CHL5250: SPECIAL TOPICS IN BIOSTATISTICS

The Diabetes Control and Complication Trial/Epidemiology of Diabetes Interventions and Complications Study: Survival Analysis with Longitudinal Data

January 7, 2019

Fangming Liao

Biostatistics
Dalla Lana School of Public Health
University of Toronto

Contents

1 Introduction

The Diabetes Control and Complication Trial (DCCT) was a multi-center, randomized clinical trial designed to compare intensive with conventional diabetes therapy regarding their effects on the development and progression of the early complications of type 1 diabetes [1]. A total of 1441 patients were recruited at 29 centers from 1983 through 1989. In June 1993, after an average follow-up of 6.5 years (range from 3 to 9), the independent data monitoring committee determined that the study results warranted terminating the trial.

At the beginning of the study, patients with type 1 diabetes are split into two cohorts: the primary prevention cohort, where patients don't have retinopathy at the base line, and the goal was to delay the development of retinopathy, and in the secondary intervention cohort are patients with mild retinopathy, where the goal was to slow the progression of retinopathy. Then within each cohort the patients are randomly assigned to intensive therapy or conventional therapy.

At the end of DCCT, all patients were instructed in intensive treatment, and all participants were invited to join the observational Epidemiology of Diabetes Interventions and Complications (EDIC) study, which subsequently monitored 93% of patients from DCCT study [2].

As Dr. Charlie Keown-Stoneman gave us an introduction to survival analysis on Nov.7th lecture, we found it interesting to assess the survival probability of intensive therapy compared with conventional therapy during the DCCT affected the incidence of renal function over 30 years of follow-up, whereas early survival analysis has been done for the DCCT by The Diabetes Control and Complications Trial Research Group [3].

2 Dataset

2.1 The DCCT/EDIC Dataset

To measures kidney functions, albumin excretion rate (AER) and estimated glomerular filtration rate (eGFR) were used. AER measures protein albumin in patients' urine, whereas larger amounts occur in the urine of patients with kidney disease, and it was measured every year during DCCT measured every other year during the EDIC study. On the other hand, Serum creatinine was measured every year throughout the DCCT and the EDIC study, and the results were used to estimate glomerular filtration rate with the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula.

The dataset for the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study was in longitudinal format where patients' indices were repeatedly recorded during the DCCT/EDIC study. There were 34,564 observations and 25 variables in the original dataset, for this study we only kept variables that were needed, and *Table 1* gives a description for these variables,

Variable	Description
DCCTYEAR	Years since the beginning of DCCT study.
AER	Measure of Albumin Excretion Rate.
CKD_GEF	Estimated glomerular filtration rate.
CONV	Binary indicator where 1 indicates Conventional Therapy.

Table 1: Description for variables

2.2 Preparing the Longitudinal Data

Through the DCCT and EDIC study, we did a survival analysis where considering the first time when a patient's AER greater than 30 mg per 24 hours on two consecutive visits as an event, which is called Microalbuminuria, and the time to that event is years since the patient participated in DCCT study. Patients who dropped out or never have their AER greater than 30 during the DCCT and EDIC study are considered censoring.

Technically, we find the first time when a patient's AER is greater than 30 and consider that time as the time to the incident. Thus, we transform the longitudinal data into a dataset where each observation contains a patient's ID and survival time, life status at the end of the study and treatment group during DTTC and EDIC, demonstrated as in *Table 2*.

Patient ID	Survival time	Status	Treatment
patient1	t1	0	1
patient2	t2	1	0
...	...	...	...

Table 2: Dataset for survival analysis. Status and Treatment are indicator variables, in which 0 represents alive and treated with conventional therapy, and survival time in years.

We processed the data similarly for eGFR, whereas eGFR less than 60 mL/min per $1.73m^2$ is considered as an event.

3 Methodology

By transforming data into the format above, we are now able to draw the Kaplan-Meier Survival curve (life table), and the analysis and plotting were done using *R* packages *survival* [4] and *survminer* [5].

As Charlie stated [6], for discrete time, the hazard at time t is the probability of the incident at time t given survival up to time t, and the hazard function is defined as:

$$\alpha(t) = P(T = t | T \geq t)$$

To test the survival probability difference between two treatment groups, log-rank test was employed where the null hypothesis assumes the two groups having identical survival and hazard functions. A log-rank test for 2 groups (A and B) is based on the below log-rank statistic [7] [8]

$$LR = \frac{U^2}{V} \sim \chi(1)$$

where

$$U = \sum_{i=1}^{T} w_{t_i}(o_{t_i}^A - e_{t_i}^A), \quad V = Var(U) = \sum_{i=1}^{T} (w_{t_i}^2 \frac{n_{t_i}^A n_{t_i}^B d_i (n_{t_i} - o_{t_i})}{n_{t_i}^2 (n_{t_i} - 1)})$$

and

- t_i for $i = 1, ..., T$ are possible event times,
- n_{t_i} is the overall risk set size on the time t_i,
- $n_{t_i}^A$ is the risk set size on the time t_i in group A,
- $n_{t_i}^B$ is the risk set size on the time t_i in group B,
- o_{t_i} is the overall observed events on the time t_i,
- $o_{t_i}^A$ is the observed events on the time t_i in group A,
- $o_{t_i}^B$ is the observed events on the time t_i in group B,
- e_{t_i} number of overall expected events in the time on the time t_i,
- $e_{t_i}^A$ number of expected events in the time on the time t_i in group A,
- $e_{t_i}^B$ number of expected events in the time on the time t_i in group B,

- w_{t_i} is a weight for the statistic

Combining the Kaplan-Meier Survival curve and log-rank test statistic, we would be able to drew preliminary conclusion about the effect of the intensive therapy compared with the conventional therapy in terms of renal benefit in patients with type 1 diabetes.

