Bakhshi, Andisheh (2011) *Modelling and predicting patient recruitment in multi-centre clinical trials.*

MSc(R) thesis

http://theses.gla.ac.uk/3301/

Modelling and Predicting Patient Recruitment in Multi-centre Clinical Trials

Andisheh Bakhshi

A Dissertation Submitted to the
University of Glasgow
For the degree of
Master of Science

College of Science and Engineering
School of Mathematics and Statistics
University of Glasgow

December 2011

Abstract

One of the main concerns in multi-centre clinical trials is how to enrol an adequate number of patients during a specific period of time. Accordingly, the sponsors are keen to minimise the recruitment time for cost effectiveness purposes.

This research tended to concentrate on forecasting the patients' accrual time for the pre-arranged number of sample size by simulating an on-going trial. The method was to model the data from the recruitment frequency domain and apply the estimations derived from the frequency domain to predict the time domain.

Whereas previous papers did not concentrate on variations of recruiting over centres, this research assumed that patient arrivals followed the Poisson process and let the parameter of the process vary as a Gamma distribution. Consequently, the Poisson-gamma mixed distribution was confirmed as the promising model of the frequency domain. Then with the help of the relationship between the Poisson process and the exponential distribution, accrual time was predicted assuming that the waiting time between patients followed the Gamma-exponential distribution.

As the result of the project, a trial was simulated based on the estimated values derived from completed trials. The first part of the prediction estimated the expected average number of patients per centre per month in an on-going trial. The second part, predicted the length of time (in months) to enrol specific number of patients in the simulated trial.

Contents

Abstract ii

List of tables vi

List of figures viii

Acknowledgement xi

Declaration xii

1. Literature and developing method

1.1. Introduction 1

1.2. Reviewing literatures 2

1.2.1. Summary of various papers 3

1.2.2. Overall conclusion based on reviewing papers 11

1.3. Aims 12

2. Recruitment data description

2.1. Software platforms 15

2.2. Frequency domain 15

2.3. Time domain 23

3. Modelling in recruitment data

3.1. Modelling the frequency domain 27

3.1.1. Distribution of the recruitment frequency data.....28

3.1.2. Parameter estimation.....30

3.1.3. Goodness of fit test (GoF).....38

3.1.4. Discussion.....45

3.2. Modelling in time domain.....46

3.2.1. Parameter estimation.....48

3.3. Time and Frequency parameter correlation test.....49

3.4. A proposed solution to the time analysis problem.....52

4. Modelling in parameters

4.1. Parameters estimated from recruitment frequency domain.....55

4.2. Parameter estimation from recruitment time domain.....58

4.3. Discussion.....62

5. The improved model and predictions

5.1. Issues in modelling and solutions.....64

5.2. Parameter estimation.....65

5.3. Correlation test for the transformed parameters in frequency domain.....66

5.4. Prior distribution of parameters.....69

5.5. Random effect meta-analysis.....73

5.6. Posterior estimations for the selected trial.....76

5.7. Prediction of patient recruitment.....80

5.7.1. Prediction in frequency domain.....81

5.7.2. Prediction in time domain.....88

Conclusion and discussion 100

Appendix1 103

Appendix2 109

References 112

List of tables

2.1 General information about the 18 completed clinical trial studies....16

3.1 Estimated parameters from fitting Negative Binomial to the frequency data....32

3.2 Estimated parameters from Method of Moments....36

3.3 Negative Binomial Goodness of Fit test to the recruitment data (Frequency domain)..38

3.4 Estimated parameters from fitting the Pearson VI distribution to the time domain data....48

3.5 SAS output of the correlation test between the ALPHA values in Frequency and Time domain....51

3.6 SAS output of the correlation test between the BETA values in Frequency and Time domain....52

4.1 Estimated values of parameters from Negative Binomial and within trial correlation coefficient....57

4.2 SAS output for the location test of the within trial correlation between parameters in frequency domain....57

4.3 SAS output of the between correlation coefficient of parameters among all the trials in the frequency domain....58

4.4 Estimated values of parameters from Pearson VI distribution and within trial correlation coefficient....61

4.5 SAS output for the location test of the within trial correlation between parameters in time domain 61

4.6 SAS output of the between correlation coefficient of parameters among all the trials in the time domain 62

5.1 Estimated transformed parameters from fitting Negative Binomial to the frequency data and including time-spans to the model 66

5.2 Table of the correlation-coefficient and covariance between estimated parameters *k* and *mu* from Negative Binomial with adding time-spans in the model 67

5.3 SAS output of the within trial correlation coefficient test between parameters k & mu 68

5.4 SAS output of the correlation test between parameters k & mu among the trials 68

5.5 SAS output of the normality test of the parameter k 70

5.6 SAS output of the normality test of the parameter mu 70

5.7 SAS output of the normality test of Ln (k) 71

5.8 Estimated parameters mu and Ln(k) with their within-trial standard errors and variances 72

5.9 SAS results of the random effects meta-analysis for parameter *k* 74

5.10 SAS results of the random effects meta-analysis for parameter Ln (k) 75

5.11 SAS results of the random effects meta-analysis for parameter mu 76

5.12 The four recruitment data set made from the selected trial in four time intervals monitoring 78

5.13 Summary of the estimated parameters and the posterior values for the selected trial in four time intervals 79

5.14 Summary table of the predicted accrual times based on the 1st, 2nd and 3rd month data 99

List of figures

2.1 Histograms of the frequency data with the fitted Negative Binomial curves and the Kernel density curves in studies....17-22

2.2 Probability histograms of patients' accrual times (months) in centres and estimated Kernel density curves....24-26

3.1 Maximum Likelihood contour plots for parameters in the studies....34-35

3.2 Comparing the Maximum Likelihood Estimation and Method of Moments in estimating shape parameter α by fitting NB distribution to the frequency data....37

3.3 Comparing the Maximum Likelihood Estimation and Method of Moments in estimating scale parameter β by fitting the Negative Binomial distribution to the frequency data....37

3.4 QQ-Plots and the Empirical cumulative density functions to test the suitability of the Negative Binomial to the frequency data....39-45

3.5 Scatter plot of the estimated ALPHA values in Frequency and time domain....50

3.6 Scatter plot of the estimated BETA values in Frequency and time domain 51

4.1 Scatter plot of estimated α and β from NB distribution in frequency domain....55

4.2 Scatter plot of α and β with 1*Standard Error.....56

4.3 Scatter plot of Alpha and beta in time domain....59

4.4 Scatter plot of ALPHA and BETA estimated from Pearson VI distribution with 1*Standard Error.....60

5.1 Line graph of the average number of recruited patients per centre per month in the completed trial througout the four monitoring intervals.....83

5.2 Average number of recruited patients per centre per month based on one month monitoring.....84

5.3 Average number of recruited patients per centre per month based on two months monitoring.....85

5.4 Average number of recruited patients per centre per month based on three months monitoring.....86

5.5 Average number of recruited patients per centre per month based on four months monitoring.....87

5.6 Line graph of the mean number of patients arrived to the centres per month in the real trial, marginal estimation, posterior estimation and the simulated trial.....88

5.7 Line graph of the recruiting trend througout the trial.....91

5.8 Pie chart of the share of each time interval in recruiting patients.....91

5.9 Histogram of the simulated accrual time for 16 patients based on one month follow up.....93

5.10 Histogram of the predicted accrual time of 36 patients based on the assumption that 16 patients had already arrived into the centres in the past month.....94

5.11 Histogram of the predicted accrual time of 80 patients based on the assumption that 16 patients had already arrived into the centres in the past month.....95

5.12 Histogram of the predicted accrual time of 136 patients based on the assumption that 16 patients had already arrived into the centres in the past month.....96

5.13 Histogram of the predicted accrual time of 44 patients based on the assumption that 52 patients had already arrived into the centres in the past two months 97

5.14 Histogram of the predicted accrual time of 100 patients based on the assumption that 52 patients had already arrived into the centres in the past two months 98

5.15 Histogram of the predicted accrual time of 56 patients based on the assumption that 96 patients had already arrived into the centres in the past three months 99

Acknowledgement

Many special thanks go to my supervisor Professor Stephen Senn for his help and support throughout the project.

