Survival Analysis on the Long Term Survival Following Hospitalization for an Acute Myocardial Infarction Using Cox Proportional Hazard and Accelerated Time Failure Models

Shune Kawaoto

May 15th, 2025

Abstract

This paper looks at long-term survival following hospitalization for acute myocardial infarction (AMI), using data from the Worcester Heart Attack Study (WHAS). The analysis focuses on how survival is influenced by age, sex, cardiogenic shock, and congestive heart failure. Both Cox proportional hazards and accelerated failure time models are used to evaluate the data. Model assumptions are checked for each approach, and results are interpreted in terms of clinical relevance.

1 Introduction

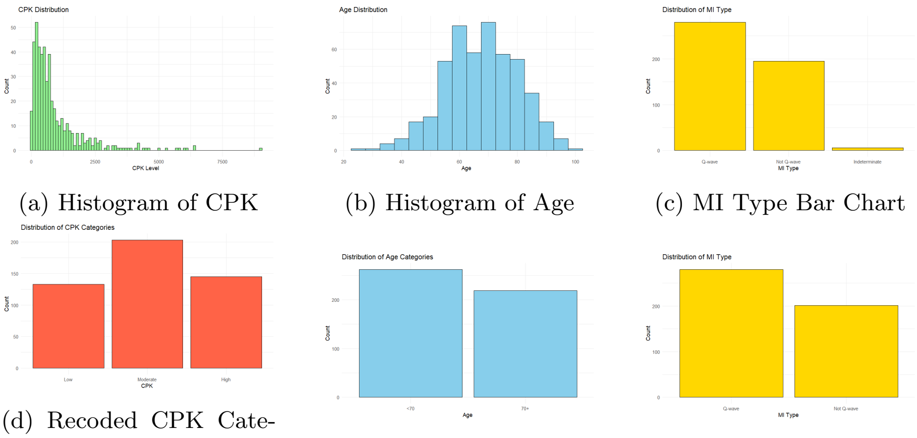
Acute myocardial infarction (AMI) is one of the leading causes of death, especially among older adults. Understanding which factors influence long-term survival after AMI is important for improving patient care and outcomes. This study uses data from the Worcester Heart Attack Study (WHAS) to examine how survival is affected by patient age, sex, presence of cardiogenic shock, congestive heart failure (CHF), MI order, and MI type. Survival analysis methods—both semi-parametric (Cox PH) and parametric (AFT)—are applied to evaluate these effects, compare modeling approaches, and identify which predictors are most strongly tied to risk.

2 Model Development

2.1 Pre-processing

The WHAS dataset contains demographic, clinical, and follow-up survival data for patients hospitalized due to acute myocardial infarction (AMI). To begin, exploratory data analysis was performed on all of the variables of interest. Continuous variables included `age`, creatine phosphokinase (`cpk`), and follow-up time. These were examined using histograms. Follow-up time is also the response variable in the analysis. Categorical variables such as `sex`, cardiogenic shock (`sho`), congestive heart failure (`chf`), myocardial infarction order (`miord`), and myocardial infarction type (`mitype`) were summarized using frequency tables.

No missing values required imputation. However, for the purpose of making interpretations of future models clearer, the two continuous explanatory variables were made into discrete ones. Age was grouped into a binary category: patients younger than 70 and those 70 and older. CPK was grouped into low, moderate, and high levels using clinical cutoffs. Additionally, the `mitype` variable originally included a small "Indeterminate" category; to improve model stability, this group was merged with the "Not Q-wave" category.



(a) Histogram of CPK

(b) Histogram of Age

(c) MI Type Bar Chart

(d) Recoded CPK Categories

(e) Recoded Age Groups

(f) Recoded MI Type

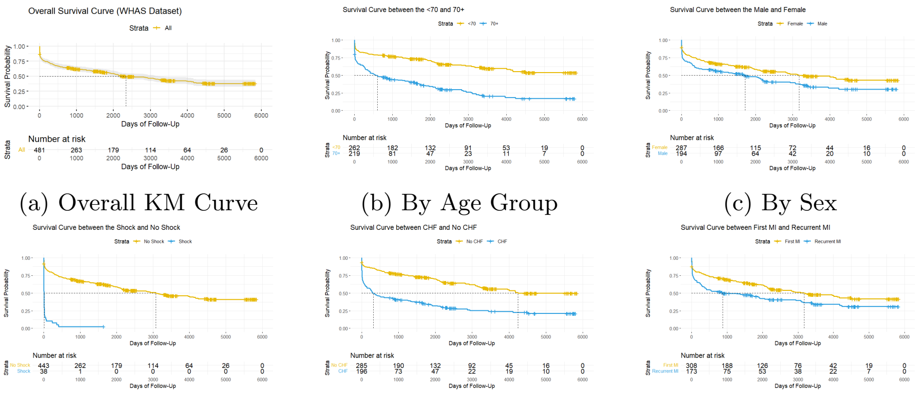
Figure 1: Exploratory histograms and categorical groupings for WHAS variables before and after recoding.

