BIOS 6646: Final Project Report

Frequentist and Bayesian Accelerated Failure Time (AFT) Models: Hospitalization Stays in Cystic Fibrosis Patients

Cole Hoffman

2024-12-06

Introduction

Cystic fibrosis (CF) is a genetic disorder caused by mutations that impair the function of the cystic fibrosis transmembrane conductance regulator (CFTR) protein. CF affects multiple organ systems, including the liver, pancreas, lungs, intestines, and reproductive organs. While rare, a German study confirmed that CF is diagnosed in roughly 1 in 3300 infants.[1] Diagnosis is typically multifaceted, relying on clinical evidence (often through newborn screening), identification of disease-causing mutations, and confirmation of CFTR-gene dysfunction through a "sweat test."[1]

The prognosis for patients with CF has significantly improved over the past few decades. In the 1950s, the median lifespan was just a few months; however, by 2017, it has risen to approximately 40 years.[1] Despite these medical advancements, CF remains a challenging condition to treat, with many patients frequently requiring hospitalization, sometimes as often as 2-6 times per year. While hospitalizations can occur for various reasons, most are due to pulmonary exacerbations, with stays often lasting 7 to 14 days.[2] This underscores the extensive burden CF inflicts on both patients and the healthcare system and serves as the main motivation for this project.

This project examines how certain patient characteristics and disease factors affect the length and frequency of hospitalizations among CF patients. There are two primary research aims: (1) how do these factors affect the length of first hospital stay among CF patients, and (2) to examine how they influence the time until patients are hospitalized again after their first discharge.

To address these aims, the project utilizes data from a comprehensive CF patient data registry comprised of three primary datasets. The first dataset includes all recorded hospital stays for CF patients admitted for pulmonary exacerbation. The second dataset features collected cultures, and the third dataset includes patient demographics and important disease factors contributing to CF diagnosis. These data were collected from 2002 with the most recent observations from 2020.

Methods

For the first research aim, the time-to-event variable is defined as the number of days in the first hospital stay. Patients whose first hospital stay exceeds 14 days will be treated as right censored in this analysis. For the second research aim, the time-to-event variable measures the number of days from discharge after the first hospital stay to readmission for the second hospital stay. The common covariates across both aims include patient sex (male or female), age at admission, and three key CF-related diagnosis factors: meconium ileus, respiratory illness, and failure to thrive. While many

factors are associated with a CF diagnosis, these three have been identified in the literature as principal drivers of CF diagnosis.[1]

Of the 851 subjects with available demographics data, 824 were diagnosed with CF. The remaining 27 subjects were diagnosed with CFTR-related gene disorders that did not meet the criteria for a CF diagnosis and were excluded from this analysis. Among the 824 CF subjects, 433 had at least one hospital stay on record. Of these, 392 individuals experienced the "event" for Aim 1, with first hospital stays lasting 14 days or less. The remaining 41 individuals were classified as right censored with hospital stays exceeding 14 days. Table 1 below provides demographic information for these patients.

Table 1: Demographics for Hospitalized CF Patients

	Female	Male	Overall
	(N=220)	(N=213)	(N=433)
Age at Admission			
Mean (SD)	10.4 (6.60)	10.5 (7.31)	10.4 (6.95)
Median [Min, Max]	10.4 [0.0520, 32.8]	11.1 [0.110, 41.1]	10.6 [0.0520, 41.1]
Meconium Ileus			
Yes	38 (17.3%)	35 (16.4%)	73 (16.9%)
No	182 (82.7%)	178 (83.6%)	360 (83.1%)
Respiratory Symptoms			
Yes	31 (14.1%)	29 (13.6%)	60 (13.9%)
No	189 (85.9%)	184 (86.4%)	373 (86.1%)
Failure to Thrive			
Yes	29 (13.2%)	27 (12.7%)	56 (12.9%)
No	191 (86.8%)	186 (87.3%)	377 (87.1%)

Of all CF patients with at least one hospitalization, 324 had at least two hospital stays. Among these, 159 experienced a second hospitalization within 365 days of discharge from their first hospital stay, representing the "event" of interest for Aim 2. The remaining 165 individuals were classified as right censored, as they were not hospitalized again within 365 days of their initial discharge. Table 2 below provides demographic information for the patients included in the analysis for Aim 2.

Table 2: Demographics for CF Patients with 2 or more hospitalizations

	Female	Male	Overall
	(N=172)	(N=152)	(N=324)
Age at Admission			
Mean (SD)	10.5 (6.64)	9.53 (6.61)	10.1 (6.64)
Median [Min, Max]	10.6 [0.0520, 32.8]	10.0 [0.110, 28.4]	10.4 [0.0520, 32.8]
Meconium Ileus			
Yes	31 (18.0%)	31 (20.4%)	62 (19.1%)
No	141 (82.0%)	121 (79.6%)	262 (80.9%)
Respiratory Symptoms			
Yes	27 (15.7%)	24 (15.8%)	51 (15.7%)
No	145 (84.3%)	128 (84.2%)	273 (84.3%)
Failure to Thrive			
Yes	20 (11.6%)	18 (11.8%)	38 (11.7%)
No	152 (88.4%)	134 (88.2%)	286 (88.3%)

To assess covariate effects across both aims, Kaplan-Meier curves and Schoenfeld residual plots were generated to test the proportional hazards assumption (see Aim 1 Proportional Hazards checks in Appendix). Overlapping survival functions suggested a violation to the assumption, leading to the use of parametric accelerated failure time (AFT) models. These models assess the log-survival time as a linear function of the covariates and were fit assuming exponential, Weibull, lognormal, and log-logistic baseline hazard distributions in both frequentist and Bayesian frameworks. For the Bayesian models, weakly informative priors were placed: Normal$(0, 100)$ for intercepts and coefficients, Gamma$(1, 1)$ for the Weibull shape parameter, and lognormal$(0, 2)$ for the lognormal scale parameter. All Bayesian models ran with 4 chains and 2000 iterations to ensure adequate sampling of the posterior distributions and stable hazard ratio estimates. Hazard ratios for the Weibull AFT models were reparametrized into the proportional hazards framework and reported in Tables 3 and 4. Model performance was assessed with Akaike Information Criterion (AIC) for frequentist models and Watanabe-Akaike Information Criterion (WAIC) for Bayesian models, with results in Table 5. Of note, WAIC is not reported for the Bayesian log-logistic model since this cannot be directly interpretable in R.

Diagnostic plots were analyzed to evaluate model fits. Cumulative hazard plots for all AFT frequentist models assessed the best-fitting parametric distribution for the baseline hazard (see Aim 1: AFT Model Diagnostics in Appendix) . Bayesian AFT models underwent additional diagnostic checks, including posterior predictive checks, trace plots, autocorrelation plots, and R-hat convergence diagnostics (see Bayesian Diagnostics in Appendix). All analyses were performed in R Version 4.3.3, with reproducible code available in the Code Appendix.

Results

For Aim 1, all predictors in the frequentist Weibull AFT model showed statistically significant hazard ratio estimates ($p < 0.05$), except for meconium ileus (Table 3). This indicates that age at admission, sex, and the presence of respiratory illness or failure to thrive significantly impact the length of the first hospital stay among CF patients. For every additional year of age, the likelihood

of a hospital stay greater than 14 days increased by 2% (95% CI: (0.969, 0.990), p = 0.008). Male subjects experienced 22.9% shorter hospital stays compared to females (95% CI: (1.054, 1.434), p = 0.049). Respiratory illness increased the chance of longer stays by 52.2% (95% CI: (0.364, 0.629), p < 0.001), while failure to thrive increased the chance by 36.8% (95% CI: (0.472, 0.846), p = 0.021). Hazard ratio estimates from the Bayesian Weibull AFT model showed slight differences, particularly for sex, where males had a 4.5% increased likelihood of longer first hospital stays (95% credible interval: (0.666, 1.247)). However, the Bayesian model's 95% credible intervals suggest statistical significance except for sex, whose interval crossed 1.

Table 3: Weibull AFT Model Results: Length of First Hospital Stay (Aim 1)

	Frequentist			Bayesian		
Variable	Hazard Ratio	95% Confidence Interval	P-Value	Hazard Ratio	95% Credible Interval	R-Hat
Age at Admission	0.980	(0.969, 0.990)	0.008	0.955	(0.936, 0.974)	1.00
Sex (Male)	1.229	(1.054, 1.434)	0.049	0.911	(0.666, 1.247)	1.00
Meconium Ileus	0.803	(0.653, 0.986)	0.118	0.465	(0.302, 0.716)	1.00
Respiratory Illness	0.478	(0.364, 0.629)	< 0.001	0.136	(0.063, 0.292)	1.00
Failure to Thrive	0.632	(0.472, 0.846)	0.021	0.260	(0.119, 0.570)	1.00

For Aim 2, the hazard ratio point estimates for the frequentist Weibull AFT model are mostly statistically insignificant ($p > 0.05$) (Table 4), with age at admission being the only statistically significant predictor from first discharge to readmission. For every one-year increase in age, the likelihood of readmission within one year increases by 3.4% (95% CI: (0.985, 1.084), p = 0.014). While sex, meconium ileus and respiratory illness contribute to a lower likelihood of readmission, failure to thrive is a risk factor for earlier readmission. In the Bayesian Weibull AFT model, the hazard ratio estimates differ, but the effects are largely consistent with those from the frequentist model. Meconium ileus and failure to thrive are statistically insignificant, with 95% credible intervals including 1, while age, sex, and respiratory illness remain statistically significant.

Table 4: Weibull AFT Model Results: Time Until 2nd Hospitalization (Aim 2)

	Frequentist			Bayesian		
Variable	Hazard Ratio	95% Confidence Interval	P-Value	Hazard Ratio	95% Credible Interval	R-Hat
Age at Admission	1.034	(0.985, 1.084)	0.014	1.073	(1.053, 1.095)	1.00
Sex (Male)	0.861	(0.473, 1.566)	0.372	0.727	(0.610, 0.867)	1.00
Meconium Ileus	0.691	(0.300, 1.591)	0.114	0.915	(0.738, 1.136)	1.00
Respiratory Illness	0.995	(0.431, 2.296)	0.984	0.642	(0.488, 0.845)	1.00
Failure to Thrive	1.154	(0.418, 3.185)	0.615	1.173	(0.874, 1.573)	1.00

For Aim 1, the log-logistic outperformed all other frequentist AFT models based on AIC, though the lognormal model also appears to reasonably fit the data for modeling the length of the first hospital stay. In the Bayesian AFT framework, the lognormal model performs substantially better than the Weibull and exponential models. For Aim 2, both AIC and WAIC suggest that the lognormal model best fits the data for time from initial discharge to readmission. The log-logistic and Weibull models perform similarly well in the frequentist framework, with the exponential model being clearly inferior.

Table 5: AFT Model Performance Results for Aims 1 and 2

	Aim 1: Length of first Hospital Stay		Aim 2: Time from Discharge to Readmission	
Model Type	AIC	WAIC	AIC	WAIC
Exponential	2650.690	474.538	2407.779	2697.038
Weibull	2595.374	420.346	2320.276	2608.904
Lognormal	2450.727	370.761	2291.518	2565.090
Log-Logistic	2412.904	NA	2300.775	NA

Discussion

The lognormal model showed strong fit for modeling the length of first hospital stay, outperforming the Weibull and exponential AFT models, and fit well for time from discharge to readmission. Most predictors were significant for the length of the first hospital stay, with males typically experiencing shorter stays, although this effect is reversed in the Bayesian Weibull AFT model. The presence of disease factors were associated with longer stays. For time from initial discharge to readmission, most predictors were statistically insignificant in the frequentist Weibull AFT model, but some became significant in the Bayesian model. Discrepancies between the frequentist and Bayesian estimates, with narrower 95% credible intervals in the Bayesian model, suggest that the latter may be more robust to data noise. These discrepancies may also be attributed to the specification of the implemented priors in the Bayesian models, as the hazard ratio estimates from both AFT frameworks would ideally be similar. Aim 2 showed higher WAIC values than Aim 1, potentially due to fewer observations (159 subjects) or wider data distribution in time from discharge to readmission.

In terms of model diagnostics, global testing confirmed violations of the proportional hazards assumption, supporting AFT model use. Cumulative hazard plots suggested superior performance of the Weibull and lognormal models compared to the exponential model, though the Weibull QQ-plot indicated poor model fit. Bayesian diagnostics showed stable trace plots, low autocorrelation, and relatively good fit in posterior predictive plots for the lognormal model, but the Weibull and exponential models showed obvious misspecification.

Future analyses could explore additional demographics and disease factors, such as race and genotype status, as predictors for hospital stay length and frequency in CF patients. However, some racial groups were underrepresented in this dataset, with approximately 95% of the patients identified as Caucasian (see Table 1 in Appendix). Additionally, many genotypes were missing or unknown, complicating the analysis. It may also be valuable to explore time until first hospitalization, though missing data for older adult patients – particularly those born before the 1980s – poses a challenge since data collection had not started until 2002, resulting in 20 or more years of absent hospitalization records. For this type of analysis, proper handling of left truncated data is essential, as neglecting this would lead to significant bias.

Conclusion

Cystic fibrosis (CF) is a complex disease affecting multiple organ systems, and hospitalization durations vary widely among those affected. Disease factors such as meconium ileus, respiratory illness, and failure to thrive are significant predictors of longer first hospital stays. However, these factors do not significantly predict time to readmission. Age, on the other hand, is a significant predictor of both shorter first hospital stays and earlier readmission. The impact of sex on hospitalization patterns appears less clear. Further research is needed to better understand the

clinical significance of these variables in shaping hospitalization patterns for CF patients.

References

(1) Naehrig, S., Chao, C.-M., & Naehrlich, L. (2017). Cystic fibrosis. Deutsches Ärzteblatt International, 114(33–34), 564–574. https://doi.org/10.3238/arztebl.2017.0564

(2) Stephenson, A., Hux, J., Tullis, E., Austin, P. C., Corey, M., & Ray, J. (2011). Higher risk of hospitalization among females with cystic fibrosis. Journal of Cystic Fibrosis, 10(2), 93–99. https://doi.org/10.1016/j.jcf.2010.10.005

Appendix

Dataset: Hospital Admissions

```
## # A tibble: 6 x 7
##      mrn admit_date          days_admitted reasons_for_admit discharge_date
##    <dbl> <dttm>                      <dbl> <chr>             <dttm>
## 1 231022 2003-06-16 00:00:00            14 Pulmonary Exacer~ 2003-06-30 00:00:00
## 2 231022 2004-02-14 00:00:00            13 Pulmonary Exacer~ 2004-02-27 00:00:00
## 3 231022 2004-08-18 00:00:00            14 Pulmonary Exacer~ 2004-09-01 00:00:00
## 4 231022 2005-09-21 00:00:00            13 Pulmonary Exacer~ 2005-10-04 00:00:00
## 5 231022 2006-05-04 00:00:00            14 Pulmonary Exacer~ 2006-05-18 00:00:00
## 6 231022 2006-12-22 00:00:00            14 Pulmonary Exacer~ 2007-01-05 00:00:00
## # i 2 more variables: days_elapsed <dbl>, visit_number <int>
```

Dataset: Collected Cultures

```
## # A tibble: 6 x 28
##      mrn encounter_date      bacterial_culture_done date_of_bacterial_culture
##    <dbl> <dttm>              <lgl>                  <dttm>
## 1 840766 2011-06-15 00:00:00 TRUE                   2011-06-15 00:00:00
## 2 840766 2011-09-14 00:00:00 TRUE                   2011-09-14 00:00:00
## 3 869155 2010-06-09 00:00:00 TRUE                   2010-06-09 00:00:00
## 4 489172 2010-08-09 00:00:00 TRUE                   2010-08-09 00:00:00
## 5 478401 2010-04-01 00:00:00 TRUE                   2010-04-01 00:00:00
## 6 478401 2010-07-20 00:00:00 TRUE                   2010-07-20 00:00:00
## # i 24 more variables: specimen_type <chr>, results <chr>,
## #   methicillin_sensitive_staph_aureus <lgl>, mrsa <lgl>, h_flu <lgl>,
## #   p_aeruginosa_mucoid <lgl>, p_aeruginosa_non_mucoid <lgl>,
## #   burkholderia <lgl>, burk_species <chr>, burk_species_other <chr>,
## #   othermicroorganisms <chr>, a_xylosoxidans <lgl>, s_maltophilia <lgl>,
## #   other_bacteria <chr>, aspergillus_any_species <lgl>,
## #   candida_any_species <lgl>, scedosporium_species <lgl>, ...
```

