiley, New York.

De Uña Álvarez, J. (2004). Nonparametric estimation under length-biased sampling and Type I censoring: A moment based approach. *Annals of the Institute of Statistical Mathematics*, 56:667–681.

Drummer, T. D. and McDonald, L. L. (1987). Size bias in line transect sampling. *Biometrics*, 28:13–21.

Feuerverger, A. and Hall, P. (2000). Methods for density estimation in thick-slice versions of Wicksell's problem. *Journal of the American Statistical Association*, 95:545–546.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6:13–25.

Gilbert, P. B., Lele, S. R., and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, 86:27–43.

Helzner, E., Scarmeas, N., Cosentino, S., Tang, M., Schupf, N., and Stern, Y. (2008). Survival in Alzheimer disease: A multiethnic, population-based study of incident cases. *Neurology*, 71:1489–1495.

Huang, C.-Y. and Qin, J. (2012). Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *Journal of the American Statistical Association*, 107:946–957.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*, Second Edition. Academic Press, New York.

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society, Series A*, 154:371–412.

Kvam, P. (2008). Length bias in the measurements of carbon nanotubes. *Technometrics*, 50:462–467.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*, Second Edition. Wiley, Hoboken, NJ.

Leiva, V., Sanhueza, A., and Angulo, J. M. (2009). A length-biased version of the Birnbaum–Saunders distribution with application in water quality. *Stochastic Environmental Research and Risk Assessment*, 23:299–307.

Mandel, M. and Fluss, R. (2009). Nonparametric estimation of the probability of illness in the illness-death model under cross-sectional sampling. *Biometrika*, 96:861–872.

Matsui, Y., Tanizaki, Y., Arima, H., Yonemoto, K., Doi, Y., Ninomiya, T., Sasaki, K., Iida, M., Iwaki, T., Kanba, S., et al. (2009). Incidence and survival of dementia in a general population of Japanese elderly: The Hisayama study. *Journal of Neurology, Neurosurgery & Psychiatry*, 80:366–370.

Mölsä, P. K., Marttila, R., and Rinne, U. (1986). Survival and cause of death in Alzheimer's disease and multi-infarct dementia. *Acta Neurologica Scandinavica*, 74:103–107.

Morgenthaler, S. and Vardi, Y. (1986). Choice-based samples: A non-parametric approach. *Journal of Econometrics*, 32:109–125.

Neyman, J. (1955). Statistics — Servant of all science. *Science*, 122:401–406.

Nowell, C., Evans, M. A., and McDonald, L. (1988). Length-biased sampling in contingent valuation studies. *Land Economics*, 64:367–371.

Nowell, C. and Stanley, L. R. (1991). Length-biased sampling in mall intercept surveys. *Journal of Marketing Research*, 28:475–479.

Qin, J. and Shen, Y. (2010). Statistical methods for analyzing right-censored length-biased data under the Cox model. *Biometrics*, 66:382–392.

Raina, P., Wolfson, C., Kirkland, S. A., Griffith, L. E., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C. M., Hogan, D., et al. (2009). The Canadian Longitudinal Study on Aging (CLSA). *Canadian Journal on Aging*, 28:221–229.

Rouah, F. and Wolfson, C. (2001). A recommended method for obtaining the age at onset of dementia from the CSHA database. *International Psychogeriatrics*, 13:57–70.

Shen, H. and Cook, R. J. (2013). Regression with incomplete covariates and left-truncated time-to-event data. *Statistics in Medicine*, 32:1004–1015.

Stern, Y., Tang, M.-X., Albert, M. S., Brandt, J., Jacobs, D. M., Bell, K., Marder, K., Sano, M., Devanand, D., Albert, S. M., et al. (1997). Predicting time to nursing home care and death in individuals with Alzheimer disease. *Journal of the American Medical Association*, 277:806–812.

Terwilliger, J. D., Shannon, W. D., Lathrop, G. M., Nolan, J. P., Goldin, L. R., Chase, G. A., and Weeks, D. E. (1997). True and false positive peaks in genomewide scans: Applications of length-biased sampling to linkage mapping. *The American Journal of Human Genetics*, 61:430–438.

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987). A note on the product limit estimator under right censoring and left truncation. *Biometrika*, 74:883–886.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10:616–620.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, 76:751–761.

Vardi, Y. and Zhang, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *The Annals of Statistics*, 20:1022–1039.

Walsh, J. S., Welch, H. G., and Larson, E. B. (1990). Survival of outpatients with Alzheimer-type dementia. *Annals of Internal Medicine*, 113:429–434.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86:130–143.

Wei, L. (1980). A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *Journal of the American Statistical Association*, 75:634–637.

Wicksell, S. D. (1925). On the size distribution of sections of a mixture of spheres. *Biometrika*, 17:84–99.

Wolfson, C., Wolfson, D. B., Asgharian, M., M'Lan, C.-É., Østbye, T., Rockwood, K., and Hogan, D. B., for the Clinical Progression of Dementia Study Group (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344:1111–1116.

Xie, J., Brayne, C., and Matthews, F. E. (2008). Survival times in people with dementia: Analysis from population based cohort study with 14 year follow-up. *British Medical Journal*, 336:258–262.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, 56:601–614.