To narrow the pool of variables, individual log-rank tests were conducted for each covariate using a Cox model with a single predictor. These were compared against the null model using likelihood ratio tests, with significance determined at $\alpha = 0.05$.

Table 1: Log-likelihood Tests for Single Covariate Cox Models Compared to the Null Model

Variable	$\mathbf{-2 \log \hat{L}}$	**Chi-Square**	**df**	**p-value**
None	2839.997	–	–	–
`age_group`	2771.585	68.412	1	1.33×10^{-16}
`sex`	2830.803	9.194	1	0.00243
`cpk_group`	2838.196	1.801	2	0.406
`sho`	2739.873	100.124	1	1.43×10^{-23}
`chf`	2773.534	66.464	1	3.56×10^{-16}
`miord`	2828.180	11.817	1	0.000587
`mitype`	2838.447	1.550	1	0.213

As shown in Table 1, the significant univariate variables included `age_group`, `sex`, `sho`, `chf`, and `miord`. Their corresponding Kaplan-Meier survival curves are displayed in Figure 2. Age over 70, presence of shock, and having CHF were all associated with visibly worse survival. Differences by sex and MI order were not as aparent as the previous three variables, but CPK levels and MI type all had a greater p-value than $\alpha = 0.05$ so these were removed from any further analysis.

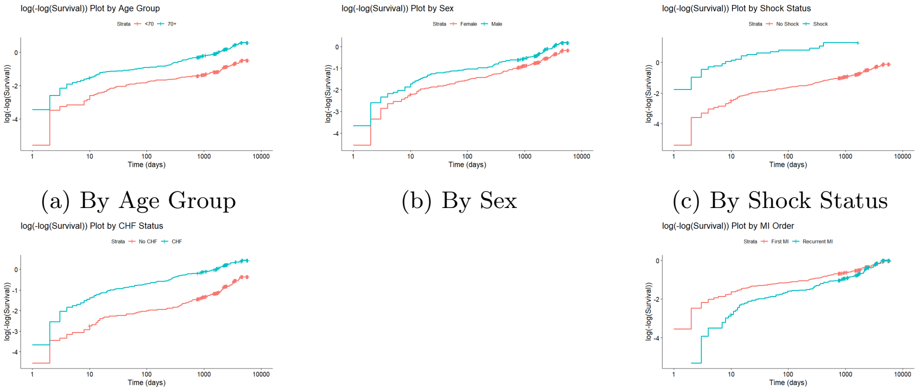


(a) Overall KM Curve (b) By Age Group (c) By Sex

(d) By Shock Status (e) By CHF Status (f) By MI Order

Figure 2: Kaplan-Meier survival curves by clinical and demographic variables.

To verify that the Cox proportional hazards model was appropriate, log-log survival plots for each variable was constructed. If the log(-log(S(t))) curves are visibly and clearly parallel, then this indicates whether the proportional hazards (PH) assumption is reasonable.



(a) By Age Group (b) By Sex (c) By Shock Status

(d) By CHF Status (e) By MI Order

Figure 3: Log(-log) survival curves used to assess the proportional hazards assumption.

Variables such as `age_group`, `sho`, and `chf` showed roughly parallel log-log curves. `Sex` also demonstrated acceptable parallelism, especially if focused on long-term survival. However, `miord` showed clear violations of the PH assumption due to the curves converging over time. Because of this, `miord` was excluded altogether from further multivariate modeling, in both Cox proportional hazard models and accelerated time failure models.