I acknowledge ICON clinical Research Limited for sponsoring the project and providing data. I am also indebted to Dr Alan Philips for his invaluable comments and technical support.

I appreciate Prof Andrew Grieve’s help in finding an orthogonal parameterisation.

I like to express my thanks to Prof. Adrian Bowman, Daniel Molinari and Reza Anari for their help and advices in programming R and SAS.

I am also grateful to Dr Vladimir Anisimov for providing me with his papers and that I was able to benefit from his research results in my dissertation.

Help and support from my family in Iran is highly appreciated.

Last not least, I like to express my sincere gratitude to Dr Reza Katebi whose love, support and constant confidence helped me to complete this study. Also, I owe my 6-year-old son, Farhang, for being wonderfully understanding.

Declaration

I have prepared this thesis myself; no section of it has been submitted previously as part of any application for a degree. I carried out the work reported in it, except where otherwise stated.

Andisheh Bakhshi

Chapter 1
Literature and Approach

1.1. Introduction

A clinical trial is an experiment in which human subjects are exposed to medical treatment to understand the effects of the treatment for their well-being [9]. One of the main concerns in multi-centre clinical trials is to enrol sufficient numbers of patients during a specific period.

The numbers of patients recruited over time may reflect the intended study sample size but also the capabilities of participating centre. Some clinical trial delays are due to inadequate accrual of patients [11] but most are because of delay in recruitment.

Investigators need robust and reliable statistical tools to deal with the stochastic variations occurring in patient accrual rates. The recruitment rate may also vary depending on several factors such as the time of the year, the capacity of the centre [7], the number of staff [1], the popularity of the centre and the nature and the population of the local area. Consequently, with good planning tools, researchers would be able to reach a reasonable sample size rather than a sample size that could not be achieved during the trial period. Efficient recruitment planning tools can also reduce the number of centres required in clinical trials.

Recruitment time, however, is a fundamental aspect to the success of planning multi-centre clinical trials and appears to be even more important compared to the recruitment sample size. Due to it being costly, recruitment time is the biggest obstacle in, for example, drug development trials. On the one hand, in order to compete with other treatments and prevent loss of sales [1], the recruitment period should be shorter. If the recruitment time is too long, it can cost the sponsor revenue since a competing treatment may be released to the market first [9]. On the other hand, an inadequate number of patients may reduce the power of the trial [9]. Very often, in real trials, the number of patients that can be recruited is over estimated and the length of time to accomplish the trial is under estimated [8].

The primary objective of this thesis is to propose a method to predict patient recruitment across a fixed time length as well as modelling accrual time for pre-arranged sites and patient number.

1.2. Reviewing literatures

In reviewing the literature the aim was to identify publications dealing with the modelling of patient recruitment in order to see first, whether any suitable methods are available with easy implementation and second, if not, what further work would be needed.

In order to gain some overall knowledge of the requirements of analysing multi-centre clinical trials; the starting point was chapter 14 of ***Statistical Issues in Drug Development*** by Senn [1]. In the next step, to find the other currently existing statistical approaches in modelling patient enrolment, English language papers were searched for: ‘recruitment’, ‘clinical trial’, ‘predicting’, ‘recruitment time’, ‘patient recruitment’, ‘multicentre’, ‘multi-centre’, ‘accrual periods', 'recruitment period', 'modelling' and 'enrolment'. To find the source journals, I used the Google® search engine. Salience was determined by reviewing the abstract identified. Then, using the server of the University of Glasgow, I was able to download the salient full texts. In fact, after I registered for EndNote®, an alternative approach was possible and consequently, I had access to more papers through PubMed®. The references at the end of each paper were also employed to identify relevant papers. In addition, in response to an email sent to a leading authority in the field, Vladimir Anisimov (who is an honorary Professor in the School of Mathematics and Statistics at University of Glasgow and Senior Director in Research Statistics Unit at GlaxoSmithKline), I was kindly sent some of his papers as well as chapter 25 [4] of his book.

Among the numerous papers found, not all were of relevance for our research. I retained those papers focusing directly upon modelling patient recruitment, predicting the patient enrolment or predicting recruitment period for clinical trials.

1.2.1. Summary of various papers

The papers were summarised by grouping them chronologically by the first paper by the lead author. The most important publications which encompass the topic were those by Williford et al [6], Haidich et al [7], Carter et al [8,9], Abbas et al [10], Anisimov et al [2-5] and Gajewski et al [11]. This summary briefly describes the main statistical approaches proposed by the authors, followed by the application of the models suggested. Furthermore, an overall summary is provided at the end of this review. The reader, requiring a basic overview, may wish to skip to the following descriptions of individual papers.

Williford et al [6]

Williford et al [6] propose a negative binomial distribution as a mixture of Poisson distributions for modelling the recruitment data. They discuss that although the basic assumption, usually made when enrolling patients, is that the recruitment rate is constant; this is not always true in real trials. This is evidenced from several NIH (National Institutes of Health) funded multi-centre clinical trials [15-17]. Nevertheless, based on a constant recruitment rate, the number of patients enrolled in a trial is normally supposed to arise from a Poisson distribution. After testing the adequacy of the Poisson model, they suggest the parameter of the Poisson distribution, which is the average of recruitment rates, varies based on a gamma distribution. The outcome is a Negative Binomial.

This is an early paper using a Poisson-gamma mixture. This approach is discussed in more detail under Anisimov below.

Haidich et al [7]

In a particular database lunched by the AIDS Clinical Trial Group (ACTG) including a large set of over 700 trials with overall 120000 enrolments, Haidich et al [7] examined whether quarterly patients enrolment could be modelled to predict the number of new studies as well as the recruitment time and the effect of large trials. The relationship between the performance of large studies and enrolment, which accelerated over time, were significant in multivariate autoregressive modelling applied in SPSS. Observing the current trend and its variations was potentially beneficial for predicting future recruitment time.

They then showed that for a fixed sample size, the recruitment rate differs significantly between different months, seasons and centres as well as being influenced by launching large studies. The capacity of centres and the sample size play important roles in the number of patients arriving to participate in the trial. The effect of large studies, however, increases the overall patient intake, which is due to the huge demands on enrolling patients.

Carter et al [8, 9]

In a particular multi-centre effectiveness study, Carter et al [8] proposed three methods. In another paper [9], they propose the stochastic process to support theoretically the final method. In fact, with the help of the Poisson process and simulation, they provide a model to estimate the recruitment time. At the beginning of the trial, Carter seeks the basic historical information of each centre through questionnaires including the total number of subjects which have been entered in the centres [8].

