Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Luyu Zhang Date

Analysis of Time to Recurrent Pulmonary Exacerbation: A Review of Three Statistical Approaches

By

Luyu Zhang MPH

Department of Biostatistics and Bioinformatics

_____ [Chair's signature]

Jose Binongo, PhD
Committee Chair

_____ [Member's signature]

George Cotsonis, MS
Committee Member

Analysis of Time to Recurrent Pulmonary Exacerbation: A Review of Three Statistical Approaches

By

Luyu Zhang
Bachelor of Science Kansas State University 2012
Master of Science University Illinois at Urbana-Champion 2014

Thesis Committee Chair: Jose Binongo, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
in Master of Public Health in Biostatistics
2018

Abstract

Analysis of Time to Recurrent Pulmonary Exacerbation: A Review of Three Statistical Approaches

By Luyu Zhang

Recurrent data are not unheard of in biomedical studies. Individuals can experience multiple events of the same type during the study period. Three common statistical methods for recurrent outcomes are reviewed in this thesis. These methods are count regression, standard Cox proportional hazards (PH) regression, recurrent event survival regression. This study examines the mathematical underpinnings of these three methods; the limitations of each of the three approaches are also discussed. Two models, Poisson and negative binomial, are discussed in count regression. Both of them model the total number of interest events over the study period, but the negative binomial is a more flexible model than the Poisson model due to an additional parameter. The standard Cox PH model analyzes the time to the first event and ignores repeated events. Finally, the proportional intensity model, Prentice, Williams, and Peterson (PWP) total time model and PWP gap time model are discussed in recurrent event survival regression. They all account for the repeated events, but the two PWP models are more robust for analyzing recurrent events. The example study that we use in this paper is for testing the effectiveness of high dose vitamin D on preventing consecutive re-hospitalizations because of pulmonary exacerbations, which often happen repeatedly in adults with cystic fibrosis. Nevertheless, we found, based on results of all methods, that the vitamin D treatment does not have significant effects on preventing adults with cystic fibrosis from pulmonary exacerbations.

Analysis of time to pulmonary exacerbation: a review of three statistical approaches for recurrent outcomes

By

Luyu Zhang

Bachelor of Science Kansas State University 2012
Master of Science University Illinois at Urbana-Champion 2014

Thesis Committee Chair: Jose Binongo, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
in Master of Public Health in Biostatistics
2018

Analysis of time to pulmonary exacerbation: a review of three statistical approaches for recurrent outcomes

Table of Contents

Chapter 1 Introduction

Many models and analyses presume that individuals experience only one event, which is reasonable if the event is unrepeatable, like death. However, in many biomedical and epidemiological studies, an individual can experience repeated events during a given period. Some examples of recurrent events are women's menstrual cycle, sickness leave from work, heart attack, stroke, postoperative infection (Cumming, Kelsey, & C.Nevitt, 1990; Peduzzi, Henderson, Hartigan, & Lavori, 2002). It is very important to understand that whether there is an association between risk factors and recurrent events during the study period. (Twisk, Smidt, & de Vente, 2005) have divided statistical techniques for recurrent event data into two groups, naïve techniques and longitudinal techniques. Naive techniques ignore the existence of recurrent events or the correlation between the recurrent events within subjects. Different from naïve techniques, longitudinal techniques are characterized by taking into account whole pattern of recurrent events and their correlation within subjects. Many papers in the statistical literatures have talked about different methods of modeling recurrent events (LIM, 2000; Twisk et al., 2005; Yang et al., 2017). However, there are challenges to choose proper methods to address specific research questions. This thesis has two aims: (1) to give an overview of common statistical techniques to handle recurrent event data and (2) to provide recommendations on how to analyze recurrent event data based on specific research questions.

Three different statistical approaches and their assumptions and limitations are discussed and compared in this thesis. In chapter 2, count regression is discussed, including the Poisson regression and the negative binomial regression. Chapter 3 focuses on the Cox Proportional Hazards (PH) model, which analyzes time to the first event. In chapter 4, three

extended Cox approaches for recurrent event survival regression are illustrated. They are the Andersen-Gill (AG) model, Prentice, Williams, and Peterson (PWP) total time model, and the PWP gap time model. Chapter 5 summarizes the results of three statistical approaches using data from a recent Vitamin D study (DISC) for enhancing the immune system among patients with cystic fibrosis. The study aimed to investigate the efficacy of high dose vitamin D on preventing pulmonary exacerbations in adults with cystic fibrosis (CF). Chapter 6 provides recommendations in regard to conducting analysis of recurrent event data.