4 Results

4.1 Basic Statistics

Before doing survival analysis, it is necessary to check the data quality by plotting the distribution of the response variables, AER and eGFR as in *Figure 1*.

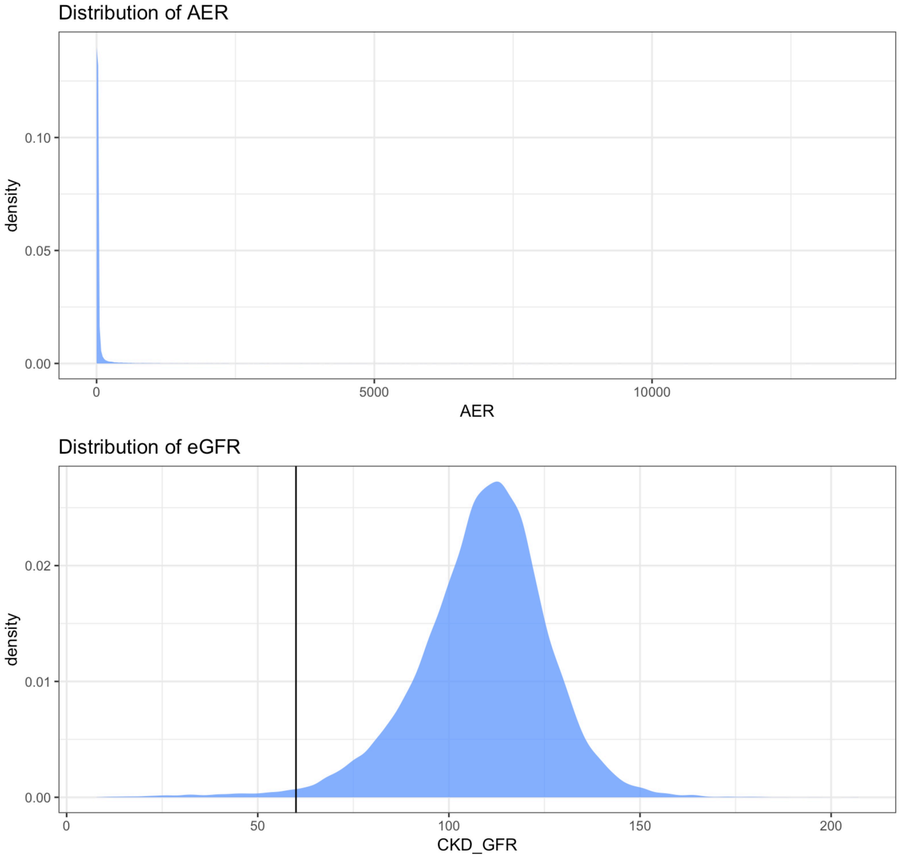


Figure 1: Distribution of albumin excretion rate (AER) and the estimated glomerular filtration rate (eGFR).

We notice that the eGFR is normally distributed, and the area on the left hand side of the vertical line in the plot indicates records that are less than 60 mL/min per 1.73m^2. However, the density for AER was extremely left skewed so can we can't drew much conclusion, then we decide to plot the density on a log-scaled histogram as in *Figure 2*.

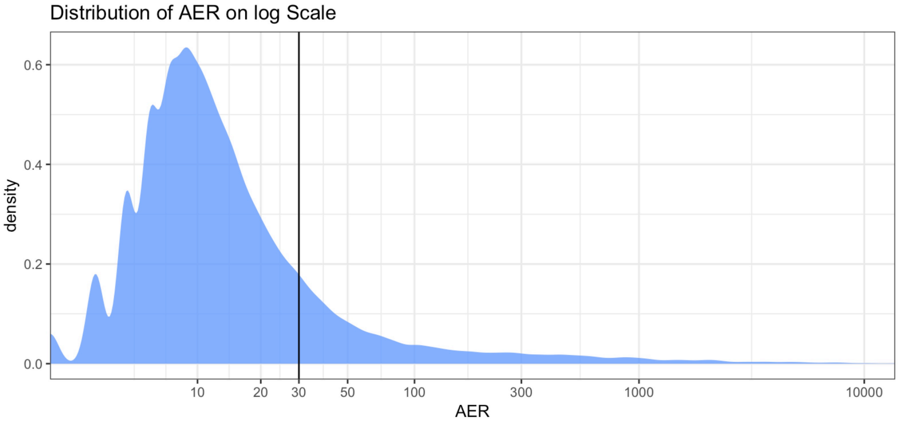


Figure 2: Distribution of albumin excretion rate (AER) on log scale.

