Frequentists and Bayesian methods to incorporate recruitment rate stochasticity at the design stage of a clinical trial

---

Master Thesis in Biostatistics (STA495)

by

Pilar Pastor
23-733-975

supervised by

PD Dr. Malgorzata Roos

Zurich, month year

Frequentists and Bayesian methods to incorporate recruitment rate stochasticity at the design stage of a clinical trial

Pilar Pastor

Version April 9, 2025

Contents

Preface

Howdy!

Pilar Pastor
June 2025

Chapter 1

Introduction

Why, what and how...

Chapter 2

Methodology

2.1 Definitions

The general notion of **Recruitment** in this Master Thesis refers to the number of patients (Counts) at the Eligibility, or Enrollment, or Randomization, or Statistical Analysis stage in Figure 2.1. Figure 2.1 is a schematic representation of a PRISMA study flow chart inspired by a real PRISMA study flow chart in Figure 2.2. We define **Accrual** as cumulative recruitment.

The **Target Population** is a specific group within the broader population, defined by attributes relevant to the research question. This group is focused on criteria that match the study's goals (Willie, 2024). Defining the target population allows researchers to refine their objectives and recruitment methods to align with the study's aims.

The **Eligibility** criteria are the specific requirements that individuals must meet to participate in a study. Eligible patients will be selected from the target population. Inclusion criteria specify the conditions that allow individuals to participate in the trial, particularly focusing on the medical condition of interest. Any other factors that limit eligibility are classified as exclusion criteria (Van Spall *et al.*, 2007), conditions or circumstances that disqualify potential participants (Food *et al.*, 2018).

In clinical trials, **Enrollment** refers to the formal process of registering participants into a study after they have met all eligibility criteria and provided informed consent. This process includes verifying that each participant satisfies the inclusion and exclusion criteria outlined in the study protocol (National Institute of Allergy and Infectious Diseases, 2021). It is important to distinguish between recruitment and enrollment. Recruitment involves identifying and inviting potential participants to join the study, whereas enrollment occurs after these individuals have been screened, consented, and officially registered into the trial (Frank, 2004).

Once enrolled, participants are assigned to specific treatment groups or interventions as defined by the study design. The most common practice is **Randomization**. In clinical research, randomization is the process of assigning participants to different treatment groups using chance methods, such as random number generators or coin flips (Lim and In, 2019). Randomized controlled trials (RCTs) are considered the most effective method for preventing bias in the evaluation of new interventions, drugs, or devices. (Van Spall *et al.*, 2007).

In clinical research, **Statistical Analysis** involves applying statistical methods to collect, summarize, interpret, and present data derived from clinical studies. This process is essential for evaluating the safety, efficacy, and overall outcomes of medical interventions, ensuring that conclusions drawn are both reliable and valid (Panos and Boeckler, 2023). Not all participants who are randomized may be included in the final statistical analysis due to protocol deviations of patients not adhering to the protocol (Rehman *et al.*, 2020), missing data (Shih, 2002) or loss-to-follow-up, some participants may become unreachable or withdraw consent during the study, resulting in missing outcome data (Nüesch *et al.*, 2009).

The number of patients decreases at each stage of a clinical study, from defining the target

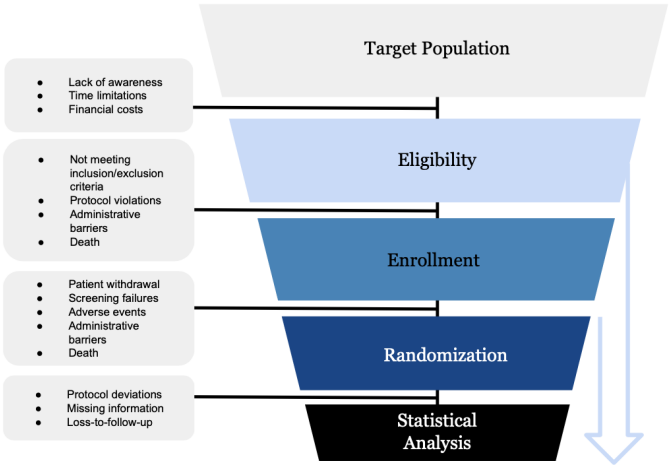

Figure 2.1: Patient recruitment and leakage at each stage of a clinical study (Piantadosi and Meinert, 2022; Whelan *et al.*, 2018; Bogin, 2022).

population to final statistical analysis, see Figure 2.3. This process is known as patient leakage (Desai, 2014), alternative terms are attrition or retention.

Figure 2.1 generalizes the notion of patient leakage found in trial profiles such as the one found in Figure 2.2. Figure 2.1 outlines the various stages of a clinical trial and analyzes the key factors contributing to patient attrition as they transition from one stage to the next.

Eligibility criteria narrow down participants, and enrollment further reduces numbers as only those meeting strict criteria are registered. Randomization assigns individuals to treatment groups, but some may later be excluded due to protocol deviations, missing data, or loss to follow-up.

2.2 Uncertainty and models for counts and time

There are two types of uncertainty, aleatory and epistemic (O'Hagan, 2006). The **Aleatory Uncertainty** reflects randomness that is inherent, irreducible and unpredictable in nature. **Epistemic Uncertainty** arises primarily from limited or imperfect knowledge about the parameters of a statistical model and can reflect fluctuations of the parameter. Obtaining more or better information about the parameter typically reduces the epistemic uncertainty.

Let us denote

- $T = time$
- $C = counts$
- $\lambda = \frac{C}{T}$

We define **Recruitment Rate** $\lambda = \frac{C}{T}$ at which patients are collected, measured as persons per unit of time, where **Rate** is understood as a ratio in which the numerator and denominator are incremental differences (Piantadosi, 2024):

$$\lambda = \frac{\Delta C}{\Delta T} = \frac{C_1 - C_0}{T_1 - T_0} = \frac{C_1 - 0}{T_1 - 0} = \frac{C_1}{T_1}$$

Methods in Tables 2.1, 2.2 and 2.3 are applicable to each level of recruitment in Figures 2.1 and 2.3.

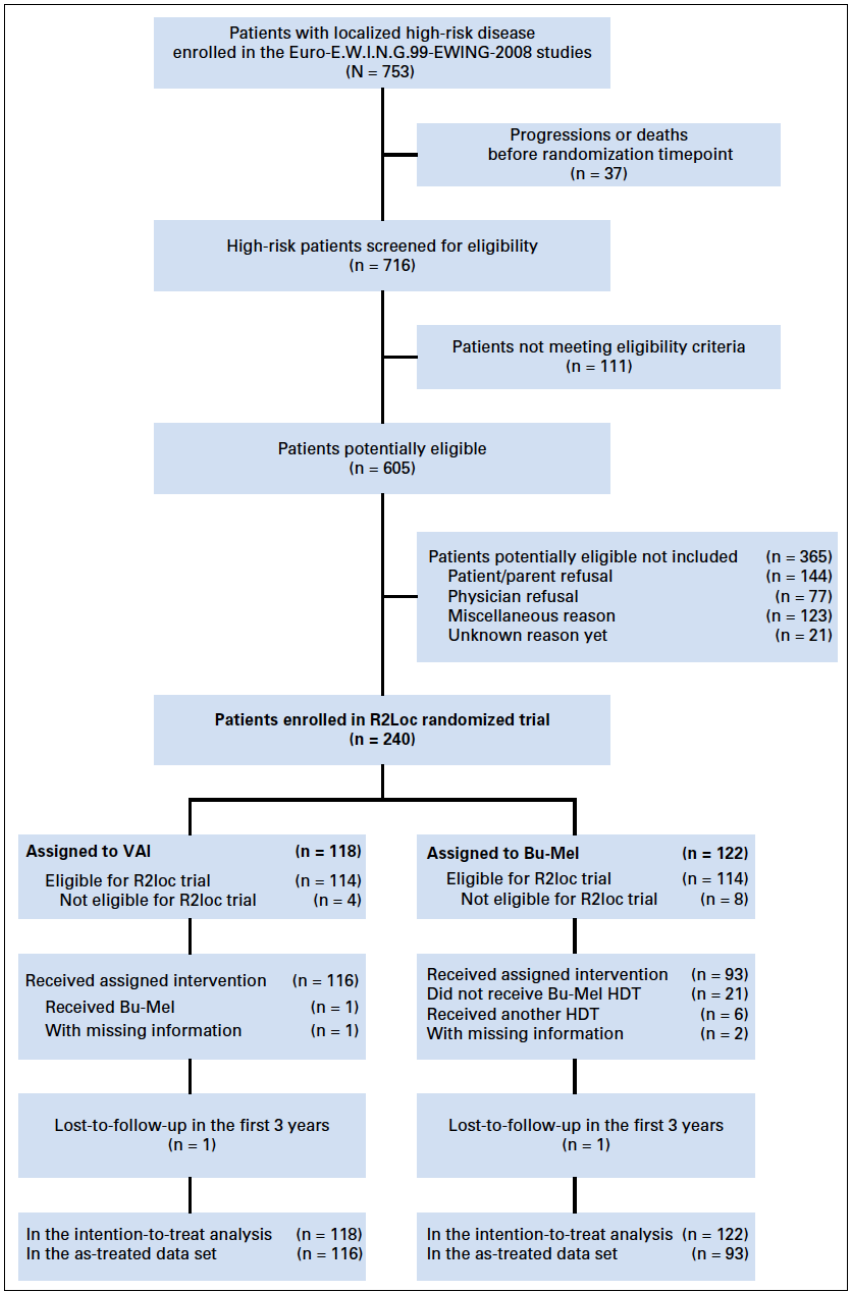

Figure 2.2: PRISMA study flow-chart (Whelan *et al.*, 2018).

Methods	Counts	Expectation	Variance	Aleatory	Epistemic
Expectation	$C = \lambda$	λ	0	No	No
Poisson	$C \sim \text{Po}(\lambda)$	λ	λ	Yes	No
Poisson-Gamma	$C \sim \text{Po}(\Lambda);\ \Lambda \sim \text{G}(\alpha, \beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha(\beta+1)}{\beta^2}$	Yes	Yes

Table 2.1: Moments and aleatory and epistemic uncertainty of recruitment in one unit of time recruitment covered by different models for counts.

2.3 Counts: Model based on Expectations

If we fix the duration of a study at time T and we expect that we collect C patients until T, we deterministically predict the recruitment rate per one unit of time (without taking into consideration any uncertainty) to be $\hat{\lambda} = \frac{C}{T}$ (Carter, 2004).

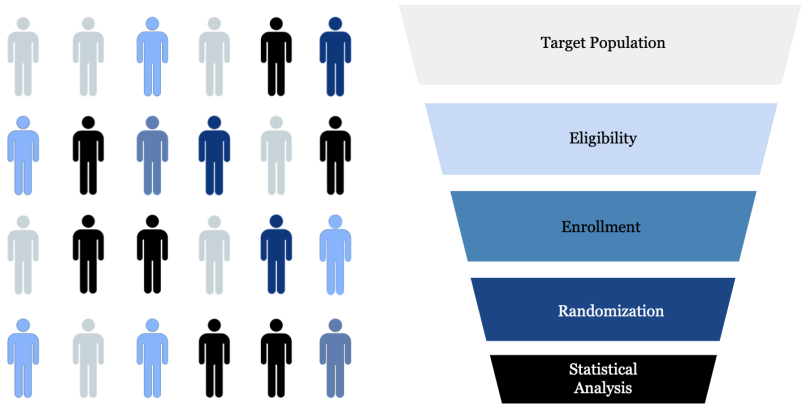

Figure 2.3: Visual representation of patient recruitment and leakage at each stage of a clinical study (Piantadosi and Meinert, 2022; Whelan *et al.*, 2018; Bogin, 2022).