2.2 Cox Proportional Hazards Model

To identify the best set of predictors for modeling time to death, the following four primary covariates `age`, `sex`, `sho` (shock), and `chf` (congestive heart failure) were considered. For the variable selection process, the best-subset selection was initially applied by fitting all possible combinations of the candidate variables and ranking them by log-likelihood test with a χ^2 critical value of 3.84 (corresponding to $\alpha = 0.05$). Then, forward stepwise selection was used with the likelihood ratio test again being the selection criteria to determine whether each added variable provided a statistically meaningful improvement in model fit.

The variable selection process identified the model containing `age`, `sho`, and `chf` as optimal, while also achieving the lowest AIC among all combinations. This result was supported by the stepwise procedure, which added `sho` first, followed by `age`, and finally `chf`, with no other variables improving the model beyond the threshold.

Table 2: Final Cox Proportional Hazards Model Summary

Variable	Log-Likelihood	$-2\log(\hat{L})$	AIC
Intercept Only (null model)	-1419.999	2839.997	2839.997
age + sho	-1343.529	2687.058	2691.058
age + sho + chf (final model)	-1334.640	2669.280	2675.280

The final model is expressed as:

$$\lambda(t \mid x) = \lambda_0(t) \exp\left(\beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{SHO} + \beta_3 \cdot \text{CHF}\right)$$

Before finalizing the model, the proportional hazards assumption was previously verified for each variable by inspecting complementary log-log (log(-log(S(t)))) plots. The curves for `age`, `sho`, and `chf` appeared approximately parallel, supporting their inclusion in the Cox framework. One candidate variable, `miord` (myocardial infarction order), violated the proportional hazards assumption due to diverging curves. Although stratification could have addressed this issue, `miord` was entirely removed in order to maintain consistency with parametric AFT models, where stratification is not easily implemented or interpreted.

Given the strong fit of variables and satisfaction of model assumptions, the Cox proportional hazards model with `age`, `sho`, and `chf` provides a statistically and clinically meaningful framework to determine survival time after experiencing AMI. This model will also compared to an accelerated time failure model in the subsequent sections.

2.3 Accelerated Failure Time Model

To complement the semi-parametric Cox model, fully parametric survival models were fit under the Accelerated Failure Time (AFT) framework. Unlike Cox models, AFT models describe how covariates affect the survival time directly, rather than the hazard rate. Three nested distributions were fitted:

- The **exponential model** assumes a constant hazard rate over time.
- The **Weibull model** generalizes the exponential by allowing monotonic hazard trends.
- The **generalized gamma model** allows the hazard function to be non-monotonic and accommodates skewed survival distributions.

To determine the most appropriate distribution, we used likelihood ratio tests (LRTs) to compare models in a nested framework. The LRT statistic is calculated as:

$$\text{LRT} = 2(\log \hat{L}_{\text{full}} - \log \hat{L}_{\text{reduced}})$$

where $\hat{L}$ is the log-likelihood and the degrees of freedom reflect the difference in model complexity.

Table 3: Likelihood Ratio Tests Comparing AFT Models

Comparison	LRT Statistic	df	p-value
Weibull vs Exponential	326.33	1	< 0.0001
Generalized Gamma vs Weibull	6.82	1	0.009
Generalized Gamma vs Exponential	333.15	2	< 0.0001

In each likelihood ratio test, the reduced model is determined by the distribution that is nested in the others. For example, when comparing exponential to Weibull, exponential is ”nested” inside Weibull so it is considered to be the reduced model in this case. Similarly, Weibull is nested inside the generalized Gamma distribution, so the Weibull will be the reduced model in that case. And finally, the exponential is treated as the reduced model when compared to the generalized Gamma.

Each likelihood ratio test produced statistically significant results, indicating improvement in fit. While the jump from exponential to Weibull was substantial, the improvement from Weibull to generalized gamma was more modest but still statistically meaningful ($p = 0.009$). Therefore, the generalized gamma distribution was selected as the final AFT model.

The final model takes the following log-linear form:

$$\log(T) = \beta_0 + \beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{SHO} + \beta_3 \cdot \text{CHF} + \sigma \cdot \epsilon$$

where T is survival time and ϵ is a random error term following a generalized gamma distribution. The $\beta's$ and σ are parameters to be estimated.

While the Cox model estimates hazard ratios (HR), the AFT model estimates time ratios (TR). A time ratio less than 1 implies shorter survival time compared to the reference group, offering an intuitive interpretation on the time scale:

- Patients with `shock` had a time ratio of 0.351, meaning their expected survival time was 64.9% shorter than those without shock.
- Patients with `CHF` had a time ratio of 0.597, corresponding to a 40.3% reduction in survival time.
- Each additional year of `age` was associated with a 3.7% decrease in expected survival time (TR = 0.963).