Seeing from this plot, although there is a long right tail for the distribution, it seems normally distributed with a mean of about 10, and the area on the right hand side of the vertical line indicates records that are larger than 30.

4.2 Survival Analysis

By using *R* packages *survival* [4] and *survminer* [5], we plotted the Kaplan-Meier Survival curve for both intensive and conventional treatment groups as in *Figure 3*.

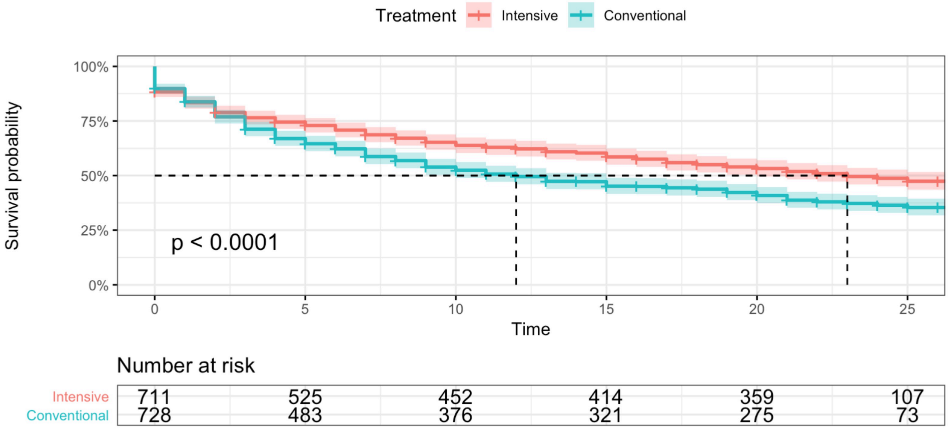


Figure 3: Kaplan-Meier Survival curve for AER > 30mg/24 hours

From the graph we notice that although the survival probabilities decrease in both groups, the slope in conventional group is steeper, which indicates that the survival probability drops slower in the intensive group.

By using log-rank test, with a p-value less than 0.0001, we have strong evidence to say that the survival distribution between treatments is significantly different. In other words, the intensive therapy yields more renal benefits than the conventional therapy.

Similarly, the Kaplan-Meier survival curve for eGFR (<60 mg) was also drawn and the significance of different was also computed using log-rank test, as in *Figure 4*.

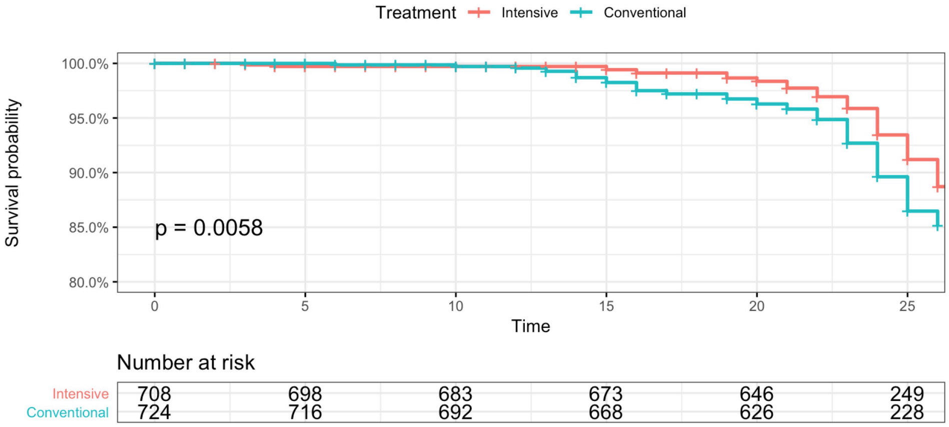


Figure 4: Kaplan-Meier Survival curve for eGFR < 60 mL/min per 1.73m^2

The results here are similar as in study of AER, as the slope in conventional group is also steeper, and the p-value for log-rank test is 0.0058, which strongly indicates that the intensive therapy yields more renal benefits than the conventional therapy.

Furthermore, the difference between two groups does not look significantly during the first 10 years, which is the period of DCCT. But the gap between two curve became bigger after the year of 15, which means the renal benefit of intensive therapy in terms of eGFR does not appear until the EDIC study.

5 Discussion

By using survival analysis we found that the intensive therapy has significantly more renal benefits than the conventional therapy through the DCCT and EDIC study. This result agrees with what has been found only during the DCCT study [3], as well as a recent research over the 30-year follow-up [9] while they use 40mg as a threshold of microalbuminuria.

As I'm currently analyzing the effect of single nucleotide polymorphisms (SNPs) on renal functions of the same group of patients using linear mixed models, and about to conduct a genome-wide association study (GWAS) under the supervision of Dr. Andrew Paterson as the Hospital of Sick Children, using survival analysis to compare the effect of SNPs, instead of effect of treatment as in this study, might be an alternative way to check the results produced by the joint models.

References

[1] The DCCT Research Group The diabetes control and complications trial (DCCT): design and methodologic considerations for the feasibility phase. Diabetes. 1986;35:530-545. doi: 10.2337/diab.35.5.530.