One model considered an unconditional approach, which is, to divide the whole sample size by the total expected recruitment rate in individual centres. This method would be feasible only if all the centres initiate recruitment simultaneously. In reality, an unconditional approach cannot be applied if there are gaps between starting timesamongst participating centres [8]. To deal with this variation, Carter suggests a conditional approach depending on the recruitment rate in each centre and the length of the time each site recruits patients for.

In the two methods discussed above, since the number of participants in each time period and in each centre is assumed to be fixed, the variation of the estimated rate is not taken into account [9]. Carter has demonstrated in his third method to deal with the variation in the average number of patients over time at each centre that patients are assumed to enter the trial according to a Poisson process. Carter concludes that a Poisson distribution can be applied to model the recruiting probabilities according to the rate assumptions [9]. This approach is mostly useful when a clinical trial is designed to accrue the requisite number of patients in a finite time [8].

Overall, they recommend applying the conditional approaches for most multi-centre trials. However, in the case of uncertainty in the recruitment rate, a Poisson method is suggested [8].

Abbas et al [10]

Abbas et al indicate that the Markov model simulation has mostly been used in clinical trials for other purposes than patient recruitment such as evaluation of the benefits of treatments. However, in their paper, they use a Monte Carlo Markov simulation model for modelling recruitment patterns. In the interest of calculating the recruitment time, depending on the availability and accessibility of the data, Abbas et al [10] consider either continuous or discrete time variables. In order to increase the validity of the model, they regard the continuous time variables if the investigator knows that patients arrive one by one over time. In contrast, if the only available information is the average rate of arrivals in a particular time, the recruitment time needs to be considered as a discrete variable.

In the first simulation method (SM1)[10], with a constant pre-arranged probability *p*, not all the patients arriving to the centre are recruited. In fact, *p* plays the role of a filter to prevent patients passing from arrival phase to recruited phase. However, it is still assumed that the recruitment time is a continuous variable. There is also a random variable *R*, which is generated from a uniform distribution between 0 and 1, to be compared with *p* to decide if a patient is recruited (*R*<=*p*). As a result, the remaining time is highly related to *p*, such that the duration increases when the probability p decreases. Due to containing a probabilistic factor, the simulation needs to be repeated many times in order to calculate the variability of *t*. The simulation would stop when it reaches the target sample size. The next simulation method (SM2) considers the case of having no idea about what *p* might be. It, then, can be produced randomly from the uniform distribution. This is the only difference between SM1 and SM2. In this case, it is expected to reach the same average recruitment time as what we get with constant *p*=0.5! As a result, in a special case the two last methods could be merged in a particular case.

In the third simulation model (SM3), Abbas et al apply a discrete time variable, in which a group of patients arrive during a period of time *T*. The ultimate goal is to estimate the number of enrolled patients in each period of time. The patients who fail to be recruited (if *R*<*p*) are contacted again in the ensuing trial. The process continues throughout the time period, and the number of patients is recorded. Simulation model 3 fixes the length of recruitment time and estimates the mean number of patients at the end of the period. However, in SM4 (fourth simulation model), the objective is to calculate the expected delay of recruitment or remaining time. Such that the patients are recruited with a controlling fixed p but they are not replaced from the population.

Anisimov et al [2-5]

Since a robust method is required to plan for enrolling patients in multi-centre trials, a model is sought to deal with the uncertainties observed in practice. Vladimir V. Anisimov [2] and

Anisimov et al [3] propose a Poisson-Gamma approach for modelling patient recruitment. It is assumed that patients arrive in the centres independently [4, 5]. Therefore, the Poisson process is the most common and suitable distribution for the rate of patients recruiting in centres [1]. In a Poisson distribution, the average rate of patient arrival, which is the parameter of the model, is unchanged. However, due to some differences among centres such as size, number of staff, number of patients in the area and type of centre, the rate may vary among centres [2, 3]. Therefore, Anisimov et al [3, 5] let the rates vary, as samples coming from a Gamma distribution. That means that the simulation is performed in two steps, first a sample of rates of size N (N is the number of centres) is taken from a Gamma distribution. The patients are then enrolled according to a Poisson process with these sample rates [5]. This aspect of modelling leads to the use of an empirical Bayesian approach since the recruitment rate can be a random variable in the 'prior' distribution. The Gamma distribution is particularly convenient because it is a distribution of non-negative variables and the mean of a Poisson is, of course, non-negative. With this technique the number of centres with zero or few participants could be evaluated as well [2, 5]. The Poisson –Gamma recruitment model has been validated for large centres (>20) through GlaxoSmithKline (GSK) studies. Moreover, the techniques, to predict the additional number of centres to be added to complete the trial, are also suggested [3].

In order to estimate the shape parameter α of the recruitment model with a Gamma distributed rate, supposing that all the centres initiated at the same time, Anisimov [5] uses three estimation techniques: the Method of Maximum Likelihood, the Least Squares and the Method of Moments. Then he runs a Monte Carlo simulation method to compare the three estimation methods. All three methods, he concludes, are similar for large sample sizes and centres. The number of centres plays a more essential role in effective parameter estimation, though.

Regarding the analysis of recruitment time, Anisimov et al [4] have also considered three enrolment policies. One of which is competitive time, in which the enrolment stops whenever the total number of patients reaches the pre-arranged sample size. The second policy is the balanced time in which all the centres should reach a fixed number of patient arrivals. According to calculations, the recruitment time in this case can be longer that the competitive time. Opposed to the two former analyses, in the third approach, in which there are restrictions on the number of patients recruited by centres, Anisimov et al have not come to a closed solution. However, they conclude that it should be something between competitive times and balanced times [4]. Nevertheless, the real trial may not be flexible to incorporate into a particular model once the trial has started.

Gajewski et al [11]

In order to calculate the patient recruitment rate in a clinical trial, Byron J. Gajewski et al [11] propose a Bayesian approach. With the help of Bayesian posterior predictive distributions derived from prior knowledge, they provide a model to estimate the average waiting time between each patient. Hence, the overall recruiting time would be calculated for a fixed sample size. In addition, if the arrival process is to stop after a particular time, the model can predict the expected number of patients by the end of the trial.

Gajewski et al suppose that the waiting time is exponentially distributed with parameter θ (the parameterisation has not been mentioned in the paper!) and a conjugate prior distribution for

θ is the inverse gamma distribution. In a particular case, the prior distribution is supposed to be an Inverse Gamma distribution with parameters *nP* and *TP*, in which *n* is the prior sample size, *T* is the waiting time and *P* is the weighting factor for accuracy of *n* and *T* according to the historical information. In an example, Gajewski et al [11] compare three methods of predicting the waiting time and the total length of the trial. The first is only to consider the information at the beginning of the study. Hence, the average waiting time is distributed as an *Inverse Gamma (nP,TP)*. The second approach relies solely on observed data from the ongoing trial. The third suggestion, however, is to take advantage of both the prior information and the observations from the ongoing trial. As a result, this third approach culminates in an Inverse Gamma distribution in which the shape parameter is *nP* plus the number of observed patients in the current trial and the scale parameter is *TP* plus the time period, in which they have experienced the real recruitment. The authors conclude that the posterior with informative prior estimation (third approach) shows the faster rate of prediction compared to applying only the information at the outset of the study or using the observed data only.

It should be noted that if the number of events in a given interval follows the Poisson distribution, then inter-arrival times lead to an exponential distribution. Thus, the basic model in this approach can be regarded as being the time domain equivalent of the frequency domain model considered by Anisimov.

