Evaluation of Methods for the Statistical Analysis of Exacerbations in Cystic Fibrosis

Xinyun Dai

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Committee:

Noah Simon

Nicole Mayer Hamblett

Program Authorized to Offer Degree:

Biostatistics - Public Health

University of Washington

Abstract

Evaluation of Methods for the Statistical Analysis of Exacerbations in Cystic Fibrosis

Xinyun Dai

Chair of the Supervisory Committee:
Noah Simon
Department of Biostatistics

Reducing pulmonary exacerbations (PEx) is an important goal in the treatment of Cystic Fibrosis (CF). Endpoints based on PEx are relatively commonly used in the evaluation of new therapies. There are two different ways that one might consider modeling exacerbations: We can model an overall rate (count of PEx in a patient), or model a frequency (time-to-recurrent PEx events). These two endpoints necessitate the use of different types of models.

In this study, we evaluate 7 methodologies for modeling the relationship between treatment and PEx in CF: 4 in the Poisson-modeling family for count outcome, and 3 in the Proportional hazard modeling family for time-to-event outcome. In Chapter 1, we review the theoretical framing of all 7 models, including the assumptions of each model. In Chapter 2 we conduct simulated experiments, where I evaluate and compare the power and type 1 error rates of all 7 models. We

consider scenarios with different amounts of over-dispersion, sample sizes, and censoring schemes. Finally, in Chapter 3, we apply all 7 models to two previous CF trials, and summarize findings on the performance of these methods.

Our results show that for count outcome, Poisson regression is not reliable when the outcome is over-dispersed. Other models from the Poisson-modeling family that deal with over-dispersion all worked well in every scenario considered. We additionally note that the negative binomial model is more computationally efficient than the similar Poisson GLMM (with Gaussian random effects). Furthermore, the negative binomial model explicitly accounts for between-person variability: In contrast the quasi-Poisson only implicitly accounts for that variability. Thus we recommend use of the negative binomial model for modeling PEx as a count-based outcome. For modeling PEx as a time-to-event outcome, we found that results from the Cox-PH model and from the two recurrent event model (AG model and PWP methods) may differ if treatment effect is heterogeneous over first vs recurrent events. In particular, in the two CF trials, it appears that effects on the first event were vastly different than effects on later events. Choice of model, in this case, should be based on two things. First, the scientific question should be considered: Is the investigator interested in the effect of treatment on the first event (in which case we recommend the standard Cox PH model); or the effect averaged over recurrent events as well? (In which case we suggest use of PWP methodology). Second, one must consider the effect of confounding and post-randomization intervention: Patients receive treatment after their first exacerbation, if this is given differently between arms (based on an effect of treatment on first exacerbation), then the standard estimate of effect on later events may be biased. If one expects this bias to be severe, then likely only time-to-first event should be analyzed.

Future work related to this thesis could study effects from administrative decisions on method performance. For example, there are multiple candidate definitions of PEx, and it would be useful to study if choice of PEx definition might affect relative performance of our 7 models.

This thesis offers an overview of some common methods that are appropriate for CF endpoints, and hopefully can help researchers in model selection for statistical analysis in CF trials in the future.

Contents

Introduction

Pulmonary exacerbations (PEx) occur commonly in patients with cystic fibrosis (CF). Indicators of a PE include increased cough and major decrease in lung function. While the definition of a PE is not completely agreed upon (Keene et al. 2008), they generally indicate worsening of the disease. Therefore, decreasing the number of PEx has become an important treatment goal for CF patients.

The selection of an appropriate endpoint is an important part of CF trial analysis. There are several candidate endpoints that might be considered in a CF trial studying the effect of treatment on PE. Most commonly, CF studies have used the time to first PE as the primary endpoint. It is also possible to use an endpoint that considers multiple PE events such as frequency of PEx or rate of PEx: These are usually the preferred primary outcome for other pivotal trials (outside of CF) in the pulmonary division at the FDA. However, in CF pivotal trials, multiple events are rarely considered for the primary endpoint. Studying multiple events as the endpoints may have advantages (for example, making better use of all available data). In cases where treatment reduces the number of PEx in patient who experiences multiple events, an endpoint that takes these multiple events into account may also be more clinically meaningful.

Based on the choice of endpoint, different statistical methods can be used to measure the effect of treatment on PEx of patients with CF. For studies that wish to consider recurrent PE events, "event rate" is generally used as the primary endpoint. In this case Poisson regression is usually recommended, however it does not account for heterogeneity between patients with the same measured covariates. In contrast, an improved model such as Poisson regression with an overdispersion correction (quasi-Poisson model) is better for accounting for departures from the mean-variance relationship assumed by a Poisson model. Further improved models such as the Poisson Generalized Linear Mixed models or negative binomial regression can be more appropriate choices because they explicitly model between-person

variability.

Rather than using exacerbation rate as the primary endpoint, many studies use time-to-first PE event. In such cases it is common to use the Cox Proportional Hazards model for statistical analysis (Cox 1972). However, the Cox-PH model only considers the first event. In clinical trials involving patients with multiple PEx, events following the first will be ignored. If one is interested in modeling the hazard of pulmonary exacerbation, but would like to take into account multiple events, there are alternative models: Andersen-Gill model (Andersen and Gill 1982) and Prentice-Williams-Peterson (PWP) (Prentice, Williams, and Peterson 1981). The Andersen-Gill model is a simple extension of the Cox-PH model that takes recurrent events into account. However, it is based on a strong assumption that all PE events are independent where the hazard of experiencing an event remains the same at any time. The independence assumption may be violated in CF trials: patients who have had an event may be more likely to have another event, or may be less likely to have another one because of the protective nature of the PEx treatment. The PWP method is a stratified adaptation of the Cox model that allows the hazard at time t for the jth event to be conditional on previous events.

The purpose of this thesis is to review these seven analysis methods and compare their effectiveness, with respect to Type I error control and power. This thesis is divided into three chapters. The first will discuss all seven models' characteristics, merits, and assumptions, in particular considering application to CF trials. The second chapter will empirically compare these methods in a number of simulated experiments. In the third chapter, we will study the impact of these statistical methods on clinical trial results using data from two previous clinical trials in CF. The first study was conducted from February 2007 to July 2009 at 40 CF care centers in the United States and Canada, studying the effect of Azithromycin on pulmonary function in patients with CF uninfected with Pseudomonas aeruginosa (Saiman et al. 2010). The second study was conducted in 2018 at 45 CF care centers in the United States, studying Azithromycin for early Pseudomonas infection in CF (Hamblett et al. 2018).

1. Review of Methods

Pulmonary exacerbations (PEx) are episodes of acute worsening of symptoms. PEx have become a key clinical efficacy outcome measure in cystic fibrosis (CF) trials, and decreasing the number of PEx has become an important treatment goal (Hamblett et al. 2013). In CF trials, many endpoints can be selected to study treatments' effects on PEx. In the past, time to the first event was generally used as the endpoint. Focusing on the first PE event has advantages: The endpoint is not impacted by treatment received as a result of the PE, and PE duration does not need to be defined. It also has the obvious disadvantages that focusing only on the first PE event does not make full use of the data. The time until the first PE event may also be less clinically meaningful than evaluating if treatment reduces the number of PEx in a patient.

Multiple events, though rarely used in previous CF trials, are usually the preferred primary outcome for other pivotal trials in the pulmonary division at the FDA. Since some CF patients develop more than one event, multiple events are also valuable endpoints that can make better use of our data.

There are usually two ways of looking at multiple events. The first is to consider the PE rate, where we focus on the count of PE events each patient has experienced. In this case, we will look at treatment effect in terms of a rate ratio. We could alternatively look at the frequency of PEx, where we focus on the time until each of the multiple events occurs. In this case, we will look at treatment effect in terms of a hazard ratio. In this overview, we will be looking at various statistical methods that have been used or could be used to measure the impact of treatment on cystic fibrosis with endpoints that consider multiple PEx. These methodologies can be divided into two families: the Poisson regression family that focuses on the PE rate, including the Poisson regression model, Quasi-Poisson model, Poisson GLMM model, and negative binomial model; and the Proportional Hazards Model family that focuses on the frequency of PEx, including Cox-PH model, Andersen-Gill model,

and Prentice-Williams-Peterson models. We will review their characteristics and discuss each one's merits as well as issues involved.

1.1 Poisson Regression

Let us first consider the number of PEx as our response. The most common method to use in this situation is Poisson regression. Poisson regression is similar to regular multiple regression, except that the dependent variable Y is an observed count that is assumed to follow a Poisson distribution with rate that is a function of the covariates:

$$Pr(Y = y|\mu, t) = \frac{e^{-\mu t}(\mu t)^y}{y!}$$
$$(y = 0, 1, 2...)$$

Where μ is the mean incidence rate of a rare event per unit of exposure. In CF trials, the exposure is the period of time on the trial; we use t to represent the exposure time. In Poisson Regression, the Poisson incidence rate μ is determined by our regressor variables. Thus, the form of the model equation for Poisson regression is as following:

$$log(\mu) = log(t) + X\beta$$

In our data, Y is the count of exacerbations for a given patient, μ is the mean of Y (per unit time), t is the followup time, X is the vector of our p covariates and β is the corresponding vector of p coefficients. In a Poisson regression model, $log(t)$ is also called an offset.

To use Poisson Regression for inference, the following model assumptions are required:

1. The response, Y, is a count per unit of time or space, which follows a Poisson distribution.
2. The observations must be independent of each other.

3. By definition of Poisson distribution, the mean of a Poisson random variable must be equal to its variance.
4. The log of the mean incidence rate, $\log(\mu)$, must be a linear function of X.

In reality, however, the conditional mean and variance of our outcome rarely match. Often due to unmeasured covariates, there is additional heterogeneity between patients that we cannot account for in our regression. This causes overdispersion of our outcome, resulting in data with more variability than would be indicated by a Poisson model. A simple Poisson regression model can modified and improved to account for this mean-variance imbalance. The three methods we consider are the Poisson regression model with overdispersion correction (quasi-Poisson), Poisson linear mixed model, and the negative binomial model.

1.2 Quasi-Poisson model

As we noted, it is common that the conditional variance of our PE count is larger than expected from a Poisson regression. This is known as "overdispersion." If we do not adjust for overdispersion, we will use artificially small standard errors and give artificially small p-values for our model coefficients. This may result in erroneous conclusions. In the context of count data, we might consider a more general mean-variance relationship, where we imagine that the variance is proportional to the mean:

$$Var(Y) = \phi\mu$$

In a quasi-Poisson model, we do not assume that the dependent variable Y follows a specified distribution, we only assume that Y has E(Y)=μ and Var(Y)=$\phi\mu$, where ϕ is called a dispersion parameter. If ϕ=1, we obtain the mean-variance equivalence relationship. If ϕ>1, there is overdispersion. In the overdispersed case, when estimating the coefficients, instead of maximizing the likelihood function derived from the Poisson distribution, we instead

use the Poisson quasi-likelihood. For more information on the Poisson quasi-likelihood see McCullagh (1983). The quasi-Poisson model is modeled similarly to a regular Poisson model:

$$log(\mu) = log(t) + X\beta$$

Where again Y is the count of exacerbations for a given patient, μ is the mean of Y (per unit time), log(t) is the followup time, X is the vector of our p covariates and β is the corresponding vector of p coefficients.

The quasi-Poisson model requires assumptions similar to that of Poisson regression but does not assume that Y follows a specific distribution and does not need the mean to be equal to the variance of our outcome:

1. The response, Y, is a count per unit of time or space.
2. The observations must be independent of each other.
3. The log of the mean incidence rate, log(μ), must be a linear function of X.

The overdispersion correction term in the quasi-Poisson model is a generic correction, but the between-patient variability is not an explicit part of the model. To explicitly model this variability, one might use methods such as the Poisson GLMM or negative binomial model, which will be discussed next.

1.3 Poisson Generalized Linear Mixed Model

The Poisson Generalized Linear Mixed Model (Poisson GLMM model) is a generalization of linear mixed effects regression that allows the linear model to be related to the response variable via a log link. It is an extension of the Poisson GLM that includes both fixed and random effects. The general form of this model (with n observations) is given by:

$$log(\mu) = log(t) + X\beta + Z\gamma$$

Where μ is an n vector containing the means of our outcomes Y (per unit time); t is the exposure time; X is an $n \times p$ matrix with p fixed-effect covariates; β is a $p \times 1$ column vector of the fixed-effect regression coefficients; Z is an $n \times q$ design matrix for the q random effects, and γ is a $q \times 1$ column vector of the random effects (Bruin 2011). We usually assume that γ is drawn from a mean zero normal distribution with some prespecified covariance matrix.

In CF trials where we consider between-person variability, we use a simple form of the Poisson GLMM model:

$$log(\mu) = log(t) + X\beta + z$$

$$or$$

$$\mu = t \times exp(X\beta) \times exp(z)$$

$$z \sim N(0, \sigma^2)$$

Where z is a vector random effects (modeling between-person variability) and follows a normal distribution. Thus exp(z) follows a log-normal distribution.

To use a Poisson GLMM, we require some assumptions similar to that of a regular Poisson model. In addition, we add some flexibility to account for between-person variability (with the random effect), but make an assumption on that shape of the distribution of that effect.

1. The response, Y, is a count per unit of time or space.
2. The observations must be independent of each other.
3. The log of the average incidence rate, log(μ), must be a linear function of X.
4. Conditional on the random effects the response follows a Poisson distribution.
5. The random effects come from a normal distribution on the additive scale (or a log-normal on the multiplicative).

In a CF trial, it is likely that each person has personal differences (not related to measured covariates) that contribute to their chance of developing PEx. These cannot be modeled with fixed effects and thus a regular Poisson regression cannot account for them. However

in a Poisson GLMM, personal differences will get estimated as an additional random effect. This results in a correction for overdispersion (and ensures appropriate confidence interval widths and p-values). While the Poisson GLMM is great for explaining the between-person variabilities, it is often more computationally intensive compared to other methods that allow for overdispersion (Breslow and Clayton 1993).

1.4 Negative Binomial Model

Another approach to modeling overdispersion in count data is to use a negative binomial regression model. Like the Poisson GLMM, this is an extension to a simple Poisson regression model that adds a multiplicative random effect z to account for heterogeneity.

$$\mu = t \times exp(X\beta) \times z$$

$$z \sim Gamma\,(\theta, \theta)$$

The random effect z is assumed to have a gamma distribution, $z \sim Gamma(\theta, \theta), \theta > 0$ (shape and rate). E(z)=1 and Var(z)=θ^{-1}. This is quite similar to our Poisson GLMM model, only now our multiplicative random effect comes from a Gamma distribution rather than log-normal. This is primarily done for computational reasons: The multiplicative convolution of Poisson and gamma distributions is precisely a negative binomial distribution. So this is

referred to as negative binomial regression. For completeness, we give the derivation below.

$$h(Y = y|\mu, \theta) = \int_0^\infty \frac{e^{-\mu z}(\mu z)^y}{y!} \times \frac{\theta^\theta}{\Gamma(\theta)} z^{\theta-1} e^{-\theta z}\, dz$$

$$= \frac{\theta^\theta}{y! \times \Gamma(\theta)} \times \mu^y \times \int_0^\infty e^{-(\mu+\theta)z} \times z^{y+\theta-1}\, dz$$

$$= \frac{\theta^\theta}{y! \times \Gamma(\theta)} \times \mu^y \times \frac{1}{\mu+\theta}^{y+\theta} \times \int_0^\infty e^{-(\mu+\theta)z} \times [(\mu+\theta)z]^{y+\theta-1}\, d(\mu+\theta)z$$

$$= \frac{\Gamma(\theta+y)}{y! \times \Gamma(\theta)} \times \frac{\theta}{\theta+\mu}^{\theta} \times \frac{\mu}{\mu+\theta}^{y}$$

The mean of the negative binomial distribution is $E[Y|\mu, \theta] = \mu$, and the variance is $Var[Y|\mu, \theta] = \mu(1 + \frac{\mu}{\theta})$. We see that the variance is always larger than the mean and allows for overdispersion.