Methods	Counts	Expectation	Variance	Aleatory	Epistemic
Expectation	$C(t) = \lambda t$	λt	0	No	No
Poisson	$C(t) \sim \text{Po}(\lambda t)$	λt	λt	Yes	No
Poisson-Gamma	$C(t) \sim \text{Po}(\Lambda t);\ \Lambda \sim \text{G}(\alpha, \beta)$	$t\frac{\alpha}{\beta}$	$t\frac{\alpha(\beta+t)}{\beta^2}$	Yes	Yes

Table 2.2: Moments and aleatory and epistemic uncertainty in accrual until t covered by different models for counts.

Methods	Time	Expectation	Variance	Aleatory	Epistemic
Expectation	$T(c) = c/\lambda$	c/λ	0	No	No
Erlang	$T(c) \sim \text{G}(c, \lambda)$	c/λ	c/λ^2	Yes	No
Gamma-Gamma	$T(c) \sim \text{G}(c, \Lambda);\ \Lambda \sim \text{G}(\alpha, \beta)$	$c\frac{\beta}{\alpha-1}$	$\frac{c\beta^2(c+\alpha-1)}{(\alpha-1)^2(\alpha-2)}$	Yes	Yes

Table 2.3: Moments and aleatory and epistemic uncertainty of recruitment covered by different models for time having a fixed sample size c.

2.3.1 Expected recruitment in one unit of time

$C = \text{E}C = \text{E}\lambda = \lambda$
$\text{Var}(C) = \text{Var}(\lambda) = 0$

As we can see in Table 2.1

2.3.2 Expected accrual at time point t

Assuming that recruitments in each unit of time are independent of each other we have:

$$C(t) = \text{E}(\underbrace{C + \ldots + C}_{t \text{ times}}) = \text{E}(\lambda t) = \lambda t$$

$$\text{Var}(C(t)) = \text{Var}(\underbrace{C + \ldots + C}_{t \text{ times}}) = t\text{Var}(\lambda) = 0$$

Both the expected accrual and its zero-variance are recorded in Table 2.2 and visualized in Figure 2.4 and Figure 2.7.

2.3.3 Criticism

This is a simple deterministic method based on a linear extrapolation of the constant expected recruitment also called "First Order Recruitment Model" (FORM) (Comfort, 2013). It is also

referred to as the *unconditional approach* and its main limitation is the lack of a mechanism to account for known sources of variation in the rate (Carter *et al.*, 2005).

The model based on expectations is overly simplistic and fails to account for changes in center recruitment or the regulatory environment (Barnard *et al.*, 2010). Therefore, stochastic models that incorporate randomness in the recruitment process are more suitable than the widely used deterministic approach (Zhang and Long, 2012).

2.4 Counts: Model based on Poisson Process

One way to incorporate variation in the mean number of participants per day is by assuming that participants are recruited according to a known probability distribution, such as the Poisson distribution (Carter, 2004). This approach emulates trial accrual using a random number generator and records the time required to reach the target sample size over multiple iterations (Carter *et al.*, 2005).

The resulting distribution helps estimate the probability of completing accrual within a given time frame and assess variability in accrual time. This method is particularly useful when a trial has a fixed duration, as it allows researchers to determine the necessary number of clinical centers and monthly recruitment rate to achieve a high probability (e.g., 80%) of completing enrollment on time (Carter *et al.*, 2005).

The Poisson distribution $C \sim \text{Po}(\lambda)$ allows us to explain the recruitment of patients. It is a discrete variable that expresses the probability of a given number of events (in our case, patient recruitment) occurring in a fixed unit interval of time. We assume that these events occur with a known constant rate λ and are independent of each other.

$$\begin{aligned} \text{P}[C = c] &= \frac{\lambda^c}{c!} e^{-\lambda} \\ c &= 0, 1, 2, \ldots \end{aligned}$$

One important property from the Poisson distribution is that it is infinitely divisible (Held and Bové, 2014). If $X_i \sim \text{Po}(\lambda_i)$ for $i = 1, \ldots, n$ are independent, then, $\sum_{i=1}^{n} X_i \sim \text{Po}\Big(\sum_{i=1}^{n} \lambda_i\Big)$.

2.4.1 Recruitment in one unit of time

The recruitment of patients in one unit of time follows $C \sim \text{Po}(\lambda)$ and the expectation and variance are:

$$\begin{aligned} \text{E}C &= \lambda \\ \text{Var}(C) &= \lambda \end{aligned}$$

As we can see in Table 2.1

2.4.2 Accrual at time point t

At time point t, the accrual follows $C \sim \text{Po}(\lambda t)$. Using the infinitely divisible property from the Poisson applicable to independent random variables, $\underbrace{\text{Po}(\lambda) + \cdots + \text{Po}(\lambda)}_{t \text{ times}} = \text{Po}(\lambda t)$. We assume that the recruitment of patients in t unit time intervals is independent from another. As we can see in Table 2.2, the expectation and variance are the following:

$$\begin{aligned} \text{E}C(t) &= \lambda t \\ \text{Var}(C(t)) &= \lambda t \end{aligned}$$

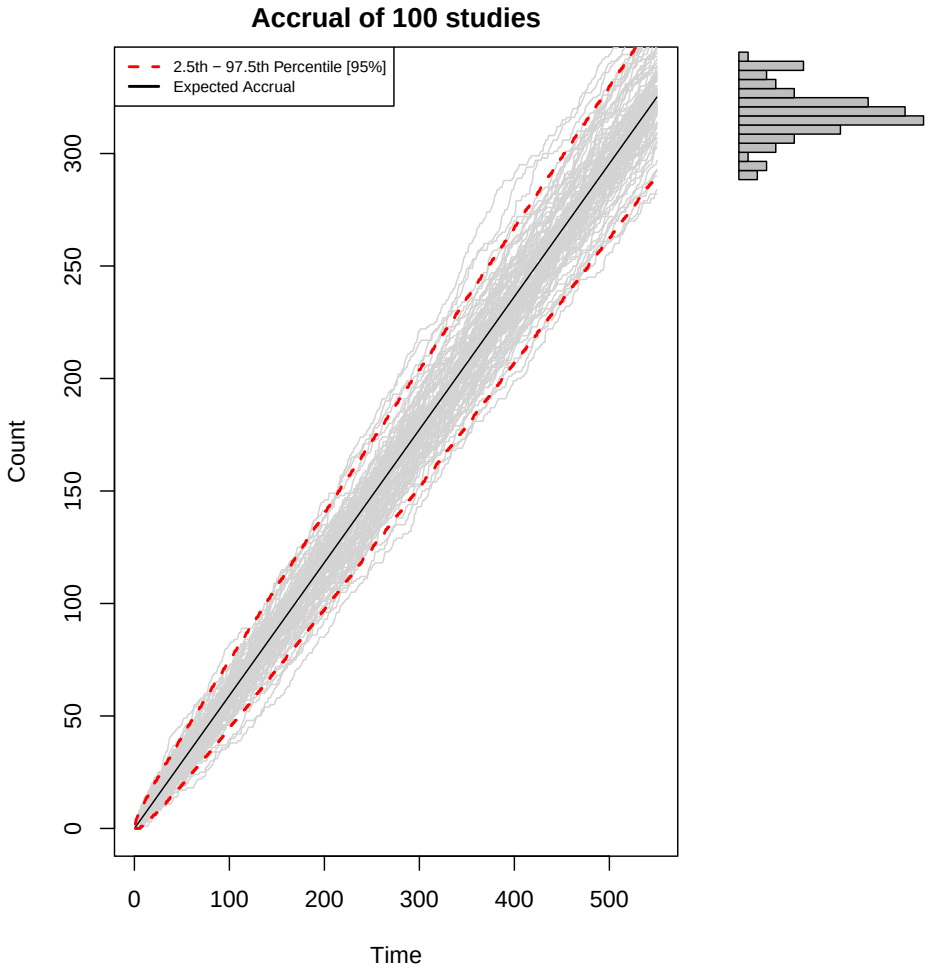

Figure 2.4: Poisson-distributed counts with $\lambda = 0.591$ per day and uncertainty range. The black line represents the point estimate of the expected accrual from section 2.3, while the red dashed lines indicate Poisson's 95% aleatory uncertainty. The histogram illustrates the distribution of observed counts in 100 studies at time $t = 550$ days (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

For example, if we assume $\lambda = 0.591$ per day and $t = 550$, we can show the accrual of 100 different studies in Figure 2.4 and the histogram at $t = 550$ days. The exact distribution at $t = 550$ is provided in Figure 2.5 and the Cummulative Distribution Function (CDF) in Figure 2.6. The uncertainty bands based on the exact quantiles are displayed in Figure 2.7.

2.4.3 Criticism

This model was used by Carter (2004) and Carter *et al.* (2005). Although this model accounts for aleatory uncertainty, the recruitment rate is assumed to be constant for the entire period of time. Therefore, an alternative method that accounts for varying recruitment rates over time is necessary.

2.5 Counts: Negative Binomial model derived from Poisson-Gamma model

The basic Poisson model does not account for variations in recruitment rates or uncertainties in rate estimates (Mountain and Sherlock, 2022). To address this Anisimov and Fedorov (2007)

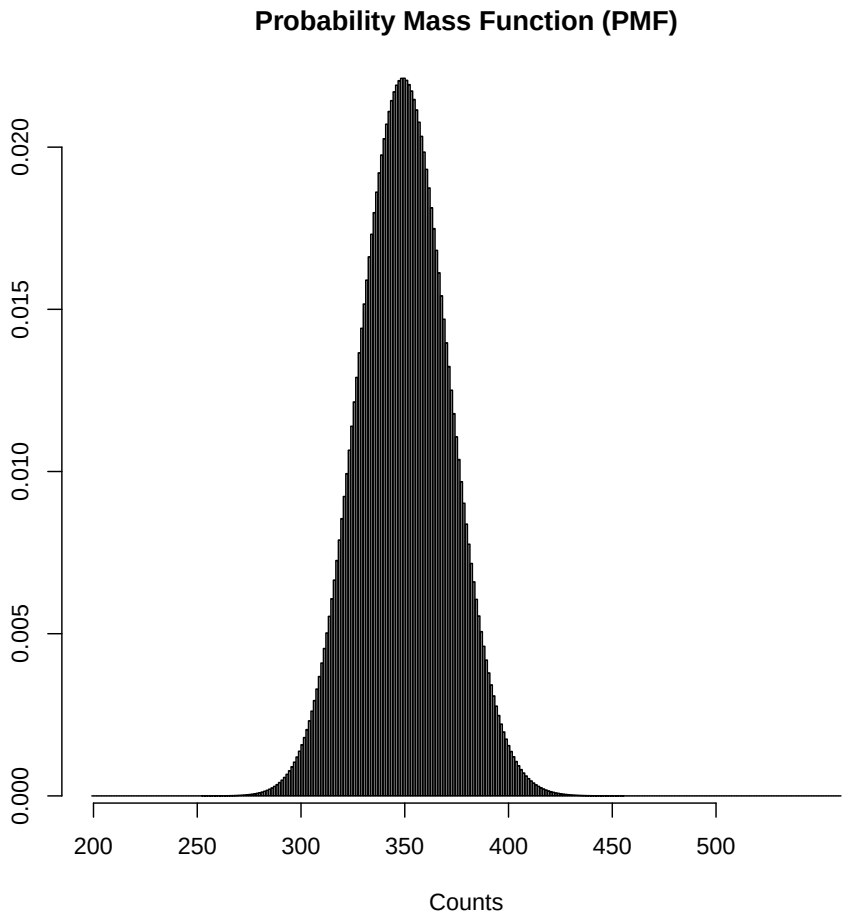

Figure 2.5: Probability Mass Function (PMF) of Poisson-distributed counts: This bar plot represents the probability mass function (PMF) of counts ranging from 200 to 500, using a Poisson distribution $\text{Po}(\lambda t)$ with a rate parameter $\lambda = 0.591$ per day at time $t = 550$ days.

propose a random effects model where recruitment follows a homogeneous Poisson process with rates drawn from a gamma distribution, a probabilistic model with Poisson-Gamma mixture distribution. In this approach, rates are treated as random variables with a prior gamma distribution, whose parameters are estimated using current recruitment data. This allows for a posterior distribution of rates to be used for recruitment prediction in an empirical Bayesian framework.