The consistency between HRs and TRs shows the reliability of these findings. While Cox models emphasize instantaneous risk, AFT models work in terms of the overall acceleration or deceleration of time to event, offering an alternative and complementary perspective. The fact that the same variables emerged as significant across both parametric and semi-parametric methods further validates the performance of the selected model.

3 Results

Table 4: Comparison of Cox and Generalized Gamma AFT Model Estimates

Variable	HR (Cox)	95% CI	Time Ratio (AFT)	95% CI	p-value
`age`	1.034	(1.022, 1.045)	0.963	(0.951, 0.974)	< 0.001
`shoShock`	6.717	(4.503, 10.020)	0.351	(0.270, 0.457)	< 0.001
`chfCHF`	1.825	(1.381, 2.412)	0.597	(0.484, 0.737)	< 0.001

Table 4 presents a side-by-side comparison of the estimates from the Cox proportional hazards model and the Generalized Gamma AFT model for the three selected covariates: `age`, `sho`, and `chf`.

Across both models, all covariates were found to be statistically significant with $p < 0.001$. In the Cox model, each additional year of age was associated with a 3.4% increase in the hazard of death (HR = 1.034, 95% CI: 1.022–1.045), while the AFT model estimated a time ratio of 0.963, indicating a 3.7% decrease in expected survival time per year. These interpretations seem consistent: as age increases, patients are both more likely to die sooner and more quickly.

For `sho`, patients who experienced cardiogenic shock had a hazard 6.72 times greater than those who did not (95% CI: 4.50–10.02). The AFT model revealed a time ratio of 0.351, implying that shock patients survived only 35.1% as long as those without shock.

Similarly, those with `chf` had a hazard ratio of 1.825 (95% CI: 1.381–2.412), meaning an 82.5% higher risk of death at any point in time. The corresponding time ratio of 0.597 in the AFT model suggests a 40.3% reduction in survival time for CHF patients.

The results of each model support each other: variables associated with higher hazards in the Cox model correspond to lower time ratios in the AFT model. This clearly indicates that older age, cardiogenic shock, and left heart failure complications are strong and consistent factors for determining survival following acute myocardial infarction.

4 Conclusion

This analysis evaluated long-term survival following acute myocardial infarction using data from the Worcester Heart Attack Study. Two distinct modeling approaches were applied: the semi-parametric Cox proportional hazards model and the parametric accelerated failure time model. Both used and determined the same three variables of age, cardiogenic shock, and left heart failure complications, as significant predictors of mortality.

Personally, the most important takeaway is the impact of cardiogenic shock: patients who experienced shock had dramatically reduced survival prospects. Heart failure was also a strong negative prognostic factor. Age, while less extreme, consistently reduced survival both by increasing hazard and shortening expected lifespan. I expected Age to be the most obvious significant variable, but I was not expecting the impacts of cardiogenic shock and heart failure to be severe to this extent.

What stands out is the agreement between the Cox and AFT models. Despite their different assumptions and interpretations, they were consistent in results and interpretations. This consistency provides reliable confidence in the findings. The models not only fit the data well but also support intuitive results, such as, older patients and those with more severe complications tend to do worse in surviving after acute myocardial infarction.

5 References

- Hosmer, D.W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* Wiley.
- WHAS data courtesy of Dr. Robert J. Goldberg, University of Massachusetts Medical School.

