Stockholms universitet

Geometric Rate Regression for Summarizing the Occurrence of Events

Vilma Härkönen

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:17
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2018:17**
http://www.math.su.se

Geometric Rate Regression for Summarizing the Occurrence of Events

Vilma Härkönen*

June 2018

Abstract

In this thesis we present geometric rates in survival analysis and two different types of regression models to estimate them: quantile regression and generalized linear models. With the latter we estimated the instantaneous geometric rate and the instantaneous geometric odds models. We used data from a Swedish prospective cohort study among patients at Intensive Care Units to fit an instantaneous geometric odds model to estimate the risk of death within different renal disease groups. From this we observed that the risk of death was at the highest in the beginning of the study. The risk of death is approximately the same for all of the patients with different renal diseases except for the patients with acute-on-chronic kidney disease who had the highest risk of death.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: harkonen.vilma@gmail.com. Supervisor: Ola Hössjer Disa Hansson.

Sammanfattning

I denna uppsats introducerar vi geometriska intensiteten för överlevnadsanalys och två olika typer av regressions modeller för att estimera dem. Det första tillvägagångssättet är att använda sig av kvantilregression. Ett annat sätt att estimera geometriska intensiteten är med hjälp av generaliserade lineära modeller. Med denna metod får vi momentana geometriska intensitet och momentana geometrisk odds-modellerna. Vi använder data från en svensk prospektiv kohortstudie bland patienter på akuten och anpassar en momentan geometrisk odds modell för att estimera risken att dö för patienter med olika njursjukdomar. Från detta såg vi att risken att dö är högst i början av studien. Risken att dö var ungefär samma för alla patientier med olika njursjukdomar förutom dem patienterna med akut-på-kronisk njursjukdom som hade den högsta risken att dö.

Contents

List of Figures

List of Tables

1 Introduction

In this paper we use data from a study at Karolinska Institutet, which is prospectively collected from the Swedish intensive care registry and other registries [13]. The study concluded 103 363 adult patients and we have randomly extracted 2000 of them. With predictors such as age and survival time, we fit an instantaneous geometric odds model and illustrate how the geometric rates change over time.

1.1 Background

The survival time is the time from the first observation to the event of interest. The probability for an individual to survive beyond time t is the survival function $S(t)$, which is often given as the proportion of the observations that have not failed at time t of the total amount of observations. The difference between analysis of survival data and other types of data is that the survival data does not include the time to failure for all observations. The individuals that do not fail before the study ends are called censored observations [12]. One application of this is for example studying time to death: some of the individuals included in the study will be alive when the study ends, so called censored observations. Survival functions are very popular to use in medical research when estimating time to some event.

The Kaplan-Meier method is a common non-parametric way of estimating the survival function. This method is based on calculating the probability of event A, the individual has not failed at time t, and event B, the individual will survive at time $t+1$. Therefore the multiplicative rule of probability is used to calculate the Kaplan-Meier curve; $P(A \cap B) = P(A)P(B|A)$. The advantage with the Kaplan-Meier method is that it only includes the times at which the failure occurs, not the time points in between [12].

Another popular way to model risk of failure/time to an event is the hazard function. This is the risk per unit of time that the event occurs at time t given that it has not occurred before that, for instance the risk rate to die at time t given that the individual is alive at time t. Therefore, the hazard function gives the instantaneous death rate for an individual that has survived to time t. The hazard function is defined as $h(t) = f(t)/S(t)$, where $f(t)$ is the density function of the life length [6].

1.1.1 Geometric rates

Geometric rates are often used in economic development and demography when calculating the compound annual growth rate of wealth and population, respectively. However, this method is not common in survival analysis. The following are some examples from studies where the geometric rate was used to calculate the growth rate.

The compound annual growth rate was used when studying the growth of online learning [2]. The research group studied the growth in students taking at least one online course in higher education institutes in the U.S between 2002 and 2006. For the largest

institutions the compounding annual growth rate was $(1\,387\,982/586\,122)^{1/4}-1 \approx 24.1\%$. This means that the number of students on average increased by 24.1% every year.

In another research report [14], the investigators studied the effect of financial inclusion programs on poor households in India in 2007-2012. The research included gender dimension and studied the difference between households represented by females and males. The compound annual growth rate was used when calculating the annual growth rate of income. The result was that the annual percentage change on income for females due to the effect of the financial inclusion programs was $(15\,023/10\,179)^{1/5}-1 \approx 8.10\%$. The corresponding percentage for males was $(19\,128/15\,271)^{1/5}-1 \approx 4.61\%$. The effect of the financial inclusion program was therefore almost twice as large for women than for men.

1.1.2 Incidence rates

The incidence rate is defined as the number of events divided by the person-time which is the sum of the observation times.

For example, in [3] incidence rates were used to estimate risk of heart failure in person-years. The incidence rate of heart failure was obtained by dividing the number of cases of incident heart failure by the sum of person-years of the individuals in the study that did not have heart failure previously. The incidence rate of heart failure was calculated for different age categories every five-year interval. The incidence rate for individuals of age 55-59 years was $4/2\,888.6 \approx 0.0014$ per person-year. The incidence rate for persons of age 90 or higher was $86/1\,813.5 \approx 0.0474$ per person-year. This implies that the risk of heart failure is approximately 34 times higher for the latter group than for the first.

The incidence rate was also used in another study [15] to estimate the risk of gastrointestinal stromal tumors. The incidence rates were age-adjusted which means that it was weighted according to the proportions of individuals in the respective age groups. The overall age-adjusted incidence rate between 1992-2000 was 0.68 per 100 000 person-years. The incidence rate for white people was 0.60 per 100 000 person-years which is a bit lower than the overall. For black people the incidence rate was higher than the overall, 1.16 per 100 000 person-years. The results suggest that the risk of gastrointestinal stromal tumor is approximately 93% higher for black persons than for white.

It is important to notice that the incidence rate and the geometric rate are not the same and should not be interpreted as such. This follows from the definition of these two concepts. For instance, for the time variabel $T \sim \exp(\lambda)$ the survival function is $S(t) = \exp(-\lambda t)$, from this it follows that the geometric rate will be $1 - \exp(-\lambda)$, whereas the incidence rate λ. [4]

1.2 Outline

In Section 2, we introduce the geometric rates in survival analysis and give an overview of the theory behind the geometric rates. We also discuss about the software used in this thesis. We present our data and state our research question in Section 3. In addition,

we fit several instantaneous geometric odds models to the data and choose the one with the best fit by the lowest AIC score. The instantaneous geometric rates for the different renal diseases are also illustrated in this section. In Section 4, we discuss our results and future work. The link programs to calculate the log and logit link functions and their derivatives are presented in Appendix A 6. Furthermore, the complete summaries of the instantaneous geometric odds models, obtained in Section 3, are shown in Appendix B 7.

2 Method

2.1 Geometric rates

In this section we introduce the theory of instantaneous geometric rates and how they can be estimated with generalized linear models. Moreover, we give a short summary of the theory behind the instantaneous geometric rates.

2.1.1 Generalized linear models

According to [1], generalized linear models are often preferable when the outcome variable is not normally distributed. These models are identified by three main components: a random component, a systematic component and a link function.

The random component specifies the outcome variable, Y, with independent observations, $(y_1, ..., y_N)$ and their distribution. The probability density function or mass function for the outcome from a distribution that belongs to the natural exponential family is

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i))$$

where $Q(\theta_i)$ is the natural parameter and the value of parameter θ_i varies for $i = 1, ..., N$.

The systematic component specifies the predictor variables in the model. The linear predictor can be written as

$$\eta_i = \sum_{j=0}^{k} \beta_j x_{ij},$$

where x_{ij} is the value of predictor j for observation i and normally x_{ij} is 1 for the intercept, β_0.

The link function describes the interaction between the linear predictor and the mean of the outcome variable. Let $\mu_i = \mathbb{E}(Y_i)$ be the mean of the outcome variable. The systematic component η_i is linked to the mean by $\eta_i = g(\mu_i)$. Hence

$$g(\mu_i) = \sum_{j=0}^{k} \beta_j x_{ij},$$

for $i = 1, ..., N$.

The canonical link is a link function that transforms the mean to the natural parameter. For some regression models the canonical link function is the logit function, given by

$$Q(\theta) = \log\left(\frac{\theta}{1-\theta}\right).$$

This is because the natural parameter is the log odds for a binary Bernoulli variable with expected value θ. This is the *logit* of θ. Regression models that use a canonical logit link are called logistic regression models.

When the outcome is binary, we also denote the mean of the outcome variable's distribution as $E(Y) = P(Y = 1) = \pi(x)$. The regression model for this type of outcome, the logistic regression model, is then given by

$$\pi(x) = \frac{\exp(\beta_0 + \sum_j \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_j \beta_j x_{ij})},$$

where $j = 1, ..., k$ and $i = 1, ..., N$. As mentioned above, the canonical link for this type of model is the logit link.

The systematic component can be generalized as

$$\eta_i = g(\mu_i) = \sum_j s_j(x_{ij}), \tag{1}$$

where $s_j(\cdot)$ is a smooth function of covariate j.

2.1.2 Restricted cubic splines

Restricted cubic splines are widely used in survival analysis to estimate the smooth functions in (1) [8]. The restricted cubic splines with k knots, $t_1 < ... < t_k$, are defined as:

$$C(u) = \beta_0 + \beta_1 u + \sum_{j=1}^{k-2} \theta_j C_j(u),$$

where $C_1(u)...C_{k-2}(u)$ are cubic terms. These are given by

$$C_j(u) = \max(0, u - t_j)^3 - \frac{\max(0, (u - t_{k-1})^3)(t_k - t_j)}{(t_k - t_{k-1})} + \frac{\max(0, (u - t_k)^3)(t_{k-1} - t_j)}{(t_k - t_{k-1})},$$

for $j = 1, ..., k - 2$.

