UNIVERSITÉ DU QUÉBEC À MONTRÉAL

A COMPARISON OF PAST EPIDEMICS THROUGH CONDITIONAL SURVIVAL FUNCTIONS

DISSERTATION

PRESENTED

AS PARTIAL REQUIREMENT

TO THE MASTERS IN MATHEMATICS

BY

PARASTOO SEPIDBAND

FEBRUARY 2017

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

COMPARAISON D' ÉPIDÉMIES DU PASSÉ, À PARTIR DE FONCTIONS DE SURVIE CONDITIONNELLES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

PARASTOO SEPIDBAND

FEVRIER 2017

ACKNOWLEDGEMENT

I would like to place my sincere gratitude to my supervisor, Sorana Froda, for helping me during my studies at UQAM, for her great advice, for her patience, support, trust and encouragement.

I am grateful to the department of mathematics and statistics of UQAM, for their support and help. I also thank all the professors of the mathematics and statistics department of UQAM, especially professors with whom I took courses. I am also thankful to the informatics technician of the department, Giesele Legault, for her patience in solving technical problems.

I thank Hadi Bigdely, for reading and correcting my thesis many times. I appreciate my lovely family in Canada, my husband, Hadi, for his vast support, help, patience and love, my adorable daughter, Yasmine, who was born during my studies. She gave me lots of hope, motivation and energy. I am grateful to my wonderful parents, who always encourage me to learn and study. I also thank my siblings and my friends who always give me their support and motivation.

I would love to dedicate this draft, to my dear close and supportive cousin, Amin, who passed away during writing this thesis, due to brain cancer.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

RÉSUMÉ

Dans ce mémoire de maîtrise, nous faisons une analyse de la grippe (Influenza Like Illness, ILI) à partir des données qui sont disponibles sur le site des "Centers for Disease Control and Prevention" (CDC), aux États-Unis. En utilisant ces données, nous développons une approche en analyse de survie en considérant des patients qui ont un test positif à la grippe. De plus, nous traitons les données comme censurées par intervalle, et nous appliquons des méthodes de censure par intervalle pour estimer la fonction de survie. Les estimateurs de la fonction de survie sont utilisés pour comparer certaines saisons de grippe, différents groupes d'âge et des régions. En utilisant l'estimateur Kaplan-Meier et en appliquant des méthodes de censure par intervalle, les estimateurs de la fonction de survie sont différents, mais pour ce qui est des résultats des tests d'hypothèses, les conclusions des tests de *log-rank* respectifs sont identiques dans la plupart des cas.

Mots-clés: données CDC ILI , fonction de survie, Kaplan-Meier, censure par intervalle, comparer les fonctions de survie, test de *log-rank*.

ABSTRACT

In this M. Sc. thesis, we analyze the Influenza Like Illness (ILI) data available on the site of the Centers for Disease Control and Prevention (CDC), USA. Using this data, we develop a survival approach by considering positive flu tested patients. Moreover, we treat the data as interval censored, and we apply interval censoring methods to estimate the survival functions. Further, the survival estimators are used to compare some flu seasons, different age groups and regions. The survival estimates are different when applying Kaplan-Meier at the reported event time and applying interval censoring, but in hypothesis testing the conclusion of the respective log-rank tests are the same in most cases.

Key words: CDC ILI data, survival function, Kaplan-Meier, interval censoring, comparing survival functions, log-rank test.

INTRODUCTION

The main purpose of this thesis is to introduce an innovative approach to a study of the flu data presented by the CDC (Centers for Disease Control and Prevention). We mainly use the data found through a tool called "FluView Interactive", available on the CDC (U.S.) website. Different types of datasets are available to be downloaded throughout this tool. For example, the most important data-set used in our analysis is the U.S. national and regional numbers of people with influenza like illness (ILI) for seasons from $1997 - 1998$ until now. These are visiting patients at the GP practices that participate. Related available datasets on FluView Interactive are the regional number of ILI for 10 regions of U.S. and the distribution by flu types for some other subjects.

The aim is to create, study and compare some survival functions by considering the number of people who get sick every week (ILI cases) among the patients of a network of about 2000 surveillance clinics. Since the CDC is not keeping track of the people who visit the participating clinics over time, we consider only the reported ILI cases in our analysis; then, a cohort data is created using patients with ILI in order to apply survival analysis techniques. To estimate the survival function, a fixed time τ_0 (typically 30 or 52 weeks) is picked and we work conditionally, i.e. we consider some $S(t|\tau_0)$ conditional survival function for the time T to develop the flu symptoms.

The number of patients with ILI is reported once a week, and thus the exact

time of getting flu is not known; therefore, one can consider the event time to be interval censored. In the available datasets, the flu time is reported at the end of the week and this could be taken as the exact event time. On the other hand, as the flu could have started in the previous week, we created interval censored data; further we applied interval censoring to this data to estimate survival functions. Non-parametric estimators of survival functions, applying interval censoring methods like Turnbull (1976) are developed to estimate survival functions. Also we considered the end of the reported weeks as the exact event time and the empirical survival function was given in order to compare these functions with the survival functions using interval censoring methods. To compare survival estimates through different populations, log-rank tests are implemented.

The content of the chapters in this thesis is as follows: the sources of data available on the CDC website is presented in Chapter 1. In this Chapter, we describe as well how to create a dataset to which survival analysis can be applied. Interval censoring is discussed in Chapter 2 and we implement this method in order to answer the problem of unprecise reporting of the event time. The main objective of Chapter 3 is to compare survival functions when considering different seasons, age groups and regions. Another objective is to compare the results obtained by considering the empirical survival function and the interval censored methodology.

CHAPTER I

DATA PRESENTATION FROM A SURVIVAL ANALYSIS PERSPECTIVE

The data of our study is collected by the **Centers for Disease Control and Prevention (CDC)**, USA. The Epidemiology and Prevention Branch in the Influenza Division at CDC gathers and analyzes information on influenza activity in the United States and posts some of these data on FluView and FluView Interactive. FluView is a weekly influenza surveillance report, while FluView Interactive allows us to visualize the influenza surveillance data. The main data resource for data analysis in this thesis is available on FluView Interactive. The influenza surveillance system of U.S. is a collective production between CDC and its many partners in state, local, and territorial health departments, public health and clinical laboratories, vital statistics offices, health care providers, clinics, and emergency departments.

The information that is collected from different data sources allows the CDC to discover when and where influenza is happening. Using this information, CDC can determine influenza viruses type and can measure hospitalization and mortality caused by influenza. There exist five categories of Influenza Surveillance which are: Virologic Surveillance, Outpatient Illness Surveillance, Mortality Surveillance, Hospitalization Surveillance and Summary of the Geographic Spread of

Influenza.

FluView Interactive provides databases and some graphics for these different categories of information. In Section 1.1 some of these categories of data are introduced.

1.1 FluView Report

The CDC FluView report presents weekly influenza surveillance information through the United States. Through the FluView Interactive website, we have access to this data since $1997 - 1998$ to the current season. Some of the organizations which provide this data are: the U.S. branch of the World Health Organization (WHO), National Respiratory and Enteric Virus Surveillance System (NREVSS) collaborating laboratories and U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). We show below how some of this data is visualized on the FluView Interactive website.

An available graph for the Virologic Surveillance is the Line Chart ILINet. Line Chart ILINet is available for each flu season. In this chart, the percentage of visits for ILI is reported weekly. Figure 1.1 shows this graph, for $2014 - 2015$ flu season.

The influenza season starts on the Sunday of the week 40 of the year which falls at about the end of September and the beginning of October. Health care providers report the total number of patients and the number of patients diagnosed with influenza like illness (ILI) by age group (ILITOTAL). ILI is defined as fever (temperature of $100°F[37.8°C]$ or greater) and a cough and/or a sore throat without a known reason other than influenza.

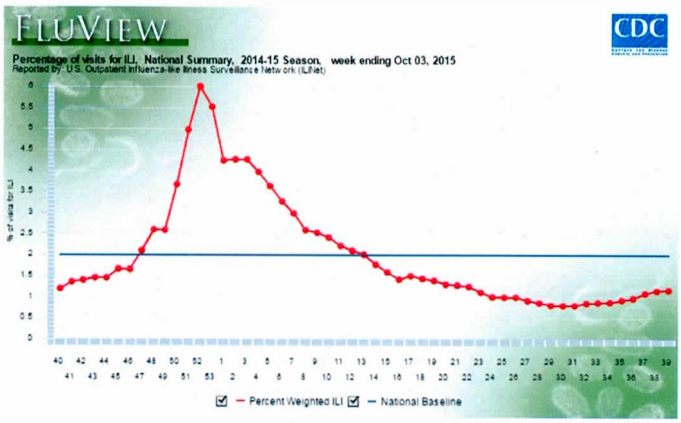

Figure 1.1 Line Chart ILINet.

In the data reported by ILINet, for each season, the variable ILITOTAL is reported weekly. Both public health and clinical laboratories, situated throughout the United States participate in the virologic surveillance for influenza, but their influenza testing technique can differ. They both provide useful information to observe influenza activity, and report age or age group of the patient, if available. Also they do examinations for positive influenza testing and influenza virus type. In the FluView Line Chart data, the number of influenza-like illness (ILI) is shown by age groups. Finally, cumulative ILI totals are provided for each season. We can look at the variable "Age" as an explanatory variable in the statistical model of our study. We come back to this issue in Chapter 2.

In Table 1.1, we can see how the FluView Line Chart data, available at the U.S. WHO/NREVSS Collaborating Laboratories and ILINet are presented. Table 1.1 shows the data on the first 10 weeks of the $2014-2015$ flu season. The first column of this table is the "Week" which starts by week 40. The next five columns are "Age groups" as follows "$0-4$" years, "$5-24$" years, "$25-49$" years, "$50-64$" years and ">64" years. In column 7, the total number of ILI is given. The

Table 1.1 First 10 weeks of ILINet data, 2014-2015.

Week	0-4	5-25	25-49	50-65	>65	ILITotal	Num of Providers.
40	2985	4078	2056	725	530	10374	1988
41	3125	4534	2209	866	563	11297	2010
42	3483	4806	2377	880	581	12127	2045
43	3486	5027	2520	906	535	12474	2055
44	3652	5172	2316	815	535	12490	2097
45	4303	5878	2421	889	611	14102	2098
46	4433	5836	2332	916	592	14109	2028
47	4765	7449	3025	1078	650	16967	2090
48	5468	6964	3401	1188	890	17911	2100
49	6206	9220	4765	1812	1244	23247	2133

number of providers is specified in the last column.

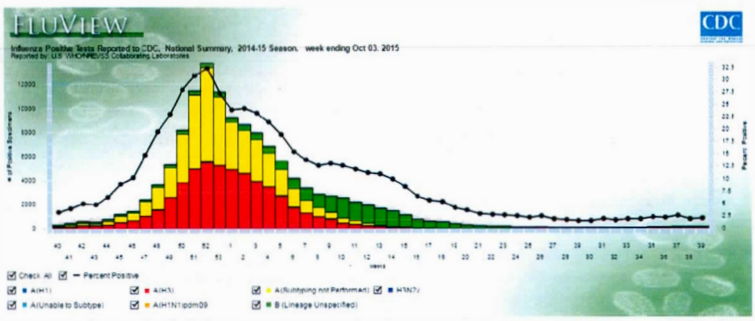

Figure 1.2 Stacked Column Chart WHO/NREVSS, 2014-2015 season.

As discussed previously, from both sources public health and clinical laboratories, useful information to monitor influenza activity is provided. Clinical laboratories do the examination to diagnose the flu and the data provided by them gives valuable information on the timing and intensity of influenza. Public health labo-

ratories do the examination on specimens to diagnose influenza virus type in each group of patients. Some of specimens from the clinical laboratories may be sent to public health laboratories to do further examinations.

For each week, they both report to the CDC the total number of examined specimens and the number positive for flu. Also "Age" or "Age Group" of the person is reported, if available. Figure 1.2 gives in 2014-2015 the FluView chart reported by the "U.S. WHO/NREVSS Collaborating Laboratories". The number of influenza positive specimens is given weekly in this chart. Public health laboratories present

Table 1.2 First 10 weeks of ILINet data, 2014-2015.

WEEK	TESTED SPECIMENS	A(H3)	A(H1N1)	A(Subtyping not Performed)	B(Lineage Unspecified)
40	9567	100	2	97	110
41	11036	125	5	149	160
42	11729	226	7	186	164
43	11385	212	5	200	127
44	11531	295	4	292	136
45	12918	506	6	489	155
46	13777	633	8	629	147
47	16166	1002	5	1233	163
48	18504	1538	8	1878	200
49	23068	2585	13	2483	249

the weekly total number of tested specimens, the number of positive flu tests, and the number of flu viruses by type. In Table 1.2 a part of this data is presented. This table shows the total number of tested specimens and some influenza types

in the first 10 weeks of the 2014 – 2015 flu season. The week number is the first column of this table and total specimens and some flu types are shown in the following columns.

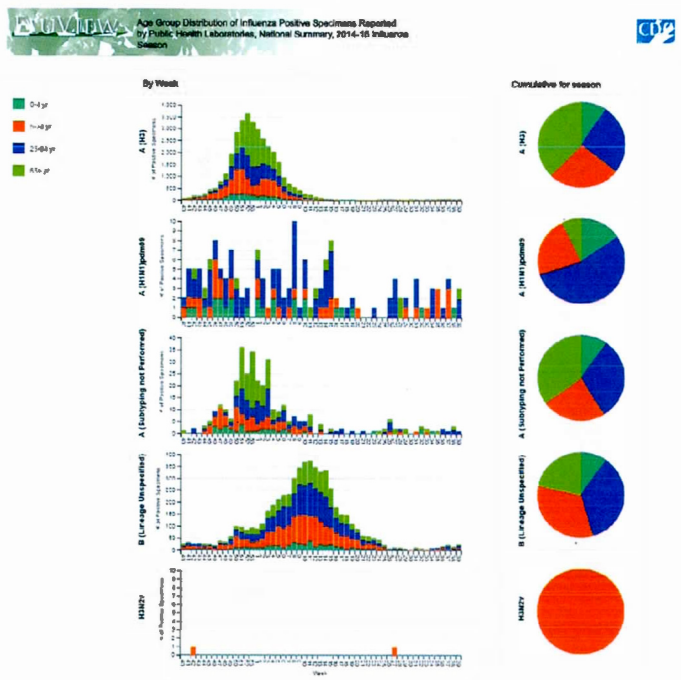

Figure 1.3 Age Group Distribution Of Influenza, 2014-15 Season.

Another graph which allows to visualize the flu data, which is provided by CDC in FluView, is the age distribution of influenza positive samples. The data is reported from public health laboratories. Figure 1.3 shows how this data is presented on the CDC site, in the 2014 – 2015 flu season. In this graph, there are 4 "age groups", namely: 0-4. In each age group, the graph shows the number of specimens by week. Virus types are indicated by different colours, as specified in the figure.

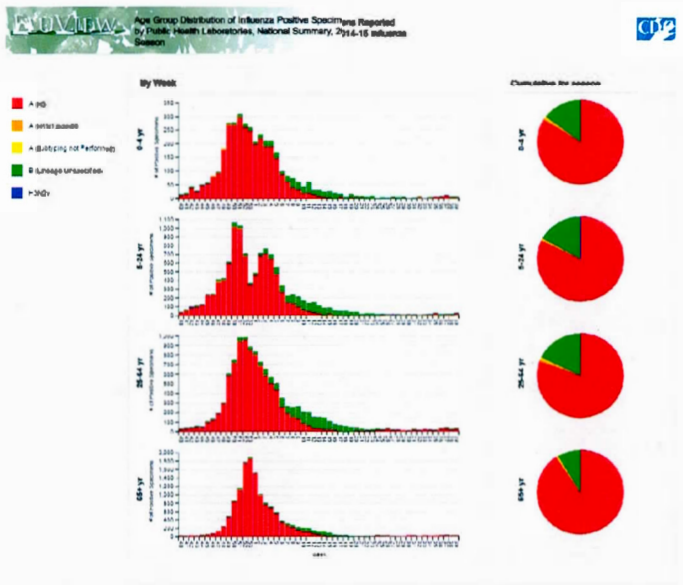

Figure 1.4 Influenza virus type distribution, 2014-15 Season.

FluView provides another visualization tool for the age group distribution of influenza positive specimens. Figure 1.4 illustrates this type of data visualization. Here the graphs are divided by "virus type". For each virus type, the number of positive specimens is showed weekly, separated by age groups.

Table 1.3 Virus View in four Age groups by season, 2014-2015.

Virus	0-4 yr	5-24 yr	25-64 yr	65+ yr
A(H3)	3113	9108	9154	12852
A(H1N1)	25	37	89	12
A(Subtyping not Performed)	33	81	108	119
B(Lineage Unspecified)	536	1846	2004	1199
H3N2v	0	2	0	0

Generally, there are five different virus types, who are: $A(H3)$, $A(H1N1)$, A(Subtyping

not Performed), B and $H3N2v$. In the reported number of positive specimens, age groups are distinguished by colours. In Figure 1.4 the corresponding colours and age groups are mentioned in detail.

Data Excel files can be downloaded through the FluView Interactive application. In Table 1.3, the data for the age group distribution on influenza positive tests is given by season. Indeed, the number of specimens with a positive test result is given by different age groups for each virus type. In this table, the given number is not the weekly number but the number of patients for the whole season $2014-2015$. For example, we can see in this table that 3113 children of age 0 to 4 years old had $A(H3)$ flu type during the $2014-2015$ flu season.

ILINet comprises more than 2900 subscribed outpatient health care providers in all 50 states, Puerto Rico, the District of Columbia and the U.S. Virgin Islands. Each year, they report more than 36 million patient visits. Health care providers around the country report data on the total number of visiting patients and the ILI total by age groups, to the CDC. Influenza positive tests data is also reported to the CDC by HHS Region. Using this data, it is possible to compare the flu in different U.S regions.

The list of 10 U.S. Regions for influenza season are:

Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont

Region 2: New Jersey, New York, Puerto Rico, and the U.S. Virgin Islands

Region 3: Delaware, District of Colombia, Maryland, Pennsylvania, Virginia, and West Virginia

Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee

Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin

Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, and Texas

Region 7: Iowa, Kansas, Missouri, and Nebraska

Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, and Wyoming

Region 9: Arizona, California, Hawaii, and Nevada

Region 10: Alaska, Idaho, Oregon, and Washington

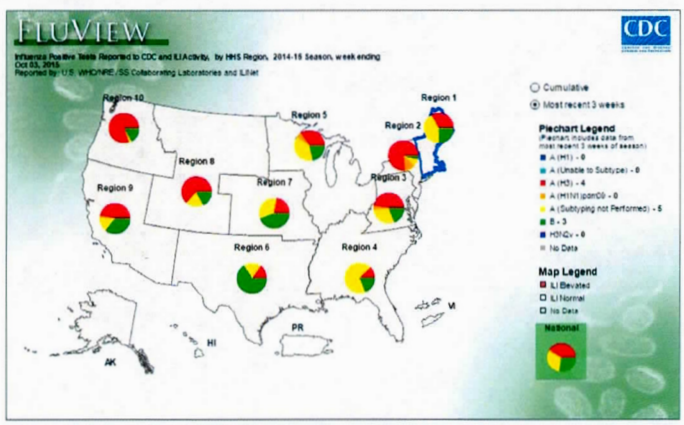

Figure 1.5 Influenza Positive tests, by Region, 2014-15 Season.

