A Statistical Analysis of the Average Waiting Time Between Flares in Lupus Patients

Lee A. Strassenburg

19 Nov 2004

Abstract

In this study I investigate the effect of race and type of lupus on the manifestation of the disease, specifically through the average number of days between recurrences of flares due to the disease. Lupus is an auto-immune disease that doctors and researchers do not completely understand. The most immediate concerns related to understanding the disease include a lack of knowledge as to what causes lupus and how to treat it. Lupus is a disease with recurring flares interspersed with periods of remission. Certain groups of lupus patients have more frequent relapses than other groups. Researchers are trying to determine what factors are significant in causing the onset of flares, and which groups of lupus patients might be more susceptible to shorter periods of time between flares, or more intense flares. The dataset I use to study lupus was collected by the Department of Nephrology at The Ohio State University. In this manuscript I explore the underlying distributions for the waiting times between flares for each of the groups I am comparing: African-Americans versus Caucasians, and renal versus non-renal lupus patients. Because of the large number of censored datapoints, survival analysis is the most readily available method to analyze the waiting time between flares. In order to understand the data more completely, I use both nonparametric and parametric methods to compare and to estimate the differences between African-Americans and Caucasians and between renal and non-renal lupus patients. Specifically, I use Kaplan-Meier curves and the logrank test to compare the survival rates between groups, the likelihood ratio test to compare estimates under the assumption that lupus flares are a Poisson Process, and the Kolmogorov-Smirnov goodness-of-fit test to determine which distribution most closely resembles the waiting time between flares in lupus patients. With this information, doctors will be better equipped to monitor the progress of lupus patients and to recommend further treatment and follow-ups.

1 Introduction

1.1 History of Lupus

Lupus is a widespread auto-immune disease for which doctors and researchers can determine neither a cause nor an effective treatment. In affected lupus patients, rather than protect the body from disease and foreign materials, the immune system attacks itself, destroying tissues and organs including the joints, kidneys, heart, lung, brain, blood, or skin. Symptoms range from mild to life-threatening, though it is most common for only one or two organs in a given patient to be affected. While there are three basic types of lupus, the most common is Systemic lupus, affecting seventy percent of lupus patients. Of those patients with Systemic lupus, about half of patients experience severe symptoms in a major organ, while the other half of patients have a more mild version that does not affect any of the previously mentioned organs (for more information, visit www.lupus.org).

The Lupus Foundation of America used a nationwide telephone survey to estimate that approximately 1.5 million Americans are affected by some form of lupus. While it can affect men and women of all races and ages, lupus occurs ten to fifteen times more frequently among women than men, and two to three times more frequently in African Americans, Hispanics, Asians, and Native Americans than Caucasians. Although scientists believe there is a genetic pre-disposition to lupus, only ten percent of people with lupus have a parent or sibling who will develop the disease, and only five percent of children born to people with lupus will develop the disease.

Systemic lupus erythematosus (SLE) is characterized by 'flares' of activity interspersed with periods of remission. Flares are marked by the physical worsening of symptoms as well as heightened biological indicators such as increased levels of specific hormones. In this study, we follow renal flares, the flares which affect the kidney. Renal activity affects approximately fifty percent of lupus patients.

1.2 Data

In a longitudinal study conducted by The Ohio State University Medical Center Department of Nephrology, each of seventy-six patients reported for check-ups every two months. During their scheduled visit, a doctor recorded whether the patient was in a state of remission or whether the patient was experiencing a flare. By tracking the number of days between flares, we hope to determine whether there is a statistically significant difference between the average number of days between flares for renal versus non-renal lupus patients and for African-American versus Caucasian lupus patients. In order to better understand the affect race and type of lupus have on the patients, we not only collect and analyze data on the number of days between flares, but we also need to determine the underlying distribution for waiting times between flares for each group of patients. Analyzing the differences in the number of days between flares gives an understanding of which groups of people are at a higher risk of having frequent recurrences of flares.

Of the seventy-six total patients, thirty were African-American and forty-six were Caucasian; forty-eight had renal lupus and twenty-eight had non-renal lupus (see table 1). Some patients experienced multiple flares during the observation period while others did not experience any flares during the observation period. In addition, because the

patients entered the study at different dates, the total number of observation days varies across groups and from one patient to another.

Race	Type of Lupus	Number of Patients	Uncensored Flares	Censored Flares
African-American	Renal	17	14	28
African-American	Non-Renal	10	0	42
Caucasian	Renal	28	10	39
Caucasian	Non-Renal	16	3	23

Table 1: Data Summary

One problem with the data analysis is that patients experiencing flares were treated in order to give relief from pain and discomfort. In treating the patients, there may be a chance that the probability of a subsequent flare decreased. Thus the underlying distribution would be affected by doctor intervention, and the data are not be completely independent.

2 Survival Analysis

Survival analysis is the study of the time until a specified event. In the medical setting, 'event' often refers to death or to the relapse of a disease. In studying lupus, we consider a flare as the desired 'event' of interest, and thus we study the time between flares. Survival analysis is a useful tool not only because of its ability to describe skewed data (some time intervals are extremely large) but also because of its ability to handle censored data.

Censored data occurs when the actual time between events is either longer or shorter than the observed time. If the actual time between events is longer than the observed time, we call this *right-censoring*, and if the actual time between events is shorter than the observed time, we call this *left-censoring*. An example of a right censored data point is one that measures the number of days a cancer patient survives after undergoing some form of treatment. The data will be uncensored if person dies while under observation. The data will be right censored if the person either moves out of state or the study ends before the person dies. In the case of one of these two latter events, the actual time until death is greater than the number of days the patient was observed under the study. An example of a left censored study is one in which researchers are trying to determine the average age teenagers first used marijuana. In a questionnaire, each person might be asked at what age he or she first used marijuana. If the person has tried marijuana and could remember at what age he or she tried it, then the data is uncensored. If the person only remembered that the first time was sometime before the age of twenty, say, the data would be left censored. The only thing we know in the latter case is that the true age at which he or she first used marijuana is less than twenty years, and thus the data is left censored.

In the lupus study, when the data are censored it is always right-censored because the amount of time between flares is always longer than the observed number of days. There are censored data points for all patients excluding those who are both in the middle of a flare on the first day of the study and in the middle of a flare on the last day

of observation. The censored observations occur by the nature of the disease because we know there will be subsequent flares after the end of the study regardless of the length of the study, and we know the patients experienced more than one flare prior to the beginning of the study. Unfortunately, the exact time of both the flare prior to the beginning of the study and the flare immediately following the conclusion of the study are unknown. For example, although we know every lupus patient had a flare at some point prior to the first day of observation, we did not record the date of the flare and thus we only know that the true period of time between the previous flare and the first observed flare is greater than the number of days between the first day of observation and the first observed flare (see figure 1). For example, flare 1 occurs before the observation period, and thus the time between the first and second flare is considered censored data. Because both flares 2 and 3 occur during the observation period, the waiting time between these two flares is known and is considered uncensored data. Similar to flare 1, flare 4 is censored because the actual amount of time between the last observed flare (here, flare 3) and the ssubsequent flare (flare 4) is unknown. We only know that the time between flares 3 and 4 is longer than the time between flare 3 and the end of the observation period. Thus the censored data are right censored.