Negative binomial regression requires assumptions similar to that of the Poisson GLMM:

1. The response, Y, is a count per unit of time or space.
2. The observations must be independent of each other
3. The log of the average incidence rate, $\log(\mu)$, must be a linear function of X.
4. Conditional on the random effects the response follows a Poisson distribution.
5. The random effects come from a gamma distribution on the multiplicative scale.

Like the Poisson GLMM, negative binomial regression also explicitly accounts for random effects from personal differences. Both models offer greater flexibility than the standard Poisson model, but negative binomial model is computationally more efficient.

1.5 Cox Proportional Hazards model

Another way to analyze our data is to consider the time to first PEx as the endpoint. The Cox proportional hazards model is the most common approach to assess a treatment effect with time-to-event outcome between two or more groups (Cox P-H model)(Ozga, Kieser, and Rauch 2018). In clinical trials where time to first event is the primary endpoint, the Cox P-H model is often the first model to consider.

Let T denote a random variable representing event time, and $\lambda(t)$ denote the so-called "hazard". Recall that the "hazard" at time t is $\lambda(t) = lim_{\Delta t->0} \frac{Pr(t \leq T < t+\Delta t | t \leq T)}{\Delta t}$, where Δt stands for a short time period. In a Cox P-H model, the hazard for an individual i is modeled as:

$$\lambda_i(t) = \lambda_0(t) exp(X_i \beta)$$

$$i = 1, ..., n$$

Here $\lambda_0(t)$ is a common baseline hazard and $\lambda_i(t)$ is the hazard for individual i to experience an event at time t. In addition, X_i is a vector of covariates for the ith individual.

To use the Cox P-H model for inference, the following assumptions are required:

1. Proportional hazards assumption: The hazard ratio for any two individuals is assumed to be constant over time. For example, the hazard ratio for patient A with $X_1 = 1$ compared to patient B with $X_1 = 0$ would be $\exp(\beta_1)$ at every time.
2. The log hazard ratio (between any pair of patients) must be a linear function of their covariates, X.
3. The patients must be independent of each other.

The Cox model is quite common in CF studies that use PE as an endpoint. Unfortunately, the Cox P-H model potentially makes inefficient use of data in these studies. CF patients

usually have multiple events, but the Cox P-H model ignores all events after the first. We explore extensions to the Cox P-H model that more efficiently incorporate recurrent events.

1.6 Andersen Gill model

The first time-to-event model that includes recurrent events is the Andersen Gill model (A-G model). The A-G model is a common model for recurrent events and is an extension of the Cox P-H model (Ozga, Kieser, and Rauch 2018). Let k_i denote the number of observed events for individual i. The hazard function of the ith individual for the jth event at time t is written as:

$$\lambda_{ij}(t) = I_{ij}(t)\lambda_0(t)exp(X_{ij}\beta)$$

$$i = 1, ..., n,\ \ j = 1, ..., k_i$$

As in the Cox P-H model, $\lambda_0(t)$ is a common baseline hazard, and X_{ij} is a vector of covariates at time t. $I_{ij}(t)$ is a risk indicator, taking value 1 if the ith individual is under observation (at risk) at time t, and value 0 if not. This is different from in the Cox PH model when a person experienced an event, that person will be no longer at risk ($I_{ij}(t) = 0$). However, a person will remain at risk ($I_{ij}(t) = 1$) through the study unless the person is no longer under observation (death or censoring) (Castañeda and Gerritse 2010).

The A-G model requires similar assumptions to those required by the Cox P-H model:

1. Proportional hazards assumption.
2. The log hazard ratio (between any pair of patients) must be a linear function of their covariates, X.
3. The patients must be independent of each other.
4. The hazard of experiencing an event at time t since study entry does not change based on the number of previous events a patient has experienced.

The 4th assumption is a strong assumption which implies that the recurrent events need to be independent of each other. If this strong assumption is not satisfied, the A-G model may give invalid inference. Observations through the counting process are usually assumed to be independent of each other. However, in cases where an exacerbation's occurrence may relate to the individual's exacerbation history, this required independence assumption may be violated and other methods should be used.

1.7 Prentice-Williams-Peterson models

One might alternatively consider the class of Prentice-Williams-Peterson methods (PWP models). There are two models in this class, the PWP total time model (PWP-TT model) and the PWP gap time model (PWP-GT model). These methods are stratified Cox-based approaches. The PWP-TT model uses, as event times, the time since study entry (total time scale) while the PWP-GT model uses the time since the previous event (gap time scale).

In the PWP-TT model the hazard for an individual i for the jth event as:

$$\lambda_{ij}(t) = \lambda_{0j}(t)exp(X_{ij}\beta)$$

$$i = 1, ..., n,\ j = 1, ..., k_i$$

In the PWP-GT model the hazard is modeled as:

$$\lambda_{ij}(t) = \lambda_{0j}(t - t_{j-1})exp(X_{ij}\beta)$$

$$i = 1, ..., n,\ j = 1, ..., k_i$$

In both cases, it can be seen that the underlying model is similar to the common Cox model but is stratified by number of events. Here, k_i indicates the numbers of events that occur for person i (Prentice, Williams, and Peterson 1981). For each recurrent event $j = 1, \ldots, k_i$, a

separate hazard function is modeled with its own baseline hazard λ_{0j} but shared regression parameter β. Thus, the hazard for a recurrent event can be influenced by previous events. An individual is at risk for the jth event only if the individual experienced a previous (j-1) st event.

Since PWP models are also based on proportional hazards, PWP models require similar assumptions to the standard Cox PH model:

1. Proportional hazards assumption.
2. The log hazard ratio (between any pair of patients) must be a linear function of their covariates, X.
3. The patients must be independent of each other.

PWP models do not rely on the strong assumption of independence of events (which the A-G model requires) since they stratify the data by event to allow baseline hazards to vary according to event number. Thus, in trials where a person's previous experiences of PEx may result in different risk for another PE, PWP is likely a more appropriate method.

However, the PWP models suffer from an issue involving small risk sets. This issue occurs if, for some large number of events (stratum), there are only a few patients remaining in the risk set. In this case, the hazard function being modeled in these strata will be less reliably estimated, and this may lead to inferential problems. In these cases, we usually need to drop event numbers that are too large. We assume that in most of these cases, since patients with many events are rare, dropping these events should not substantively affect our conclusions. However, since we are not making use of all of our data, this still becomes a limitation of PWP models.

2 The Impact of Statistical Methods on Simulated Data

To better understand how the seven models perform in different scenarios, we ran simulated experiments. We considered different scenarios: with well-behaved Poisson outcome, and overdispersed outcome with log-normal or gamma distributed multiplicative random effects. We also studied how missingness affects the analysis results.

2.1 Data simulation methods

We used three covariates in our simulation study. The first variable is age group. To generate the age group variable, we first generated the variable age from a uniform distribution from 1 to 18 years old. Any patient with age <= 3 is categorized as in age group 1, 3 < age <= 6 is age group 2, 6< age <= 12 is age group 3, and finally any patient with age > 12 is age group 4. The second variable, sex, is generated as a binary variable from a Bernoulli distribution with p=0.5. Finally, the variable treatment is generated from a Bernoulli distribution with p=0.5. As we were considering a randomized clinical trial, treatment was generated independently of covariates.

We simulated PE events from a Poisson process with a separate per day rate for each individual. We calculated this per day rate using the three covariates we simulated. The rate for the ith patient's Poisson process is calculated as $rate_i = r \times exp(\beta_1 \times age\ group_i + \beta_2 \times sex_i + \beta_3 \times trt_i) \times \epsilon_i$, where (ordinal) age group is treated as a continuous variable and r is the baseline hazard rate (we use r=0.002 through our data simulations). The ϵ_i are [potential] random effects: They are only included in specific overdispersion scenarios. Coefficient values of $\beta_1 = 0.2$ and $\beta_2 = 0.4$ were used in our simulations. To assess the Type I error rates and power to detect an effect of treatment, we simulated with β_3 ranging from 0 to 0.8, with an increment = 0.1. For each β_3, we ran 5000 simulations. For each scenario we will be talking about below, we simulated data for five different sample sizes: $n = 100, 200, 300, 400$, and 500 (these reflect sample size observed in contemporary CF trials). All power calculations

are based on the nominal significance level of $\alpha = 0.05$.

We considered three censoring scenarios meant to resemble cases we might observe in actual data. The first is an administrative censoring only scenario (which we will refer to as the "non-missingness scenario"). We simulated each patient with the same total observed time of 180 days. For the second case we include censoring completely at random (MCAR). For this MCAR scenario, a random censoring time (C) was generated for each patient. The patient was included in the trial for $min(C, 180)$: Any patients for which this was less than 180 were considered censored. C was generated from the same exponential distribution for each patient with rate=0.001. This rate was selected to obtain about 20% [non-administratively] censored observations. For the third scenario we considered censoring at random (MAR). For this MAR scenario, we generated C for a given patient using a Poisson process with a rate specific to that patient: In particular we used the rate calculated for that patient's PE-event process divided by 6 (this was selected to give about 20% [non-administrative] censoring). In this way, the missingness status of a patient will depend on his or her covariates but not otherwise on the outcome. As before, each patient was observed on the trial until time $min(C, 180)$. In these scenarios, for our Poisson-like models an offset was used: $log(180)$ for administrative-only censoring, and $log(min(C, 180))$ for the other scenarios.

In these simulated experiments with homogeneous treatment effect, we expect Cox-PH to have lower power because it only takes into account the first event while other models also include following events.

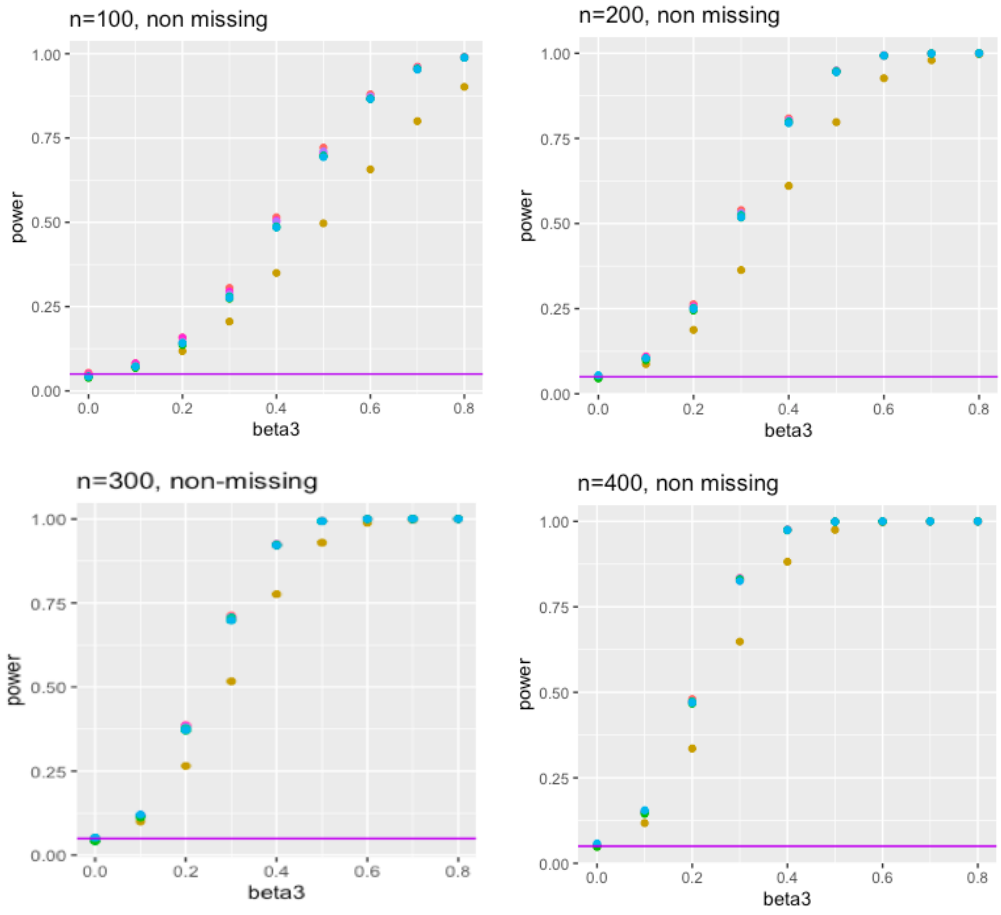
2.2 Data simulation results

2.2.1 No overdispersion

We first ran simple simulations to see how these models perform on data without overdispersion. The rate the for ith patient of the Poisson process was calculated as $rate_i = r \times exp(\beta_1 \times age\ group_i + \beta_2 \times sex_i + \beta_3 \times trt_i)$.

The power plots we obtained for the three missing scenarios are as follows:

Non-missing



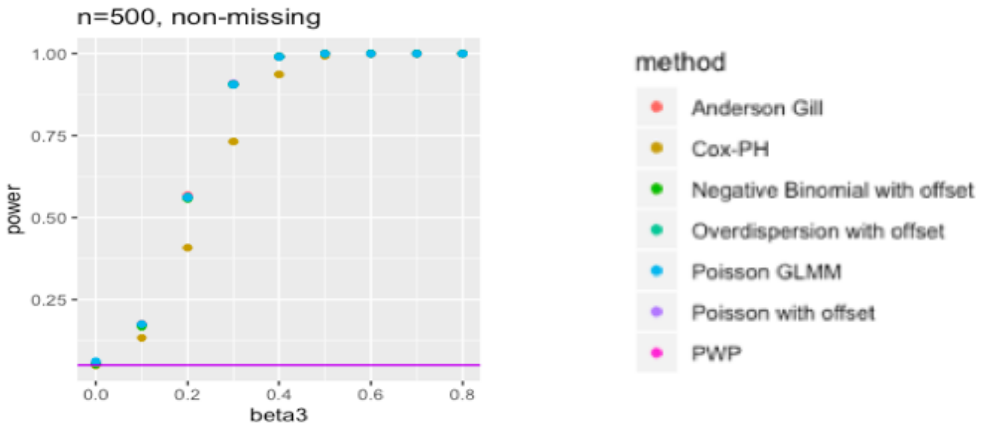
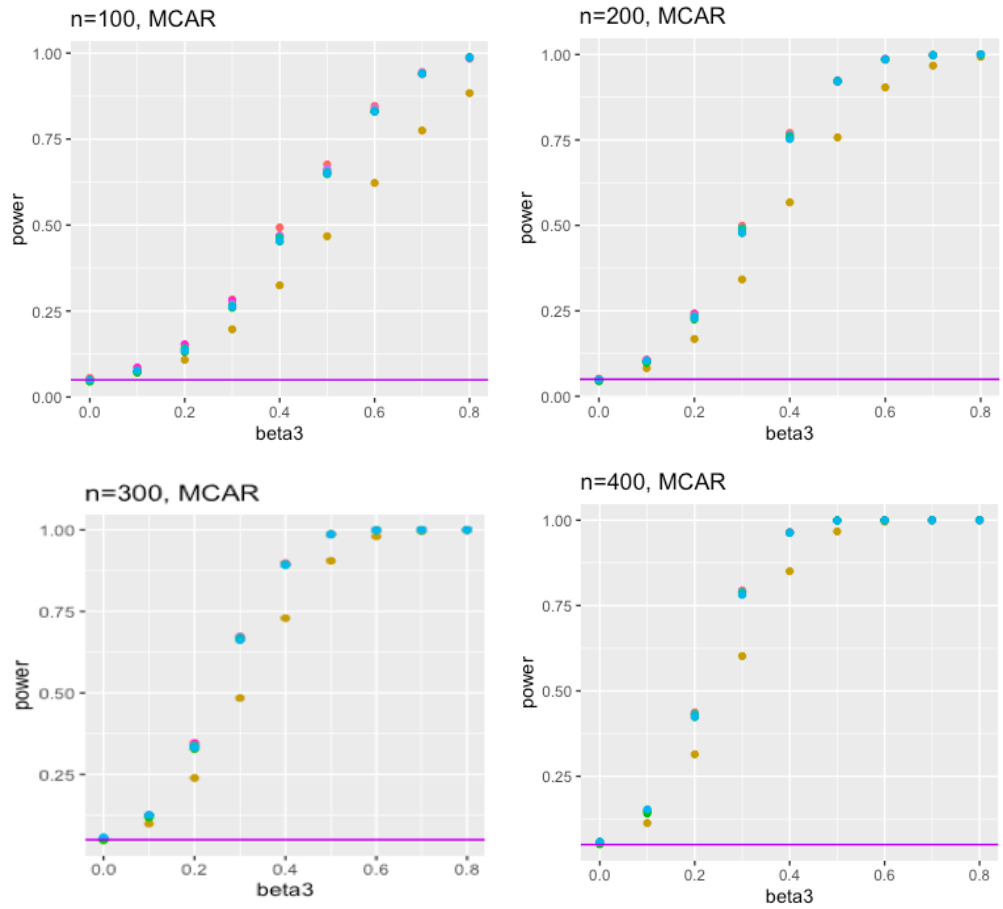