As an example, influenza positive tests reported to CDC and ILI-Activity, season $2014-15$, by 10 U.S. regions, is displayed in the map, given in Figure 1.5.

In Section 1, we discussed different types of flu data, reported by CDC which is available in FluView. We have studied some of this data as described in the following sections.

1.2 A study based on Survival Analysis

1.2.1 Limitations of the CDC statistics

In the data bases which are available at the CDC, the flu disease is studied for the whole United States population. The number of studied people (Total Patients) and the number of people diagnosed by ILI (ILITotal), is accessible weekly in the data reported by collaborating laboratories and ILINet. This data is re-weighted as on the example given below (they take into account the population size), for the flu season $2014 - 2015$ is shown in Table 1.4.

Table 1.4 First 5 weeks of ILINet data, 2014-2015.

Week	ILITotal	Total	Num.Providers	% UnweightedILI	% WeightedILI
40	10374	847743	1988	1.224	1.182
41	11297	838183	2010	1.348	1.347
42	12127	851592	2045	1.424	1.391
43	12474	868755	2055	1.436	1.450
44	12490	862144	2097	1.449	1.444

In this table, besides ILITotal and Total (total patients), Num.Providers (Number of Providers), the proportion of people diagnosed with ILI (% UnweightedILI) and (% WeightedILI) are given weekly. This proportion is the ratio of ILITotal over Total Patients. Unweighted ILI is also called "percentage of visits for ILI", and it is shown in the FLUVIEW for every week. Figure 1.1 demonstrates the percentage of visits for ILI by week for the season $2014 - 2015$. As presented in the Figure 1.1, the flu season starts at week 40 of each year. To study each flu season, since we need to start the time (week) from 1, we correspond week 40 to

1, week 41 to 2 and etc.

In Figure 1.1, we can study the percentage of people diagnosed with ILI through weekly visits to the provider clinics. As it is clear in this figure, in the week 52 (week 11 of the flu season), the percentage of visits for ILI has its maximum value. Indeed, Figure 1.1, presents the percentage of visits for ILI through different weeks of the indicated season. As noted on the CDC website, such proportions cannot be compared between seasons and regions, as they vary due to many factors. In particular the age distributions can be very different among GP practices.

In this thesis, we propose to address this problem and consider the cohort of people who visit the clinics over time, and eventually get the flu in one given season. This way, when comparing proportions, it is only the make up of the diseased cohort that matters, and this makes sense. Once the cohort is created, the following study can be made: Over fixed time intervals (for example after every week), what is the proportion of people who "survive" or have not gotten the flu yet? Among "survivors" or non-sick ones, what is the ratio of people who get sick? In this context, in the population which we study and at a given time, people who do not still get the flu are called survivors at that time. When the same people are studied over a specified flu season, the time to get flu for this population by week will be from 1 to 52 weeks, when every one fell ill by the end of the season.

1.2.2 Creating a cohort data

ILITotal is the number of people with influenza like illness among "Total Patients (Total)", where "Total Patients" are the people who go to the provider clinics each week. We would like to concentrate on a fixed group of people who visit the

provider clinic but if we count all patients who visit clinics, we may count some people more than once. When someone goes to the clinic who doesn't have the flu, he/she may visit the clinic again. Also among these people, many of them do not have the experience of getting flu at all in the studied season.

To prevent having these problems, for a specified season, a way out is to count all given weekly ILITotal in that season. If someone is diagnosed by influenza, he/she will not get it again for a given year. Therefore we cannot count sick people more than once in each flu season. Since ILITOTAL is reported weekly, we can look at it as the number of events occurring every week for a given season. In other words, we take ILITotal as the size of new population where all get influenza by the end of a specific flu season. In this given season, we gather all the people who got influenza and this is a group of individuals with a common property, i.e. getting flu by the end of the indicated season, or less than 52 weeks.

This new group of individuals, who have shared together the event of getting the flu, during a particular flu season forms a "cohort" followed for a year. For any flu season, we can create this new population using the ILITotal variable. Following this cohort data through time, we can study the proportion of the ones who "survived" at the end of each week, in different age groups, different U.S regions, different flu seasons and other available factors. Still, in most years, both methods give the same week for the maximum percentage of ILI cases.

Using this cohort data, we can compute the percentage of ILI cases by week, by taking the proportion of ILITotal of the week among all the people who got flu in the whole season. In Figure 1.6 we note that in Season 2014-2015, the ILI per-

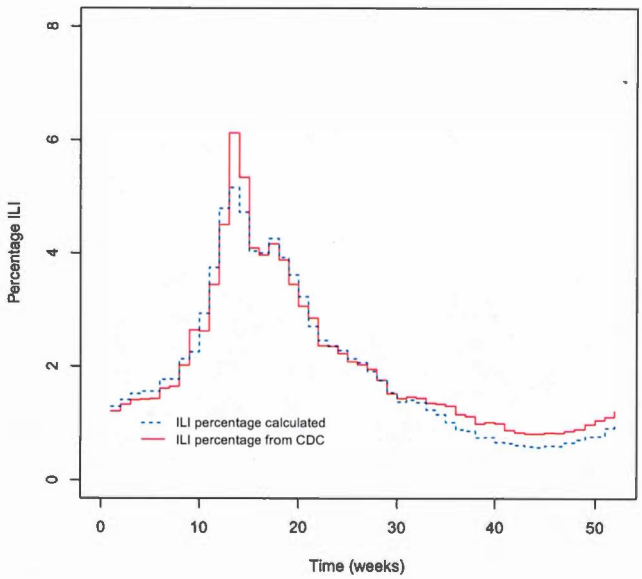

Figure 1.6 Comparison of given ILI percentage as reported by the CDC and calculated as the ILI weekly percentage among all ILI cases in the 2014-15 Season.

centage as given at the CDC site is almost equal to the ILI percentage computed from the cohort data, but this is not always the case.

1.3 Survival and hazard function, basic notions

Survival analysis involves the modelling of time to event data. Death or failure, for example, getting sick in this context, is considered an "event" in the survival analysis literature. This statistical method is defined as a set of methods for analyzing data where the outcome variable is the time until an event of interest occurs. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? How do particular circumstances or

characteristics increase or decrease the probability of survival?

The survival function is the probability that a subject survives past the time T. A survival function of the random variable T is defined as

$$S(t) = P(T > t). \tag{1.1}$$

In Figure 1.7 and Figure 1.8, we can see general examples of a survival function S. If T is a continuous variable, its survival function should behave like the one in Figure 1.7. As for Figure 1.8, we can consider that it represents an estimate of such survival functions.

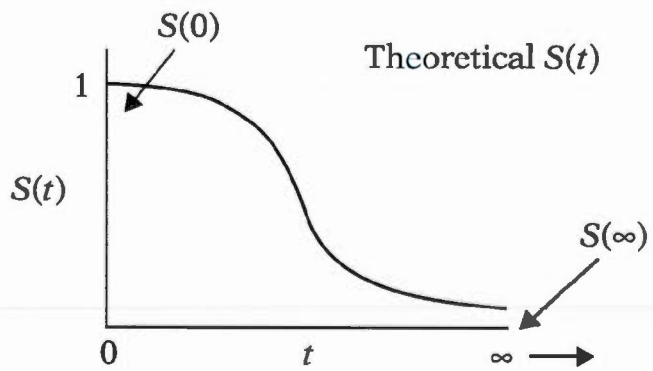

Figure 1.7 Survival curve in theory, Kleinbaum & Klein (2006).

At the beginning of the study, there exists no event, so the estimated survivals starts from roughly 1. If by the end of the study all individuals experienced the event $\hat{S}(t)$ will descend to 0, otherwise, $\hat{S}(t)$ is still positive and undefined beyond this point.

One fundamental quantity in survival analysis is the hazard function, denoted by $h(x)$. This function is defined as:

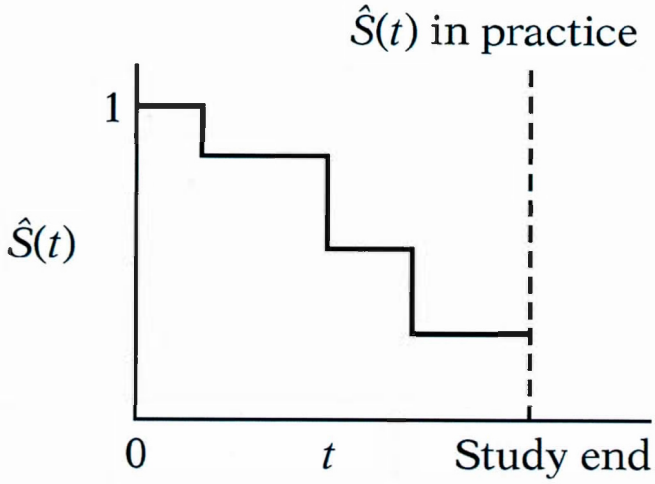

Figure 1.8 Survival curve in practice, Kleinbaum & Klein (2006).

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq T < x + \Delta x | T \geq x]}{\Delta x}. \tag{1.2}$$

The hazard function provides the instantaneous potential to occur the event, per unit time given the condition of survival up to time t, Klein & Moeschberger (2005). The hazard function concentrates on failing, which is in contrast to the survivor function. The hazard is a rate, rather than a probability and its values range between zero and infinity. The following equation, (1.4) shows the relationship between survival and hazard function. If T is a continuous random variable of probability density function $f(x)$, then its survival function is

$$S(x) = \int_x^{\infty} f(t)dt, \tag{1.3}$$

while its hazard function is

$$h(x) = f(x)/S(x) = -d\ln[S(x)]/dx. \tag{1.4}$$

Further, $H(x)$ or the cumulative hazard function is defined as $H(x) = \int_0^x h(u)\,du$.

Therefore, if T is continuous then

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)du\right]. \tag{1.5}$$

An estimate of the cumulative hazard function $H(t)$ is the "Nelson Aalen" estimator, which is defined up to the largest observed time of study as follows:

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}, \tag{1.6}$$

where d_i is the number of events at the event time $t_i, i = 1, \ldots, m$, and n_i or $n(t_i)$ is the number of susceptibles at t_i (i.e. the number of individuals at risk just before time t_i). An estimate of the hazard rate at an event time t_i is given by $h(t_i) = d_i/n_i$.

1.4 Unconditional and Conditional survival functions

In general, if T is a continuous random variable, its survival function is decreasing from 1 to 0. Survival curves are available in many different types but they have some common properties. They all are monotone, non-increasing functions. The value of survival functions are $S(0) = 1$ and this value approaches zero when the time x approaches infinity.

In the reported data, the variable of interest in our study is presented as the number of cases or ILI (defined according to a specified protocol), among number of visits. As mentioned above, in this study, the data provided by health agencies is not cohort data, but rather a cohort we have created out of this data. As explained in Section 1.2.1, this cohort is formed by the people who have experienced ILI, as reported by the provider clinics.

Thus, in the present treatment, the main idea is to use only the reported cases

in the analysis and treat them as a cohort (since in principle they cannot come twice for the same condition), in other words to work conditionally. This approach comes to considering a conditional survival function

$$S(t|\tau_0) = Pr(T > t|T \leq \tau_0). \tag{1.7}$$

where τ_0 is a fixed time (typically 30 or 52 weeks). The conditional survival function (1.7) is a proper survival function for a new variable $\tilde{T}$, where

$Pr(\tilde{T} > t) = S(t|\tau_0)$.

In what follows, we introduce an example of a conditional survival function and compare it with the unconditional survival function. Consider the Weibull distribution and its survival function, $S(t) = \exp(-\lambda t^{\alpha})$, where $\lambda > 0$ and $\alpha > 0$. In Figure 1.9, a survival curve of Weibull distribution and its conditional survival curve are illustrated, For the case where $\alpha = 3$ and $\lambda = 0.00208$. The conditional survival curve is defined as:

$$S(t|\tau_0 = 8) = Pr(T > t|T < 8).$$

The value of this conditional probability can be calculated as the following

$$\frac{Pr(T > t, T \leq 8)}{Pr(T \leq 8)} = \frac{Pr(t < T \leq 8)}{Pr(T \leq 8)} = \frac{Pr(T > t) - Pr(T > 8)}{1 - Pr(T > 8)} = \frac{S(t) - S(8)}{1 - S(8)}.$$

In Figure 1.9, we present $S(t)$ and $S(t|\tau_0)$.

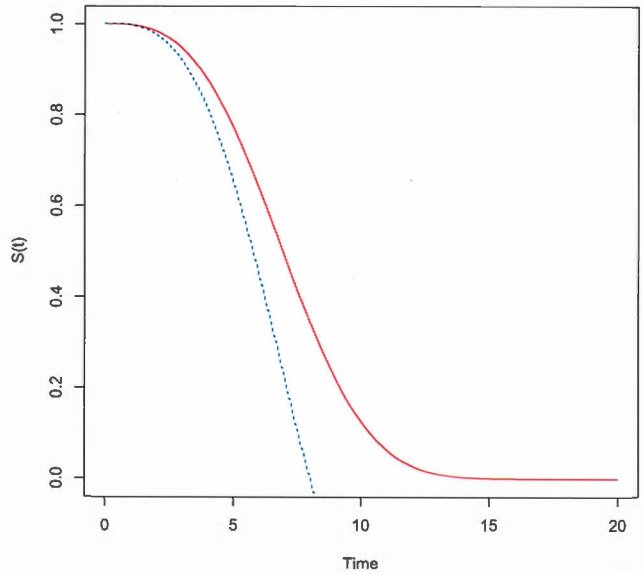

Figure 1.9 Weibull Survival functions for $\alpha = 3, \lambda = 0.00208$ (red dotted curve) and conditional survival curve $S(t|\tau_0 = 8)$ (blue dashed curve).

1.5 Estimating the survival functions

The aim of this section is to explain how can one estimate survival functions, for the flu data provided by U.S. health agencies, considering the reported week of getting the flu, as the event's time. By assuming these weeks as the event time, we can estimate the survival function for the data of our study.

In this part, after a brief review of some definitions, we illustrate the concepts by considering a specific flu season "$2014 - 15$"; further, a non-parametric estimate of the survival function, using the R software, is calculated.

The objective of this section is to describe a non-parametric estimate of the *survival function.* When all event times are exactly known, an obvious estimate of $S(t)$ would be the empirical survival function,

$$\tilde{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{t_i > t\}, \tag{1.8}$$

where I is the indicator function that takes the value 1 if the condition in braces is true and 0 otherwise. Clearly this estimator is the proportion of alive (people who have not experienced the event), at time t. We say these people survived at time t, as defined in the previous section.

Kaplan and Meier (1985) extended this survival estimation to a specific type of missing information, namely *censored* data (see section 2.1). This estimator is known as the **product limit** or **Kaplan-Meier estimator** and can be computed in the case where there is no missing information as follows. Let

$$t_1 < t_2 < \ldots < t_m$$

represent ordered times of events; let d_i be the number of events at t_i, and let n_i be the number of subjects at risk at t_i. In other words, n_i is the number of people who experienced no event, survivors just before t_i, or the number at risk at time t_i. The Kaplan-Meier estimator is the nonparametric maximum likelihood estimate of $S(t)$:

$$\hat{S}(t) = \prod_{i:t_i<t} (1 - \frac{d_i}{n_i}) \tag{1.9}$$

and it can be shown that $\hat{S}(t) = \tilde{S}(t)$ when there is no missing information. The idea behind the estimator is the following. Surviving to time t means you should survive to t_1; further, you should survive from t_1 to t_2, given that you already survived to t_1, and so on. There is no event between t_{i-1} and t_i, so the probability of an event between these times is zero. The conditional probability of

having the event at t_i given that there is no event right before this time, can be estimated by d_i/n_i. The conditional probability of surviving to time t_i is the complement $1 - d_i/n_i$. The overall unconditional probability of surviving to time t is obtained by multiplying the conditional probabilities for all event times up to t.

In our data, since there is no censoring, the Kaplan-Meier estimator or product limit estimator equals the empirical survival estimate. In what follows we use either Kaplan-Meier estimator or the empirical survival estimate interchangeably.

1.5.1 An example of empirical survival estimates for complete data

The empirical survival curve is a step function which has jumps at event times. Figure 1.10 represents the survival curve to study the evolving of ILI over time. The estimated survival function is defined on the created cohort data using the reports of ILINet (Table 1.4), for the flu season $2014 - 2015$.

In the following table (Table 1.5), we can see the value of some survival estimates in this flu season. The first column is the time of the event i.e. the week of reporting flu. The following columns are the number at risk, the number of events and the last column gives the survival estimates at the end of each week, in the given season.

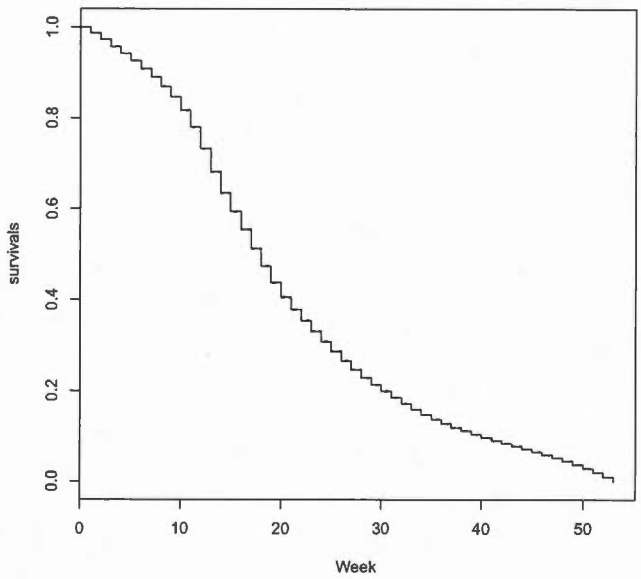

Figure 1.10 Empirical Survival Estimate for ILINet data 2014-2015.

Table 1.5 Some values of survival estimates of ILINet data 2014-2015.

Week	At risk	Event	Survival
40	794902	10374	0.987
41	784528	11297	0.973
42	773231	12127	0.958
43	761104	12474	0.942
44	748630	12490	0.926
⋮	⋮	⋮	⋮
52	16027	8026	0.01
53	8001	8001	0.00

CHAPTER II

INTERVAL CENSORING METHODS AND HOW TO APPLY THEM IN THE CDC FLU DATA

In many applied fields such as medicine, biology, epidemiology, public health, engineering and economics, we need to analyse time to event data. Time to event data sets usually contain censored observations. The data is censored when the exact event time for an observation is not known and we only know that the event happened in a certain period of time. One of the possible types of censoring is right-censoring. Right-censoring emerges when all we know about the event is that it occurred after a given time. Left-censoring arises when the event time of the observation is prior to the time of observation. Interval censoring, which is the general case of censoring, appears in a case where the event is known only to have taken place in some interval.