The restricted cubic splines have continuous first and second derivatives. They are also linear in the tails for $u < t_1$ and $u > t_k$. The restricted cubic spline is a linear function with respect to the parameters $\beta_0, \beta_1, \theta_j$ [8].

The number of knots and their positions have to be defined. Usually for survival data the recommended number of knots is three to five. For a covariate with range 0 to 100, the common locations for the knots are at $\{5, 50, 95\}$, $\{5, 25, 75, 95\}$ and $\{5, 25, 50, 75, 95\}$ percentiles for 3,4 and 5 knots respectively. According to [8], using these quantiles is recommended because it makes the data analysis more objective and comparable.

2.1.3 Quantile regression

In a typical regression model we want to estimate rates of change in the mean of the response variable distribution. The idea of quantile regression is to fit regression curves to different quantiles of this distribution. Quantile regression describes more completely the relationship between predictor variables, especially when the variances are heterogeneous. This is when predictor variables cause a change in the mean as well as a change in the variance of the distribution of the response variable [5].

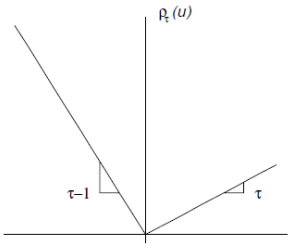

Figure 1: *Quantile regression ρ function.* Figure from [9] p.6

As stated in [9], the random variable, X, is described by its probability distribution $F(x) = P(X \leq x)$ and the pth quantile of X is given by

$$F^{-1}(p) = inf\{x : F(x) \geq p\}$$

for $0 < p < 1$. The median of the distribution of X is $F^{-1}(1/2)$. The quantiles are derived from the following optimization problem: assume that a loss function is given by

$$\rho_p(u) = u(p - I(u < 0))$$

for $0 < p < 1$, it is then of interest to find a value $\hat{x}$ that minimizes the loss. The loss function is shown in Figure 1. This implies that the following expectation is minimized

$$E[\rho_p(X - \hat{x})] = (p-1)\int_{-\infty}^{\hat{x}}(x - \hat{x})dF(x) + p\int_{\hat{x}}^{\infty}(x - \hat{x})dF(x). \tag{2}$$

After differentiating (2) with respect to $\hat{x}$, we have

$$0 = (1-p)\int_{-\infty}^{\hat{x}} dF(x) - p\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - p.$$

It is known that $F(x)$ is monotone and thus the solution is given by any element of $\{x : F(x) = p\}$. In order to obtain a unique solution put $\hat{x} = F^{-1}(p)$, otherwise the solution is the interval of the pth quantile.

The optimization problem gives rise to solutions for more general problems. The sample mean solves the following equation

$$\min_\mu \sum_{i=1}^{n}(y_i - \mu)^2.$$

When the conditional mean $\mathbb{E}(Y|x)$ is expressed as $\mu(x) = x'\beta$, the parameter β can be estimated with

$$\min_\beta \sum_{i=1}^{n}(y_i - x_i'\beta)^2,$$

where $x_i = (x_{i1}, \ldots, x_{ip})^T$. Analogously, the pth quantile, $\alpha(p)$ solves the minimization problem

$$\min_\alpha \sum_{i=1}^{n} \rho_p(y_i - \alpha).$$

Hence, the pth conditional quantile function is defined as $Q_y(p|x) = x'\beta(p)$ and in order to estimate $\beta(p)$ the quantile regression problem is formulated as

$$\min_\beta \sum_{i=1}^{n} \rho_p(y_i - x_i'\beta). \tag{3}$$

The solution to the quantile regression problem in (3) is referred to as the regression quantile $\hat{\beta}(p)$.

For instance, a linear regression model for a sample with one covariate and independent and identically distributed errors is

$$y_i = \beta_0 + x_i\beta_1 + u_i.$$

The quantile functions y_i are

$$Q_y(p|x) = \beta_0 + x_i\beta_1 + F_u^{-1}(p),$$

where F_u is the distribution function of the errors.

2.1.4 Quantile regression for geometric rates

As suggested in [4], let T be a continuous time variable on the positive real line. The geometric rate over the time interval $(0, t)$ is

$$g(0, t) = 1 - S(t)^{1/t} \tag{4}$$

where $S(t)$ is the survival function of T. We can interpret the geometric rate as the average probability of the event in interest per unit of time over the time interval $(0, t)$.

If the proportion of events that occurs in time $t \in (0, \infty)$ is $p \in (0, 1)$, then $p = P(T \leq t) = 1 - S(t)$. The geometric rate over the proportion interval $(0, p)$ is then

$$g(0, p) = 1 - (1 - p)^{1/Q(p)}$$

where $Q(p) = t$ is the quantile function of T. The interpretation of this is similar to the interpretation of (4), but instead of time interval we obtain the geometric rate as a function of the quantile p.

When p is fixed, the function $1 - (1 - p)^{1/t}$ is monotonically decreasing in t for $t > 0$. This implies that $g(0, p)$ is the $(1 - p)$ -quantile of the transformed time variable $T^* = 1 - (1 - p)^{1/T}$, i.e.

$$P(T^* \leq g(0, p)) = 1 - p. \tag{5}$$

Let $\beta_p \in R^k$ be a vector of regression coefficients and $x \in R^k$ a set of predictor variables. The conditional geometric rate is

$$g(0, p|x) = x'\beta_p.$$

By estimating the $(1-p)$-quantile of the transformed time variable $T_i^* = 1-(1-p)^{1/T_i}$ with ordinary quantile regression, we can estimate the regression coefficients, given the observations t_i (of T_i) and x_i, $i = 1, ..., n$. The components of the coefficient vector β_p can be interpreted as the change in the geometric rate for one-unit increase in the corresponding components of the predictor variables, x. This interpretation is analogous to interpretation of the regression coefficient for other regression methods.

It might be of interest to model the coefficient differences between two sets of covariates x_0 and x_1 or the rate ratios between the covariates. The first is easily obtained by

$$g(0, p|x_1) - g(0, p|x_0) = (x_1 - x_0)'\beta_p.$$

Suppose that $\log(g(0, p|x)) = x'\gamma_p$ for a vector $\gamma_p \in R^k$. Then the rate ratio is given by

$$\frac{g(0, p|x_1)}{g(0, p|x_0)} = \exp[(x_1 - x_0)'\gamma_p].$$

Similarly to the equation (5) the transformation of the time variable $\log[1-(1-p)^{1/t}]$ is monotonically decreasing in t for $t > 0$. The coefficient vector γ_p can therefore be estimated by estimating the $(1-p)$-quantile of $T_i^* = \log[1-(1-p)^{1/T_i}]$ for $i = 1, \ldots, n$. [4]

2.1.5 Instantaneous geometric rates

The geometric rate between two different times t_1 and t_2, $0 < t_1 < t_2 < +\infty$ is

$$g(t_1, t_2) = 1 - \left(\frac{S(t_2)}{S(t_1)}\right)^{\frac{1}{t_2 - t_1}}$$

where $S(t_1)$ and $S(t_2)$ are evaluations of the survival function at time points t_1 and t_2.

The instantaneous geometric rate is the geometric rate over decreasing time intervals $(t, t + \Delta t)$. This is given by

$$\begin{aligned} g(t) &= \lim_{\Delta t \to 0^+} \left[1 - \left(\frac{S(t+\Delta t)}{S(t)}\right)^{1/\Delta t}\right] \\ &= \lim_{\Delta t \to 0^+} \left[1 - \exp\left(\frac{\log S(t+\Delta t) - \log S(t)}{\Delta t}\right)\right] \\ &= 1 - \exp\left(\frac{\partial \log S(t)}{\partial t}\right) \\ &= 1 - \exp\left(-\frac{f(t)}{S(t)}\right) \\ &= 1 - \exp(-h(t)) \end{aligned} \tag{6}$$

where $f(t)$ is the probability function of T and $h(t) \equiv f(t)/S(t)$ the hazard function. The instantaneous geometric rate in (6) corresponds to the instantaneous probability of the event of interest per unit of time. [7]

2.1.6 Instantaneous geometric rates as GLM

As suggested in [7] the instantaneous geometric rate can be estimated via generalized linear models by using non-standard link functions. Two models are presented in [7]: the proportional instantaneous geometric rate model and the proportional instantaneous geometric odds model.

Let $t_i, i = 1, ..., n$ be n possibly censored observations of the time variable and d_i be the event indicator, where d_i takes value 0 for a censored observation and 1 for an event. Furthermore, let $x_i = (x_{1i}, ..., x_{qi})'$ be a vector of predictor variables and $\beta = (\beta_1, ..., \beta_q)'$ be an unknown parameter vector. The proportional instantaneous geometric rate model is then given by

$$g_i(t|x_i) = g_0(t) \exp(x_i'\beta). \tag{7}$$

Taking the logarithm of (7), the following is obtained

$$\log(g_i(t|x_i)) = \log(g_0(t)) + x_i'\beta. \tag{8}$$

By taking the logarithm of the instantaneous geometric rate in (6), we obtain

$$\log[1 - \exp(-h_i(t))|x_i] = s(t;\gamma) + x_i'\beta, \tag{9}$$

where $s(t;\gamma)$ is a smooth parametric function dependig on a vector of unknown parameters $\gamma = (\gamma_1, ..., \gamma_r)'$.

By dividing each individual's follow-up into intervals, the baseline log instantaneous geometric rate via smooth function $s(t;\gamma)$ can be modelled with cubic splines based on r knots. Let then t_{ij} be the length of the jth time interval, the time at risk, of the ith individual. In addition, let d_{ij} be the event indicator, where $d_{ij} = 1$ if the event occurs for individual i in interval j and $d_{ij} = 0$ otherwise.