Building on this, a Poisson-Gamma model can be implemented by randomly generating recruitment rates using a Gamma distribution and plugging them into a Poisson process. This model can be used to make two types of projections: a point prediction estimating when the expected number of events will reach a specific sample size and a Bayesian interval prediction based on the generation of future accrual and event dates. This method enables flexible forecasting for any milestone at any calendar time (Bagiella and Heitjan, 2001).

2.5.1 Recruitment in one unit of time

Let $C|\Lambda \sim \text{Po}(\Lambda)$ and $\Lambda \sim \text{G}(\alpha, \beta)$. Where Λ represents the "recruitment proneness" in the unit of time.

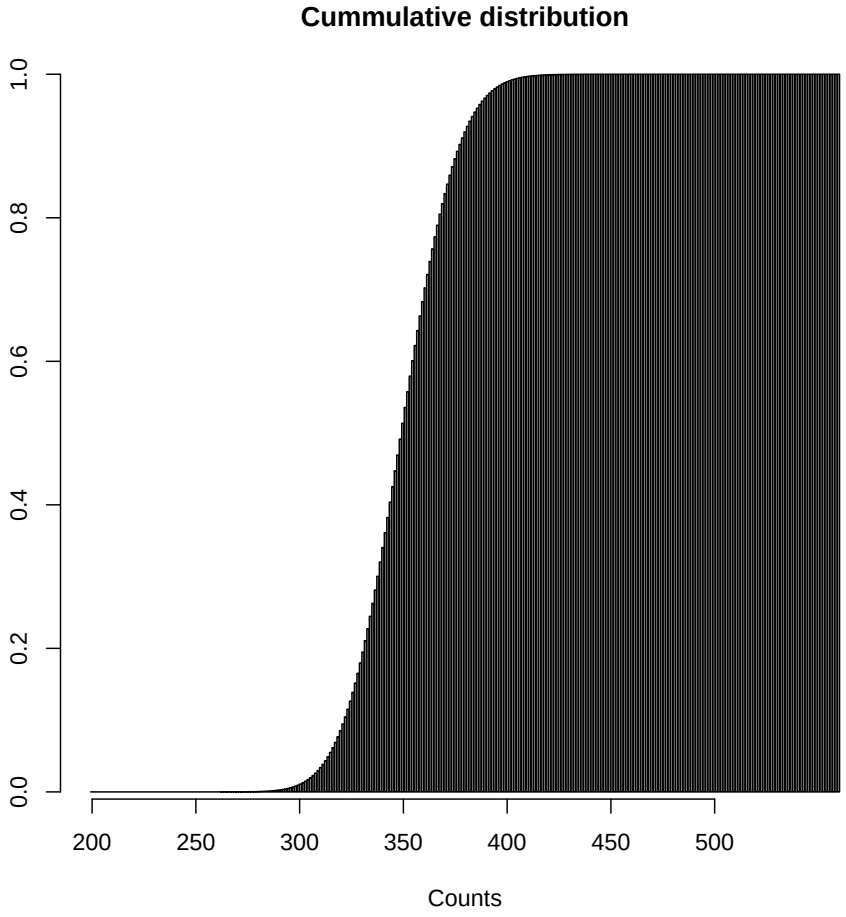

Figure 2.6: Cumulative Distribution Function (CDF) of Poisson-distributed counts: The bar plot illustrates the cumulative probability distribution for counts within the range of 200 to 500, using a Poisson $\mathrm{Po}(\lambda t)$ distribution with a rate parameter $\lambda = 0.591$ per day at time $t = 550$ days.

$$
\begin{aligned}
p(c) &= \int_0^\infty p(c|\lambda)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^c \exp(-\lambda)}{c!}\left[\lambda^{\alpha-1}\exp(-\beta\lambda)\frac{\beta^\alpha}{\Gamma(\alpha)}\right]d\lambda \\
&= \frac{\beta^\alpha}{c!\Gamma(\alpha)}\int_0^\infty \lambda^{\alpha+c-1}\exp(-\lambda)\exp(-\lambda\beta)d\lambda \\
&= \frac{\beta^\alpha\Gamma(\alpha+c)}{c!\Gamma(\alpha)(\beta+1)^{\alpha+c}}\underbrace{\int_0^\infty \frac{(\beta+1)^{\alpha+c}}{\Gamma(\alpha+c)}\lambda^{\alpha+c-1}\exp(-(\beta+1)\lambda)d\lambda}_{=1} \\
&= \beta^\alpha\binom{\alpha+c-1}{\alpha-1}\left(\frac{1}{\beta+1}\right)^{\alpha+c} \\
&= \binom{\alpha+c-1}{\alpha-1}\left(\frac{1}{\beta+1}\right)^{c}\left(\frac{\beta}{\beta+1}\right)^{\alpha} \\
c &= 0, 1, 2, 3, \ldots
\end{aligned}
$$

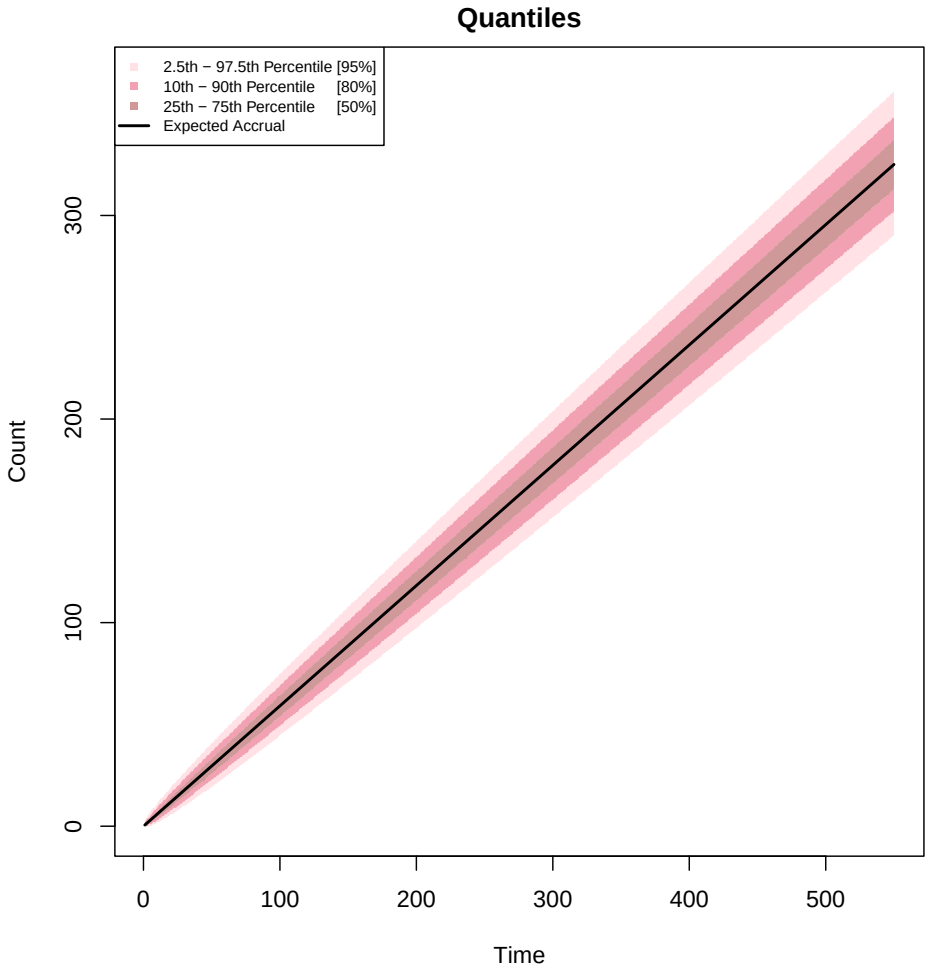

Figure 2.7: Predicted uncertainty bands for Poisson process with $\lambda = 0.591$ per day. The black line represents the expected accrual, while the red shaded regions indicate aleatory uncertainty: the dark red band spans the interquantile range (25th - 75th percentiles), the lighter red band cover the 10th - 90th percentile range and the light red the 2.5th - 97.5th percentile range (Spiegelhalter *et al.*, 2011).

Thus, $C \sim \text{NBin}\left(\alpha, \frac{\beta}{\beta+1}\right)$

The parameter λ in the integral represents the expected recruitment rate per unit of time for an individual and this recruitment rate is assumed to vary from individual to individual as it is generated by a $\text{G}(\alpha, \beta)$ distribution (Johnson *et al.*, 2005).

Using the expressions of iterated expectation and variance (Held and Bové, 2014) and the expectation and variance from the respective random variables $C|\Lambda \sim \text{Po}(\Lambda)$ and $\Lambda \sim \text{G}(\alpha, \beta)$, we have that:

$$\text{E}C = \text{E}_\Lambda[\text{E}_C(C|\Lambda)] = \text{E}_\Lambda[\Lambda] = \alpha/\beta$$

$$\begin{aligned}\text{Var}(C) &= \text{Var}_\Lambda[\text{E}_C(C|\Lambda)] + \text{E}_\Lambda[\text{Var}_C(C|\Lambda)] \\ &= \text{Var}_\Lambda[\Lambda] + \text{E}_\Lambda[\Lambda] \\ &= \alpha/\beta^2 + \alpha/\beta = \frac{\alpha(\beta+1)}{\beta^2}\end{aligned}$$

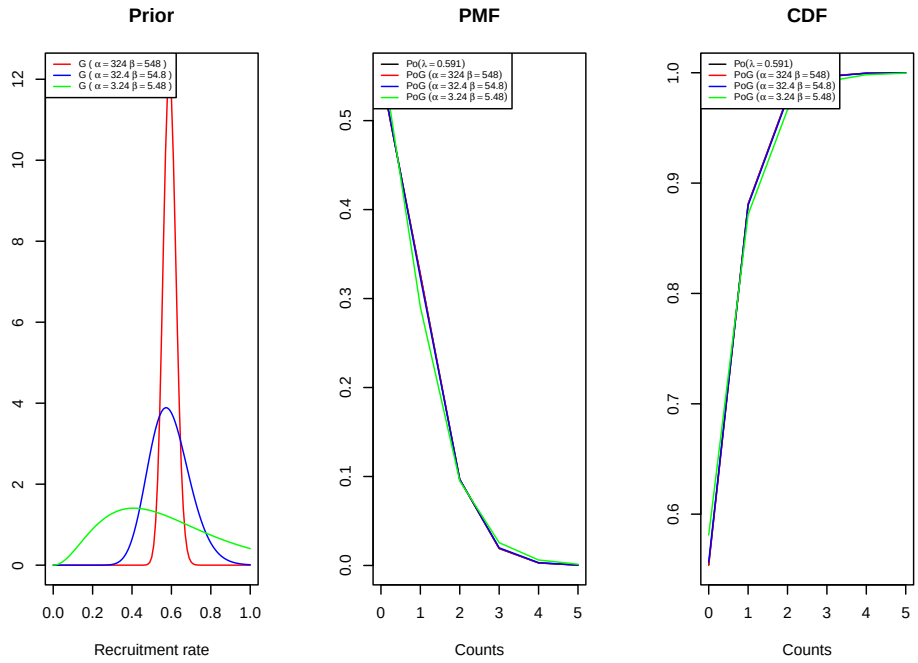

Figure 2.8: Sensitivity analysis between Poisson distribution with $\lambda = 0.591$ and Negative Binomial changing parameters of Gamma prior that maintain same expectation $\frac{\alpha}{\beta} = 0.591$ with $t = 1$.