1.2.2. Overall conclusions based on reviewing papers

Based on reviewing the papers, a number of issues can be identified as follows:

1.2.2.1. General appropriate approach to modelling patient recruitment

Almost none of the approaches discussed above focused deeply on variations in recruiting among centres. They did not concentrate on the variations in recruiting over time either. Also, it is obvious that in real trials the rates of patient intake cannot be fixed. As a result, a tentative conclusion is that the Poisson-Gamma model is a promising approach. The Poisson-Gamma model lets the rates vary as random variables from Gamma distributed samples. The patient arrivals are then assumed to follow the Poisson distribution with the average rate being derived from Gamma distribution. The model originally was proposed by Williford et al in the context of clinical trials [6] but has been extensively developed by Anisimov [2-5].

1.2.2.2. Variations and risk factors that should be considered to optimise the model

Anisimov emphasised that recruitment rates are not fixed in real trials. Haidich [7] concluded the various factors such as time of the year and the influences of large sample size are significant. Gaps between centres such as the differences in capacity and the popularity of the centres and the numbers of staff, which Senn [1] highlights, should be taken into account.

1.2.2.3. Clarity and feasibility of the applying model

In his discussion, Abbas [10] does not consider a statistical model for the waiting times between patients. Hence, it is not clear which statistical distribution the time follows for the suggested simulation models.

In all his approaches, Carter [8, 9] assumes the average patients' arrival rates are fixed although he considers the variation regarding the initiation time. The conditional approach has not been sufficiently clarified enough to be practically applicable for large clinical trials.

Haidich's time series method is primarily designed to analyse the risk factors of the recruitment period focusing upon the reasons for delays rather than on a mature model for recruiting patients.

Gajewski [11], in contrast, proposes a Bayesian approach for the waiting time, which is more practical and easier to apply. However, Anisimov [2-5] expands the statistical model with the interest of feasibility and supporting the idea theoretically as well. The Poisson-Gamma distribution is more reliable to apply; but it is complicated compared to Gajewski's approach.

1.3. Aims

From reviewing the literature, possible modelling approaches to predict the patient recruitment prior to the start of the study were identified. The next step for the on-going trial, however, would be to predict the patient recruitment and accrual times with respect to the frequency domain[1] and the time domain[2] respectively. Therefore, it is proposed that to fulfil these aims for the present project the future approaches are adopted:

- ✓ Initially, the plan is to apply the theory of Vladimir Anisimov to model the frequency data as well as the time data. But in the case of any practical issue in modelling the recruitment data the aim is to apply an alternative model for the recruitment sample data.
- ✓ The models in the literature will be applied to identify a suitable model for the frequency and time domain.
- ✓ Test for the suitability of the fitted distributions.
- ✓ Use the alternative methods in case of unsuitable fitted models.
- ✓ Estimate the parameters based on Maximum Likelihood Estimation.
- ✓ Check the possible correlations between parameters in case they are dependent in order to make the modelling process easier.
- ✓ Model the estimated parameters to gain the prior distribution of parameters
- ✓ Apply Bayesian forecasting to predict patient recruitment founded upon predictive distribution and simulation

[1] Frequency domain is the number of patients that are recruited in centres in a clinical trial.

[2] Time domain is the length of time (in the current research project the unit of time was set to months) to recruit patients in clinical centres.

Chapter 2
Recruitment Data Description

The first step in modelling the patient recruitment in multi-centre clinical trials is to identify and collect some possible prior information from previous trials. The aim is to employ previous data and observations in clinical trials to forecast the required patient arrivals for a given trial even though the patients are not recruited yet. The forecast may not be exact but it will produce enough information to set a plan based on these historical data and experiences; otherwise, the trial experiment would be very intuitive and inaccurate. For this reason, ICON Clinical Research, the sponsor of this MSc project, provided data from completed studies in Excel spread sheet format. Each trial dataset consists of site codes and country information, number of patients' arrivals in clinical centres; and start and finish recruitment dates in centres.

Table 2.1 illustrates the general feature of the historical data provided by ICON. It includes 18 completed trials (column 1) with the number of clinical centres in each trial (column 2). Next (column 3) is the total number of recruited patients in the studies accompanied by the minimum (column 4) and maximum (column 5) number of patients in the centres in each trial.

The last two columns summarise the minimum and the maximum length of time (in months) that have taken to recruit patients in the trials.

2.1. Software platforms

All the analyses were performed in Statistical Analysis Software (SAS®) version 9.2 and package R 2.12.1. R is freely downloadable from http://www.r-project.org/, but SAS is licensed and was only accessible from the university computers. Microsoft Excel was also used for making new data sheets and some graphical applications.

2.2. Frequency domain

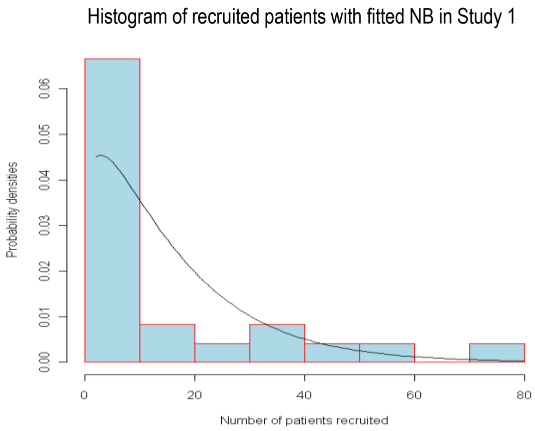
Although the studies vary a lot regarding the number of clinical centres and patients (table 2.1), the general features of the Kernel density curves [21] in the frequency domain follow a similar statistical distribution (figures 2.1). A Kernel density curve is a graphical and non-parametric method of estimating the probability density functions. Kernel density curves and histograms are closely related. In a histogram, the horizontal axis is divided into bins, which should cover the whole range of the data. For a kernel density, each point is allocated to a normal kernel density. The individual kernels are added up to make the kernel estimate. The advantage of the kernel density is its smoothness compared to the histogram.

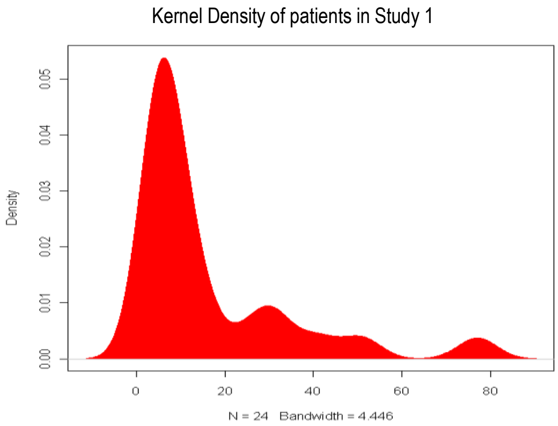
R produces the kernel density curves in the MASS package (Appendix 1). In SAS, however, 'kernel' option in the UNIVARIATE procedure superimposes the kernel density curve on the histogram. Many different kernel densities are possible but just the default one has been applied in this project.