According to [7], from equation (9) the following link function, $l(\cdot)$, is obtained

$$\eta_{ij} \equiv l(\mu_{ij}) = \log\left[1 - \exp\left(-\frac{\mu_{ij}}{t_{ij}}\right)\right], \tag{10}$$

where μ_{ij} is the expected value of d_{ij}, which is assumed to belong to a distribution from the exponential family. After some calculations the following derivatives of (10) are obtained

$$\begin{aligned} \mu &= l^{-1}(\eta) = -t\log(-\exp(\eta) + 1), \\ \partial\mu/\partial\eta &= t\exp(\eta)(-\exp(\eta) + 1)^{-1}, \\ \partial^2\mu/\partial\eta^2 &= t\exp(\eta)(\exp(\eta) - 1)^{-2}. \end{aligned} \tag{11}$$

Moreover, the proportional instantaneous geometric odds model is introduced in [7]. It is given by

$$\frac{g_i(t|x_i)}{1 - g_i(t|x_i)} = \frac{g_0(t)}{1 - g_0(t)} \exp(x_i'\beta). \tag{12}$$

This can be written as equation (9) by taking the logit of (6)

$$\text{logit}[1 - \exp(-h_i(t))|x_i] = s(t;\gamma) + x_i'\beta. \tag{13}$$

From (13), the second nonstandard link function is obtained

$$\eta_{ij} \equiv l(\mu_{ij}) = \text{logit}\left[1 - \exp\left(-\frac{\mu_{ij}}{t_{ij}}\right)\right]. \tag{14}$$

Similarly, after some calculations the following derivatives for the link function are obtained

$$\begin{array}{rcl} \mu & = & l^{-1}(\eta) = t\log(1+\exp(\eta)), \\ \partial\mu/\partial\eta & = & t\exp(\eta)(1+\exp(\eta))^{-1}, \\ \partial^2\mu/\partial\eta^2 & = & t\exp(\eta)(1+\exp(\eta))^{-2}. \end{array} \tag{15}$$

The difference between (7) and (12) is that in the first equation the estimated exponentiated coefficients can be interpreted as the instantaneous geometric rate ratios. In (12) the exponentiated coefficients can be interpreted as the instantaneous geometric odds ratios. [7]

2.2 Software

We perform our data analysis with Stata software that is commonly used in medical statistics. The build-in commands such as `stset`, `glm` and `predict` are used in the data analysis. In addition we use two user-written programs, `rcsgen` and `logit_igr`. The first one [10] is used to generate the restricted cubic splines for the survival time distribution. The latter, presented in Section 6, is used to calculate the logit link function and its derivatives for the instantaneous geometric odds model. The purpose of the program is to calculate the derivatives of the link function fast and effectively. The link function for the instantaneous geometric rate model is also presented in Section 6. The derivatives of the link functions are calculated in equations (11) and (15).

3 Data analysis

In this section we describe some applications of the instantaneous geometric odds model to estimate the risk of death. We used data from a cohort study [13] at Karolinska Institute. The purpose of that research was to study long-term mortality and end-stage renal disease incidence in patients with and without chronic kidney disease.

3.1 The research question

We want to estimate the risk of death within different renal disease groups with age, sex and observation time as covariates. Our main interest is to study if there are any differences in the odds of death between the groups and how the odds of death fluctuates with time.

3.2 Description of data

The data was prospectively collected from Swedish Intensive Care registry, SIR, and other Swedish national health registries such as the Swedish cause of death register and the national patient register. The Swedish renal register was used to collect data on individuals with end-stage renal disease before and after the Intensive Care Unit (ICU) admission. The observations are from January 1, 2005 to December, 31 2011 meaning that the total follow-up time is 7 years. The study included all first ICU admissions of patients older than 18 years. The researchers excluded patients that miss disease severity scores, intervention codes and diagnosis codes for acute kidney injury. The total amount of patients included in the study was 103 363. The Swedish personal identification number was used to identify the patients in the different registries. [13]

We have randomly extracted 2000 observations from this study. There are five variables in our data set where the binary outcome variable is death. The outcome variable follows a Poisson distribution.

The first three variables are age, sex, and the number of days the patients were observed until death or censoring. The covariate age is continuous and it obtains values within a range between 18 and 98 years. The binary predictor variable sex obtains value 1 when the patient is a female and 0 for a male. The maximum days of observation or censoring is 2 454, which is approximately 6 years and 9 months.

The fourth variable is a categorical variable with five levels. The different levels in this variable determine the disease of the patient. The first level is acute kidney injury, AKI, with 111 observations. The second level is only chronic kidney disease and it includes 53 observations. The third level has the smallest amount of observations, 22, and it includes patients with acute-on-chronic kidney disease, AOC. End-stage renal disease, ESRD, is the fourth level with 25 observations. The patients with no renal disease belong to the last level and it consists of 1789 observations. The fifth level is used as control and is the reference level of the categorical variable in our analysis.

3.3 Analysis and results

We start by preparing the data set for our analysis. First we declare the data set as survival data and set death as the failure event. Because we are interested in the risk of death yearly, not daily, we transform the survival times from days to years. We split the follow-up time of each patient into intervals with length of one week. Then we generate a new variable that is the time at risk within each interval. We proceed by generating the restricted cubic splines for the observed survival times with 4 knots. The knots are placed at the minimum, maximum, 33th percentile and 67th percentile of the survival times' distribution. This generates three new variables that we call RCS1, RCS2 and RCS3 in our analysis. These are the first, second and third quantiles of the survival time distribution, respectively.

We are interested in finding a suitable instantaneous geometric odds model to our data with generalized linear models and therefore we use the `logit_igr` link function, that was introduced in section 2.2.

The first model, Model 1, includes only the categorical variable `renalgroup` where the group with no renal disease is the reference level. The parameter estimates, confidence intervals and Akaike's Information Criterion (AIC) of the model are shown in Table 1. Complete summaries of the results are found in Section 7. As expected all of the coefficients for the different levels in the categorical variable are positive meaning that the odds of death is higher for the patients with a renal disease compared with those with no renal disease. The highest risk of death is for patients having acute-on-chronic kidney disease. The odds ratio for this group is $\exp(2.245) \approx 9.4$ which indicates that the odds of death is 9.4 times higher for patients with acute-on-chronic kidney disease than for patients with no renal disease. The AIC for this model is 11076.3.

Next, in Model 2, we add the restricted cubic spline terms RCS1, RCS2 and RCS3 to Model 1. The coefficient estimates for the restricted cubic spline terms are difficult to interpret. We notice that the coefficient estimates for the renal diseases, except for only chronic kidney disease, are slightly smaller in this model than in the first model; the effect they have on the outcome variable has decreased. As in the first model, all of the coefficient estimates are significant despite the ESRD. The second model gives a bit better fit to the data due to the smaller AIC.

	Model 1	Model 2	Model 3	Model 4	Model 5
AKI	1.188***	0.984***	0.991***	0.578**	0.573**
	[0.855,1.522]	[0.580,1.388]	[0.585,1.398]	[0.155,1.001]	[0.148,0.998]
Chronic only	1.337***	1.504***	1.511***	0.726*	0.718*
	[0.895,1.778]	[0.953,2.055]	[0.960,2.062]	[0.147,1.305]	[0.138,1.298]
Acute-on-chronic	2.245***	1.674***	1.691***	1.554**	1.541**
	[1.392,3.099]	[0.788,2.560]	[0.803,2.579]	[0.596,2.512]	[0.577,2.504]
ESRD	0.611	0.555	0.576	0.779	0.770
	[-0.0284,1.250]	[-0.192,1.301]	[-0.175,1.328]	[-0.0227,1.581]	[-0.0301,1.571]
RCS1		-26.06***	-26.08***	-26.65***	-26.65***
		[-28.52,-23.61]	[-28.53,-23.63]	[-29.24,-24.07]	[-29.23,-24.07]
RCS2		-694.4***	-694.8***	-709.7***	-709.5***
		[-766.4,-622.3]	[-766.7,-622.8]	[-785.2,-634.1]	[-785.1,-634.0]
RCS3		22.28***	22.29***	22.76***	22.76***
		[19.93,24.63]	[19.94,24.64]	[20.30,25.22]	[20.30,25.22]
Sex			0.0764		-0.0520
			[-0.133,0.286]		[-0.273,0.169]
Age				0.0615***	0.0617***
				[0.0540,0.0691]	[0.0541,0.0694]
Constant	-1.618***	2.782***	2.752***	-1.016***	-1.008***
	[-1.705,-1.531]	[2.492,3.073]	[2.447,3.056]	[-1.516,-0.515]	[-1.512,-0.504]
N	188399	188399	188399	188399	188399
AIC	11076.3	9705.7	9707.2	9351.1	9352.9

95% confidence intervals in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

AKI Acute Kidney Injury *ESRD* End-stage renal disease

Table 1: *Summary of parameter estimates, confidence intervals and values of Akaike's Information Criterion (AIC) for five different instantaneous geometric odds ratio models*

Model 3 includes also the sex of the patient as a covariate. However, we see from Table 1 that the coefficient estimate for this covariate is not significant. The odds ratio for sex is approximately 1.08 implying that the odds of death is 8% higher for a female patient than for a male. The 95% confidence interval for the odds ratio is $(0.88, 1.33)$. When the confidence interval includes 1, there is no significance at the 5% significance level, so we cannot be sure of the effect the covariate has on the outcome and the estimated effect is also very small. In this model the coefficient estimates are very close to the ones in Model 2. AIC is higher in this model compared to Model 2, 9707.2 respective 9705.7. We therefore conclude that Model 3 fits the data worse than Model 2.