2.1 Missing information in time to event data

In this section we recall some concepts of various categories of censoring, such as right-censoring, left-censoring and interval-censoring.

2.1.1 Interval-censoring, the general case

A general case of censoring happens when the only known information we have about the event time T is to be in an interval of time, $[L, R]$, where $L \leq T \leq R$; L is the left endpoint and R is the right endpoint of the censoring interval. Interval-censoring can occur in a clinical trial, for example, when patients are assessed only at periodic or pre-scheduled follow-ups. When all sequences of time are independent of the event time we say that the censoring is non-informative.

Exactly observed, right and left-censored data are special cases of interval-censored data. When the event occurs exactly at the moment of a visit, an exact survival time $T_i = L_i = U_i$ is observed. If $R = \infty$, then the event time is in this interval $[L, \infty)$ and the event time is right-censored. In the case $L = 0$, the failure time is not observed and it lies in $(0, R)$, therefore the failure time is a left-censored observation. Consequently interval-censoring is a generalization of left and right-censoring.

As noted in Law & Brookmeyer (1992), when dealing with interval censored data, a common approach in practice is to assume that the event occurred at the end (or beginning or midpoint) of each interval, and then apply methods for standard time to event data. The authors note that this approach can lead to invalid inferences, and in particular will tend to underestimate the standard errors of the estimated parameters. In order to see the difference of survival estimates using different approaches, in Chapter 3 we calculated survival functions at the midpoints of time intervals and using interval censoring methods (Figure 3.1).

2.2 Why censoring is needed in the survival data of our study

As mentioned in Chapter one, when we put together all the people who got the flu each week (all cases), we have a new population of individuals who all get sick by the end of each flu season (October 3th). The main idea of this thesis is to estimate the survival function for this cohort data provided by **ILINET**. When this conditional survival function is estimated, we can compare it in different situations.

In this study and in the data illustration of this thesis, the event of interest is the time to catch the flu. But in the reported data of health care providers and clinics, the exact time of getting the flu or the exact failure time is not observed. The only known time is the reported time (by the GP), which is the end of the week where the patient visited the clinic. In this section, we discuss this problem of reporting the event-time and a solution for that is given.

2.2.1 Problem with reporting the event time

Indeed, the event of interest or getting sick in the data of our study is reported once a week. All we know about the reported event time is the week where a new case was observed (went to visit a doctor). But for each case, we don't have the information on the exact starting time of the disease. By assuming the week of reporting cases as the event-time, for each individual, the event is considered to happen in this reported week. But the start of the disease can be in the previous week of reporting. Therefore, the information conveyed by health agencies and clinics, does not provide the exact time of getting the flu. In other words, the reported time of an event in the data of our study is not a precise event-time.

Collaborating laboratories consider Sunday as the start day of the week in their reports. Influenza can last for about two weeks. Usually when people get flu they go to the clinic in their first week of disease but not the first day. For example, if a patient gets sick on Sunday, it is very probable that he/she goes to the clinic before next Sunday. But when someone gets flu on Friday and goes to the clinic on the next Sunday or after that, his reported event time of getting flu will be set in the week where he/she went to the clinic. In this case, the week of going to the clinic is a week after the week of getting the flu. Therefore, the event of interest can occur in the same week of reporting it or in a week before going to the clinic.

If we consider reported weeks as the time of event, the estimated survival functions could be biased. The aim is to consider this existing imprecision in the available reported event time.

2.2.2 A given solution for reporting the event time

We assume that the duration of flu is over two weeks. Considering the time t as the reported time of visit, the real failure time (measured in weeks), is in this interval: $[t-1, t]$. Since in this study the event-time or the time of getting flu is in an interval, we have interval censored data. In this thesis and in the data analysis here, interval censoring methods are applied to estimate the survival functions (with or without covariates).

As mentioned before, interval censored data arises when a failure time t is not observed, but can only be determined to lie in an interval obtained from a sequence of observed times. In this study, the real event time is either in the reported

week or in a preceding week. For example if the observed week of getting flu (event-time) for person A is week 14, the actual week of starting flu (t) is either during week 13 or during week 14. In other words, the real event-time is in this interval of time: $[13, 14]$. So $13 \leq t \leq 14$. Existence of interval censoring in the reported "event-time" is now clear. If t is the time where one gets the flu, by considering the time interval $[t-1, t]$, we can apply interval censoring to this data.

From now on, in all statistical analyses that are used, the method of interval censoring is applied. Indeed, we changed the variable of event-time in our data to a time interval variable. In the following sections after giving some essential concepts and definitions, we apply a nonparametric method to estimate survival functions in the presence of interval censoring.

2.3 Nonparametric survival estimation

In Section 1.4, the Kaplan-Meier estimator of the survival function was introduced (equation 1.11), and it applies to right censored data. In the following section, we want to estimate the survival function in the presence of interval censored time data. As previously discussed, for the Kaplan-Meier estimator, we use a nonparametric procedure here, which is an initial investigation tool. First it is needed to describe the survival time and then the survival function can be estimated. The main factors or the covariates that are used in our survival models are qualitative and with few levels; thus, the quantitative variables can be categorized. For example, age can be classified into three or four categories such as 0 to 5 years, 5 to 10 years and so on, Giolo (2004).

If the event of interest is not observed for all individuals, an indicator variable

for censoring should be defined. Some lines of a typical data-set are presented in Table 2.1, to illustrate how a data-set should be organized for an analysis in R.

Table 2.1 Rows of an example of a data-set.

left	right	therapy	censored
5	11	1	1
12	27	1	1
0	14	1	0
18	NA	1	0
19	NA	0	0
10	16	0	1
21	32	0	1
⋮	⋮	⋮	⋮

The example shown in Table 2.1, is a general form of an interval censored data-set. In this data-set, it is assumed that for each individual the event took place between an upper and lower time limit. The upper limit "NA" means that the upper limit of this time interval is not available, or the data is right censored. If time is measured in weeks, the time interval $[18, NA)$ means that the event of interest for this observation happened after the week 18. This observation can be presented in the form $[18, \infty)$ as well. The same type of censoring is applied to $[19, NA)$. For the other observation $[0, 14]$, as mentioned in Section 2.1.2, the event time is left censored. In this toy, the censoring indicator variable is assumed to be known. Also there is a treatment variable (1 or 0).

In this section, an analog to the Product-Limit estimator of the survival function for interval censored data is presented. This estimator has been suggested by

Turnbull (1976), and we use the algorithm described in Giolo (2004) and Klein & Moeschberger (2005).

2.3.1 An algorithm for Turnbull's method to estimate survival functions

The following algorithm shows the step by step Turnbull's method to estimate a survival function, in the case of interval censored data.

Step 0: Let $(L_i, U_i]$ $(i = 1, \ldots, n)$, be the n observed time intervals. Note that the event times of interest are in these n intervals. We put all the times L_i and U_i together and order them. So let $0 = \tau_0 < \tau_1 < \cdots < \tau_m$ are the values $\{L_1, U_1, L_2, U_2, \ldots, L_n, U_n\}$ in increasing order. Note that in some cases, $L_i = L_i'$ or $L_i = U_i', i \neq i'$. Therefore, for the ith observation we define a weight α_{ij} as follows

$$\alpha_{ij} = \begin{cases} 1, & \text{if } (\tau_{j-1}, \tau_j] \subseteq (L_i, U_i] \\ 0, & \text{otherwise,} \end{cases}$$

where $j = 1, \ldots, m$. This weight α_{ij} specifies if the event which occurs in $(L_i, U_i]$ could have happened at time τ_j. The algorithm counts of the following steps,

Step 1: Calculate the probability of an event occurring at time τ_j, denoted by p_j:

$$p_j = \hat{S}(\tau_{j-1}) - \hat{S}(\tau_j), \qquad j = 1, \ldots, m \leq n. \tag{2.1}$$

If we apply the definition of survival function to estimate $\hat{S}(\tau_{j-1})$ and $\hat{S}(\tau_j)$, then p_j estimates π_j

$$\pi_j = P(X > \tau_{j-1}) - P(X > \tau_j) = P(\tau_{j-1} < X \leq \tau_j). \tag{2.2}$$

Since the event does not occur at the exact time τ_j, the probability of an event occurring at this time is calculated by using the time interval $(\tau_{j-1}, \tau_j]$, which contains τ_j.

Step 2: We estimate the number of events which occurred at τ_j, denoted by d_j Turnbull (1976)

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ik} p_k}; \qquad j = 1, \ldots, m. \tag{2.3}$$

In what follows, we explain why d_j can estimate the number of events that occurred at τ_j.

Let $D_i = \sum_{k=1}^{m} \alpha_{ik} p_k$, where $i = 1, \ldots, n$. For each i, D_i gives the sum of probabilities of possible events which occurred in $(L_i, U_i]$. To estimate the number of events that happen at time τ_j, first we fix the time interval $(\tau_{j-1}, \tau_j]$, and then look for all $(L_i, U_i], i = 1, \ldots, n$ which contain $(\tau_{j-1}, \tau_j]$. For each i, p_j/D_i is the proportion of events that happen at time τ_j when τ_j belongs to the interval $(L_i, U_i]$; the weighted sum of all these proportions estimates the number of events happening at time τ_j. Note that d_j is not necessarily an integer.

Step 3: Determine the estimated number at risk at time τ_j, denoted by Y_j:

$$Y_j = \sum_{k=j}^{m} d_k. \tag{2.4}$$

The people who are at risk at τ_j are the ones who have not yet experienced the event until the time τ_j. Therefore, the number of people at risk at time τ_j equals the sum of the number of events which occurred at time τ_j and at $t > \tau_j$.

Step 4: Compute the updated Product-Limit estimator using the number of events and the number at risk at time τ_j computed respectively in Steps 2 and 3; namely:

$$\hat{S}(\tau_j) = \prod_{i:\tau_i \leq \tau_j} \frac{Y_i - d_i}{Y_i}. \tag{2.5}$$

If the updated estimate of $S(.)$, $\hat{S}(\tau_j)$, is close to the previous estimate of $S(.)$ for all τ_j's, stop the algorithm, otherwise repeat Steps 1-3, using the updated estimate of $S(.)$.

In the following part, a small example of interval censored data is considered in order to illustrate the previous algorithm. On this example after computing K-M estimates, we apply the steps of Turnbull's method algorithm, to calculate estimators of the survival function.

2.3.2 Using Turnbull's method in an illustration

Example 2.3.1. *In this small example* 9 *intervals of time are available. The time intervals in the form of* $(L_i, U_i], 1 \leq i \leq 9$ *are as follows:*

$$(2,5];\ (2,5];\ (3,4];\ (1,3];\ (5,7];\ (5,7];\ (5,7];\ (3,6];\ (7,9]$$

Each event of interest happens in an interval, so each interval corresponds to an event, but we don't have the information for the exact time of any event.

If we calculate the K-M estimator of the survival function assuming that some event of interest happens at all given (L_i's and U_i's) times, the following result would be obtained:

Figure 2.2 gives the empirical survival curve for this example with 9 time intervals. The K-M estimates of survivals are shown in Table 2.2.

The aim is to illustrate Turnbull's method, when applied to this example. We

Table 2.2 Empirical survival estimates at the limits of time intervals in Example 2.3.1:

Time	Survival
1	0.9444
2	0.8333
3	0.6667
4	0.6111
5	0.3333
6	0.2778
7	0.0556
9	0.0000

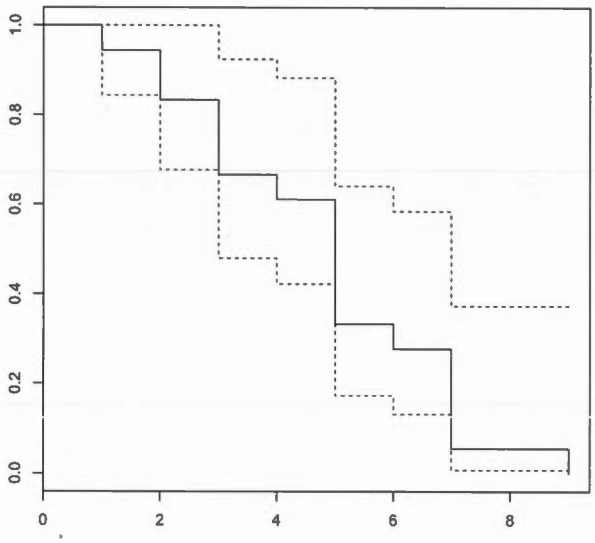

Figure 2.1 Empirical Survival curve and confidence bands for Example 2.3.1.

need to apply all five algorithm steps shown previously, to compute Turnbull's non-parametric estimate of the survival function. In this computing, we need to choose some initial estimates of the survival functions like those which are shown in Table 2.2.

Step 0 Let

$$\tau_0 = 0,\ \tau_1 = 1,\ \tau_2 = 2,\ \tau_3 = 3,\ \tau_4 = 4,\ \tau_5 = 5,\ \tau_6 = 6,\ \tau_7 = 7,\ \tau_8 = 9.$$

be the grid of times which includes all the points L_i and U_i, $i = 1, 2, \ldots, 9$. We compute the K-M estimators where each τ_j is an event time. All the α_{ij}'s are shown in the following matrix $A = [\alpha_{ij}]$. Since $i = 1, 2, \ldots, 9, j = 1, 2, \ldots, 8$, we have $n = 9$ and $m = 8$. Consequently the matrix A is 9×8. After introducing the matrix A, we will compute some of α_{ij}'s.

$$A = [\alpha_{ij}] = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The time interval $(\tau_0, \tau_1] = (0, 1]$ is in none of the given time intervals $(L_i, U_i], i = 1, \ldots, n$. Therefore $\alpha_{i1} = 0$ for all $i = 1, \ldots, n$. The next time interval $(\tau_1, \tau_2] =$

$(1,2]$ is just in $(L_4, U_4] = (1,3]$, so in the second column of the matrix A, all the α_{i2}'s are equal to 0 except $\alpha_{42} = 1$. Other α_{ij}'s can be computed similarly.

Step 1 Probability of an event occurring at time τ_j, $j = 1, \ldots, 8$. In this step we need to use K-M survival estimates in Table 2.2, namely:

$$
\begin{aligned}
p_1 &= S(0) - S(1) = 1 - 0.9444 = 0.0556, \\
p_2 &= S(1) - S(2) = 0.9444 - 0.8333 = 0.1111, \\
p_3 &= S(2) - S(3) = 0.8333 - 0.6667 = 0.1666, \\
p_4 &= S(3) - S(4) = 0.6667 - 0.6111 = 0.0556, \\
p_5 &= S(4) = S(5) = 0.6111 - 0.3333 = 0.0.2778, \\
p_6 &= S(5) - S(6) = 0.3333 - 0.2778 = 0.0.0555, \\
p_7 &= S(6) - S(7) = 0.2778 - 0.0556 = 0.2222, \\
p_8 &= S(7) - S(8) = 0.0556 - 0.000 = 0.0.0556.
\end{aligned}
$$

Step 2 Further, we compute the number of pseudo events d_j which occurred at τ_j, $j = 1, \ldots, 8$. To compute this, first we should calculate D_i's, $i = 1, \ldots, n$. In the following, D_i's are computed using the matrix A.

$$
\begin{aligned}
D_1 &= \sum_{k=1}^{8} \alpha_{1k} p_k = p_3 + p_4 + p_5, & D_2 &= \sum_{k=1}^{8} \alpha_{2k} p_k = p_3 + p_4 + p_5, \\
D_3 &= \sum_{k=1}^{8} \alpha_{3k} p_k = p_4, & D_4 &= \sum_{k=1}^{8} \alpha_{4k} p_k = p_2 + p_3, \\
D_5 &= \sum_{k=1}^{8} \alpha_{5k} p_k = p_6 + p_7, & D_6 &= \sum_{k=1}^{8} \alpha_{6k} p_k = p_6 + p_7, \\
D_7 &= \sum_{k=1}^{8} \alpha_{7k} p_k = p_6 + p_7, & D_8 &= \sum_{k=1}^{8} \alpha_{8k} p_k = p_4 + p_5 + p_6, \\
D_9 &= \sum_{k=1}^{8} \alpha_{9k} p_k = p_8,
\end{aligned}
$$

To compute D_i, we need the multiply the row i of the matrix A by the vector $[p_j] = [p_1\ p_2\ p_3\ p_4\ p_5\ p_6\ p_7\ p_8]^\top$. For example $D_1 = [\alpha_{1j}] \times [p_j]$, and in the first row of matrix A we have $\alpha_{13} = 1,\ \alpha_{14} = 1,\ \alpha_{15} = 1$ and other α_{1j}'s are equal to 0. Therefore D_1 is computed as above. Other D_i's are computed similarly. In the following, d_j's are calculated as explained in the algorithm in 2.3.1.

$$
\begin{aligned}
d_1 &= \sum_{i=1}^{9} \frac{\alpha_{i1} p_1}{D_i} = 0 \\
d_2 &= \sum_{i=1}^{9} \frac{\alpha_{i2} p_2}{D_i} = \frac{\alpha_{42} p_2}{D_4} = \frac{p_2}{p_2 + p_3} = 0.400, \\
d_3 &= \sum_{i=1}^{9} \frac{\alpha_{i3} p_3}{D_i} = \frac{p_3}{D_1} + \frac{p_3}{D_2} + \frac{p_3}{D_4} \\
&= \frac{p_3}{p_3 + p_4 + p_5} + \frac{p_3}{p_3 + p_4 + p_5} + \frac{p_3}{p_2 + p_3} = 1.266, \\
d_4 &= \sum_{i=1}^{9} \frac{\alpha_{i4} p_4}{D_i} = \frac{p_4}{D_1} + \frac{p_4}{D_2} + \frac{p_4}{D_3} + \frac{p_4}{D_8} \\
&= \frac{p_4}{p_3 + p_4 + p_5} + \frac{p_4}{p_3 + p_4 + p_5} + \frac{p_4}{p_4} + \frac{p_4}{p_4 + p_5 + p_6} = 1.365, \\
d_5 &= \sum_{i=1}^{9} \frac{\alpha_{i5} p_5}{D_i} = \frac{p_5}{D_1} + \frac{p_5}{D_2} + \frac{p_5}{D_8} \\
&= \frac{p_5}{p_3 + p_4 + p_5} + \frac{p_5}{p_3 + p_4 + p_5} + \frac{p_5}{p_4 + p_5 + p_6} = 1.826, \\
d_6 &= \sum_{i=1}^{9} \frac{\alpha_{i6} p_6}{D_i} = \frac{p_6}{D_5} + \frac{p_6}{D_6} + \frac{p_6}{D_7} + \frac{p_6}{D_8} \\
&= \frac{p_6}{p_6 + p_7} + \frac{p_6}{p_6 + p_7} + \frac{p_6}{p_6 + p_7} + \frac{p_6}{p_4 + p_5 + p_6} = 0.742, \\
d_7 &= \sum_{i=1}^{9} \frac{\alpha_{i7} p_7}{D_i} = \frac{p_7}{D_5} + \frac{p_7}{D_6} + \frac{p_7}{D_7} \\
&= \frac{p_7}{p_6 + p_7} + \frac{p_7}{p_6 + p_7} + \frac{p_7}{p_6 + p_7} = 2.400, \\
d_8 &= \sum_{i=1}^{9} \frac{\alpha_{i8} p_8}{D_i} = \frac{p_8}{D_9} = \frac{p_8}{p_8} = 1.
\end{aligned}
$$

Since the time interval $(\tau_0, \tau_1] = (0, 1]$ is included in none of the observed time intervals $(L_i, U_i], i = 1, \ldots, 9$, the number of events at time τ_1, which is denoted by d_1, equals 0.