[2] Epidemiology of Diabetes Interventions and Complications (EDIC) Research Group. Epidemiology of Diabetes Interventions and Complications (EDIC): Design and implementation of a long-term follow-up of the Diabetes Control and Complications Trial cohort. Diabetes Care 1999;22:99-111

[3] The Diabetes Control and Complications Trial Research Group (The DCCT Trial lists) . The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. N Engl J Med 1993;329: 977-986.

[4] Therneau T (2015). A Package for Survival Analysis in S. version 2.38, <URL: https://CRAN.R-project.org/package=survival>. Terry M. Therneau, Patricia M. Grambsch (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York. ISBN 0-387-98784-3.

[5] Alboukadel Kassambara and Marcin Kosinski (2018). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.3. https://CRAN.R-project.org/package=survminer

[6] Keown-Stoneman, C. (2018). Introduction, the TARGet Kids Registry, and Previous Research in Survival Analysis [PowerPoint slides].

[7] Gehan A. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika 1965 Jun. 52(1/2):203-23.

[8] ROBERT E. TARONE, JAMES WARE; On distribution-free tests for equality of survival distributions, Biometrika, Volume 64, Issue 1, 1 April 1977, Pages 156–160, https://doi.org/10.1093/biomet/64.1.156

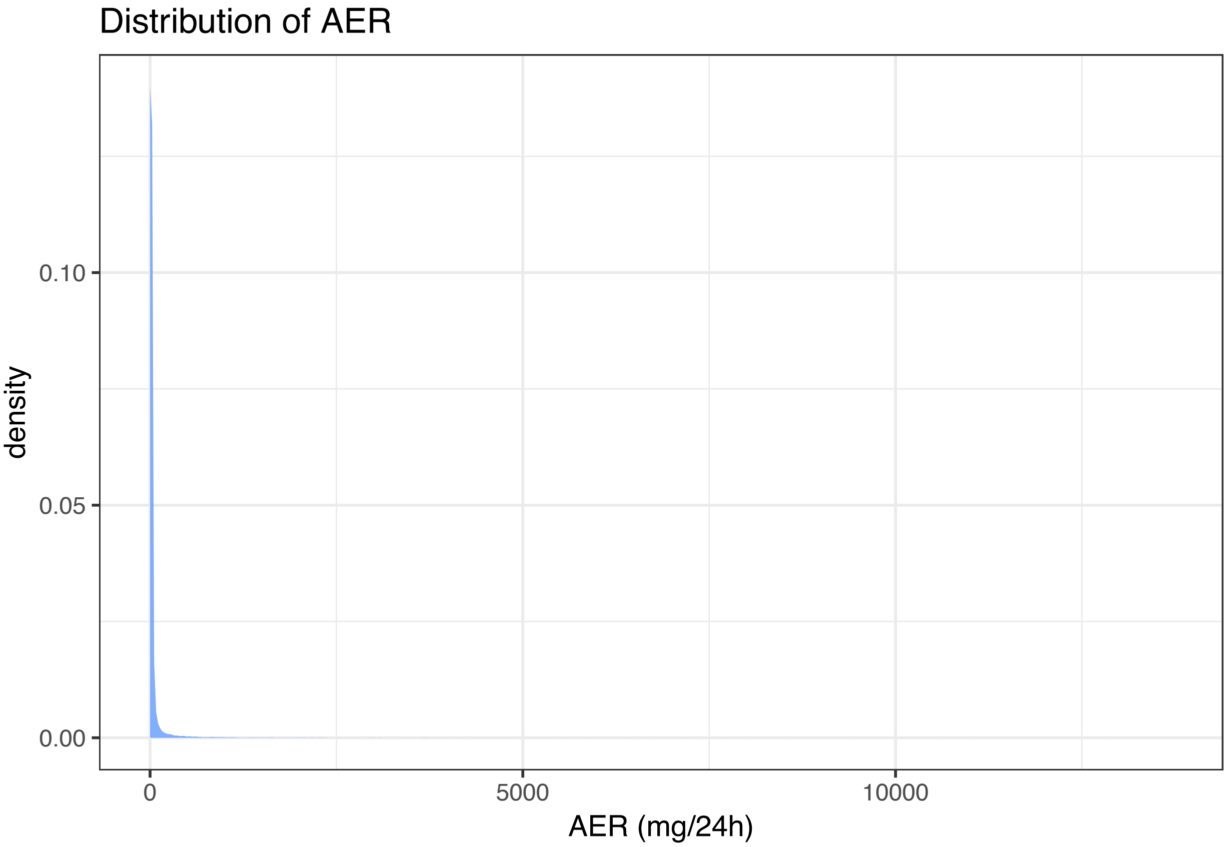
[9] de Boer IH. Kidney disease and related findings in the diabetes control and complications trial/epidemiology of diabetes interventions and complications study. Diabetes Care 2014; 37: 24–30