Dataset: Patient Demographics

```
##       mrn      dob genotype_1 genotype_2       diagnosis    sex      race
## 1 1691997 11/28/10    F508del    F508del Cystic Fibrosis   Male Caucasian
## 2 1559717  1/28/00    F508del    F508del Cystic Fibrosis   Male Caucasian
## 3 1771333  3/26/08    F508del      S945L Cystic Fibrosis Female Caucasian
## 4  707084  9/25/96    F508del      R553X Cystic Fibrosis   Male Caucasian
## 5 1468928   1/8/12     R1162X   2143delT Cystic Fibrosis Female Caucasian
## 6 1891512  5/30/16     R1162X        UNK Cystic Fibrosis   Male     Other
##   hispanic_latinx              mec_il      dx_by_nbs dx_by_family_history
## 1           FALSE                  No            Yes                FALSE
## 2           FALSE Yes - treatment unk N/A or Unknown                FALSE
## 3           FALSE                  No            Yes                FALSE
## 4            TRUE             Unknown             No                FALSE
## 5            TRUE                  No            Yes                FALSE
```

```
## 6            TRUE                   No            Yes                  TRUE
##   dx_due_to_malabsorption dx_due_to_meconium_ileus dx_due_to_rectal_prolapse
## 1                   FALSE                    FALSE                     FALSE
## 2                   FALSE                     TRUE                     FALSE
## 3                   FALSE                    FALSE                     FALSE
## 4                    TRUE                    FALSE                     FALSE
## 5                   FALSE                    FALSE                     FALSE
## 6                    TRUE                    FALSE                     FALSE
##   dx_due_to_failure_to_thrive dx_due_to_respiratory_symptoms prenatal_dx
## 1                       FALSE                          FALSE       FALSE
## 2                       FALSE                          FALSE       FALSE
## 3                       FALSE                          FALSE       FALSE
## 4                        TRUE                           TRUE       FALSE
## 5                       FALSE                          FALSE       FALSE
## 6                       FALSE                          FALSE       FALSE
##   dx_due_to_nasal_sinus_symptoms dx_due_to_liver_symptoms dx_by_dna_analysis
## 1                          FALSE                    FALSE              FALSE
## 2                          FALSE                    FALSE              FALSE
## 3                          FALSE                    FALSE              FALSE
## 4                          FALSE                    FALSE              FALSE
## 5                          FALSE                    FALSE              FALSE
## 6                          FALSE                    FALSE              FALSE
##   dx_due_to_electrolyte_imbalance other_factors_in_dx factors_in_dx_are_unknown
## 1                           FALSE                                         FALSE
## 2                           FALSE                                         FALSE
## 3                           FALSE                                         FALSE
## 4                           FALSE                                         FALSE
## 5                           FALSE                                         FALSE
## 6                           FALSE                                         FALSE
##   genotype_status
## 1      Homozygous
## 2      Homozygous
## 3    Heterozygous
## 4    Heterozygous
## 5    Heterozygous
## 6         Unknown
```

Dataset: Length of First Hospital Stay (Aim 1)

```
## # A tibble: 6 x 33
##      mrn admit_date          days_admitted reasons_for_admit discharge_date
##    <dbl> <dttm>                      <dbl> <chr>             <dttm>
## 1 231022 2003-06-16 00:00:00            14 Pulmonary Exacer~ 2003-06-30 00:00:00
## 2 315360 2003-06-05 00:00:00            14 Pulmonary Exacer~ 2003-06-19 00:00:00
## 3 358989 2003-01-02 00:00:00             3 Pulmonary Exacer~ 2003-01-05 00:00:00
## 4 386664 2003-02-19 00:00:00             8 Pulmonary Exacer~ 2003-02-27 00:00:00
## 5 405201 2005-09-01 00:00:00            11 Pulmonary Exacer~ 2005-09-12 00:00:00
## 6 412320 2004-04-06 00:00:00            10 Pulmonary Exacer~ 2004-04-16 00:00:00
```

```
## # i 28 more variables: days_elapsed <dbl>, visit_number <int>, event <dbl>,
## #   dob <date>, genotype_1 <chr>, genotype_2 <chr>, diagnosis <chr>, sex <chr>,
## #   race <chr>, hispanic_latinx <lgl>, mec_il <chr>, dx_by_nbs <chr>,
## #   dx_by_family_history <lgl>, dx_due_to_malabsorption <lgl>,
## #   dx_due_to_meconium_ileus <lgl>, dx_due_to_rectal_prolapse <lgl>,
## #   dx_due_to_failure_to_thrive <lgl>, dx_due_to_respiratory_symptoms <lgl>,
## #   prenatal_dx <lgl>, dx_due_to_nasal_sinus_symptoms <lgl>, ...
```

Dataset: Time from Initial Discharge to Readmission (Aim 2)

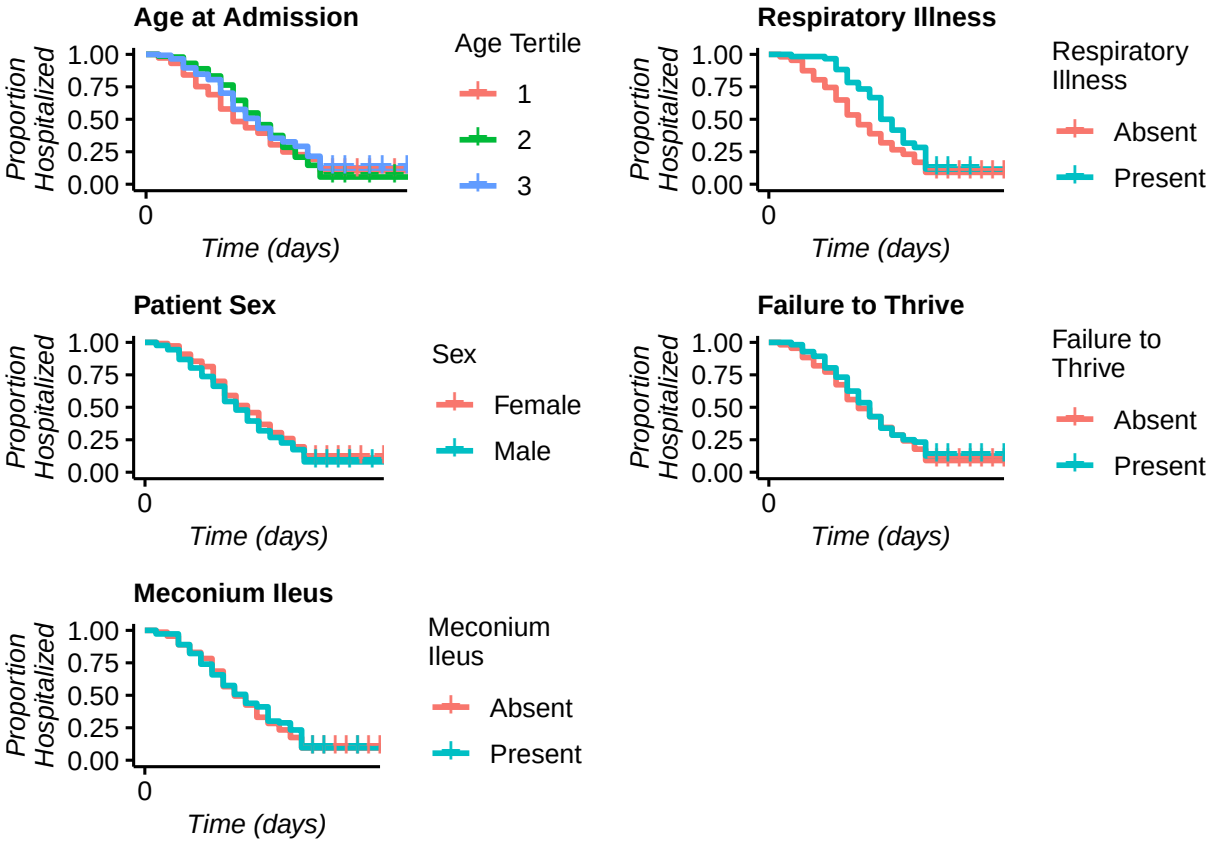
```
## # A tibble: 6 x 34
##      mrn admit_date          days_admitted reasons_for_admit discharge_date
##    <dbl> <dttm>                      <dbl> <chr>             <dttm>
## 1 231022 2003-06-16 00:00:00            14 Pulmonary Exacer~ 2003-06-30 00:00:00
## 2 315360 2003-06-05 00:00:00            14 Pulmonary Exacer~ 2003-06-19 00:00:00
## 3 358989 2003-01-02 00:00:00             3 Pulmonary Exacer~ 2003-01-05 00:00:00
## 4 386664 2003-02-19 00:00:00             8 Pulmonary Exacer~ 2003-02-27 00:00:00
## 5 412320 2004-04-06 00:00:00            10 Pulmonary Exacer~ 2004-04-16 00:00:00
## 6 442291 2004-12-20 00:00:00             8 Pulmonary Exacer~ 2004-12-28 00:00:00
## # i 29 more variables: days_elapsed <dbl>, visit_number <int>, event <dbl>,
## #   dob <date>, genotype_1 <chr>, genotype_2 <chr>, diagnosis <chr>, sex <chr>,
## #   race <chr>, hispanic_latinx <lgl>, mec_il <chr>, dx_by_nbs <chr>,
## #   dx_by_family_history <lgl>, dx_due_to_malabsorption <lgl>,
## #   dx_due_to_meconium_ileus <lgl>, dx_due_to_rectal_prolapse <lgl>,
## #   dx_due_to_failure_to_thrive <lgl>, dx_due_to_respiratory_symptoms <lgl>,
## #   prenatal_dx <lgl>, dx_due_to_nasal_sinus_symptoms <lgl>, ...
```

Table 1: Demopgrahics for Overall Population

Table 6: Demographics by Disease Type

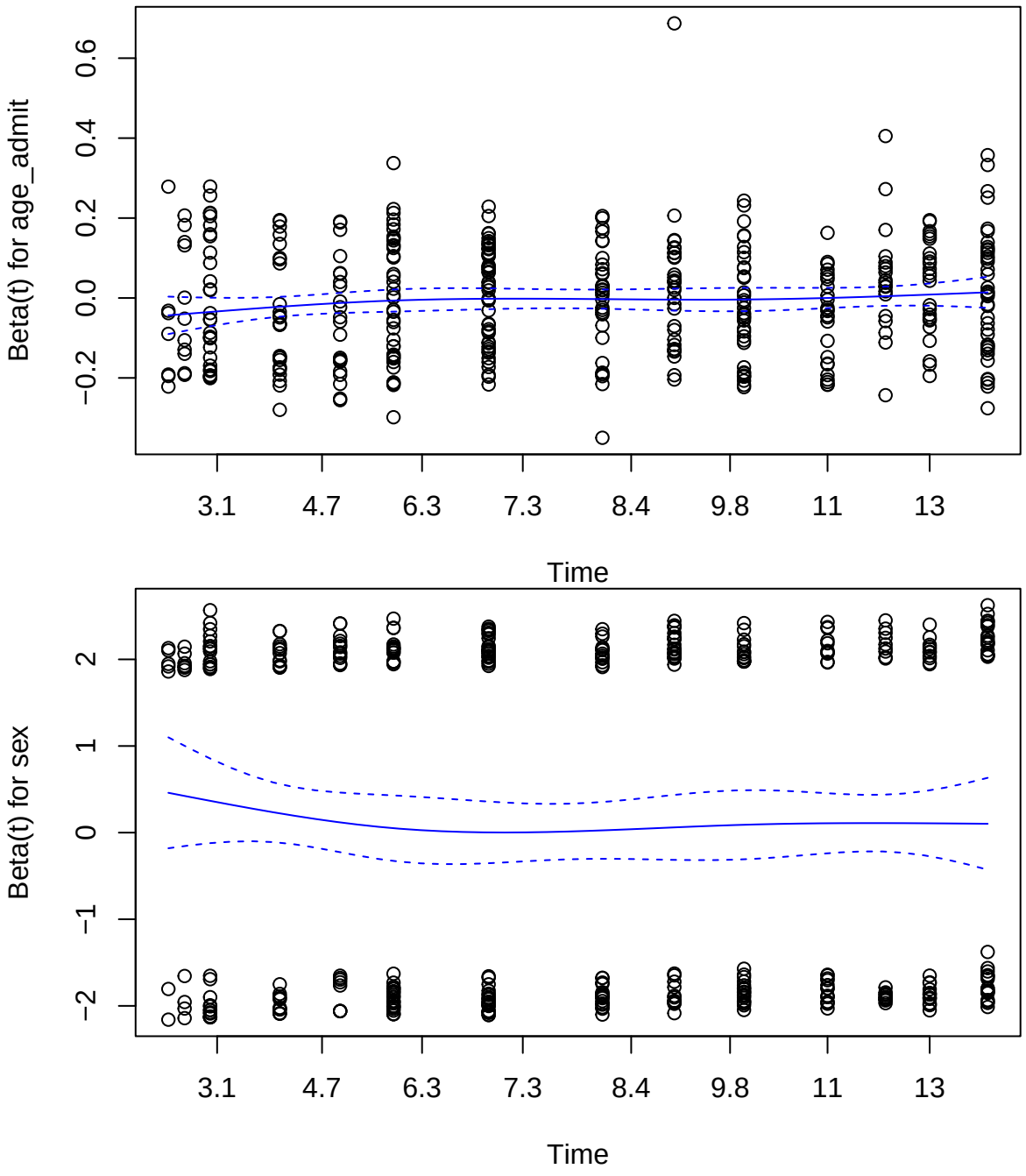
	CFTR-related disorder	CRMS	Cystic Fibrosis	Overall
	(N=6)	(N=21)	(N=824)	(N=851)
Sex				
Female	3 (50.0%)	15 (71.4%)	399 (48.4%)	417 (49.0%)
Male	3 (50.0%)	6 (28.6%)	425 (51.6%)	434 (51.0%)
Race				
Caucasian	5 (83.3%)	18 (85.7%)	774 (93.9%)	797 (93.7%)
Other	1 (16.7%)	1 (4.8%)	14 (1.7%)	16 (1.9%)
Black	0 (0%)	1 (4.8%)	12 (1.5%)	13 (1.5%)
Unknown	0 (0%)	1 (4.8%)	19 (2.3%)	20 (2.4%)
Caucasian/Native American	0 (0%)	0 (0%)	1 (0.1%)	1 (0.1%)
Indian	0 (0%)	0 (0%)	2 (0.2%)	2 (0.2%)
Native American	0 (0%)	0 (0%)	2 (0.2%)	2 (0.2%)
Hispanic/Latinx				
Yes	0 (0%)	4 (19.0%)	77 (9.3%)	81 (9.5%)
No	6 (100%)	17 (81.0%)	747 (90.7%)	770 (90.5%)
Genotype Status				
Heterozygous	5 (83.3%)	18 (85.7%)	345 (41.9%)	368 (43.2%)
Unknown	1 (16.7%)	3 (14.3%)	33 (4.0%)	37 (4.3%)
Homozygous	0 (0%)	0 (0%)	446 (54.1%)	446 (52.4%)
Meconium Ileus				
Yes	0 (0%)	1 (4.8%)	114 (13.8%)	115 (13.5%)
No	6 (100%)	20 (95.2%)	710 (86.2%)	736 (86.5%)
Respiratory Symptoms				
Yes	4 (66.7%)	4 (19.0%)	117 (14.2%)	125 (14.7%)
No	2 (33.3%)	17 (81.0%)	707 (85.8%)	726 (85.3%)
Failure to Thrive				
Yes	0 (0%)	2 (9.5%)	105 (12.7%)	107 (12.6%)
No	6 (100%)	19 (90.5%)	719 (87.3%)	744 (87.4%)

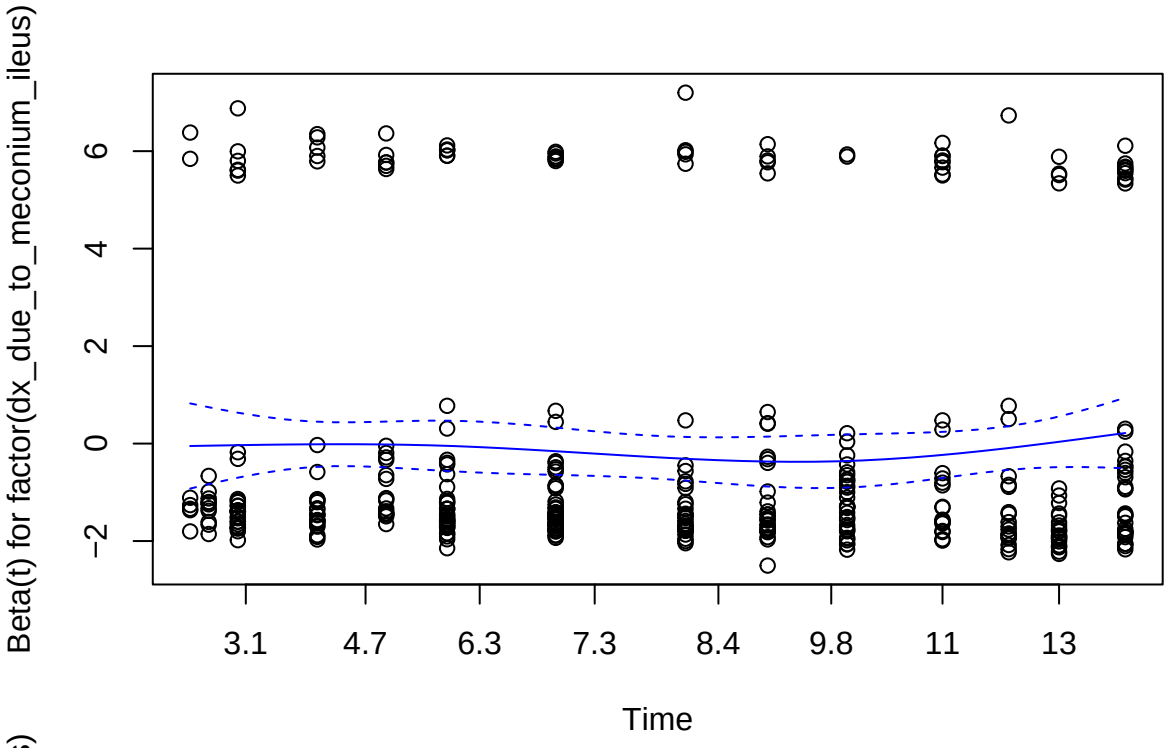
Aim 1: Cox Proportional Hazards Test - Kaplan-Meier Curves