6 Appendix (R code)

```
library(survival)
# install.packages("survminer")
library(survminer)
library(dplyr)
library(ggplot2)
library(readr)
library(corrplot)
library(tidyverse)
# install.packages("flexsurv")
library(flexsurv)
library(broom)

whas <- read_table2("whas.dat", col_names = FALSE)

colnames(whas) <- c(
  "id", "age", "sex", "cpk", "sho", "chf",
  "miord", "mitype", "year", "yrgrp",
  "lenstay", "dstat", "lenfol", "fstat"
)

whas <- whas[, 1:14]

colnames(whas) <- c(
  "id", "age", "sex", "cpk", "sho", "chf",
  "miord", "mitype", "year", "yrgrp",
  "lenstay", "dstat", "lenfol", "fstat"
)


whas <- whas %>%
  mutate(
    sex = factor(sex, levels = c(0,1), labels = c("Male", "Female")),
    sho = factor(sho, levels = c(0,1), labels = c("No Shock", "Shock")),
    chf = factor(chf, levels = c(0,1), labels = c("No CHF", "CHF")),
    miord = factor(miord, levels = c(0,1), labels = c("First MI", "Recurrent MI")),
    mitype = factor(mitype, levels = c(1,2,3), labels = c("Q-wave", "Not Q-wave", "Indetermi
  )


# Missing Values

missing_summary <- sapply(whas, function(x) sum(is.na(x)) / length(x) * 100)
missing_summary


# Summary Statistics


summary(whas)
```

```
whas %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
    median_cpk = median(cpk, na.rm = TRUE),
    mean_lenfol = mean(lenfol, na.rm = TRUE),
    death_rate = mean(fstat, na.rm = TRUE)
  )

table(whas$sex)
table(whas$sho)
table(whas$chf)
table(whas$miord)
table(whas$mitype)

# To check the survival curves between age groups and cpk groups (put them into groups) but


# Histograms
# Age
ggplot(whas, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Age Distribution", x = "Age", y = "Count")

# CPK
ggplot(whas, aes(x = cpk)) +
  geom_histogram(binwidth = 100, fill = "lightgreen", color = "black") +
  theme_minimal() +
  labs(title = "CPK Distribution", x = "CPK Level", y = "Count")

# Follow-up Time
ggplot(whas, aes(x = lenfol)) +
  geom_histogram(binwidth = 100, fill = "lightcoral", color = "black") +
  theme_minimal() +
  labs(title = "Follow-up Time Distribution", x = "Days", y = "Count")


# Bar Charts

whas$sex <- factor(whas$sex, levels = c("Male", "Female"))
whas$sho <- factor(whas$sho, levels = c("No Shock", "Shock"))
whas$chf <- factor(whas$chf, levels = c("No CHF", "CHF"))
whas$miord <- factor(whas$miord, levels = c("First MI", "Recurrent MI"))
whas$mitype <- factor(whas$mitype, labels = c("Q-wave", "Not Q-wave", "Indeterminate"))
```

```
# Bar Chart for SEX (Gender)
ggplot(whas, aes(x = sex)) +
  geom_bar(fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Gender", x = "Gender", y = "Count")

# Bar Chart for SHO (Cardiogenic Shock)
# Not balanced, so survival curve will not be included
ggplot(whas, aes(x = sho)) +
  geom_bar(fill = "tomato", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Cardiogenic Shock", x = "Shock Status", y = "Count")

# Bar Chart for CHF (Heart Failure)
ggplot(whas, aes(x = chf)) +
  geom_bar(fill = "lightgreen", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Congestive Heart Failure", x = "CHF Status", y = "Count")

# Bar Chart for MIORD (MI Order)
ggplot(whas, aes(x = miord)) +
  geom_bar(fill = "orchid", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of MI Order", x = "MI Order", y = "Count")

# Bar Chart for MITYPE (Type of MI)
ggplot(whas, aes(x = mitype)) +
  geom_bar(fill = "gold", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of MI Type", x = "MI Type", y = "Count")


# Creating Categories for Age and CPK

# Age
whas$age_group <- cut(
  whas$age,
  breaks = c(0, 70, Inf),
  right = FALSE,
  labels = c("<70", "70+")
)

# CPK
whas$cpk_group <- cut(
  whas$cpk,
  breaks = c(0, 300, 900, Inf),
  right = FALSE,
  labels = c("Low", "Moderate", "High")
)

# Merging Categories in Mitype
levels(whas$mitype)