Appendix - R codes and outputs

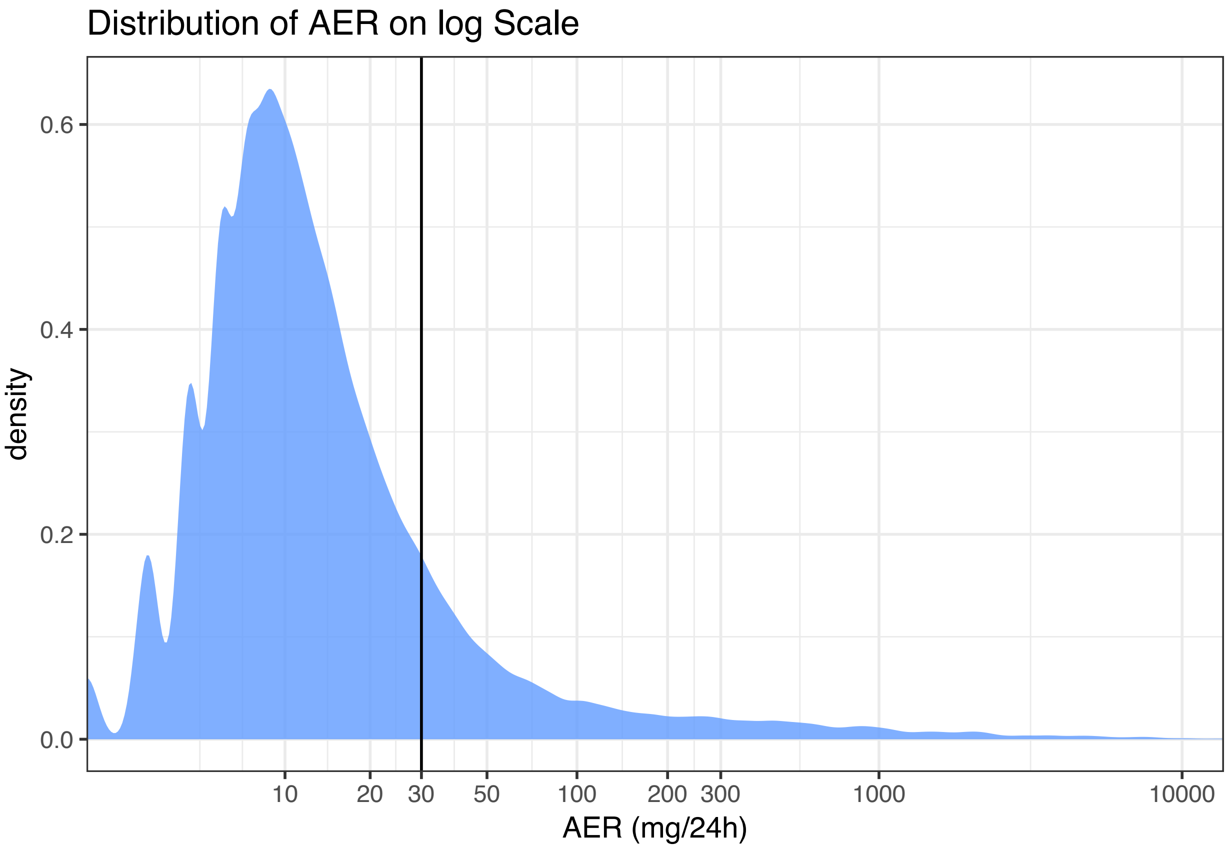
```
library(haven)
library(Hmisc)
library(ggplot2)
library(survival)
library(survminer)
library(scales)

setwd("..")
setwd("..")
setwd("..")
DCCT <- read_sas("Desktop/DCCT.sas7bdat")

### data cleaning

# remove the duplicate records of last DCCT visit
DCCT <- DCCT[DCCT$DTEDYEAR != 99,]

# creat a new variable for the time in years since the beginning of DCCT
DCCT$DCCTYEAR <- 0
# divide the quarter by 4 to get years during DCCT
DCCT$DCCTYEAR <- DCCT[DCCT$DCCTQTR <  100,]$DCCTQTR/4
# divide the DTED time by 100 and add the last visit of DCCT to get year during EDIC
for (i in 1:1441){
  if (sum(DCCT[DCCT$MASK_PAT == i,]$DTEDYEAR >= 100) != 0){
    DCCTdur <- max(DCCT[DCCT$MASK_PAT == i,]$DCCTYEAR, na.rm=T)
    rowsi <- nrow(DCCT[DCCT$MASK_PAT == i,]) # number of observation for patient i
    # get years during EDIC for patient i
    EDICdur <- DCCT[DCCT$MASK_PAT == i,]$EDICYEAR + DCCTdur
    DCCTobs <- sum(is.na(DCCT[DCCT$MASK_PAT == i,]$EDICYEAR)) # number of observation during DCCT
    DCCT[DCCT$MASK_PAT == i,][(DCCTobs+1):rowsi,]$DCCTYEAR <- EDICdur[!is.na(EDICdur)]
  }
}

### variables check
# binary variables
DCCT$CONV <- as.factor(DCCT$CONV)
DCCT <- DCCT[,c("DCCTYEAR","AER","CKD_GFR","CONV","MASK_PAT")]
str(DCCT)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    33149 obs. of  5 variables:
##  $ DCCTYEAR: num  0 1 2 3 4 5 6 7 8 9 ...
##  $ AER     : num  18.72 8.64 8.64 15.84 8.64 ...
##  $ CKD_GFR : num  139 130 137 128 135 ...
##  $ CONV    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ MASK_PAT: num  48 48 48 48 48 48 48 48 48 48 ...
```