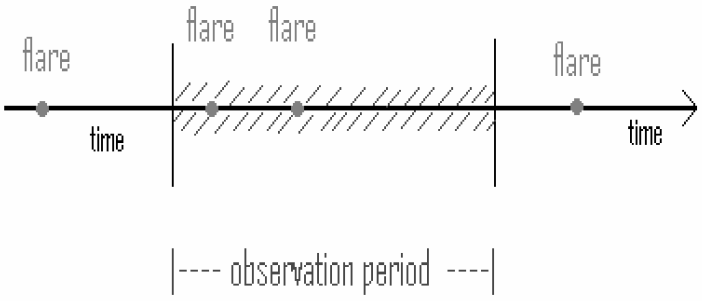

Figure 1: Censored Data

In this study we have a relatively large number of censored data points. Because flares occur infrequently in lupus patients, even if the study is run over a long period of time, like four years as this study will be run for, the number of uncensored observations remains small as compared to the number of censored observations. Nearly every patient in the study has at least one or two censored flares, but some patients only contribute one uncensored flare to the study. Furthermore, some patients do not contribute any uncensored data but rather they only contribute censored data to the study. It is important to note that we must assume censoring is non-informative about the actual number of days between flares [Frank E. Harrell, 2001], i.e., that the censor is caused by a factor independent of the observation of a flare. Thus it is important that censored data are not caused, say, by a deteriorating condition but rather because of a reason unrelated to the variable of interest. If the censor were caused because patients did not show up for their appointments when their condition was bad, then the information would be biased away from the shorter waiting times as a censored data point does not provide as much information as an uncensored data point.

2.1 Survival Analysis Notation

In order to study censored data when the variable of interest is the time until an event, it is important to understand survival functions and how they are related to known probability functions. The survival function, $S(t)$, is the probability that a flare occurs after a given time t, or that a person 'survives' free of flares until at least time t, and is given by [Frank E. Harrell, 2001]

$$S(t) = \mathrm{P}(T > t) = 1 - F(t), \tag{1}$$

where $F(t)$ is the probability of having a flare by time t. Note that $S(t) = 1$ at $t = 0$ (the probability of having a flare at some point after the study begins is one).

The hazard function, $\lambda(t)$, is related to the probability that a flare will occur in some small interval around t, and is given by [Frank E. Harrell, 2001]

$$\begin{aligned}
\lambda(t) &= \lim_{u \to 0} \frac{\mathrm{P}\{t < T \leq t + u | T > t\}}{u} \\
&= \lim_{u \to 0} \frac{\mathrm{P}\{t < T \leq t + u\}/\mathrm{P}\{T > t\}}{u} \\
&= \lim_{u \to 0} \frac{[F(t+u) - F(t)]}{u} \cdot \frac{1}{S(t)} \\
&= \frac{\partial F(t)/\partial t}{S(t)} \\
&= \frac{f(t)}{S(t)}, \qquad (2)
\end{aligned}$$

where $f(t)$ is the probability density function of T evaluated at t.

Hazard functions are an easy way to understand how the rate of failure changes over time and across groups. The hazard function is also known as the instantaneous failure rate, which is intuitive from equation 2 because the hazard function can be written as the probability of surviving at time t, $f(t)$, divided by the probability of surviving to time t or greater, $S(t)$. Note that as the proportion of censored data increases, the survival function, $S(t)$, decreases and thus the hazard rate, $\lambda(t)$ increases. As uncertainty about survival increases, it is more likely that a patient will experience an instantaneous failure.

3 Nonparametric Analysis: Kaplan-Meier Curves

3.1 Motivation for Kaplan-Meier Curves

In comparing two populations with unknown distributions, we first start by assuming no specified underlying distribution for either sample or population. The lack of parametric assumptions puts our analyses into a class of nonparametric procedures. Nonparametric analyses have a number of advantages over parametric analyses. First, nonparametric tests do not require as many assumptions as parametric tests. There are no assumptions of normal distributions or otherwise specified underlying distributions for the overall population. For example, when comparing two samples, the standard deviations donot have to be the same or be within a specified distance of each other. Secondly, nonparametric tests are often simpler and more intuitive than parametric

tests. Distribution-free analyses are more widely applicable in situations where there is uncertainty in the accuracy of the distribution in question [Dallal, 2004]. However, the price we pay for the easy application of nonparametric analyses comes as a loss of power to reject the null hypothesis even when the null hypothesis is not true.

3.2 Description of Kaplan-Meier Curves

The Kaplan-Meier estimate of the density curve is based on the idea that taking smaller and smaller intervals between observations provides more complete data, and thus as an ideal estimate one could take the limit as the interval size becomes arbitrarily small [Fisher and vanBelle, 1993]. The Kaplan-Meier curve does not require all data to be uncensored. In this sense, a patient does not have to be 'removed' if he or she has to leave the study early, and a patient does not have to be removed if he or she survives the entire study without having a 'failure' during the observation period, i.e., if the data is censored. Without the ability to handle censored data, we would have to throw away information. Patients who did not experience flares during the observation period could not be included in the analyses. Excluding censored data, however, would bias the results away from a longer period between flares, as patients with longer average time between flares are more likely to contributed censored data than those with short average time between flares. In this sense, Kaplan-Meier curves are useful estimates for reliability/survival functions where there is censored data.

The survival curve is affected by censored data in the steepness of the step size but does not determine the point at which the value of the function changes. The survival curve does not take a step down when a patient leaves the study, or is censored, but rather the total number of patients with a potential of having a 'failure' during the next time interval decreases. Thus after a patient is censored, the next failure will result in a bigger step downwards, because one individual represents a larger proportion of the people remaining.

Censoring reduces the total sample size at each step, effectively reducing the reliability of the survival curve. At each point where a patient is censored, the reliability decreases from that point onwards. By the end of the curve, if there are a significant number of censored data points, the reliability has decreased substantially, which is unfortunate because the end of the curve is the most important, representing the long-run survival rate for a given group.

3.3 Derivation of Kaplan-Meier Curves

For t, the survival time of an experimental unit, and $F(t)$ the continuous empirical cumulative density function of T, we have $\hat{S}(t) = 1 - \hat{F}(t)$, the estimated survival function. Thus the estimated survival function is equivalent to one minus the estimated cumulative density function.

The empirical cumulative distribution function, $\hat{F}(x)$, is defined by

$$\hat{F}(x) = \text{proportion of observations in the random sample} \leq x. \tag{3}$$

Intuitively, $\hat{F}(x)$ makes sense because the fraction of the sample with survival times less than x represent the sample probability of the data being less than x, which closely mimics the idea of a cumulative density function. The empirical cumulative density function produces a nonparametric density estimate that tries to adapt itself

to the data, rather than producing a density with a particular underlying parametric distribution. The empirical cumulative density function simply assigns probability $\frac{1}{n}$ to each of the n observations in a sample. If the sample comes from a population with a known parametric family, then the empirical cumulative density function will closely resemble the cumulative density function of the known distrubution.

For the ordered data $X_{(1)} < \cdots < X_{(n)}$, the sample distribution function, $F(x)$ is defined by

$$F(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n} & X_{(i)} \leq x \leq X_{(i+1)}, \quad i = 1, \ldots, n \end{cases}$$

where $X_{(n+1)} = \infty$.

For a set of survival measurements with n flares, denote the ith measurement as t_i. Let $t_{(1)} \leq t_{(2)} \leq t_{(3)} \leq \ldots \leq t_{(n)}$ denote the ordered values including censored values. For $t_{(i)}$, the time of the ith event, we can find $S(t_{(i)})$, the probability that a patient will survive past the time the ith patient survives. To find $\hat{S}(t_{(i)})$, use an iterative procedure for any value of $t_{(i)}$ that is not censored [Higgins, 2004]. We set $\hat{S}(0) = 1$, and for $t_{(1)}$, we write

$$\hat{S}(t_{(1)}) = \text{fraction of observations} > t_{(1)}. \tag{4}$$

This is intuitive because $S(t)$ is, by definition, the probability that a patient will survive until time t without experiencing a 'failure'. Thus the estimated probability that a patient will survive past the time the first patient experiences a failure will simply be the fraction of patients who have a survival time longer than time $t_{(1)}$.