Chapter 2 Count Regression

Poisson Regression

The Poisson regression model has been widely used for modeling recurrent data in survival analysis. One of the examples investigated the number of hospitalizations for patients treated with hemodialysis (HD) verses peritoneal dialysis (PD) (Habach, Bloembergen, Mauger, & Wolfe, 1995). Poisson regression was used to estimate the rate ratio comparing the annual incidence rate of hospitalization between PD and HD. They concluded that hospital admission rates were 14% higher for those who treated with PD compared with HD patients (95% confidence interval [CI], 1.13 to 1.15).

In the Poisson regression model, counts of the events are assumed to follow a Poisson distribution in a given length of time, where the probability mass function is

$$P(Y_i = y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

where λ denotes the average number of events in a given interval and also its variance, y_i can take positive integer values in the interval. The likelihood function for the Poisson model is:

$$L(\beta|y) = \prod_{i=1}^{N} P(y_i|\lambda_i) = \prod_{i=1}^{N} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

Since the events are assumed to be independent, the likelihood function is equal to the product of the probability mass functions. λ is the only parameter in Poisson distribution, which is both its mean and variance. In Poisson regression, the logarithm of the expected value of Y for subject i is modelled by a linear combination with unknown parameters.

Poisson regression is thus also called "log-linear model". When individuals are not followed in the same amount of time, an offset variable is needed:

$$\log\big(E(Y_i)\big) = \log t_i + \beta' x_i$$

where β denotes a vector of regression coefficients, x_i denotes a vector of covariates for subject i, $\log t_i$ is an offset variable for subject i.

In a few situations, the conditional variance is greater than the conditional mean. This is called overdispersion, and overdispersion is a problem when fitting a Poisson model because the Poisson distribution assumes that the mean is equal to the variance. In this case, negative binomial regression may be used, which allows a dispersion parameter α to be incorporated in the model. (Rodriguez, 2013).

Negative Binomial Regression

The negative binomial distribution may be used to model count data. Its probability mass function is given by

$$P(Y_i = y_i|\lambda, \alpha) = \frac{\Gamma(\alpha + y)}{y!\,\Gamma(\alpha)} \frac{\beta^\alpha \mu^y}{(\mu + \beta)^{\alpha+y}}$$

The distribution has two parameters: λ and α. λ is the mean or expected value of Y. α is the overdispersion parameter. When α is equal to 0, the negative binomial distribution reduces to the Poisson distribution. The likelihood function for the negative binomial is as follows:

$$L(\beta|y, X) = \prod_{i=1}^{N} p(y_i|x_i) = \prod_{i=1}^{N} \frac{\Gamma(\alpha + y)}{y!\,\Gamma(\alpha)} \frac{\beta^\alpha \mu^y}{(\mu + \beta)^{\alpha+y}}$$

Since the variance of a negative binomial distribution is greater than or equal to the variance of a Poisson distribution with the same mean, the negative binomial regression has greater

flexibility in modeling mean and variance of the outcomes than the highly restricted Poisson regression (Pedan).

Chapter 3 Cox Proportional Hazards Regression

The Cox PH model is a popular technique for analyzing survival data to explain the effect of explanatory covariates on hazard rate in epidemiological and biomedical studies. Using the standard Cox PH model, (Abadi et al., 2014) assessed the impact of different treatments for patients with diverse stages of cancer survival time. Here, survival time was defined as the time from the diagnosis of the disease to death or the end of follow-up. They found that when patients with stage I and II breast cancer, radiotherapy and chemotherapy treatments had the highest hazard; when patients with stage III and IV breast cancer, surgery treatment had the highest hazard.