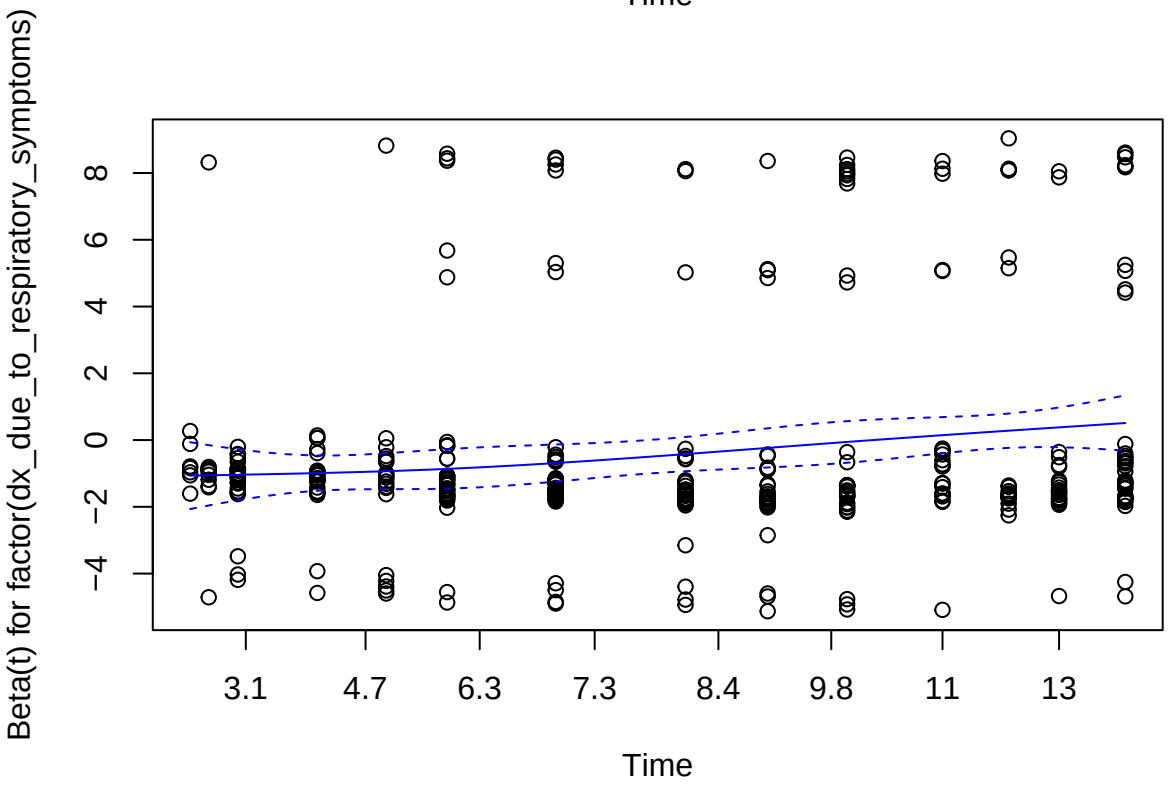


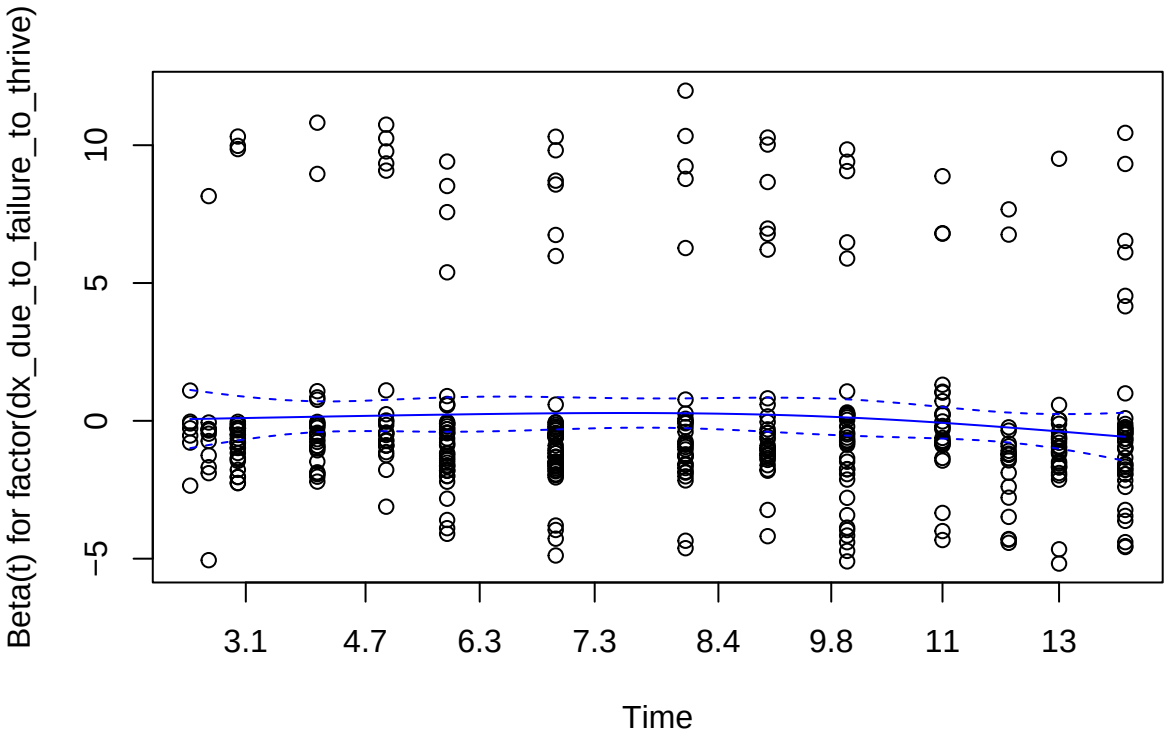
Aim 1: Cox Proprtional Hazards Global Test and Schoenfeld Residual Plots

```
##                                        chisq df       p
## age_admit                             3.9770  1 0.04613
## sex                                   0.4360  1 0.50904
## factor(dx_due_to_meconium_ileus)      0.0734  1 0.78646
## factor(dx_due_to_respiratory_symptoms) 12.9522  1 0.00032
## factor(dx_due_to_failure_to_thrive)   0.1585  1 0.69057
## GLOBAL                               16.9407  5 0.00461
```

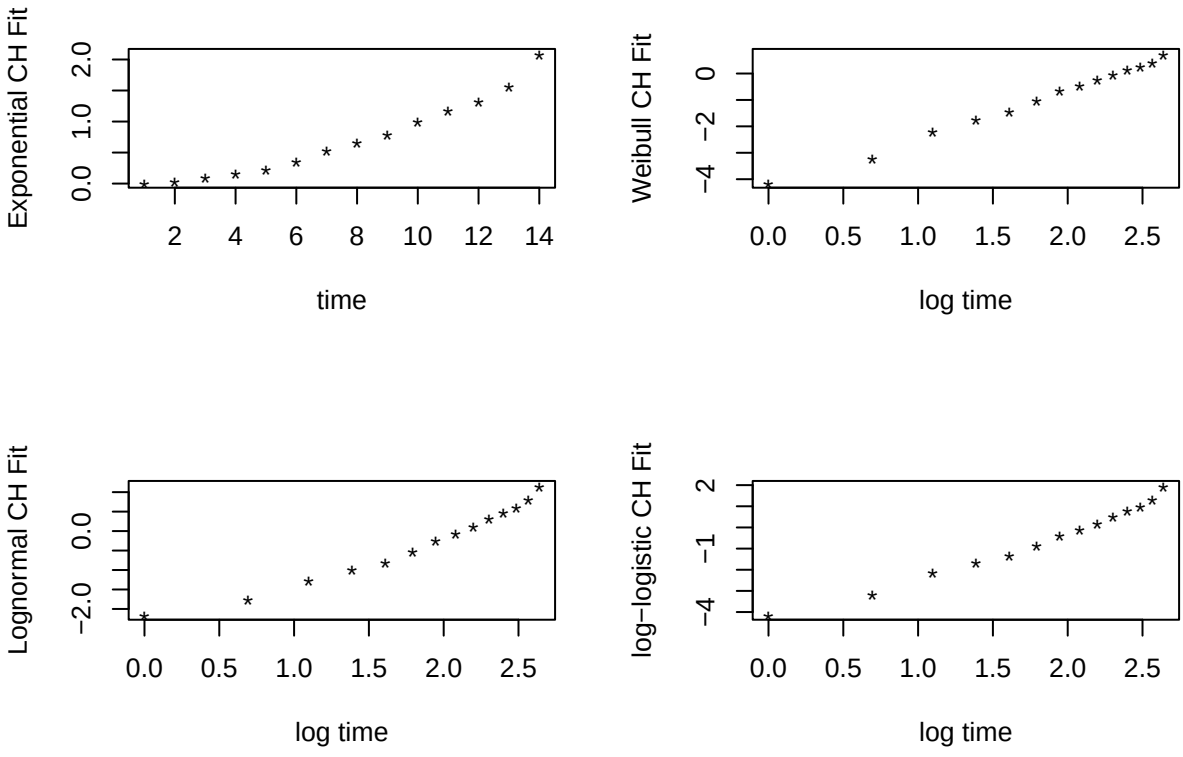

Beta(t) for age_admit
−0.2
0.0
0.2
0.4
0.6
3.1
4.7
6.3
7.3
8.4
9.8
11
13
Time
Beta(t) for sex
−2
−1
0
1
2
3.1
4.7
6.3
7.3
8.4
9.8
11
13
Time


Beta(t) for factor(dx_due_to_meconium_ileus)
−2
0
2
4
6
3.1
4.7
6.3
7.3
8.4
9.8
11
13
Time

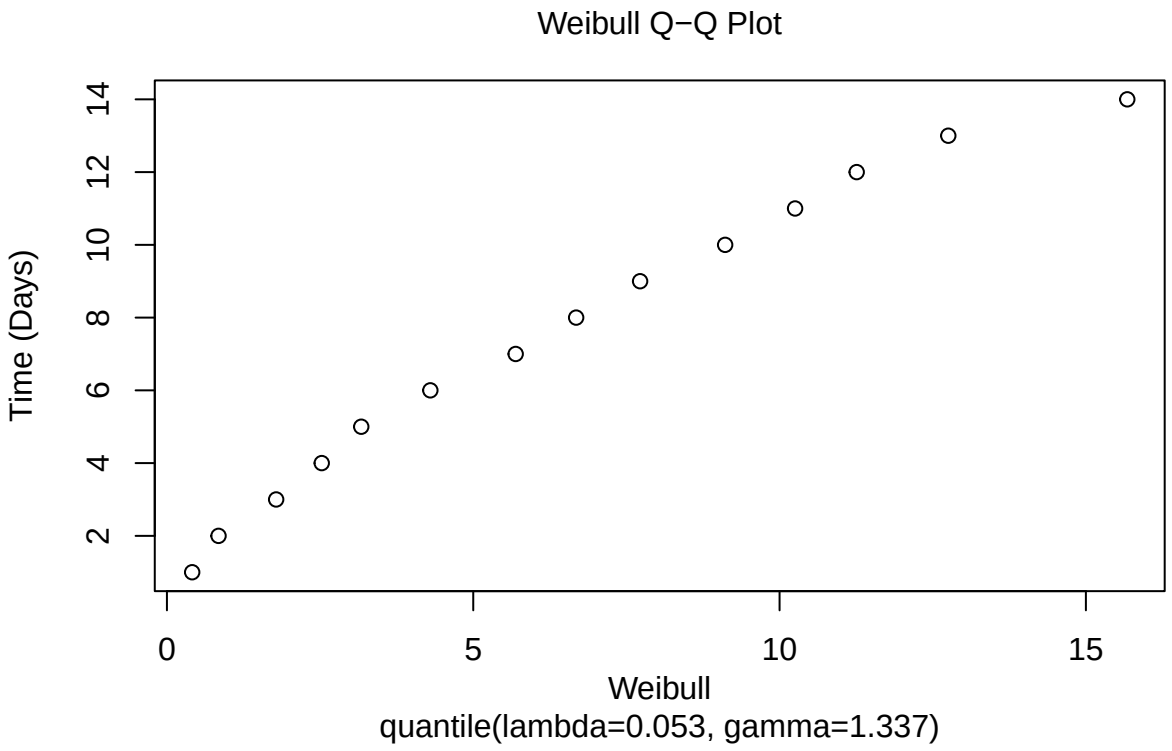

Beta(t) for factor(dx_due_to_respiratory_symptoms)
−4
−2
0
2
4
6
8
3.1
4.7
6.3
7.3
8.4
9.8
11
13
Time



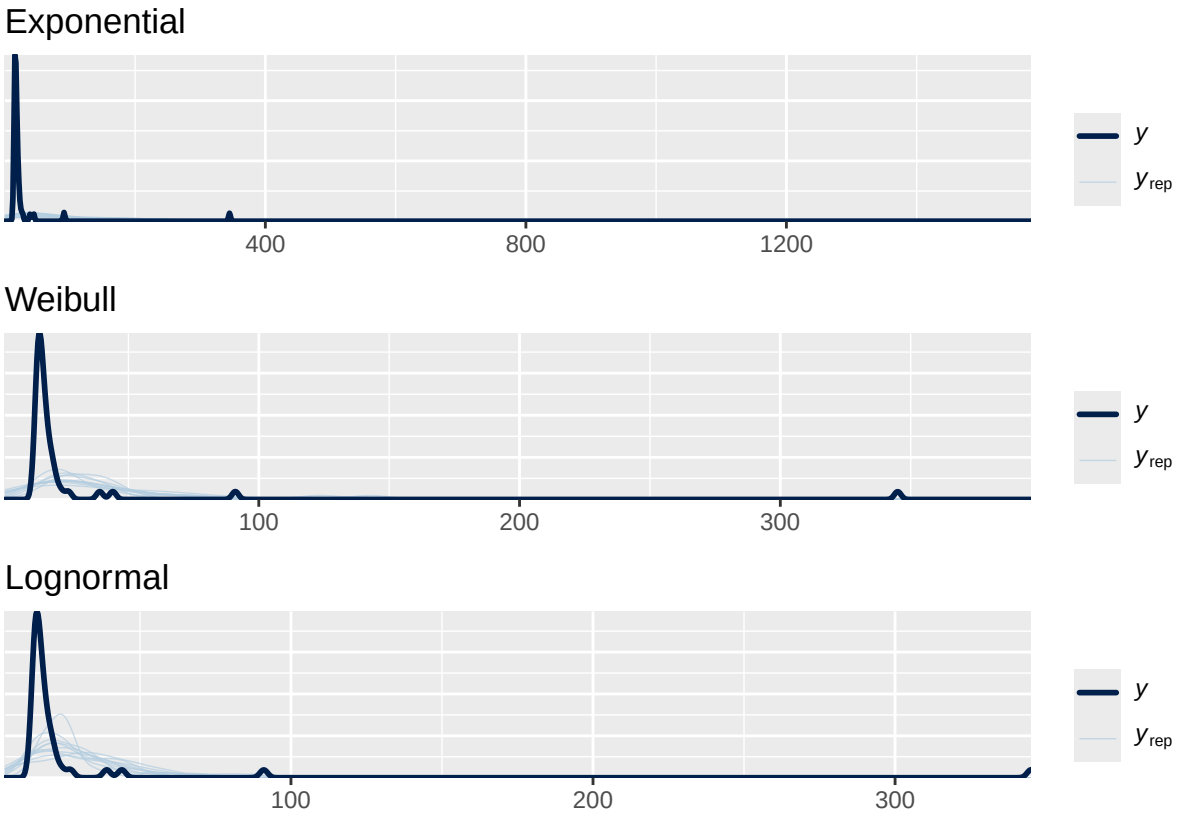
Aim 1: AFT Model Diagnostics - Cumulative Hazards Plots



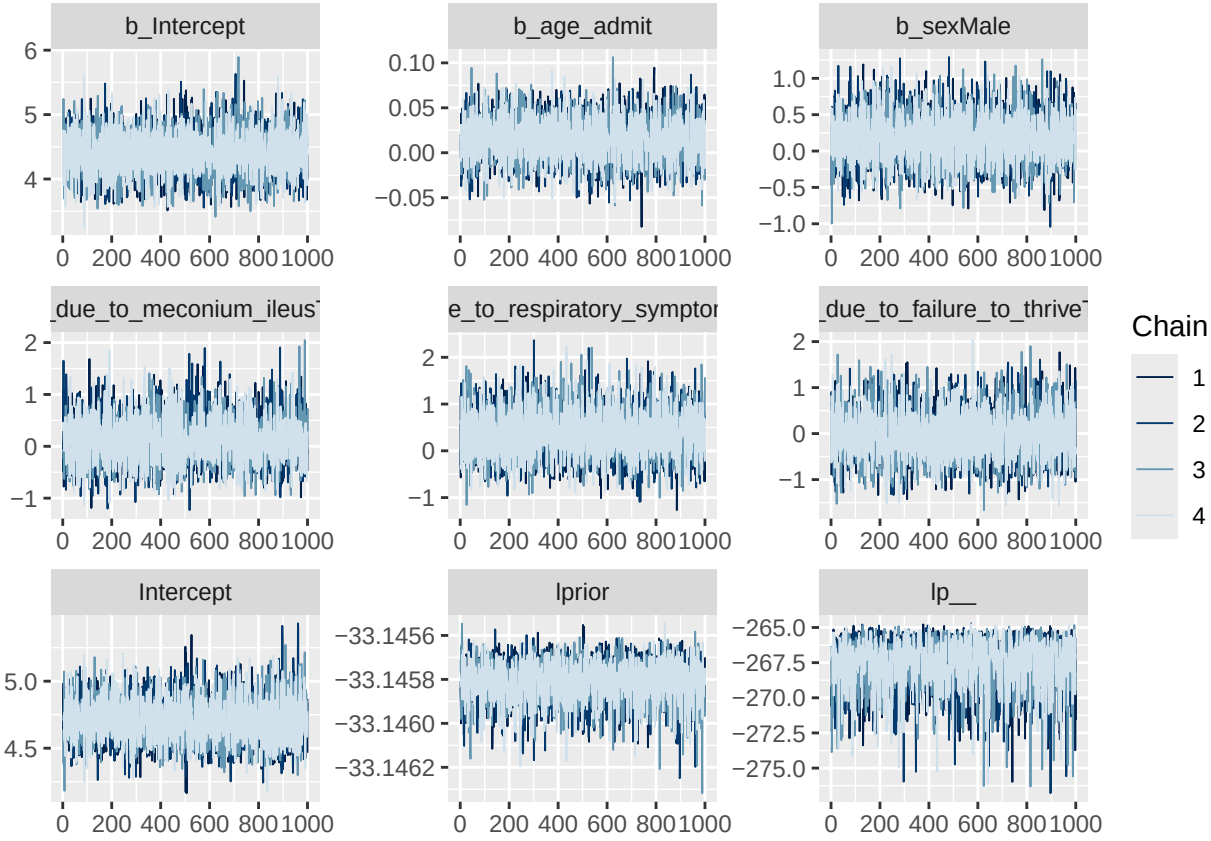
Aim 1: AFT Model Diagnostics - Weibull QQ-Plot