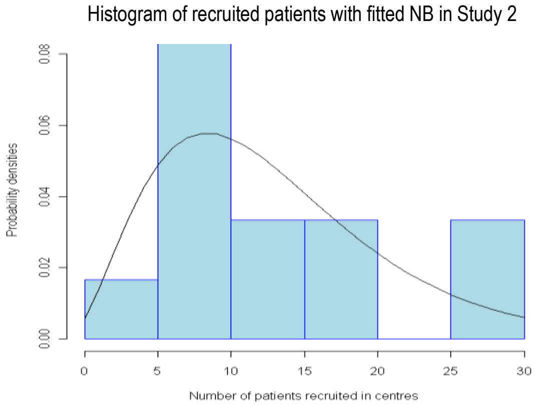
Study code	N (number of sites)	n (total number of patients)	Min number of patients in clinical centers	Maximum number of patients in clinical centers	Minimum recruitment time	Maximum recruitment time
1	24	385	2	77	7.30	14.70
2	12	152	2	28	0.40	3.13
3	24	811	4	76	0.97	18.23
4	16	80	1	17	0.97	3.00
5	25	244	2	26	0.97	9.76
6	66	796	1	48	16.23	34.80
7	110	1126	0	41	0.33	7.27
8	141	1241	0	40	0.57	14.80
9	51	927	2	67	7.30	14.70
10	150	2000	0	63	4.67	16.30
11	97	2936	0	115	3.50	7.90
12	75	546	0	20	3.27	6.47
13	270	4363	0	263	2.97	28.30
14	410	3274	0	59	3.60	35.73
15	30	103	0	9	5.23	21.80
16	92	533	0	60	1.80	12.57
17	26	549	1	76	2.60	10.03
18	60	1696	3	122	6.47	21.00

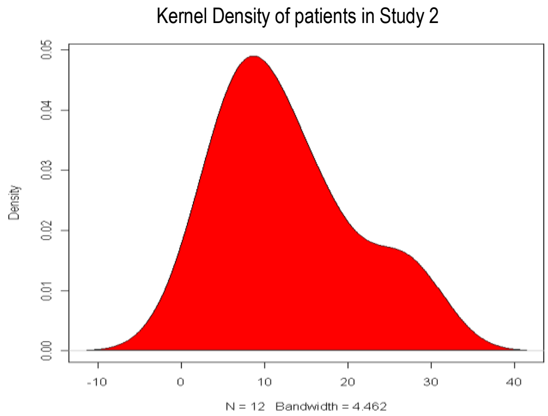
Table 2.1: General information about the 18 completed clinical trial studies

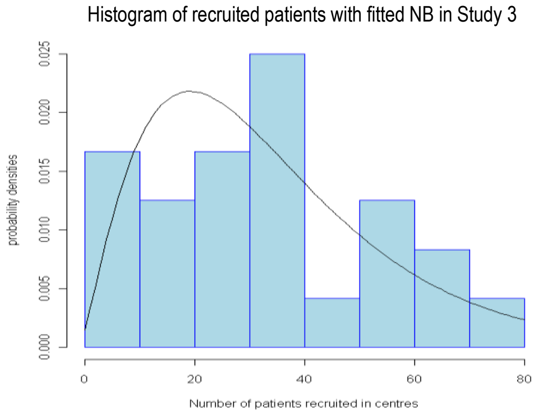
In figure 2.1, histograms of the frequency data of the 18 studies are illustrated as well, which are accompanied by the fitted Negative Binomial distributions (solid lines on the histograms).

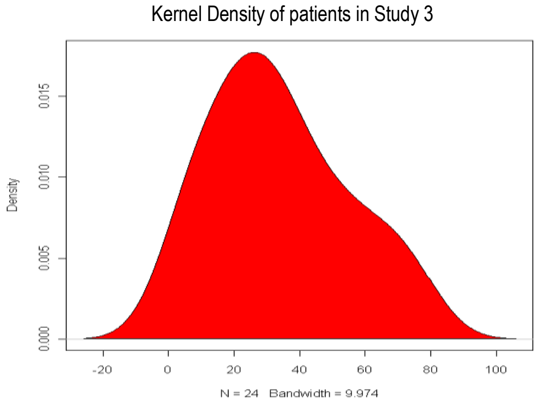

Histogram of recruited patients with fitted NB in Study 1
Probability densities
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
20
40
60
80
Number of patients recruited


Kernel Density of patients in Study 1
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
N = 24 Bandwidth = 4.446

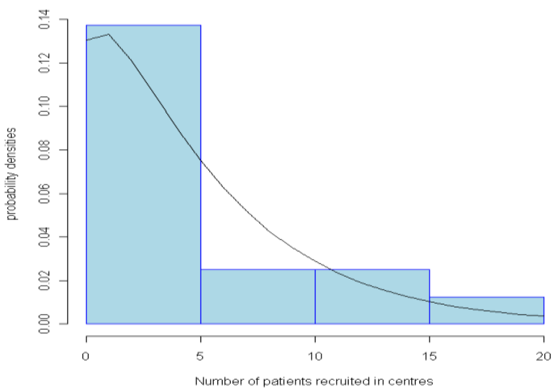

Histogram of recruited patients with fitted NB in Study 2
Probability densities
0.00
0.02
0.04
0.06
0.08
0
5
10
15
20
25
30
Number of patients recruited in centres


Kernel Density of patients in Study 2
Density
0.00
0.01
0.02
0.03
0.04
0.05
-10
0
10
20
30
40
N = 12 Bandwidth = 4.462

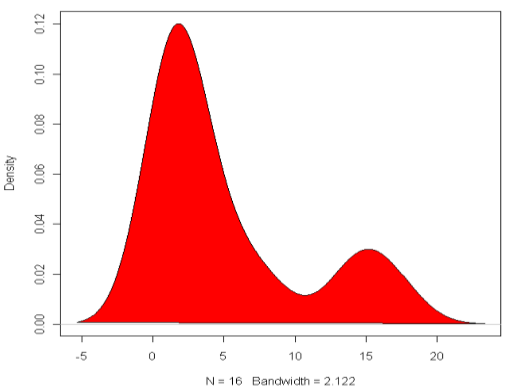

Histogram of recruited patients with fitted NB in Study 3
probability densities
0.000
0.005
0.010
0.015
0.020
0.025
0
20
40
60
80
Number of patients recruited in centres


Kernel Density of patients in Study 3
Density
0.000
0.005
0.010
0.015
-20
0
20
40
60
80
100
N = 24 Bandwidth = 9.974

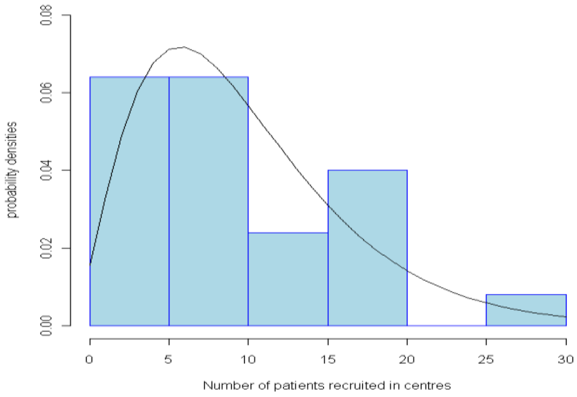
Histogram of recruited patients with fitted NB in Study 4