For $t_{(i)} < t_{(j)}$, adjacent uncensored times to failure, note that

$$\begin{aligned} S(t_{(j)}) &= P(T > t_{(j)}) \\ &= P(T > t_{(i)})P(T > t_{(j)} | T > t_{(i)}) \\ &= S(t_{(i)})P(T > t_{(j)} | T > t_{(i)}). \end{aligned} \tag{5}$$

The probability $P(T > t_{(j)} | T > t_{(i)})$ can be estimated iteratively by computing the fraction of observations, censored data excluded, greater than $t_{(i)}$ that are also greater than $t_{(j)}$. Censored data points between $t_{(i)}$ and $t_{(j)}$ must be excluded because we have no method of determining whether those patients survived longer than $t_{(j)}$. From the above equation, we can write the estimated survival function, $\hat{S}(t_j)$, using an iterative method:

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(i)}) \cdot \frac{\text{number of observations} > t_{(j)}}{\text{number of observations} \geq t_{(i)}}. \tag{6}$$

Note that if t is censored and $t_{(i)} \leq t < t_{(j)}$, where $t_{(i)}$ and $t_{(j)}$ are adjacent uncensored times, then $\hat{S}(t) = \hat{S}(t_{(i)})$.

3.4 Comparing Kaplan-Meier Curves

By the proportional hazards model assumption, if the survival curves for two groups are essentially the same, we would expect the number of flares for one group over any given interval to be proportional to the number of flares in the other group. The proportionality constant is based on the number of people at risk of having a flare in

each group. Thus if the two curves are significantly different, then they would not be proportional, but rather would have other factors influencing the survival rates.

There are two basic tests used to compare Kaplan-Meier curves: the log-rank test and the Wilcoxon test. Both sum the absolute differences between the expected number of failures (flares) and the actual number of failures (flares) at time $t_{(j)}$, for every time j. The log-rank is suitable for comparing two survivor functions when the null hypothesis is that the two Kaplan-Meier curves are the same, and the alternative hypothesis is that the hazard rate at any given time for an individual in one group is proportional to the hazard at the same time for a similar individual in the other group [Collett, 1994]. Thus the null and alternative hypotheses are as follows:

$$\begin{aligned} H_o: &\quad h_{\mathbf{z}}(t) = h_o(t) \\ H_a: &\quad h_{\mathbf{z}}(t) = g(\mathbf{z})h_o(t) \end{aligned}$$

where $\mathbf{z} = x, y, \ldots$ is a vector of one or more explanatory variables believed to affect the variable of interest, and $g(\mathbf{z})$ does not equal one. Thus in the null hypothesis the vector $\mathbf{z}$ does not affect the hazard function; as $\mathbf{z}$ changes, the hazard function remains the same. In the alternative hypothesis, as $\mathbf{z}$ changes the hazard function is multiplied by a constant dependent on what the changes in the vector $\mathbf{z}$ are but independent of the time t. If the null hypothesis is rejected then we can assume that the survivor curves are significantly different for the two groups being compared. Proportional hazard rates give a sense of difference between the two groups of interest because proportional hazard rates cause the survival curves to diverge. The group with the higher hazard rate will have less remaining patients at each given point in time and will have a larger proportion of its patients failing at that time. Thus the number of surviving patients in the group with the higher hazard rate will drop to zero much faster than the other group.

If the proportional hazards assumption does not hold, the Wilcoxon test is more suitable. We can test the assumption of proportional hazards by looking at the estimated Kaplan-Meier survival curves. Although we do not know what the actual survival curves look like, we can use the sample curves as estimates. If the two estimated survival functions do not cross, then we can assume that the true survival curves have proportional hazard functions, and we use the log-rank test to determine difference [Collett, 1994]. From the Kaplan-Meier survival functions (see figures 2 and 3), it is fair to assume that waiting times between flares in lupus patients follow the proportional hazards assumption.

The log-rank test is derived by ordering the r distinct death times for each group, Group I and Group II, as $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. At time $t_{(j)}$, there are d_{1j} and d_{2j} individuals at risk for failure (flare) in Group I and II, respectively. Provided there are no two members in the group with the same failure time, d_{1j} and d_{2j} will either be zero, or one for a given time $t_{(j)}$. Let n_{1j} and n_{2j} represent the number of individuals at risk of failure (flare) in Group I and II, respectively, at time $t_{(j)}$. At time $t_{(j)}$ we have $d_j = d_{1j} + d_{2j}$ total failures out of $n_j = n_{1j} + n_{2j}$ remaining individuals (see table 2 [Collett, 1994]).

To evaluate the null hypothesis, fix the marginal values from the Totals row in table 2 and assume (under the null hypothesis) that survival is independent of group membership. Under this assumption, both the number of failures (flares) for Group

Group	Number of deaths at $t_{(j)}$	Number surviving beyond $t_{(j)}$	Number at risk just before $t_{(j)}$
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Table 2: Number of deaths at the jth failure time in each of two groups

I and Group II at time $t_{(j)}$ and the number of individuals who survive beyond time $t_{(j)}$ in Group I and Group II, can be determined from the value of d_{1j} alone. Thus we only need to consider the value for d_{1j} and can determine the values for d_{2j} and the remainder of table 2 from the value for $d1j$.

Regard the value of d_{1j} as a random variable, D_{1j}, which can take any value from zero to the minimum of d_j and n_{1j}. Then we know D_{1j} follows the hypergeometric distribution [Collett, 1994], where the probability that the number of failures (flares) in Group I takes the value of d_{1j} is

$$\mathrm{P}[D_{1j} = d_{1j}] = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}. \tag{7}$$

The mean of the hypergeometric random variable D_{1j} is given by

$$E[D_{1j}] = e_{1j} = n_{1j} \cdot \frac{d_j}{n_j}, \tag{8}$$

where e_{1j} is the expected number of individuals who have a failure (flare) at time $t_{(j)}$ in Group I. [Note that the expected value is appealing because it is intuitive. Under the null hypothesis, the probability of a failure at time $t_{(j)}$ does not depend on which group the individual belongs to because the hazard rates are the same for all times t. Thus the probability of failure (flare) at time $t_{(j)}$ is simply the number of individuals at risk of failure divided by the total number of individuals, $\frac{d_j}{n_j}$. The number of individuals expected to fail in Group I is just the probability of failure for an individual (regardless of group), multiplied by the number of individuals in Group I, n_{1j}.]

In order to calculate the overall deviation between the actual data and the expected data, simply sum the differences $d_{1j} - e_{1j}$ over the total number of failures for each of the two groups. Thus the statistic of interest becomes

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}) = \sum_{j=1}^{r} d_{1j} - \sum_{j=1}^{r} e_{1j}, \tag{9}$$

with $E[U_L] = 0$, since $E[(D_{1j})] = e_{1j}$. Note that U_L depends solely on d_{1j} and does not require d_{2j} to be included in its formulation because d_{2j} can be written in terms of d_{1j}. In other words, with the knowledge of the data in Group I, the data in Group II can be determined using table 2. Under the assumption that death times are independent, the variance of U_L is just the sum of the variances of d_{1j}, represented by

$$\mathrm{var}(D_{1j}) = v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}, \tag{10}$$

and the variance of the statistic U_L is

$$\text{var}(U_L) = \sum_{j=1}^{r} v_{1j} = V_L. \tag{11}$$

Furthermore, it can be shown that under H_o U_L has an approximate normal distribution when the number of death times is large [Collett, 1994]. It follows that

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1). \tag{12}$$

Because the square of a normal distribution is distributed as chi-squared, we have

$$\frac{U_L^2}{V_L} \sim \chi_1^2. \tag{13}$$

And thus under H_o we can use the chi-squared tables to determine the probability of our observed data.