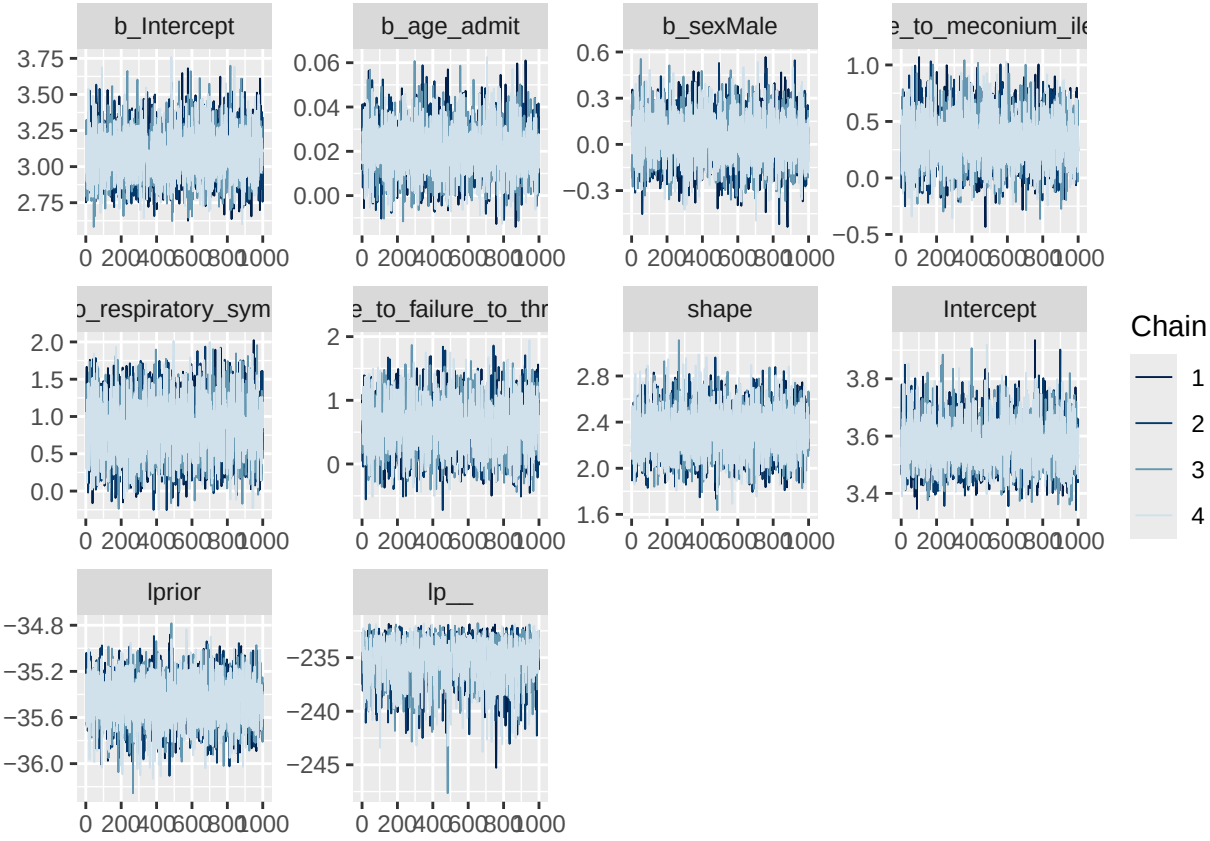
Aim 1: Bayesian AFT Diagnostics - Posterior Predictive Plots



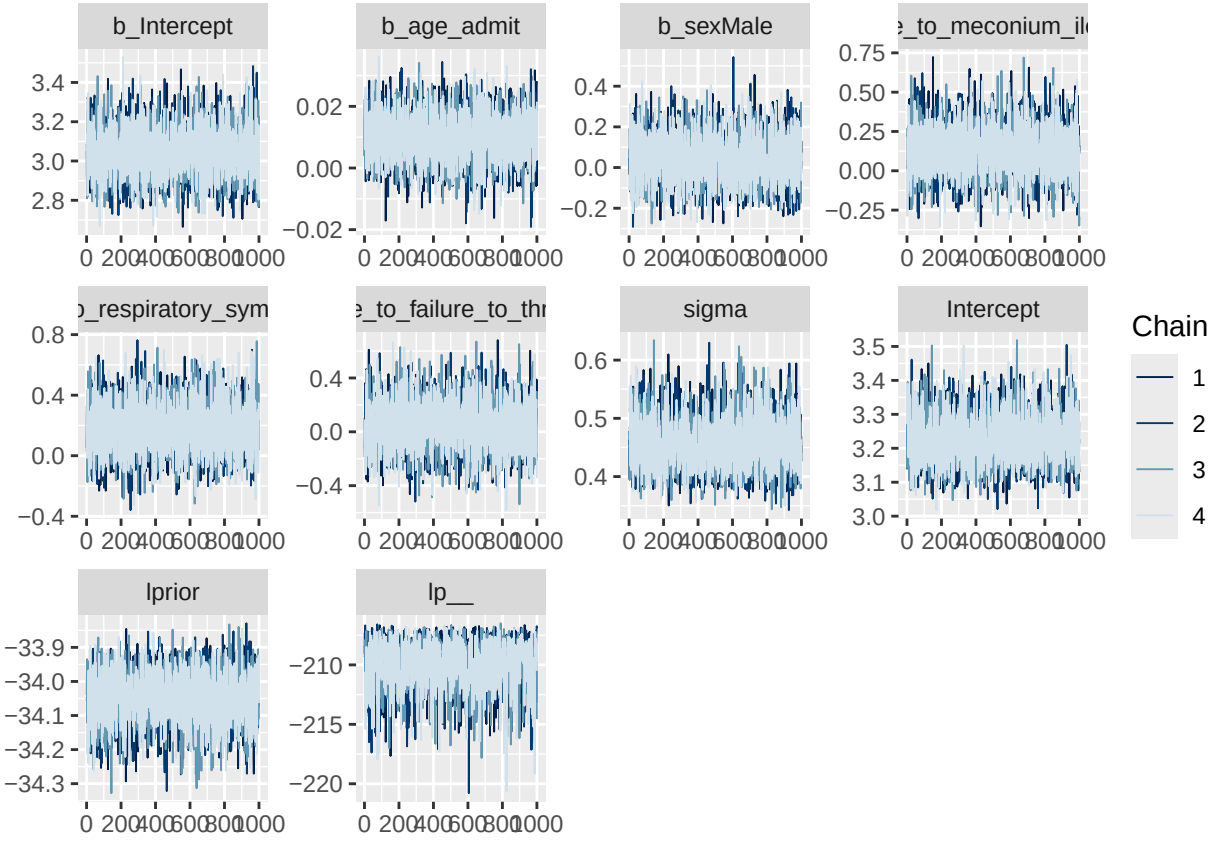
Aim 1: Bayesian AFT Diagnostics - Trace Plots for Exponential AFT Model



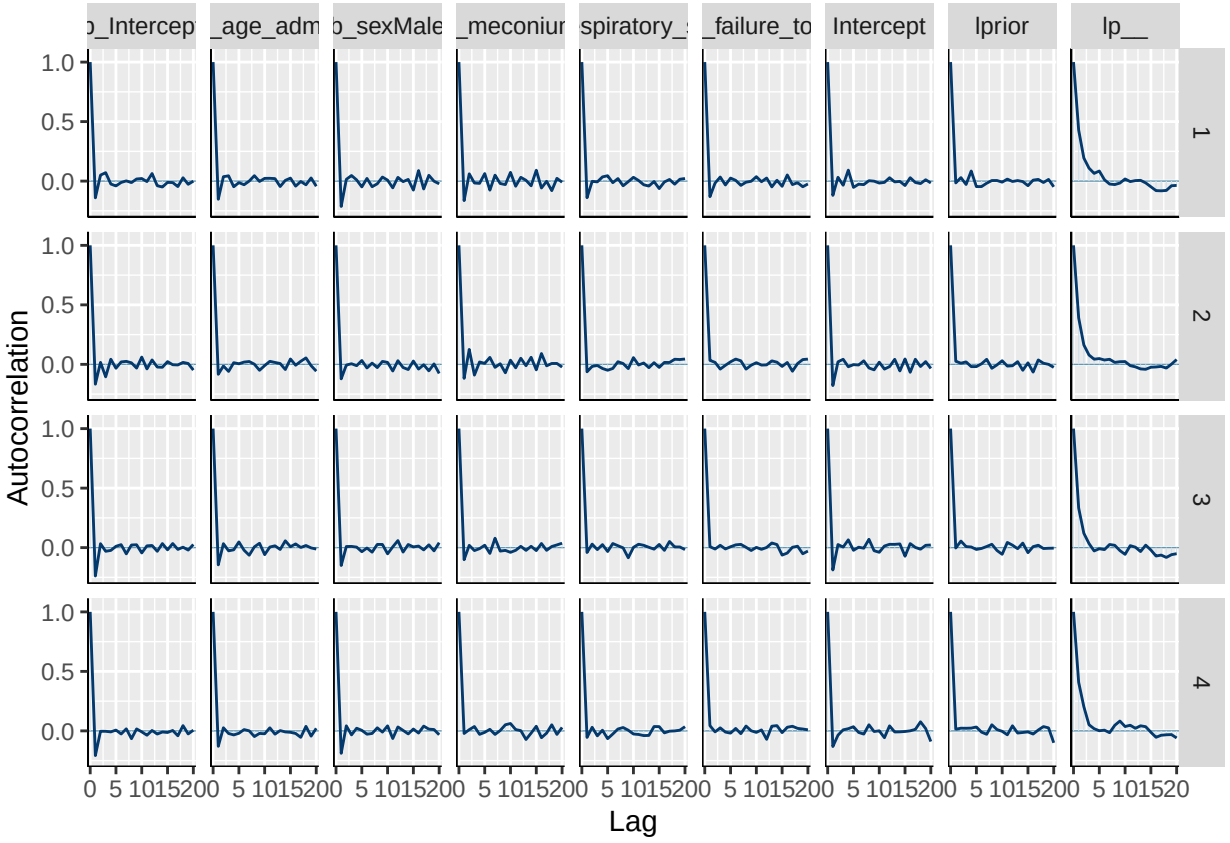
Aim 1: Bayesian AFT Diagnostics - Trace Plots for Weibull AFT Model



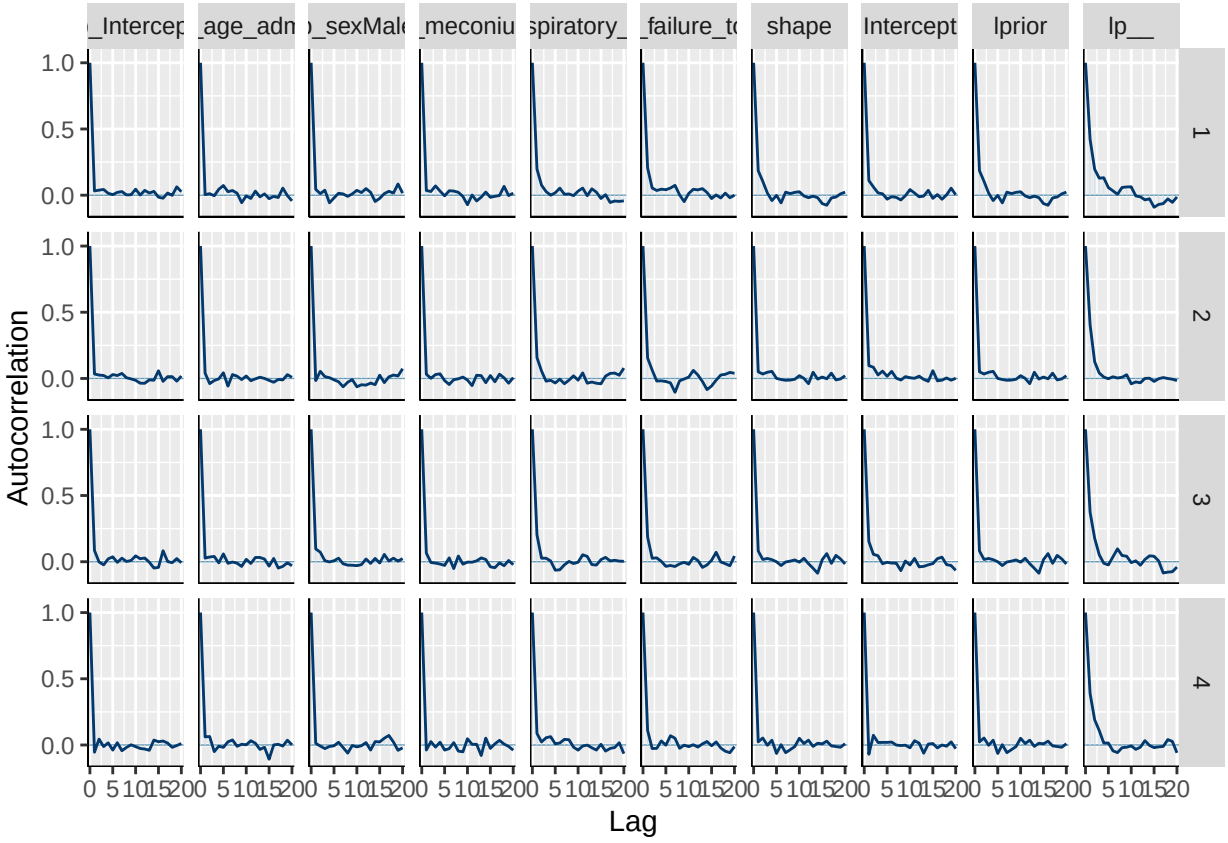
Aim 1: Bayesian AFT Diagnostics - Trace Plots for Lognormal AFT Model



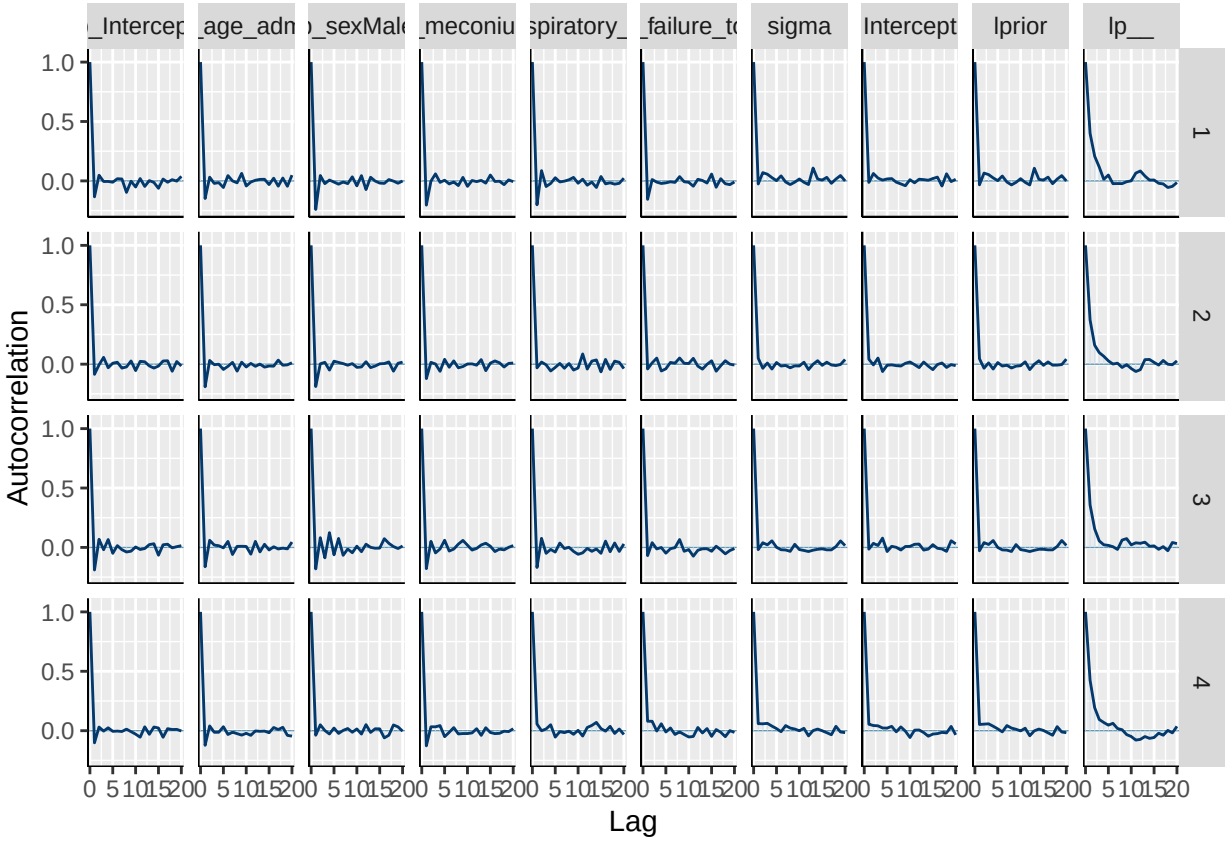
Aim 1: Bayesian AFT Diagnostics - Autocorrelation Plots for Exponential AFT Model