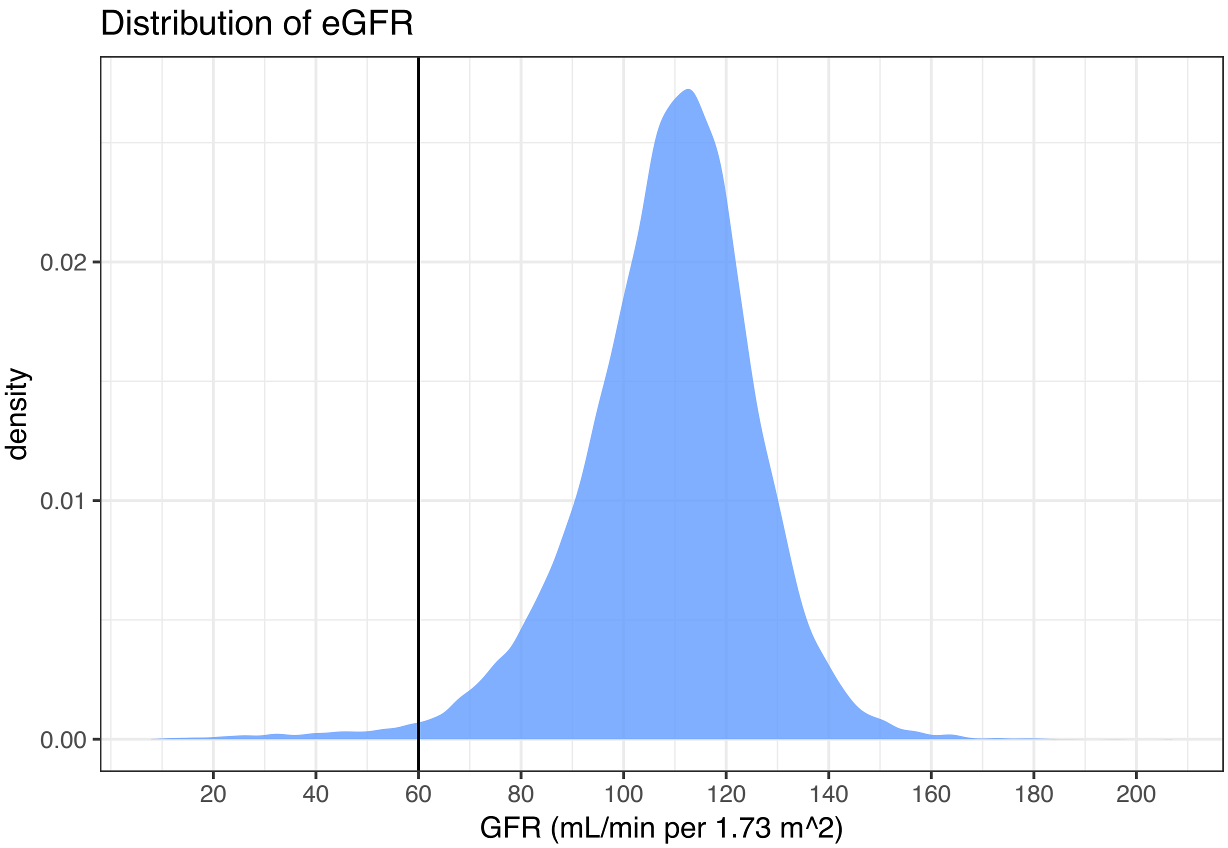
```
# plot histogram for continuous variables
# histogram for AER
ggplot(DCCT, aes(x=AER)) +
  ggtitle("Distribution of AER") +
  theme(plot.title = element_text(size=30)) +
  xlab("AER (mg/24h)") +
  stat_density(fill="#619CFF", alpha=0.8) +
  theme_bw()
```