To compute d_2, we need the second column of A and the second time interval $(\tau_1, \tau_2] = (1, 2]$. We look at all $(L_i, U_i]$'s where $i = 1, \ldots, m$ and the interval of our interest $(1, 2]$ is just in $(L_4, U_4]$. For that reason, in the second column of the matrix A, $\alpha_{42} = 1$ and all other elements are equal to 0. The proportion of events of $(1, 2]$ occurring in the time interval $(L_4, U_4]$ equals to $\frac{p_2}{D_4} = 0.400$ and the event of interest is not happening in other $(L_i, U_i]$'s. Therefore d_2 or the number of events in $(1, 2]$ is equal to $\frac{p_2}{D_4} = 0.400$.

Step 3 The estimated number at risk at time τ_j, $j = 1, \ldots, 8$.

$$Y_1 = \sum_{k=1}^{8} d_k = 9, \qquad Y_2 = \sum_{k=2}^{8} d_k = 9, \qquad Y_3 = \sum_{k=3}^{8} d_k = 8.60,$$
$$Y_4 = \sum_{k=4}^{8} d_k = 7.333, \qquad Y_5 = \sum_{k=5}^{8} d_k = 5.97, \qquad Y_6 = \sum_{k=6}^{8} d_k = 4.143,$$
$$Y_7 = \sum_{k=7}^{8} d_k = 3.40, \qquad Y_8 = \sum_{k=8}^{8} d_k = 1.$$

Step 4 The updated Product-Limit estimator using the data found in Steps 2 and 3 is:

$$\hat{S}(\tau_1) = \frac{Y_1 - d_1}{Y_1} = 1, \qquad \hat{S}(\tau_2) = \hat{S}(\tau_1) \times \frac{Y_2 - d_2}{Y_2} = 0.956,$$
$$\hat{S}(\tau_3) = \hat{S}(\tau_2) \times \frac{Y_3 - d_3}{Y_3} = 0.815, \qquad \hat{S}(\tau_4) = \hat{S}(\tau_3) \times \frac{Y_4 - d_4}{Y_4} = 0.663$$
$$\hat{S}(\tau_5) = \hat{S}(\tau_4) \times \frac{Y_5 - d_5}{Y_5} = 0.460, \qquad \hat{S}(\tau_6) = \hat{S}(\tau_5) \times \frac{Y_6 - d_6}{Y_6} = 0.378,$$
$$\hat{S}(\tau_7) = \hat{S}(\tau_6) \times \frac{Y_7 - d_7}{Y_7} = 0.111, \qquad \hat{S}(\tau_8) = \hat{S}(\tau_7) \times \frac{Y_8 - d_8}{Y_8} = 0.$$

As it is clear in the result of Step 4, the updated estimates $(\hat{S}(\tau_j))$'s are different from the previous estimates of survival functions (K-M estimates in (2.1)), for all τ_j's. So we have to redo the process (Steps $1-3$ of algorithm), using the updated estimates of survival $(\hat{S}(\tau_j))$. Given in step 4, this process should be continued until the difference of updated survivals with the previous survivals are less than some pre specified value, e.g. (10^{-3}).

Let $(\hat{S}_{jk}(\tau_j))$'s, $1 \leq j \leq 8$, show the survival estimates for Example 2.3.1, after repeating the algorithm k times. The following results are survival estimators $(\hat{S}_{j4}(\tau_j))$'s and $(\hat{S}_{j5}(\tau_j))$'s after repeating respectively 4 and 5 times the explained algorithm.

The forth updated survival estimator is:

$$\hat{S}_{14}(\tau_1) = 1, \quad \hat{S}_{24}(\tau_2) = 0.989, \quad \hat{S}_{34}(\tau_3) = 0.820,$$
$$\hat{S}_{44}(\tau_4) = 0.537, \quad \hat{S}_{54}(\tau_5) = 0.472, \quad \hat{S}_{64}(\tau_6) = 0.335,$$
$$\hat{S}_{74}(\tau_7) = 0.111, \quad \hat{S}_{84}(\tau_8) = 0.$$

The fifth updated survival estimator is:

$$\hat{S}_{15}(\tau_1) = 1, \quad \hat{S}_{25}(\tau_2) = 0.993, \quad \hat{S}_{35}(\tau_3) = 0.816,$$
$$\hat{S}_{45}(\tau_4) = 0.518, \quad \hat{S}_{55}(\tau_5) = 0.476, \quad \hat{S}_{65}(\tau_6) = 0.317,$$
$$\hat{S}_{75}(\tau_7) = 0.111, \quad \hat{S}_{85}(\tau_8) = 0.$$

The difference between $(\hat{S}_{j4}(\tau_j))$ and $(\hat{S}_{j5}(\tau_j))$ for all j's $(1 \leq j \leq 8)$, are still not less than 10^{-3}. Therefore the process (Steps $1-3$ of algorithm), should be continued. The repeating process is too long to be done by hand. In Giolo (2004)'s paper, an "R" function called *Turnbull* is created to estimate Turnbull's estimate of survival functions. By using this "R" function *Turnbul*, we get the final estimate

as follows:

$$\hat{S}(0) = 1, \qquad \hat{S}(1) = 1, \qquad \hat{S}(2) = 1,$$
$$\hat{S}(3) = 0.797, \qquad \hat{S}(4) = 0.509, \qquad \hat{S}(5) = 0.509,$$
$$\hat{S}(6) = 0.111, \qquad \hat{S}(7) = 0.111, \qquad \hat{S}(8) = 0.$$

2.3.3 Using the midpoint method in Example 2.3.1

In this Section we take the midpoint of each of the nine intervals of time and consider them as the time of the event; further, we find the estimator of the survival functions.

Let $t_1, t_2, \ldots, t_9$ be the midpoints of the intervals defined in Example 2.3.1, namely:

$$t_1 = 3.5, \quad t_2 = 3.5, \quad t_3 = 3.5, \quad t_4 = 2.0, \quad t_5 = 6.0,$$
$$t_6 = 6.0, \quad t_7 = 6.0, \quad t_8 = 4.5, \quad t_9 = 8.0.$$

In the following "R" output we can see the K-M estimator of the survival function at the midpoint time.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
0.0	9	0	1.00		1.00	1.00
2.0	9	1	0.889	0.105	0.7056	1.000
3.5	8	3	0.556	0.166	0.3097	0.997
4.5	5	1	0.444	0.166	0.2141	0.923
6.0	4	3	0.111	0.105	0.0175	0.705
8.0	1	1	0.000	NaN	NA	NA

In Example 2.3.1, the event time is not known and we just know that each event happens in a given interval. If we consider the midpoint of each interval as the

event time, we can compute the K-M estimates of survival at these midpoints. This result of K-M using midpoints is very different from Turnbull's estimate of the survival function using this interval censoring method.

In the following section we compare two types of estimators, using Turnbull's method and the midpoint method.

2.4 Comparing survival estimates in data with disjoint and overlapping intervals of time

The steps of Turnbull's survival estimate are not always different from those of the K-M's at the midpoint of the time intervals in different types of examples. In some cases, the result of survival estimates using interval censoring methods is the same as survival estimates which consider midpoints of intervals as the event time of interest. In what follows, two types of time intervals are discussed which lead to different results in comparing the two methods of interval censoring and midpoint.

2.4.1 The case of disjoint time intervals

In some situations, time intervals are disjoint intervals and they do not overlap. Indeed there is no common point in every two different intervals of time. The flu data provided by CDC is an example of such disjoint time intervals. In this data, the clinics report the data weekly (Sunday to Saturday) and there is no common day in two different weeks.

Remark 2.4.1. *In a time to event data-set, if different time intervals are disjoint intervals, then the survival estimate using Turnbull's (1976) method has the same*

jumps as the Kaplan-Meier estimate at the midpoints of the time intervals.

Proof. Note that this remark is not applicable to the right and left censored intervals, because when an event time is right or left censored, it is in the form of (L_i, ∞) or $(0, U_i)$ for the right and left censored time intervals respectively. Consequently, if there are more than one interval in this form then the intervals will overlap. Also pay attention that different time intervals are disjoint, but more than one event can happen in each time interval.

To prove the remark, first we apply Turnbull's method a general form of survival data-set with disjoint time intervals. Then we compute the empirical survival f at the midpoints of the time intervals in this general form of data-set, and compare the results.

Let the disjoint time intervals be $(L_i, U_i]$, $1 \leq i \leq n$. When we apply "Step 0" of the algorithm in Section 2.3.1, proposed by Giolo (2004) for Turnbull's method, since the time intervals do not overlap, the constructed time interval $(\tau_{j-1}, \tau_j]$ is included in one of $(L_i, U_i]$ only if it is exactly one of $(L_i, U_i]$'s. Indeed $(\tau_{j-1}, \tau_j] \subseteq (L_i, U_i]$ if and only if $(\tau_{j-1}, \tau_j] = (L_i, U_i]$. For each i, this condition of $(\tau_{j-1}, \tau_j] \subseteq (L_i, U_i]$ is valid only once. Therefore, in the matrix $A = [\alpha_{ij}]$ $(1 \leq i \leq n,\ 1 \leq j \leq m)$, there is only one 1 in each row and the other elements of that row are 0. If more than one event happens in the kth interval $(\tau_{k-1}, \tau_k]$, there will be l_k rows corresponding to $(\tau_{k-1}, \tau_k]$ in the matrix A, where l_k is the number of events which happen in the interval $(\tau_{k-1}, \tau_k]$.

Now we check the columns of the matrix A. All elements of the first column are 0, because the first interval $(0, \tau_1]$ is not one of the $(L_i, U_i]$ intervals. In the second column, as we discussed, at least one of the intervals $(L_i, U_i]$, $1 \leq i \leq n$ is equal to $(\tau_1, \tau_2]$. The number of events in the first interval is l_1, so $(L_i, U_i] = (\tau_1, \tau_2]$

repeats l_1 times. Therefore in this second column of A, 1 appears l_1 times, and all other elements are equal to 0. By the same argument, in the kth column ($k > 1$) of A, 1 appears l_{k-1} times and all other elements are 0.

In the "Step 1" of the algorithm in 2.3.1, the probability of an event occurring at time τ_j can be calculated by p_j, similar to the equation (2.1).

In the "Step 2" of Giolo's algorithm in 2.3.1, the goal is to estimate the number of events that occurred at τ_j by d_j similar to equation (2.2). So $d_j = \sum_{i=1}^{n} \alpha_{ij} p_j / D_i$, where $D_i = \sum_{k=1}^{m} \alpha_{ik} p_k$.

Calculation of the denominators (D_i)*'s*: As pointed above, for each row i of matrix A, all elements α_{ij} are equal to 0, except one of them. So in $D_i = \sum_{k=1}^{m} \alpha_{ik} p_k$, there is only one value 1 of α_{ij}, corresponding to some $(\tau_{j-1}, \tau_j]$. Therefore, the value of D_i is equal to one of the p_j probabilities.

Calculation of the d_j*'s:* The number of events in the jth interval $(\tau_{j-1}, \tau_j]$, is equal to l_j. In this summation $d_j = \sum_{i=1}^{n} \alpha_{ij} p_j / D_i$, the coefficient α_{ij} is equal to 1, l_j times and α_{ij} is equal to 0 for the rest ($n - l_j$ times). In other words, in the column j of the matrix A, there are l_j rows where the value of α_{ij} is equal to 1, and all other α_{ij}'s are equal to 0. In order to calculate d_j, we look at the interval $(\tau_{j-1}, \tau_j]$, and in "Step 1", the probability matched to this interval is p_j. Moreover, if $\alpha_{ij} = 1$, then the denominator $D_i = p_j$. The following conclusion can be achieved:

$$d_j = \frac{p_j}{p_j} + \ldots (l_j \; times) \cdots + \frac{p_j}{p_j} = l_j.$$

So the number of events in the jth interval is l_j, an integer as in the K-M estimate. It is the same number of events if we suppose that the event happens at the midpoint or at the end point of the interval. So this gives the same estimator of survival as Kaplan-Meier. K-M estimator at time τ_j is: $\hat{S}(\tau_j) = \prod_{\tau_i \leq \tau_j} \frac{Y_i - d_i}{Y_i}$. We

know that Y_j is the number at risk and d_j is the number of events at the time τ_j or in the j-th interval. In this calculation we do not repeat Steps $0-3$ and we stop because the updated survivals do not change and are the same as the K-M estimators (since the number of events and numbers at risk are the same). □

In the following example, a survival data-set with disjoint time intervals is given. By computing survival functions using these two methods in this example, we can compare Turnbull's method with K-M at the midpoints of the time intervals, as Remark 2.4.2 can be applied to this example.

Example 2.4.1. *In this hypothetical example, we want to study the survival function in a small dataset. Assume that one studies egg white (albumen) allergy on* 10 *children which are followed in order to see when they stop reacting to albumen (event time). These children are about the same age and they all stopped the reaction to albumen by the end of the study. The following data-set shows the date at which the albumen allergy test was negative for each kid.*

Table 2.3 Data for albumen allergy example.

Id	1	2	3	4	5	6	7	8	9	10
Year	2002	2002	2003	2004	2004	2004	2005	2006	2006	2006

The allergy test is done once a year and at the beginning of the study years. When the allergy test is negative for a child, it means his body developed the antibody for this allergy in the previous months of doing the test. So if A shows the year of study, the exact time to stop this reaction is in the interval $[A-1, A]$. Thus, we should consider an interval of time for each event of interest. Each of the following intervals make a correspondence between time intervals and the above dates.

$$(1,2];\ (1,2];\ (2,3];\ (3,4];\ (3,4];\ (3,4];\ (4,5];\ (5,6];\ (5,6];\ (5,6]$$

In these corresponding intervals, 2 corresponds to 2002, so $[1, 2]$ corresponds to $[2001, 2002]$. This data-set is similar to the flu data-set provided by CDC. For each child, we do not know the exact event time, so there exists interval censoring in this study. Also the data-set is a set of disjoint time intervals.

Empirical survival estimates at the midpoints of time intervals:

We used the R packages **survival** and **KMsurv** to calculate K-M at the midpoint of each interval. The version of R used in this calculation is R 3.1.1 which gave the following output:

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1.5	10	2	0.8	0.126	0.587	1.000
2.5	8	1	0.7	0.145	0.467	1.000
3.5	7	3	0.4	0.155	0.187	0.855
4.5	4	1	0.3	0.145	0.116	0.773
5.5	3	3	0.0	NaN		

In general, whether we consider the left end point, right end point or middle point of time intervals, as the event-time of interest, this does not change the jumps of the empirical estimator of survivals in these disjoint time intervals.

Turnbull's method:

In order to compute survival functions using Turnbull's method, we need to calculate the initial value of survival probabilities. We put all the points coming from time intervals together and we calculate the K-M estimate of survivals. The initial value of the survival functions are in the following table of output of R:

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	20	2	0.90	0.067	0.778	1.000
2	18	3	0.75	0.097	0.582	0.966
3	15	4	0.55	0.111	0.370	0.818
4	11	4	0.35	0.107	0.193	0.636
5	7	4	0.15	0.080	0.053	0.426
6	3	3	0.00	NaN		

Now we can apply the Turnbull (1976) method to this data-set to compute this non-parametric estimate of the survival function.

Step 0

$$\tau_0 = 0, \tau_1 = 1, \tau_2 = 2, \tau_3 = 3, \tau_4 = 4, \tau_5 = 5, \tau_6 = 6;$$

τ_j's for $1 \leq j \leq 6$, are the grid of times which includes all the points L_i and U_i for $i = 1, \ldots, 10$. All the α_{ij}'s are shown in the following matrix:

$$A = [\alpha_{ij}] = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and $1 \leq i \leq 10$, $1 \leq j \leq 6$.

Step 1 The probability of an event occurring at time τ_j, $j = 1, \ldots, 6$ is given by

$$
\begin{aligned}
p_1 &= S(0) - S(1) = 1 - 0.90 = 0.1, \\
p_2 &= S(1) - S(2) = 0.9 - 0.75 = 0.15, \\
p_3 &= S(2) - S(3) = 0.75 - 0.55 = 0.2, \\
p_4 &= S(3) - S(4) = 0.55 - 0.35 = 0.2, \\
p_5 &= S(4) = S(5) = 0.35 - 0.15 = 0.2, \\
p_6 &= S(5) - S(6) = 0.15 - 0.0 = 0.15.
\end{aligned}
$$

Step 2 The number of events which occurred at τ_j, $j = 1, \ldots, 6$ is:

$$
\begin{aligned}
d_1 &= \sum_{i=1}^{10} \frac{\alpha_{i1} p_1}{\sum_{k=1}^{6} \alpha_{ik} p_k} = 0, \\
d_2 &= \sum_{i=1}^{10} \frac{\alpha_{i2} p_2}{\sum_{k=1}^{6} \alpha_{ik} p_k} = \frac{\alpha_{12} p_2}{\sum_{k=1}^{6} \alpha_{1k} p_k} + \frac{\alpha_{22} p_2}{\sum_{k=1}^{6} \alpha_{2k} p_k} = \frac{p_2}{p_2} + \frac{p_2}{p_2} = 2, \\
d_3 &= \sum_{i=1}^{10} \frac{\alpha_{i3} p_3}{\sum_{k=1}^{6} \alpha_{ik} p_k} = \frac{p_3}{p_3} = 1, \\
d_4 &= \sum_{i=1}^{10} \frac{\alpha_{i4} p_4}{\sum_{k=1}^{6} \alpha_{ik} p_k} = \frac{p_4}{p_4} + \frac{p_4}{p_4} + \frac{p_4}{p_4} = 3, \\
d_5 &= \sum_{i=1}^{10} \frac{\alpha_{i5} p_5}{\sum_{k=1}^{6} \alpha_{ik} p_k} = \frac{p_5}{p_5} = 1, \\
d_6 &= \sum_{i=1}^{10} \frac{\alpha_{i6} p_6}{\sum_{k=1}^{6} \alpha_{ik} p_k} = \frac{p_6}{p_6} + \frac{p_6}{p_6} + \frac{p_6}{p_6} = 3.
\end{aligned}
$$

Step 3 The estimated number at risk at time τ_j, $j = 1, \ldots, 6$ is:

$$
\begin{aligned}
Y_1 &= \sum_{k=1}^{6} d_k = 10, \qquad Y_2 = \sum_{k=2}^{6} d_k = 10, \qquad Y_3 = \sum_{k=3}^{6} d_k = 8, \\
Y_4 &= \sum_{k=4}^{6} d_k = 7, \qquad Y_5 = \sum_{k=5}^{6} d_k = 4, \qquad Y_6 = \sum_{k=6}^{6} d_k = 3.
\end{aligned}
$$

Step 4 The updated Product-Limit estimator using the Pseudo data found in Steps 2 and 3 is

$$\hat{S}(\tau_1) = \frac{Y_1 - D_1}{Y_1} = 1, \qquad \hat{S}(\tau_2) = \hat{S}(\tau_1) \times \frac{Y_2 - D_2}{Y_2} = 0.8,$$
$$\hat{S}(\tau_3) = \hat{S}(\tau_2) \times \frac{Y_3 - D_3}{Y_3} = 0.7, \qquad \hat{S}(\tau_4) = \hat{S}(\tau_3) \times \frac{Y_4 - D_4}{Y_4} = 0.4,$$
$$\hat{S}(\tau_5) = \hat{S}(\tau_4) \times \frac{Y_5 - D_5}{Y_5} = 0.3, \qquad \hat{S}(\tau_6) = \hat{S}(\tau_5) \times \frac{Y_6 - D_6}{Y_6} = 0.$$

If we repeat the process, we will have the same survival estimates, because the survival estimates computed by product limit estimator (2.4), do not depend on the initial value of the survival function. Indeed, in Step 2, we can see that the actual probabilities (p_i)'s do not have any effect in the formulas that give the (d_i)'s.