Aim 1: Bayesian AFT Diagnostics - Autocorrelation Plots for Weibull AFT Model

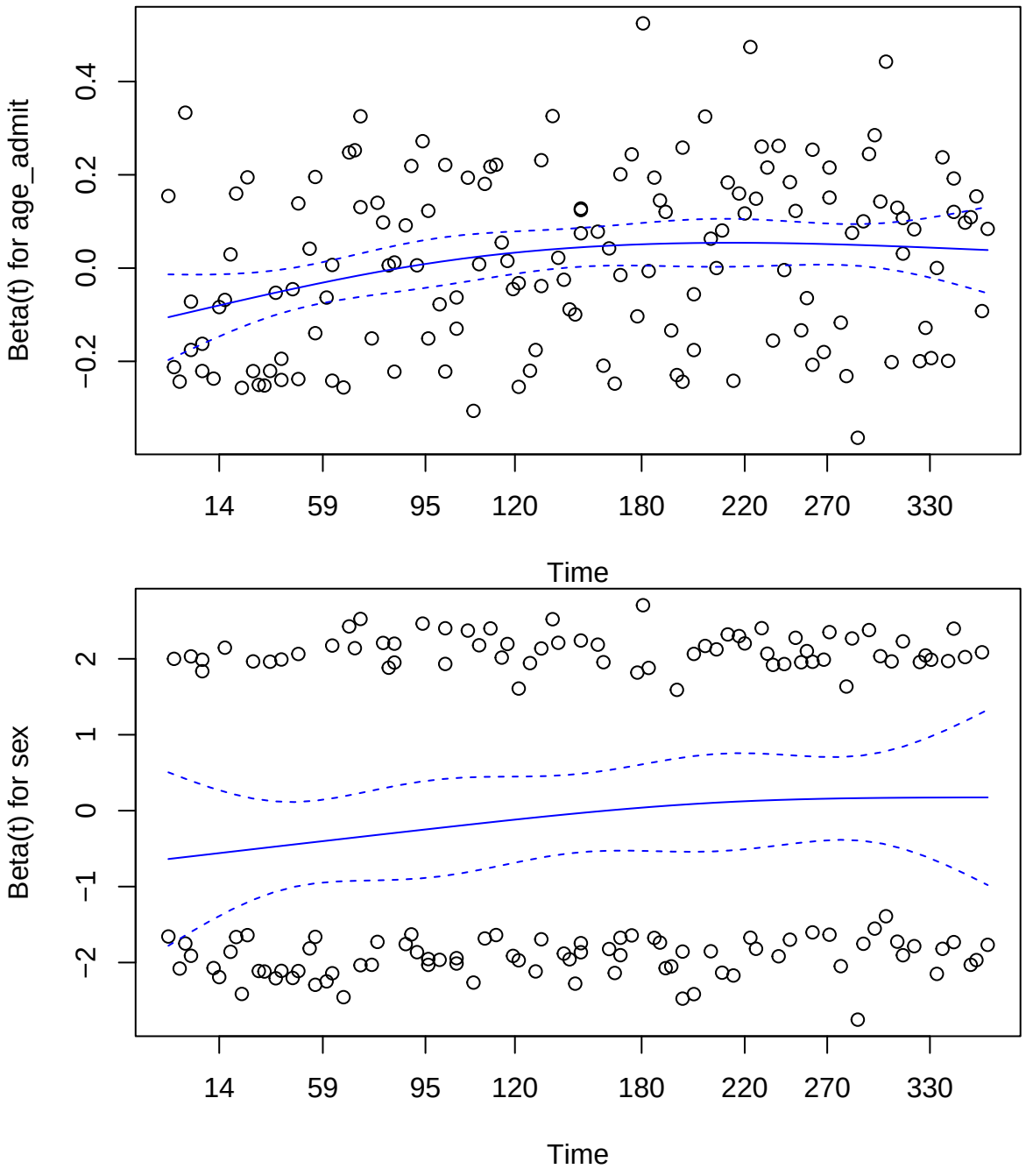


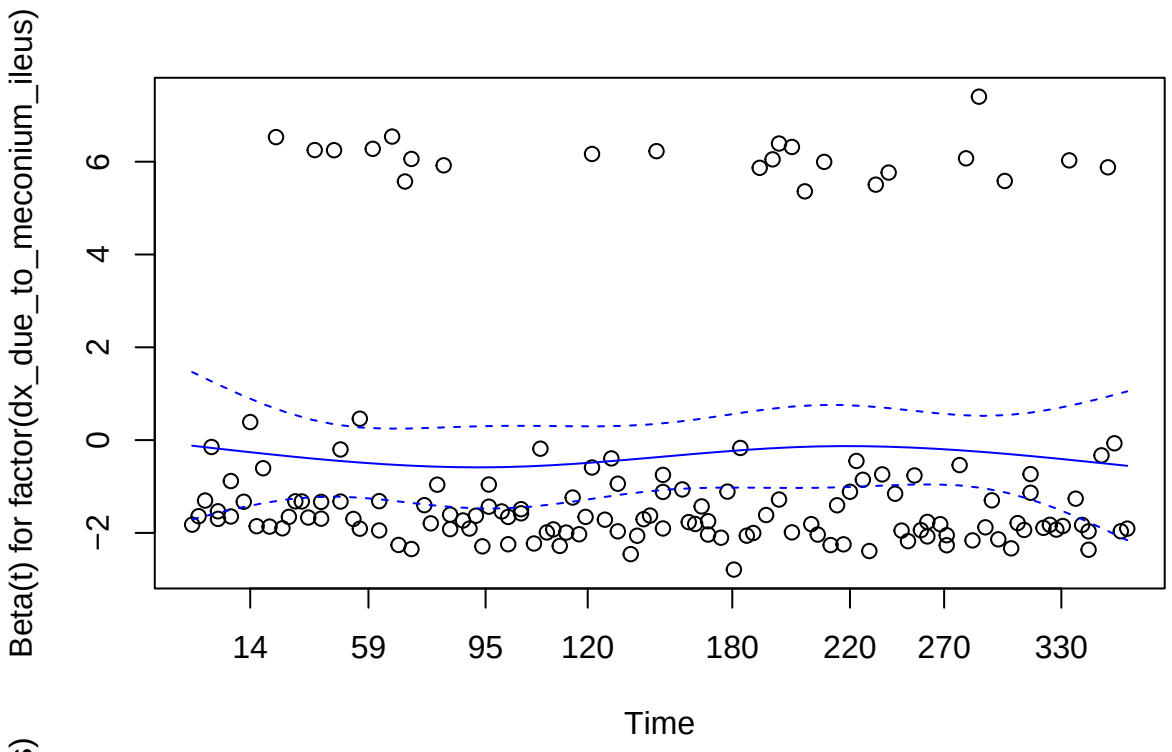
Aim 1: Bayesian AFT Diagnostics - Autocorrelation Plots for Lognormal AFT Model

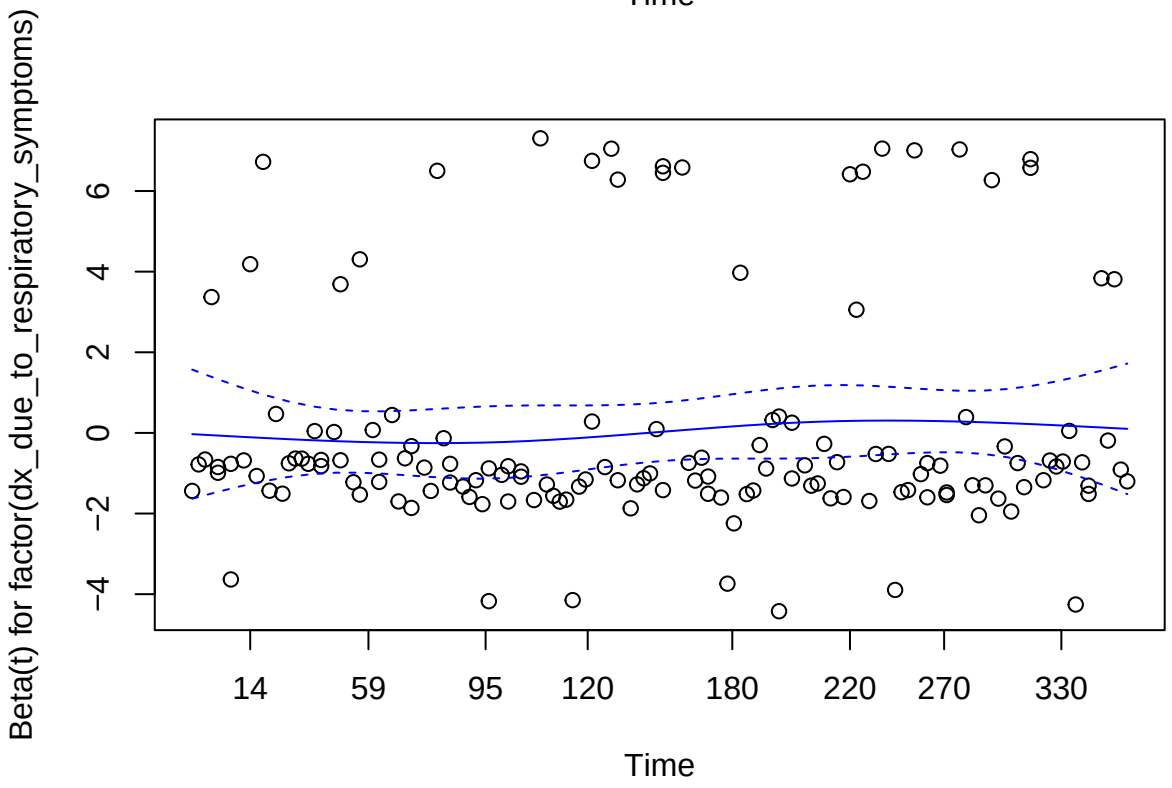


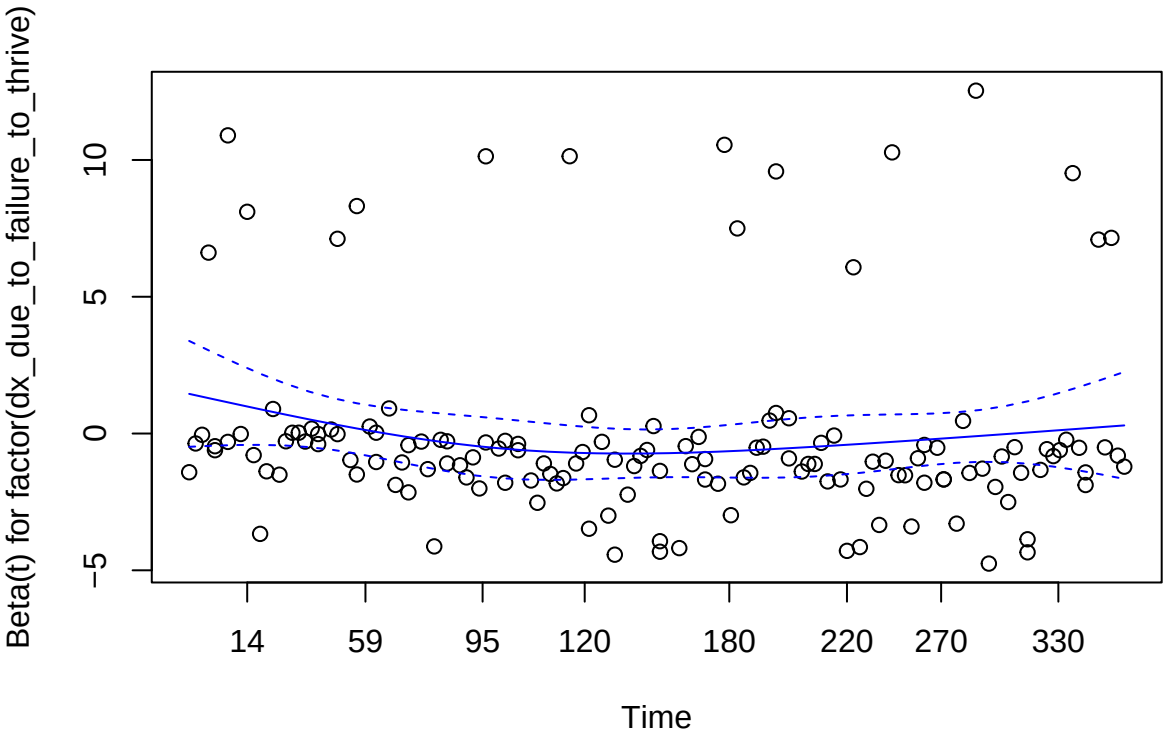
Aim 2: Checking Proportional Hazards Assumptions

```
##                                          chisq df      p
## age_admit                              8.76212  1 0.0031
## sex                                    1.67563  1 0.1955
## factor(dx_due_to_meconium_ileus)       0.10476  1 0.7462
## factor(dx_due_to_respiratory_symptoms) 0.98401  1 0.3212
## factor(dx_due_to_failure_to_thrive)    0.00189  1 0.9653
## GLOBAL                                11.79513  5 0.0377
```

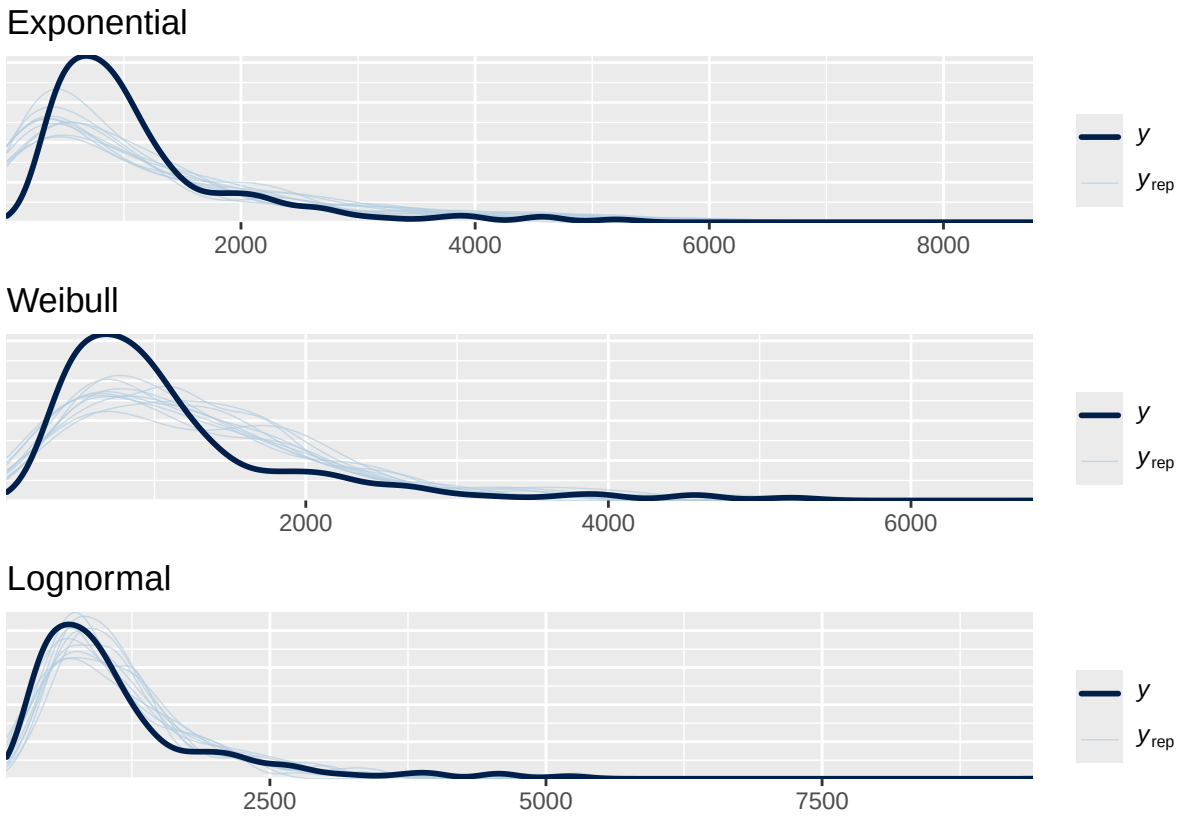

Beta(t) for age_admit
−0.2
0.0
0.2
0.4
14
59
95
120
180
220
270
330
Time
Beta(t) for sex
−2
−1
0
1
2
14
59
95
120
180
220
270
330
Time


Beta(t) for factor(dx_due_to_meconium_ileus)
6
4
2
0
−2
14
59
95
120
180
220
270
330
Time