The standard Cox PH model is a technique for modeling the hazard rate of outcome variable, which is the time to the first event. The hazard rate is an immediate risk of failure at a given period, and the function which describes how hazard rate changes over time is called the hazard function. The survival time of an individual is assumed to follow its hazard function, which can be expressed as follows for the ith individual:

$$h_i(t) = h(t, X_i) = h_0(t)\exp(X_i'\beta) = h_0(t)\exp\left(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right)$$

where $h_0(t)$ is an arbitrary and unspecified baseline hazard function, $X_i = (X_{i1}, X_{i2} \ldots, X_{ip})$ denotes the vector of explanatory variables for subject i, and the Xs are time-independent predictors, which means the values of these predictors for a given person do not change over time. $\beta = (\beta_1, \beta_2 \ldots, \beta_p)$ is the vector of unknown estimators of explanatory variables, and this vector is assumed to be the identical for all individuals. Moreover, survival times are recorded for individuals who experience an event before the end of follow-up time, and those who do not experience the event or drop from the study during follow-up time are considered as being censored. Therefore, an individual either experiences the event of

interest or is censored in the study. The Cox model uses survival time and censorship as a two-variable outcome.

The likelihood function of the Cox model is a partial likelihood function rather than a complete likelihood function because only probabilities for individuals who fail are considered, ignoring probabilities for individuals who are censored. The partial likelihood is as follows:

$$L(\beta) = L_1 . L_2 \dots L_3 = \prod_{j=1}^{k} L_j(\beta)$$

Where L_j denotes that given survival up to time j, the likelihood of falling at this time. A subset of individuals who are at risk at the jth failure time is called risk set, $R(t_{(f)})$. The risk set will get smaller when the failure time increases.

In addition, the Cox PH model requires that the ratio of two hazards remains constant over time; that the hazard ratio stays the same over time is called the proportional hazards (PH) assumption. For example, if the hazards cross, the PH assumption is violated, so a Cox PH model is not appropriate. There are many approaches for assessing the PH assumption. A popular approach is to use the Schoenfeld residuals. The main idea behind this test is that the Schoenfeld residuals for a particular covariate will not be related with survival time when the PH assumption holds for that covariate.

Chapter 4 Recurrent Event Survival Regression

In recurrent events analysis, an individual is at risk for the same event of interest throughout the study. Several statistical modeling techniques have been developed for the analysis of recurrent time-to-event data. In this thesis, the focus is on three modeling techniques including the Andersen-Gill (AG) model, the Prentice, Williams, and Peterson (PWP) total time model and the PWP gap time model.

The Andersen-Gill model

The AG model is a generalized the Cox PH model for analyzing independent increment in the number of events based on the theory of counting processes (Andersen & Gill, 1982). It also referred to as the proportional intensity model. It is based on the assumption that the outcome is the time from randomization (study entry) to the time the event occurs at time t, but whether previous occurred events or not is ignored. Recurrent events are assumed to be independent, and time increment between events are uncorrelated (Amorim & Cai, 2015; Ozga, Kieser, & Rauch, 2018). All events have a common baseline hazard function, and the same parameter is estimated for all predictors. However, if the assumption does not hold, robust estimation is widely used for adjusting the correlation among outcomes on the same subject. This technique adjusts the estimated variance of regression coefficients. Therefore, the robust variance estimator permits hypothesis testing (Liang & Zeger, 1986).

The Prentice, Williams, and Peterson total time model

The PWP total time is a stratified Cox based approach based on the each sequential event during the follow-up time (Prentice, Williams, & Peterson, 1981). Similar to the AG

model, the time scale is the time from study entry, but the PWP total time uses the actual times of the two events. Therefore, the baseline hazard function changes in the subsequent events. This model evaluates the effect of a covariate for a specific event since the study entry. The hazard function for PWP total time model is as follows:

$$h_{ij}(t) = h_{0j}(t)\exp\left(X'_{ij}\beta_j\right)$$

This equation describes the hazard for an individual $i, i = 1, \dots, n$, and for their jth recurrent event $j = 1, \dots, k_1, k_1 \leq k$. k is number of distinct observed event times. All individuals are at risk for the first stratum, but only those who had an event in the previous stratum are at risk for the next one. Therefore, the stratum has fewer events than its previous stratum, and the late few strata may contain very small number of events. In this situation, if their baseline hazards are very similar, some events can be combined to the same strata (Yang et al., 2017).