There are two different interpretations of the Negative Binomial, Failure-Based and Count-based.

2.5.2 Failure-Based

1. The Negative Binomial $X \sim \text{NBin}(r, \pi)$ models the number of **failures** before achieving a fixed number of **successes** in a sequence of Bernoulli trials.

2. Parametrization:

   - n: Number of successes to be achieved (fixed).
   - p: Probability of success in each trial
   - The random variable X represents the number of failures before achieving n successes.

3. Probability Mass Function (PMF) (R Core Team, 2024):

$$\begin{aligned} \text{P}(X = x) &= \frac{\Gamma(x+n)}{\Gamma(x)x!} p^n (1-p)^x, \\ x &= 0, 1, 2, \ldots, n > 0 \\ 0 &< p \leq 1 \end{aligned}$$

   where x is the number of failures.

4. Interpretation: In a sequence of independent binary trials with constant probability p of observing a *non-recruited* patient, X is the number of *recruited* patients observed at the time that x *non-recruited* patients are observed (Meeker *et al.*, 2017).

With respect to the parameters, $n > 0$ represents the number of successes until the experiment is stopped. The success probability in each experiment is represented by $p \in [0, 1]$. In R the

functions `_nbinom(..., size = n, prob = p)` relate to the random variable $X-r$, the number of successes (as opposed to the number of trials) until r successes have been achieved (Held and Bové, 2014).

$$EX = \frac{r(1-\pi)}{\pi}$$
$$Var(X) = \frac{r(1-\pi)}{\pi^2}$$

Since we will be using the Count-Based interpretation of the Negative Binomial (Hilbe, 2011), our parametrization relates to R with $n = \alpha$ and $p = \frac{\beta}{\beta+1}$.

2.5.3 Count-Based

1. The Negative Binomial $X \sim \text{NBin}\left(\alpha, \frac{\beta}{\beta+1}\right)$ can also be seen as a Poisson-Gamma mixture, where the observed count data follows a Poisson distribution with a mean that itself follows a Gamma distribution, $C|\Lambda \sim \text{Po}(\Lambda)$ and $\Lambda \sim \text{G}(\alpha, \beta)$.

2. Parametrization:

   - $\mu = \frac{\alpha}{\beta}$: Mean of the distribution (expected number of occurrences).
   - α: Dispersion parameter, controlling the variance.

3. Alternative formulation of the PMF (Hilbe, 2011):

$$\text{P}(X = c) = \binom{\alpha + c - 1}{\alpha - 1}\left(\frac{\mu}{\beta+\mu}\right)^c\left(\frac{\alpha}{\alpha+\mu}\right)^\alpha,$$
$$c \geq 0$$

   where c is the counts.

4. Interpretation: This model is used to represent "recruitment proneness". The parameter μ represents the expected number of recruitments in a study (Johnson *et al.*, 2005).

In the count-based method we take into account the overdispersion of the data. When the variability in the observed data is greater than what is expected.

$$\text{E}X = \mu$$
$$\text{Var}X = \mu + \frac{\mu}{\beta}$$

How do we get to our formulation of the PMF:

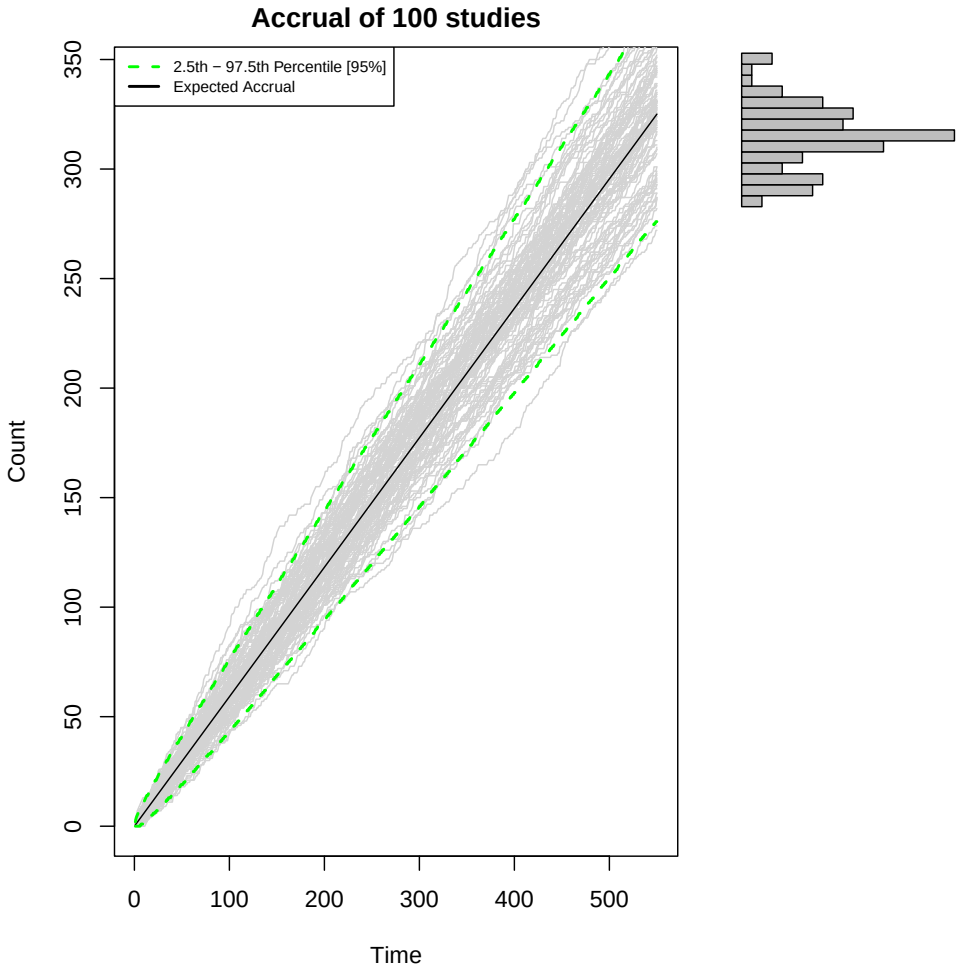

Figure 2.9: Poisson-Gamma ($\alpha = 324, \beta = 548$) distributed counts with $\mu = 0.591$ per day and uncertainty range. The black line represents the point estimate of the expected accrual from Section 2.3, while the red dashed lines indicate Poisson-Gamma 95% aleatory and epistemic uncertainty. The histogram illustrates the distribution of observed counts in 100 studies at time $t = 550$ days (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

$$
\begin{aligned}
\mathrm{P}(X = c) &= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\mu}{\alpha + \mu} \right)^{c} \left(\frac{\alpha}{\alpha + \mu} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\alpha/\beta}{\alpha + \alpha/\beta} \right)^{c} \left(\frac{\alpha}{\alpha + \alpha/\beta} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\alpha/\beta}{\alpha\beta/\beta + \alpha/\beta} \right)^{c} \left(\frac{\alpha}{\alpha\beta/\beta + \alpha/\beta} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{\alpha}{\alpha\beta + \alpha} \right)^{c} \left(\frac{\beta\alpha}{\alpha\beta + \alpha} \right)^{\alpha} \\
&= \binom{\alpha + c - 1}{\alpha - 1} \left(\frac{1}{\beta + 1} \right)^{c} \left(\frac{\beta}{\beta + 1} \right)^{\alpha}
\end{aligned}
$$

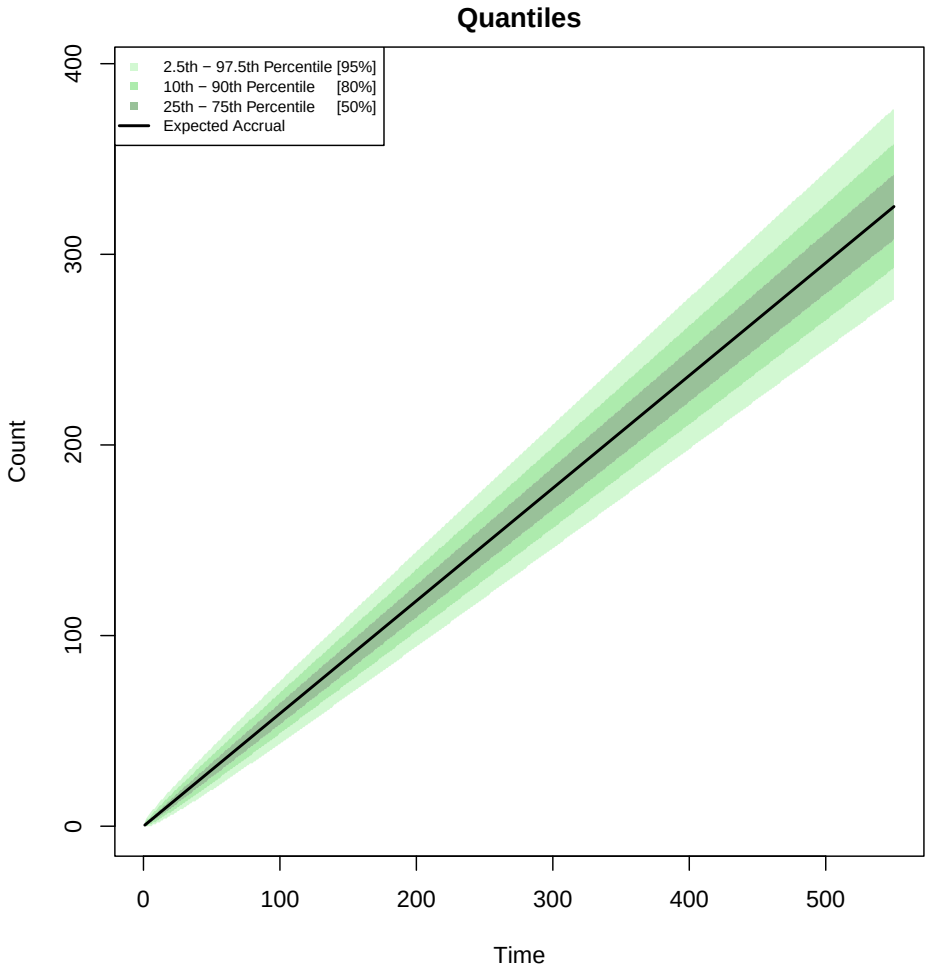

Figure 2.10: Predicted uncertainty bands for Poisson-Gamma process with $\mu = 0.591$ per day. The black line represents the expected accrual, while the green shaded regions indicate aleatory and epistemic uncertainty: the dark green band spans the interquantile range (25th - 75th percentiles), the lighter green band cover the 10th - 90th percentile range and the light green the 2.5th - 97.5th percentile range (Spiegelhalter *et al.*, 2011).

2.5.4 Sensitivity Analysis

In Figures 2.11 and 2.8, the Poisson distribution captures aleatory uncertainty, while the Gamma prior represents epistemic uncertainty. The Poisson-Gamma distribution incorporates both types of uncertainty. The sensitivity analysis, also shown in Figures 2.11 and 2.8, highlights that while the expectation remains unchanged, smaller parameter values lead to greater overall uncertainty due to the increased variance introduced by the Gamma distribution. The smaller the β, the greater the variance.

$$\text{Var}(\text{G}(\alpha, \beta)) = \frac{\alpha}{\beta^2} = \frac{\text{E}(\text{G}(\alpha, \beta))}{\beta}$$

In the Poisson-Gamma model we can interpret α as the parameter that represents the sample size and β represents time.