Figure 2.1.1: Power plots, no overdispersion, non-missing

Missing completely at random



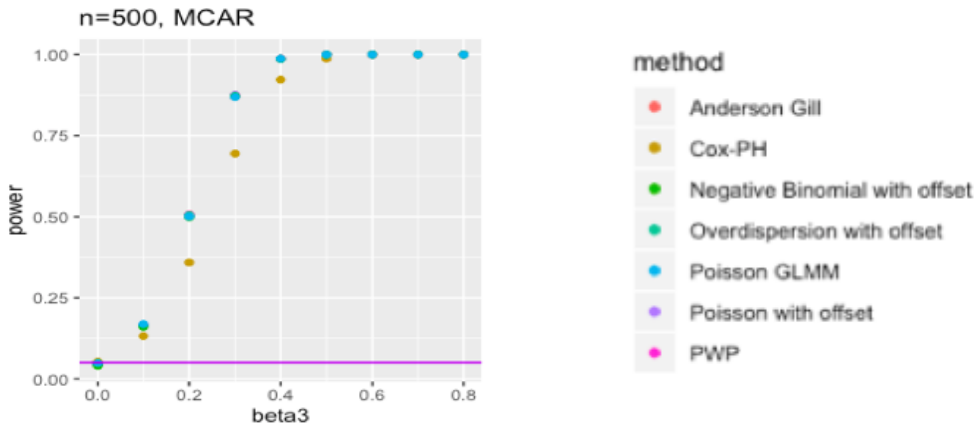
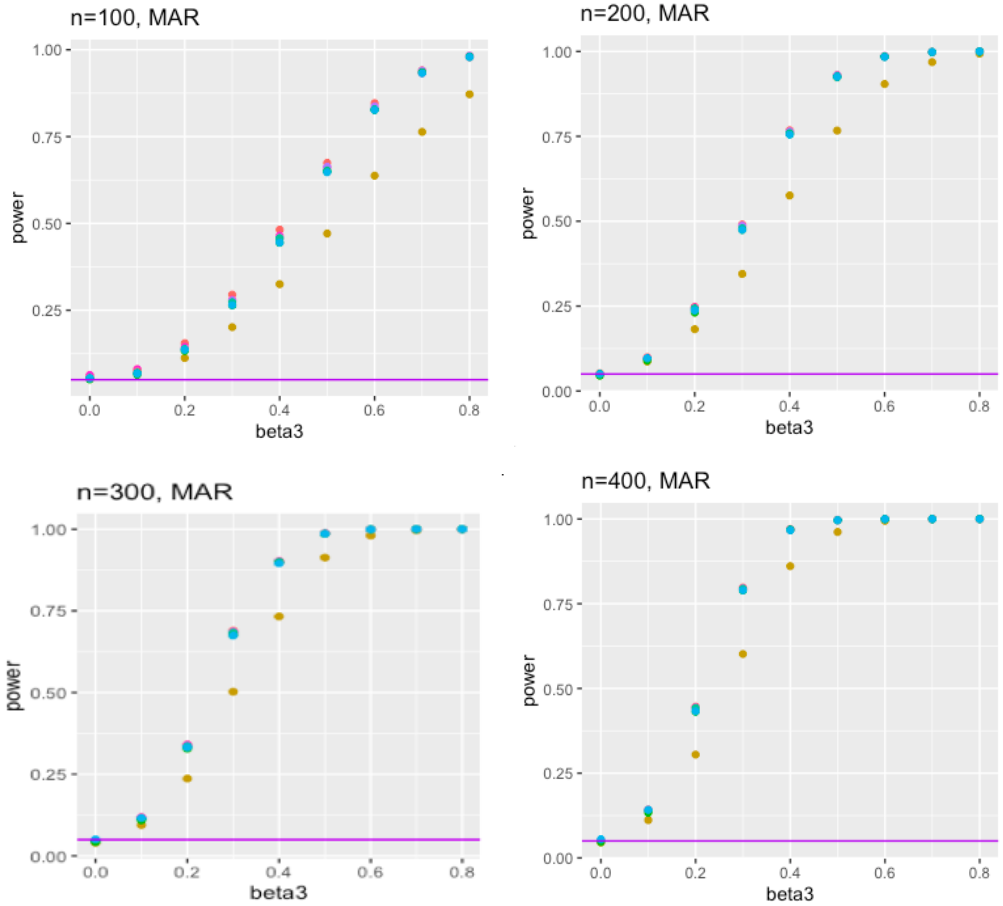


Figure 2.1.2: Power plots, no overdispersion, MCAR

Missing at random



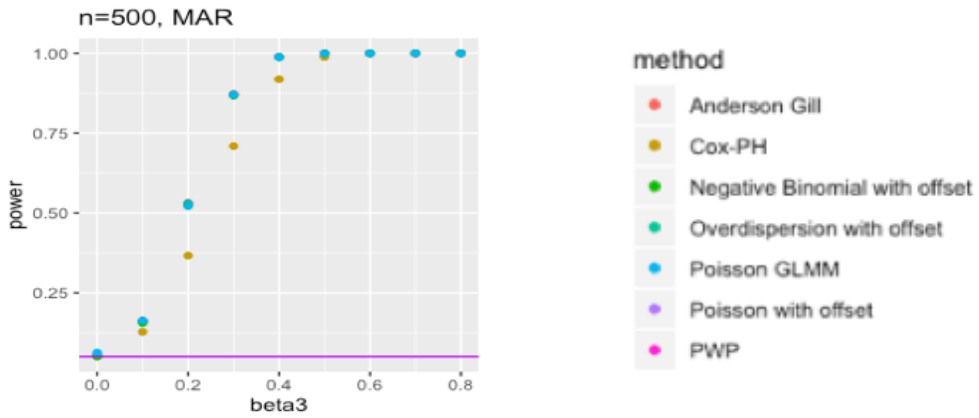


Figure 2.1.3: Power plots, no overdispersion , MAR

Although the more complex censoring scenarios both seem to have lower power than that of the administrative-only censoring, the differences are small. Thus, these models appear to work well for small to moderate amounts of censoring.

From all of these plots, we see that Type I error of each of the seven models is close to the nominal level (around 0.05) and power increases as β_3 increases. Among all examples, Cox-PH model is somewhat less powerful than all other models. For the non-missing simulation where the sample size is 200, and $\beta_3 = 0.6$, for example, power of the Cox-PH model is about 0.63 while it is at least 0.75 for other models (Figure 2.1.1). Other models perform very similarly to each other. The Cox-PH model's poor performance is likely because it ignores information from events other than the first. Next, we check how these models perform on overdispersed data.

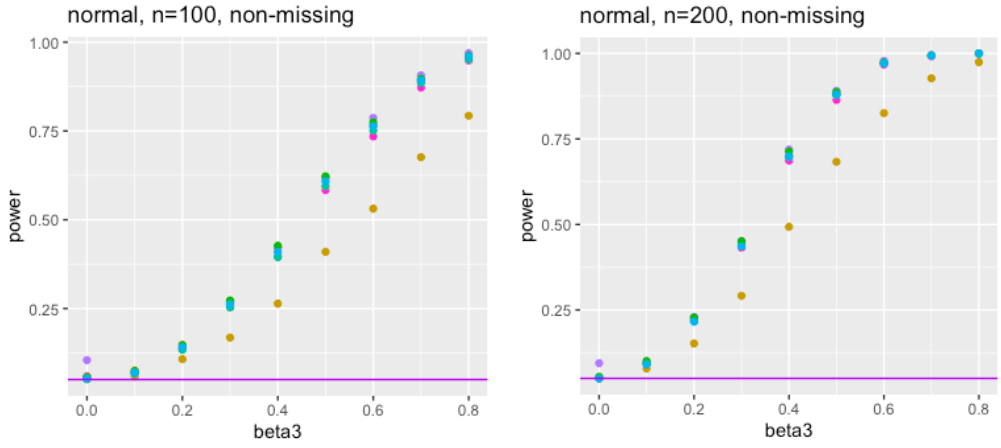
2.2.2 Overdispersion: log-normal distributed random effects

In order to compare model performance in overdispersed data, we now modify our simulations to include a multiplicative frailty term for each patient's rate when generating exacerbation events. We first considered the case where the frailty terms are generated from a log-normal distribution. The rate for the ith patient is calculated as $rate_i = r \times exp(\beta_1 \times age\ group_i +$

$\beta_2 \times sex_i + \beta_3 \times trt_i + \epsilon_i)$ where $\epsilon \sim N(0, 0.6)$. As with the non-overdispersed data simulation process, we evaluated power for each of seven models in the three censoring scenarios.

We found that there is an inflation in the Type I error-rate for the regular Poisson regression in all censoring scenarios when we have over-dispersed data with log-normal random effects. In addition to making Poisson regression generally an unreliable tool here, this failure to control type-I error gives Poisson regression an unfair advantage when evaluating its power. To level the playing field, we adjusted the nominal significance level for Poisson regression (making it more stringent), to make sure that when calculating power it was using a cutoff that actually controlled type I error at the 0.05 level. To do this, in each scenario, we ran our simulation with $\beta_3 = 0$ and found the nominal significance cutoff such that actually 5% of the p-values calculated in our simulations are below it. We then used that number as the "adjusted significance cutoff" for power calculations. The power plots we obtained with this adjustment are as follows:

Non-missing



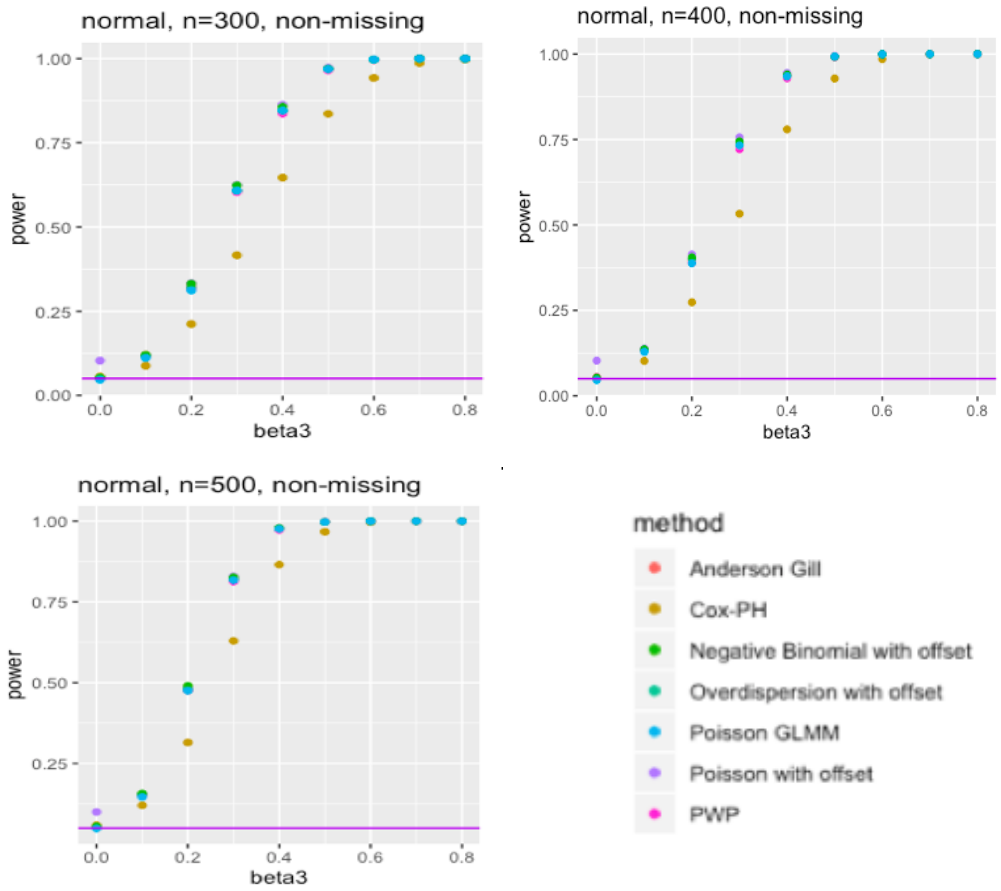
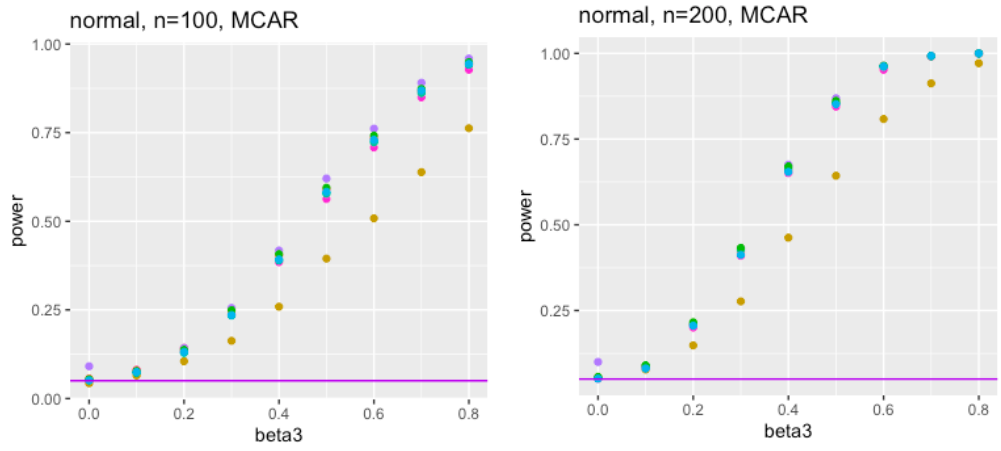


Figure 2.2.2: Power plots, overdispersed with log-normal random effects, non-missing