The PWP gap time model

The PWP gap time model is similar to the PWP total time model, but the time of PWP gap time model is reset to zero after each event has occurred (Prentice et al., 1981). The time starts from the previous event (or study entry for the first stratum) instead of study entry until the next event. The hazard function for PWP gap time model is as follows:

$$h_{ij}(t) = h_{0j}\left(t - t_{j-1}\right)\exp\left(X'_{ij}\beta_j\right)$$

where the hazard for individual $i\ i = 1, \dots, n$, and for the jth event $j = 1, \dots, k_1, k_1 < k$. . k is number of distinct observed event times. Similar to the PWP total time, it can be seen that each event has a separate hazard function with its own baseline hazard and regression parameter. Only individuals who had the previous event will be at risk for the next event (Ozga et al., 2018).

All of these three recurrent survival models are commonly used in recurrent event survival analysis. Yang et al. (2017) applied all three models for recurrent event analysis in a cohort study of Chronic Kidney Disease (CKD). This study was interested in hospitalizations due to a cardiovascular event, which occurred repeatedly in patients with CKD. They concluded that recurrent event analyses provided more flexibility and insight, compared to the standard Cox PH model.

Chapter 5 The DISC Study

Overview of Study Design

Vitamin D deficiency is a common problem among patients with CF. Since current studies show that there is an association between vitamin D status and risk of pulmonary exacerbations, the purpose of the DISC study was to determine whether high dose vitamin D would improve the time to the next pulmonary exacerbation requiring re-hospitalization. This study was a double-blind, randomized, multicenter clinical trial of adults with CF. This study was conducted at five United State Cystic Fibrosis Foundation Care Centers including Emory University Hospital, Atlanta, GA, The University of Alabama Hospital at Birmingham, Birmingham, AL, Case Western Reserve University and Rainbow Babies and Children's Hospital, Cleveland, OH, University of Iowa and University of Iowa Hospitals and Clinic, Iowa City, IA, and the University of Cincinnati, Cincinnati, OH. 91 Adults with CF at these centers were randomized to either treatment and placebo group. The treatment group was given 250,000 IU vitamin D within 72 h of admission and re-dosed 50,000 IU of vitamin every other week from 3 months after randomization. Subjects were followed for one year. The clinical outcome was retrieved from the electronic medical record including outpatient and re-hospitalization because of pulmonary exacerbation and adverse events (Tangpricha et al., 2017).

Data Description of the DISC Study

91 adults with CF were randomized to treatment and placebo groups. 46 subjects were assigned to vitamin D treatment group, and 34 (74.0%) of them had the event of interest during the follow-up period. 45 subjects were assigned to the placebo group, and 39

(86.7%) of them experienced the event of interest during the follow-up period. Subjects were followed for 365 days, and the median follow up time was 122 days. Subjects who were lost to follow up were assigned the last day of interview as their event time, and those who did not have the event of interest at the end of the study were assigned 365 days as their event time. 20 out of 91 subjects (22.0%) never had the event of interest; 26 of the subjects (28.5%) had only one event of interest; 45 of the subjects (49.5%) had more than one event over the study period.

Moreover, 25 percentile of survival time is 77 days; 50 percentile of survival time is 127 days; 75 percentile of survival time is 275 days. In figure 1, there is no visible difference in the Kaplan Meier (KM) curves between the vitamin D group and the placebo group. As time passed, the two curves declined and both stayed together, suggesting that no beneficial effect of vitamin D over the placebo group.

All tests of hypotheses are two-sided and use a 0.05 level of significance. SAS® software 9.4 (SAS Institute Inc., Cary, NC) is used in the data analyses.