In the fourth model we add age as a predictor variable to Model 2. We notice from Table 1 that the coefficient estimate for the predictor variable age is significant and the odds ratio is approximately 1.06 meaning that the odds of death rises by 6% as the age increases. All the other coefficient estimates are significant in this model except from ESRD. The odds ratio for ESRD is 2.18 and the 95% confidence interval is $(0.98, 4.86)$. Because the confidence interval includes 1, the effect the covariate has on the outcome is not significant at the 5% level. The coefficient estimate for AKI is 0.578 which implies that the odds ratio is 1.78. This suggests that the odds of death is 78% higher for a patient with acute kidney injury than for a patient with no renal disease. AIC of this model is the smallest so far, 9351.1.

Model 5 is the saturated model that includes all predictor variables. The coefficient estimate for the covariate sex is negative, signaling that the risk of death is higher for men than for women. In fact, the odds ratio is 0.94 which implies that the risk of death decreases by 6% if the patient is a female. This is the opposite effect compared to Model 3. However, the 95% confidence interval for this odds ratio is $(0.76, 1.18)$ which includes 1 and therefore the effect of the covariate is insignificant and small. AIC for this model is 9352.9 which is higher than for Model 4. Thus we choose Model 4.

In order to illustrate the instantaneous geometric odds for the different renal diseases, we calculate the linear prediction of the link function. Age has to be set to a fixed value, otherwise we obtain the instantaneous geometric odds for all ages, which makes the outcome difficult to interpret. We start with the average age, which is 59, and obtain Figure 2.

We notice from Figure 2 that the odds of death for all groups is at the highest in the beginning of the study and then it decreases drastically during the first half year. After the odds reach their minimum at around six months from the study begin, they start to increase modestly. Although after three years the odds of death starts to decrease again. The odds of death is highest for those who have acute-on-chronic renal disease. For the other renal diseases the risk of death is approximately the same and not surprisingly, it is the lowest for the patients without any renal disease.

Figures 3 and 4 show the instantaneous geometric odds for patients of ages 22 and 84 respectively, which are the 5th and 95th percentiles of the age's distribution. Notice that the curves for the instantaneous geometric odds shift upwards when the patients get older and downwards for the younger patients. The shift is very intuitive and is explained by the positive effect age has on the outcome. For instance, the instantaneous

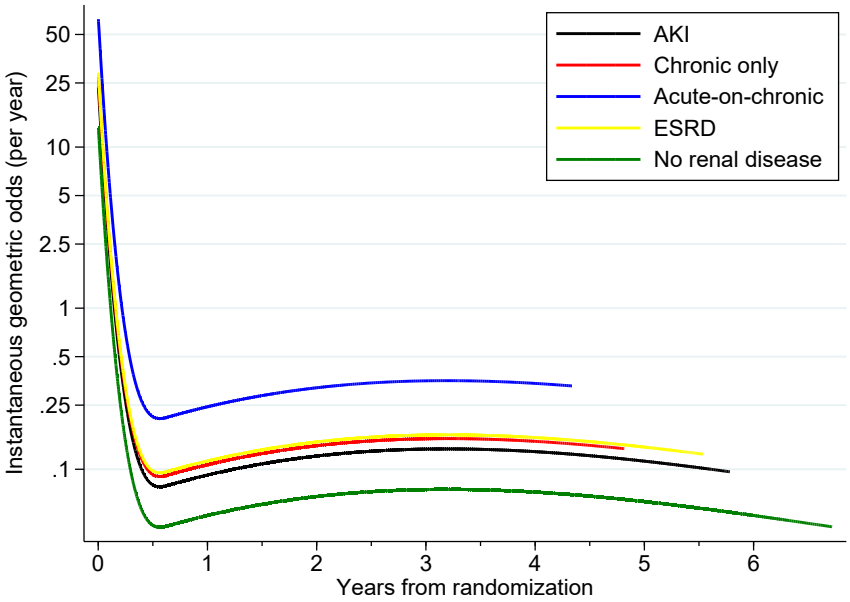

Figure 2: *The instantaneous geometric odds for renal diseases at age 59, as a function of follow-up time in years. The vertical axis in on a log scale.*

geometric odds for a person with acute-on-chronic disease after one year is approximately 0.025 at age 22. The odds of death then increases to around 0.25 at age 59 and to 1.1 at age 84.

We proceed by adding an interaction term to our instantaneous geometric odds model to assay the interplay between the categorical variable and the restricted cubic splines of the time variable. It is in our interest to study different combinations of the interaction terms in order to choose the model that fits data the best. We start by introducing an interaction term between RCS1 and the categorical variable, in order to obtain Model 6. The coefficient estimates for the covariates are shown in Table 2. We have excluded the coefficient estimates for the interaction terms from the table because they are impossible to interpret. The complete summaries are found in Section 7. The coefficient estimate for chronic only kidney disease is lower than in Model 4, that we chose earlier. The odds of death for a patient with chronic kidney disease is approximately 20% higher than for a patient with no renal disease. In Model 4 the the odds of death was approximately 107% higher. The coefficient estimates for the other renal diseases are higher in this model compared to Model 4 meaning that the positive effect the diseases have on the odds of death is higher in this model. The only interaction term that is significant is between RCS1 and acute-on-chronic kidney disease.

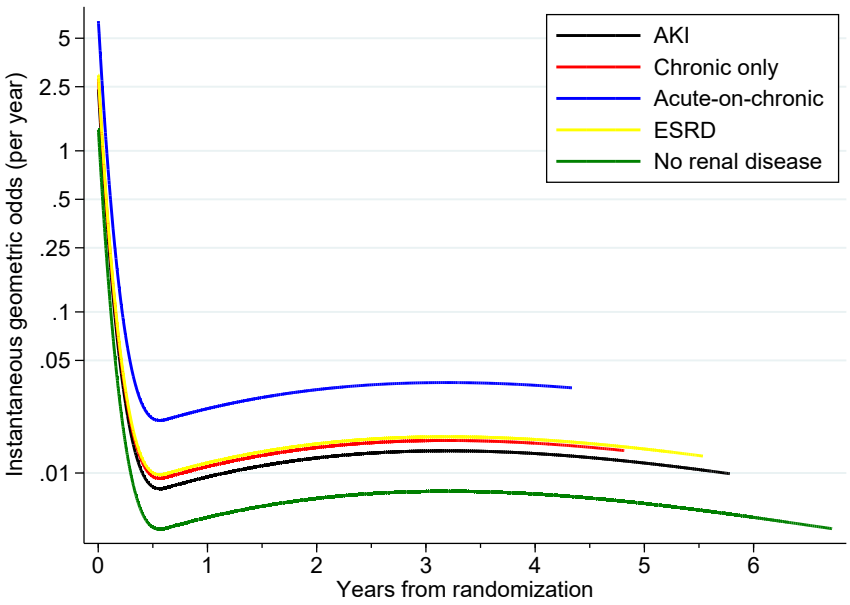

Figure 3: *The instantaneous geometric odds for renal diseases at age 22, as a function of follow-up time in years. The vertical axis in on a log scale.*

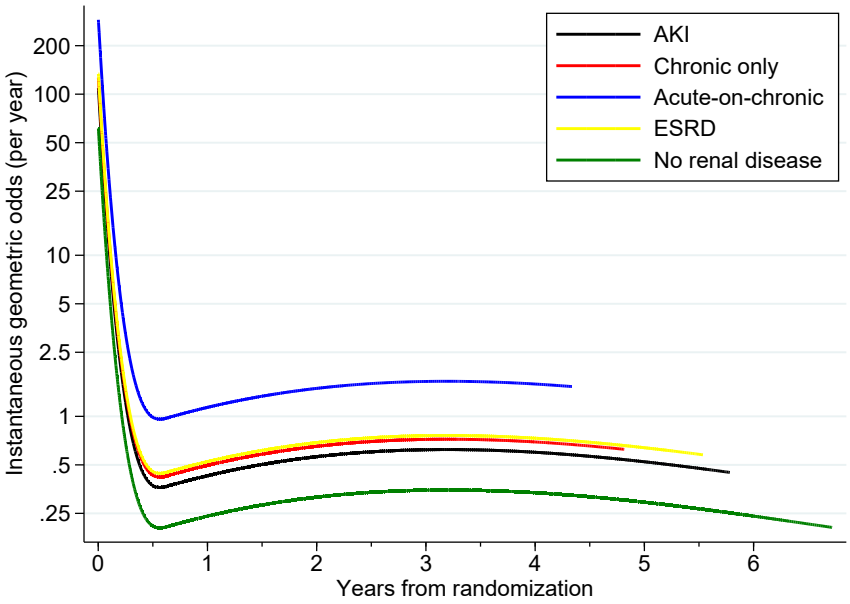

Figure 4: *The instantaneous geometric odds for renal diseases at age 84, as a function of follow-up time in years. The vertical axis in on a log scale.*