Missing completely at random



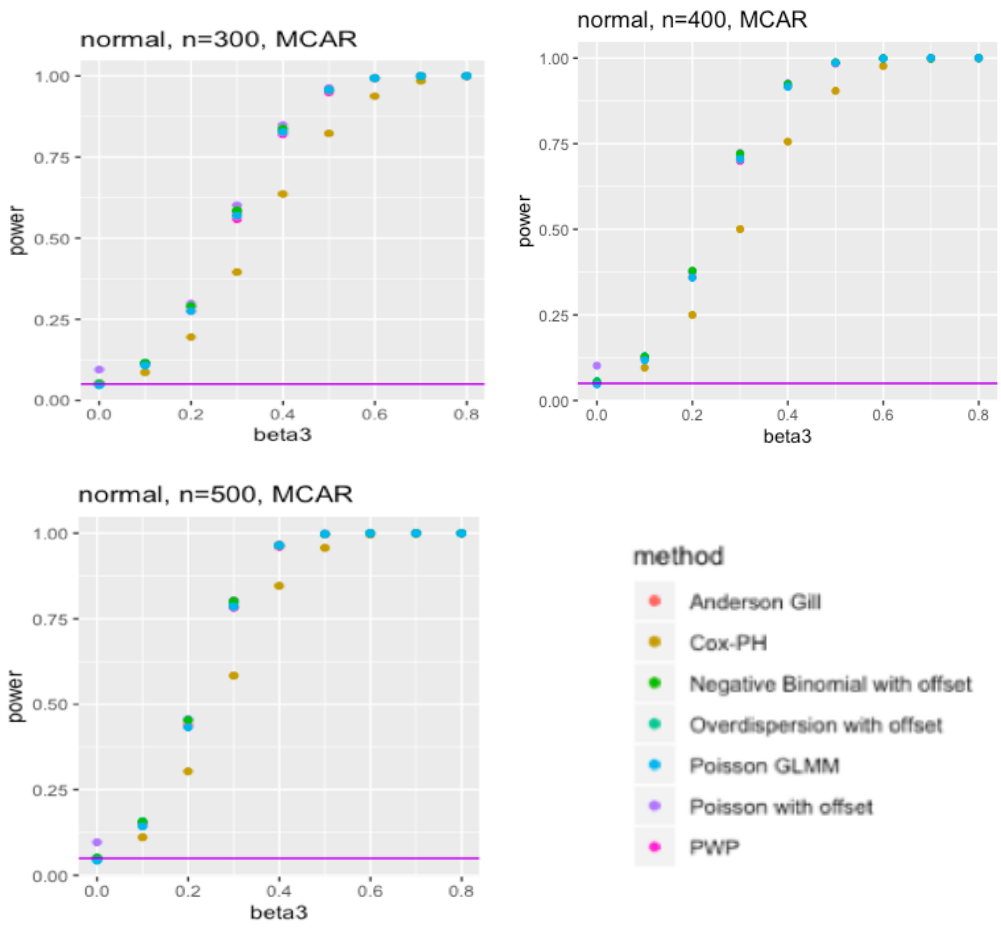
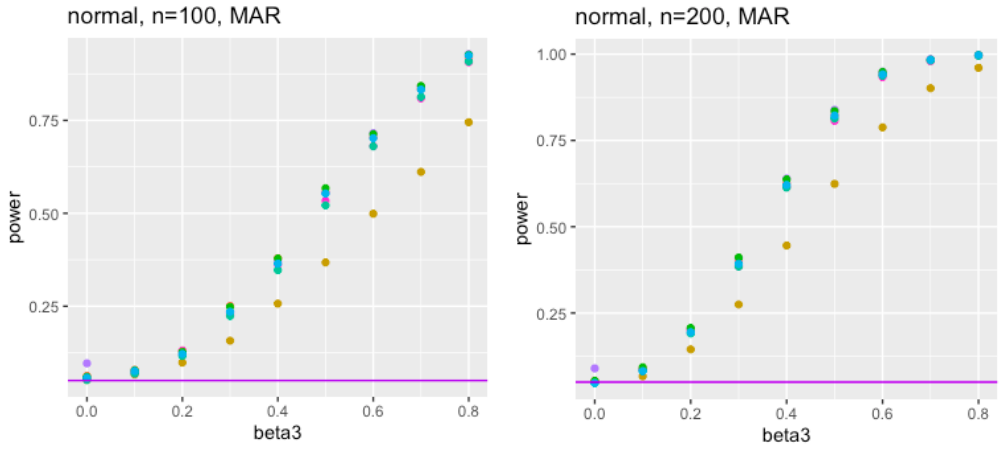


Figure 2.2.2: Power plots, overdispersed with log-normal random effects, MCAR

Missing at random



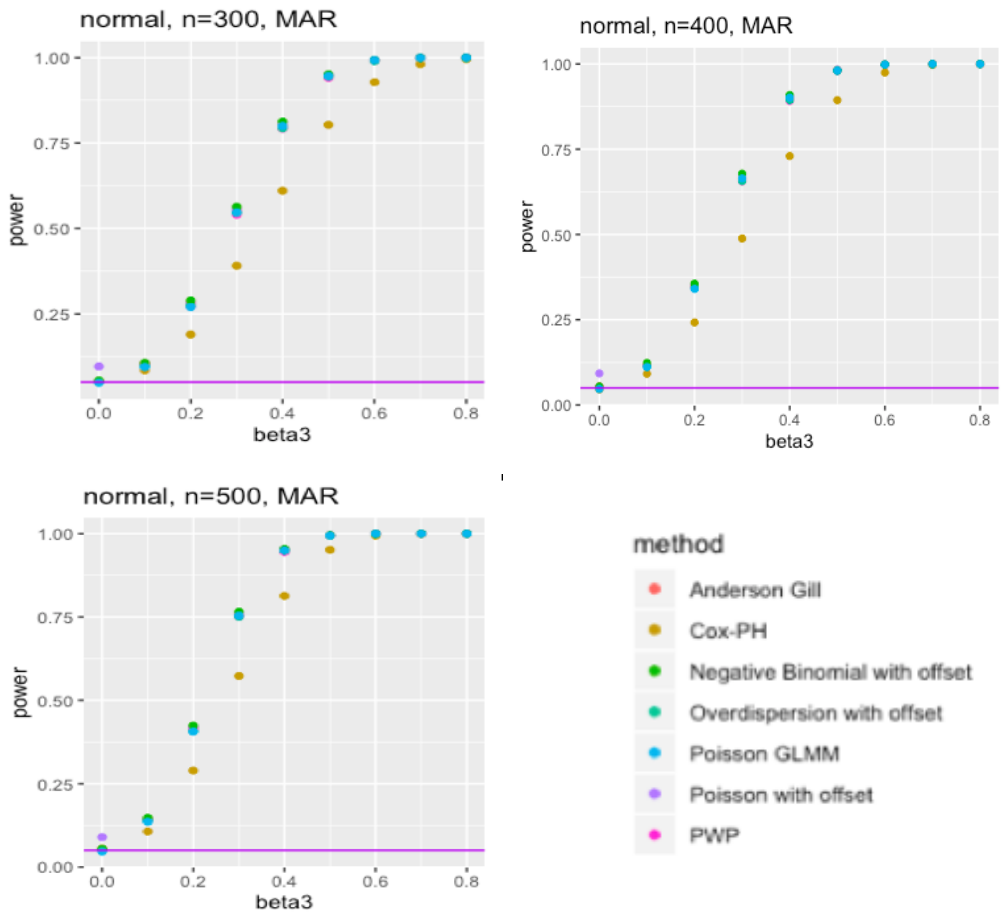


Figure 2.2.3: Power plots, overdispersed with log-normal random effects, MAR

As shown in Figure 2.2.1-2.2.3, we see that after accounting for inflated type I error, Poisson regression has no better power than any other method (and would be unreliable to use in practice). As before we see that Cox-PH performs poorly with respect to power. Other models work similarly to each other, and do not have issues with inflating Type I error-rates. This is somewhat interesting as the negative binomial model is slightly misspecified (it has the wrong form for the random effects), while the log-normal GLMM model is precisely correctly specified. In addition, as we saw with non-overdispersed data, there is a slight decrease in power when we include additional censoring. However, this difference is small, and all models appear to accommodate missing data well.

Now we consider a similar scenario with overdispersion where we use gamma distributed

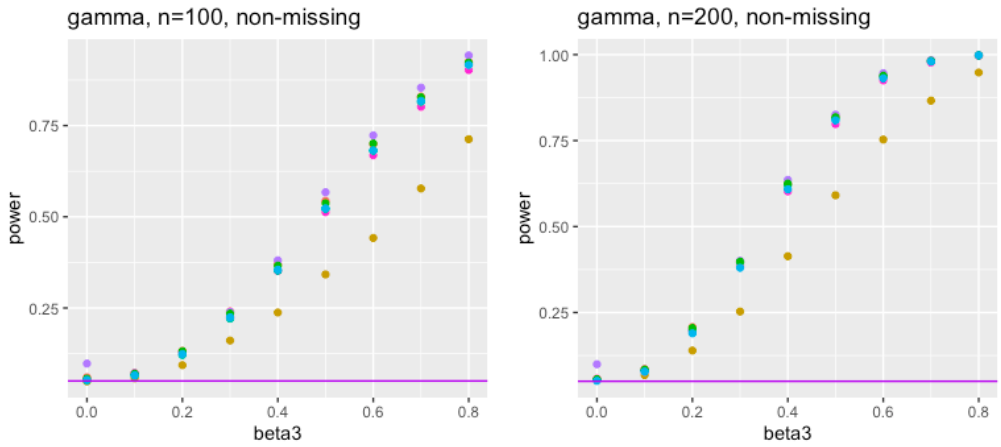
multiplicative random effects rather than log-normal distributed. We note that, in the following scenario, the negative-binomial model is precisely correctly specified.

2.2.3 Overdispersion: Gamma distributed random effects

Now instead of multiplicative log-normal random effects, we consider another scenario where the multiplicative random effects follow a gamma distribution. The rate for ith patient is calculated as $rate_i = r \times exp(\beta_1 \times age\ group_i + \beta_2 \times sex_i + \beta_3 \times trt_i) \times \epsilon_i$ where $\epsilon \sim$ Gamma$(2, 2)$ (shape=2 and rate=2).

We evaluated power and type I error rate for the seven models in the three missing scenarios. As with overdispersed data with log-normal distributed multiplicative random effects, we found that regular Poisson regression has inflated Type I error in all scenarios. As in the previous section (Section 2.2.2), we recalibrated our cutoff for Poisson regression to ensure a fair comparison. We obtained the following results:

Non-missing



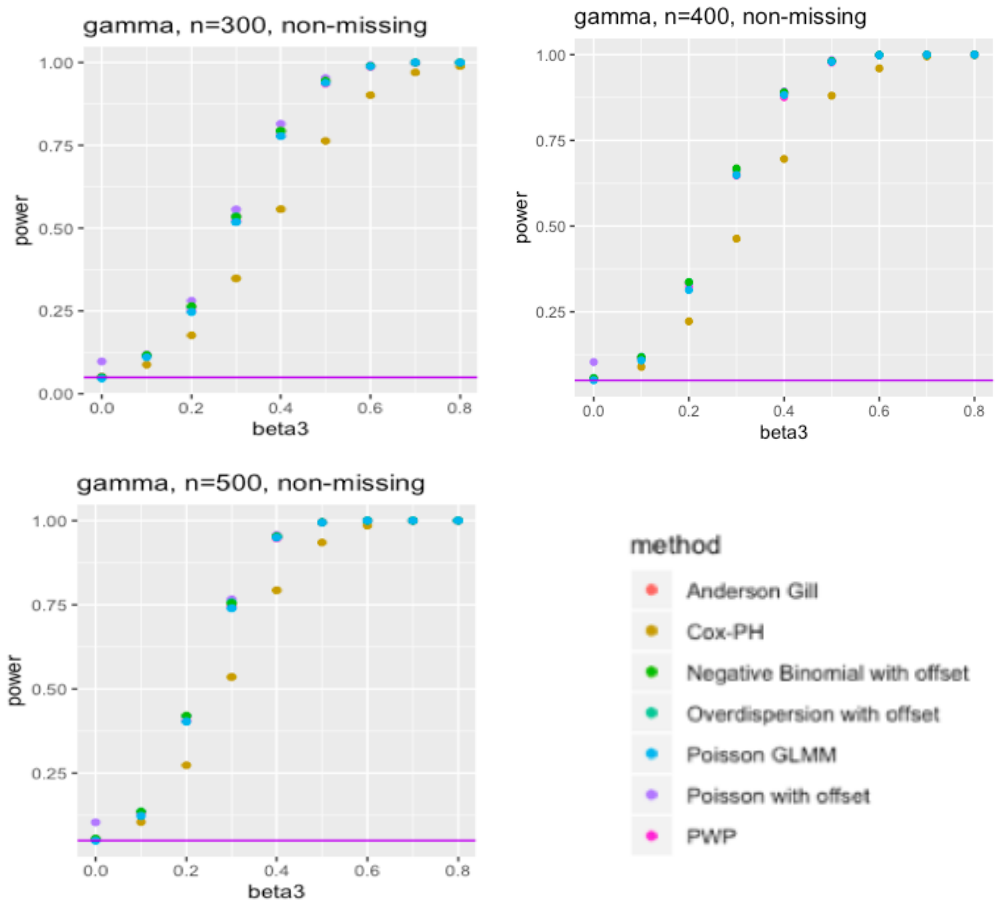
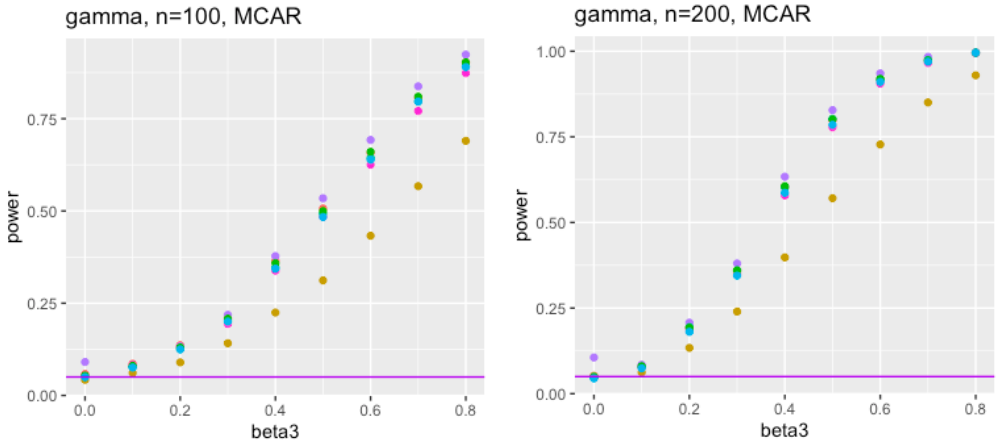


Figure 2.3.1: Power plots, overdispersed with gamma random effects, non-missing

Missing completely at random



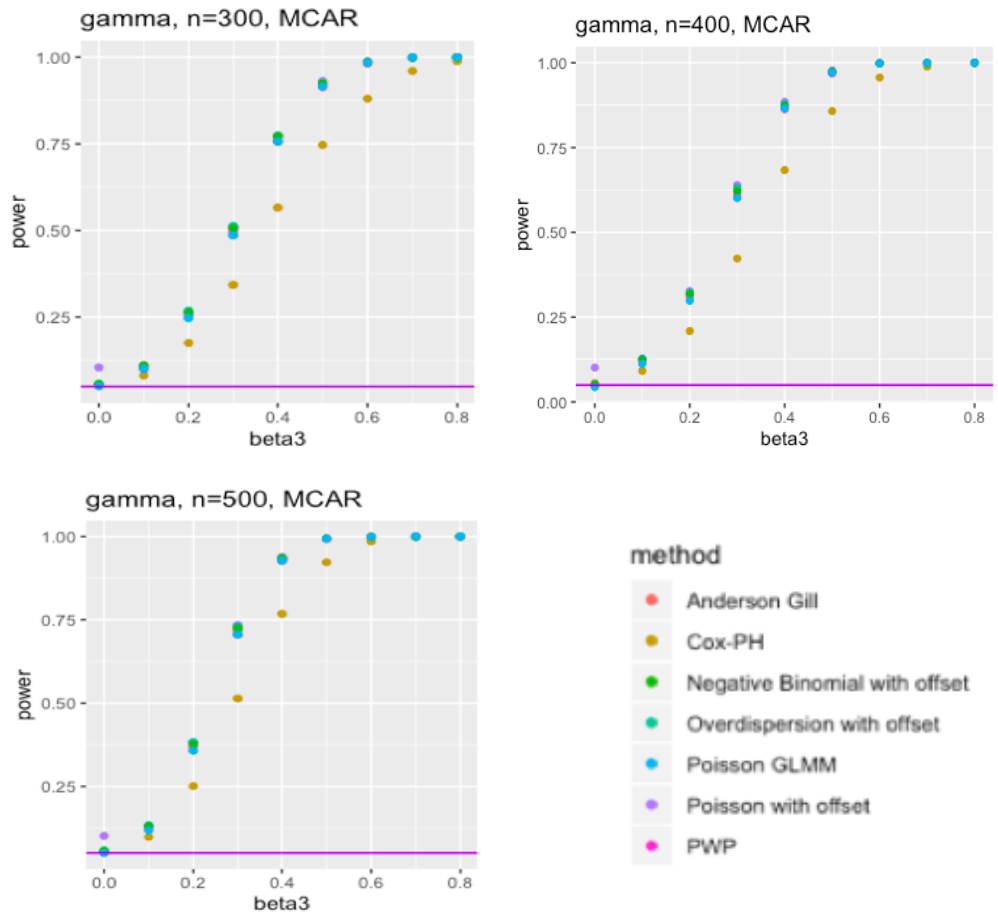
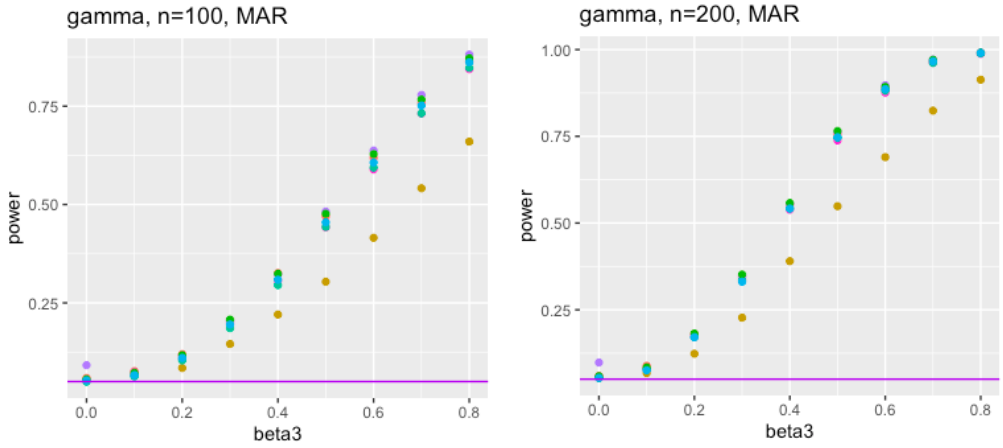