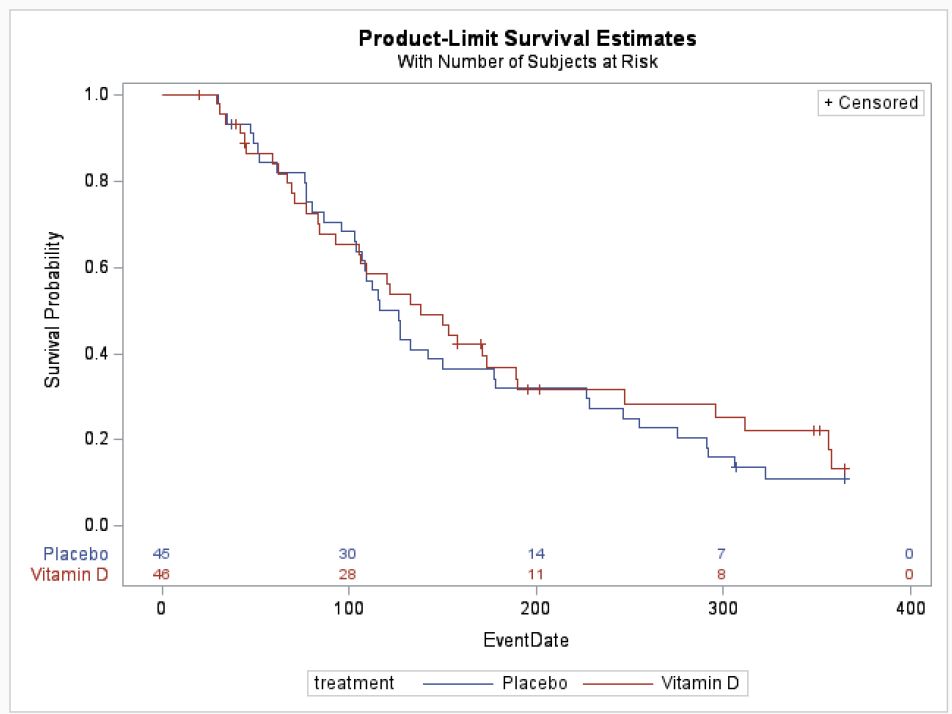

Figure 1: Kaplan Meier curves for Vitamin D (in red) and placebo (in blue) groups

Count regression analysis

A Poisson regression model was built to assess the vitamin D effect on recurrent pulmonary exacerbation hospitalizations. Since this was a multisite study, the interaction between treatment and site needed to be checked. None of the interaction termsturned out to be statistically significant. Therefore, the interaction terms were dropped, leaving only treatment in the model. From "Analysis of Parameter Estimates" table (figure 2), treatment was not statistically significant (p=0.95), which indicated that for the incidence rate of pulmonary exacerbation of CF was similar for both vitamin D and placebo groups. But because value/DF is higher than 1, there may be overdispersion in the data

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	89	175.6497	1.9736
Scaled Deviance	89	89.0000	1.0000
Pearson Chi-Square	89	912.8871	10.2572
Scaled Pearson X2	89	462.5510	5.1972
Log Likelihood		-35.3367	
Full Log Likelihood		-182.5184	
AIC (smaller is better)		369.0367	
AICC (smaller is better)		369.1731	
BIC (smaller is better)		374.0584	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.0443	0.1449	-5.3283	-4.7603	1211.92	<.0001
treatment	Vitamin D	1	-0.0125	0.2103	-0.4247	0.3996	0.00	0.9525
Scale		0	1.4048	0.0000	1.4048	1.4048		

Figure 2. Output from Poisson regression

Next, the negative binomial distribution was considered. As mentioned above, the negative binomial may provide a better fit to overdispersion data, as it allows an extra dispersion parameter. From "Analysis of Parameter Estimates" table (figure 3), treatment was not statistically significant (p=0.97). Because the 95% confidence interval for the dispersion parameter does not contain 0, the negative binomial distribution may be a better fit to the data than the Poisson distribution.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	89	127.1475	1.4286
Scaled Deviance	89	89.0000	1.0000
Pearson Chi-Square	89	828.0876	9.3044
Scaled Pearson X2	89	579.6403	6.5128
Log Likelihood		-45.7381	
Full Log Likelihood		-178.1207	
AIC (smaller is better)		362.2413	
AICC (smaller is better)		362.5172	
BIC (smaller is better)		369.7739	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.0275	0.1599	-5.3408	-4.7141	988.90	<.0001
treatment	Vitamin D	1	0.0077	0.2314	-0.4459	0.4612	0.00	0.9735
Dispersion		1	0.3058	0.1706	0.1024	0.9129		