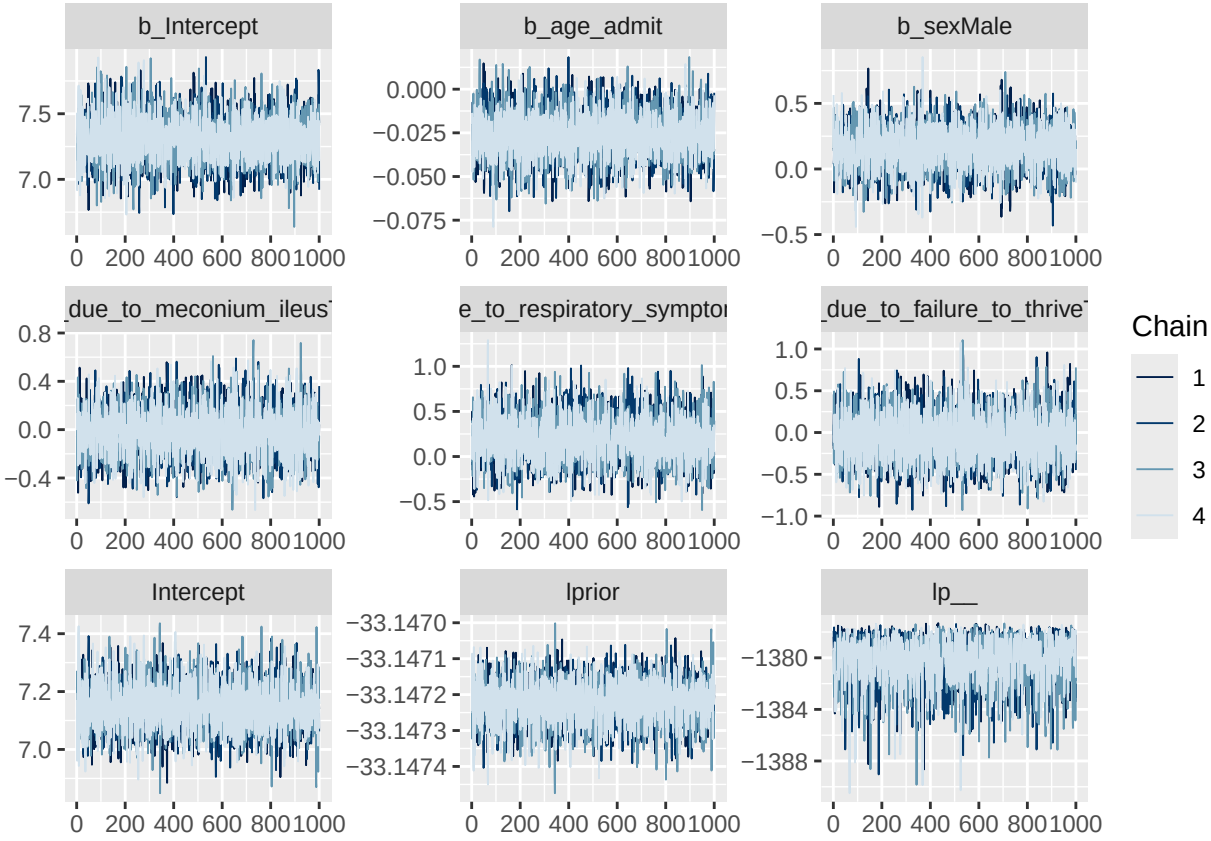

Beta(t) for factor(dx_due_to_respiratory_symptoms)
6
4
2
0
−2
−4
14
59
95
120
180
220
270
330
Time



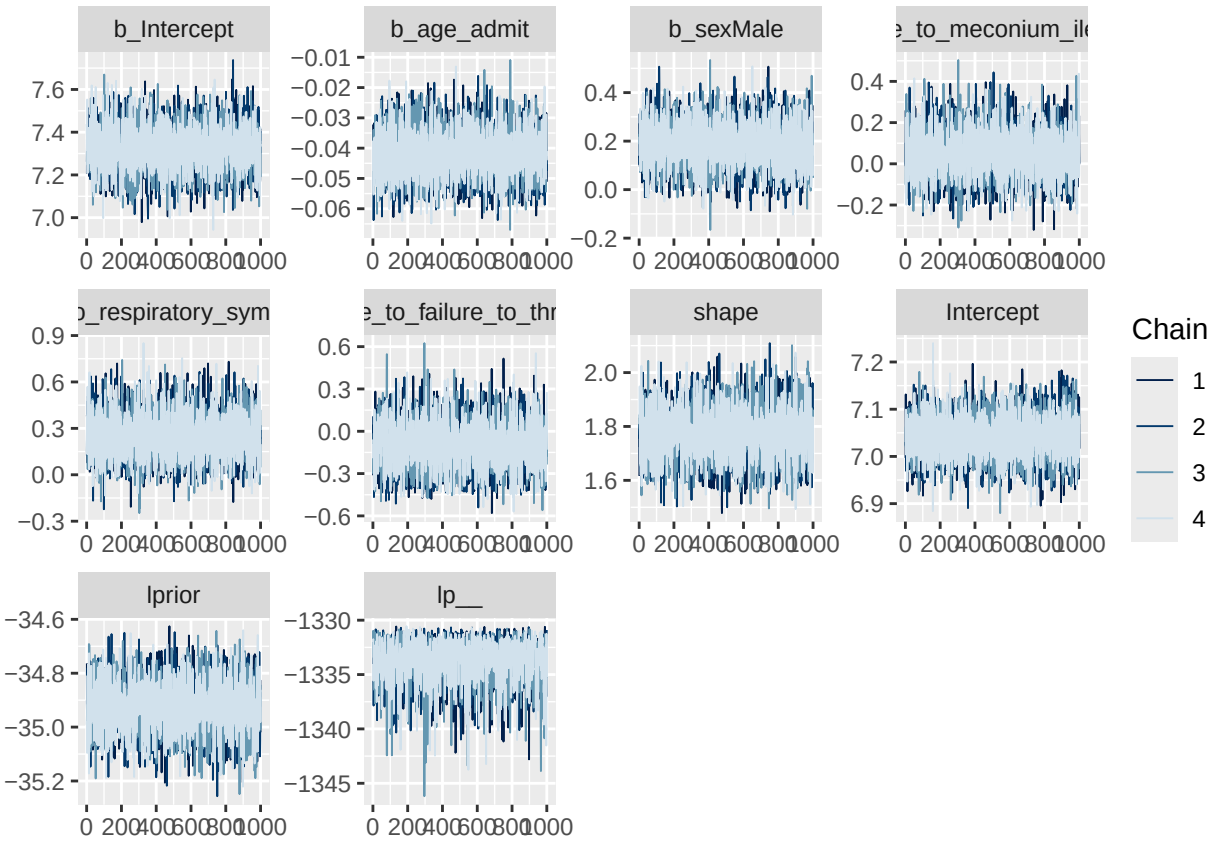
Aim 2: Bayesian AFT Diagnostics - Posterior Predictive Plots



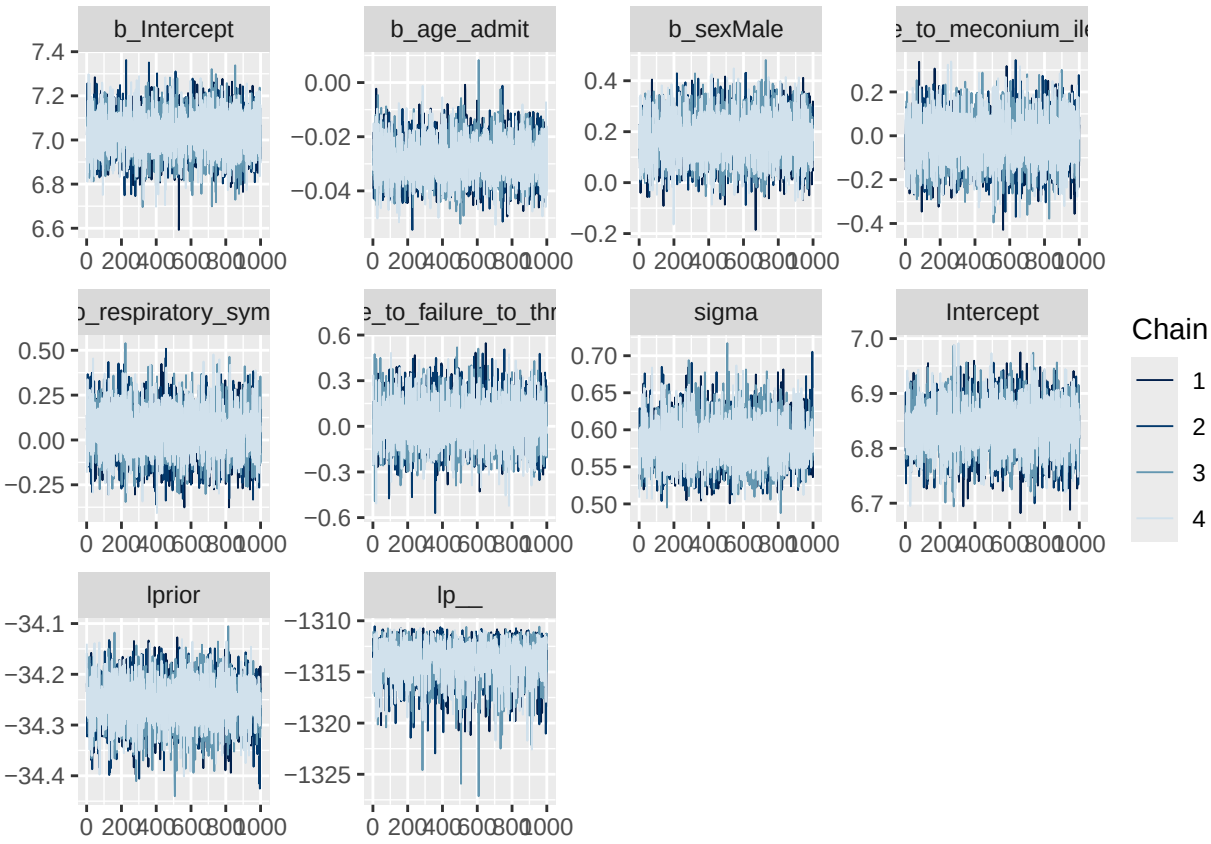
Aim 2: Bayesian AFT Diagnostics - Trace Plots for Exponential AFT



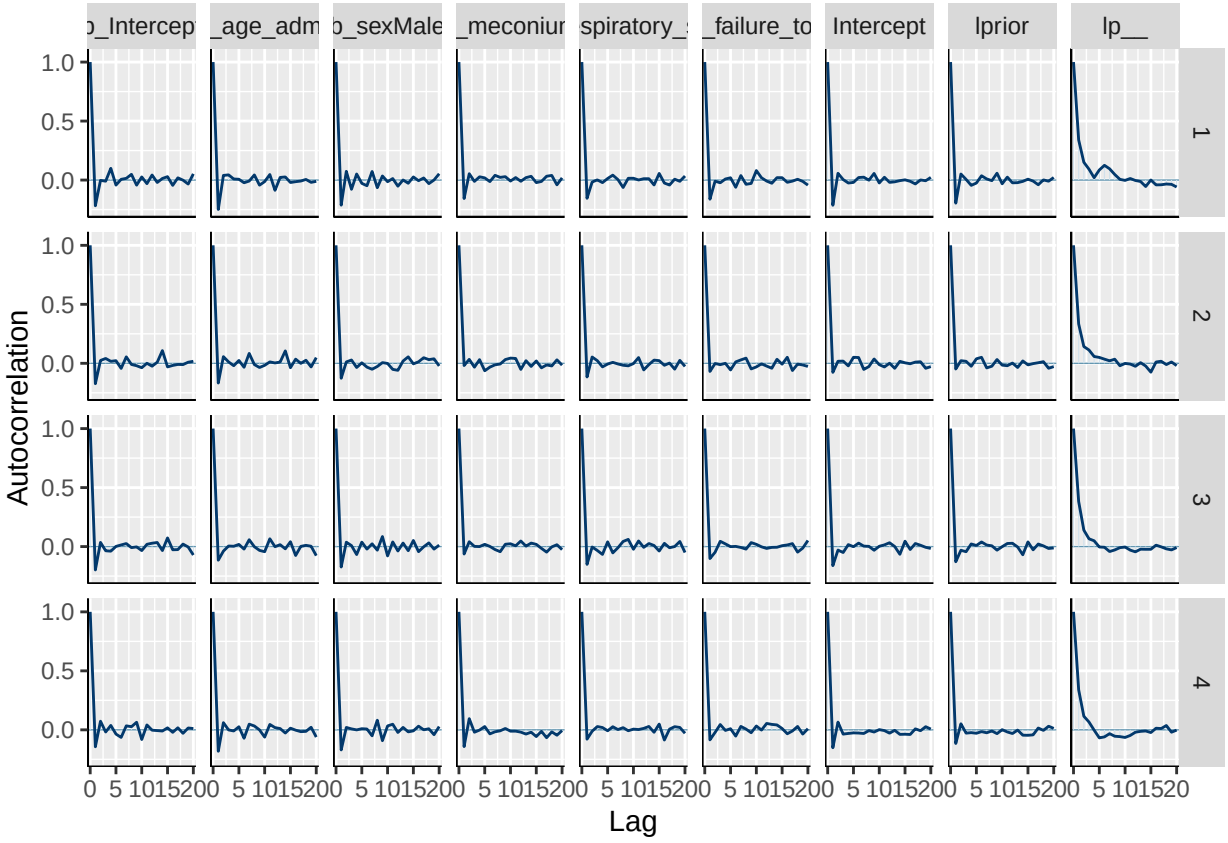
Aim 2: Bayesian AFT Diagnostics - Trace Plots for Weibull AFT



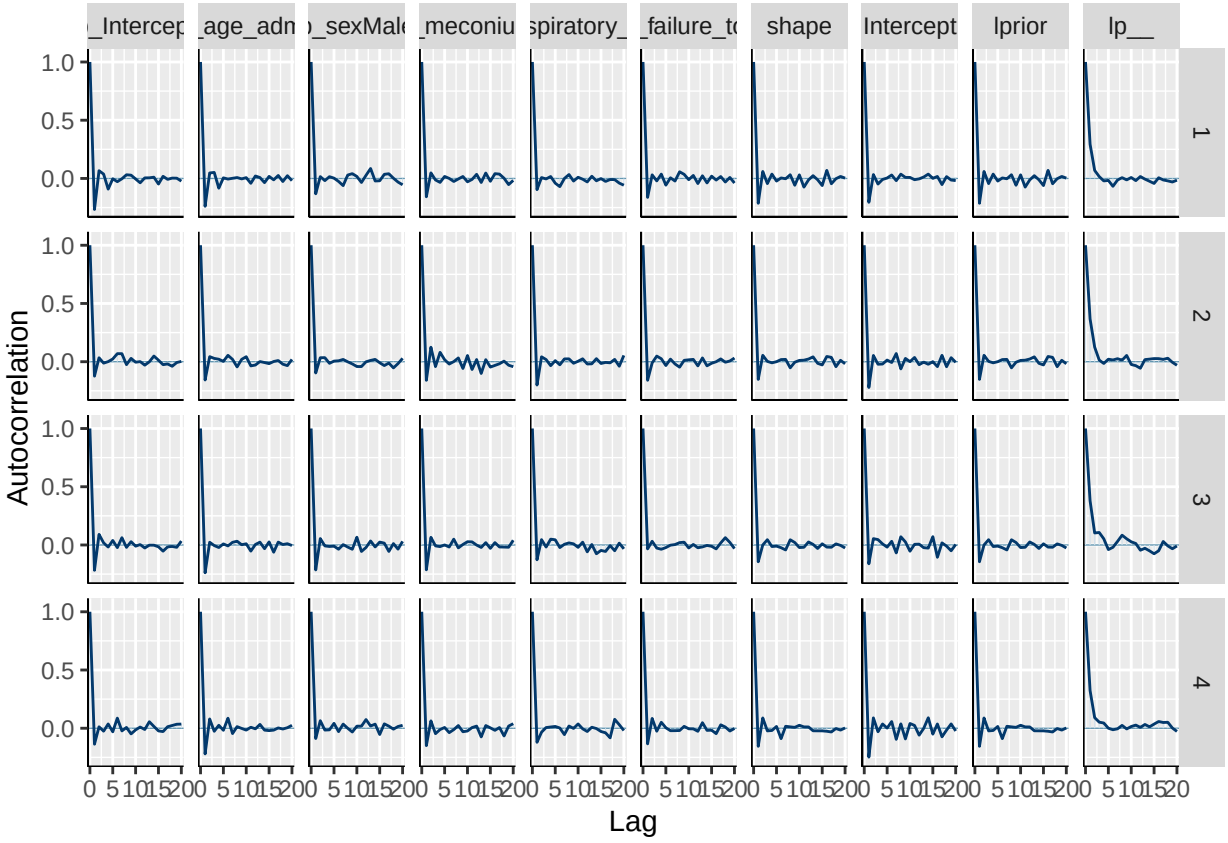
Aim 2: Bayesian AFT Diagnostics - Trace Plots for Lognormal AFT



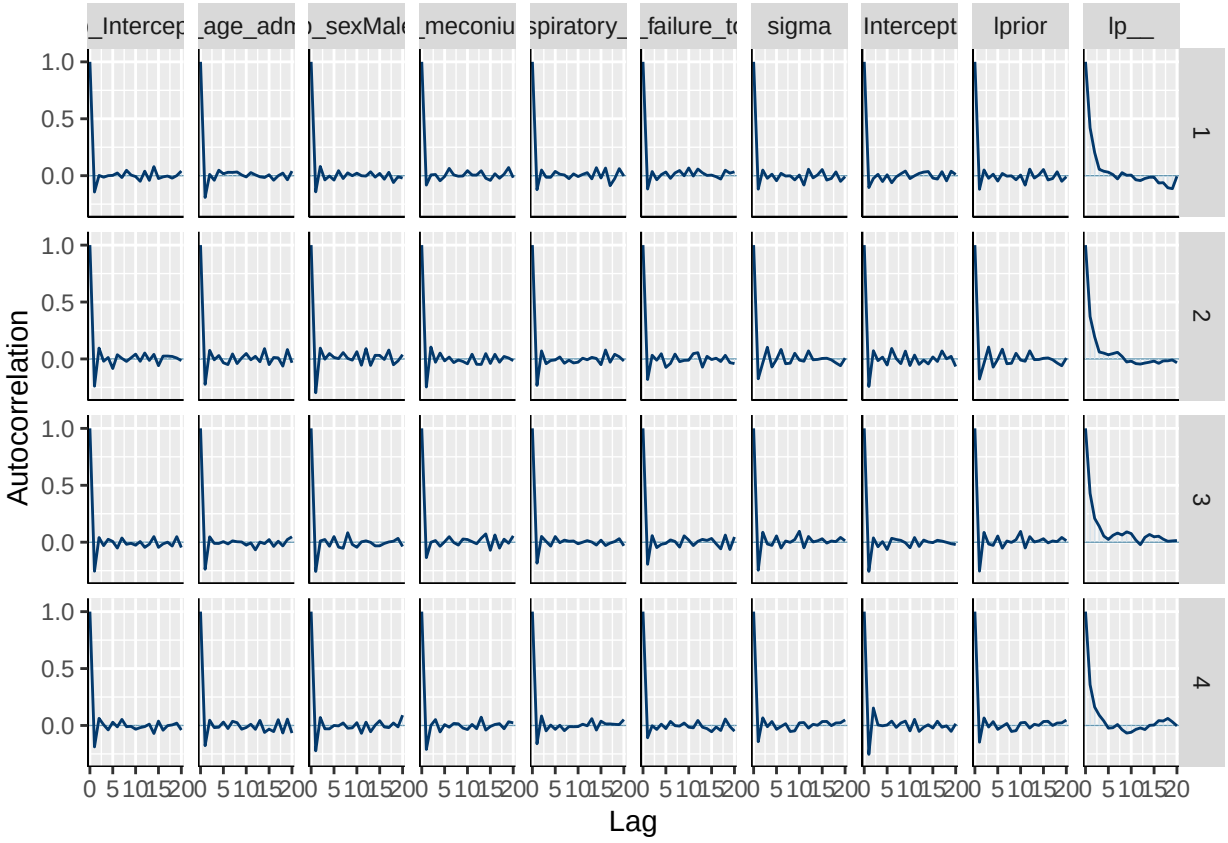
Aim 2: Bayesian AFT Diagnostics - Autocorrelation Plots for Exponential AFT