Figure 2.3.2: Power plots, overdispersed with gamma random effects, MCAR

Missing at random



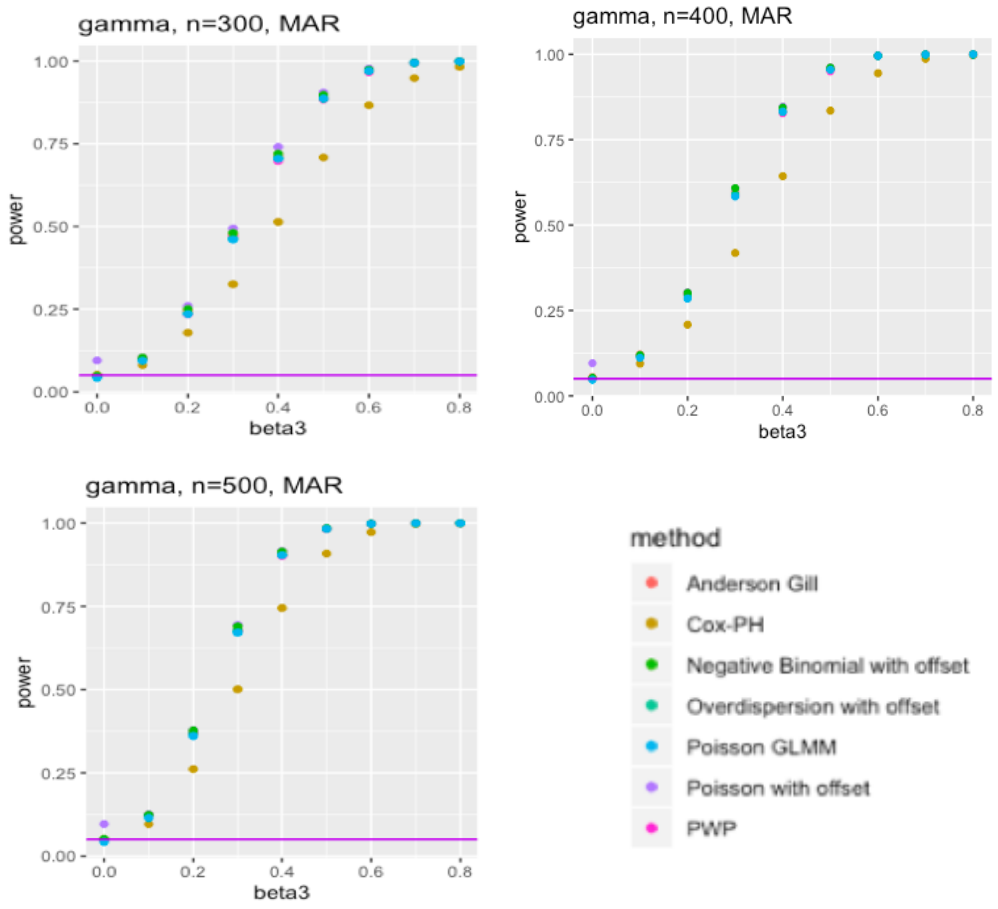


Figure 2.3.2: Power plots, overdispersed with gamma random effects, MAR

Findings from overdispersed data with gamma distributed random effects are very similar to what we saw with the log-normal distributed random effects. We see an inflated Type I error-rate from the regular Poisson regression again. With the recalibration, simple Poisson regression has no better power than other methods that actually account for overdispersion. The Cox-PH model still performs the worst among all models. In this scenario, it is interesting that all Poisson-regression-based overdispersion correcting methods perform similarly: The *correctly specified* negative binomial model performs no better than the others.

Similar to what we observed for both overdispersed data with log-normal random effects and non-overdispersed data, power in the scenarios with censoring is lower than that in the non-missing case. However the differences are again not large.

2.3 Discussion of data simulations

In general, we found that when there is [non-administrative] censoring (MCAR or MAR), the power of all methods was decreased as compared to administrative-censoring only. However, the differences are usually small in the relatively mild scenarios we considered.

For over-dispersed data, for both types of random effects we used, Poisson regression has substantially increased Type I error-rates. This inflated Type I error-rate means that Poisson regression is likely unreliable for conducting inference. All 3 of our Poisson-regression-based overdispersion correction methods worked well in both random-effects scenarios (with regard to controlling Type I error, and power).

We also found that Cox-PH regression for time-to-first-event performed comparatively poorly: While type 1 error rates were controlled, it had substantially less power in all scenarios. However, it is worth remembering that in these scenarios we simulated a homogeneous effect of treatment on all exacerbations. In examples where this is heterogeneous, it may be that engaging with time-to-first event is a stronger competitor. In particular, this is the scenario that we see in the next chapter when we reanalyze two CF studies.

3 The Impact of Statistical Methods on Clinical Trial Results: Re-analysis of Two CF Trials

In this chapter, we will apply the 7 models we reviewed and studied in the first two chapters to two Cystic Fibrosis studies. The two datasets we will be using are from the AZ trial, a randomized controlled trial studying effect of Azithromycin on pulmonary function in patients with CF uninfected with Pseudomonas aeruginosa; and the OPTIMIZE trial, a randomized trial studying Azithromycin for early Pseudomonas infection in CF.

3.1 AZ trial

3.1.1 Background

The AZ trial was a multicenter, randomized, double-blind placebo-controlled trial conducted from February 2007 to July 2009 at 40 CF care centers in the United States and Canada (Saiman et al. 2010). The AZ trial was conducted to study if azithromycin treatment improves lung function and reduces pulmonary exacerbations in pediatric CF patients uninfected with Pseudomonas aeruginosa. In the original study, the primary endpoint was change in forced expiratory volume (FEV), while numbers of PEx and time until PEx were among exploratory endpoints. Patients were randomized 1:1 stratified on age group and CF center combinations (Saiman et al. 2010).

The following table gives characteristics of patients enrolled on the AZ trial:

Table 3.1.1. Characteristics of Participants According to Treatment Group: AZ Trial

	Azithromycin (n=131)	Placebo (n=129)
Follow-up time [mean (sd)] (days)	166.31 (25.54)	168.56 (24.56)
Age at randomization [mean (sd)] (years)	10.7 (3.25)	10.6 (3.10)
Age Group [no(%)]		
6-12 y	91 (69%)	91 (71%)
>12-18 y	40 (31%)	38 (29%)
Sex [no(%)]		
Female	54 (41%)	59 (46%)
Male	77 (59%)	70 (54%)
PEx numbers [no(%)]		
0	103 (79%)	79 (61%)
1	21 (16%)	38 (30%)
2	4 (3%)	9 (7%)
3	3 (2%)	3 (2%)
≥1	28 (21%)	50 (39%)

From Table 3.1.1, we see that there are slightly more male than female patients, and substantially more patients in the 6-12 years old age group (age group 1). Most patients did not have any exacerbations. However, the percentage of patients who had at least one event appears higher among those who received placebo than those who received the treatment. In our statistical analysis, we fit all models assessing the relationship between treatment and our PEx endpoint adjusting for both age group and sex as this is how the covariates were adjusted for in the original trial analysis.

3.1.2 Statistical analysis

We first analyzed the data from this trial using a simple Poisson model: It indicates that treatment has a significant effect on rate of exacerbation. The exponentiated coefficient of treatment is 0.6 (95% CI 0.40, 0.88) (Table 3.1.2). This indicates that, among patients of the same sex, who are in the same age group, those who recieve azithromycin are expected to experience roughly 40% fewer exacerbations than those who receive placebo.

Table 3.1.2.Poisson Family Methods and Results: AZ Trial

Method	Endpoint	P-value*	Goodness of fit results	Effect size: exp $(\beta)^*$ (95% CI)
Poisson Regression (regular)	PEx number	$0.01 < 0.05$	p=0.1268(dispersion)	0.6 (0.40,0.88)
Poisson GLMM	PEx number	$0.019 < 0.05$		0.59 (0.37,0.91)
Quasi-Poisson Regression	PEx number	$0.017 < 0.05$		0.6 (0.39,0.91)
Negative Binomial Regression	PEx number	$0.017 < 0.05$	p=0.96(deviance)	0.6 (0.39,0.91)

*for treatment

Next, we ran an overdispersion test to identify if a simple Poisson model can fit the data well. The null hypothesis being tested is that the true dispersion parameter is equal to 1 (indicating that there is no overdispersion). The alternative hypothesis is that the true dispersion parameter is greater than 1, indicating that there is overdispersion. We ran the test using the function dispersiontest from R package AER (Kleiber and Zeileis 2020). This function implemented a regression-based test for overdispersion (Cameron and Trivedi 1990), where the null hypothesis being tested is that H_0: $\text{Var}(y_i)=\mu_i$ and the alternative hypothesis is H_α: $\text{Var}(y_i)=\mu_i + \phi \times g(\mu_i)$ ($\text{g}(\mu_i)$ is a specified function of μ_i). In particular, a t-test

is used to evaluate if $\phi = 0$. The p-value given was 0.1268 and the dispersion estimate is 1.11, suggesting that we do not have significant evidence to reject the null hypothesis (Table 3.1.2). Thus, it is likely that there is minimal overdispersion, and the simple Poisson regression model should perform well on the AZ data.

We compared the results of simple Poisson regression to those from our models with overdispersion correction: Poisson GLMM, Quasi-Poisson and negative binomial regression. Since our data are not substantially over-dispersed, we expect all of these three methods to perform similarly to a simple Poisson regression: We expect our mixed effect models will estimate multiplicative random effects near 1, and the Quasi-Poisson model will calculate a dispersion parameter near 1. Table 3.1.2 shows us the results from these three methods: Coefficient estimates are similar (to each other and the simple Poisson model). To check the goodness of fit for the negative binomial, we used the deviance goodness of fit test, where the null hypothesis is that this model is a good fit for our data, and the alternative hypothesis is that this model is not a good fit for our data. Results show that negative binomial model also works fine (p-value=0.96, Table 3.1.2). For the Poisson GLMM, we plot out the random effects extracted from the Poisson GLMM model (Figure 3.1.1). We see that most patients have multiplicative random effect terms close to 1. These results match our expectations.

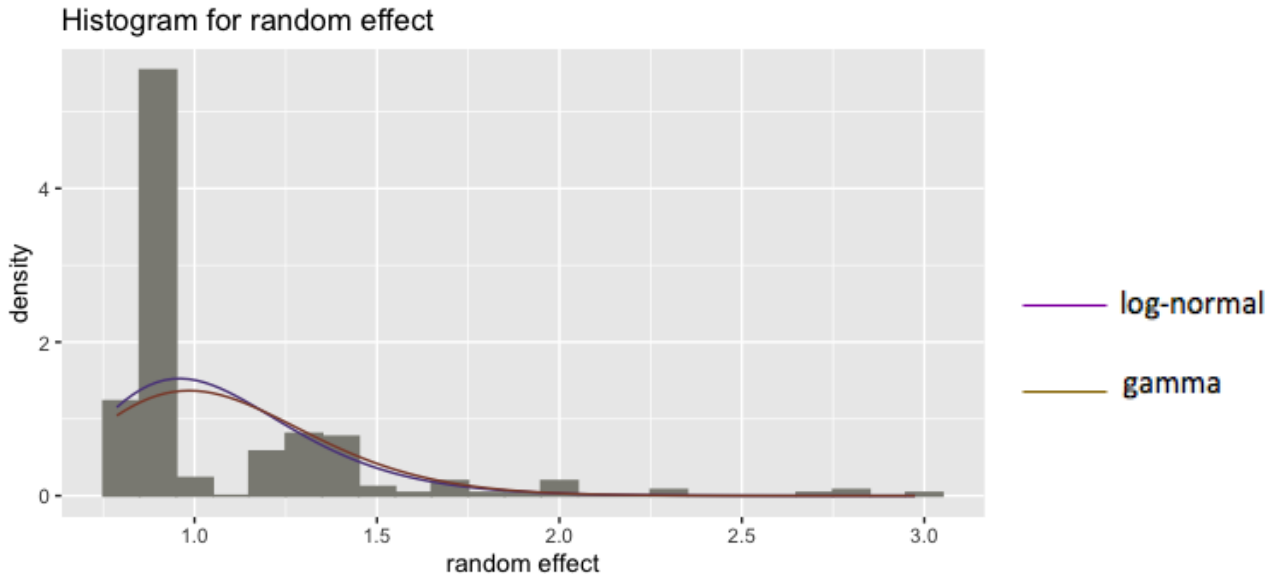


Figure 3.1.1 Random effects from Poisson GLMM (AZ) and its distribution

In this case, with nearly no overdispersion, all 4 models gave very similar results.