Figure 3. Output of Negative Binomial Regression

Cox Proportional Hazards regression analysis

Results of the analysis of the Cox PH model are shown in figure 4. Based on the last table "Analysis of Maximum Likelihood Estimates," the hazard ratio between vitamin D group and placebo group is 0.87 (95% CI: 0.55-1.37), suggests no difference in the hazard rate between the two groups. It thus appears that there is no beneficial effect of giving high dose vitamin D for reducing the hazard of pulmonary exacerbation of CF. The proportional hazards (PH) assumption was also checked; the Schoenfeld residuals do not appear to be related to survival time (p-value=0.58), indicating no major violation of the PH assumption.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	551.413	551.037
AIC	551.413	553.037
SBC	551.413	555.328

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.3755	1	0.5400
Score	0.3752	1	0.5402
Wald	0.3735	1	0.5411

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
treatment	1	0.3735	0.5411

Analysis of Maximum Likelihood Estimates										
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
treatment	Vitamin D	1	-0.14370	0.23514	0.3735	0.5411	0.866	0.546	1.373	treatment Vitamin D

Figure 4. Output from the Cox PH model

The PWP total time analysis

The PWP total time model is chosen for analyzing recurrent events based on some criteria. First of all, the AG model may not be useful when the baseline risk changes from event to event. For example, people with CF who had the first pulmonary exacerbation may be more likely to have the second, the third or the fourth pulmonary exacerbations, so the assumption of independent events may not be satisfied. Second, if researchers are interested in study the time to rehospitalizations since study entry between vitamin D group and placebo group for the adult with CF, it is more meaningful to use the PWP total time model by comparing the time from study entry to events rather than interval time between events.

Results of the analysis of the PWP total time model are shown in figure 5. Results obtained from the PWP total time model and the standard Cox PH model for the time to

the first event are similar. The treatment coefficient is not significantly different from 0, which indicates a treatment difference in survival time. The hazard ratio is 0.99 (95% CI: ???). The conclusion is that there is no beneficial effect of high dose vitamin D on the total time in any of the pulmonary exacerbations and re-hospitalization.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1119.517	1119.508
AIC	1119.517	1121.508
SBC	1119.517	1124.723

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.0097	1	0.9217
Score (Model-Based)	0.0097	1	0.9217
Score (Sandwich)	0.0108	1	0.9173
Wald (Model-Based)	0.0097	1	0.9217
Wald (Sandwich)	0.0108	1	0.9173

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
tr	1	-0.01486	0.14313	0.947	0.0108	0.9173	0.985	0.744	1.304

Figure 5. Output from the PWP total time model

Chapter 6 Discussion and Conclusion

Three statistical approaches that are often used in recurrent data analysis are reviewed in this thesis. They are count regression, the standard Cox PH models and stratified Cox models. Depends on the type of data and interesting research questions, a proper test is selected for analysis. The Poisson regression and the negative binomial regression are fully parametric models, and the total number of count of events is assumed to follow either the Poisson distribution or negative binomial distribution. The Poisson regression assumes that conditional mean and conditional variance are identical, which can lead to the problem of overdispersion. In this case, negative binomial regression is recommended, as it introduces an additional parameter allowing overdispersion.

The standard Cox PH model is semiparametric, which has an unspecified baseline hazard function. It is used to examine the association between exposure and the time to the first event. This thesis also discusses three different stratified Cox PH models, which are usually related to the event mechanism. The AG model is rarely used in practice because it hardly assumes that the baseline risk stays constant as a result of previous events. The PWP total time and gap time models are more robust for recurrent events. They consider the event sequence in the analysis. The difference between the two PWP models is that the PWP total time model considers the time from study entry, but the PWP gap time model resets time to zero after each occurred event, so the time starts from the previous event the next event.

When applied to the DISC study, all of the statistical procedures discussed in this thesis suggest that the treatment variable does not have an impact on reducing the risk occurrence of pulmonary exacerbation.

Reference

Abadi, A., Yavari, P., Dehghani-Arani, M., Alavi-Majd, H., Ghasemi, E., Amanpour, F., & Bajdik, C. (2014). Cox Models Survival Analysis Based on Breast Cancer Treatments. *Iranian Journal of Cancer Prevention, 7*(3), 124-129.

Amorim, L. D., & Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol, 44*(1), 324-333. doi:10.1093/ije/dyu222

Andersen, P., & Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *Ann Stat, 10*, 1100–1120.