Comparing survival curves through Turnbull's method and empirical survival estimates at the midpoint:

The result of survival estimates using two different methods, is presented in Figure 2.3.

In the first part of this section, the following conclusion was achieved. Applying interval censoring methods does not change the result of the Kaplan-Meier estimate at the midpoint, in the disjoint interval-censored data. In the next part, overlapping time intervals are studied through an example.

2.4.2 Overlapping time intervals

In some interval-censored survival datasets, time intervals are overlapping. In the following example we have a dataset with overlapping time intervals.

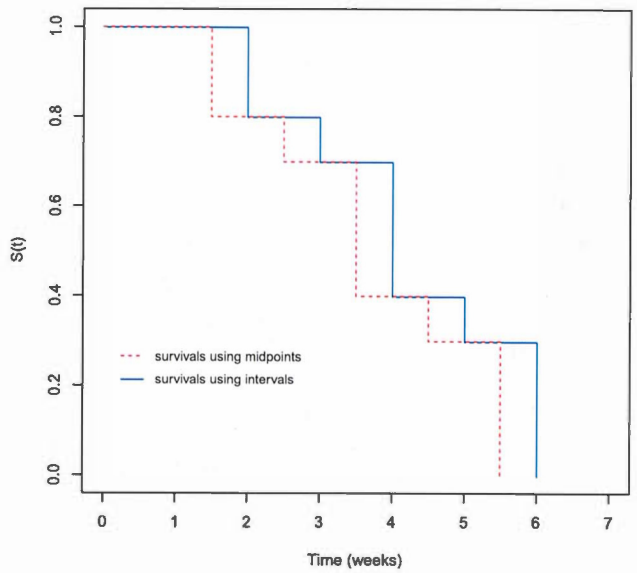

Figure 2.2 Comparing survival curves by applying two different methods to the data in Example 2.4.1.

Example 2.4.2. *Consider the study of the egg white (albumen) allergy, similar to Example* 2.4.1. *Assume* 10 *children are in the data-set. An allergy clinic starts this study in* 2008, *and it only does the test in June and December every year. The clinic asks parents to bring their children for the allergy tests every six months (every June and December), however not all the children visit the clinic at the assigned dates. Therefore, the exact time of stopping allergy symptoms for these* 10 *children is not known. The following simulated intervals are given by date:*

id	pre-test	negative-test
1	Jun2008	Dec2009
2	Jun2008	Dec2009
3	Jun2009	Dec2010
4	Jun2010	Dec2011

5	Jun2010	Dec2011
6	Jun2010	Dec2011
7	Jun2011	Dec2012
8	Jun2012	Dec2013
9	Jun2012	Dec2013
10	Jun2012	Dec2013

In the above data-set, the variable "negative-test" is the date at which the allergy test was negative. And the variable "pre-test" is the test prior to the negative-test. The real time to stop having the allergy symptoms for each child is a time between the pre-test date and the negative-test date. So here the event-time of interest is interval censored. In order to study this interval-censored data, we can correspond numbers to the given dates, as follows. We correspond 0.5 to June 2008 and 1 to the December 2008. Also 1.5 corresponds to June 2009 and 2 corresponds to December 2009 and so on. The result of this correspondence is in the following R output. Note that "cens" indicates censoring.

	left	right	cens
1	0.5	2	1
2	0.5	2	1
3	1.5	3	1
4	2.5	4	1
5	2.5	4	1
6	2.5	4	1
7	3.5	5	1
8	4.5	6	1
9	4.5	6	1
10	4.5	6	1

Similar to the previous example, we compute the survival estimation using midpoint and Turnbull's method.

Comparing survivals using Turnbull's method and midpoint method:

In this example, since time intervals are overlapping, we can not use the Remark 2.4.1. Survival estimates using interval censoring methods and the midpoint time value could be different. Survival curves using both methods are shown in Figure 2.3. The blue full line is the survival curve using Turnbull's method and the red

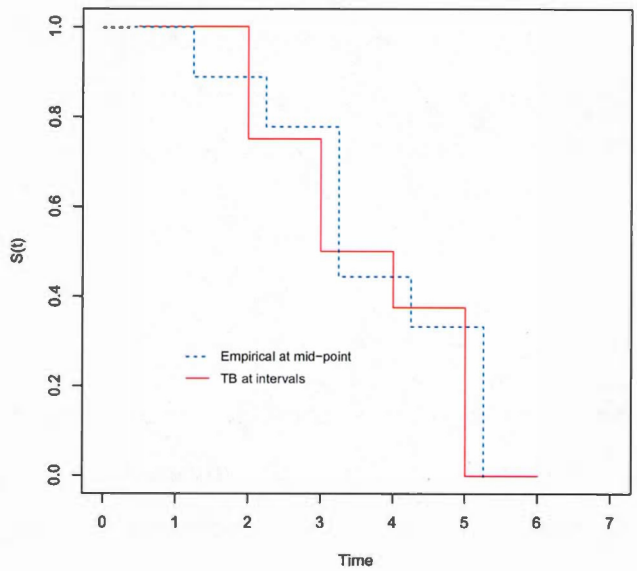

Figure 2.3 Survival curves applying two different methods to the data in Example 2.4.2.

dotted line is the survival curve using midpoints of time intervals in this example. It is clear in Figure 2.3 that these two survival curves are different. In the following table, the survival estimates in the midpoint of above time intervals are given.

As it is shown in Table 2.4, the result of survival estimates at midpoint is different

Table 2.4 Survival estimates using two methods, Example 2.4.2.

Time (mid-point)	0	1.25	2.25	3.25	4.25	5.25
Empirical Survival estimates	1.0	0.889	0.778	0.444	0.333	0.0
TB using time intervals	1.0	1.0	0.750	0.500	0.375	0.0

from estimate of survivals using Turnbull's method.

Different result of survivals using different methods in Example 2.4.2, indicates that survival estimates using midpoint and Turnbull's method are not the same, in interval-censored overlapped data-set. Indeed, overlapping may be used as a condition for an interval-censored data, to have interval-censored survival estimates different from K-M at left or middle points of time intervals.

2.5 Comparison of two groups of survival data

In some situations, the survival time is available for more than one group of individuals. The easiest way to compare the survival times of two different groups is to plot the corresponding estimates of the two survival functions on the same axes.

Figure 2.4 shows the survival curve for two groups of women with breast cancer, which is explained in Example 2.11 of Collett (2015). The survival times for these women is grouped according to whether or not sections of a tumour were positively stained with HPA. Figure 2.4 shows that the estimated survival function for women with negatively stained tumours is always greater than that for women with positively stained tumours. It means that for women with negatively stained tumours, the estimated probability of survival is higher than for women with posi-

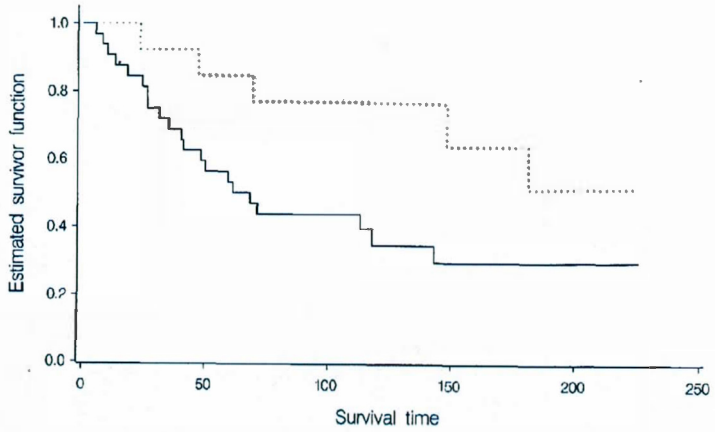

Figure 2.4 Kaplan-Meier survival function for women with tumours that were positively stained (solid line) and negatively stained (dotted line).

tively stained tumours, Collett (2015). In general, the observed difference between the two survival curves can correspond to a real difference or, alternatively, the difference between estimated survivals of the two groups can be interpreted as a chance variation. So we say that there is no real difference between the survival functions of the two groups. A better criterion to compare the survival functions is to use a hypothesis testing procedure.

In the next chapter, we consider comparing the survival functions of different flu seasons or of different age groups in the same season, etc. In these comparisons, the following is the question of our interest: when can we claim that two or more survival curves are the same? To compare two groups of survival data, there are many different methods which can be used. In this Section, the non-parametric procedure, *log-rank*, first proposed by Nathan Mantel (1967) will be described. Also, whenever we are studying time to event data, and the time may only be known partially, the *log-rank* test is a useful nonparametric test to compare survival distributions. In the case that the time to event is interval censored,

there are several approaches. One can apply the log-rank test, by considering the midpoint of the interval as the event time and do the usual right censored weighted log-rank tests; see Law & Brookmeyer (1992).

In the following sections, 2.5.1 and 2.5.2, the popular testing method of log-rank will be described. In 2.5.3, interval censored data is considered and an alternative method to the log-rank test. In the analysis of our interval censored data, an R package, called *interval* is used to perform a weighted log-rank test.

2.5.1 The log-rank test for two groups

As discussed above, in some situations, it is necessary to analyze whether two survival curves are identical. For this purpose, doing a statistical test is needed. In this section, the construction of log-rank test is illustrated in the following Collett (2015).

Since we are comparing two groups of survival data (Group 1 and Group 2), the death times in each group should be known. Consider there exist m distinct death times, in both groups, as $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. At each time $t_{(j)}$, for $j = 1, \ldots, j = m$, d_{1j} and d_{2j} are the number of individuals who die in Group 1 and Group 2, respectively. Let n_j be the total number at risk just before the time $t_{(j)}$. Also let n_{1j} and n_{2j} be the number at risk just before $t_{(j)}$ in Group 1 and Group 2 respectively. It is clear that if d_j is the total number of deaths at time $t_{(j)}$, then we have $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$. In Table 2.5 all these numbers are summarized.

We suppose that the null hypothesis (H_0), is no difference in survival experience of people who are in the two groups. One solution to test the given null hypothesis

Table 2.5 Necessary data for computing log-rank test statistic for two groups of individuals.

Group	No. of deaths at $t_{(j)}, j = 1, \ldots, m$	No. surviving beyond $t_{(j)}$	No. at risk just before $t_{(j)}$
1	d_{1j}	n_{1j} - d_{1j}	n_{1j}
2	d_{2j}	n_{2j} - d_{2j}	n_{2j}
Total	d_j	$n_j - d_j$	n_j

is to define a deviation, as the difference of observed number of people in the two groups who die at each death time and the expected number of deaths under H_0. To define this statistic, we can combine all the deviations over each of the death times.

Assume that the null hypothesis is true, and indeed the survival is independent of group. In Table 2.5, we consider all the marginal totals to be fixed (we work conditionally), therefore all four entries in this table are just determined by the value of d_{1j} (the only random variable in the table), which is the number of deaths in Group 1, at time $t_{(j)}$. Then, d_{1j}, as a random variable can vary between 0 and $\min(d_j, n_{1j})$ and has the hypergeometric distribution with parameters $\{d_j, n_{1j}, n_j\}$. Indeed, the probability that the number of deaths in Group 1 takes the value d_{1j} is given in the following equation:

$$\frac{\begin{pmatrix} d_j \\ d_{1j} \end{pmatrix} \begin{pmatrix} n_j - d_j \\ n_{1j} - d_{1j} \end{pmatrix}}{\begin{pmatrix} n_j \\ n_{1j} \end{pmatrix}} \tag{2.6}$$

where $\begin{pmatrix} d_j \\ d_{1j} \end{pmatrix} = \frac{d_j!}{d_{1j}!(d_j - d_{1j})!}$, $j = 1, \ldots, m$.

The expected number of people who die at $t_{(j)}$ in Group 1, (e_{1j}), should be the mean of the random variable d_{1j}. By applying the formula of hypergeometric random variable d_{1j}'s mean, we conclude that, $e_{1j} = \frac{n_{1j}d_j}{n_j}$. Under H_0, the probability of death at time $t_{(j)}$, does not depend on which group it happens and it is d_j/n_j. By multiplying this probability with the number of people in each group, one gets the expected number of deaths in that group at time $t_{(j)}$.

To define the statistic, we need to combine all the given numbers in Table 2.5 for all $t_{(j)}$'s, $1 \leq j \leq m$ to find a comprehensive measure of the deviation of the observed values of d_{1j} from their expected values. The easiest way to give this overall measure is to sum up the gaps $d_{1j} - e_{1j}$ over the total number of death times, m, in both groups. The overall measure of deviations is given by

$$U_L = \sum_{j=1}^{m}(d_{1j} - e_{1j}) = \sum_{j=1}^{m} d_{1j} - \sum_{j=1}^{m} e_{1j} \tag{2.7}$$

The U_L is the statistic we are looking for, which equals the difference between the total observed and expected numbers of deaths in Group 1. Since $E(d_{1j}) = e_{1j}$, the mean of this statistic will be zero. Furthermore, one assumes that number of deaths, (d_{1j})'s, are independent and the variance of U_L is as follows

$$var(U_L) = \sum_{j=1}^{m} var(d_{1j}) = \sum_{j=1}^{m} v_{1j} = V_L, \tag{2.8}$$

where $var(d_{1j}) = v_{1j}$ and $var(U_L) = V_L$. Since d_{1j} has a hypergeometric distribution, the variance v_{1j} is calculated as follows

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}. \tag{2.9}$$

When the number of death times is not too small, it is possible to show that U_L has an approximate normal distribution. Therefore, $U_L/\sqrt{V_L}$ has an approximate standard normal distribution, $N(0,1)$. The square of a standard normal random variable has a chi-squared distribution with 1 degree of freedom, denoted χ^2_1, so we have the following

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1) \Rightarrow \frac{U_L^2}{V_L} \equiv \chi^2_1. \tag{2.10}$$

This method in which the information of 2 by 2 tables is combined, is known as the Mantel-Haenszel procedure and in the case of survival data the test based on this statistic is known as the log-rank test. This name comes from the fact that the test statistic can be related to the ranks of survival times in the two groups, and the resulting test statistic is based on the logarithm of the Nelson-Aalen estimator of the cumulative hazard rate function.

In the statistic $W_L = U_L^2/V_L$, all the deviations of observed survival times and their expected values under the null hypothesis (of no difference between two groups) are summarized. A larger value of the statistic is stronger evidence against the null hypothesis. Since the test statistic W_L under the null hypothesis has approximately a chi-squared distribution with one degree of freedom, the associated p-value with the test statistic can be found in the chi-squared distribution function. The log-rank test is a large sample chi-square test that is base on an overall comparison of K-M curves by comparing the estimated hazard rates.

The methods of weighted comparison of hazard functions can be extended to the comparison of $k \geq 2$ groups. We need to test the following hypotheses:

$$\begin{aligned} &H_0 : h_1(t) = h_2(t) = \cdots = h_k(t), \quad \text{for all } t \leq T \\ &H_a : \text{at least one of the } h_j(t)\text{'s is different for some } t \leq T, \end{aligned} \tag{2.11}$$

where T is the largest time such that all the groups have at least one subject at

risk. In the data analysis of this thesis, we did not apply the extension of the log-rank test to several groups.

As noted in Collett (2015), the log-rank test is more appropriate when the hazard functions are proportional. In a more general setting, one can use the Gehan (1965) generalization of Wilcoxon test. This Wilcoxon test is based on

$$U_W = \sum_{j=1}^{m} n_j(d_{1j} - e_{1j}), \tag{2.12}$$

where, d_{1j} and e_{1j} are defined as in Section 2.5.1. In U_W, each difference $d_{1j} - e_{1j}$ is weighted by n_j, the total number of individuals at risk at time t_j. The variance of the statistic U_W is given by

$$U_W = \sum_{j=1}^{m} n_j^2 v_{1j}, \tag{2.13}$$

where v_{1j} is given by equation (2.9). The Wilcoxon test statistic is reduced to

$$W_W = U_W^2 / V_W, \tag{2.14}$$

which has an approximate chi-squared distribution with one degree of freedom, when the null hypothesis is true. Both U_W and U_L are weighted sums of $(d_{1j} - e_{1j})$, $j = 1, \ldots, m$. While the log-rank test gives a constant weight of approximately 1, the statistic U_W gives less weight to the differences corresponding to small n_j value in the tail.

2.5.2 Using the log-rank test for interval censored data

The main idea of this section is to present an extension of the log-rank test discussed in Section 2.5.1, to the interval censored data. In Section 2.5.1, we described the numbers needed to compute the log-rank test for two groups of individuals (Table 2.5). The idea is to find a substitute for these numbers in order

to compute the log-rank test for comparing two groups, when the data is interval censored. Finkelstein (1986) proposes such an extension that is related to the development is Section 2.3.