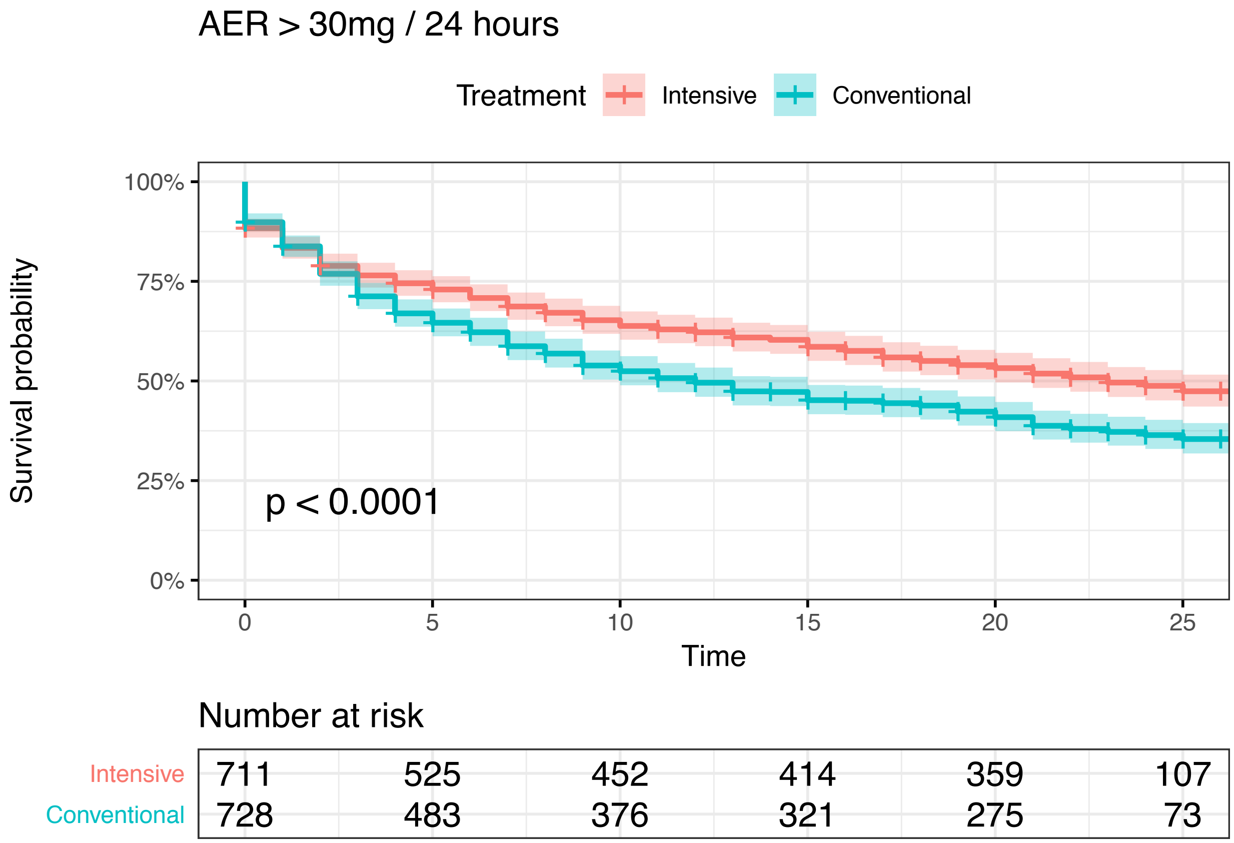
```
theme_update(plot.title = element_text(hjust = 0.5))
# histogram of eGFR on log scale
ggplot(DCCT, aes(x=AER)) +
  ggtitle("Distribution of AER on log Scale") +
  theme(plot.title = element_text(size=30)) +
  xlab("AER (mg/24h)") +
  stat_density(fill="#619CFF", alpha=0.8) +
  scale_x_continuous(breaks=c(0,10,20,30,50,100,200, 300, 1000, 10000), trans="log1p", expand=c(0,0)) +
  geom_vline(xintercept = 30) +
  theme_bw()
```