Cumming, R. G., Kelsey, J. L., & C.Nevitt, M. (1990). Methodologic issues in the study of frequent and recurrent health problems falls in the elderly. *Annals of Epidemiology, 1*(1), 49-56. doi:https://doi.org/10.1016/1047-2797(90)90018-N

Habach, G., Bloembergen, W., Mauger, E., & Wolfe, R. (1995). Hospitalization among United States dialysis patients: Hemodialysis versus peritoneal dialysis. *J Am Soc Nephrol, 5*, 1940–1948.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika, 73*(1), 13-22.

LIM, P. J. K. L. L.-Y. (2000). SURVIVAL ANALYSIS FOR RECURRENT EVENT DATA: AN APPLICATION TO CHILDHOOD INFECTIOUS DISEASES. *STATISTICS IN MEDICINE, Statist*(19), 13-33. doi:https://doi.org/10.1002/(SICI)1097-0258(20000115)19:1<13::AID-SIM279>3.0.CO;2-5

Ozga, A. K., Kieser, M., & Rauch, G. (2018). A systematic comparison of recurrent event models for application to composite endpoints. *BMC Med Res Methodol, 18*(1), 2. doi:10.1186/s12874-017-0462-x

Pedan, A. Analysis of Count Data Using the SAS® System *Statistics, Data Analysis, and Data Mining.* Retrieved from http://www2.sas.com/proceedings/sugi26/p247-26.pdf

Peduzzi, P., Henderson, W., Hartigan, P., & Lavori, P. (2002). Analysis of randomised controlled trials. *Epidemiol Rev, 24*, 26-38.

Prentice, R., Williams, B., & Peterson, A. (1981). On the regression analysis of multivariate failure time data. *Biometrika, 68*, 373-379.

Rodrıguez, G. (2013). Models for Count Data With Overdispersion. Retrieved from http://data.princeton.edu/wws509/notes/c4a.pdf

Tangpricha, V., Smith, E. M., Binongo, J., Judd, S. E., Ziegler, T. R., Walker, S., . . . Alvarez, J. A. (2017). The Vitamin D for Enhancing the Immune System in Cystic Fibrosis (DISC) trial: Rationale and design of a multi-center, double-blind, placebo-controlled trial of high dose bolus administration of vitamin D3 during acute pulmonary exacerbation of cystic fibrosis. *Contemp Clin Trials Commun, 6*, 39-45. doi:10.1016/j.conctc.2017.02.010

Twisk, J. W., Smidt, N., & de Vente, W. (2005). Applied analysis of recurrent events: a practical overview. *J Epidemiol Community Health, 59*(8), 706-710. doi:10.1136/jech.2004.030759

Yang, W., Jepson, C., Xie, D., Roy, J. A., Shou, H., Hsu, J. Y., . . . Chronic Renal Insufficiency Cohort Study, I. (2017). Statistical Methods for Recurrent Event

Analysis in Cohort Studies of CKD. *Clin J Am Soc Nephrol, 12*(12), 2066-2073. doi:10.2215/CJN.12841216