3.5 Results from Comparing Kaplan-Meier Curves

Assuming no specified underlying distribution, we can use Kaplan-Meier curves to test for a difference in the survival curves for African-Americans versus Caucasians and for renal versus non-renal lupus patients. In testing for a difference between races, there were a total of twenty-six flares between the two groups, of which fourteen were from the African-American group, and twelve were from the Caucasian group (see table 3).

Group	Number of Flares	Number Censored	Mean Days (Biased)	Std Error
African-American	14	42	356.68	17.605
Caucasian	12	59	619.69	32.439
Combined	26	101	584.81	26.198

Table 3: Survival Data Summary for African-Americans versus Caucasians

In testing the null hypothesis that the hazard rates for both African-Americans (Group I) and Caucasians (Group II) are the same, versus the alternative hypothesis that the hazard rates for African-Americans and Caucasians are proportional, the p-value for the log-rank test is 0.1789 (chi-squared value of 1.8064). A p-value greater than 0.05 indicates that there is no significant difference between the survival curves for African-Americans and Caucasians (see figure 2). Note that in figure 2 a "1" represents African-Americans and a "2" represents Caucasians. Thus we fail to reject the null hypothesis that there is no difference between the two groups. Therefore we conclude that that the hazard functions for African-Americans and Caucasians are not significantly different.

In testing the difference in survival curves between renal and non-renal lupus patients, there were twenty-six total flares, of which twenty-four were associated with renal lupus patients and only two were associated with non-renal patients (see table 4).

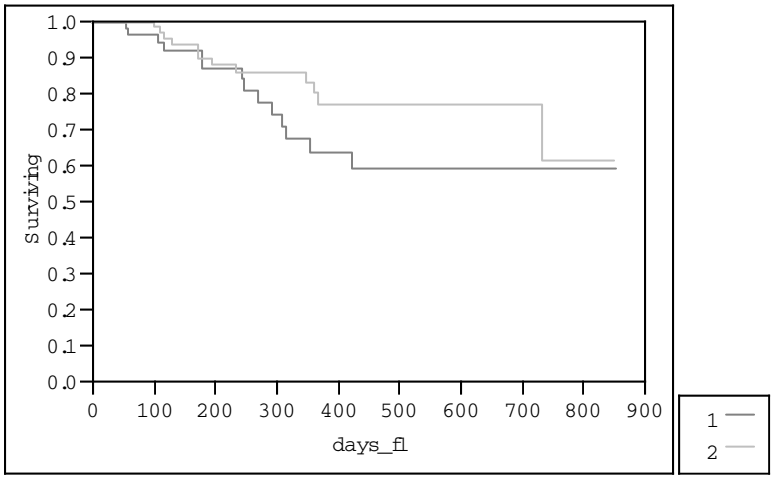

Figure 2: Kaplan-Meier Survival Curves Comparing Race

Group	Number of Flares	Number Censored	Mean Days	Std Error
Renal	24	67	544.14	32.765
Non-Renal	2	34	193.59	3.360
Combined	26	101	584.81	26.198

Table 4: Survival Data Summary for Renal versus Non-Renal Lupus Patients

The log-rank test to compare the Kaplan-Meier curves for renal and non-renal lupus patients tests the null hypothesis that the hazard rates for renal lupus patients (Group I) and non-renal lupus patients (Group II) are the same versus the alternative hypothesis that the hazard rates are proportional. A p-value of 0.0124 indicates that at a significance level of 0.05 the survival curves are significantly different (see figure 3). Note in figure 3 that a "1" represents renal lupus patients, and a "2" represents non-renal lupus patients. The null hypothesis is rejected in favor of the alternative hyptohesis. Thus we conclude that the hazard functions for renal and non-renal lupus patients are proportional.

Because Kaplan-Meier curves are a nonparametric method of analyzing data, there are few assumptions associated with them. Thus Kaplan-Meier curves are a useful tool for the initial investigation of a dataset. The lack of distribution, however, prevents us from coming up with estimates or confidence intervals, and therefore we cannot make any statements regarding the average number of patients who survive to a given point in time in one group versus the average number of patients surviving until the same time in a second group.

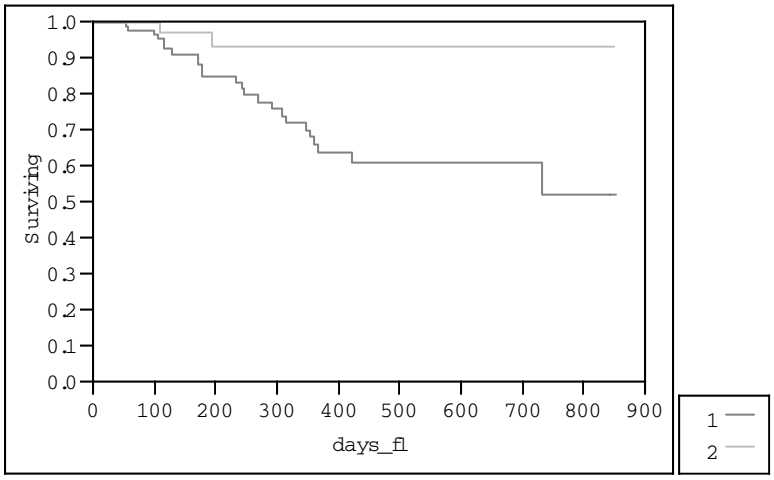

Figure 3: Kaplan-Meier Survival Curves to Compare Type of Lupus

4 Parametric Analysis: the Kolmogorov-Smirnov Goodness-of-Fit Test

4.1 Motivation for Kolmogorov-Smirnov Test

The basic disadvantage of nonparametric analyses is the fact that they are distribution-free. Recall that the Kaplan-Meier survival curves are not useful for providing estimates, predictions, or for making confidence intervals for the average number of flares over a given time interval. In order to make further statements and conclusions about the differences between the waiting times between flares for African-American and Caucasian lupus patients and between renal and non-renal lupus patients, we need to use parametric analyses.

Without assuming an underlying distribution, there are no parameters with which to describe the data or to make estimates and other quantatative statements about the data. The Kolmogorov-Smirnov goodness-of-fit test provides a means with which to test the data against a number of specified distributions, including normal, exponential, lognormal, and gamma. If the data fit a specified distribution, it can be used to find confidence intervals and make predictions about the variable of interest.

Furthermore, nonparametric tests are less powerful tests than parametric tests. A less powerful test means the test has a weaker ability to find deviations from the null hypothesis, even when the null hypothesis is not true. It has been said that "the more assumptions you make, the less data you need." Therefore, in making assumptions you must be sure to justify the assumption that your data follow the assumptions.

4.2 Description of Kolomogorov-Smirnov Test

The Kolmogorov-Smirnov goodness-of-fit test compares the observed empirical cumulative distribution function with the cumulative distribution function expected under the null hypothesis. The Kolmogorov-Smirnov test statistic, D, comes from the maximum difference in probability between the observed density and the expected density

across all x values. If D is too large then the null hypothesis that the two distributions are the same is rejected.

The Kolmogorov-Smirnov test does not need to be adjusted for different underlying cumulative probability distributions but can be used on data following any known or unknown distribution [e-Handbook of Statistical Methods, 2004]. Thus the conclusion is not affected by the actual underlying population distribution. Limitations of the Kolmogorov-Smirnov test include that it can only be applied to continuous distributions, it tends to be more sensitive at the center of the distribution than in the tails, and more seriously, that if that location, scale, and shape parameters are estimated from the data, the critical region of the Kolmogorov-Smirnov test (the area under the curve where the null hypothesis would be rejected) is no longer valid but must be determined by simulation [e-Handbook of Statistical Methods, 2004].