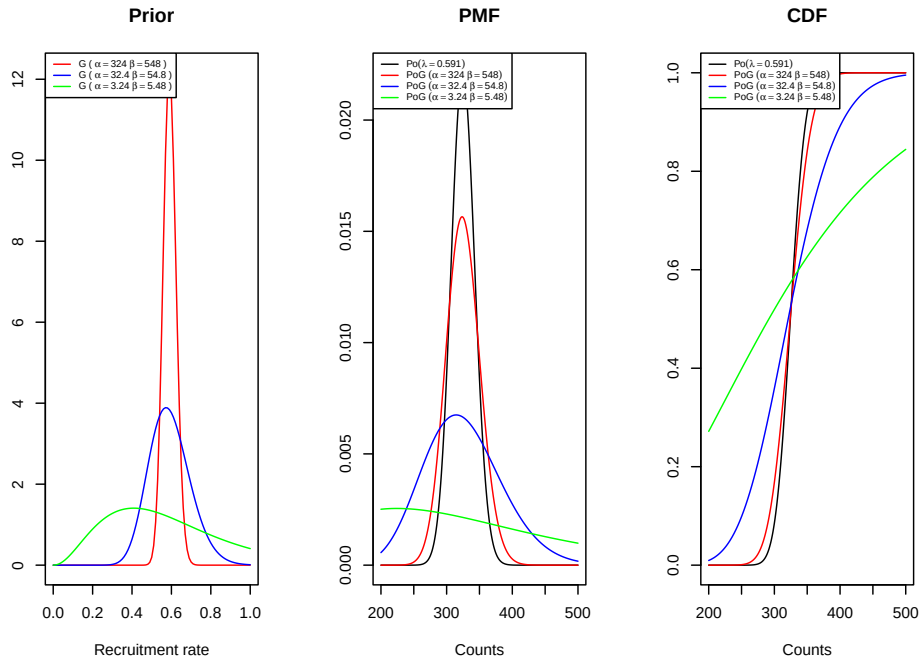

Figure 2.11: Sensitivity analysis between Poisson distribution with $\lambda = 0.591$ and Poisson-Gamma changing parameters of Gamma prior that maintain same expectation $\frac{\alpha}{\beta} = 0.591$ with $t = 550$.

2.5.5 Accrual at time point t

Let $C(t)|\Lambda \sim \text{Po}(\Lambda t)$ and $\Lambda \sim \text{G}(\alpha, \beta)$. By the infinitely divisible property of the Poisson distribution we have that $\underbrace{\text{Po}(\Lambda) + \cdots + \text{Po}(\Lambda)}_{t \text{ times}} = \text{Po}(\Lambda t)$.

$$
\begin{aligned}
p(c) &= \int_0^\infty p(c|\lambda t)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{(\lambda t)^c \exp(-\lambda t)}{c!}\left[(\lambda)^{\alpha-1}\exp(-\beta\lambda)\frac{\beta^\alpha}{\Gamma(\alpha)}\right]d\lambda \\
&= \frac{\beta^\alpha t^c}{c!\Gamma(\alpha)}\int_0^\infty \lambda^{\alpha+c-1}\exp(-\lambda t)\exp(-\lambda\beta)d\lambda \\
&= \frac{\beta^\alpha \Gamma(\alpha+c)t^c}{c!\Gamma(\alpha)(\beta+t)^{\alpha+c}}\underbrace{\int_0^\infty \frac{(\beta+t)^{\alpha+c}}{\Gamma(\alpha+c)}\lambda^{\alpha+c-1}\exp(-(\beta+t)\lambda)d\lambda}_{=1} \\
&= \beta^\alpha t^c \binom{\alpha+c-1}{\alpha-1}\left(\frac{1}{\beta+t}\right)^{\alpha+c} \\
&= \binom{\alpha+c-1}{\alpha-1}\left(\frac{t}{\beta+t}\right)^c\left(\frac{\beta}{\beta+t}\right)^\alpha \\
c &= 0, 1, 2, 3, \ldots
\end{aligned}
$$

Thus, $C(t) \sim \text{NBin}\left(\alpha, \frac{\beta}{\beta+t}\right)$

We will be focusing on the count-based interpretation of the Negative Binomial distribution. We can relate this interpretation to the more popular failure-based interpretation by seeing α

as n, the number of successes. In our case, the number of patients we wish to recruit. And, $p = \frac{\beta}{\beta+1}$ which is the probability of success, as the probability of recruiting.

Using the expressions of iterated expectation and variance (Held and Bové, 2014) and the expectation and variance from the respective random variables $C(t)|\Lambda \sim \text{Po}(\Lambda t)$ and $\Lambda \sim \text{G}(\alpha, \beta)$, we have that:

$$\text{E}(C(t)) = \text{E}_\Lambda[\text{E}_{C(t)}(C(t)|\Lambda)] = \text{E}_\Lambda[\Lambda t] = t\alpha/\beta$$

$$\begin{aligned}\text{Var}(C(t)) &= \text{Var}_\Lambda[\text{E}_{C(t)}(C(t)|\Lambda)] + \text{E}_\Lambda[\text{Var}_{C(t)}(C(t)|\Lambda)] \\ &= \text{Var}_\Lambda[\Lambda t] + \text{E}_\Lambda[\Lambda t] \\ &= t^2\alpha/\beta^2 + t\alpha/\beta = \frac{t\alpha(\beta+t)}{\beta^2}\end{aligned}$$

Therefore, we clearly have overdispersion $\text{Var}(C(t)) > \text{E}(C(t))$. This can be easily seen because:

$$\begin{aligned}\text{Var}(C(t)) &= \text{E}(C(t))\frac{\beta+t}{\beta} \\ &= \text{E}(C(t))\left(1 + \frac{t}{\beta}\right)\end{aligned}$$

2.6 Comparison of models for the accrual of counts

As we can see in Figures 2.12 and 2.14, we are taking into account more uncertainty in our model when we us the Poisson-Gamma model with a varying λ than we do in the Poisson process where λ is fixed. Hence, the color-coding, red for only aleatory and green, aleatory and epistemic. In reality we expect fluctuations of recruitment rates, therefore, the Poisson-Gamma model is more realistic.

Figures 2.12 and 2.14 assume that the recruitment rates do not vary much by assuming $\text{G}(\alpha = 324, \beta = 548)$ as prior. In contrast, Figures 2.13 and 2.15, assume a less informative prior $\text{G}(\alpha = 32.4, \beta = 54.8)$ showing a more pronounced discrepancy between the Poisson and the Poisson-Gamma process.

2.7 Generation of Poisson-Gamma model

There are two ways with which we can generate the λ in the $\text{PoG}(\alpha, \beta)$ model and we will prove in this Section how they are both equivalent.

2.7.1 Version 1

We generate a random vector of λ values of length $M = 10^5$ with $\text{G}(\alpha, \beta)$ and a vector of counts at each time point, of length t with $\text{Po}(\lambda)$. For the final counts we sum cumulatively (cumsum) the counts at t so that we get the accrual of patients at time t of each study.

$$\begin{aligned}\lambda^m &\sim \text{G}(\alpha, \beta) \\ C(t)^m &\sim \text{Po}(\lambda^m t) = \underbrace{\text{Po}(\lambda^m) + \cdots + \text{Po}(\lambda^m)}_{t \text{ times}} \\ m &= 1, \ldots, M\end{aligned}$$

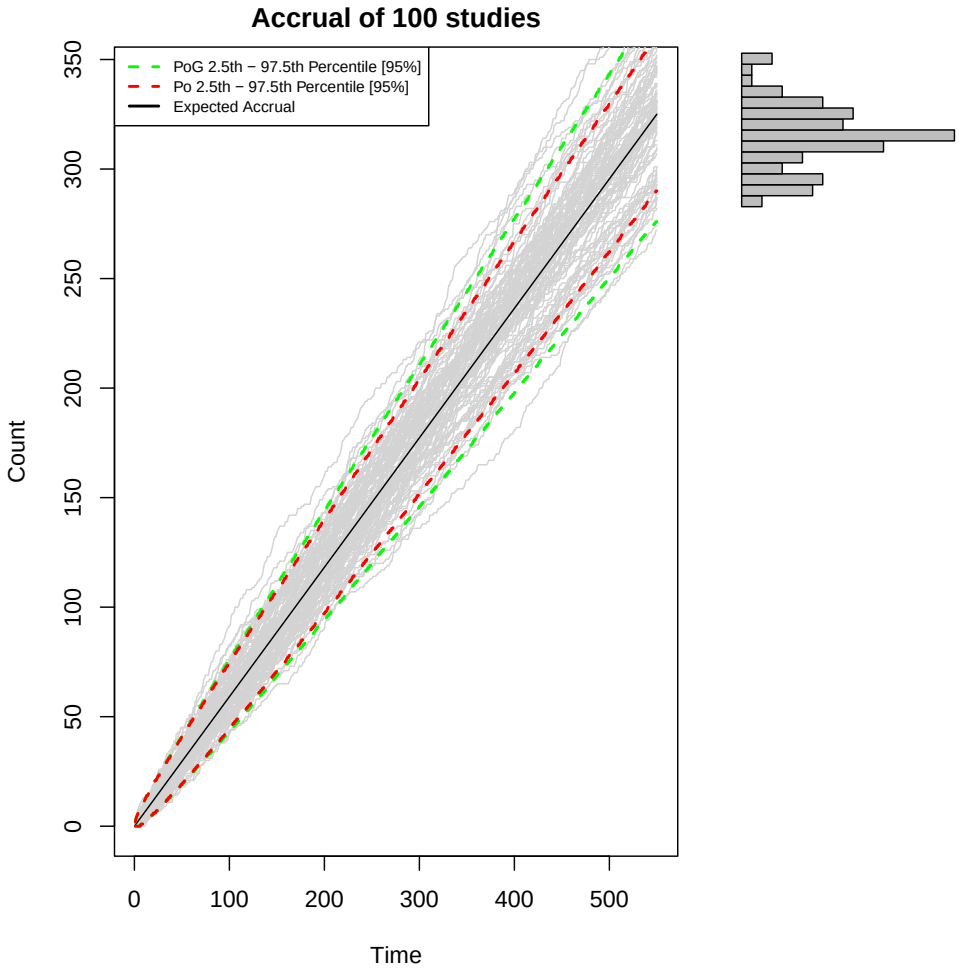

Figure 2.12: Comparison of the 95 % uncertainty range taken into consideration in the Poisson-Gamma ($\alpha = 324, \beta = 548$) model as opposed to the Poisson $\lambda = 0.591$ (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

2.7.2 Version 2

At each time point 1 to t, we generate a $\lambda^{(i)}$ with $\mathrm{G}(\alpha, \beta)$ and a count with $\mathrm{Po}(\lambda^{(i)}t)$. For the final counts we sum cumulatively (cumsum) the counts at t so that we get the accrual of patients at time t of each study.

$$\begin{aligned} \lambda^{(i)m} &\sim \mathrm{G}(\alpha, \beta) \\ C(i)^m &\sim \mathrm{Po}(\lambda^{(i)m}t) \\ i &= 1, \ldots, t \\ m &= 1, \ldots, M \\ C(t)^m &= \sum_{i=1}^{t} C(i)^m \end{aligned}$$

As we can see in Figure 2.16, the histogram of these two versions overlap and therefore we can conclude that both these ways of generating λ are equivalent.