# [1] "Q-wave" "Not Q-wave" "Indeterminate"
# First convert to character to safely recode
whas$mitype <- as.character(whas$mitype)

# Merge "Indeterminate" into "Not Q-wave"
whas$mitype[whas$mitype == "Indeterminate"] <- "Not Q-wave"

# Convert back to factor
whas$mitype <- factor(whas$mitype, levels = c("Q-wave", "Not Q-wave"))


# mitype has very few in the "Indeterminate" category --> merge to "Not Q-wave"


# Replotting the New Variables

# Bar Chart for Age
ggplot(whas, aes(x = age_group)) +
  geom_bar(fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Age Categories", x = "Age", y = "Count")

# Bar Chart for CPK
ggplot(whas, aes(x = cpk_group)) +
  geom_bar(fill = "tomato", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of CPK Categories", x = "CPK", y = "Count")

# Bar Chart for MITYPE (Type of MI)

ggplot(whas, aes(x = mitype)) +
  geom_bar(fill = "gold", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of MI Type", x = "MI Type", y = "Count")


# Log Rank test to see which groups have significant survival differences.

# Variables to test
```

```
vars <- c("age_group", "sex", "cpk_group", "sho", "chf", "miord", "mitype")

# Fit null model
null_model <- coxph(Surv(lenfol, fstat) ~ 1, data = whas)
null_loglik <- logLik(null_model)[1]
null_neg2loglik <- -2 * null_loglik

# Container for results
lrt_results <- data.frame(
  Variable = character(),
  Neg2LogL = numeric(),
  ChiSq = numeric(),
  df = numeric(),
  p_value = numeric(),
  stringsAsFactors = FALSE
)

# Loop through each variable and compute test
for (v in vars) {
  formula <- as.formula(paste("Surv(lenfol, fstat) ~", v))
  model <- coxph(formula, data = whas)
  model_loglik <- logLik(model)[1]
  model_neg2loglik <- -2 * model_loglik

  chisq <- 2 * (model_loglik - null_loglik)
  df <- attr(logLik(model), "df") - attr(logLik(null_model), "df")
  p <- pchisq(chisq, df = df, lower.tail = FALSE)

  lrt_results <- rbind(lrt_results, data.frame(
    Variable = v,
    Neg2LogL = round(model_neg2loglik, 3),
    ChiSq = round(chisq, 3),
    df = df,
    p_value = signif(p, 3)
  ))
}

# Add the null model row
lrt_results <- rbind(
  data.frame(Variable = "None", Neg2LogL = round(null_neg2loglik, 3), ChiSq = NA, df = NA, p
  lrt_results
)


print(lrt_results)
```

```
# CPK and mitype are not significantly different in survival between their groups


# Kaplan Meier Survival Curves

surv_object <- Surv(time = whas$lenfol, event = whas$fstat)

fit_overall <- survfit(surv_object ~ 1, data = whas)
ggsurvplot(
  fit_overall,
  data = whas,
  conf.int = TRUE,
  risk.table = TRUE,
  surv.median.line = "hv",
  xlab = "Days of Follow-Up",
  ylab = "Survival Probability",
  title = "Overall Survival Curve (WHAS Dataset)",
  palette = c("#E7B800"),
  ggtheme = theme_minimal(base_size = 14)
)

fit_age <- survfit(surv_object ~ age_group, data = whas)
ggsurvplot(
  fit_age,
  data = whas,
  risk.table = TRUE,
  surv.median.line = "hv",
  legend.labs = c("<70", "70+"),
  xlab = "Days of Follow-Up",
  ylab = "Survival Probability",
  title = "Survival Curve between the <70 and 70+",
  palette = c("#E7B800", "#2E9FDF"),
  ggtheme = theme_minimal()
)

fit_sex <- survfit(surv_object ~ sex, data = whas)
ggsurvplot(
  fit_sex,
  data = whas,
  risk.table = TRUE,
  surv.median.line = "hv",
  legend.labs = c("Female", "Male"),
  xlab = "Days of Follow-Up",
  ylab = "Survival Probability",
  title = "Survival Curve between the Male and Female",
  palette = c("#E7B800", "#2E9FDF"),
  ggtheme = theme_minimal()
)


fit_sho <- survfit(surv_object ~ sho, data = whas)
ggsurvplot(
  fit_sho,
  data = whas,
  risk.table = TRUE,
  surv.median.line = "hv",
  legend.labs = c("No Shock", "Shock"),
  xlab = "Days of Follow-Up",
  ylab = "Survival Probability",
  title = "Survival Curve between the Shock and No Shock",