4.3 Derivation of Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov goodness-of-fit test for a single random variable tests the hypothesis that the cumulative density function for the observed data, $F(x)$, follows a specified distribution, $F_0(x)$. And thus we have

$$\begin{aligned} H_o: &\quad F(x) = F_o(x) \\ H_a: &\quad H_o \text{ is not true} \end{aligned}$$

where $F_o(x)$ is some known continuous distribution with known parameters. The test statistic uses the empirical cumulative distribution function with sample size n, as defined in section 3.3.

The Kolmogorov-Smirnov test statistic is based on the maximum of the absolute value of the difference between the empirical cumulative density function assuming the alternative hypothesis and the cumulative density function assuming the null hypothesis. Thus D_n is given by [Hollander and Wolfe, 1999]

$$D_n = \sup_{-\infty<x<\infty} |F(x) - F_o(x)|, \tag{14}$$

so,

$$D_n = \max_{1\le i\le n} \left\{ \max \left[\frac{i}{n} - F_o(x_{(i)}), F_o(x_{(i)}) - \frac{i-1}{n} \right] \right\}, \tag{15}$$

where $\dfrac{i}{n} = F(x_{(i)})$.

Thus D_n is close to zero when the null hypothesis is true, and is large when the alternative hypothesis is true [Hollander and Wolfe, 1999].

4.4 Results from Kolmogorov-Smirnov Tests

To determine the underlying distribution of the waiting times between flares, we examine the uncensored data only. Censored data can occur for a number of reasons including patients leaving the study (censored data after the last observed flare), the timing with which a patient enters the study (censored data before the first observed flare), and a long period of time between relapses (patients who do not have any observed flares). The implications of the manner in which censored data can bias results

is discussed in section 4.4.2. It is not reasonable to assume that the uncensored data and the censored data have the same underlying distribution, and as we are interested in the amount of time between flares, we are interested in the distribution underlying the uncensored data only.

4.4.1 The lognormal, normal, and exponential distributions

A histogram and boxplot of the waiting times between flares suggests that the data may follow a lognormal distribution (see figures 4 and 5).

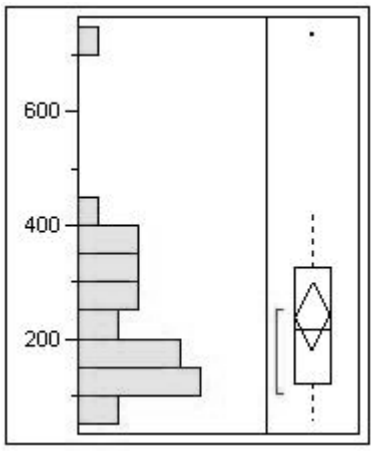

Figure 4: Waiting Times Between Flares, in days

We use the Kolmogorov-Smirnov goodness-of-fit test to determine whether the true times between flares does in fact follow a lognormal distribution. In testing the null hypothesis that the underlying population is distributed lognormally versus the alternative hypothesis that the population follows some other distribution, we have:

$$H_o: \quad F(x) = \text{ lognormal cumulative density function}$$
$$H_a: \quad H_o \text{ is not true}$$

The Kolmogorov-Smirnov goodness-of-fit test against a lognormal distribution yields a p-value of 0.1500. Thus we conclude that a lognormal distribution reasonably fits the data (see figure 5).

However, because the null hypothesis for the Kolmogorov-Smirnov test is that the data does fit the distribution in question, power is often too low to accurately reject the null hypothesis and therefore it is not uncommon for a given dataset to fit a number of distributions. The Kolmogorov-Smirnov test, like other hypothesis tests, will more accurately reject a specified distribution than prove that the distribution does in fact fit. Two common distributions for biological data are the normal distribution and the exponential distribution. Although the data does not look at all normal, it is important to ensure that the Kolmogorov-Smirnov test does in fact fail to reject the normal distribution as a potential true distribution (see figure 6).

From figure 6 it is apparent that the true distribution underlying the times between flares is not normally distributed. The goodness-of-fit test yields a p-value of 0.0049,

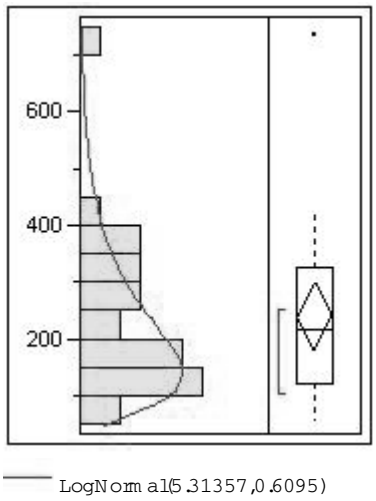

Figure 5: Waiting Times Between Flares with Lognormal Fit (p = 0.1500)

and thus the normal distribution is rejected. We conclude that the waiting times between flares is not distributed normally.

Similarly, from the histogram with the exponential model fit (7), it is intuitively obvious that the times between flares does not follow an exponential distribution (see figure 7).

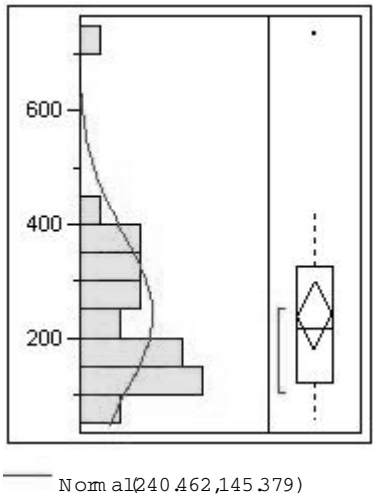

Figure 6: Waiting Times Between Flares with Normal Fit (p = 0.0049)

The fit for the exponential distribution is important to check statistically, because the exponential distribution has further implications. Unfortunately, the theories utilizing the memoryless property and other unique aspects of the exponential distribution cannot be utilized because the Kolmogorov-Smirnov test yields a p-value of 0.0100. Therefore, we reject the null hyptohesis that the waiting times between flares in lupus

patients follow an exponetial distribution, and conclude that the exponential distribution does not fit the data.

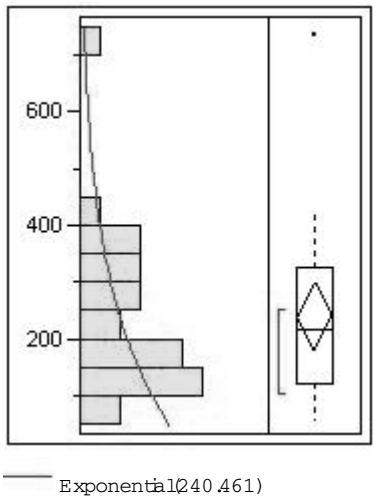

Figure 7: Waiting Times Between Flares with Exponential Fit (p = 0.0100)

4.4.2 Assumptions for Kolmogorov-Smirnov Tests

Only two assumptions must be met in order to use the Kolmogorov-Smirnov goodness-of-fit test: continuity and a distribution with completely specified parameters. Because the waiting time for the lupus study is measured in days, we consider it a continuous variable, and thus the first assumption is met. Since the Kolmogorov-Smirnov test is less sensitive in the tails of the distribution, and most of the data for the lupus study is in the tails of the distribution (the waiting time between flares appear to be distributed lognormally), the Kolmogorov-Smirnov test may not be completely accurate. The test statistic for the Kolmogorov-Smirnov test may suggest that the null hypothesis cannot be rejected at the 95 percent confidence level when it should be rejected, or alternatively the null hypothesis may be rejected with greatest confidence than the true test under known parameters would imply. Because the data for the number of days between flares is skewed and the vast majority of the flares occur in the interval between zero and four-hundred days since the last flare, the Kolmogorov-Smirnov test statistic may not be performing at the optimal level. Additionally, because the parameters of the distribution being tested are not well defined, i.e., the parameters are being estimated from the data, simulation should ideally be used in order to determine the correct confidence level with which to reject the distribution being tested.