We now move from considering exacerbation rate to examining hazard using proportional-hazard modeling methods (Cox-PH, AG, and PWP methods). We first examine our assumptions about proportionality of hazards. To do this, in each stratum (Age group and sex combination), we examine the Kaplan-Meier curves for time to first exacerbation in the two treatment groups. The results are given in the following figures:

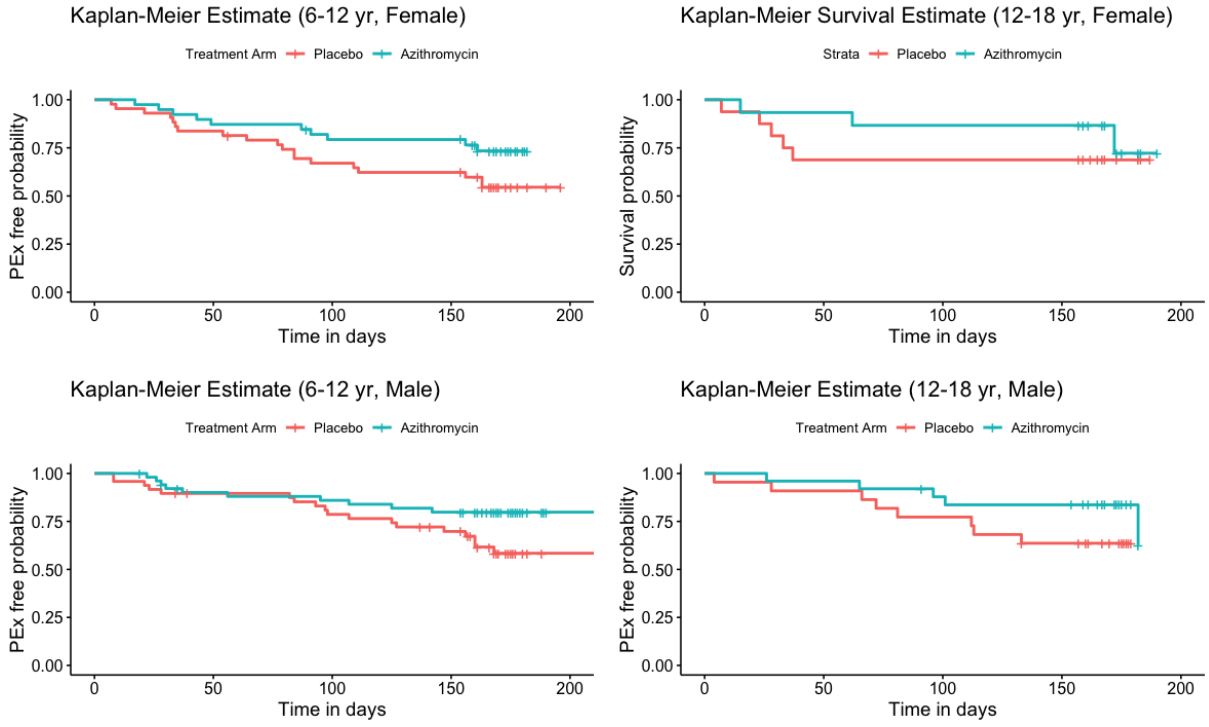


Figure 3.1.2 Kaplan Meier Curves for time to first event, stratified by age group and sex combinations

From Figure 3.1.2 we see that among patients who are male and in age group 1, the Kaplan-Meier curves for the two treatment groups appear to cross at one point. Visually, it is unclear if this crossing is "real" or just a function of the variability of KM curves constructed based on limited data. To quantify evidence of departure from the proportional hazards assumption

we used the cox.zph function from the R package survival (Therneau et al. 2020). The function implements a standard test for goodness of fit for proportional hazards (Harrell 1986). Here the null hypothesis is that proportional hazards is satisfied. We found p-value=1 (Female, 6-12 years); p-value =0.36 (Female, 12-18 years); p-value=0.23 (Male, 6-12 years); p-value=0.55 (Male, 12-18 years). These results indicate that we do not have strong evidence of a departure from proportional hazards (though this may be due to a limited sample size within each stratum).

Results for the Cox-PH model are shown in Table 3.1.3 (denoted as Cox PH 1). Adjusting for sex and stratified on age group, the exponentiated estimated coefficient of treatment (estimated hazard ratio) is 0.5 and the p-value ($p=0.003<0.05$) shows statistical significance. This indicates that a patient who receives azithromycin has roughly 50% less hazard of experiencing their first exacerbation at any time t compared to an identical patient who receives placebo.

Table 3.1.3. Cox PH Family Methods and results: AZ Trial

Model	Endpoint	P-value*	Effect size: HR and 95% CI
Cox PH 1	Time from start to first event	$0.003 < 0.05$	0.50 (0.31,0.79)
Cox PH 2	Time from start to second event	0.62	1.27 (0.79,2.06)
Cox PH gap	Time from first event to second event	0.76	1.16 (0.45,3.01)
Andersen Gill	Time from start until each event	$0.01 < 0.05$	0.57 (0.36,0.90)
PWP-TT	Time from start until each event	$0.03 < 0.05$	0.60 (0.38,0.95)
PWP-GT	Time from the previous until the next event	$0.02 < 0.05$	0.59 (0.38,0.93)

*for treatment

Since some patients had more than one exacerbation, we also wanted to understand the effect of treatment beyond the first exacerbation. First, we checked if azithromycin is also effective in reducing the hazard of having exacerbations (including and beyond the first) using the A-G

model, assuming that the time interval between exacerbations are conditionally uncorrelated. Results from the A-G model show that the treatment is also associated with reduction in hazard of experiencing all exacerbations, with p-value=0.01 $<$ 0.05, and hazard ratio of 0.57 (Table 3.1.3). This suggests that patients receiving azithromycin have, on average, $\sim$ 43% less hazard of experiencing an exacerbation at any time t compared to those who received placebo.

As we noted in the first chapter, the A-G model makes a strong assumption regarding independence of events for a given patient. To identify if our finding is robust, we also fit PWP models — this allows us to use a different baseline hazard for each event-number. We applied both the PWP gap time (PWP-GT) model and the PWP total time (PWP-TT) model. As we discussed earlier, PWP methods may encounter a small-risk-set issue if, for some large event-number, there are few patients remaining in the risk set. In this trial, patients had at most 3 exacerbations, and the numbers of patients who had 3 events was sufficiently large (6 patients, from Table 3.1.1), so we felt it was acceptable to use of the available data with PWP methods. Results of both PWP models are shown in Table 3.1.3. The p-value for the treatment variable is 0.02$<$0.05 and 0.03$<$0.05 for the PWP-TT model and the PWP-GT model, respectively. This also suggests that the treatment is effective for reducing the hazard of experiencing an exacerbation (averaged all events). Results from the PWP-GT model and the PWP-TT models are very similar to each other. Both models suggest that a patient receiving azithromycin has about 40% less hazard of experiencing any exacerbation at any time t compared to an identical patient who received placebo.

From the results of all three survival methods we reviewed above, we found that the estimated effect sizes in the PWP and AG models are slightly smaller than those from the Cox PH model. We examine two Kaplan Meier plots using time until the second event occurs instead of the first event to understand the specific effect of treatment on later events. The first plot takes the start time to be time 0 (time of randomization), while the second plot takes the start time to be the ending time of the first exacerbation. Patients who have never

experienced any event will be considered left truncated and will not be included. Patients who only experience 1 event will be considered censored. Our goal is to compare these second event-time curves to the first event time to visually examine if there is a homogeneous or heterogeneous effect of treatment.

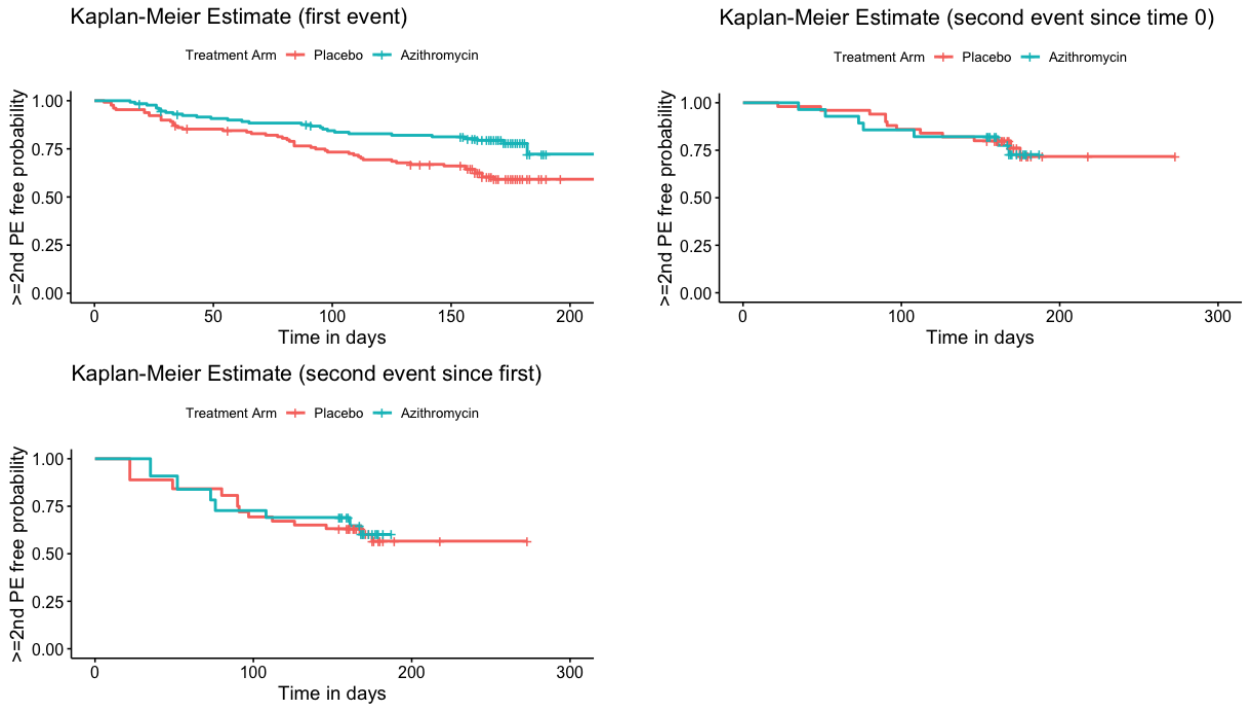


Figure 3.1.3 Kaplan Meier curves with different endpoints

The results from Figure 3.1.3 appear to indicate an attenuation of the treatment effect for events past the first. In particular curves are both closer together, and in some cases there is even a reversal of the trend (that would indicate a deleterious effect of treatment). We suspect that reversal is merely due to stochasticity of our estimates, however this still indicates a decreased positive effect of treatment. To quantitatively evaluate these effects, we fit Cox-PH models regressing time to second event on treatment. We found that the estimated hazard ratios in the Cox PH 2 and the Cox PH gap models are larger than 1, which is different from the Cox PH 1 model (Table 3.1.3). P-values for the treatment variable in the second and the third model are, however, not statistically significant, so this could easily be due to stochasticity.

We see that estimates of hazard ratio of experiencing an exacerbation in all three survival models (Cox-PH, A-G, and PWP methods) are similar. Based on our needs, we may choose different models for this endpoint. If we are not only interested in assessing the relationship between the treatment and the time until the first event, but rather in understanding the treatment's effects on the frequency of PEx, we would recommend using the two recurrent event models: These two models take all information (recurrent events) into account.

3.2 The OPTIMIZE Randomized Trial.

3.2.1 Background

The OPTIMIZE trial was a multicenter, double-blind, randomized, placebo-controlled, 18-month trial in children with CF, 6 months to 18 years of age, with early Pseudomonas aeruginosa (Pa) (Hamblett et al. 2018). The OPTIMIZE trial was conducted to test the hypothesis that the addition of azithromycin to tobramycin inhalation solution in children with cystic fibrosis and early Pa decreases the risk of pulmonary exacerbation and prolongs the time to Pa recurrence. Patients were randomized 1:1 based on randomization algorithm (Han, Enas, and McEntegart 2009) including the age group variable (Hamblett et al. 2018).

The following table gives characteristics of patients enrolled in OPTIMIZE:

Table 3.2.1. Characteristics of Participants According to Treatment Group: OPTIMIZE Trial

	Azithromycin (n=110)	Placebo (n=111)
Follow-up time [mean (sd)] (days)	319.41 (196.92)	299.33 (202.59)
Age at randomization [mean (sd)] (years)	6.58 (5.23)	6.35 (4.93)
Age Group [no(%)]		
≥6 mo to 12 y	39 (35.5%)	41 (36.9%)
>3 mo to 6 y	22 (20%)	23 (20.7%)
>6 yr to 12 y	30 (27.3%)	29 (26.1%)
>12 y	19 (17.3%)	18 (16.2%)
Sex [no(%)]		
Female	55 (50%)	49 (44%)
Male	55 (50%)	62 (56%)
PEx numbers [no(%)]		
0	67 (60.9%)	53 (47.7%)
1	24 (21.8%)	38 (34.2%)
2	9 (17.3%)	13 (18.0%)
3	4 (3.6%)	4 (3.6%)
4	2 (1.8%)	1 (0.9%)
5	2 (1.8%)	0 (0.0%)
6	2 (1.8%)	1 (0.9%)
7	0 (0.0%)	0 (0.0%)
8	0 (0.0%)	1 (0.9%)
≥1	43 (39.1%)	58 (52.3%)

Patients are roughly balanced on baseline covariates (Table 3.2.1). In the group of patients who received treatment, most patients do not have any exacerbations (60.9%). In the group of patients who received placebo, roughly half the patients had at least one exacerbation

(52.3%), compared to substantially fewer than half on the intervention arm (39.1%). In all of the models we fit for the OPTIMIZE trial, we evaluate the effect of treatment on PEx adjusting only for age group, since the original study did not adjust for sex.

3.2.2 Statistical analysis

We first evaluate potential overdispersion from a simple Poisson regression model using the overdispersion test discussed above. This resulted in a dispersion estimate of 1.46 with a corresponding p-value of $0.002<0.05$. This suggests that our outcome is over-dispersed. Thus simple Poisson regression should perform relatively poorly, and it will be of interest to use a model that accounts for overdispersion. This was verified by running a goodness-of-fit test for a simple Poisson model, resulting in a p-value of 8.23e-05 (<0.05), which indicates a poor fit.

In contrast, the negative binomial model provides a much better fit to the data: A deviance based goodness-of-fit test gives a p-value of 0.64. Unfortunately, there is no deviance-based goodness-of-fit test for Quasi-Poisson regression (Glen 2014), as only a mean model is specified. For the Poisson GLMM, we plot the random effects of each patient from the Poisson GLMM model fit (Figure 3.2.1). Unlike in the AZ trial, here the distribution of random effects is quite diffuse, with many far from 1 (Figure 3.2.1). This confirms that there is substantial between-person variability in the OPTIMIZE trial. We visually compare goodness-of-fit of Gamma and log-normal distributions to these random effects — both appear to fit similarly well, suggesting that the Poisson GLMM and negative binomial models should have very similar behavior here.

This trial exemplifies the importance of overdispersion correction when modeling count data. In addition, results on this dataset mirror what we saw in the previous chapter: All 3 methods that correct for overdispersion give very similar results. Below we give results from all of the Poisson-family models (including reported treatment effects and goodness-of-fit)

Table 3.2.2.Methods and Results: OPTIMIZE Trial

Method	Endpoint	P-value*	Goodness of fit results	Effect size:$exp(\beta)$* and 95% CI
Poisson Regression (regular)	PEx number	0.26	p=8.23e-05(dispersion)	0.84 (0.63,1.13)
Poisson GLMM	PEx number	0.27		0.80 (0.54,1.18)
Quasi-Poisson Regression	PEx number	0.4		0.84 (0.57,1.26)
Negative Binomial Regression	PEx number	0.33	p=0.64(deviance)	0.81 (0.57,1.21)

*for treatment

Results from Table 3.2.2 show that all models suggest a $\sim$ 20% reduction in number of exacerbations on treatment, however they do not provide statistically significant evidence of such an effect ($p > 0.05$ from all models). This agreement suggests that treatment may only have a minimal (or no) effect on number of exacerbations.