	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
AKI	0.825**	0.723**	0.707**	0.822*	0.698*	0.805*	3.001***
	[0.254,1.396]	[0.246,1.200]	[0.241,1.173]	[0.0913,1.553]	[0.0598,1.337]	[0.0863,1.523]	[1.253,4.750]
Chronic only	0.180	0.407	0.452	0.215	0.237	0.209	0.639
	[-0.687,1.047]	[-0.305,1.119]	[-0.235,1.139]	[-0.868,1.297]	[-0.728,1.202]	[-0.855,1.273]	[-1.277,2.556]
Acute-on-chronic	2.836***	2.460***	2.388***	2.438	1.988*	2.325*	10.01**
	[1.236,4.435]	[1.084,3.836]	[1.054,3.722]	[-0.0103,4.887]	[0.275,3.701]	[0.0126,4.638]	[3.345,16.67]
ESRD	1.187	0.944	0.910	1.761*	1.503	1.724	3.100
	[-0.173,2.547]	[-0.0994,1.988]	[-0.0862,1.907]	[0.000620,3.521]	[-0.0161,3.022]	[-0.00634,3.454]	[-0.611,6.810]
RCS1	-26.55***	-26.60***	-26.61***	-26.58***	-26.67***	-26.60***	-25.26***
	[-29.14,-23.96]	[-29.18,-24.01]	[-29.20,-24.02]	[-29.18,-23.97]	[-29.27,-24.07]	[-29.20,-24.00]	[-27.94,-22.58]
RCS2	-707.1***	-708.3***	-708.8***	-708.2***	-710.4***	-708.8***	-670.0***
	[-782.7,-631.5]	[-784.0,-632.7]	[-784.4,-633.1]	[-784.3,-632.2]	[-786.5,-634.3]	[-784.9,-632.8]	[-748.4,-591.7]
RCS3	22.68***	22.72***	22.73***	22.72***	22.79***	22.74***	21.48***
	[20.21,25.14]	[20.25,25.18]	[20.27,25.20]	[20.24,25.20]	[20.30,25.27]	[20.26,25.22]	[18.92,24.03]
Age	0.0614***	0.0615***	0.0615***	0.0614***	0.0615***	0.0614***	0.0608***
	[0.0538,0.0689]	[0.0539,0.0690]	[0.0539,0.0690]	[0.0539,0.0690]	[0.0540,0.0691]	[0.0539,0.0690]	[0.0533,0.0683]
Constant	-1.031***	-1.029***	-1.028***	-1.036***	-1.025***	-1.034***	-1.138***
	[-1.531,-0.530]	[-1.529,-0.528]	[-1.529,-0.528]	[-1.538,-0.533]	[-1.527,-0.523]	[-1.536,-0.531]	[-1.644,-0.631]
N	188399	188399	188399	188399	188399	188399	188399
AIC	9350.8	9351.3	9351.2	9357.3	9357.1	9357.1	9343.7

95% confidence intervals in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

AKI Acute Kidney Injury *ESRD* End-stage renal disease

Table 2: *Summary of parameter estimates, confidence intervals and values of Akaike's Information Criterion (AIC) for seven different instantaneous geometric odds ratio models*

We also include interaction terms between RCS2 and RCS3 and the categorical variable in order to obtain Model 7 and Model 8, respectively. The coefficient estimates for the covariate age are the same as in Model 4. Here as well, the coefficient estimates for the renal diseases are higher than in Model 4, despite the coefficient estimate for chronic only kidney disease. In Model 7, the only significant interaction term is between acute-on-chronic kidney disease and RCS2. The only significant interaction term in Model 8 is between acute-on-chronic kidney disease and RCS3. The AIC scores of these models are higher than for Model 4 meaning that these models have worse fit to data. Although the difference in the AIC scores is very small.

Next, we introduce one more interaction term to the models and obtain Model 9, Model 10 and Model 11. We have interaction terms between RCS1 and the categorical variable as well as RCS2 and the categorical variable in Model 9. In Model 10 the interaction terms are between the categorical variable and RCS2 and RCS3. Model 11 includes interaction terms between the renal diseases and RCS1 and RCS3. None of the interaction terms in these three models are significant. The AIC scores of these models are larger than the AIC scores for the models with only one interaction term and therefore they have worse fit.

Finally, we add third interaction term and obtain Model 12 that therefore includes interaction terms between all of the restricted cubic spline terms and the categorical variable. We choose this model because it has the lowest AIC score. This model has even a better fit to data than Model 4, the model with no interactions, that we chose earlier due to its lower AIC score. The interaction terms between the all of the restricted cubic spline terms and acute-on-chronic kidney disease are significant. Also the interaction terms between the restricted cubic spline terms and acute kidney injury are significant. Other interaction terms are not significant. The estimated coefficient for acute kidney injury is 3.001 which implies that the odds of death for a patient with this disease is approximately 20 times higher than for a patient with no renal disease. The odds of death for a patient with acute-on-chronic kidney disease is approximately 22 247 times higher than for a patient with no renal disease. The estimated coefficient for age is close to the estimated coefficient in Model 4.

We perform Wald-test (see [1] p.11) to test if at least one of the 12 coefficient estimates of the interaction terms of Model 12 are significantly different from zero. The test statistic $Z_w = 23.97$ is obtained and we compare it against a chisquare distributed random variable χ^2_{12} with 12 degrees of freedom. We get the $p-$value $P(\chi^2_{12} > Z_w) = 0.0205$, thus we can reject the null hypothesis that no interaction term is different from zero. Hence we choose the model with the interaction terms, Model 12.

Because we are not able to interpret the coefficients for the interaction terms between the restricted cubic spline terms and the renal diseases, we illustrate the instantaneous geometric odds ratio (IGOR) as a function of follow-up time. We predict the IGOR using the chosen model and calculate 95% confidence interval for the IGOR. Figure 5 shows the IGOR for the patients with acute kidney injury. The IGOR varies quite a lot and we see the odds of death is at its highest in the very beginning of the study. The confidence interval is wide in the very beginning but it quickly gets narrower. After

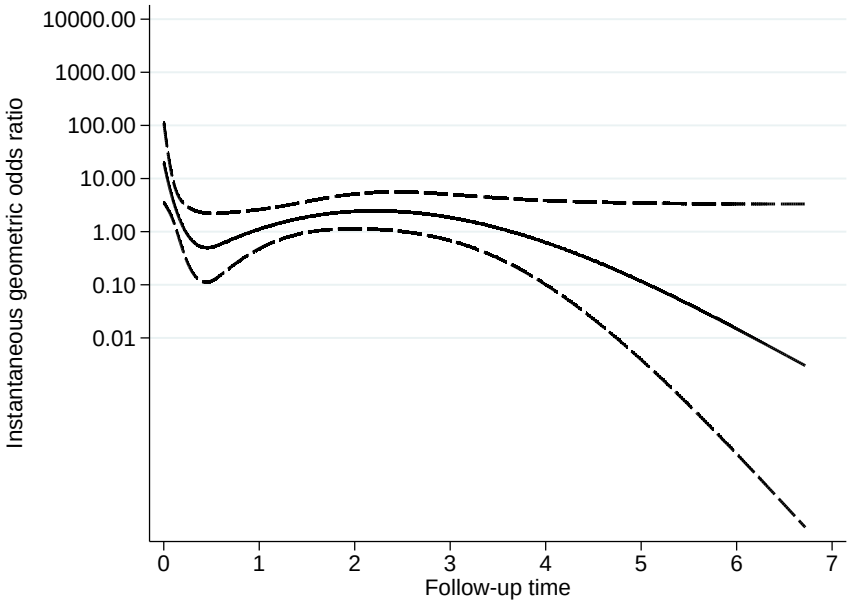

Figure 5: *Instantaneous geometric odds ratio for AKI with 95% confidence intervals for Model 12. The vertical axis is on a log scale.*

four years of observation the confidence intervals get wider again and hence the IGOR becomes more uncertain.

The IGOR for chronic kidney disease is shown in Figure 6. We notice that the IGOR decreases during the first months but then starts to increase steadily. The confidence intervals become really wide after the first three years. Hence the actual IGOR after four years is uncertain. The lowest risk for death for patients with chronic kidney disease is at around half year after study begins.

In figure 7 we see the IGOR and its 95% confidence intervals for the patients with acute-on-chronic kidney disease. The follow-up time for this group is at most one year and 7 months because after that the IGOR is very uncertain. This is due to the small amount of observations in this group. We see that the IGOR is really high, approximately 10 000, in the beginning of the study and it drops rapidly to values below 1 during the first half year. After that it increases again, meaning that the odds of death rises. Just before one-year follow-up the IGOR starts to decrease again. This means that the highest risk of death is during the first year and if the patient survives the first year, the odds of survival is higher.

The IGOR and confidence intervals for end-stage renal disease are illustrated in Figure 8. The confidence intervals are quite wide, which is partially explained by the small number of observations. As seen in Figure 8, the odds of death is highest in the beginning and decreases during the first half year. After two years the odds of death starts to increase. This means that if the patient survives the first half year the odds of

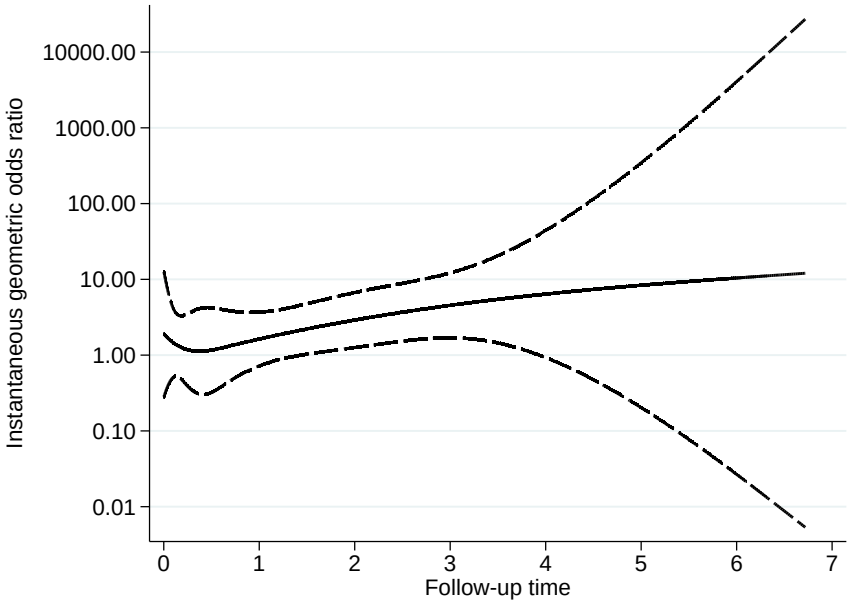

Figure 6: *Instantaneous geometric odds ratio for chronic kidney disease with 95% confidence intervals for Model 12. The vertical axis is on a log scale.*

survival the following two years is constant but after that the odds of survival decreases.