By the nature of the data in a longitudinal study, the event of interest occurs repeatedly, and thus multiple observations can be contributed by a single patient. Some patients have multiple flares (uncensored data) and some have multiple censored observations, whereas other patients may only have a single censored observation and may not contribute any uncensored observations. Thus the number of data points (both uncensored and censored) for each group depends heavily on the parameters specific

to each unique individual. While censored data is not simple to handle statistically, excluding the censored data causes bias in the estimates and intervals for the mean waiting time between flares. Patients with naturally longer waiting times between flares are more likely to contribute large number of censored data points since it requires a longer period of time to observe a flare. Excluding censored data would bias the estimate for the average waiting time between flares away from the longer averages, and towards a shorter length of time.

Not only is the assumption of independence violated between groups because of the number of flares an individual contributes can vary, but also we cannot be sure the underlying distributions are identical from one individual to the next. In using the Kolmogorov-Smirnov test, we assume that the dependence has a relatively small overall effect, and that the parameters do not vary significantly within various races, types of lupus, and other factors that are not distinguished between including gender, age, socioeconomic class, et cetera. We cannot say for sure how the assumption of independence affects the overall results of the Kolmogorov-Smirnov test because we do not know how to describe the dependence relationship. However, it seems plausible that removing the dependence assumption would put less weight on multiple contributions from the same patient, and thus the high frequencies for shorter waiting times would decrease to some extent, thus flattening the overall distribution out to look less like a lognormal distribution and more like a normal distribution, causing the p-value for each distribution test to decrease, and the certainty of the results of the Kolmogorov-Smirnov test to decrease.

5 Poisson Processes

5.1 Motivation for Poisson Processes

Under the assumption that waiting times between flares are distributed exponentially, we have the ability to test for differences between the average waiting times between two populations. Namely, we can test for differences in the length of time between flares for African-American lupus patients versus Caucasian lupus patients and between renal and non-renal lupus patients. It might be intuitive to assume that the event of a flare occurring follows a Poisson process. Flares are rare events that occur with relatively low frequency, though the probabilities for such occurrences are relatively high. Lupus patients do not expect to experience a flare on a daily, weekly, or even monthly basis, but flares can occur at any time probabilistically speaking. Under the assumption that flares follow a Poisson processes, we assume that the distribution of inter-arrival times between flares is exponential with some rate λ, and thus, intuitively, the average rate of the occurrence of flares is $\frac{1}{\lambda}$. Poisson processes allow us to test for differences in the number of flares over a given period of time. For instance, we can test for differences in annual rates of flares within each group.

5.2 Description of Poisson Processes

A stochastic process $\{N(t), t \geq 0\}$ is considered a counting process if $N(t)$ represents the total number of "events" that occur by time t. Counting processes might include the total number of people who enter a store over some interval t, the number of children

born in a given hospital by some time t, or the number of home runs a baseball player hits by a given time t in the game [Ross, 2002]. In studying lupus, we consider a flare to be the "event" of interest. In studying flares in lupus patients, it is important that the counting process we are interested in has both independent and stationary increments. Independent simply means that the time of the next flare is not dependent on the total number of flares the patient has already had, and stationary means that the probability of having a flare over a given interval depends only on the length of the interval and not on what point in time the interval occurs.

By the independence and stationary properties of increments in a Poisson process, we know that the waiting time between flares is independent and identically distributed. From this we can compare the rates of occurrence of flares between groups. Thus we can use hypothesis testing to compare the overall rate of flares between African-Americans and Caucasians and between renal and non-renal patients.

5.3 Derivation of Poisson Processes

A counting process $\{N(t), t \geq 0\}$ is a *Poisson process with rate* λ, $\lambda > 0$ if the following hold:

(i) $N(0) = 0$.
(ii) The process has independent increments.
(iii) The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t \geq 0$

$$P\{N(t+s) - N(s) = n\} = e^{\lambda t}\frac{(\lambda t)^n}{n!}, \qquad n = 0, 1, \ldots \tag{16}$$

From the third condition, it follows that a Poisson process has stationary increments and that $E[N(t)] = \lambda t$. Thus, λ is the rate of the process.

For waiting time distributions, denote the time of the first event by T_1, such that for $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and the nth events. Then the sequence $T_n, n = 1, 2, \ldots$ is known as the sequence of inter-arrival times. To determine the distribution of T_n note that the event $T_1 > t$ occurs if and only if no events occur in the interval from time zero until time t. Thus,

$$\mathrm{P}\{T_1 > t\} = \mathrm{P}\{N(t) = 0\} = e^{\lambda t}, \tag{17}$$

and T_1 has an exponential distribution with mean $\frac{1}{\lambda}$. Because we know a Poisson process has independent and stationary increments, we know that

$$\begin{aligned} \mathrm{P}\{T_2 > t | T_1 = s\} &= \mathrm{P}\{0 \text{ events in} (s, s+t] | T_1 = s\} \\ &= P\{0 \text{ events in} (s, s+t]\} \\ &= e^{\lambda t}. \end{aligned} \tag{18}$$

Thus we can conclude that T_2 is an exponential random variable with mean $\frac{1}{\lambda}$, and furthermore, that T_2 is independent of T_1. By continuing the above argument using the properties of stationary and independent increments, in conjunction with the memorylessness property for exponentially distributed random variables, we have the proposition that the times between flares, $T_n, n = 1, 2, \ldots$, are independent and identically distributed exponential random variables with mean $\frac{1}{\lambda}$.

5.4 Comparing Rates: The Likelihood Ratio Test

5.4.1 Description of Likelihood Ratio Test

The likelihood ratio test is a goodness-of-fit test used to compare two hierarchical nested models, where the null hypothesis consists of specified values of the alternative hypothesis. The likelihood ratio test calculates the likelihood that the observed sample would occur under the null hypothesis as compared to the likelihood of the observed data under the alternative hypothesis. The likelihood ratio test statistic is most easily understood in the case of a discrete random variable with probability mass function $f(\mathrm{x}|\theta)$. The numerator of the ratio measures the maximum probability of the observed sample as computed under the parameters in the null hypothesis. The denominator of the likelihood ratio statistic measures the maximum probability of the observed sample over all possible parameters. Thus the likelihood ratio statistic calculates the number of times more likely the data is under the null hypothesis as compared to the alternative hypothesis. The likelihood ratio test statistic is large if the numerator is large relative to the denominator, i.e., if the specified null hypothesis provides parameters for which the data is extremely likely. Alternatively, the likelihood ratio test is small if the denominator is much larger than the numerator, i.e., if there exist some parameters in the alternative hypothesis space for which the observed data are far more likely than for any parameter in the null hypothesis space. It then follows that the null hypothesis is rejected for small likelihood ratio statistics.