Aim 2: Bayesian AFT Diagnostics - Autocorrelation Plots for Weibull AFT



Aim 2: Bayesian AFT Diagnostics - Autocorrelation Plots for Lognormal AFT



Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# libraries
library(readxl) # loading Excel datasets
library(janitor) # cleaning dataset variables names
library(tidyverse) # general coding style
library(survival) # analyzing survival data
library(survminer)
library(table1) # creating table ones
library(lubridate) # for handling date formatted data entries
library(brms) # Bayesian survival models
library(posterior) # for extracting draws from bayesian models for plotting trace plots
library(bshazard) # plotting B-splines on hazard to determine hazard shape
library(bayesplot) # for examining Bayesian diagnostic plots
library(kableExtra) # for making tables
library(knitr) # for knitting tables
library(patchwork) # for grouping the posterior predictive plots


# demographics dataset
demo.dat <- read.csv("~/Desktop/BIOS 6646 - Survival/Final Project/Data_Demographics_revised_n

# cultures dataset
cultures.dat <- read_excel("~/Desktop/BIOS 6646 - Survival/Final Project/For Brandie_All Cultu

# Pulmonary exacerbations admits dataset
admits.dat <- read_excel("~/Desktop/BIOS 6646 - Survival/Final Project/For Brandie_All Admits

# clean variable names
demo.dat <- janitor::clean_names(demo.dat)
cultures.dat <- janitor::clean_names(cultures.dat)
admits.dat <- janitor::clean_names(admits.dat)


##############################################################################

### STAGE 1: DATA CLEANING OF EXISTING DATASETS

## (1) Cleaning the demographics dataset
# add a variable denoting genotype status
demo.dat <- demo.dat %>% mutate(genotype_status = case_when(
  genotype_1 == "" & genotype_2 == "" ~ "Unknown",
  (genotype_1 == "" | genotype_2 == "") ~ "Unknown",
  is.na(genotype_1) | is.na(genotype_2) ~ "Unknown",
  (genotype_1 == "UNK" | genotype_2 == "UNK") ~ "Unknown",
```

```
  genotype_1 == genotype_2 & genotype_1 != "" & genotype_2 != "" ~ "Homozygous",
  genotype_1 != genotype_2 ~ "Heterozygous",
  TRUE ~ "unknown"
))

# fix those blank 'race' entries
demo.dat <- demo.dat %>% mutate(race = case_when(race == "" ~ "Unknown",
                                                 is.na(race) ~ "Unknown",
                                                 TRUE ~ race))
#table(demo.dat$race) # verify that the variable manipulation worked


## (2) Cleaning the 'admits.dat' dataset
# order hospital admit date in chronological order by MRN
admits.dat <- admits.dat %>% arrange(mrn, admit_date)

# NOTE: 3 admits/hospital stays with negative elapsed days spent in the hospital (discharge da
# omit these 3 hospital stays from the analysis
admits.dat <- admits.dat %>% filter(days_admitted >= 0)

# create a 'discharge_date' variable
admits.dat <- admits.dat %>%
  dplyr::mutate(discharge_date = admit_date + lubridate::days(days_admitted))
  #dplyr::mutate(discharge_date = admit_date + days_admitted * 86400) # multiply by seconds in

# determine number of days until next hospital admission from the previous one
admits.dat <- admits.dat %>%
  arrange(mrn, admit_date) %>%
  group_by(mrn) %>%  # Group by patient_id
  mutate(days_elapsed = ifelse(lead(admit_date) < discharge_date, 0, as.numeric(lead(admit_date
         days_elapsed = ifelse(days_elapsed > 10000, days_elapsed / 86400, days_elapsed)) %>%
  ungroup()
# note: a few observations had patients with >1 hospitalization at the same time. Those negati
# note: About 180 observations were still calculated as seconds, and thus had days_elapsed in
# I corrected these observations by:
#   (1) Setting the few observations with lead admit dates earlier than the discharge date to
#   (2) Converting those lengthy days_elapsed numbers from seconds into days (dividing by 8640

# number the hospital visits by unique MRN
admits.dat <- admits.dat %>%
  dplyr::group_by(mrn) %>%
  dplyr::mutate(visit_number = row_number()) %>%
  ungroup()


################################################################################
```

```r
### STAGE 2: FORMING THE AIM 1 DATASET: CF First Hospitalizations

# Now, we will shift our focus to formulating the CF dataset. This will involve:
# (1) Restricting our attention to the 824 patients diagnosed with CF
# (2) Merging their hospital admissions and demographics data together
#   a. demo.dat: 851 observations (1 row per subject)
#   b. cultures.dat: 23485 observations
#   c. admits.dat: 2557 observations

# Focus on simplifying admit.dat down to first hospital admission only
first.admit.dat <- admits.dat %>%
  dplyr::group_by(mrn) %>%
  summarise(first_admit_date = min(admit_date)) # N=478 unique MRNs

# merge with demo.dat
merged.dat <- demo.dat %>%
  left_join(first.admit.dat, by = "mrn") # keeps all patients, even those with no stays

# convert date format of DOB
merged.dat$dob <- lubridate::mdy(merged.dat$dob) # N=851 unique subjects

# filter to only include CF diagnosed patients
merged.dat <- merged.dat %>% filter(diagnosis == "Cystic Fibrosis") # N=824 subjects

# calculate time to first hospitalization (days? weeks?)
merged.dat <- merged.dat %>%
  dplyr::mutate(time_to_first_hospitalization = as.numeric(difftime(first_admit_date, dob, uni
                event = ifelse(is.na(first_admit_date), 0, 1))

# lastly, there appears to be 7 people with DOB entered with futuristic dates
# this is not possible, and would make sense to ask the clinical investigator,
# or to remove them from the analysis.
#invalid.dob <- merged.dat[as.Date(merged.dat$dob) >= as.Date("2024-11-26"), ] # anything afte
merged.dat <- merged.dat[as.Date(merged.dat$dob) < as.Date("2024-11-26"), ] # filter out those
# now, there are N=817 unique subjects

##############################
# some quick notes:
# (1) 478 patients have had at least one hospital admission/stay.
# (2) In merging these patients with the patients in the demographics dataset, we have 851 tot
#   (a) This means of the 478 patients with at least one hospital stay, 478 - 431 = 47 individ
# (3) Of the 851 total patients, 824 have CF. The remaining 27 patients diagnosed with other d
# (4) Calculated time until first hospitalization (from birth).
#   a. Research shows that most individuals with CF are diagnosed within the first 2 years of
#   b. Research also shows that CF patients can be hospitalized for up to 2 weeks, and hospita
#
##############################
```

```r
##############################
#
# NEXT STEPS/POSSIBLE ANALYSES:
# (1) Once patients are admitted to the hospital; what factors contribute to a longer hospital
#   (a) Perhaps draw the line at 14 days/2 weeks. Of the 2554 hospital admit stays, 353 of the
#     (i) Would be good to determine their age at first discharge/hospitalization.
#   (b) Time to event: Time until discharge from first hospital visit
# (2) - Of those patients that were hospitalized at least once, how long until they experience
#   (a) Time to event: Time until 2nd hospitalization
#     (i) Would be helpful to know their age at 2nd hospital admission/discharge.
#
##############################


# this dataset is purely for investigating length of hospital stay for 1st hospitalization
first.stay.dat <- admits.dat %>% filter(visit_number==1)
first.stay.dat$event <- ifelse(first.stay.dat$days_admitted <= 14, 1, 0)
# note: of the 478 patients with at least 1 hospital stay, 428 of them have a
# first hospital stay less than or equal to 14 days.

# convert demo.dat MRN to numeric
demo.dat$mrn <- as.numeric(demo.dat$mrn) # for merging datasets

# merge with demographics dataset
merged.first.dat <- first.stay.dat %>%
  inner_join(demo.dat, by = "mrn")
# inner join keeps the patients with all demographics and hospital admits data
# N=434 unique MRNs/individuals

merged.first.dat <- merged.first.dat %>%
  dplyr::mutate(dob = lubridate::mdy(dob)) # all 434 DOB's formatted correctly

# calculate age at admit date
merged.first.dat <- merged.first.dat %>%
  dplyr::mutate(age_admit = as.numeric(difftime(admit_date, dob, units = "days")) / 365.25) %>%
  filter(age_admit >= 0) %>% # remove one patient with a birth date in the year 2063 (?)
  dplyr::mutate(age_tertile = ntile(age_admit, 3))

# note: Of the 433 patients, 392 of them were hospitalized for <=14 days
# this means 41 patients were hospitalized >14 days (censored)

##############################################

#unmatched_mrn <- first.stay.dat %>%
  #filter(!mrn %in% demo.dat$mrn) # 44 umatched MRNs with no DOB information.
```

```
################################################################################

### STAGE 3: MAKE TABLE 1 FOR AIM 1

# create table1 for patients with at least 1 hospitalization
# create table1 labels
label(merged.first.dat$sex) <- "Sex"
label(merged.first.dat$age_admit) <- "Age at Admission"
label(merged.first.dat$dx_due_to_meconium_ileus) <- "Meconium Ileus"
label(merged.first.dat$dx_due_to_respiratory_symptoms) <- "Respiratory Symptoms"
label(merged.first.dat$dx_due_to_failure_to_thrive) <- "Failure to Thrive"

# consider the 3 diagnoses (CF, CRMS, CFTR-related disorder)
aim1.table1 <- table1(~age_admit + dx_due_to_meconium_ileus + dx_due_to_respiratory_symptoms +
       data = merged.first.dat, caption = "Demographics for Hospitalized CF Patients")


################################################################################

### STAGE 4: MODEL DIAGNOSTICS FOR AIM 1
### Note: include these plots in Appendix section of report

# compile KM-plots into a list
aim1.km.plots <- list()

# basic KM fit
km.fit1 <- survfit(Surv(days_admitted, event) ~ age_tertile,
                   data = merged.first.dat, type = c("kaplan-meier"))

aim1.km.plots[[1]] <- ggsurvplot(km.fit1, data = merged.first.dat, xlab = "Time (days)",
                                 ylab = "Proportion\nHospitalized", title = "Age at Admission"
                                 xlim = c(0,20), legend = "right",
                                 legend.title = "Age Tertile",
                                 legend.labs = c("1", "2", "3"),
                                 font.main = c(10, "bold"), font.x = c(10, "italic"),
                                 font.y = c(10, "italic"), font.tickslab = 10)

km.fit2 <- survfit(Surv(days_admitted, event) ~ sex,
                   data = merged.first.dat, type = c("kaplan-meier"))

aim1.km.plots[[2]] <- ggsurvplot(km.fit2, data = merged.first.dat,
                                 xlab = "Time (days)",ylab = "Proportion\nHospitalized",
                                 title = "Patient Sex", xlim = c(0,20),
                                 legend = "right", legend.title = "Sex",
                                 legend.labs = c("Female", "Male"),
                                 font.main = c(10, "bold"), font.x = c(10, "italic"),
                                 font.y = c(10, "italic"), font.tickslab = 10)
```

```r
km.fit3 <- survfit(Surv(days_admitted, event) ~ dx_due_to_meconium_ileus,
                   data = merged.first.dat, type = c("kaplan-meier"))

aim1.km.plots[[3]] <- ggsurvplot(km.fit3, data = merged.first.dat, xlab = "Time (days)",
                                 ylab = "Proportion\nHospitalized", title = "Meconium Ileus",
                                 xlim = c(0,20), legend = "right",
                                 legend.title = "Meconium\nIleus",
                                 legend.labs = c("Absent", "Present"),
                                 font.main = c(10, "bold"), font.x = c(10, "italic"),
                                 font.y = c(10, "italic"), font.tickslab = 10)

km.fit4 <- survfit(Surv(days_admitted, event) ~ dx_due_to_respiratory_symptoms,
                   data = merged.first.dat, type = c("kaplan-meier"))

aim1.km.plots[[4]] <- ggsurvplot(km.fit4, data = merged.first.dat, xlab = "Time (days)",
                                 ylab = "Proportion\nHospitalized", title = "Respiratory Illnes
                                 xlim = c(0,20), legend = "right",
                                 legend.title = "Respiratory\nIllness",
                                 legend.labs = c("Absent", "Present"),
                                 font.main = c(10, "bold"), font.x = c(10, "italic"),
                                 font.y = c(10, "italic"), font.tickslab = 10)

km.fit5 <- survfit(Surv(days_admitted, event) ~ dx_due_to_failure_to_thrive,
                   data = merged.first.dat, type = c("kaplan-meier"))

aim1.km.plots[[5]] <- ggsurvplot(km.fit5, data = merged.first.dat, xlab = "Time (days)",
                                 ylab = "Proportion\nHospitalized", title = "Failure to Thrive
                                 xlim = c(0,20), legend = "right",
                                 legend.title = "Failure to\nThrive",
                                 legend.labs = c("Absent", "Present"),
                                 font.main = c(10, "bold"), font.x = c(10, "italic"),
                                 font.y = c(10, "italic"), font.tickslab = 10)

# quick check for distribution of hospital stay admission 14 days or less
#hist(merged.first.dat$days_admitted)
#filtered.stays <- merged.first.dat %>% filter(days_admitted <= 14) # 392 out of 433
#hist(filtered.stays$days_admitted)


# CoxPH models
cox.fit1 <- coxph(Surv(days_admitted, event) ~ age_admit + sex + factor(dx_due_to_meconium_ileu
                  data = merged.first.dat, method = "efron")
#summary(cox.fit1)

# examine whether proportional hazards is met
ph.test <- cox.zph(cox.fit1)
```

```r
# exponential AFT model fit
exponential.fit1 <- survreg(Surv(days_admitted, event) ~ age_admit + sex + dx_due_to_meconium_

# weibull AFT model fit
weibull.fit1 <- survreg(Surv(days_admitted, event) ~ age_admit + sex + dx_due_to_meconium_ileus

# lognormal AFT model fit
lognormal.fit1 <- survreg(Surv(days_admitted, event) ~ age_admit + sex + dx_due_to_meconium_il

# log-logistic AFT model fit
loglogistic.fit1 <- survreg(Surv(days_admitted, event) ~ age_admit + sex + dx_due_to_meconium_

# examine AIC and BIC values across the models
#AIC(exponential.fit1, weibull.fit1, lognormal.fit1, loglogistic.fit1) # AIC

# interpret the Weibull AFT model parameters
#summary(weibull.fit1)
# interpret these in the PH framework (do math. store results in table.)
# weibull scale: e^(-2.20168/.748) = 0.053
# weibull shape: 1/.748 = 1.337


# examine model diagnostics. - which one fits the best?
# use null model fit, which is 'cox.fit1.'
km.cox.fit1 <- survfit(cox.fit1, type = "aalen")

# extract time and survival time
time <- summary(km.cox.fit1)$time
surv.val <- summary(km.cox.fit1)$surv

# extract cumulative hazards for each AFT model type
H <- -log(surv.val) # Exponential
logtime <- log(time) # need to use log-scaled time in some plotting cases
logH <- log(H) # weibull
probH <- qnorm(1-exp(-H)) # lognormal
logeH <- log(exp(H)-1) # log-logistic

# include plots in appendix


# goodness of fit Weibull model (QQ-plot)
#summary(weibull.fit1)

#weibull.fit1$coef[1] # extract intercept
#weibull.fit1$scale # extract scale
# Weibull scale re-parameterization: exp(-mu\sigma)
scale1 <- exp(-weibull.fit1$coef[1]/weibull.fit1$scale)
```

```r
shape1 <- 1/summary(weibull.fit1)$scale # gamma = 1/shape
a1 <- shape1
b1 <- scale1 ^ (-1/shape1)
cdf1 <- 1 - surv.val

# include Weibull QQ-plot in Appendix


# Define the model formula (same as in survreg)
bayes.formula1 <- bf(days_admitted | cens(event) ~ age_admit + sex + dx_due_to_meconium_ileus +
    dx_due_to_respiratory_symptoms + dx_due_to_failure_to_thrive)

# Define priors (adjust these priors based on your knowledge or experimentation)
weakinf_priors <- c(prior(normal(0, 100), class = "b"),         # Priors for regression coeffic
                    prior(normal(0, 100), class = "Intercept"), # Prior for intercept
                    prior(gamma(1, 1), class = "shape"))       # Prior for Weibull shape parame


#inf_priors <- c(prior(normal(0, 2), class = "b"), # priors for regression coefficients
             #prior(normal(0, 5), class = "Intercept"), # Prior for intercept
             #prior(gamma(3, 2), class = "shape") ) # reflects shape parameter k=1.33 roughly