In Section 2.3.1, an algorithm to estimate survival functions, when the time is interval censored is given. In this algorithm, in each time interval we need to calculate a "pseudo number of events". In Table 2.5, we can replace the number of events d_j at $t_{(j)}$, by the pseudo number of events at $\tau_{(j)}$, as given by Equation (2.3). In order to avoid confusion, let's denote by d'_j the pseudo number of events and by n'_j the pseudo number at risk at time τ_j. Thus, the pseudo number at risk n'_j, just before τ_j is given by $n'_j = \sum_{k=j}^{m} d'_j$.

We apply the same notation as in Section 2.5.1 in order to comparing two groups. Consequently, in Group 1 and Group 2 respectively, d'_{1j} and d'_{2j} are the pseudo number of events at τ_j and n'_{1j} and n'_{2j} are the pseudo number at risk just before τ_j. We can replace the numbers in Table 2.5 by these pseudo numbers to compute the proportional hazards model for interval censored failure time data. In Finkelstein (1986), it is shown that, as the time intervals $(\tau_{j-1}, \tau_j]$ decrease, and the pseudo number of events is small relative to the pseudo number at risk, the score statistic resembles the usual log-rank test statistic, i.e.

$$U = \sum_{j=1}^{m} (d'_{1j} - d'_j n'_{1j}/n'_j) \tag{2.15}$$

In the next chapter, in our analysis, we applied a variant of this method as introduced in Fay & Shaw (2010).

CHAPTER III

ILLUSTRATIVE EXAMPLES

In this chapter, we apply interval censoring as studied in Chapter 2 to the influenza surveillance data, available on the CDC website. Out of the ILI surveillance data, we created a cohort by considering only the affected subjects (case-cohort). Therefore, we can apply standard survival analysis methods to this cohort data. Since different data-sets are available, for each one that is picked, we can create a cohort and then we can estimate its survival function. Also we need to create a specific interval censored event time to apply interval censoring (discussed in Section 3.1).

The main part of the analysis in this chapter is to compare the estimated survival curves. To estimate survival functions, we compute the empirical survival estimate at the reported flu times and then we take into account the flu time is interval censored and we use methods like Turnbull's method (2.3.1) to estimate survival functions in the interval censored case. The reason for estimating the survival functions using these two methods is to see whether their result is different and to assess influence of interval censoring on the analysis. Indeed, first we want to illustrate how the treatment of the data proposed in Chapter 2 changes the estimates of the survival probabilities. Further, one can apply hypothesis testing and compare survival functions across seasons, regions, etc. Another crucial com-

parison is across age groups, as one expects differences in survival according to age.

In order to perform this analysis, we had to arrange our survival data as to have overlapping time intervals and this is explained in Section 3.1. In our analysis we focus on the most recent flu seasons.

The first comparison that one can apply is to compare survival functions across different flu seasons, and this is done in Section 3.2. A priori, one can think that in different seasons the results could be very different, because there are many factors that can change. For example, the number of people in different age groups is not equal in two different seasons; flu types may also change. As mentioned above, ILI data in each season is given by age groups and regions. Comparing seasons just among people of a fixed age group or among people who live in a predetermined region, should lead to a better comparison. Indeed, by adding conditions like being in a specified age group, we are restricting people in the comparison to have more similar survival functions (we control for age). Therefore, in Section 3.3, we compare survival functions of two recent seasons, namely $2014-2015$ and $2013-2014$, across specific age groups.

A third type of comparison is to consider specific contiguous age groups (Section 3.4). Since some interesting results are obtained in this section when comparing age groups $25-49$ and $50-64$ in season $2014-2015$, we decided to compare these age groups in some other seasons as well.

A last type of analysis (Section 3.5), is to compare survival functions among different regions of U.S. in the recent flu season $2014-2015$. Given graphs of percentage

of ILI for all these regions we picked up a few regions that seemed to be more similar in order to compare their survival functions through a testing procedure. Also in Section 3.6, we compare survival functions in two age groups $25-49$ and $50-64$ in all 10 regions of U.S. for the most recent flu season $2014-2015$. In Section 3.7, the comparison of survival functions among all available flu types in the recent season $2014-2015$ is studied.

3.1 Treatment of the raw data: Creating overlapping time intervals

In this section, we study the survival function on the flu data provided by *CDC*, which was presented in the first Chapter. In Table 3.1, all needed variables to estimate survival probabilities are presented for the first 10 weeks. We added the variable "Pre-week" to our data-set, which is the previous week of reported event time. We need to add "Pre-week" in order to apply interval censoring to our flu data, as discussed in Section 2.2.

Since the week 40 is the first week of study in each flu season, we can consider it 1 in the data of our study. So $[0,1]$ corresponds to $[39,40]$, $[1,2]$ corresponds to $[40,41]$ and so on. In the following "R" output, we can see the corresponding data for the first 10 weeks.

Time Interval	Number of Events	Censure
[0, 1]	10374	1
[1, 2]	11297	1
[2, 3]	12127	1
[3, 4]	12474	1
[4, 5]	12490	1
[5, 6]	14102	1

[6, 7]	14109	1
[7, 8]	16967	1
[8, 9]	17911	1
[9, 10]	23247	1

As we see in this part of the data shown above, time intervals of our survival flu data are disjoint intervals. Due to Remark 2.4.1, interval censoring methods do not change the K-M survival estimates at the reported event time (right end point of time intervals).

Table 3.1 First 10 weeks of the data provided by ILINet, 2014-2015.

Year	Week	Pre-week	ILITotal
2014	40	39	10374
2014	41	40	11297
2014	42	41	12127
2014	43	42	12474
2014	44	43	12490
2014	45	44	14102
2014	46	45	14109
2014	47	46	16967
2014	48	47	17911
2014	49	48	23247

Moreover, as noted above, the event could have occurred before the week where it is reported. So we propose to consider time intervals in the form of $[a, b]$, so that the event time is between a and b. Therefore, following the same argument as the one in Section 2.2, we consider a as one week before the reported event time, and b as the week where the event is reported. To illustrate, we give a and b and the

number of events that occurred in the time interval $[a, b]$ by considering the first 5 weeks of this data-set. Such overlapping time intervals in the form $[a, b]$ are used in this thesis to apply Turnbull's method. In the following intervals the number 0 indicates 39^{th} week, 1 presents 40^{th} week and 2 presents 41^{th} week and etc. The "R" output of our data treatment is the following:

a	b	event	cens
0	1	10374	1
0	2	11297	1
1	3	12127	1
2	4	12474	1
3	5	12490	1

We apply Turnbull's method to the flu data provided by ILINet, for the flu season $2014-2015$. Using the other method, we apply K-M at midpoints of time intervals $[a, b]$. Midpoints of $[a, b]$'s are the reported event time i.e. the reported week, so K-M at midpoints are survival estimates at the reported event time. In Figure 3.1, survival curves using both methods of Turnbull and midpoint, are presented. Indeed in this figure, we can see the difference of survival estimates at reported event time with applying interval censoring method to this data-set.

In Figure 3.1, the red lined curve shows survival curve, obtained by interval censoring method of Turnbull. In this method, we assume the exact event time happens in $[a, b]$. In Figure 3.1, the blue dotted line represents the survival curve at the reported weeks as the exact event time. As we see in Figure 3.1, by considering the reported weeks as the time of interest, we underestimate survivals. In the following table, we can see survival estimates for the first 10 weeks of $2014-2015$ flu season, calculated by R. 1.000002383091

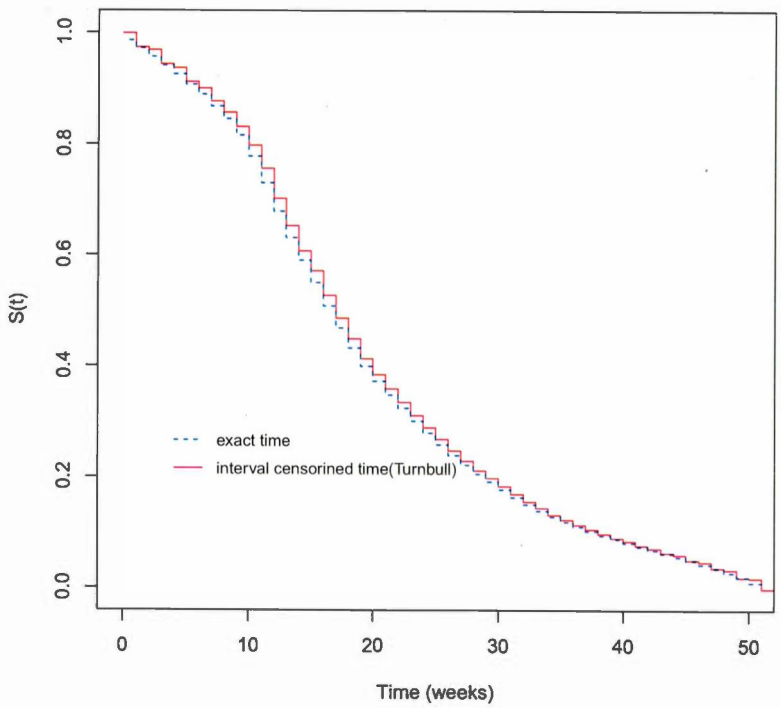

Figure 3.1 Comparing survival estimates in the flu season 2014-2015, using an empirical survival function and interval censoring method.

In Table 3.2, TB indicates Turnbull's estimate of survival and K-M indicates the empirical survival estimates at the right point of given $[a, b]$ time intervals as the exact reported event time.

In our analysis we need the R package "interval" introduced by Fay & Shaw (2010), manely the functions *icfit* and *ictest*.

Table 3.2 First 10 weeks survival estimates, 2014-2015, by two methods TB and K-M.

a	b(exact time)	TB	K-M
0	1	1.00	0.9868
0	2	0.9745	0.9725
1	3	0.9703	0.9571
2	4	0.9445	0.9412
3	5	0.9378	0.9225
4	6	0.9133	0.9075
5	7	0.9015	0.8896
6	8	0.8778	0.8680
7	9	0.8581	0.8453
8	10	0.8325	0.8157

3.2 Comparing different flu seasons

The question of interest in this section is "how the percentage of ILI changes in different years (flu seasons)?" First, as an example of comparing flu seasons, we look at "Line Chart ILINet" which is available in FluView. Figure 3.2 shows Line Chart ILINet of four recent flu seasons $(2011 - 2012)$ to $(2014 - 2015)$. Line Chart ILINet shows the percentage of visits for ILI through weeks, by graphic. Note that the given percentage of ILI is for all regions and age groups of each flu season together.

In CDC website, the Line Chart ILINet is given since $(1997 - 98)$ flu season up to now. By looking at the percentage of ILI in different flu season, some seasons seem to be similar. Line Chart ILINet for seasons of $(2012 - 2013)$, $(2013 - 2014)$ and $(2014 - 2015)$ have their maximum (highest percentage of ILI) at almost the same

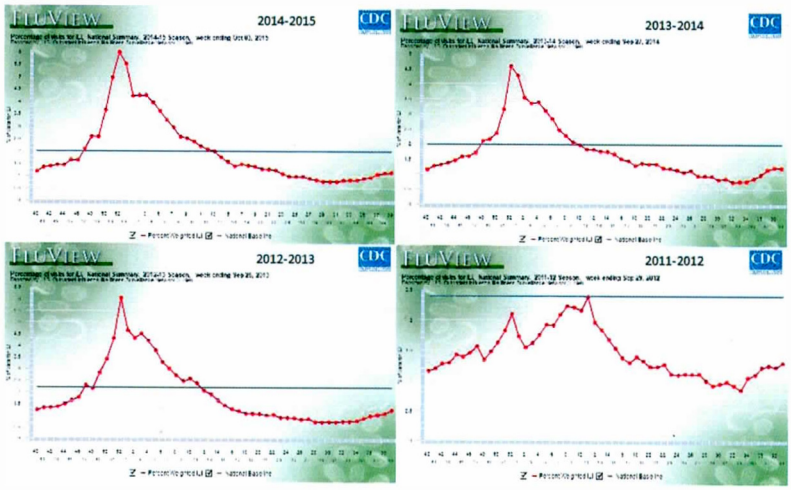

Figure 3.2 ILINet graphic, Percentage of visits for ILI, seasons 2011-2012 to 2014-2015.

week (52). Indeed for these three seasons the maximum number of events happen during roughly the same week of the year, which is the last week in December. On the other hand, the ILI activity for the flu season $(2011 - 2012)$ seems to be very different.

Table 3.3 Test result to compare percentage of visits for ILI, season 2013-2014 and 2014-2015.

t-test	p-value
1.3991	0.1652

We applied the related tests to compare two recent flu seasons and we find that they are identical. The result of t-test is given in the Table 3.3. Also we can look at the Line Chart ILINet of two recent flu seasons together.

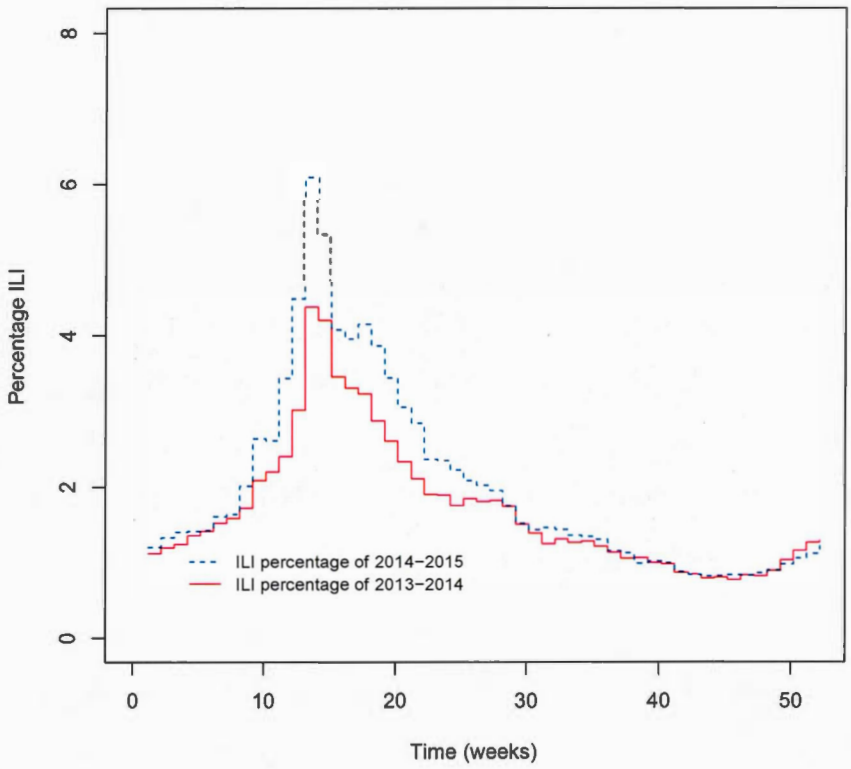

Figure 3.3 Percentage (weighted) of visits for ILI, seasons 2013-2014 and 2014-2015.

Figure 3.3 indicates the percentage of ILI for two different seasons 2014-2015 and 2013-2014. Figure 3.3 gives the impression that two distributions of ILI percentage are similar. In Section 3.2.1, our interest is in comparing flu seasons through their survival curves.

As noted in the first Chapter (Section 1.4.2), the survival function considered in our study is not the classical survival function. Since we are not following the same people over time, as explained in the first chapter, only the reported cases

are used in the analysis and they are treated as a cohort. In order to estimate the survival function, a fixed time τ_0 (typically 30 or 52 weeks) is picked and we work conditionally. The conditional survival function $S(t|\tau_0) = Pr(T > t|T \leq \tau_0)$ (Equation 1.10), is computed.

3.2.1 Comparing two recent flu seasons through their survival curves

Based on the the result of above mentioned statistical test, Table 3.3, the percentage of visits due to ILI in $(2014-2015)$ and $(2013-2014)$ flu seasons are the same. The interest here is to compare these two recent flu seasons, through their survival curves. First we compute the empirical survival function (K-M) at the given event time (reported week).

Figure 3.4 shows empirical survival curves for these two flu seasons, when the weekly reported event time is considered to be the event time of interest. In Figure 3.4, these two seasons' survival curves seem to intersect at some time point. Looking at these two K-M curves, we can not decide weather these two curves are the same or differ. We need to do a statistical test in order to compare them. In the next step, we will apply the method of Turnbull.

Turnbull's algorithm was shown in the second Chapter and in Table 3.2, we compared this method to the empirical survival estimates at the midpoints for flu season $(2014-2015)$. In this step of comparing flu seasons through their survival curves, the algorithm of Turnbull *(T-B)* as described in Giolo (2004) and Klein & Moeschberger (2005) is applied to estimate the survival curve. We present in Figure 3.5, the *T-B* survival curves of two recent flu seasons $(2014-2015)$ and $(2013-2014)$.

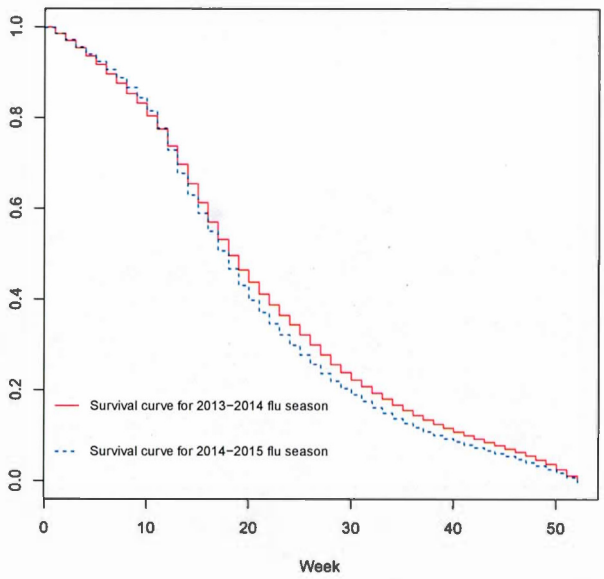

Figure 3.4 Empirical survival Estimates for two seasons, 2014-2015 and 2013-2014.

Also we apply interval censoring methods using the R package, *interval*, where the result is shown in Figure 3.6. When survival distributions are compared, looking at their curves does not provide us the answer to the question of whether they are statistically equivalent or not. In the following, some statistical tests will be applied to the data in either of two cases, first with considering interval censoring and second one in considered without interval censoring.

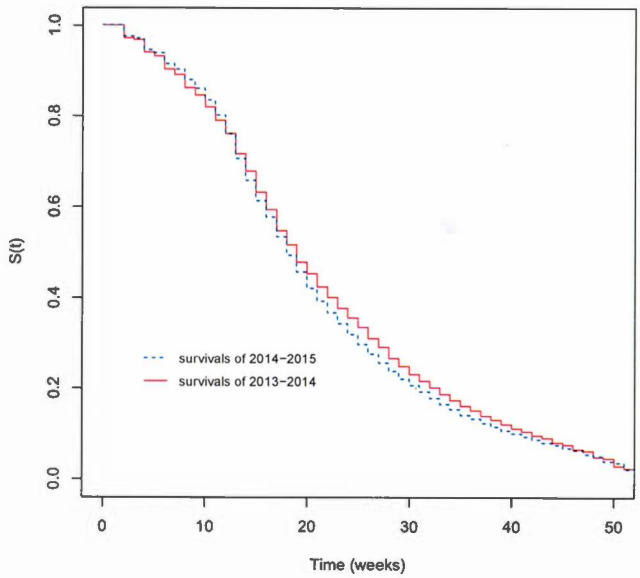

Figure 3.5 Survival curves using Turnbull's method for two seasons, 2014-2015 and 2013-2014.