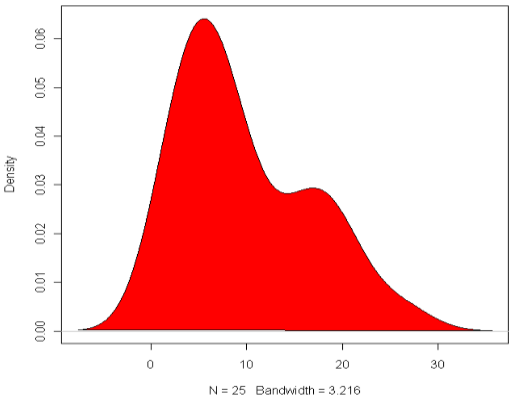
Kernel Density of patients in Study 4



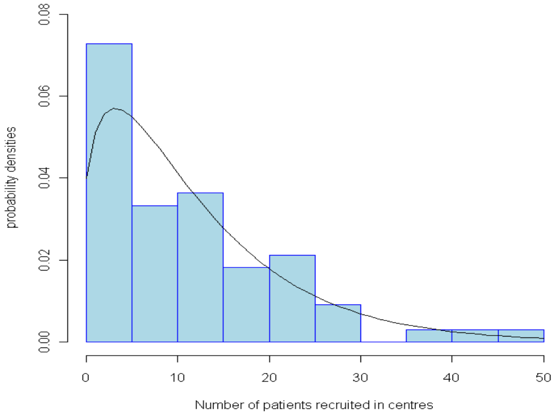
Histogram of recruited patients with fitted NB in Study 5



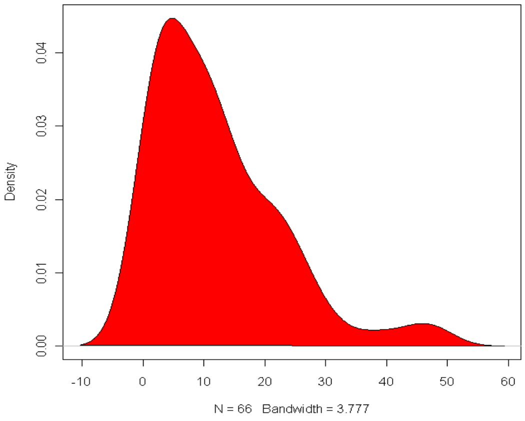
Kernel Density of patients in Study 5



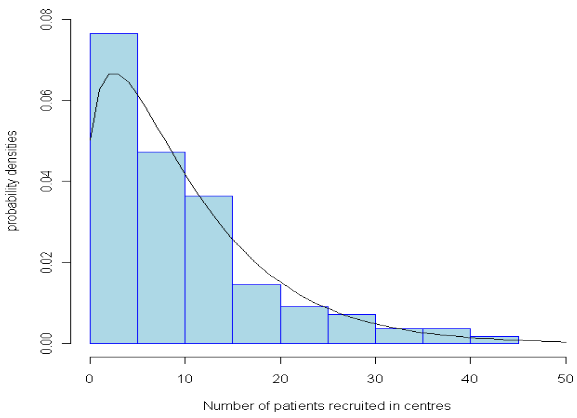
Histogram of recruited patients with fitted NB in Study 6



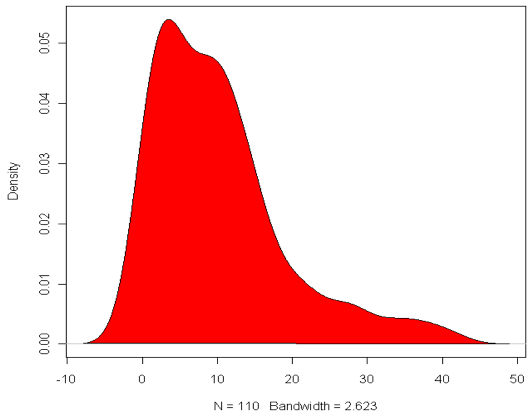
Kernel Density of patients in Study 6



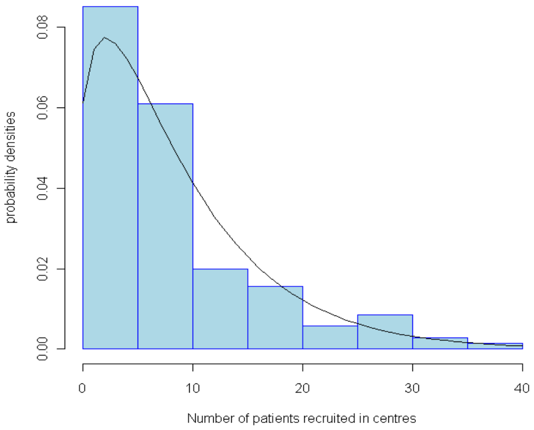
Histogram of recruited patients with fitted NB in Study 7



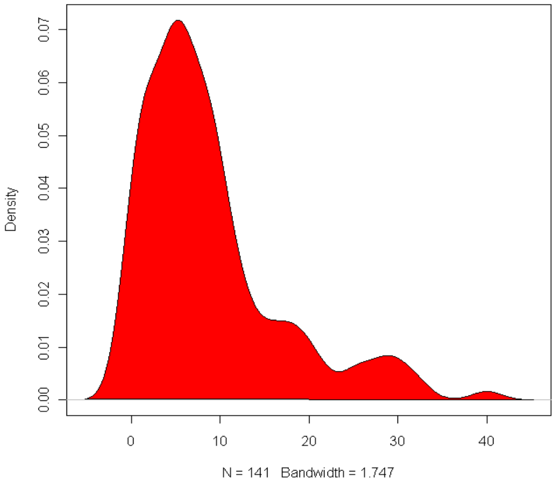
Kernel Density of patients in Study 7



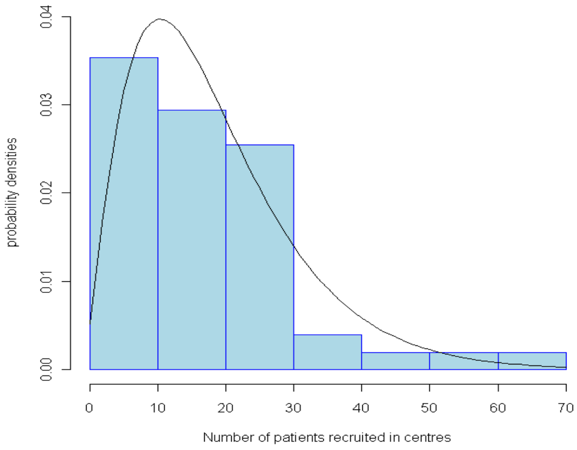
Histogram of recruited patients with fitted NB in Study 8