  palette = c("#E7B800", "#2E9FDF"),
  ggtheme = theme_minimal()
)


fit_chf <- survfit(surv_object ~ chf, data = whas)
ggsurvplot(
  fit_chf,
  data = whas,
  risk.table = TRUE,
  surv.median.line = "hv",
  legend.labs = c("No CHF", "CHF"),
  xlab = "Days of Follow-Up",
  ylab = "Survival Probability",
  title = "Survival Curve between CHF and No CHF",
  palette = c("#E7B800", "#2E9FDF"),
  ggtheme = theme_minimal()
)


fit_miord <- survfit(surv_object ~ miord, data = whas)
ggsurvplot(
  fit_miord,
  data = whas,
  risk.table = TRUE,
  surv.median.line = "hv",
  legend.labs = c("First MI", "Recurrent MI"),
  xlab = "Days of Follow-Up",
  ylab = "Survival Probability",
  title = "Survival Curve between First MI and Recurrent MI",
  palette = c("#E7B800", "#2E9FDF"),
  ggtheme = theme_minimal()
```

```
)

# Checking the Cox PH Model Assumption (if the log(-log(survival)) are parallel then assumpt

# By sho
fit_sho <- survfit(Surv(lenfol, fstat) ˜ sho, data = whas)
ggsurvplot(
  fit_sho,
  fun = "cloglog",
  title = "log(-log(Survival)) Plot by Shock Status",
  xlab = "Time (days)",
  ylab = "log(-log(Survival))",
  risk.table = FALSE,
  legend.labs = c("No Shock", "Shock")
)

# By CHF
fit_chf <- survfit(Surv(lenfol, fstat) ˜ chf, data = whas)
ggsurvplot(
  fit_chf,
  fun = "cloglog",
  title = "log(-log(Survival)) Plot by CHF Status",
  xlab = "Time (days)",
  ylab = "log(-log(Survival))",
  risk.table = FALSE,
  legend.labs = c("No CHF", "CHF")
)

# By MI Order
fit_miord <- survfit(Surv(lenfol, fstat) ˜ miord, data = whas)
ggsurvplot(
  fit_mitype,
  fun = "cloglog",
  title = "log(-log(Survival)) Plot by MI Order",
  xlab = "Time (days)",
  ylab = "log(-log(Survival))",
  risk.table = FALSE,
  legend.labs = c("First MI", "Recurrent MI")
)

# By Age Group
fit_age <- survfit(Surv(lenfol, fstat) ˜ age_group, data = whas)
ggsurvplot(
  fit_age,
  fun = "cloglog",
  title = "log(-log(Survival)) Plot by Age Group",
  xlab = "Time (days)",
  ylab = "log(-log(Survival))",
  risk.table = FALSE,
  legend.labs = c("<70", "70+")
)

# By Sex
fit_age <- survfit(Surv(lenfol, fstat) ~ sex, data = whas)
ggsurvplot(
  fit_age,
  fun = "cloglog",
  title = "log(-log(Survival)) Plot by Sex",
  xlab = "Time (days)",
  ylab = "log(-log(Survival))",
  risk.table = FALSE,
  legend.labs = c("Female", "Male")
)

# Everything except miord meets assumptions, now run Cox PH Models and do Variable Selection

# Variable Selection
vars <- c("age", "sex", "sho", "chf")

model_results <- list()
counter <- 1

null_model <- coxph(Surv(lenfol, fstat) ~ 1, data = whas)
model_results[[counter]] <- list(
  model = "Intercept Only",
  n_vars = 0,
  logLik = logLik(null_model)[1],
  neg2logLik = -2 * logLik(null_model)[1],
  AIC = AIC(null_model)
)
counter <- counter + 1

for (i in 1:length(vars)) {
  combos <- combn(vars, i, simplify = FALSE)
  for (combo in combos) {
    # Build formula
    formula <- as.formula(paste("Surv(lenfol, fstat) ~", paste(combo, collapse = " + ")))

    # Fit Cox model
    fit <- coxph(formula, data = whas)

    # Save model info
    model_results[[counter]] <- list(
      model = paste(combo, collapse = " + "),
      n_vars = length(combo),
      logLik = logLik(fit)[1],
      neg2logLik = -2 * logLik(fit)[1],
      AIC = AIC(fit)
    )
    counter <- counter + 1
  }
}

model_table <- do.call(rbind, lapply(model_results, as.data.frame))
model_table <- model_table %>% arrange(n_vars, neg2logLik)
model_table

# Stepwise selection based on -2LL improvement
stepwise_ll_select <- function(data, time, status, candidates, threshold = 3.84) {
  selected <- c()         # Final selected variables
  remaining <- candidates # Variables to consider

  # Start with null model
  null_model <- coxph(as.formula(paste("Surv(", time, ",", status, ") ~ 1")), data = data)