3.2.2 Two different statistical tests for comparing survival functions in different seasons

Similar to the previous section, first we compute the K-M at the reported event time for two different seasons and then we apply the log-rank test to compare them. The following "R" output is the log-rank test result of comparing $(2013-2014)$ and $(2014-2015)$ flu seasons.

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
season=2013-14 668021   668021   698029      1290      2698
season=2014-15 787492   787492   757484      1189      2698

 Chisq= 2698  on 1 degrees of freedom, p==0
```

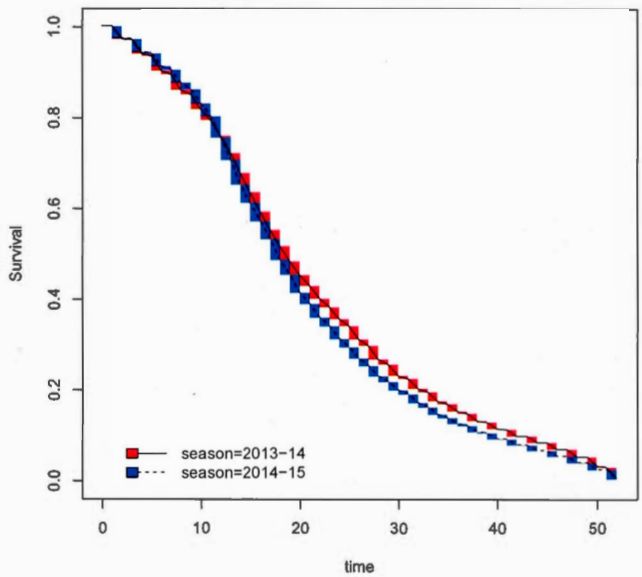

Figure 3.6 Survival curves using the *interval* "R" package for two seasons, 2014-2015 and 2013-2014.

The obtained p-value confirms that the two K-M survival curves are different. Using the log-rank test, the Kaplan-Meier survival curves in two flu seasons $(2013-2014)$ and $(2014-2015)$ are different. Now we apply interval censoring to the data of these seasons and do the test again. Interval package and icfit and ictest are used to do the asymptotic log-rank two sample test. Here is the "R" output:

```
Asymptotic Logrank two-sample test (permutation form)

data:  Surv(left, right, type = "interval2") by season
Z = -52.3132, p-value < 2.2e-16
alternative hypothesis: survival distributions not equal
```