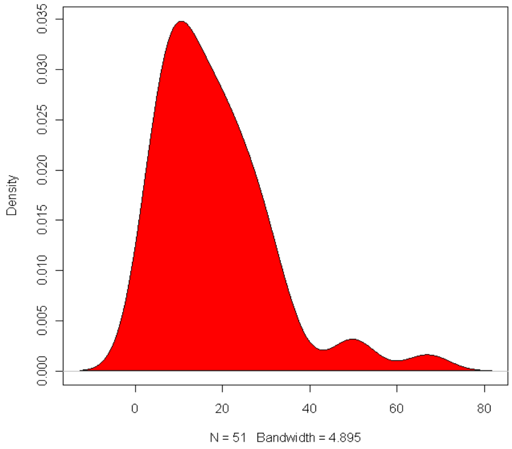
Kernel Density of patients in Study 8

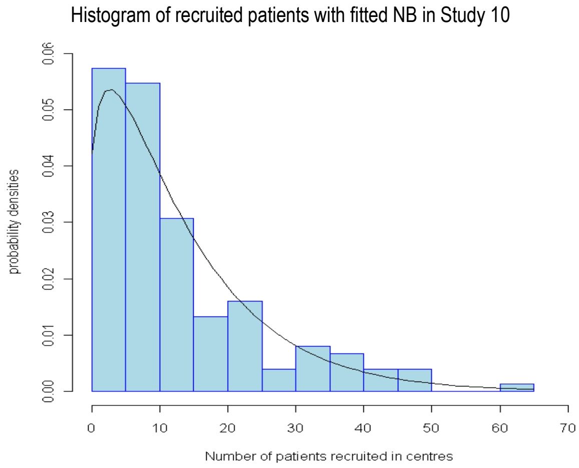


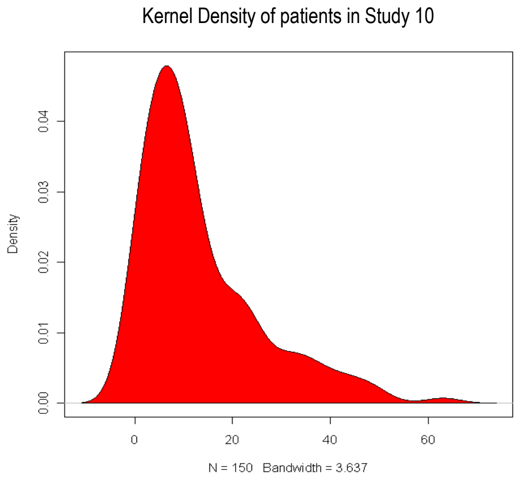
Histogram of recruited patients with fitted NB in Study 9

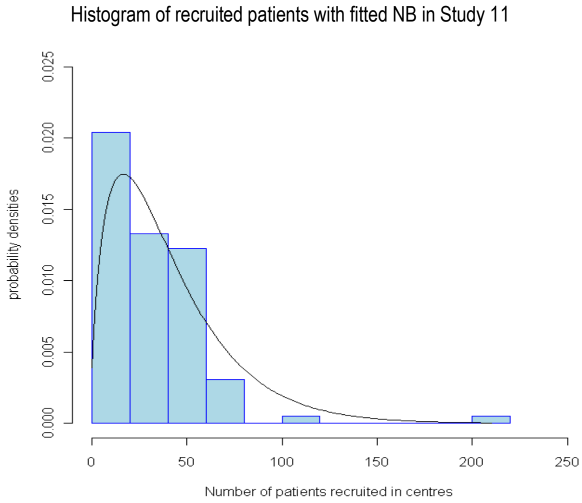


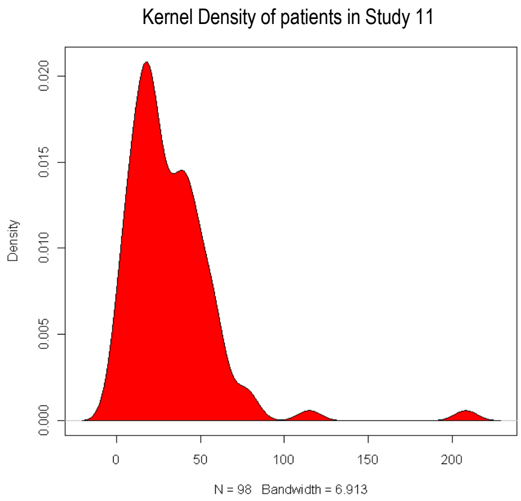
Kernel Density of patients in Study 9

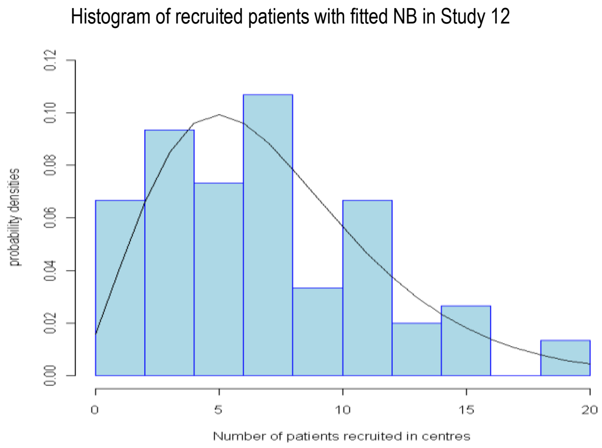


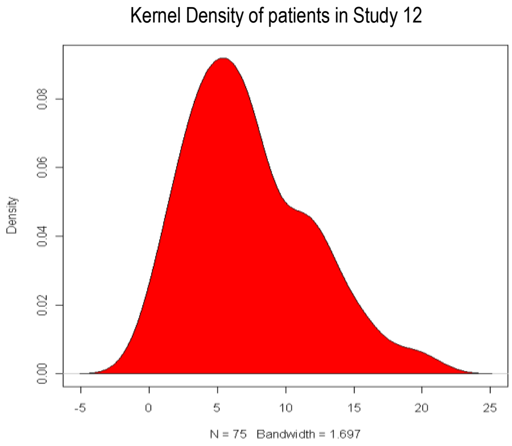

Histogram of recruited patients with fitted NB in Study 10
probability densities
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
10
20
30
40
50
60
70
Number of patients recruited in centres

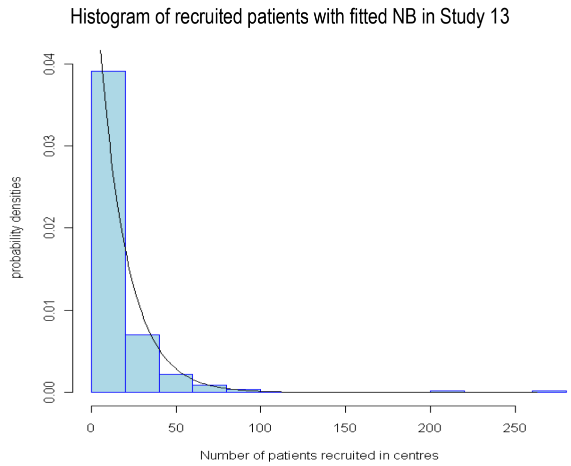

Kernel Density of patients in Study 10
Density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
N = 150 Bandwidth = 3.637


Histogram of recruited patients with fitted NB in Study 11
probability densities
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200
250
Number of patients recruited in centres

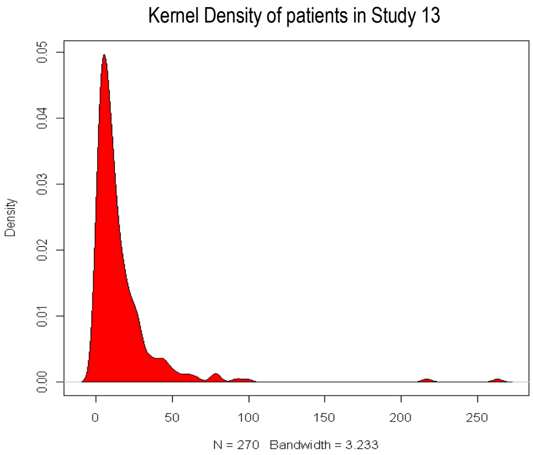

Kernel Density of patients in Study 11
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
N = 98 Bandwidth = 6.913


Histogram of recruited patients with fitted NB in Study 12
probability densities
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
Number of patients recruited in centres