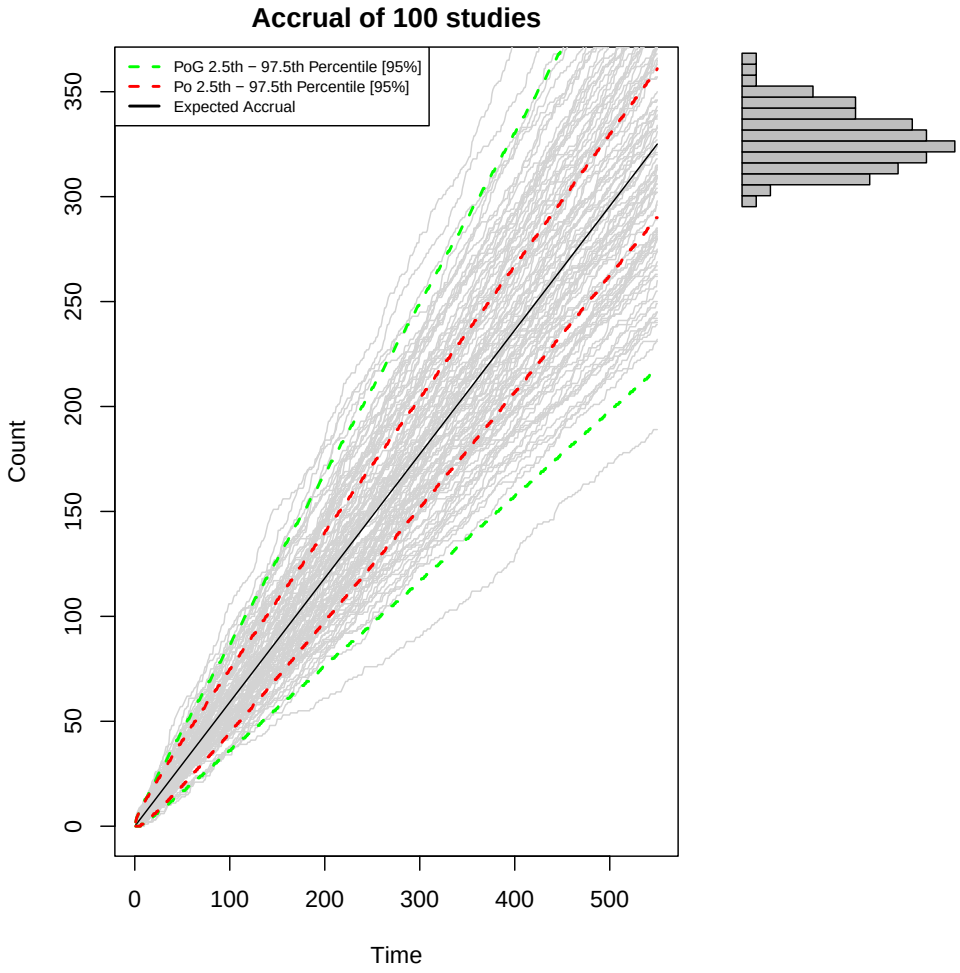

Figure 2.13: Comparison of the 95 % uncertainty range taken into consideration in the Poisson-Gamma ($\alpha = 32.4, \beta = 54.8$) model as opposed to the Poisson $\lambda = 0.591$ (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

2.8 Time: Model based on Expectations

If we fix the sample size of a study at c and we expect the recruitment rate to be λ, we deterministically predict the time planned for the study (without taking into consideration any uncertainty) to be $\hat{T} = \frac{c}{\lambda}$ (Bagiella and Heitjan, 2001).

Regarding the expectation and variance in this framework:

$T = ET = \mathrm{E}(c/\lambda) = c/\lambda$ and $\mathrm{Var}(T) = \mathrm{Var}(c/\lambda) = 0$

As we can see in Table 2.3.

2.9 Time: Model based on Erlang distribution

Let $T(c)$ denote the waiting time until c objects are recruited. For a fixed sample size c, assuming we have a fixed recruitment rate λ, $T(c) \sim \mathrm{G}(c, \lambda)$. Since c is an integer, we can use the additivity property from the Gamma distribution applicable to independent random variables, $\underbrace{\mathrm{Exp}(\lambda) + \cdots + \mathrm{Exp}(\lambda)}_{c \text{ times}} = \mathrm{G}(c, \lambda)$. Moreover, this distribution is also called the Erlang distribution and can be denoted as $\mathrm{Erlang}(c, \lambda)$. As we can see in Table 2.3, the expectation and variance are the following:

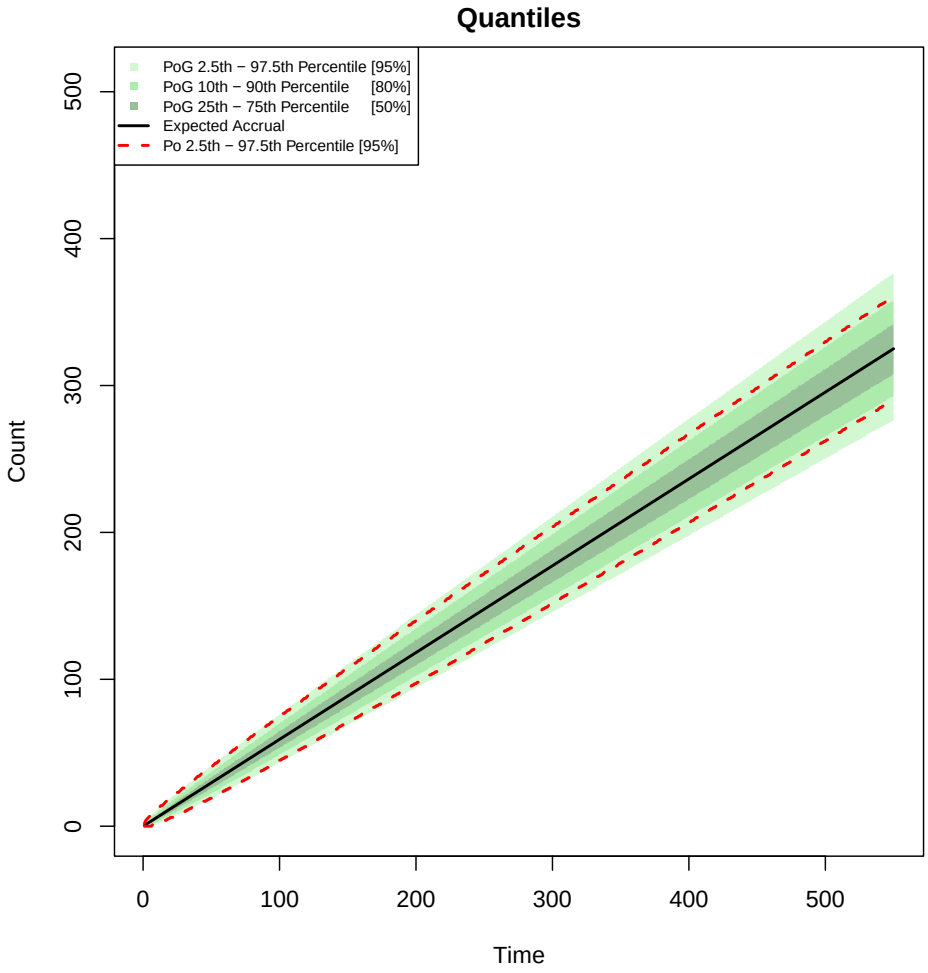

Figure 2.14: Comparison of the theoretical quantiles in the Poisson-Gamma ($\alpha = 324, \beta = 548$) model and the 95 % uncertainty range of the Poisson $\lambda = 0.591$ (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

$$\mathrm{E}T(c) = c/\lambda$$
$$\mathrm{Var}(T(c)) = c/\lambda^2$$

The Erlang model explains only the aleatory uncertainty of the waiting time until c objects have been recruited.

2.10 Time: Beta-Prime derived from Gamma-Gamma model

To take into account the epistemic uncertainty of recruitment rates and the aleatory uncertainty of waiting times, a Gamma-Gamma model is recommended (Bagiella and Heitjan, 2001), $T(c)|\Lambda \sim \mathrm{G}(c, \Lambda)$ and $\Lambda \sim \mathrm{G}(\alpha, \beta)$.

Using the expressions of iterated expectation and variance (Held and Bové, 2014), the expectation and variance from the respective random variables $T(c)|\Lambda \sim \mathrm{G}(c, \Lambda)$ and $\Lambda \sim \mathrm{G}(\alpha, \beta)$, and the fact that when $\Lambda \sim \mathrm{G}(\alpha, \beta)$ then, $\frac{1}{\Lambda} \sim \mathrm{IG}(\alpha, \beta)$ with:

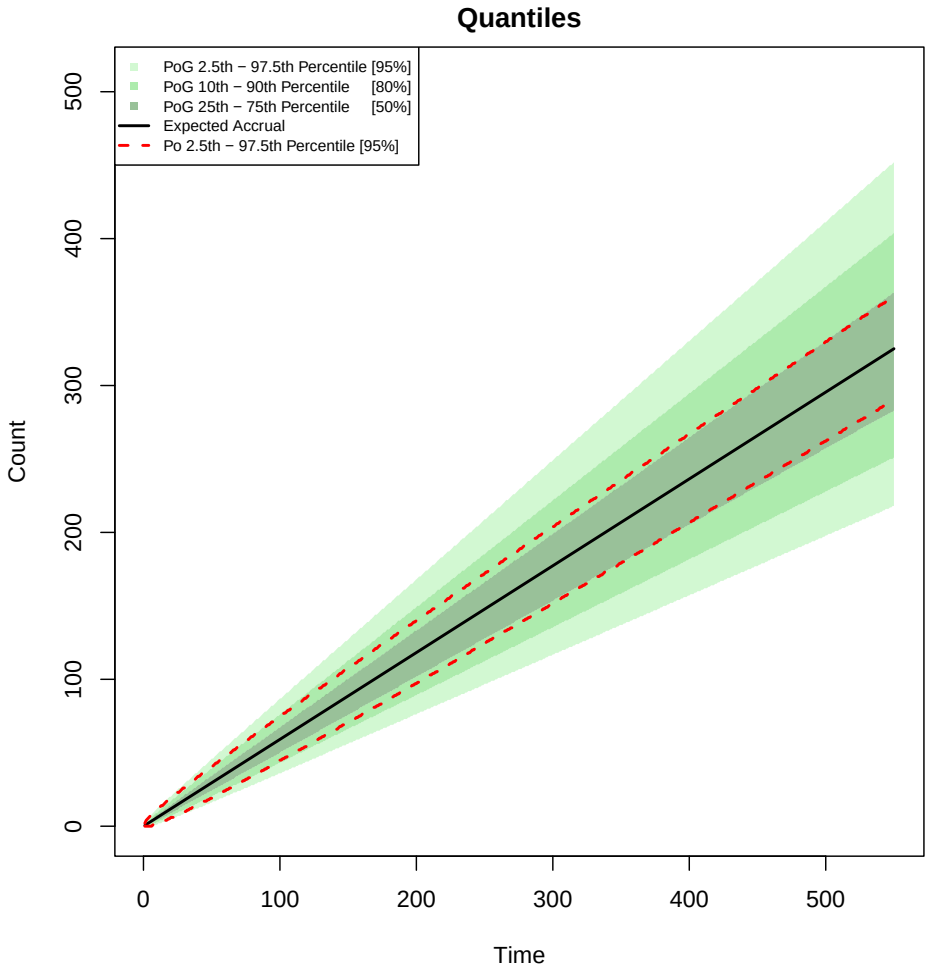

Figure 2.15: Comparison of the theoretical quantiles in the Poisson-Gamma ($\alpha = 32.4, \beta = 54.8$) model and the 95 % uncertainty range of the Poisson $\lambda = 0.591$ (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

$$\mathrm{E}\left[\frac{1}{\Lambda}\right] = \mathrm{E}(\mathrm{IG}(\alpha, \beta)) = \frac{\beta}{\alpha - 1}$$
$$\mathrm{Var}\left[\frac{1}{\Lambda}\right] = \mathrm{Var}(\mathrm{IG}(\alpha, \beta)) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)}$$