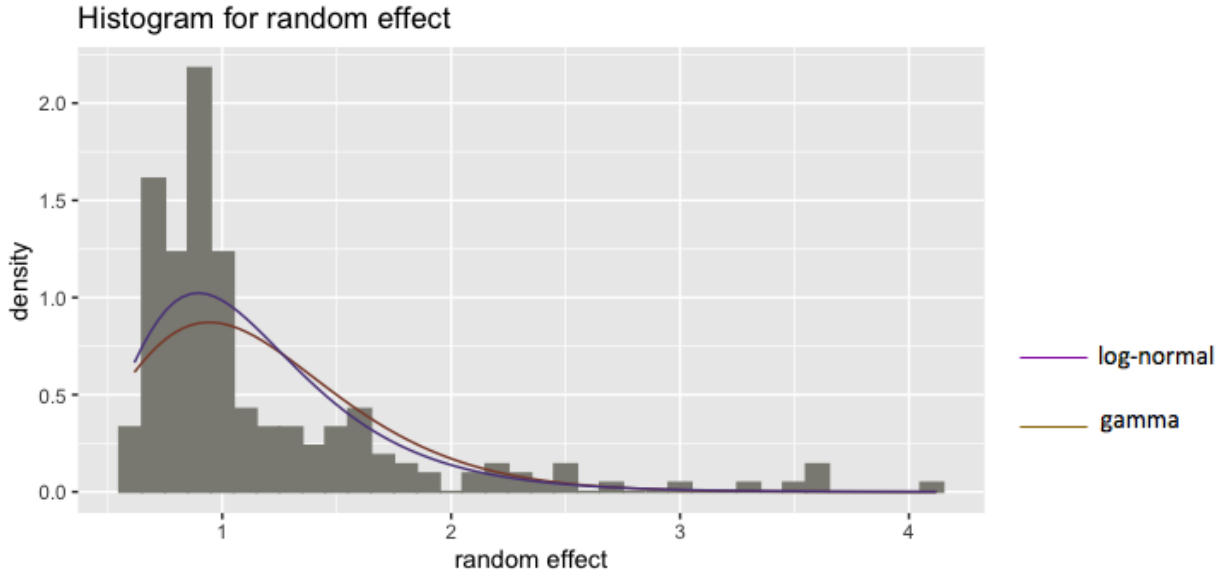


Figure 3.2.1. Random effects from Poisson GLMM (OPTIMIZE)

We now move on and consider proportional-hazard modeling methods. Again, we first check if the proportional hazard assumption is satisfied. In each level of age group, we examine the Kaplan-Meier curves for time to first exacerbation in the two treatment arms:

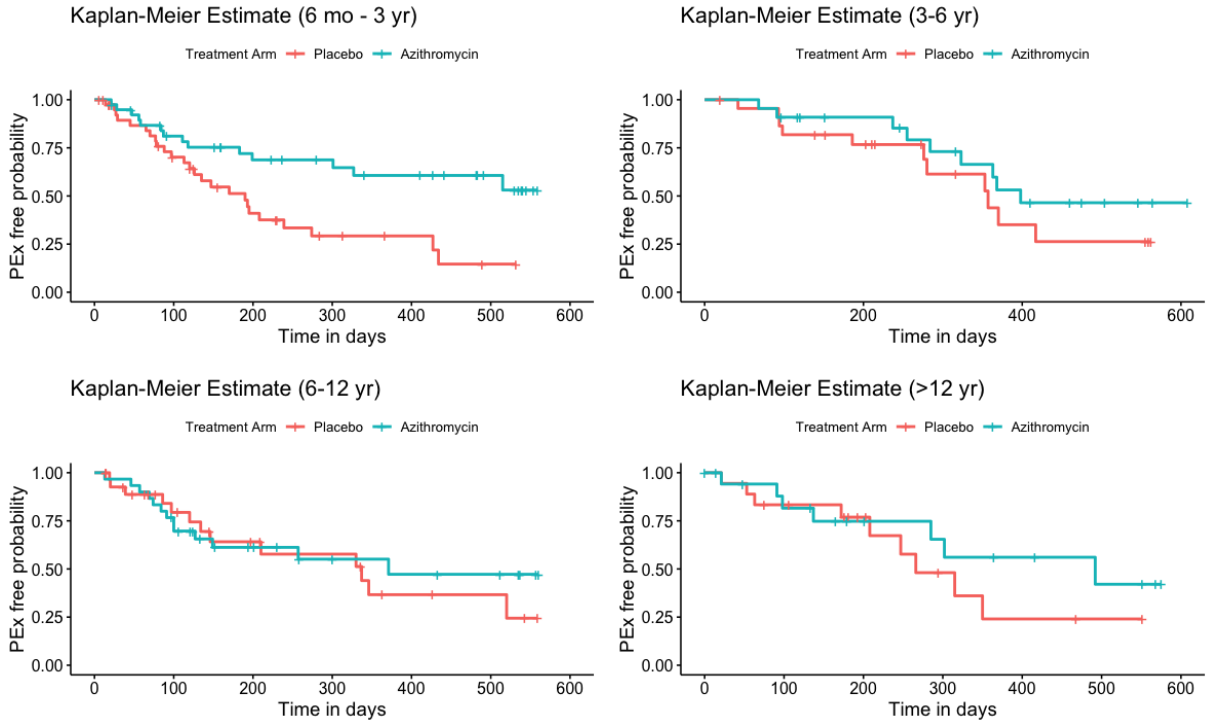


Figure 3.2.2 Kaplan Meier Curves until the first event, stratified on age group

From Figure 3.2.2 we see that among patients who are in age groups 3 and 4 (6-12 years, and older than 12 years), the Kaplan-Meier curves for the two treatment groups cross. As before, we cannot tell from the plot if this just a function of the variability of KM curves based on limited data. We quantitatively evaluate the proportional hazards assumption by running a test of proportional hazards (Harrell 1986) and find that we do not have sufficient evidence to reject proportional hazards in any stratum: p-value=0.18 (6 months - 3 years); p-value =0.9 (3-6 years); p-value=0.25 (6-12 years); p-value=0.59 (>12 years). We note though that this may be due to our limited number of observations in each stratum.

Results of the Cox-PH model for the first event are shown in Table 3.2.3 (Cox PH 1). Stratified on age group, the exponentiated estimated coefficient of treatment is 0.56 and the p-value ($p=0.004<0.05$) shows statistical significance. This indicates that patients receiving azithromycin have a $\sim$ 44% reduction in hazard of experiencing their first exacerbation at time t as compared to those who receive placebo.

Table 3.2.3. Cox PH Model with Different Endpoints: OPTIMIZE Trial

Model	Endpoint	P-value*	Effect size: HR and 95% CI
Cox PH 1	Time from start to first event	$0.004 < 0.05$	0.56 (0.37,0.83)
Cox PH 2	Time from start to second event	0.16	1.59 (0.84,3.01)
Cox PH gap	Time from first event to second event	0.58	0.80 (0.36,1.78)
Andersen Gill	Time from start until each event	0.36	0.84 (0.57,1.22)
PWP-TT	Time from start until each event	0.26	0.85 (0.59,1.22)
PWP-GT	Time from the previous until the next event	0.24	0.83 (0.60,1.15)

*for treatment

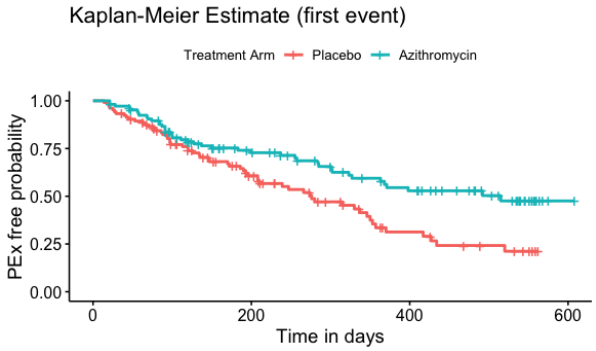
In addition, to incorporate later events, we used an A-G model. As a reminder, this assumes that the time intervals between exacerbations are conditionally uncorrelated.

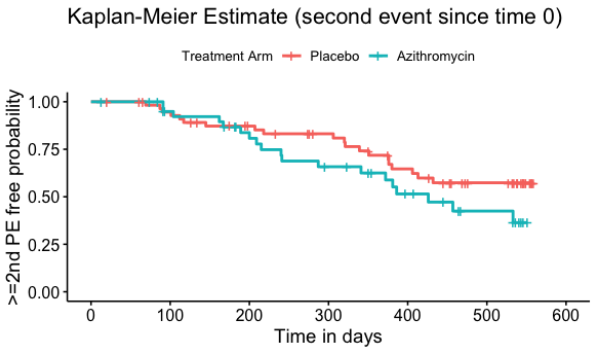
Results from the A-G model, however, indicate that the treatment is not associated with a

change in the hazard of experiencing exacerbations, with p-value=0.35 >0.05 (Table 3.2.3). This suggests that although the treatment is effective at reducing the hazard of experiencing the first event, the treatment is not effective in reducing the hazard of having a following event. As this method averages over all events, it leads to some concern that, in fact, treatment increases risk of a later event.

Next, we try both a PWP-GT model and a PWP-TT model. As a reminder this allows us to use a different baseline hazard for each event-number. Since PWP methods may encounter small-risk-set issues, we removed any events after the fourth (only 9 patients experienced 4 or more events). The results from these two methods agree with the results from the AG model and did not identify a significant effect of treatment; p = 0.26, and 0.24 for PWP-TT and PWP-GT models respectively (Table 3.2.3).

Because we found that the estimated effect sizes from PWP and AG models are much smaller than that of the Cox PH model (and no longer significant), we would like to identify what is happening with events after the first. To better understand what is going on, we examined Kaplan-Meier plots: The first plot takes the start time to be time 0 (time of randomization), and the second plot takes the start time to be termination of the first exacerbation. Patients who have never experienced any event are considered left truncated and are not be included. Patients who only experience 1 event are considered censored.





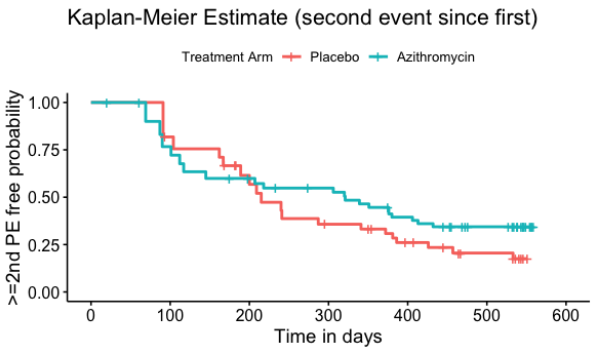


Figure 3.2.3 Kaplan Meier curves with different endpoints

From Figure 3.2.3, we see potential violation of proportional hazards (crossing of the curves). In addition, we see a potential reversal in the effect of treatment for the second event (that patients on the intervention arm may, at some points, have increased hazard). To quantitatively evaluate this, we fit Cox-PH models regressing time to second event on treatment, accounting for age group. For the first model (Cox-PH 2), we set the start time as the patient's randomization time. For the second model (Cox-PH GAP) we left-truncate observations at the time of the first exacerbation. Cox-PH GAP shows a much attenuated effect from the original Cox-PH model for the first event (Cox-PH 1). Cox-PH 2 has a point estimate with a reversed effect. Further investigations may be needed to understand this: It may be due to post-randomization factors such as degree of treatment in response to the first event. Neither of these models attains statistical significance.

From Figure 3.2.3 and Table 3.2.3, we found that the relationship between treatment and exacerbations beyond the first exacerbation is unclear in the OPTIMIZE trial. If we are more interested in understanding the treatment's effects on overall frequency of PEx (including later events), it seems most sensible to use the recurrent event models. However, as we see here, these result in a very different effect estimate than we found from evaluating only time to the first event via a Cox-PH model.

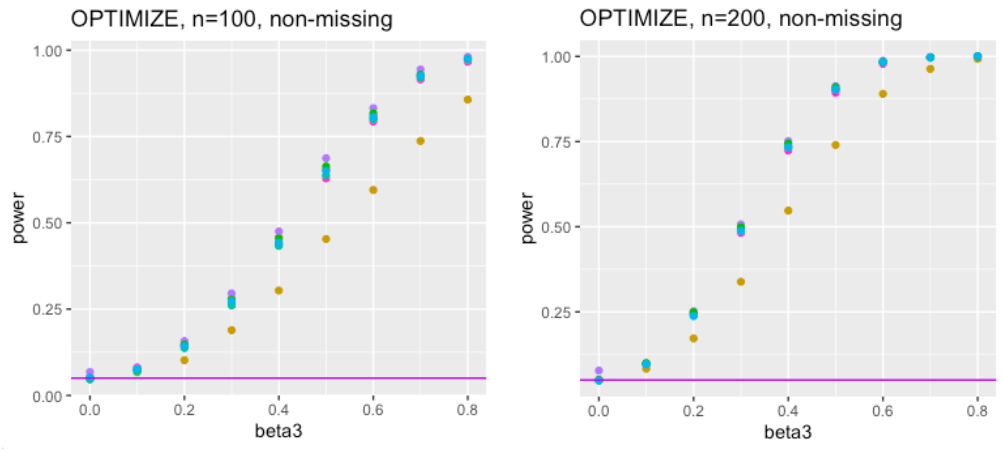
3.2.3 Data simulation revisited: overdispersion with random effects from the OPTIMIZE trial

We now consider a data simulation with random effects extracted from the OPTIMIZE trial. We examine how this longer-tailed distribution of random effects will affect the performance of our methods. By considering this in simulation where we know the parameters used to generate the data, we can more easily evaluate the performance of our methods.

We generate the data precisely as in the overdispersion simulations from Chapter 2. The only difference is that to generate random effects we sample with replacement from the estimated random effects obtained from our Poisson GLMM fit to the OPTIMIZE trial data. We again consider 3 censoring scenarios, and a nominal significance level of $\alpha = 0.05$.

Similar to what we saw on over-dispersed data in Chapter 2, we found inflation in the Type I error-rate of simple Poisson regression. As we discussed in Chapter 2, an inflated Type I error-rate can unfairly lead to increased power. To combat this, we recalibrated the cutoffs for Poisson regression as in Chapter 2 to control type 1 error rates and put all methods on an even footing for comparing power.

Non-missing



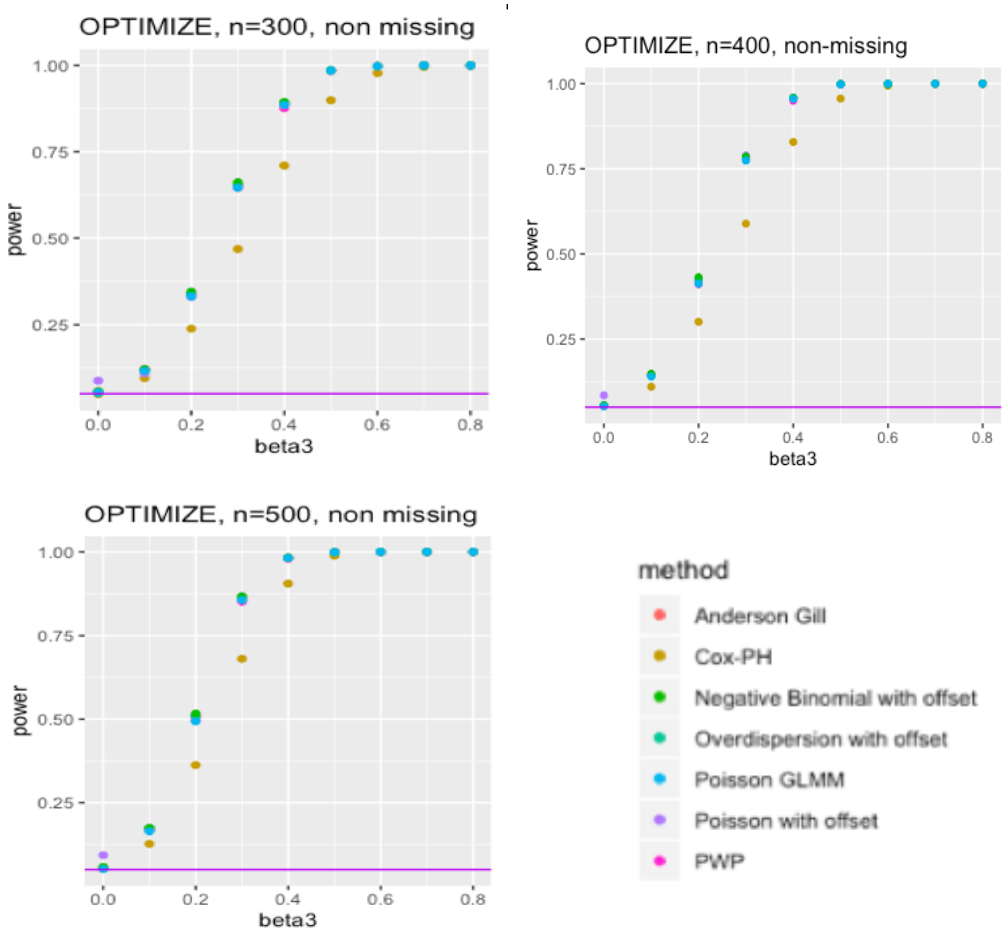
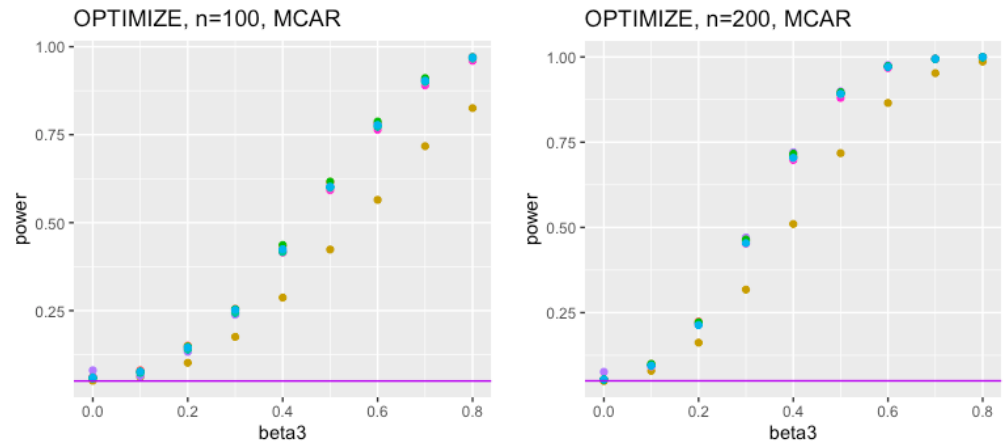