# Fit the Bayesian survival model using brms and weakly informative priors
# Exponential Bayes fit
exp.bayes.fit1 <- brm(formula = bayes.formula1, data = merged.first.dat,
                      family = exponential(),
                      prior = c(prior(normal(0, 100), class = "b"),
                                prior(normal(0, 100), class = "Intercept")),
                      cores = 4, chains = 4, iter = 2000, warmup = 1000)

# Weibull Bayes fit
weib.bayes.fit1 <- brm(formula = bayes.formula1, data = merged.first.dat,
                       family = weibull(),
                       prior = c(prior(normal(0, 100), class = "b"),
                                 prior(normal(0, 100), class = "Intercept"),
                                 prior(gamma(1, 1), class = "shape")),
                       cores = 4, chains = 4, iter = 2000, warmup = 1000)

# Lognormal Bayes fit
lognorm.bayes.fit1 <- brm(formula = bayes.formula1, data = merged.first.dat,
                          family = lognormal(),
                          prior = c(prior(normal(0, 100), class = "b"),
                                    prior(normal(0, 100), class = "Intercept"),
                                    prior(lognormal(0, 2), class = "sigma")),
                          cores = 4, chains = 4, iter = 2000, warmup = 1000)

invisible(bayes.formula1)
```

```r
invisible(weakinf_priors)
invisible(exp.bayes.fit1)
invisible(weib.bayes.fit1)
invisible(lognorm.bayes.fit1)

#bayes_fit <- brm(formula = bayes.formula1, data = merged.first.dat, family = weibull(), prior

# Check model summary
#summary(exp.bayes.fit1)
#summary(weib.bayes.fit1)
#summary(lognorm.bayes.fit1)
#loo_compare(waic(exp.bayes.fit1), waic(weib.bayes.fit1), waic(lognorm.bayes.fit1))

# extracting WAIC values from models (use this instead of elapsed/SE differences)
#waic(exp.bayes.fit1)$waic # 474.538
#waic(weib.bayes.fit1)$waic # 420.346
#waic(lognorm.bayes.fit1)$waic # 370.761


# create data frame of results:
fit1.df <- data.frame(Term = c("Age at Admission", "Sex (Male)", "Meconium Ileus", "Respiratory
             hr.freq = c(0.980, 1.229, 0.803, 0.478, 0.632),
             ci.freq = c("(0.969, 0.990)", "(1.054, 1.434)", "(0.653, 0.986)", "(0.364, 0.629)"
             p = c("0.008", "0.049", "0.118", "< 0.001", "0.021"),
             hr.bayes = c(0.955, 0.911, 0.465, 0.136, 0.260),
             ci.bayes = c("(0.936, 0.974)", "(0.666, 1.247)", "(0.302, 0.716)", "(0.063, 0.292)
             rhat = c(rep("1.00", 5)))
colnames(fit1.df) <- c("Variable", "Hazard Ratio", "95% Confidence Interval", "P-Value", "Hazar


################################################################################

### STAGE 5: CREATING DATASET FOR AIM 2 (Time until 2nd hospitalization)
# filter out first and second hospitalization data
second.admit.full.dat <- merged.first.dat %>%
  filter(!is.na(days_elapsed)) %>%
  dplyr::mutate(hosp_event = ifelse(days_elapsed <= 365, 1, 0))

# make 'hosp_event' variable numeric
second.admit.dat <- second.admit.full.dat %>% filter(days_elapsed > 0)

# now include table1 for aim1
t1kable(aim1.table1) %>%
  kable_styling(latex_options = "HOLD_position")


# make aim2 table1
```

```r
aim2.table1 <- table1(~age_admit + dx_due_to_meconium_ileus + dx_due_to_respiratory_symptoms +
       data = second.admit.full.dat, caption = "Demographics for CF Patients with 2 or more hos

# now include table1 for aim2
t1kable(aim2.table1) %>%
  kable_styling(latex_options = "HOLD_position")


# make presentation-worthy table for aim1 results
kbl(fit1.df, align = 'lcccccc', caption = "Weibull AFT Model Results: Length of First Hospital
  kable_styling(full_width = FALSE, position = "center", font_size = 12,
                bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = c("scale_down", "HOLD_position")) %>%
  add_header_above(c(" " = 1, "Frequentist" = 3, "Bayesian" = 3)) %>%
  column_spec(1, border_right = T) %>%
  column_spec(4, border_right = T)


# data frame for AIC/WAIC values
#AIC(exponential.fit1, weibull.fit1, lognormal.fit1, loglogistic.fit1) # AIC

fit1.compare.df <- data.frame(Model = c("Exponential", "Weibull", "Lognormal", "Log-Logistic")
                              aic = c(2650.690, 2595.374, 2450.727, 2412.904), # extracted fro
                              waic = c(474.538, 420.346, 370.761, NA)) # extracted from waic()
colnames(fit1.compare.df) <- c("Model Type", "AIC", "WAIC")

# combine this with aim 2 metrics in 1 table (later in code).


# fit Cox model and assess PH assumptions
cox.fit2 <- coxph(Surv(days_elapsed, hosp_event) ~age_admit + sex + factor(dx_due_to_meconium_
                  data = second.admit.dat, method = "efron")
#summary(cox.fit1)

# examine whether proportional hazards is met
ph.test2 <- cox.zph(cox.fit2)


# exponential AFT model fit
exponential.fit2 <- survreg(Surv(days_elapsed, hosp_event) ~ age_admit + sex + dx_due_to_meconi

# weibull AFT model fit
weibull.fit2 <- survreg(Surv(days_elapsed, hosp_event) ~ age_admit + sex + dx_due_to_meconium_

# lognormal AFT model fit
lognormal.fit2 <- survreg(Surv(days_elapsed, hosp_event) ~ age_admit + sex + dx_due_to_meconiu
```

```r
# log-logistic AFT model fit
loglogistic.fit2 <- survreg(Surv(days_elapsed, hosp_event) ~ age_admit + sex + dx_due_to_meconi

# examine AIC and BIC values across the models
#AIC(exponential.fit2, weibull.fit2, lognormal.fit2, loglogistic.fit2) # AIC

# interpret the Weibull AFT model parameters
#summary(weibull.fit2)
# interpret these in the PH framework (do math. store results in table.)
# weibull scale: e^(-7.95742/.1.82) = 0.013
# weibull shape: 1/.748 = 0.549


# Define the model formula (same as in survreg)
bayes.formula2 <- bf(days_elapsed | cens(hosp_event) ~ age_admit + sex + dx_due_to_meconium_ile
    dx_due_to_respiratory_symptoms + dx_due_to_failure_to_thrive)

#inf_priors <- c(prior(normal(0, 2), class = "b"), # priors for regression coefficients
             #prior(normal(0, 5), class = "Intercept"), # Prior for intercept
             #prior(gamma(3, 2), class = "shape") ) # reflects shape parameter k=1.33 roughly

# Fit the Bayesian survival model using brms and weakly informative priors
# Exponential Bayes fit
exp.bayes.fit2 <- brm(formula = bayes.formula2, data = second.admit.dat,
                      family = exponential(),
                      prior = c(prior(normal(0, 100), class = "b"),
                                prior(normal(0, 100), class = "Intercept")),
                      cores = 4, chains = 4, iter = 2000, warmup = 1000)

# Weibull Bayes fit
weib.bayes.fit2 <- brm(formula = bayes.formula2, data = second.admit.dat,
                       family = weibull(),
                       prior = c(prior(normal(0, 100), class = "b"),
                                 prior(normal(0, 100), class = "Intercept"),
                                 prior(gamma(1, 1), class = "shape")),
                       cores = 4, chains = 4, iter = 2000, warmup = 1000)

# Lognormal Bayes fit
lognorm.bayes.fit2 <- brm(formula = bayes.formula2, data = second.admit.dat,
                          family = lognormal(),
                          prior = c(prior(normal(0, 100), class = "b"),
                                    prior(normal(0, 100), class = "Intercept"),
                                    prior(lognormal(0, 2), class = "sigma")),
                          cores = 4, chains = 4, iter = 2000, warmup = 1000)

# hide weird text output in PDF knitted document
invisible(bayes.formula2)
```

```r
invisible(exp.bayes.fit2)
invisible(weib.bayes.fit2)
invisible(lognorm.bayes.fit2)

#bayes_fit <- brm(formula = bayes.formula1, data = merged.first.dat, family = weibull(), prior

# Check model summary
#summary(exp.bayes.fit2)
#summary(weib.bayes.fit2)
#summary(lognorm.bayes.fit2)
# Rhat = 1 across all coefficients, suggesting convergence is met.

# obtain WAIC values
#waic(exp.bayes.fit2)$waic # 2697.038
#waic(weib.bayes.fit2)$waic # 2608.904
#waic(lognorm.bayes.fit2)$waic # 2565.090


# create data frame of results:
fit2.df <- data.frame(Term = c("Age at Admission", "Sex (Male)", "Meconium Ileus", "Respiratory
             hr.freq = c(1.034, 0.861, 0.691, 0.995, 1.154),
             ci.freq = c("(0.985, 1.084)", "(0.473, 1.566)", "(0.300, 1.591)", "(0.431, 2.296)"
             p = c("0.014", "0.372", "0.114", "0.984", "0.615"),
             hr.bayes = c(1.073, 0.727, 0.915, 0.642, 1.173),
             ci.bayes = c("(1.053, 1.095)", "(0.610, 0.867)", "(0.738, 1.136)", "(0.488, 0.845)
             rhat = c(rep("1.00", 5)))
colnames(fit2.df) <- c("Variable", "Hazard Ratio", "95% Confidence Interval", "P-Value", "Hazar

# make presentation-worthy table
kbl(fit2.df, align = 'lccccc', caption = "Weibull AFT Model Results: Time Until 2nd Hospitaliz
  kable_styling(full_width = FALSE, position = "center", font_size = 12,
                bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = c("scale_down", "HOLD_position")) %>%
  add_header_above(c(" " = 1, "Frequentist" = 3, "Bayesian" = 3)) %>%
  column_spec(1, border_right = T) %>%
  column_spec(4, border_right = T)


# data frame for AIC/WAIC values
#AIC(exponential.fit2, weibull.fit2, lognormal.fit2, loglogistic.fit2) # AIC
#loo_compare(waic(exp.bayes.fit2), waic(weib.bayes.fit2), waic(lognorm.bayes.fit2)) # WAIC

fit2.compare.df <- data.frame(aic = c(2407.779, 2320.276, 2291.518, 2300.775), # extracted fro
                              waic = c(2697.038, 2608.904, 2565.090, NA)) # extracted from WAI
colnames(fit2.compare.df) <- c("AIC", "WAIC")

# metrics table column bind with 1st aim
```

```r
fit.metrics <- cbind(fit1.compare.df, fit2.compare.df)

# make presentation-worthy table
kbl(fit.metrics, align = 'lcccc', caption = "AFT Model Performance Results for Aims 1 and 2") %
  kable_styling(full_width = FALSE, position = "center", font_size = 12,
                bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = c("scale_down", "HOLD_position")) %>%
  add_header_above(c(" " = 1, "Aim 1: Length of first Hospital Stay" = 2,
                     "Aim 2: Time from Discharge to Readmission" = 2)) %>%
  column_spec(1, border_right = T) %>%
  column_spec(3, border_right = T)


# first few rows of original hospital admissions dataset
head(admits.dat)


# first few rows of original cultures dataset
head(cultures.dat)


# first few rows of demographics dataset
head(demo.dat)


# first few rows of length of hospital stay data (Aim 1)
head(merged.first.dat)


# first few rows of Aim2 dataset
head(second.admit.dat)


################################################################################

# TABLE 1 (overall population)

# use 'demo.dat' demographics dataset to create these
# create table1 labels
label(demo.dat$genotype_status) <- "Genotype Status"
label(demo.dat$sex) <- "Sex"
label(demo.dat$race) <- "Race"
label(demo.dat$hispanic_latinx) <- "Hispanic/Latinx"
label(demo.dat$dx_due_to_meconium_ileus) <- "Meconium Ileus"
label(demo.dat$dx_due_to_respiratory_symptoms) <- "Respiratory Symptoms"
label(demo.dat$dx_due_to_failure_to_thrive) <- "Failure to Thrive"
```

```r
# consider the 3 diagnoses (CF, CRMS, CFTR-related disorder)
table1.overall <- table1(~sex + race + hispanic_latinx + genotype_status + dx_due_to_meconium_

t1kable(table1.overall) %>%
  kable_styling(latex_options = "HOLD_position")

################################################################################


# Aim 1 K-M plots: arrange plots in one figure
arrange_ggsurvplots(aim1.km.plots, ncol = 2, nrow = 3)


ph.test # global PH assumptions test
plot(ph.test, resid = TRUE, se = TRUE, col = "blue") # Schoenfeld resids plots


par(mfrow=c(2,2)) # place the plots in a 2x2 matrix for presentation
plot(time, H, type = "p", pch = "*", xlab = "time", yla = "Exponential CH Fit")

plot(log(time), logH, type = "p", pch = "*", xlab = "log time", ylab = "Weibull CH Fit")

plot(log(time), probH, type = "p", pch = "*", xlab = "log time", ylab = "Lognormal CH Fit")

plot(log(time), logeH, type = "p", pch = "*",xlab="log time", ylab="log-logistic CH Fit")


# plot QQ-plot
qqplot(qweibull(cdf1, a1, b1), time, xlab="Weibull
quantile(lambda=0.053, gamma=1.337)", ylab = "Time (Days)")
mtext("Weibull Q-Q Plot\n")


exp.bayes1.ppcheck <- pp_check(exp.bayes.fit1) +
  ggtitle("Exponential")

weib.bayes1.ppcheck <- pp_check(weib.bayes.fit1) +
  ggtitle("Weibull")

lognorm.bayes1.ppcheck <- pp_check(lognorm.bayes.fit1) +
  ggtitle("Lognormal")

# arrange PP checks plots
combined.aim1.ppchecks <- (exp.bayes1.ppcheck / weib.bayes1.ppcheck / lognorm.bayes1.ppcheck)

# print plots
combined.aim1.ppchecks
```

```r
# lognormal model does the best job of overlaying the observed outcomes with the
# simulated Bayesian model fit outcomes
# yrep: indicates modeled outcomes
# y: actual, observed outcomes

# examine trace plots of model fits
# start with exponential fit
exp.bayes.fit1.draws <- posterior_samples(exp.bayes.fit1)
mcmc_trace(as.array(exp.bayes.fit1), pars = names(exp.bayes.fit1.draws))


# then perform on weibull fit
weib.bayes.fit1.draws <- posterior_samples(weib.bayes.fit1)
mcmc_trace(as.array(weib.bayes.fit1), pars = names(weib.bayes.fit1.draws))


# lastly, perform on lognormal fit
lognorm.bayes.fit1.draws <- posterior_samples(lognorm.bayes.fit1)
mcmc_trace(as.array(lognorm.bayes.fit1), pars = names(lognorm.bayes.fit1.draws))

# notes:
# lprior: Indicates how likely the parameter values are under the specified prior distribution.
# If it fluctuates, it might indicate improper priors, or some numerical instability in evalua
# lp_: total log posterior probability density for the model, combining the prior, likelihood,
# Summarizes the overall "fit" of the model in terms of the posterior density.


# assess autocorrelation plots
# exponential AFT
mcmc_acf(as.array(exp.bayes.fit1), pars = names(exp.bayes.fit1.draws))


# weibull AFT
mcmc_acf(as.array(weib.bayes.fit1), pars = names(weib.bayes.fit1.draws))


# lognormal AFT
mcmc_acf(as.array(lognorm.bayes.fit1), pars = names(lognorm.bayes.fit1.draws))


ph.test2 # global test for PH assumption
plot(ph.test2, resid = TRUE, se = TRUE, col = "blue") # Schoenfeld resids plots


# bayesian survival model diagnostics
exp.bayes2.ppcheck <- pp_check(exp.bayes.fit2) + ggtitle("Exponential")
```

```r
weib.bayes2.ppcheck <- pp_check(weib.bayes.fit2) + ggtitle("Weibull")
lognorm.bayes2.ppcheck <- pp_check(lognorm.bayes.fit2) + ggtitle("Lognormal")
# lognormal model does the best job of overlaying the observed outcomes with the
# simulated Bayesian model fit outcomes
# yrep: indicates modeled outcomes
# y: actual, observed outcomes

# arrange PP checks plots
combined.aim2.ppchecks <- (exp.bayes2.ppcheck / weib.bayes2.ppcheck / lognorm.bayes2.ppcheck)

# print plots
combined.aim2.ppchecks


# examine trace plots of model fits
# start with exponential fit
exp.bayes.fit2.draws <- posterior_samples(exp.bayes.fit2)
mcmc_trace(as.array(exp.bayes.fit2), pars = names(exp.bayes.fit2.draws))


# then perform on weibull fit
weib.bayes.fit2.draws <- posterior_samples(weib.bayes.fit2)
mcmc_trace(as.array(weib.bayes.fit2), pars = names(weib.bayes.fit2.draws))


# lastly, perform on lognormal fit
lognorm.bayes.fit2.draws <- posterior_samples(lognorm.bayes.fit2)
mcmc_trace(as.array(lognorm.bayes.fit2), pars = names(lognorm.bayes.fit2.draws))

# notes:
# lprior: Indicates how likely the parameter values are under the specified prior distribution
# If it fluctuates, it might indicate improper priors, or some numerical instability in evalua
# lp_: total log posterior probability density for the model, combining the prior, likelihood,
# Summarizes the overall "fit" of the model in terms of the posterior density.


# assess autocorrelation plots
# exponential AFT
mcmc_acf(as.array(exp.bayes.fit2), pars = names(exp.bayes.fit2.draws))


# weibull AFT
mcmc_acf(as.array(weib.bayes.fit2), pars = names(weib.bayes.fit2.draws))


# lognormal AFT
mcmc_acf(as.array(lognorm.bayes.fit2), pars = names(lognorm.bayes.fit2.draws))
```