5.4.2 Neyman-Pearson Lemma

The Neyman-Pearson theory is the "classical" hypothesis test. It circumvents the dependence of type I and type II errors by fixing type I error to be less than some pre-specified type I error rate, α. Once α is fixed, you look for the test statistic that maximizes the power of the test, $1-\beta$, and thus minimizes type II error, β. A test is considered most powerful for a simple null hyptohesis $\theta = \theta_o$ against a simple alternative hypothesis $\theta = \theta_1$ if the power of the test at $\theta = \theta_1$ is a maximum [Miller and Miller, 2004]. In order to create a test statistic that gives a test with the most power for a fixed α, use likelihoods. Denoting the null and alternative likelihoods by L_o and L_1, respectively, for a population of size n we have

$$L_o = \prod_{i=1}^{n} f(x_i;\theta_o) \text{ and } L_1 = \prod_{i=1}^{n} f(x_i;\theta_1).$$

Intuitively, it seems reasonable that $\frac{L_o}{L_1}$ would be small for points inside the critical region (where the alternative hypothesis is considered to be true) and would be small for points outside the critical region (where the null hypothesis is considered to be true). By the Neyman-Pearson Lemma, we are guaranteed a most powerful critical region [Miller and Miller, 2004].

Neyman-Pearson Lemma 5.1 *If C is a critical region of size α and k is a constant such that*

$$\frac{L_o}{L_1} \leq k \text{ inside } C$$

and

$$\frac{L_o}{L_1} \geq k \text{ outside } C$$

then C is a most powerful critical region of size α for testing $\theta = \theta_o$ against $\theta = \theta_1$.

Proof. The proof for the discrete case is similar to the proof for the continuous case, and thus only the continuous case will be presented here. Suppose that C is a critical region of size α satisfying the Neyman-Pearson Lemma and that D is another critical region of size α. Thus,

$$\int \cdots \int_C L_o dx = \int \cdots \int_D L_o dx = \alpha$$

where dx is notation short for $dx_1, dx_2, \ldots, dx_n$, and the multiple integrals are taken over the respective n-dimensional regions C and D. Because C can be written as the union of the disjoint sets $C \cap D$ and $C \cap D'$, and D is the union of the disjoint sets $C \cap D$ and $C' \cap D$, we can write

$$\int \cdots \int_{C \cap D} L_o dx + \int \cdots \int_{C \cap D'} L_o dx = \int \cdots \int_{C \cap D} L_o dx + \int \cdots \int_{C' \cap D} L_o dx = \alpha$$

and hence

$$\int \cdots \int_{C \cap D'} L_o dx = \int \cdots \int_{C' \cap D} L_o dx$$

Since $L_1 \geq L_o/k$ inside C and $L_1 \leq L_o/k$ outside C, it follows that

$$\int \cdots \int_{C \cap D'} L_1 dx \geq \int \cdots \int_{C \cap D'} \frac{L_o}{k} dx = \int \cdots \int_{C' \cap D} \frac{L_o}{k} dx \geq \int \cdots \int_{C' \cap D} L_1 dx$$

and thus

$$\int \cdots \int_{C \cap D'} L_1 dx \geq \int \cdots \int_{C' \cap D} L_1 dx$$

Also, we can write

$$\int \cdots \int_C L_1 dx = \int \cdots \int_{C \cap D} L_1 dx + \int \cdots \int_{C \cap D'} L_1 dx \geq \int \cdots \int_{C \cap D} L_1 dx + \int \ldots \int_{C' \cap D} L_1$$

$$= \int \ldots \int_D L_1 dx$$

and hence

$$\int \cdots \int_C L_1 dx \geq \int \cdots \int_D L_1 dx.$$

Thus the probability of committing a type II error in the critical region C is less than or equal to the corresponding probability for any other critical region of size α.

The likelihood ratio test follows immediately as an extension of the Neyman-Pearson Lemma for composite hypotheses. In extending the hypotheses to be more complicated statements that cover an entire space, the likelihood ratio test is not always the most powerful test and therefore does not have a formal proof, but rather is 'proved' through logical reasoning.

5.4.3 Derivation of Likelihood Ratio Test for composite hypotheses

If $X_1, X_2, \ldots, X_n$ is a random sample from a population with probability density function $f(x|\theta)$, where θ may be a vector, the likelihood function is defined as

$$L(\theta|x_1, \ldots, x_n) = L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{19}$$

For Θ, the entire parameter space, the likelihood ratio test is defined as follows

$$\lambda(\mathrm{x}) = \frac{\sup_{\Theta_o} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}. \tag{20}$$

All likelihood ratio tests must satisfy a rejection region of the form reject $\mathbf{x} : \lambda(\mathbf{x}) \leq c$, where c is a real number satisfying $0 \leq c \leq 1$. Thus we have $\mathrm{P}(\lambda(\mathbf{x} \leq c|H_o) = \alpha$. When $\sup_{\Theta_o} L(\theta|\mathbf{x})$ is small relative to $\sup_{\Theta} L(\theta|\mathbf{x})$, the likelihood ratio test statistic $\lambda(x)$ is close to zero, and thus we assume that some parameter in θ-space is significantly more likely than the null-space. When $\lambda(x)$ is close to zero, we reject the null hypothesis in favor of the alternative. Alternatively, when $\sup_{\Theta_o} L(\theta|\mathbf{x})$ is large relative to $\sup_{\Theta} L(\theta|\mathbf{x})$, the test statistic $\lambda(x)$ is close to one and we assume there is no θ parameter that yields a higher likelihood function than the null-space and thus we do not have enough evidence to reject the null hypothesis.

5.5 Results from the Likelihood Ratio Test

Under the assumption that flares in lupus patients are a Poisson process, likelihood ratio tests can be used to compare the annual rates of flares across groups where f is the poisson probability mass function. Rates are convenient not only because they are intuitive and easily understood by both statisticians and laymen, but also because the analysis of rates eliminates any question regarding the different start dates for observation (in traditional studies involving censored data, all subjects begin their observation period on the same day) and accounts for the various forms of censored data in a longitudinal study (the period before the first observed flare, after the last observed flare, and those patients who did not have any observed flares).

Upon initial observation (see table 5), the rates appear to be different between renal and non-renal lupus patients, but it is not evident whether or not there is a statistical difference in the average number of flares per year between African-American and Caucasian lupus patients.

Type/Race	African-American	Caucasian
Renal	0.8734	0.7019
	(28 flares, 11699 days)	(31 flares, 16118 days)
Non-Renal	0.3800	0.5351
	(5 flares, 4803 days)	(11 flares, 7503 days)

Table 5: Annual Rate of Flares (flares per year)

Using the likelihood ratio test to test for a difference in the annual rate of flares between African-Americans and Caucasians, a p-value of 0.6144 prevents us from rejecting the null hypothesis that there is no difference in the annual rate of flares between

African-Americans and Caucasians. Thus we conclude that there is no evidence of a significant difference between African-Americans and Caucasians in the average number of flares per year. For the hypothesis test that renal lupus patients and non-renal lupus patients have the same annual rate of flares, the p-value is 0.0696. Thus at a significance level of 0.05, we cannot reject the null hypothesis that the annual rates between renal and non-renal lupus patients are the same. Under the assumption that flares in lupus patients follow a Poisson process, neither of the two groups have statistically significantly different annual rates. However, the p-value for the difference between annual rates for renal and non-renal lupus patients is close to 0.05, indicating that further research should be done with a larger sample size in order to increase the power of the test.

Similar to the previously mentioned violations in assumptions, the assumptions of independence and identical distributions are violated. In addition, the average number of days between flares did not fit the exponential distribution, and therefore, we cannot assume that flares in lupus patients follow a Poisson process.

6 Discussion

Both nonparametric and parametric analyses were used to investigate the differences between the average number of days between flares for African-American versus Caucasian lupus patients and for renal versus non-renal lupus patients. In testing the null hypothesis that both groups have the same survival curve against the alternative hypothesis that the two groups have different survival curves, the nonparametric logrank test was applied to Kaplan-Meier curves to determine that there was not enough evidence to reject the null hypothesis between races, while the null hypothesis for type of lupus was rejected. Thus we conclude that the Kaplan-Meier survival curve for African-American lupus patients is not significantly different from the survival curve for Caucasians. However, we also conclude that the Kaplan-Meier survival curve for renal lupus patients is different from that of non-renal patients.