Figure 3.2.4: Power plots, overdispersed with random effects, non-missing

Missing completely at random



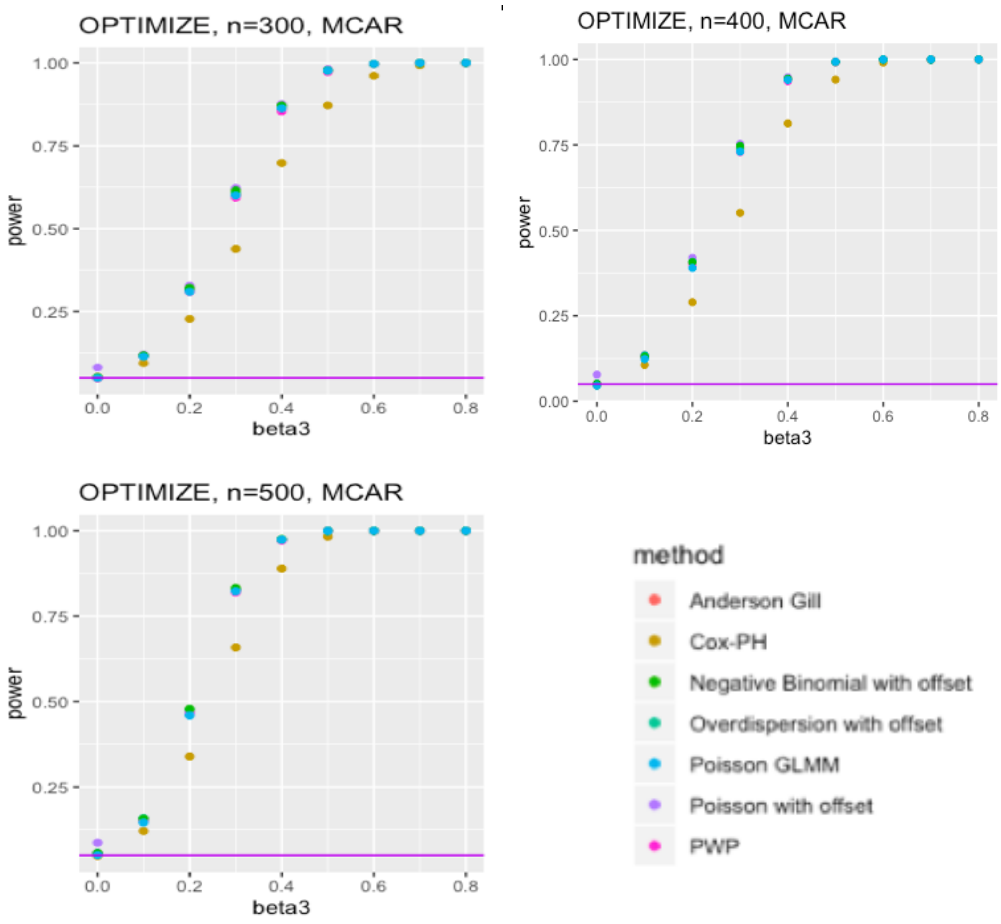
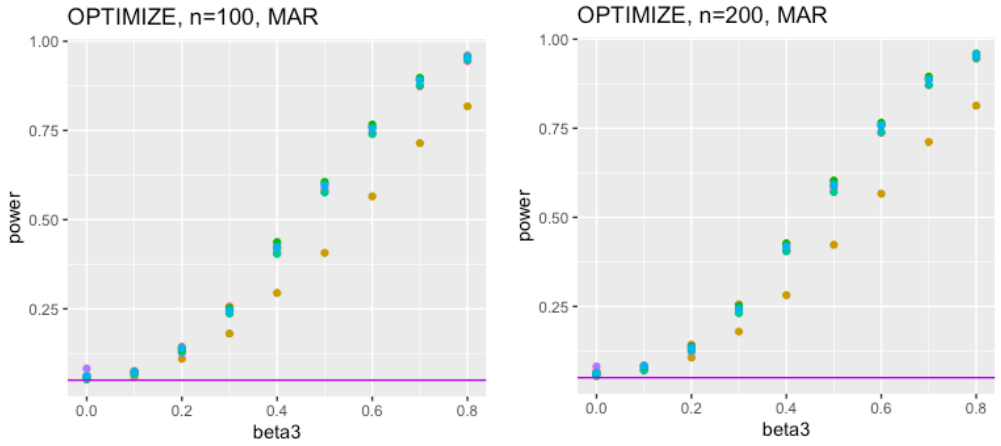


Figure 3.2.5: Power plots, overdispersed with random effects, MCAR

Missing at random



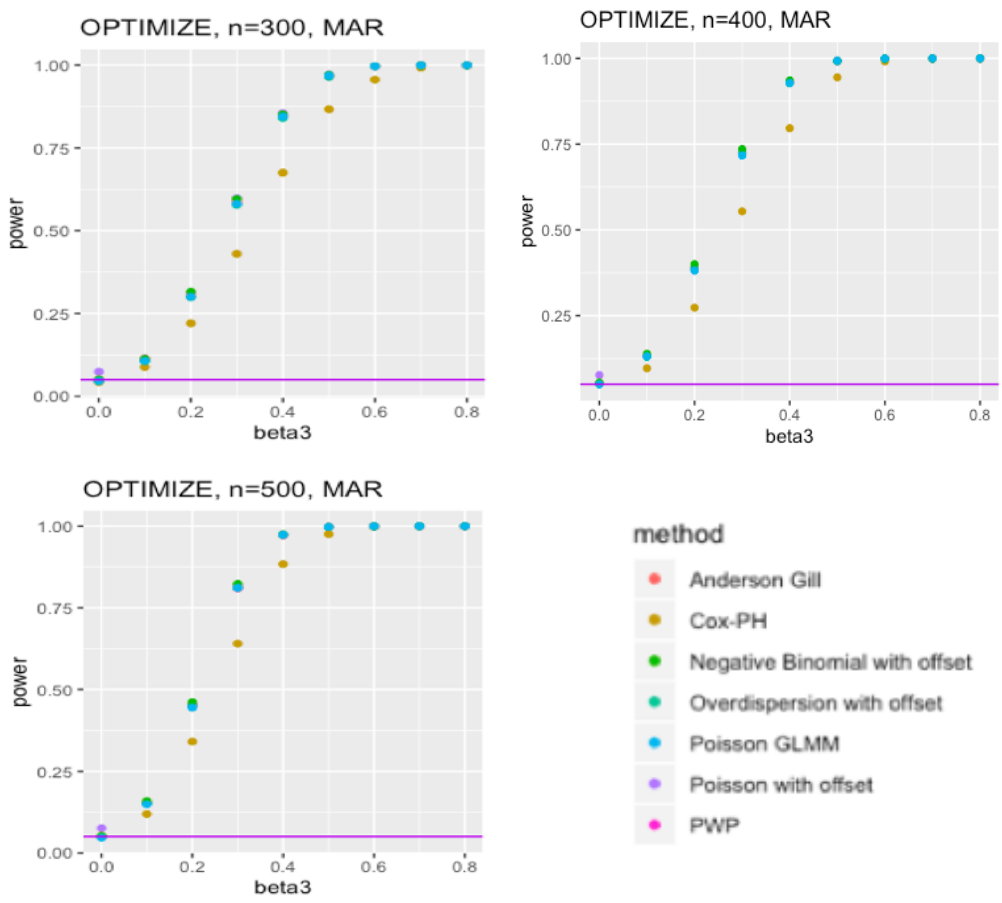


Figure 3.2.6: Power plots, overdispersed with random effects, MAR

The results from this simulation study are extremely similar to what we saw in Chapter 2. As noted before, Poisson regression has inflated type 1 error rates; all Poisson-family methods that deal with overdispersion perform quite well (and similarly); and Cox-PH regression struggles with power when it ignores later events.

The lower power from Cox-PH regression here comes from our assumption of homogeneity of the treatment effect across events. In the case of the OPTIMIZE trial that appears not to be the case. Further simulations under treatment effect heterogeneity may be warranted.

4 Discussions and Conclusions

In this thesis, we examined 7 different methodologies for modeling the relationship between treatment and pulmonary exacerbation in Cystic Fibrosis. We evaluated the advantages and disadvantages of these methods via theoretical study, simulation, and reanalysis of 2 contemporary trials. Here we will summarize our findings.

There are two possible general ways to engage PEx outcome in our models: One can treat the outcome as a count, and model the rate, or one can treat the outcome as a point-process and model the hazard. The two approaches lead to different methodologies: Treating PEx as a count leads us to use Poisson-like models; while treating PEx as a process leads us to use Cox-PH-like (time-to-event) models.

We evaluated 4 methodologies in the Poisson-modeling family: Poisson GLM, Quasi-Poisson GLM, Poisson GLMM, and the negative binomial model. The three latter models are extensions of the Poisson GLM, developed to deal with overdispersion. We found that those three extensions performed at least as well as Poisson regression in every scenario (including those without overdispersion): In all scenarios Type-1 error rates were controlled, and power was as good as Poisson regression. In contrast, Poisson regression had inflated Type-1 error rates in all scenarios with overdispersion. All three overdispersion-correcting methods performed very similarly. In some scenarios, the negative binomial model or the Poisson GLMM might be preferable because they allow one to explicitly model random effects while quasi-Poisson model does not. Additionally, in scenarios where computational resources are a concern, we suggest a negative binomial model as it is less computationally burdensome to fit.

If one decides to treat PEx as a [recurrent] time-to-event outcome, it is important to decide whether to engage with events beyond the first. When there is homogeneity of treatment effect, engaging with events beyond the first will increase power. We saw this in our simulated experiments: There the Cox-PH model has substantially lower power as compared to the AG model and PWP methods in all scenarios. When there is heterogeneity of treatment

effect, however, considering later events may decrease power. We saw this in the OPTIMIZE trial data, where we found that the behavior of the first exacerbation greatly differs from behavior/timing of later exacerbations. In particular, the Cox PH model showed statistical significance of the treatment effect, but AG and PWP methods did not (due to an attenuation of treatment effect in later events).

In addition to power considerations, there are two additional determinants that we believe are important when deciding whether to use only the first event vs recurrent events. First, one must consider the scientific question: If time until the first event is of primary interest then we believe it makes most sense to use a standard Cox-PH model which considers only the first event. If we are interested in an average hazard (averaged over all exacerbations), then we would recommend using the AG model or PWP methods. In particular, estimates of hazard ratios from the Cox-PH may be very different from those of AG model or PWP methods in cases where the treatment effect is heterogeneous. The second important consideration is more administrative (though it relates to the scientific question). It seems likely that timing on a second exacerbation is more difficult to define and measure as compared to a first, as one needs to identify, for example, when the first exacerbation ends, and if the first vs second exacerbation are actually different events (or just one long worsening). Additionally there may be some confounding induced by treatment of the first exacerbation (perhaps a less severe first exacerbation leads to less intensive treatment and potentially a sooner second exacerbation). These issues may attenuate or more generally bias estimation of the effect of treatment on later exacerbations, and this might push us towards using only time to first exacerbation as our outcome.

In a prospectively planned trial and analysis, we must choose our methodology before looking at the data. We believe that engaging with the above questions and our scientific objectives in a given study can help inform our choice of analysis.

Appendix

The data analysis of this thesis is conducted using R (R Core Team 2013), version 3.5.3. R codes used for this thesis can be found in this GitHub link:

https://github.com/xydai/generation_poisson

References

Andersen, PK, and RD Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics* 10 (4): 1100–1120. doi:10.1214/aos/1176345976.

Breslow, NE, and DG Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88 (421): 9–25. doi:10.1080/01621459.1993.10594284.

Bruin, J. 2011. "Newtest: Command to Compute New Test @ONLINE." February. https://stats.idre.ucla.edu/stata/ado/analysis/.

Cameron, AC, and PK Trivedi. 1990. "Regression-Based Tests for Overdispersion in the Poisson Model." *Journal of Econometrics* 46 (3): 347–64. doi:10.1016/0304-4076(90)90014-K.

Castañeda, J, and B Gerritse. 2010. "Appraisal of Several Methods to Model Time to Multiple Events Per Subject: Modelling Time to Hospitalizations and Death." *Revista Colombiana de Estadistíca* 33 (1): 43–61.

Cox, DR. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (2): 187–220. doi:10.1111/j.2517-6161.1972.tb00899.x.

Glen, Stephanie. 2014. "Goodness of Fit Test: What is it? from Statisticshowto.com: Elementary Statistics for the Rest of Us!" https://www.statisticshowto.com/goodness-of-fit-test/.

Hamblett, NM, GR Bogart, M Kloster, and et al. 2018. "Azithromycin for Early Pseudomonas Infection in Cystic Fibrosis." *Am J Respir Crit Care Med* 198 (9): 1177–87. doi:10.1164/rccm.201802-0215OC.

Hamblett, NM, L Saiman, LC. Lands, M Anstead, M Rosenfeld, M Kloster, L Fisher, and F Ratjen. 2013. "Impact of Acute Antibiotic Therapy on the Pulmonary Exacerbation

Endpoint in Cystic Fibrosis Clinical Trials." *Contemporary Clinical Trials* 36 (1): 99–105. doi:10.1016/j.cct.2013.06.004.

Han, Baoguang, NH Enas, and Damian McEntegart. 2009. "Randomization by Minimization for Unbalanced Treatment Allocation." *Stat Med.* 28 (27): 3329–46. doi:10.1002/sim.3710.

Harrell, FE. 1986. "SAS Supplemented Library User's Guide, Version 5, Cary." In. NC: SAS Institute Inc.

Keene, ON, PMA Calverley, PW Jones, J Vestbo, and JA Anderson. 2008. "Statistical Analysis of Exacerbation Rates in Copd:TRISTAN and Isolde Revisited." *Eur Respir J* 32 (1): 17–24. doi:10.1183/09031936.00161507.

Kleiber, Christian, and Achim Zeileis. 2020. *AER: Applied Econometrics with R.*

McCullagh, Peter. 1983. "Quasi-Likelihood Functions." *Ann. Statist.* 11 (1): 59–67. doi:10.1214/aos/1176346056.

Ozga, AK, M Kieser, and G Rauch. 2018. "A Systematic Comparison of Recurrent Event Models for Application to Composite Endpoints." *BMC Medical Research Methodology* 18 (2): 1–12. doi:10.1186/s12874-017-0462-x.

Prentice, RL, BJ Williams, and AV Peterson. 1981. "On the Regression Analysis of Multivariate Failure Time Data." *Biometrika* 68 (2): 373–79. doi:10.1093/biomet/68.2.373.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Saiman, L, M Anstead, NM Hamblett, and et al. 2010. "Effect of Azithromycin on Pulmonary Function in Patients with Cystic Fibrosis Uninfected with Pseudomonas Aeruginosa: A Randomized Controlled Trial." *JAMA* 303 (17): 1707–15. doi:10.1001/jama.2010.563.

Therneau, Terry, Thomas Lumley, Atkinson Elizabeth, and Crowson Cynthia. 2020. *Survival: Survival Analysis.*