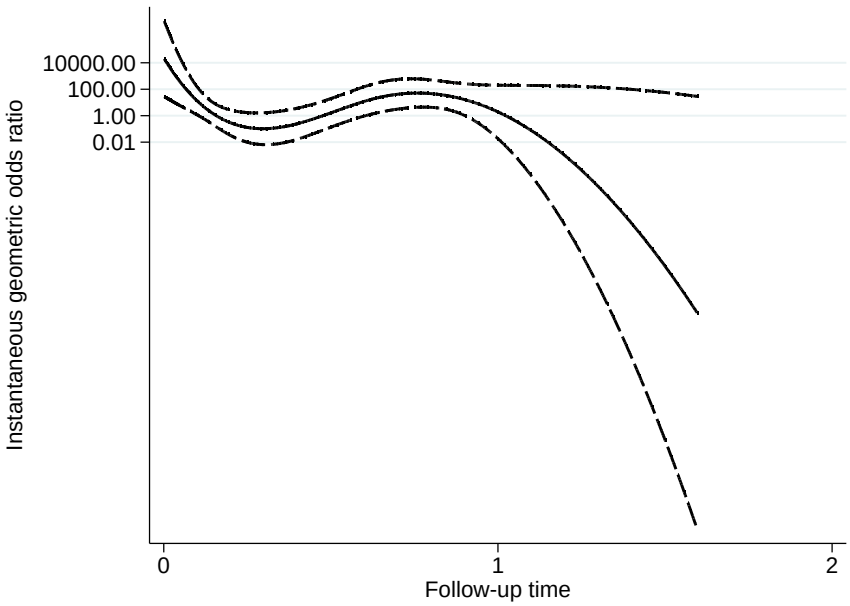

Figure 7: *Instantaneous geometric odds ratio for acute-on-chronic kidney disease with 95% confidence intervals for Model 12. The vertical axis is on a log scale.*

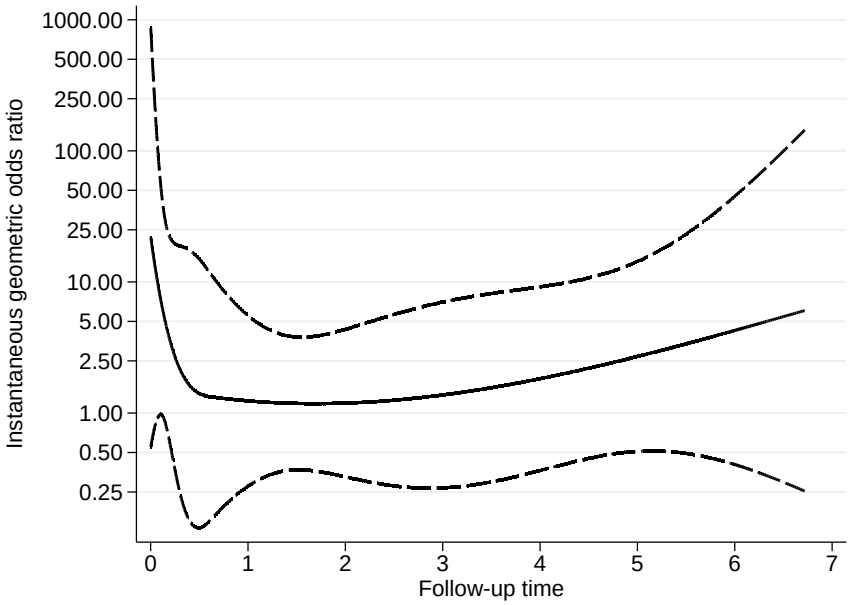

Figure 8: *Instantaneous geometric odds ratio for ESRD with 95% confidence intervals for Model 12. The vertical axis is on a log scale.*

4 Discussion

In Section 3.3 we concluded that the model with the best fit to data is Model 12. This model includes interaction terms between all the restricted cubic spline terms and the categorical variable with renal diseases. Since sex is not included as a predictor variable in this model, we can say that the sex has no effect on the odds of death for this dataset. From Figures 2-4 we saw that the patients with acute-on-renal kidney disease have the highest odds of death. The patients who have either ESRD, AKI or only chronic kidney disease have approximately the same risk of death. The first half year after the study begins is the most crucial time for the survival. The risk of death is the highest during this time but if the patient survives this time period the risk of death decreases. Figures 5-6 show that the odds of death becomes more uncertain, due to wider 95% confidence intervals, after four years for AKI and chronic kidney disease, which makes our conclusions more vague. For acute-on-chronic kidney disease the odds of death gets unsure already after the first year. Due to wide 95% confidence intervals for the odds ratio for ESRD, we cannot say with certainty how large effect the disease has on the odds of death.

4.1 Limitations

We choose to use generalized linear models instead of quantile regression in Section 3.3. The reason for this choice was that we wanted to illustrate the instantaneous geometric rate. The instantaneous geometric rate model was fitted at first but it failed to converge; the model was ill-defined. When the probability to die is close to 1, the log-likelihood function is negative and very close to 0 and therefore the maximum likelihood estimates cannot be found. Thus we decided to proceed with the instantaneous geometric odds model that uses the logit link function instead of the log link. A study on related problems with failed convergence can be found in [16].

It is important to notice that the amount of observations is relatively small, 2000, and that we have not separated the dataset into testing data and validation data. Therefore we have not tested the prediction ability of the models. Thus our conclusions should be considered with caution.

4.2 Future work

In future research it would be interesting to fit also Cox proportional hazards model, which is a very common method in survival analysis, to data and compare the results. The exponentiated coefficient estimates of this model represent the hazard ratios which indicate the effect of the covariate on the hazard rate [11]. The hazard rate describes the risk to die, but per definition it is not a probability. However, the geometric rate, given in equation (6) is the average probability to die over time interval $(0, t)$. Due to this, the geometric rate seems more appropriate method to estimate the occurrence of death. The relationship between these two are therefore of interest.

Another future aspect that could be of interest is to perform the same analysis as in Section 3.3 with simulated data instead, this in order to analyze if the simulated data gives similar results as when using original data. Knowing the true (simulated) data gives us the possibility to test how well the models really perform and to verify if our conclusions are correct.

5 Acknowledgements

First of all, I would like to express my sincere gratitude to Matteo Bottai, my supervisor at Karolinska Institutet, for your time and valuable guidance, and for introducing me to this interesting topic. Thank you Ola Hössjer and Disa Hansson, my supervisors at Stockholms University, for your wise comments and advice. I also wish to thank Andrea Discacciati for your help. Last but not least I would like to thank my boyfriend, Carlos, for your support and patience.

References

[1] Agresti, A. (2002): *Categorical Data Analysis*, 2nd edition. New Jersey ,USA. John Wiley & Sons, Inc.

[2] Allen, I. & Seaman J. (2007): *Online Nation: Five Years of Growth in Online Learning*. Sloan Consortium.

[3] Bleumink, G., Knetsch, A., Sturkenboom, M., Strausa, S., Hofmana, A., Deckers, J., Wittemana, J. & Strickera, B. (2004): Quantifying the heart failure epidemic: prevalence, incidence rate, lifetime risk and prognosis of heart failure: The Rotterdam Study. *European Heart Journal* Vol. 25(18): 1614–1619.

[4] Bottai, M. (2017): A regression method for modelling geometric rates. *Statistical Methods in Medical Research* Vol. 26(6): 2700–2707.

[5] Cade, B. & Noon, B. (2003): A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* Vol. 1(8): 412–420.

[6] Collett, D. (1994): *Modelling Survival Data in Medical Research* 1st edition. London, United Kingdom. Chapman & Hall.

[7] Discacciati, A. & Bottai, M. (2017): Instantaneous geometric rates via generalized linear models. *The Stata Journal* Vol. 17(2): 358–371.

[8] Heinzl, H. & Kaider, A. (1997): Gaining more flexibility in COX proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine* Vol. 54(3): 201-208.

[9] Koenker, R. (2005): *Quantile Regression*. Econometric Society Monagraph Series. Cambridge. Cambridge University Press.

[10] Lambert, P. (2008): rcsgen: Stata module to generate restricted cubic splines and their derivatives. Statistical Software Components S456986, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456986.html.

[11] Liu, X. (2012): *Survival Analysis: Models and Applications*. John Wiley & Sons, Inc.

[12] Pagano, M. & Gauvreau, K. (1993): *Principles of Biostatistics*. Belmont, California. Wadsworth, Inc.

[13] Rimes-Stigare, C., Bottai, M., Mårtensson, J., Martling, C. & Bell,M. (2015): Long-term mortality and risk factors for development of end-stage renal disease in critically ill patients with and without chronic kidney disease. *Critical Care*, Vol. 19: 383.

[14] Swamy, W. (2014): Financial Inclusion, Gender Dimension, and Economic Impact on Poor Households. *World Development*, Vol. 56: 1–15.

[15] Tran, T., Davila, J. & El-Serag, H. (2005): The epidemiology of malignant gastrointestinal stromal tumors: An analysis of 1,458 cases from 1992 to 2000. *American Journal of Gastroenterology* Vol 100(1): 162-168.

[16] Williamsson, T., Eliasziw, M. & and Fick, G. (2013): Log-binomial models: exploring failed convergence. *Emerging Themes in Epidemiology*, Vol 10: 14.