  base_ll <- logLik(null_model)[1]
  base_formula <- "1"

  cat("Starting Stepwise Selection\n")
  cat("--------------------------------\n")
  cat("Base model: Surv(", time, ",", status, ") ~ 1\n")
  cat("LogLik:", round(base_ll, 3), "\n\n")

  while (length(remaining) > 0) {
    best_var <- NULL
    best_ll <- base_ll
    best_chisq <- 0
    best_model <- NULL

    for (var in remaining) {
      test_formula <- as.formula(paste("Surv(", time, ",", status, ") ~", paste(c(selected,
      test_model <- coxph(test_formula, data = data)
      test_ll <- logLik(test_model)[1]
      chisq <- 2 * (test_ll - base_ll)

      if (chisq > best_chisq) {
        best_var <- var
        best_ll <- test_ll
        best_chisq <- chisq
        best_model <- test_model
      }
    }

    if (!is.null(best_var) && best_chisq >= threshold) {
      cat("Adding:", best_var, "| Chi-square:", round(best_chisq, 3), "\n")
      selected <- c(selected, best_var)
      remaining <- setdiff(remaining, best_var)
      base_ll <- best_ll
    } else {
      cat("No remaining variable exceeds chi-square threshold. Stopping.\n")
      break
    }
  }

  final_formula <- as.formula(paste("Surv(", time, ",", status, ") ~", paste(selected, colla
  final_model <- coxph(final_formula, data = data)

  cat("\nFinal model:\n")
  print(final_model$call)

  return(list(model = final_model, variables = selected))
}

# Run stepwise selection using -2LL and chi-square > 3.84
result <- stepwise_ll_select(data = whas,
                              time = "lenfol",
                              status = "fstat",
                              candidates = vars,
                              threshold = 3.84)

summary(result$model)

-2 * logLik(result$model)[1]

# age + sho + chf has the lowest AIC


# AFT Models

tidy_cox <- broom::tidy(result$model, exponentiate = TRUE, conf.int = TRUE) %>%
  rename(
    Variable = term,
    HR = estimate,
    HR_lower = conf.low,
    HR_upper = conf.high,
    p_value_cox = p.value
  ) %>%
  select(Variable, HR, HR_lower, HR_upper, p_value_cox) %>%
  mutate(across(where(is.numeric), ~ round(., 3)))

weibull_model <- survreg(Surv(lenfol, fstat) ~ age + sho + chf, data = whas, dist = "weibull

exp_model <- survreg(Surv(lenfol, fstat) ~ age + sho + chf, data = whas, dist = "exponential

gen_gamma_model <- flexsurvreg(Surv(lenfol, fstat) ~ age + sho + chf, data = whas, dist = "g

# Use Likelihood Ratio Test to compare the best
lrt_weibull_vs_exp <- 2 * (logLik(weibull_model)[1] - logLik(exp_model)[1])
lrt_gamma_vs_weibull <- 2 * (logLik(gen_gamma_model)[1] - logLik(weibull_model)[1])
lrt_gamma_vs_exp <- 2 * (logLik(gen_gamma_model)[1] - logLik(exp_model)[1])

lrt_table <- data.frame(
  Comparison = c("Weibull vs Exponential",
                 "Generalized Gamma vs Weibull",
                 "Generalized Gamma vs Exponential"),
  LRT_Statistic = c(lrt_weibull_vs_exp,
                    lrt_gamma_vs_weibull,
                    lrt_gamma_vs_exp),
  df = c(1, 1, 2),
  p_value = c(
    pchisq(lrt_weibull_vs_exp, df = 1, lower.tail = FALSE),
    pchisq(lrt_gamma_vs_weibull, df = 1, lower.tail = FALSE),
    pchisq(lrt_gamma_vs_exp, df = 2, lower.tail = FALSE)
  )
)

lrt_table

# Chi squared values are all really small --> we go with the inner most nested model --> g_g

tidy_aft <- broom::tidy(gen_gamma_model, exponentiate = TRUE, conf.int = TRUE) %>%
  filter(term != "(Intercept)") %>%
  rename(
    Variable = term,
    Time_Ratio = estimate,
    TR_lower = conf.low,
    TR_upper = conf.high,
    p_value_aft = p.value
  ) %>%
  select(Variable, Time_Ratio, TR_lower, TR_upper, p_value_aft) %>%
  mutate(across(where(is.numeric), ~ round(., 3)))
```

```
comparison_cox_exp <- left_join(tidy_cox, tidy_aft, by = "Variable")

comparison_cox_exp

# Generalized Gamma is best among AFT models
```