Appendix

```
libname x 'H:\Thesis';

* Read in the data;
PROC IMPORT OUT= WORK.DISC
            DATAFILE= "H:\Thesis\DISC_data_all.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="DISC_data_all$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

*cleaning data;

*data_all;
data x.data_all ;
      set work.disc;

      if EventDate > 365 and EventType = 'LTFU' then do
            EventDate = 365;
      end;
      if EventDate > 365 and EventType ne 'LTFU' then delete;

      *if EventType = "Death" then delete;

      if Group="A" then treatment="Vitamin D";
      else treatment="Placebo";

      keep ID site EventDate EventType Observed treatment;
      run;

*check that everything loaded correctly;
proc contents data=x.data_all;
run;
proc print data=x.data_all;
run;
proc freq data=x.data_all;
table treatment ;
run;

proc sort data=x.data_all;
      by ID;
      run;

*first_event;
data x.firstevent;
set x.data_all;
by ID;
if first.id;
proc print;run;
```

```
******************
Data description
******************;
proc freq data=x.firstevent;
table Observed * treatment/ missing NOPERCENT  NOROW  NOCOL ;
run;

proc freq data=x.DISC_poisson;
table number_event * treatment/ missing NOPERCENT  NOROW  NOCOL ;
run;

*mean follow-up time;
proc means data=x.firstevent median;
var EventDate;
run;

*median survival time;
 ods graphics on;
proc lifetest data=x.firstevent plots=(survival(atrisk));
      time EventDate*Observed(0);
      *strata Treatment;
   run;
   ods graphics off;

******************
Possion regression
******************;

*count number of events;
data x.DISC_poisson;
      set x.data_all;
      by ID;
      if first.ID then number_event= -1;
      number_event + 1;
      if last.ID then do;
      log_EventDate= log (EventDate);
            output;
            end;
      proc print; run;

proc sort data=x.DISC_poisson;
      by treatment;
run;

proc freq data=x.DISC_poisson;
      by treatment;
      tables number_event;
run;

/*check interaction;

proc genmod data=x.DISC_poisson;
      class treatment(ref='Placebo') site / param=ref;
      model number_event = treatment | site / offset= log_EventDate
                                                                  dist=
poisson
                                                  link= log dscale;
```

```
      run;*/

*remove site and interaction terms;
proc genmod data=x.DISC_poisson;
      class treatment(ref='Placebo') / param=ref;
      model number_event = treatment / dist= poisson
                                                    link= log dscale
                                                    offset=
log_EventDate;
      run;

data pvalue;
df = 89; chisq = 175.6497; pvalue = 1 - probchi(chisq,
df); run;
proc print noobs; run;

*****************************
Negative Binomial regression
*****************************;

/*proc genmod data=x.DISC_poisson;
      class treatment(ref='Placebo') site / param=ref;
      model number_event = treatment | site / offset= log_EventDate
                                                  dist= nb
                                                  link= log dscale;
      run;*/

*nb regression, treatment only;
proc genmod data=x.DISC_poisson;
      class treatment(ref='Placebo') / param=ref;
      model number_event = treatment / offset= log_EventDate
                                       dist= nb
                                       link= log dscale;
      run;

data pvalue;
df = 89; chisq = 127.1475; pvalue = 1 - probchi(chisq,
df); run;
proc print noobs; run;

*check zero-inflated;
ods graphics  / width=4in height=3in border=off;
proc sgplot data = x.DISC_poisson;
  histogram number_event /binwidth=1;
run;
ods graphics off;

*******************
   Cox PH Model
*******************;
*Check interaction term treatment*site;
proc lifetest data= x.firstevent method=km
                                          plots=survival
(strata=overlay);
      time EventDate*observed(0);
      strata treatment;
      run;
```

```
proc lifetest data= x.firstevent method=km
                                                plots=survival
(strata=overlay);
      time EventDate*observed(0);
      strata site;
      run;

*model 1 contains only treatment;
proc phreg data=x.firstevent;
      class treatment(ref='Placebo');
      model EventDate*observed(0)=treatment / rl;
run;

*model 2 contains treatment, site and interaction term;
proc phreg data=x.firstevent;
      class treatment(ref='Placebo')site;
      model EventDate*observed(0)=treatment | site/ rl;

run;


****Assessing the PH assumption;
*run model and output Schoenfeld residuals;
proc phreg data=x.firstevent;
      class treatment(ref='Placebo');
      model EventDate*observed(0)=treatment ;
      output out=results RESSCH=rtreatment;
      run;
proc print data=results;
      run;
data events;
      set results;
      if observed=1;
      run;

*create rank variable;
proc rank data=events out=ranked
ties=mean; var EventDate;
ranks timerank; run;
proc print data=ranked; run;

*correlate rank variable and Schoenfeld residuals;
proc corr data=ranked nosimple; var rtreatment;
with timerank;
run;

******************
  Recurrent Events
******************;
proc sort data=x.data_all;
      by ID eventdate;
      run;

data DISC2(drop = SITE EventType);
retain id interval Observed start stop tr;
```

```
rename EventDate=stop;
set x.data_all;
by id eventdate;
start=lag(eventdate);
if first.id then start=0;
if first.id then interval=1;
else interval+1;
if treatment = "Vitamin D" then tr=1;
else tr=0;
run;

proc print data = DISC2; run;

proc phreg data=DISC2 covs(aggregate);
model (start,stop)*observed(0) = tr / rl;
strata interval;
id id;
run;
```