```
                    n Score Statistic*
season=2013-14 668021          -29866.46
season=2014-15 787492           29866.46
* like Obs-Exp, positive implies earlier failures than expected
```

This result and the p-value less than 2.210^{-16}, shows that one reaches the same conclusion in both cases, weather one applies a test to interval censored event time data or to the exact event time data.

3.2.3 Controlling for age when comparing flu seasons

In the previous section, survival curves of two recent flu seasons were compared, and their difference was found significant. In two different flu seasons there are many different components which are different and can affect the survival function. For instance, people of two seasons are not the same, flu types are different, the number of people in each age group is not the same. If the comparison of seasons is just among patients who are in a specified age group, then the comparison can make sense.

We study two recent flu seasons $2014-2015$ and $2013-2014$, by age group; since there are five different age groups, we have five comparisons to perform. Table 3.4 gives the results of the statistical tests, log-rank on the original data and IC log-rank for the interval censored data.

According to the result of the test statistic and its p-value for each comparison, even limiting our comparison among people within the same age group, leads to a significant difference between two recent flu seasons. In addition, we compared these seasons, only among people of one specific region. One concludes that the

Table 3.4 Tests for comparing survival functions of two recent flu seasons by age.

Age group	Test Method	tests Statistic	p-Values
$0-4$	Log-Rank	$\chi^2 = 533$	$p \simeq 0$
	IC Log-Rank	$z = -23.2544$	$p < 2.2e-16$
$5-24$	Log-Rank	$\chi^2 = 2473$	$p \simeq 0$
	IC Log-Rank	$z = -50.2784$	$p < 2.2e-16$
$25-49$	Log-Rank	$\chi^2 = 53.6$	$p = 2.42e-13$
	IC Log-Rank	$z = -7.2855$	$p = 3.204e-13$
$50-64$	Log-Rank	$\chi^2 = 22.6$	$p = 2.04e-06$
	IC Log-Rank	$Z = -4.7326$	$p = 2.216e-06$
$65+$	Log-Rank	$\chi^2 = 476$	$p \simeq 0$
	IC Log-Rank	$Z = -22.0749$	$p < 2.2e-16$

survival curves of the two different flu seasons $2014-2015$ and $2013-2014$ are always different. Furthermore, the comparison of survival curves, among people of a fixed age group in a given region of two different recent flu seasons results in a significant difference.

3.3 Factor age: comparing survival functions across age groups, in the same season

The other comparison that we want to make is to compare reported age groups, using their survival curves. First we need to choose a flu season to the study. In what follows, we look at a recent flu season where complete data is available.

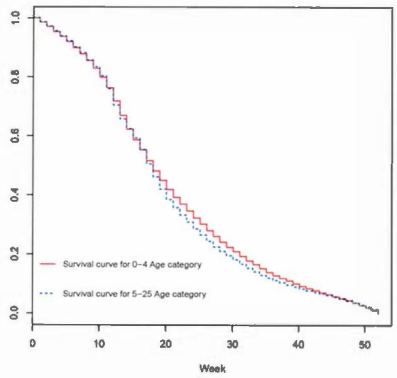

Figure 3.7 K-M survival curves of 0-4 and 5-24 age groups, flu season 2014-2015.

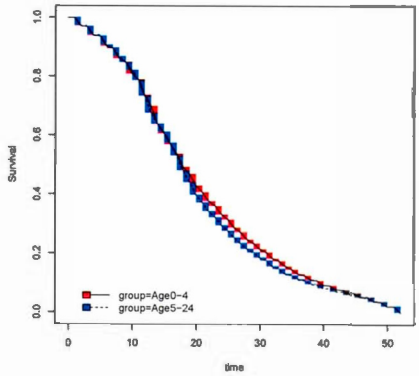

Figure 3.8 IC survival curves of 0-4 and 5-24 age groups, flu season 2014-2015.

3.3.1 Comparing contiguous age groups in the flu season 2014 – 2015

In this section the flu season (2014 – 2015) is studied. The reported age groups, as shown in Table 2.1, are available in five categories $0-4$, $5-25$, $25-49$, $50-64$, and 65+. The following question will be the interest of the study of this section. Are survival curves the same in different available age groups of a specified flu season? The first age category and the second category seem to be different conceptually. Babies and very young children are in the first age groups while children and young people are in the second age group. The immune system of individuals in these groups and also the probability of getting the flu is not the same. In Figure 3.7, K-M survival curves of the age groups $0-4$ and $5-25$, are shown. The event time is considered as the reported exact event time in K-M curves of Figure 3.7. Figure 3.8 shows estimated survival curves of the same age groups, but the event time is considered to be interval censored as described in Section 3.1. It seems that both Figures 3.7 and 3.8 give the same results when comparing the estimated survival curves, but we need to perform a hypothesis

test for each comparison and compare the results.

The value of the two log-rank test, calculated by R, when comparing $0-4$ and $5-24$ age groups are given below:

If the reported event time is considered as the event time of interest, we have:

$$\chi^2 = 352 \qquad p < 10^{-16} \tag{3.1}$$

and the result of the asymptotic log-rank test shows that the difference between two survival curves of two given age groups, is not negligible.

If interval censoring is applied to the flu data, we obtain:

$$Z = -19.0233 \ , \quad p < 2.210^{-16} \tag{3.2}$$

and the result of the asymptotic IC log-rank two-sample test confirms the results of the previous test. Therefore it seems that the survival estimates are different in two age groups $0-4$ and $5-24$ of $2014-2015$ flu season, and the results obtained through two different methods coincide.

If there exists the possibility to have similar survival curves in two different age groups, conceptually it should be between the age groups of $25-49$ and $50-64$. People in both of these age groups are adults and they are not very old or very young, and their survival functions should not be very different.

Figures 3.9 and 3.10 show comparing survival curves of two age groups $(25-49)$ and $(50-64)$. K-M survival curves at the reported event time are shown in Figure 3.9 and interval censoring (IC) is applied to the data at the basis of the survival curves in Figure 3.10. Survival curves seem very similar in these age groups, namely $(25-49)$ and $(50-64)$, using both methods of K-M at the reported event

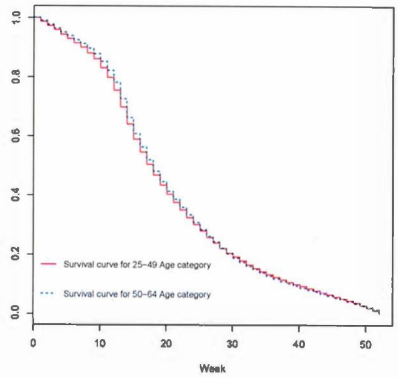

Figure 3.9 K-M survival of age groups 25-49 and 50-64, flu season 2014-2015.

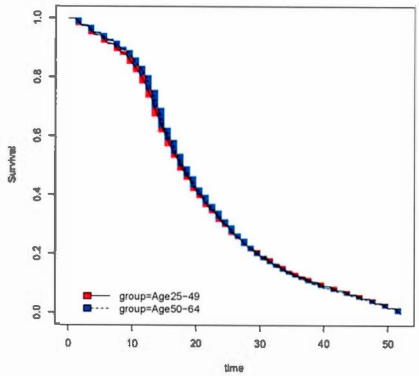

Figure 3.10 survival curves using interval package of age groups 25-49 and 50-64, flu season 2014-2015.

time and interval censoring. Doing hypothesis tests is needed to verify the possibility of not rejecting H_0.

The result of the two log-rank tests, calculated by R, to compare $(25-49)$ and $(50-64)$ age groups are given below.

If the reported event time is considered as the event time of interest, we have

$$\chi^2 = 2.5, \qquad p = 0.112. \tag{3.3}$$

The value of the log-rank test statistic is very small, 2.5 with one degree of freedom and the p-value bigger than 0.05 confirms that the difference of survival curves between given age groups is negligible.

If interval censoring is applied to the flu data, we have

$$Z = 1.5919, \qquad p = 0.1114, \tag{3.4}$$

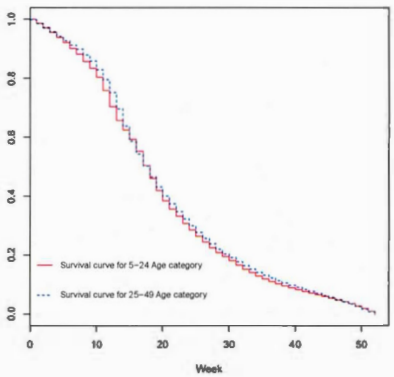

Figure 3.11 K-M survival curves of 5-24 and 25-49 age groups, season 2014-2015.

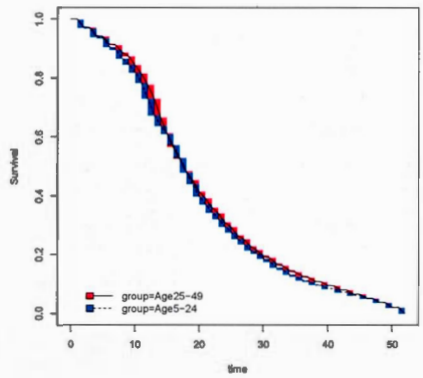

Figure 3.12 Ic survival curves of 5-24 and 25-49 age groups, season 2014-2015.

and the result of the asymptotic IC log-rank two-sample test confirms the result of the previous test. Therefore it seems that the survival estimates are very similar in the two age groups $25-49$ and $50-64$ of $2014-2015$ flu season.

Further, we compare other contiguous age groups among the five available age groups of the $2014-2015$ flu season data. First, we compare the $5-24$ and $25-49$ age groups.

Figures 3.11 and 3.12 show this comparisons of survival curves using the reported event time as the event time of interest and using interval censoring methods. Two age groups survival curves do not seem very different using both methods, but the result of statistical tests and a very small p-value confirms the difference of survival estimates, in the two mentioned age groups. In Table 3.4, the result of statistical tests using different methods is given.

Table 3.5 Statistical test results of comparing survivals in some age groups of (2014-2015) flu season

Test Method	Age groups	Test Statistic	p-value
Log-Rank	5-24 and 25-49	$\chi^2 = 165$	$p \simeq 0$
IC Log-Rank	5-24 and 25-49	$Z = 12.9737$	$p < 2.2e-16$
Log-Rank	50-64 and 65+	$\chi^2 = 18.1$	$p = 2.07e-05$
IC Log-Rank	50-64 and 65+	$Z = -4.2178$	$p = 2.467e-05$

The other comparison of survival functions across age groups which is studied here is comparing of $50-65$ and 65+ age groups. Table 3.1 shows the statistical R-results, when two different methods of K-M (survdiff) and interval censoring (ictest) applied to the $(2014-2015)$ flu season data, to compare survival curves of two age groups $50-64$ and 65+ and of the other contiguous two age groups $5-24$ and $25-49$. In Table 3.1 and in following tables, the interval censoring method (icfit), is showed by IC Log-Rank. As it is clear in the R-results from both indicated methods, showed in Table 3.4, two age groups $50-64$ and 65+ have different survival estimates in the mentioned flu season.

All survival curves through consecutive age groups of $2014-2015$ flu season data were compared in Section 3.4.1. Considering the reported event time as the event time of interest and considering the event time to be interval censored, we applied the proper log-rank test. Consequently for all contiguous age groups, survival estimates are different except for $(25-49)$ and $(50-64)$. Now the following question arises. In other flu seasons, are survival curves similar in $25-49$ and $50-64$ age groups? Finding the answer of this question is the aim of the following section.

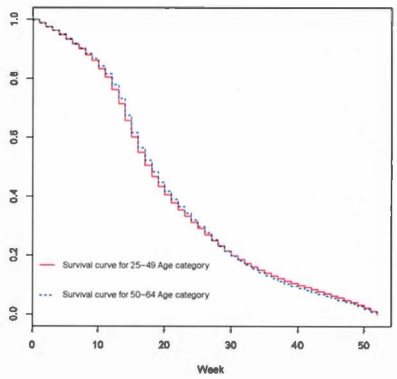

Figure 3.13 K-M survival curves of $25-49$ and $50-64$ age groups, season 2013-2014.

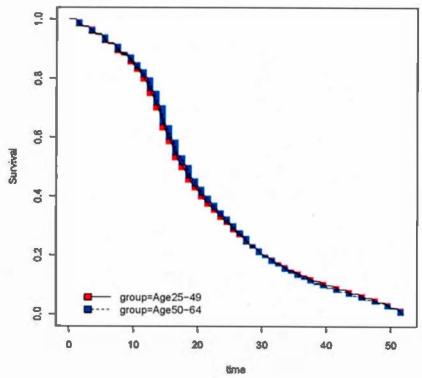

Figure 3.14 IC survival curves of $25-49$ and $50-64$ age groups, season 2013-2014.

3.3.2 Comparing adults survival in five flu seasons

In this section, we compare survival curves between the two adult's age groups $25-49$ and $50-64$.

Flu season $2013-2014$:

Figures 3.13 and 3.14 compare survival curves of age groups $(25-49)$ and $(50-64)$. Survival curves through two indicated age groups using both methods seem very similar. In Figure 3.13, survival estimates are computed at the reported event time, and in Figure 3.14 the event time of interest is considered to be interval censored. The results of log-rank tests, applied to the original data and considering interval censoring is given in the following table.

The small value of the test statistics and having a p-value moderately large confirm that we can not reject the hypothesis of equality of survival estimates in two

Table 3.6 Flu season $2013-2014$, age groups $(25-49)\&(50-64)$.

Test Method	Test Statistic	p-value
Log-Rank	$\chi^2 = 1.92$	$p = 0.166$
IC Log-Rank	$Z = 1.3829$	$p = 0.1667$

mentioned age groups of $2013-2014$ flu season. Therefore, in the $2013-2014$ flu season, for the comparison of $(25-49)$ and $(50-64)$ age groups, the same conclusion of no difference between survival functions was obtained as in the $2014-2015$ flu season.

Flu seasons $2012-2013$, $2011-2012$, $2010-2011$:

Given that survival curves look quite similar, in what follows we study the result of statistical tests. The following Table 3.7 summarizes our results for these flu seasons.

Table 3.7 Test statistic of three flu seasons, age groups $(25-49)\&(50-64)$.

Flu season	Test Method	Test Statistic	p-value
2012-2013	Log-Rank	$\chi^2 = 4.1$	$p = 0.042$
	IC Log-Rank	$Z = 1.9526$	$p = 0.05086$
2011-2012	Log-Rank	$\chi^2 = 6.1$	$p = 0.0133$
	IC Log-Rank	$Z = -2.5909$	$p = 0.009574$
2010-2011	Log-Rank	$\chi^2 = 7.9$	$p = 0.00501$
	IC Log-Rank	$Z = 2.8031$	$p = 0.005062$

For the $2012-2013$ we note that the conclusions of the log-rank and the IC log-rank tests do not coincide. Thus applying interval censoring on the data, as

explained in Section 3.1, can change the conclusions in comparing survival estimates. Equivalently, the result of comparing survivals is not always the same, when the reported event time in the data is considered to be exact or interval censored.

As for the flu season $2011-2012$ and $2010-2011$ the p-value of both tests are less than the significance level, and survival curves using both methods seem different.

Although one expects that survival curves of the two age groups $(25-49)$ and $(50-64)$ to be very similar, we cannot conclude that these two age groups have the same survival curves in all flu seasons, an interesting fact.

3.4 Comparing different U.S. regions

In the *FluView* report of *CDC*, for each season, the flu data is also available by **U.S regions**. Is the percentage of ILI comparable in diferent U.S regions? In order to answer this question it would be interesting to compare U.S regions through their survival curves. Indeed, comparison of the Line Chart ILINet or percentage of visits for ILI, through different regions will motivate us to study the survival curve in different regions.

Figure 3.15 shows the percentage of ILI by week, for all 10 U.S regions in the recent flu season $2014-2015$. In general, since the ILI percentage graph is different for different regions of U.S. (Figure 3.15), they seem to have different survival curves.

The Line Chart ILINet report of region 1 and 3 do not seem to be very different,

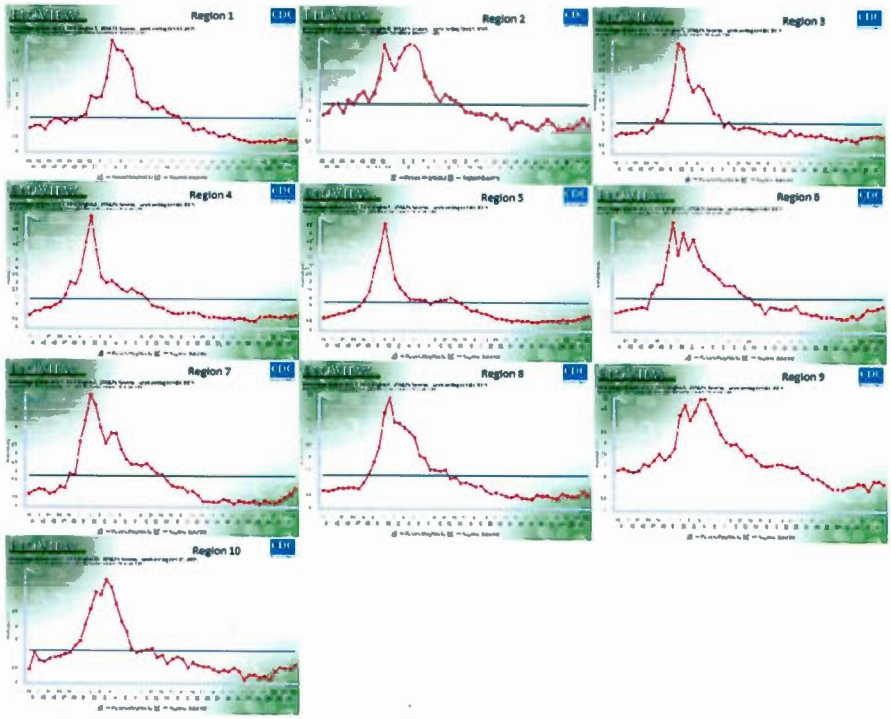

Figure 3.15 ILINet graph, through 10 different regions, 2014 – 2015 flu season.

and the same is true for regions 4 and 5. Indeed we pick these two regions 4 and 5 to study their survival curve in a given flu season 2014-2015. In Figures 3.16 and 3.17, we present the related survival estimates at the exact event time and in the interval censored data. The result of statistical tests is given to confirm the conclusion. All the analysis of this section is on the data of the recent flu season 2014 – 2015.

3.4.1 Comparing regions in the flu season 2014-2015

Regions 1 and 3:

Figures 3.16 and 3.17 give survival curves of regions 1 and 3 and they seem to

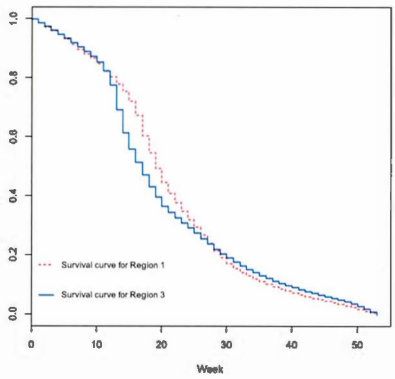

Figure 3.16 K-M survival curves of regions 1 and 3 (full line), season 2014-2015.

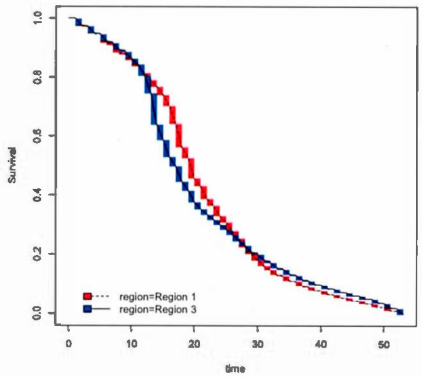

Figure 3.17 IC survival curves of regions 1 and 3, season 2014-2015.

be very different. This is not surprising, as region 1 covers northern states and region 3 southern ones.

Table 3.8 Flu season $(2014 - 2015)$, Region 1 & 3.

Test Method	Test Statistic	p-value
Log-Rank	$\chi^2 = 46.3$	$p = 1.03e - 11$
IC Log-Rank	$Z = -7.0255$	$p = 2.132e - 12$

The result of hypothesis tests, presented in Table 3.8, confirms that survival functions through region 1 and region 3 are different.

Regions 4 and 5:

Survival curves to compare regions 4 and 5 are shown in Figure 3.18 and 3.19, and they seem to be very different.

The test statistic value χ^2 and z, in Table 3.9 implies that the difference of survival

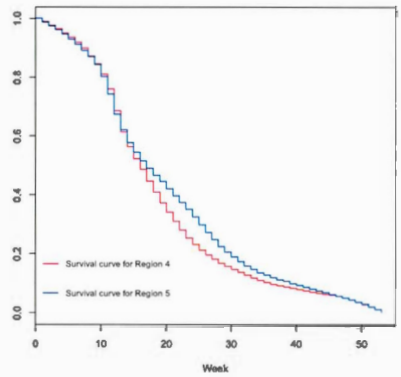

Figure 3.18 K-M survival curves of regions 4 and 5, season 2013-2014.

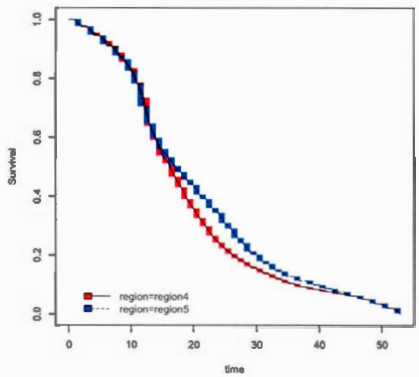

Figure 3.19 IC survival curves of regions 4 and 5, season 2013-2014.

functions in regions 4 and 5 is not negligible.

Table 3.9 Flu season $(2014-2015)$, Region 4 & 5.

Test Method	Test Statistic	p-value
Log-Rank	$\chi^2 = 472$	$p \simeq 0$
IC Log-Rank	$Z = 22.2502$	$p < 2.2e-16$

Since the survival curves through different regions seem to be very different, we expect that performing comparison tests through U.S. regions should only confirm this difference. For each region and in a given season, the CDC website provides also the number of ILI in different age groups. Using this number of events given in different age groups for a given U.S. region, we can compare survival functions through age groups, in a given region and season.

In the flu season $2014-2015$, we pick a region (region 1), and then the comparison

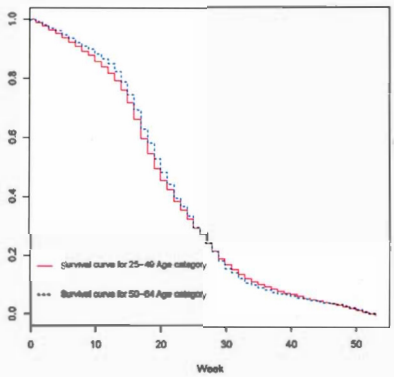

Figure 3.20 K-M survival curves of two age groups in regions 1, season 2014-2015.

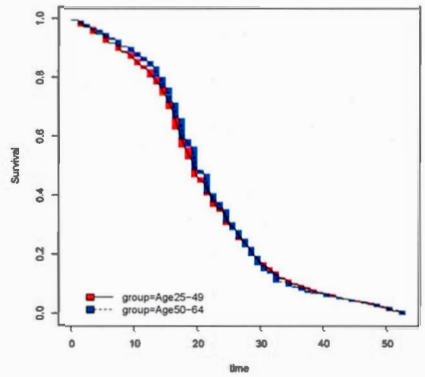

Figure 3.21 IC survival curves of two age groups in regions 1, season 2014-2015.

of survival curves among two age groups, $25-49$ and $50-64$, are studied. Figures 3.20 and 3.21 show the comparison of survival curves through the indicated age groups and they seem very similar. The test values in Table 3.10 confirms the similarity of survival curves.

Table 3.10 Region 1, Comparing Age groups $(25-49)\&(50-64)$.

Test Method	test Statistic	p-Value
Log-Rank	$\chi^2 = 1.3$	$p = 0.261$
IC Log-Rank	$Z = 1.0096$	$p = 0.3127$

The p-values in Table 3.10 confirm that the difference between survival curves in Figures 3.20 and 3.21 is not significant. To compare survival functions among the $25-49$ and $50-64$ age groups, the same result is obtained in the data of region 1 as in the national data.

In the illustrative examples of Section 3.3, we studied survival curves through age groups and found that the $25-49$ and $50-64$ age groups have similar survival functions in some flu seasons. Now, the following question is the motivation of the following section. In a given U.S. region of a fixed flu season, are survival functions of the $25-49$ and $50-64$ age groups still the same?

3.4.2 Comparing the $25-49$ and $50-64$ age groups by regions

In the given season $2014-2015$, for each of 10 U.S. regions, we compare survival functions among these specified age groups using the log-rank tests, and the result of this comparison is given in Table 3.11. Comparing these age groups in each of the 10 regions separately give us an interesting result.

As we can see in Table 3.11, when comparing the survival function of two age groups $(25-49)$ and $(50-64)$ by region, we no longer have the equality of survival functions as in the national data. In Section 3.4, we studied that the difference of survival curves between two mentioned age groups is insignificant in season 2014-2015, however in the regions 2, 4, 5 and 9, the survival curves of the two age groups are different.

3.5 Comparing flu types in season $2014-2015$

Finally, as shown in Chapter 1, the CDC is also testing the influenza strain for some individuals. Figure 3.22 shows how the given flu type graph is posted on FluView, in the $2014-2015$ season. Different colors show different available flu type of the give season. In Figure 3.22, the number of positive tests for three chosen weeks is given. For the $2014-2015$ season, available flu types are:

Table 3.11 Comparison of the survival in the $(25-49)$ and $(50-64)$ age groups by region, $2014-2015$ season.

Region	Test Method	Tests Statistic	p-values
Region 1	Log-Rank	$\chi^2 = 1.3$	$p = 0.261$
	IC Log-Rank	$z = 1.0096$	$p = 0.3127$
Region 2	Log-Rank	$\chi^2 = 16.3$	$p = 5.3e-05$
	IC Log-Rank	$z = -3.9443$	$p = 8.003e-05$
Region 3	Log-Rank	$\chi^2 = 1$	$p = 0.323$
	IC Log-Rank	$z = -1.0495$	$p = 0.2939$
Region 4	Log-Rank	$\chi^2 = 14.7$	$p = 0.000127$
	IC Log-Rank	$Z = 3.8743$	$p = 0.0001069$
Region 5	Log-Rank	$\chi^2 = 5$	$p = 0.0256$
	IC Log-Rank	$Z = 2.1302$	$p = 0.03315$
Region 6	Log-Rank	$\chi^2 = 2.4$	$p = 0.119$
	IC Log-Rank	$Z = 1.4151$	$p = 0.157$
Region 7	Log-Rank	$\chi^2 = 3.7$	$p = 0.0561$
	IC Log-Rank	$Z = 1.9336$	$p = 0.05316$
Region 8	Log-Rank	$\chi^2 = 1.5$	$p = 0.219$
	IC Log-Rank	$Z = -1.2623$	$p = 0.2068$
Region 9	Log-Rank	$\chi^2 = 7.5$	$p = 0.00629$
	IC Log-Rank	$Z = 2.9088$	$p = 0.003628$
Region 10	Log-Rank	$\chi^2 = 0.6$	$p = 0.439$
	IC Log-Rank	$Z = 0.7798$	$p = 0.4355$

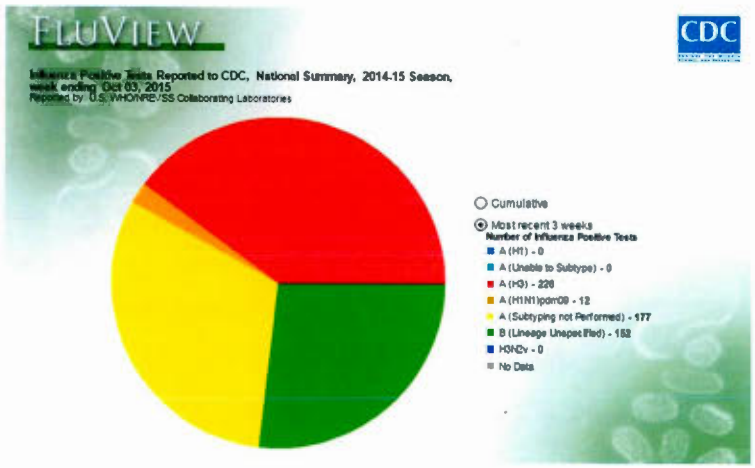

Figure 3.22 FluView report for flu type, $2014 - 2015$ season.

1. $A(H3)$

2. $A(H1N1)$

3. A(Subtyping not Performed)

4. B(Lineage Unspecified)

We characterize the above flu types by numbers 1 to 4. For example by type 1, we mean the $A(H3)$ flu type. In this section we will compare survival curves between different flu types. First, we pick two types 3 and 4 to compare their survivals. In Figure 3.22 the proportion of type 3 is in yellow and type 4 is in green. Figures 3.23 and 3.24 show the comparison of survival curves of flu type 3 and 4 using empirical survival curves at the exact event time and survival curves at interval censored time (IC survival curves), respectively. By comparing survival curves in Figures 3.23 and 3.24, it seems that survival curves of type 3 and 4 are very different, and one (type 4) is definitely above the other.

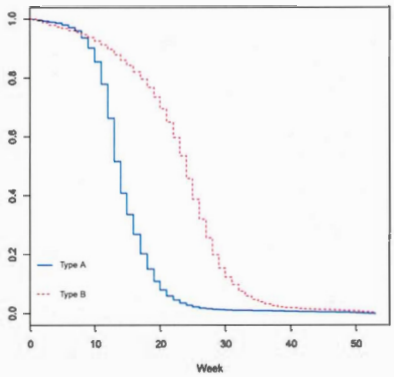

Figure 3.23 Empirical survival curves of two flu virus types (type A, lined), 2014-2015.

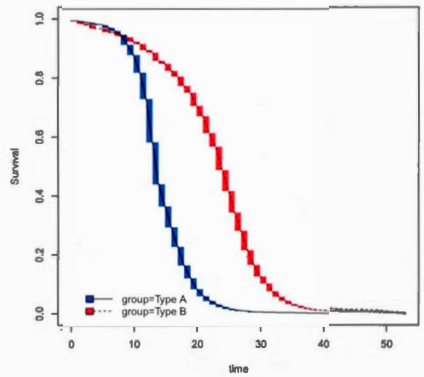

Figure 3.24 IC survival curves of two flu virus types, 2014-2015.

The statistical comparison of all available flu types in 2014-2015, using their survival estimates is given in Table 3.12. The result of statistical tests in Table 3.12 confirms the conclusion obtained from Figures 3.23 and 3.24.

The result of statistical tests and p-values in Table 3.12 shows that different flu types have very different survival curves and the difference of their survival is not negligible.

Table 3.12 Results of comparing survival by different flu types, 2014 – 2015 season.

Comparing flu types	Test Method	Tests Statistic	p-values
Type 1 and 2	Log-Rank	$\chi^2 = 180$	$p \simeq 0$
	IC Log-Rank	$z = 18.4478$	$p < 2.2e-16$
Type 1 and 3	Log-Rank	$\chi^2 = 428$	$p \simeq 0$
	IC Log-Rank	$z = -20.6042$	$p < 2.2e-16$
Type 1 and 4	Log-Rank	$\chi^2 = 19019$	$p \simeq 0$
	IC Log-Rank	$z = 140.363$	$p < 2.2e-16$
Type 2 and 3	Log-Rank	$\chi^2 = 225$	$p \simeq 0$
	IC Log-Rank	$Z = -21.4839$	$p < 2.2e-16$
Type 2 and 4	Log-Rank	$\chi^2 = 3.9$	$p = 0.0487$
	IC Log-Rank	$Z = -1.8495$	$p = 0.06439$
Type 3 and 4	Log-Rank	$\chi^2 = 22591$	$p \simeq 0$
	IC Log-Rank	$Z = 150.7147$	$p < 2.2e-16$

CONCLUSION

In this thesis, we studied ILI (Influenza Like Illness) data available on the CDC (Center for Disease Control and Prevention, USA) website. Using the ILI data, we created a cohort and considered a conditional survival function. Since the number of ILI cases is reported once per week, the exact time of getting the flu is not known and the event time was considered to be interval censored; indeed, this event can occur during the week preceding the reported one. Therefore, we considered an interval censored approach and interval censoring estimation methods were applied to this data in order to estimate the survival function; moreover, we computed the empirical survival function, by letting the reporting week as the exact event time.

Since the percentages of ILI in two recent flu seasons $2013-2014$ and $2014-2015$ seemed very similar, we applied our survival analysis methodology to the ILI data of these seasons in order to assess whether we obtain a similar conclusion or not. Computing the empirical survival estimates at the reported flu times and further applying an interval censored estimation method, we concluded that these flu seasons are different. Also, we compared survival functions of the above flu seasons, by age group, by region, and for a fixed age group in a given region. In every such case the resulting estimators of the survival functions turned out to be different. A log-rank test was used for the comparison of the survival functions at an exact event time and by considering interval censoring.

Some of the most interesting results were obtained in comparing two contiguous age groups (forming the "adults"), where we found significant differences among these groups in some regions, but not in the national data. As expected, we also found that survival functions differ dramatically across flu types. Another important finding is that applying interval censoring methods or empirical survival estimation can lead to different conclusions in some of these cases. We conclude that our estimation methodology can prove interesting for further studies of these data.

BIBLIOGRAPHY

Chowell, G. & Nishiura, H. (2008). Quantifying the transmission potential of pandemic influenza. *Physics of Life Reviews*, *5*(1), 50–77.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Fay, M. P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics*, pp. 811–822.

Fay, M. P. (1999). Comparing several score tests for interval censored data. *Statistics in Medicine*, *18*(3), 273–285.

Fay, M. P. & Fay, M. M. P. (2013). Package 'interval'. *R Project*.

Fay, M. P. & Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the interval r package. *Journal of Statistical Software*, *36*(2).

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, pp. 845–854.

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, *52*(1-2), 203–223.

Giolo, S. R. (2004). Turnbull's nonparametric estimator for interval-censored data. *Department of Statistics, Federal University of Paraná*, pp. 1–10.

Huang, J., Lee, C. & Yu, Q. (2008). A generalized log-rank test for interval-censored failure time data via multiple imputation. *Statistics in Medicine*, *27*(17), 3217–3226.

Kalbfleisch, J. D. & Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Klein, J. P. & Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kleinbaum, D. G. & Klein, M. (2006). *Survival analysis: a self-learning text*. Springer Science & Business Media.

Law, C. G. & Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in medicine*, *11*(12), 1569–1578.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, *58*(303), 690–700.

Parodi, S., Bottarelli, E. et al. (2008). Survival analysis in epidemiology: a brief introduction. *Annali della Facoltà di Medicina Veterinaria, Università di Parma*, *28*, 17–42.

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pp. 185–207.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 290–295.