6 Appendix A

To program the log link function for the instantaneous geometric rate model Discacciati and Bottai [7] use equation (11) and provide the following link program `log_igr` in Stata.

```
*! version 1.0.0 - 07dec2016
capture program drop log_igr
program define log_igr
\space version 7
\space args todo eta mu return
if `todo´ == -1 { /* Title */
global SGLM_lt "Log IGR"
global SGLM_lf "log(1-exp(-u/$SGLM_p))"
capture confirm numeric variable $SGLM_p
if _rc != 0 {
noi di as error "argument ($SGLM_p) to log_igr " /*
*/ "link function must be a numeric variable"
exit 198
}
exit
}
if `todo´ == 0 { /* eta = g(mu) */
gen double `eta´ = log(-exp(-`mu´/$SGLM_p)+1)
exit
}
if `todo´ == 1 { /* mu = g^-1(eta) */
gen double `mu´ = -$SGLM_p*log(-exp(`eta´)+1)
exit
}
if `todo´ == 2 { /* (d mu)/(d eta) */
gen double `return´ = $SGLM_p*exp(`eta´)*(-exp(`eta´)+1)^(-1)
exit
}
if `todo´ == 3 { /* (d^2 mu)/(d eta^2) */
gen double `return´ = $SGLM_p*exp(`eta´)*(exp(`eta´)-1)^(-2)
exit
}
noi di as err "Unknown call to glm link function"
exit 198
end
```

The following is the link program `logit_igr` for the instantaneous geometric odds model that uses the derivatives in equation (15).

```
*! version 1.0.0 - 07dec2016
```

```
capture program drop logit_igr
program define logit_igr
version 7
args todo eta mu return
if `todo´ == -1 { /* Title */
global SGLM_lt "Logit IGR"
global SGLM_lf "logit(1-exp(-u/$SGLM_p))"
confirm numeric variable $SGLM_p
if _rc != 0 {
noi di as error "argument ($SGLM_p) to logit_igr " /*
*/ "link function must be a numeric variable"
exit 198
}
exit
}
if `todo´ == 0 { /* eta = g(mu) */
gen double `eta´ = logit(1-exp(-`mu´/$SGLM_p))
exit
}
if `todo´ == 1 { /* mu = g^-1(eta) */
gen double `mu´ = -$SGLM_p*log((exp(`eta´)+1)^(-1))
exit
}
A. Discacciati and M. Bottai 363
if `todo´ == 2 { /* (d mu)/(d eta) */
gen double `return´ = $SGLM_p*exp(`eta´)*(exp(`eta´)+1)^(-1)
exit
}
if `todo´ == 3 { /* (d^2 mu)/(d eta^2) */
gen double `return´ = $SGLM_p*exp(`eta´)*(exp(`eta´)+1)^(-2)
exit
}
noi di as err "Unknown call to glm link function"
exit 198
end
```

7 Appendix B

This appendix includes the complete summaries of the instantaneous geometric odds models introduced in Section 3.3. Figures 9-13 show the summaries for the models without interaction terms. The summaries for the models with interaction terms are found in Figures 14-20.

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188394
                                                  Scale parameter =         1
Deviance         =  9604.310418                   (1/df) Deviance =  .0509799
Pearson          =   900504.981                   (1/df) Pearson  =  4.779903

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0587918
Log pseudolikelihood = -5533.155209               BIC             =  -2278689
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	1.188095	.1701551	6.98	0.000	.854597	1.521593
2	1.336669	.2253139	5.93	0.000	.8950619	1.778276
3	2.245257	.4355108	5.16	0.000	1.391671	3.098842
4	.6107905	.3261182	1.87	0.061	-.0283893	1.24997
_cons	-1.617833	.0443052	-36.52	0.000	-1.70467	-1.530996

Figure 9: *Summary of Model 1*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188391
                                                  Scale parameter =         1
Deviance         =  8227.692007                   (1/df) Deviance =  .0436735
Pearson          =  1029814.048                   (1/df) Pearson  =  5.466365

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0515167
Log pseudolikelihood = -4844.846004               BIC             =  -2280029
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.9839176	.2062501	4.77	0.000	.5796748	1.38816
2	1.503842	.2809672	5.35	0.000	.9531565	2.054528
3	1.673895	.4520712	3.70	0.000	.7878522	2.559939
4	.554527	.3809553	1.46	0.145	-.1921317	1.301186
_rcs1	-26.06219	1.251657	-20.82	0.000	-28.51539	-23.60899
_rcs2	-694.3605	36.74472	-18.90	0.000	-766.3788	-622.3422
_rcs3	22.27851	1.198377	18.59	0.000	19.92973	24.62728
_cons	2.782481	.1484151	18.75	0.000	2.491593	3.073369

Figure 10: *Summary of Model 2*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188390
                                                  Scale parameter =         1
Deviance         =  8227.170155                   (1/df) Deviance =  .0436709
Pearson          =  1029376.625                   (1/df) Pearson  =  5.464073

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0515245
Log pseudolikelihood = -4844.585077               BIC             =  -2280018
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.9914237	.2072329	4.78	0.000	.5852547	1.397593
2	1.511243	.2810347	5.38	0.000	.9604249	2.062061
3	1.690798	.4532088	3.73	0.000	.8025246	2.57907
4	.576194	.3834301	1.50	0.133	-.1753151	1.327703
_rcs1	-26.07708	1.250996	-20.85	0.000	-28.52899	-23.62517
_rcs2	-694.7533	36.72253	-18.92	0.000	-766.7281	-622.7784
_rcs3	22.29112	1.197644	18.61	0.000	19.94378	24.63846
female	.0763974	.1070348	0.71	0.475	-.133387	.2861818
_cons	2.751518	.1552525	17.72	0.000	2.447229	3.055807

Figure 11: *Summary of Model 3*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188390
                                                  Scale parameter =         1
Deviance         =  7871.131685                   (1/df) Deviance =   .041781
Pearson          =  1169215.083                   (1/df) Pearson  =  6.206354

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496347
Log pseudolikelihood = -4666.565842               BIC             =  -2280374
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.5783641	.2158868	2.68	0.007	.1552338	1.001494
2	.7259244	.2952664	2.46	0.014	.1472129	1.304636
3	1.553973	.4889347	3.18	0.001	.5956788	2.512268
4	.7792706	.4091904	1.90	0.057	-.0227277	1.581269
_rcs1	-26.65356	1.318652	-20.21	0.000	-29.23807	-24.06905
_rcs2	-709.6511	38.53155	-18.42	0.000	-785.1716	-634.1307
_rcs3	22.76032	1.255672	18.13	0.000	20.29925	25.22139
age	.0615109	.0038534	15.96	0.000	.0539585	.0690634
_cons	-1.015773	.2552989	-3.98	0.000	-1.51615	-.5153967

Figure 12: *Summary of Model 4*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188389
                                                  Scale parameter =         1
Deviance         =  7870.912902                   (1/df) Deviance =  .0417801
Pearson          =  1171169.227                   (1/df) Pearson  =   6.21676

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496442
Log pseudolikelihood = -4666.456451               BIC             =  -2280362
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.5727572	.2169591	2.64	0.008	.1475252	.9979891
2	.7181447	.2958258	2.43	0.015	.1383368	1.297953
3	1.540588	.491517	3.13	0.002	.5772323	2.503943
4	.7704573	.4084633	1.89	0.059	-.0301161	1.571031
_rcs1	-26.64997	1.318642	-20.21	0.000	-29.23446	-24.06547
_rcs2	-709.5398	38.52892	-18.42	0.000	-785.0551	-634.0245
_rcs3	22.75657	1.255562	18.12	0.000	20.29571	25.21743
age	.0617242	.0038987	15.83	0.000	.0540829	.0693655
female	-.0520052	.1127641	-0.46	0.645	-.2730188	.1690083
_cons	-1.007757	.2570633	-3.92	0.000	-1.511592	-.5039223

Figure 13: *Summary of Model 5*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188386
                                                  Scale parameter =         1
Deviance         =  7862.764295                   (1/df) Deviance =  .0417375
Pearson          =  1165738.871                   (1/df) Pearson  =  6.188033

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496328
Log pseudolikelihood = -4662.382148               BIC             =  -2280333
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.8250415	.2913335	2.83	0.005	.2540384	1.396045
2	.1798399	.4424308	0.41	0.684	-.6873084	1.046988
3	2.835769	.8161383	3.47	0.001	1.236167	4.43537
4	1.187232	.6939793	1.71	0.087	-.1729423	2.547407
_rcs1	-26.55085	1.31975	-20.12	0.000	-29.13751	-23.96419
renalgroup#c._rcs1						
1	-.2095532	.184558	-1.14	0.256	-.5712803	.152174
2	.4163731	.2464498	1.69	0.091	-.0666597	.8994059
3	-1.73795	.8484491	-2.05	0.041	-3.40088	-.0750202
4	-.2767886	.4070773	-0.68	0.497	-1.074645	.5210682
_rcs2	-707.1171	38.58479	-18.33	0.000	-782.7418	-631.4923
_rcs3	22.67952	1.257492	18.04	0.000	20.21488	25.14416
age	.0613773	.0038509	15.94	0.000	.0538298	.0689249
_cons	-1.030802	.2554182	-4.04	0.000	-1.531412	-.5301912

Figure 14: *Summary of Model 6*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188386
                                                  Scale parameter =         1
Deviance         =  7863.273063                   (1/df) Deviance =  .0417402
Pearson          =  1165997.569                   (1/df) Pearson  =  6.189407