In fitting a parametric distribution for the average number of days between flares in lupus patients, we used the Kolmogorov-Smirnov goodness-of-fit test to determine that the lognormal distribution fits the uncensored data most closely. In addition, the normal and exponential distributions did not fit the uncensored data well. If the occurrence of a flare in lupus patients was assumed to be a Poisson process nonetheless, we can use the likelihood ratio test to form a hypothesis test comparing the annual rate of flares for each group of patients: African-Americans versus Caucasians and renal lupus patients versus non-renal patients. In testing the null hypothesis that both groups have the same annual rate against the alternative hypothesis that the two groups have different rates, there is not enough evidence to reject the null hypothesis for either set of groups. Thus despite the fact that p-value for the test between renal and non-renal lupus patients is close to 0.05, we conclude that the average annual rates are not significantly different.

The hypothesis tests do not all show significance. However, because the number of uncensored flares is so low, we are working with a very small sample size (see table 6). Small sample sizes result in hypothesis tests with low power (i.e., low ability to reject the null hypothesis even when the null hypothesis is not true).

Therefore, it is important to not only use the results of the hypothesis tests to make

	Renal	Non-Renal	Total
African-American	20	10	30
Caucasian	28	17	45
Total	48	27	75

Table 6: Number of Patients

conclusions about the waiting time between flares for different groups of people, but it is also important to look at the trends for each group being compared in order to motivate further studies with larger sample sizes. The trends in the data are evident from the histograms. Recall the distribution of the uncensored data for the waiting times between flares (see figure 8).

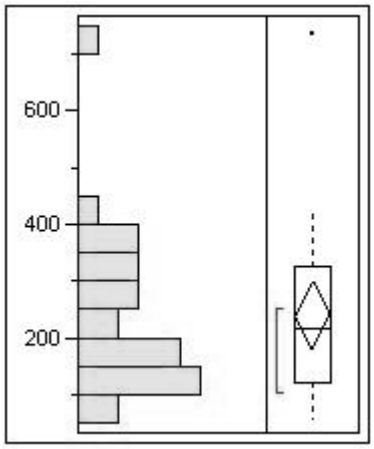

Figure 8: Waiting Times Between Flares, in days

We can section the histogram into respective groups by shading the African-American lupus patients dark and the Caucasian lupus patients a lighter color (see figure 9).

Thus it appears that the longer average waiting time between flares for Caucasian lupus patients may be caused by the one outline with an usually long period of time between flares. The Caucasian group did have more patients than the African-American group, so the apparent difference between races may simply be due to the small sample size of African-Americans which is unable to detect a difference between the two populations (see table 6). In conjuction with the additional patients, the total number of days of observation for the Caucasian patients was significantly larger than the total number of observation days for African-American patients (see table 7).

However, African-Americans may be less likely to have outliers with extremely long periods between flares because they have a shorter period between flares. As supporting evidence for the claim that the medical team at The Ohio State University did not begin recruiting patients of a specific race earlier than the other race, it is important to note that the longest observation period for an individual African-American patient was not substantially different from the longest observation period for Caucasian lupus patients

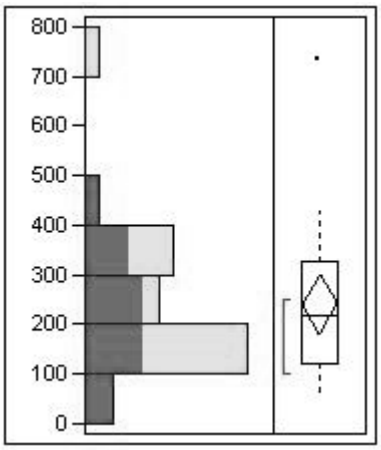

Figure 9: Waiting Times Between Flares, Grouped by Race

	Renal	Non-Renal	Total
African-American	7873	3964	11837
Caucasian	12570	6233	18803
Total	20443	10197	30640

Table 7: Total Number of Observation Days per Group

(see table 8).

If the histogram of waiting times between flares is grouped by type of lupus rather than the race of the patient, there are a few more potential problems. From figure 10 the most apparent problem in comparing the waiting time between flares for renal and non-renal lupus patients is the small number of uncensored flares for non-renal patients. [Note that in figure 10 renal lupus patients are the darker shade and non-renal lupus patients are the lighter shade.]

	Renal	Non-Renal
African-American	855	729
Caucasian	846	853

Table 8: Longest Individual Observation Period per Group

While there are fewer non-renal patients than renal patients (see table 6) and the total number of days of observation are less for non-renal patients than renal patients (see table 7), it still appears that non-renal patients have a significantly shorter waiting period between flares. Similar to the evidence that a specific race was not recruited earlier than the other, there is evidence that a specific type of lupus patient was not recruited earlier than the other type because the longest observation period for non-renal patients was similar to that for the renal patients (see table 8).

In terms of treating lupus patients, it would be useful to know if certain groups of patients were more susceptible to a higher frequency of flares than other groups

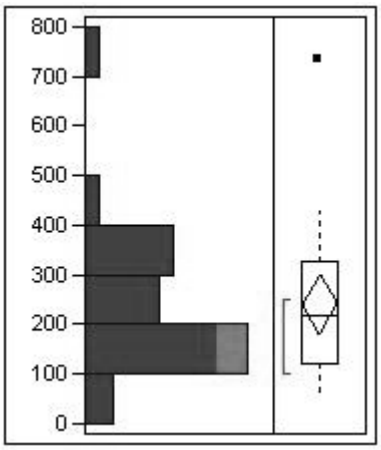

Figure 10: Waiting Times Between Flares, Grouped by Type

of patients. There is not enough evidence to support the hypothesis that African-American lupus patients experience a significantly different number of days between flares or a different annual rate of flares as compared to Caucasian lupus patients. Therefore, in practice race should not be a preliminary marker used to flag patients who may need special attention or care. However, there may be a difference in the average number of days between flares for renal versus non-renal lupus patients. Despite the fact that the distribution-free analysis found a significant difference between the hazard rates for type of lupus, the parametric analyses did not find the evidence as conclusive. Type of lupus does seem to be an important indicator as to the severity of symptoms for lupus patients. Doctors should note carefully when patients have renal lupus, as they may need more frequent check-ups and intervention to avoid severe flares.

References

[Collett, 1994] Collett, D. (1994). *Modelling Survival Data in Medical Research*. Chapman and Hall.

[Dallal, 2004] Dallal, G. E. (2004). Nonparametric statistics.

[e-Handbook of Statistical Methods, 2004] e-Handbook of Statistical Methods, N. (2004). 26 Sept 2004.

[Fisher and vanBelle, 1993] Fisher, L. D. and vanBelle, G. (1993). *Biostatistics: A Methodology for the Health Sciences*, chapter 16: Analysis of the Time to an Event: Survival Analysis. John Wiley and Sons, Inc.

[Frank E. Harrell, 2001] Frank E. Harrell, J. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

[Higgins, 2004] Higgins, J. J. (2004). *An Introduction to Modern Nonparametric Statistics*. Brooks/Cole.

[Hollander and Wolfe, 1999] Hollander, M. and Wolfe, D. (1999). *Nonparametric Statistical Methods*. Wiley and Sons, 2 edition.

[Miller and Miller, 2004] Miller, I. and Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications*. Pearson: Prentice Hall.

[Ross, 2002] Ross, S. (2002). *An Introduction to Probability Models*. Academic Press, Incorporated, 8 edition.