We have that:

$$\mathrm{E}T(c) = \mathrm{E}_\Lambda[\mathrm{E}_{T(c)}(T(c)|\Lambda)] = \mathrm{E}_\Lambda\left[\frac{c}{\Lambda}\right] = c\mathrm{E}_\Lambda\left[\frac{1}{\Lambda}\right] = c\frac{\beta}{\alpha - 1}$$

$\alpha > 1$

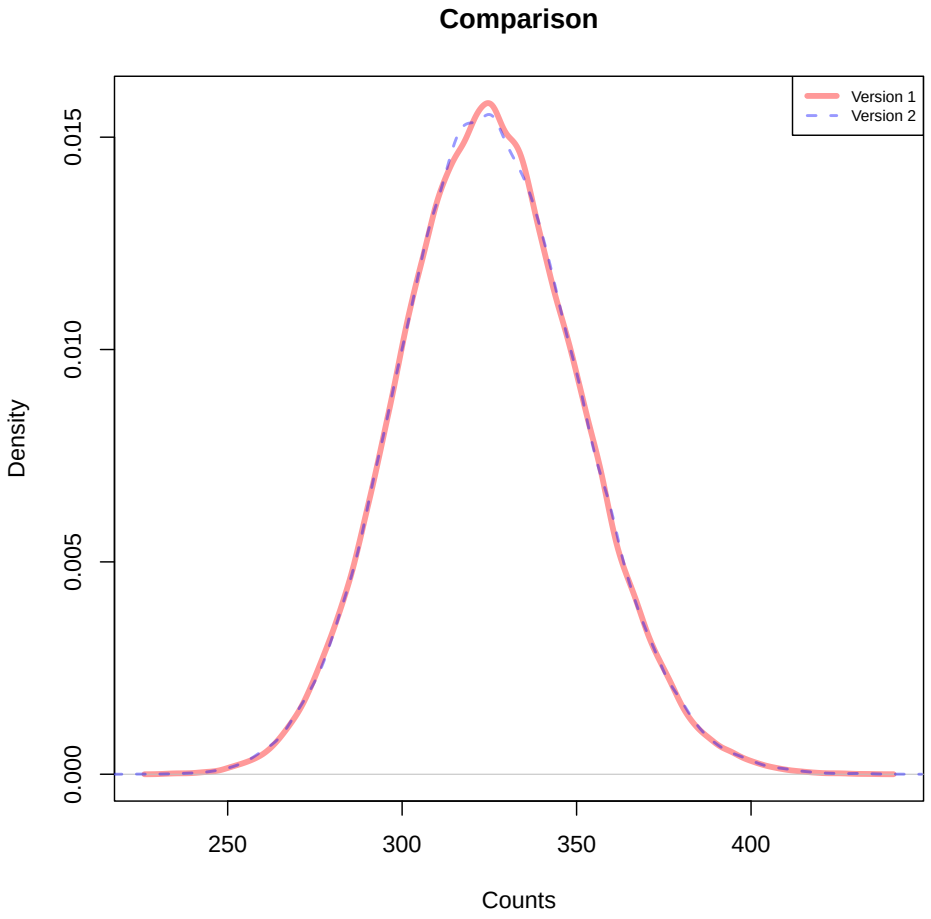

Figure 2.16: Comparison of the two versions with which we can generate a Poisson-Gamma model. Version 1: generates a vector of λ with length $M = 10^5$ using a $G(\alpha = 324, \beta = 548)$. Version 2: at each time point only one λ is generated. We compare both these versions at accrual of time $t = 550$ for $\lambda = 0.591$.

$$\begin{aligned}
\text{Var}(T(c)) &= \text{Var}_\Lambda[\text{E}_{T(c)}(T(c)|\Lambda)] + \text{E}_\Lambda[\text{Var}_{T(c)}(T(c)|\Lambda)] \\
&= \text{Var}_\Lambda\left[\frac{c}{\Lambda}\right] + \text{E}_\Lambda\left[\frac{c}{\Lambda^2}\right] \\
&= c^2\text{Var}_\Lambda\left[\frac{1}{\Lambda}\right] + c\text{E}_\Lambda\left[\frac{1}{\Lambda^2}\right] \\
&= \frac{c^2\beta^2}{(\alpha-1)^2(\alpha-2)} + \frac{c\beta^2}{(\alpha-1)(\alpha-2)} \\
&= \frac{c\beta^2(c+\alpha-1)}{(\alpha-1)^2(\alpha-2)} \\
&\alpha > 2
\end{aligned}$$

As we can see in Table 2.3. Here, we have again, a clear example of overdispersion because the variance is larger than the expectation:

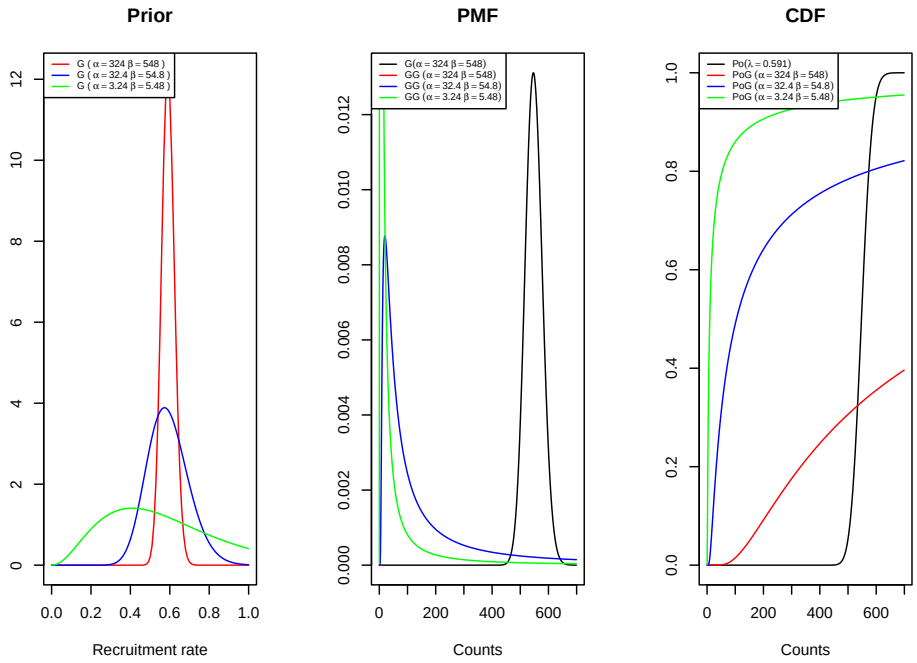

Figure 2.17: Sensitivity analysis between Poisson distribution with $\lambda = 0.591$ and Poisson-Gamma changing parameters of Gamma prior that maintain same expectation $\frac{\alpha}{\beta} = 0.591$ with $t = 550$.

$$\begin{aligned}\mathrm{Var}(T(c)) &= \mathrm{E}(T(c))\frac{\beta(c+\alpha-1)}{(\alpha-1)(\alpha-2)} \\ &= \mathrm{E}(T(c))\beta\left(\frac{c}{(\alpha-1)(\alpha-2)} + \frac{1}{\alpha-2}\right)\end{aligned}$$

For small parameter α ($\alpha > 2$) and large parameter β, we will have larger uncertainty (variance).

To compute $\mathrm{E}_\Lambda\left[\frac{1}{\Lambda^2}\right]$ we use property $\mathrm{Var}X = \mathrm{E}X^2 - (\mathrm{E}X)^2$, therefore, $\mathrm{E}X^2 = \mathrm{Var}X + (\mathrm{E}X)^2$. Thus,

$$\begin{aligned}\mathrm{E}_\Lambda\left[\frac{1}{\Lambda^2}\right] &= \mathrm{Var}\left[\frac{1}{\Lambda}\right] + \left[\mathrm{E}\frac{1}{\Lambda}\right]^2 \\ &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} + \frac{\beta^2}{(\alpha-1)^2} \\ &= \frac{\beta^2(1+\alpha-2)}{(\alpha-1)^2(\alpha-2)} \\ &= \frac{\beta^2}{(\alpha-1)(\alpha-2)}\end{aligned}$$

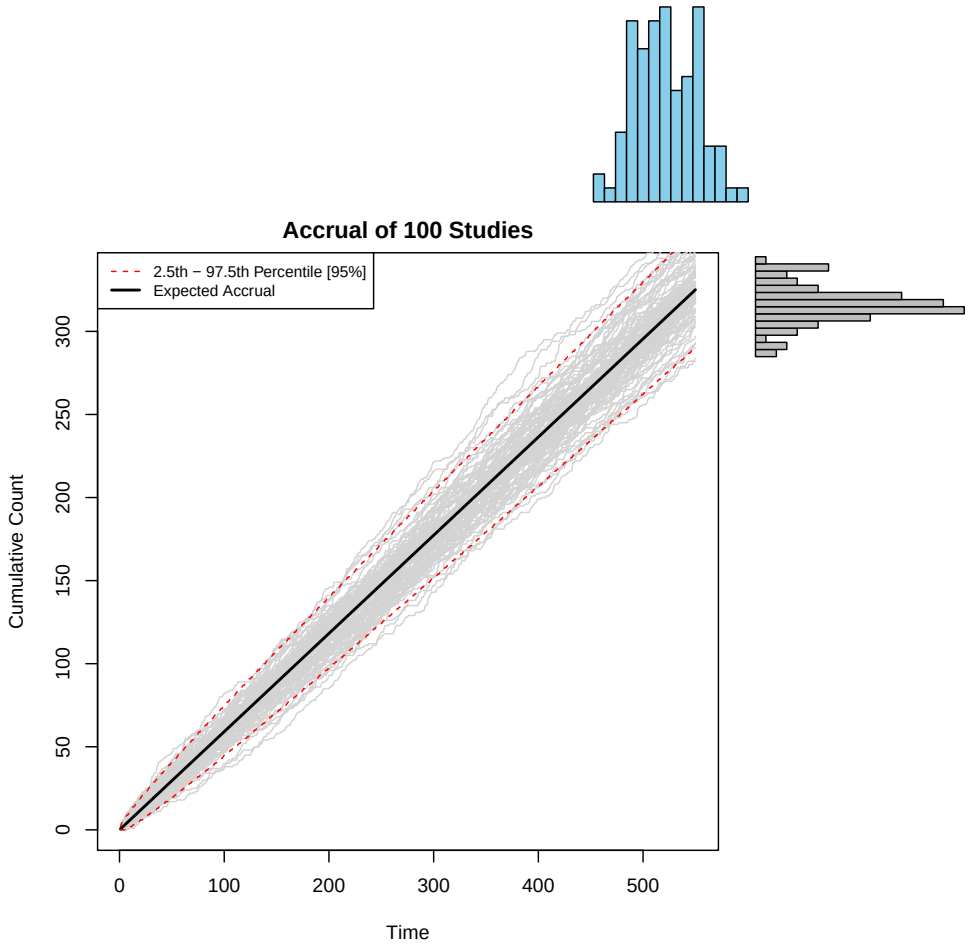

Figure 2.18: Poisson-distributed counts with $\lambda = 0.591$ per day and uncertainty range. The black line represents the point estimate of the expected accrual from section 2.3, while the red dashed lines indicate Poisson's 95% aleatory uncertainty. The histogram on the y-axis illustrates the distribution of observed counts in 100 studies at time $t = 550$ days and the histogram on the x-axis, is the distribution of time for $N = 324$ (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