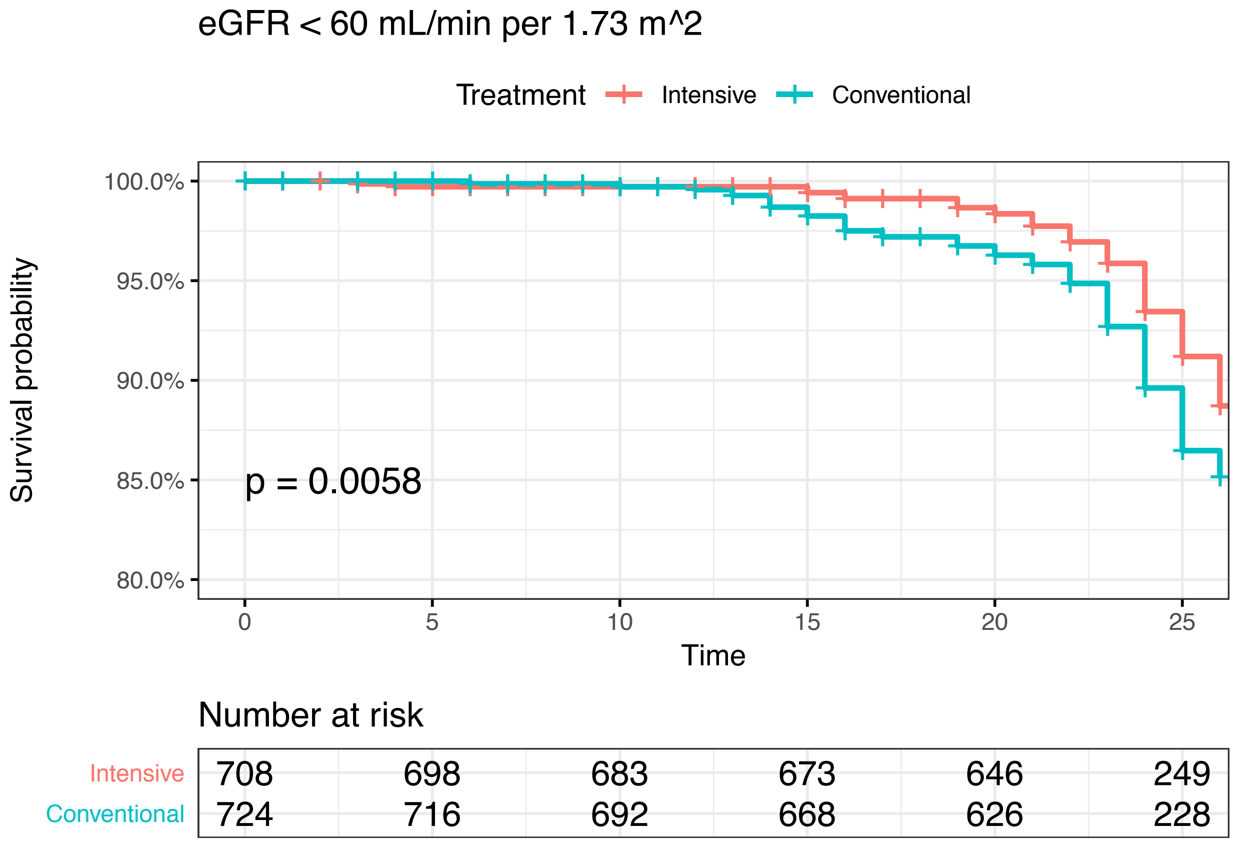
```
# histogram of eGFR
ggplot(DCCT[DCCT$CKD_GFR != 10,], aes(x=CKD_GFR)) +
  ggtitle("Distribution of eGFR") +
  theme(plot.title = element_text(size=30)) +
  xlab("GFR (mL/min per 1.73 m^2)") +
  stat_density(fill="#619CFF", alpha=0.8) +
  scale_x_continuous(breaks=seq(20,200,20)) +
  geom_vline(xintercept = 60) +
  theme_bw()
```



```
### Survival Analysis for AER30 against treatment

# generate a dataset that contains the years until a patient have AER greater than 30
MASK_ID <- c(1:1441)
SURTIME <- rep(0, 1441)
STATUS  <- rep(0, 1441)
TMT     <- rep(0, 1441) # 0 - intensive group
for (i in 1:1441){
  tempAER <- DCCT[DCCT$MASK_PAT == i,]$AER
  if (max(tempAER, na.rm = T) < 30){
    SURTIME[i] <- max(DCCT[DCCT$MASK_PAT == i,]$DCCTYEAR)
  }
  else{
    geq30 <- min(which(tempAER >= 30)) # The first time AER is greater than 30
    SURTIME[i] <- DCCT[DCCT$MASK_PAT == i,]$DCCTYEAR[geq30]
    STATUS[i]  <- 1
  }
  if (DCCT[DCCT$MASK_PAT == i,]$CONV == 1){TMT[i] <- 1}
}

AER30df <- as.data.frame(cbind(MASK_ID, SURTIME, STATUS, TMT))
AER30df$TMT <- factor(AER30df$TMT,
                     levels = c("0", "1"),
                     labels = c("Intensive", "Conventional"))
km <- survfit(Surv(SURTIME, STATUS) ~ TMT, data=AER30df)
ggsurvplot(km, data = AER30df,
           title = "AER > 30mg / 24 hours",
           # Change legends: title & labels
           legend.title = "Treatment",
           legend.labs = c("Intensive", "Conventional"),
           # Add p-value and tervals
           pval = TRUE,
           conf.int = TRUE,
           # Add risk table
           risk.table = TRUE,
           tables.height = 0.2,
           tables.theme = theme_cleantable(),
           surv.scale=c("percent"),
           ggtheme = theme_bw() # Change ggplot2 theme
)
```



```
### Survival Analysis for eGFR60 against treatment

# generate a dataset that contains the years until a patient have eGFR less 60
MASK_ID <- c(1:1441)
SURTIME <- rep(0, 1441)
STATUS  <- rep(0, 1441)
TMT     <- rep(0, 1441) # 0 - intensive group
for (i in 1:1441){
  tempGFR <- DCCT[DCCT$MASK_PAT == i,]$CKD_GFR
  if (min(tempGFR, na.rm = T) > 60){
    SURTIME[i] <- max(DCCT[DCCT$MASK_PAT == i,]$DCCTYEAR)
  }
  else{
    leq60 <- max(which(tempGFR <= 60)) # The first time eGFR is less than 60
    SURTIME[i] <- DCCT[DCCT$MASK_PAT == i,]$DCCTYEAR[leq60]
    STATUS[i]  <- 1
  }
  if (DCCT[DCCT$MASK_PAT == i,]$CONV == 1){TMT[i] <- 1}
}

GFR60df <- as.data.frame(cbind(MASK_ID, SURTIME, STATUS, TMT))
GFR60df$TMT <- factor(GFR60df$TMT,
                     levels = c("0", "1"),
                     labels = c("Intensive", "Conventional"))
km2 <- survfit(Surv(SURTIME, STATUS) ~ TMT, data=GFR60df)
ggsurvplot(km2, data = GFR60df,
           title = "eGFR < 60 mL/min per 1.73 m^2",
           ylim = c(0.8,1),
           # Change legends: title & labels
           legend.title = "Treatment",
           legend.labs = c("Intensive", "Conventional"),
           pval = TRUE,
           pval.coord = c(0, 0.85),
           # Add risk table
           risk.table = TRUE,
           tables.height = 0.2,
           tables.theme = theme_cleantable(),
           surv.scale=c("percent"),
           ggtheme = theme_bw() # Change ggplot2 theme
)
```



Intensive	708	698	683	673	646	249
Conventional	724	716	692	668	626	228