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496355
Log pseudolikelihood = -4662.636532               BIC             =  -2280333
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.7226336	.2433278	2.97	0.003	.2457199	1.199547
2	.4070262	.3630826	1.12	0.262	-.3046025	1.118655
3	2.460188	.7020606	3.50	0.000	1.084175	3.836202
4	.9443717	.5325526	1.77	0.076	-.0994122	1.988156
_rcs2	-708.3093	38.59319	-18.35	0.000	-783.9505	-632.668
renalgroup#c._rcs2						
1	1.034479	.869046	1.19	0.234	-.6688199	2.737778
2	-2.100395	1.293305	-1.62	0.104	-4.635226	.4344353
3	19.63552	9.379807	2.09	0.036	1.25144	38.01961
4	.7859929	1.724167	0.46	0.648	-2.593311	4.165297
_rcs1	-26.59569	1.320242	-20.14	0.000	-29.18331	-24.00806
_rcs3	22.71786	1.257717	18.06	0.000	20.25278	25.18294
age	.0614655	.0038535	15.95	0.000	.0539128	.0690183
_cons	-1.028792	.2553931	-4.03	0.000	-1.529353	-.5282304

Figure 15: *Summary of Model 7*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188386
                                                  Scale parameter =         1
Deviance         =  7863.238621                   (1/df) Deviance =    .04174
Pearson          =  1166178.734                   (1/df) Pearson  =  6.190368

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496353
Log pseudolikelihood =  -4662.61931               BIC             =  -2280333
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.7072573	.2376709	2.98	0.003	.241431	1.173084
2	.4515921	.3505018	1.29	0.198	-.2353787	1.138563
3	2.38819	.6807106	3.51	0.000	1.054022	3.722358
4	.9103787	.5084675	1.79	0.073	-.0861993	1.906957
_rcs3	22.73262	1.257942	18.07	0.000	20.2671	25.19815
renalgroup#c._rcs3						
1	.0407582	.0334263	1.22	0.223	-.0247561	.1062725
2	-.0805495	.0508093	-1.59	0.113	-.180134	.019035
3	1.155989	.5447245	2.12	0.034	.0883485	2.223629
4	.0260699	.0632053	0.41	0.680	-.0978102	.1499499
_rcs1	-26.61142	1.320534	-20.15	0.000	-29.19962	-24.02322
_rcs2	-708.7637	38.60017	-18.36	0.000	-784.4186	-633.1087
age	.0614846	.0038539	15.95	0.000	.0539311	.069038
_cons	-1.028246	.2554135	-4.03	0.000	-1.528847	-.5276444

Figure 16: *Summary of Model 8*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188382
                                                  Scale parameter =         1
Deviance         =  7861.327608                   (1/df) Deviance =  .0417308
Pearson          =  1169871.804                   (1/df) Pearson  =  6.210104

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496676
Log pseudolikelihood = -4661.663804               BIC             =  -2280286
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.8219799	.3728025	2.20	0.027	.0913005	1.552659
2	.2145857	.5522223	0.39	0.698	-.8677501	1.296922
3	2.438117	1.249219	1.95	0.051	-.0103075	4.886541
4	1.760888	.8981123	1.96	0.050	.0006204	3.521156
_rcs1	-26.57537	1.327643	-20.02	0.000	-29.1775	-23.97324
renalgroup#c._rcs1						
1	-.2013774	.5646719	-0.36	0.721	-1.308114	.9053593
2	.3428461	.7972119	0.43	0.667	-1.219661	1.905353
3	.1192685	4.69832	0.03	0.980	-9.089269	9.327806
4	-1.601046	1.05152	-1.52	0.128	-3.661988	.4598951
_rcs2	-708.2307	38.79618	-18.26	0.000	-784.2698	-632.1916
renalgroup#c._rcs2						
1	.0544511	2.458465	0.02	0.982	-4.764052	4.872955
2	-.389724	4.267716	-0.09	0.927	-8.754294	7.974846
3	21.09642	44.72819	0.47	0.637	-66.56922	108.7621
4	-6.604942	3.845158	-1.72	0.086	-14.14131	.9314284
_rcs3	22.71836	1.264426	17.97	0.000	20.24013	25.19659
age	.0614124	.0038534	15.94	0.000	.0538599	.0689649
_cons	-1.035653	.2562442	-4.04	0.000	-1.537882	-.5334235

Figure 17: *Summary of Model 9*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188382
                                                  Scale parameter =         1
Deviance         =   7861.09051                   (1/df) Deviance =  .0417295
Pearson          =  1174048.552                   (1/df) Pearson  =  6.232276

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496663
Log pseudolikelihood = -4661.545255               BIC             =  -2280286
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.6983455	.3257866	2.14	0.032	.0598155	1.336876
2	.2373329	.4922792	0.48	0.630	-.7275165	1.202182
3	1.987869	.8738386	2.27	0.023	.2751768	3.700561
4	1.502946	.7750174	1.94	0.052	-.0160603	3.021952
_rcs2	-710.3971	38.83299	-18.29	0.000	-786.5083	-634.2858
renalgroup#c._rcs2						
1	-.5878605	13.84505	-0.04	0.966	-27.72365	26.54793
2	-11.57018	19.38448	-0.60	0.551	-49.56306	26.4227
3	-81.97832	142.2752	-0.58	0.564	-360.8326	196.8759
4	35.8218	25.68915	1.39	0.163	-14.52801	86.1716
_rcs3	22.78584	1.265911	18.00	0.000	20.3047	25.26698
renalgroup#c._rcs3						
1	.0638308	.5263878	0.12	0.903	-.9678702	1.095532
2	.372262	.7638726	0.49	0.626	-1.124901	1.869425
3	5.794803	7.71974	0.75	0.453	-9.335609	20.92522
4	-1.355962	.949649	-1.43	0.153	-3.21724	.5053158
_rcs1	-26.667	1.327193	-20.09	0.000	-29.26825	-24.06574
age	.0615234	.0038604	15.94	0.000	.0539571	.0690896
_cons	-1.024946	.256021	-4.00	0.000	-1.526737	-.5231537

Figure 18: *Summary of Model 10*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188382
                                                  Scale parameter =         1
Deviance         =  7861.148832                   (1/df) Deviance =  .0417298
Pearson          =  1170765.887                   (1/df) Pearson  =   6.21485

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0496667
Log pseudolikelihood = -4661.574416               BIC             =  -2280286
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	.8048148	.3666065	2.20	0.028	.0862792	1.52335
2	.2089735	.5429656	0.38	0.700	-.8552195	1.273167
3	2.325473	1.180037	1.97	0.049	.0126421	4.638303
4	1.723613	.8826463	1.95	0.051	-.0063424	3.453568
_rcs1	-26.59911	1.327894	-20.03	0.000	-29.20174	-23.99649
renalgroup#c._rcs1						
1	-.1673251	.4836991	-0.35	0.729	-1.115358	.7807078
2	.3636787	.6606007	0.55	0.582	-.9310748	1.658432
3	.2317315	3.693122	0.06	0.950	-7.006655	7.470118
4	-1.35438	.9176927	-1.48	0.140	-3.153025	.4442645
_rcs3	22.73796	1.264946	17.98	0.000	20.25871	25.21721
renalgroup#c._rcs3						
1	.0095897	.0803203	0.12	0.905	-.1478352	.1670146
2	-.011099	.1393706	-0.08	0.937	-.2842604	.2620623
3	1.308114	1.947707	0.67	0.502	-2.509322	5.125551
4	-.2112937	.1238904	-1.71	0.088	-.4541145	.0315271
_rcs2	-708.8483	38.81077	-18.26	0.000	-784.916	-632.7806
age	.0614358	.0038549	15.94	0.000	.0538804	.0689912
_cons	-1.033623	.2562211	-4.03	0.000	-1.535807	-.5314392

Figure 19: *Summary of Model 11*

```
Generalized linear models                         No. of obs      =    188399
Optimization     : ML                             Residual df     =    188378
                                                  Scale parameter =         1
Deviance         =  7839.739601                   (1/df) Deviance =  .0416171
Pearson          =  1133392.158                   (1/df) Pearson  =  6.016585

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = logit(1-exp(-u/risktime))[Logit IGR]

                                                  AIC             =  .0495955
Log pseudolikelihood = -4650.869801               BIC             =  -2280259
```

_d	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
renalgroup						
1	3.001442	.8922986	3.36	0.001	1.252569	4.750315
2	.6392724	.9777821	0.65	0.513	-1.277145	2.55569
3	10.00881	3.40006	2.94	0.003	3.344812	16.6728
4	3.09968	1.893209	1.64	0.102	-.6109409	6.8103
_rcs1	-25.26437	1.367579	-18.47	0.000	-27.94478	-22.58397
_rcs2	-670.0175	39.98157	-16.76	0.000	-748.38	-591.6551
_rcs3	21.47531	1.303026	16.48	0.000	18.92143	24.02919
renalgroup#c._rcs1						
1	-19.37662	7.744286	-2.50	0.012	-34.55514	-4.198102
2	-3.153337	7.265065	-0.43	0.664	-17.3926	11.08593
3	-90.9175	30.71899	-2.96	0.003	-151.1256	-30.70937
4	-12.20633	13.22635	-0.92	0.356	-38.1295	13.71684
renalgroup#c._rcs2						
1	-564.7459	227.885	-2.48	0.013	-1011.392	-118.0996
2	-102.3272	212.032	-0.48	0.629	-517.9022	313.2479
3	-3700.425	1220.468	-3.03	0.002	-6092.498	-1308.352
4	-311.0494	375.1474	-0.83	0.407	-1046.325	424.226
renalgroup#c._rcs3						
1	18.44484	7.437244	2.48	0.013	3.868107	33.02157
2	3.321131	6.918329	0.48	0.631	-10.23854	16.88081
3	151.8225	51.34392	2.96	0.003	51.19023	252.4547
4	9.881264	12.13908	0.81	0.416	-13.9109	33.67343
age	.0607687	.0038185	15.91	0.000	.0532846	.0682529
_cons	-1.137634	.2584182	-4.40	0.000	-1.644125	-.6311441

Figure 20: *Summary of Model 12*