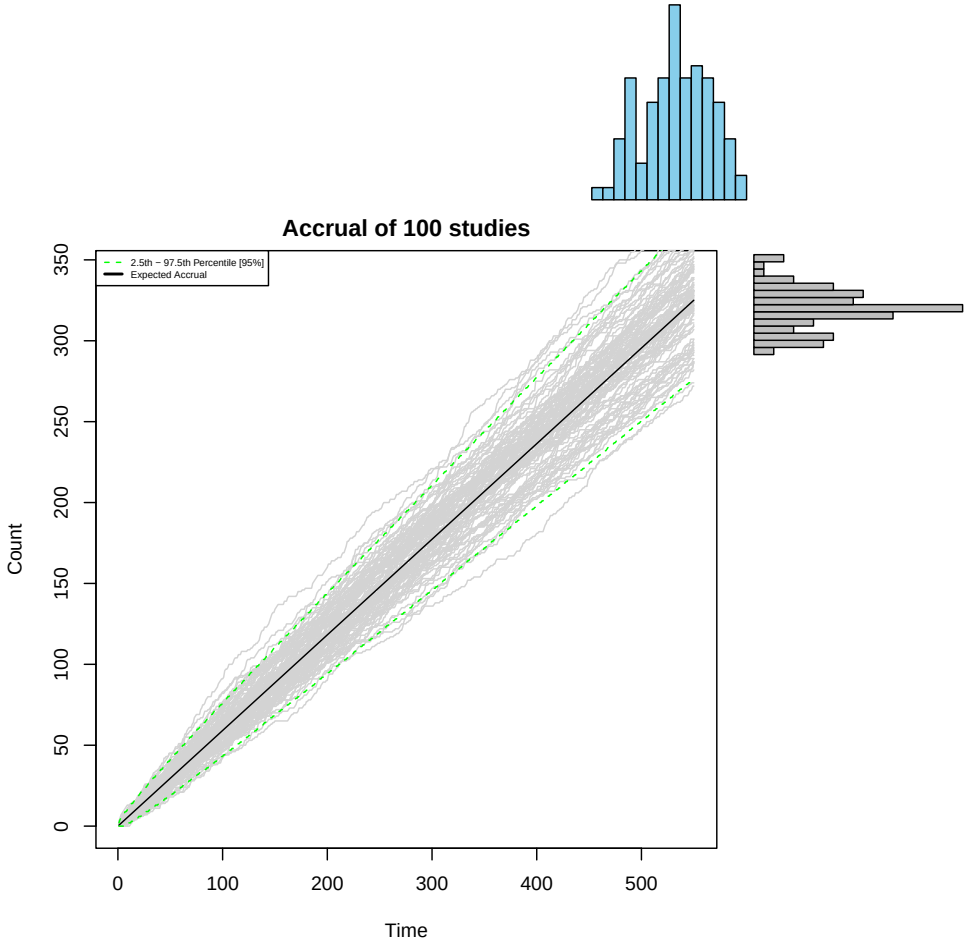

Figure 2.19: Poisson-Gamma ($\alpha = 324, \beta = 548$) distributed counts with $\mu = 0.591$ per day and uncertainty range. The black line represents the point estimate of the expected accrual from Section 2.3, while the red dashed lines indicate Poisson-Gamma 95% aleatory and epistemic uncertainty. The histogram on the y-axis illustrates the distribution of observed counts in 100 studies at time $t = 550$ days and the histogram on the x-axis, the distribution of time for $N = 324$ (Spiegelhalter *et al.*, 2011; Liu *et al.*, 2023).

Chapter 3

Results

3.1 Important questions when forecasting recruitment at the design-stage of a study

By normal approximation to the Poisson distribution $C(t) \sim \mathrm{N}(\mu = \lambda t, \sigma^2 = (\lambda t)^2)$, we know that the probability of recruiting the desired N participants is 0.5. Which means that the study has 50% chance of obtaining the desired sample size in the suggested T (Carter, 2004). We would also be assuming that the recruitment rate is constant over time.

In fact, we do not need normal approximation to see this. This can be shown with the Poisson distribution itself. We only need to specify the probability above λt, for example, 0.5. For large λ, 50% of the distribution will be below λt. With $\lambda < 1$ this is no longer the case.

This raises two questions which will be answered throughout this Master Thesis:

1. **Rate:** If T is fixed, what does the expected rate λ need to be to achieve a certain certainty of enrolling the total sample size N within the time frame T?

2. **Time:** Given a certain rate λ, how long should the recruitment period T be planned to give a confidence above 50% of recruiting the total sample size N? In Machine Learning, this confidence is aimed at 80%. In Carter (2004), at 90%.

3.2 Pros and cons of Monte Carlo's simulations

Carter suggests using Monte Carlo simulations, independent and identically distributed realizations of random variables. One clear advantage is its flexibility, as we can simulate any distribution we want. However, we must consider the following when we compute MC sampling instead of exact:

- M, the number of simulations
- Set a seed for computational reproducibility
- Monte Carlo standard errors (MCse) of estimates based on MC simulations
- Pseudo random numbers generated in R rely in the assumption that these pseudo random numbers are close to the true realizations of random variables (Held and Bové, 2014)

3.3 Counts: Comparison exact vs Monte Carlo simulations

Carter raises two important questions, enumerated in the previous section (Carter, 2004; Carter *et al.*, 2005). He suggests the use of Monte Carlo (MC) simulations for Counts. Here we investi-

gate the accuracy of his MC simulations by comparing them with exact distributions for accrual of counts introduced in Chapter 2.

In Figures **??** and **??**, we can see how for $n = 10^5$, MC sampling converges to the theoretical approaches discussed in Chapter 2.

Chapter 4

Discussion and Outlook

Chapter 5

Conclusions

Closed gaps: Graphical representation of study flow and recruitment at each stage of a clinical trial as well as leakage Unified derivations and mathematical notation of theoretical models for counts and time Properties of expectation and variance of the distributions for the models of time and counts Sensitivity analysis for models of counts and time COUNTS Extension of carter to exact distributions. We not only have aleatory but also epistemic uncertainty (currently only for counts) Not only poisson (aleatory) but poisson-gamma (aleatory and epistemic) Graphical representations of accrual process The simulation of accrual in 100 studies Exploration of two different approaches to the generation of fluctuating recruitment rates (version 1 and 2) Extension of code snippet from Carter to model counts but instead of having lambda fixed, it can vary using a poisson-gamma model

TIME Not only Erlang (aleatory) but gamma-gamma (aleatory and epistemic) Graphical representations for waiting time Extension of code snippet from Carter to model time but instead of having lambda fixed, it can vary using a Gamma-gamma model

Addressed questions provided by Carter of rate and time based on MC simulations and exact distributions for models on count and time

5.0.1 Personal Statement

NOT AI: This thesis was not written by any generative AI. It was written independently and without assistance from third parties. All sources utilized in this thesis are appropriately cited in the references

AI: During the preparation of this Master Thesis, I used [NAME OF TOOLS AND SERVICES] in order to [REASON]. After using this tool/service, I reviewed and edited the content as needed and I take full responsibility for the content of the Master Thesis.

Bibliography

Anisimov, V. V. and Fedorov, V. V. (2007). Modelling, prediction and adaptive adjustment of recruitment in multicentre trials. *Statistics in medicine*, **26**, 4958–4975. 8

Bagiella, E. and Heitjan, D. F. (2001). Predicting analysis times in randomized clinical trials. *Statistics in medicine*, **20**, 2055–2063. 9, 19, 20

Barnard, K. D., Dent, L., and Cook, A. (2010). A systematic review of models to predict recruitment to multicentre clinical trials. *BMC medical research methodology*, **10**, 1–8. 7

Bogin, V. (2022). Lasagna's law: A dish best served early. *Contemporary Clinical Trials Communications*, **26**, 100900. 4, 6

Carter, R. E. (2004). Application of stochastic processes to participant recruitment in clinical trials. *Controlled Clinical Trials*, **25**, 429–436. 5, 7, 8, 27

Carter, R. E., Sonne, S. C., and Brady, K. T. (2005). Practical considerations for estimating clinical trial accrual periods: application to a multi-center effectiveness study. *BMC medical research methodology*, **5**, 1–5. 7, 8, 27

Comfort, S. (2013). Improving clinical trial enrollment forecasts using sorm. *MJH Life Sciences*, **22**, . 6

Desai, R. (2014). *Preventing patient volume leakage in healthcare systems*. PhD thesis, University of Pittsburgh. 4

Food, Administration, D.,*et al.* (2018). Evaluating inclusion and exclusion criteria in clinical trials. In *Workshop Report. The National Press Club, Washington DC*. 3

Frank, G. (2004). Current challenges in clinical trial patient recruitment and enrollment. *SoCRA Source*, **2**, 30–38. 3

Held, L. and Bové, D. S. (2014). Applied statistical inference. *Springer, Berlin Heidelberg, doi*, **10**, 16. 7, 11, 13, 17, 20, 27

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press. 13

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. John Wiley & Sons. 11, 13

Lim, C.-Y. and In, J. (2019). Randomization in clinical studies. *Korean Journal of Anesthesiology*, **72**, 221–232. 3

Liu, J., Jiang, Y., Wu, C., Simon, S., Mayo, M. S., Raghavan, R., and Gajewski, B. J. (2023). *accrual: Bayesian Accrual Prediction*. R package version 1.4. 8, 14, 18, 19, 20, 21, 24, 25

Meeker, W. Q., Hahn, G. J., and Escobar, L. A. (2017). *Statistical intervals: a guide for practitioners and researchers*. John Wiley & Sons. 12

Mountain, R. and Sherlock, C. (2022). Recruitment prediction for multicenter clinical trials based on a hierarchical poisson–gamma model: Asymptotic analysis and improved intervals. *Biometrics*, **78**, 636–648. 8

National Institute of Allergy and Infectious Diseases (2021). Screening, enrollment, and unblinding of participants. https://www.niaid.nih.gov/sites/default/files/screening-enrollment-unblinding-of-participants.pdf. Accessed: 2025-02-10. 3

Nüesch, E., Trelle, S., Reichenbach, S., Rutjes, A. W., Bürgi, E., Scherer, M., Altman, D. G., and Jüni, P. (2009). The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. *BMJ*, **339**, . 3

O'Hagan, A. e. a. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons. 4

Panos, G. D. and Boeckler, F. M. (2023). Statistical analysis in clinical and experimental medical research: Simplified guidance for authors and reviewers. 3

Piantadosi, S. (2024). *Clinical trials: a methodologic perspective*. John Wiley & Sons. 4

Piantadosi, S. and Meinert, C. L. (2022). *Principles and Practice of Clinical Trials*. Springer Nature. 4, 6

R Core Team (2024). *The 'stats' Package: R Base Package for Statistical Functions*. R Foundation for Statistical Computing, Vienna, Austria. R package version X.Y.Z. 12

Rehman, A. M., Ferrand, R., Allen, E., Simms, V., McHugh, G., and Weiss, H. A. (2020). Exclusion of enrolled participants in randomised controlled trials: what to do with ineligible participants? *BMJ Open*, **10**, e039546. 3

Shih, W. J. (2002). Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine*, **3**, 1–7. 3

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science*, **333**, 1393–1400. 8, 11, 14, 15, 18, 19, 20, 21, 24, 25

Van Spall, H. G., Toren, A., Kiss, A., and Fowler, R. A. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: A systematic sampling review. *JAMA*, **297**, 1233–1240. 3

Whelan, J., Le Deley, M.-C., Dirksen, U., Le Teuff, G., Brennan, B., Gaspar, N., Hawkins, D. S., Amler, S., Bauer, S., Bielack, S.,*et al.* (2018). High-dose chemotherapy and blood autologous stem-cell rescue compared with standard chemotherapy in localized high-risk ewing sarcoma: Results of euro-ewing 99 and ewing-2008. *Journal of Clinical Oncology*, **36**, 3110–3119. 4, 5, 6

Willie, M. M. (2024). Population and target population in research methodology. *Golden Ratio of Social Science and Education*, **4**, 75–79. 3

Zhang, X. and Long, Q. (2012). Modeling and prediction of subject accrual and event times in clinical trials: a systematic review. *Clinical Trials*, **9**, 681–688. 7