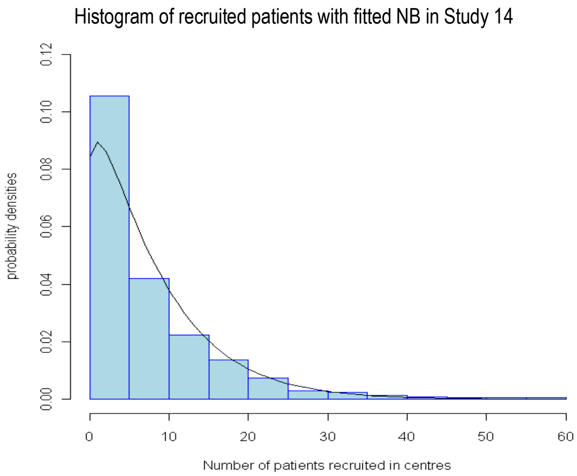

Kernel Density of patients in Study 12
Density
0.00
0.02
0.04
0.06
0.08
-5
0
5
10
15
20
25
N = 75 Bandwidth = 1.697



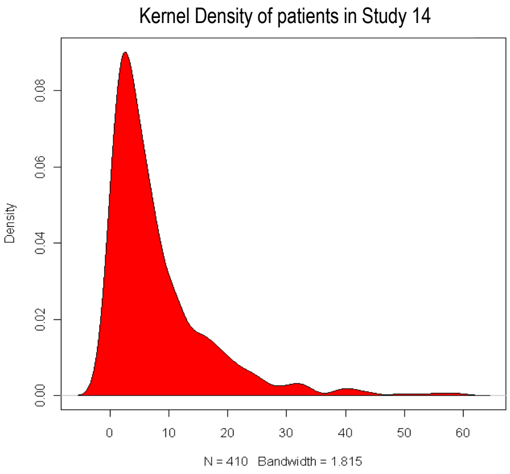
Histogram of recruited patients with fitted NB in Study 13



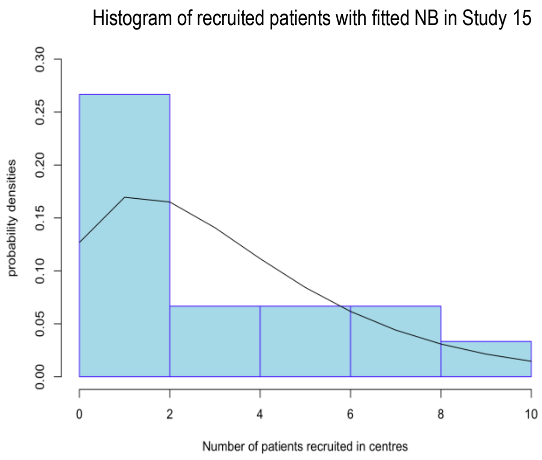
Kernel Density of patients in Study 13



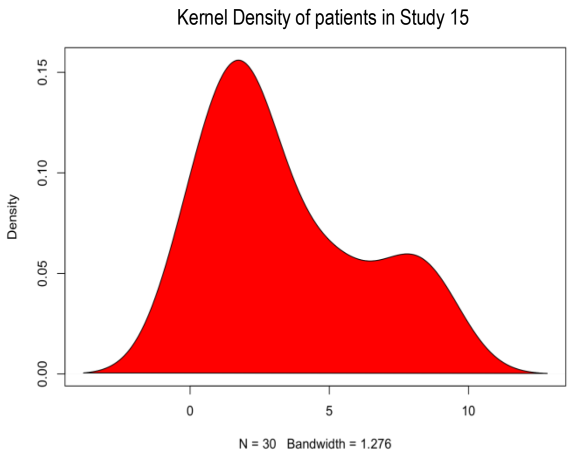
Histogram of recruited patients with fitted NB in Study 14



Kernel Density of patients in Study 14



Histogram of recruited patients with fitted NB in Study 15



Kernel Density of patients in Study 15

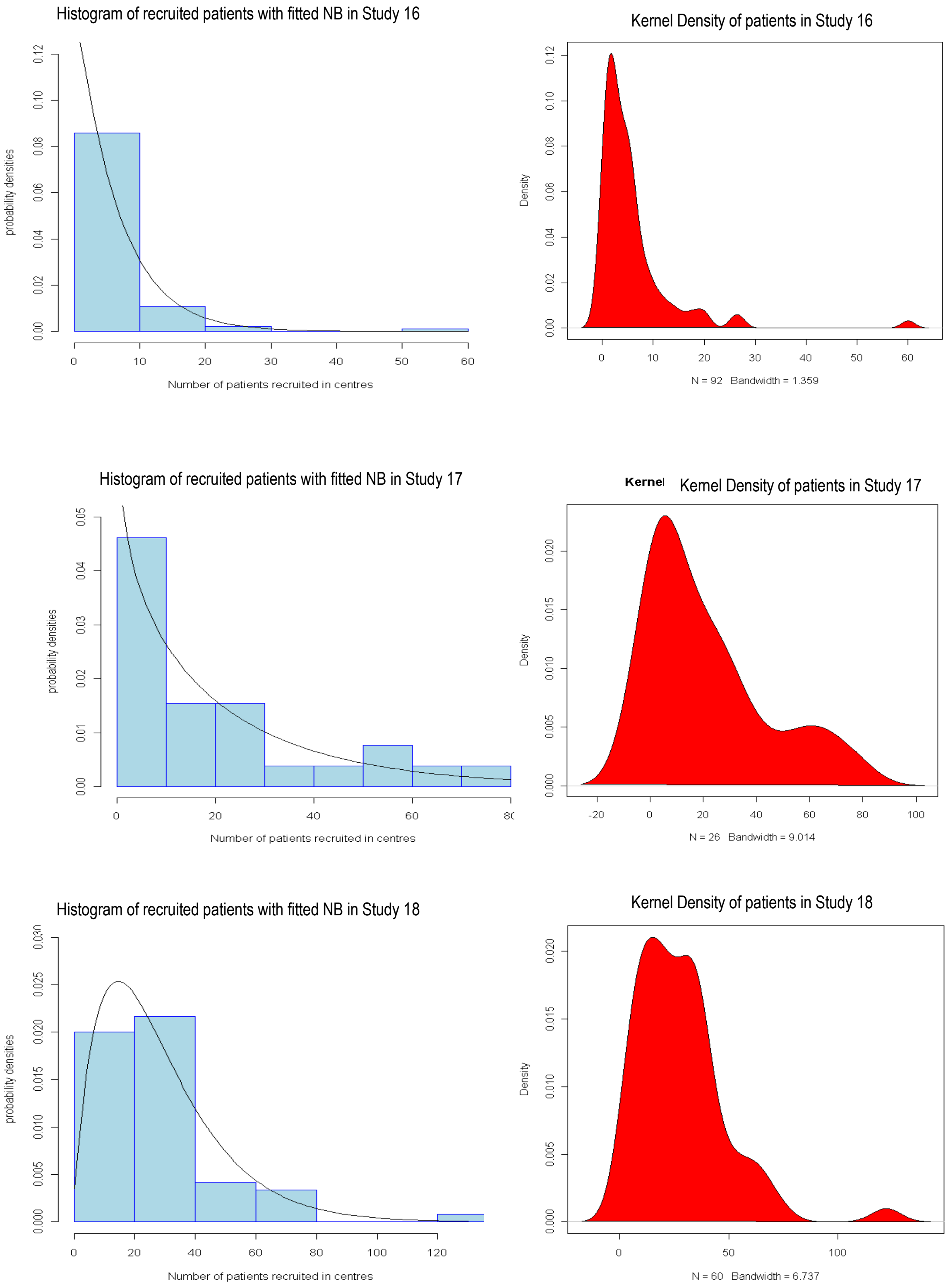


Figure 2.1: Histograms of the frequency data with the fitted Negative Binomial curves (left) and the Kernel density curves (right) in studies

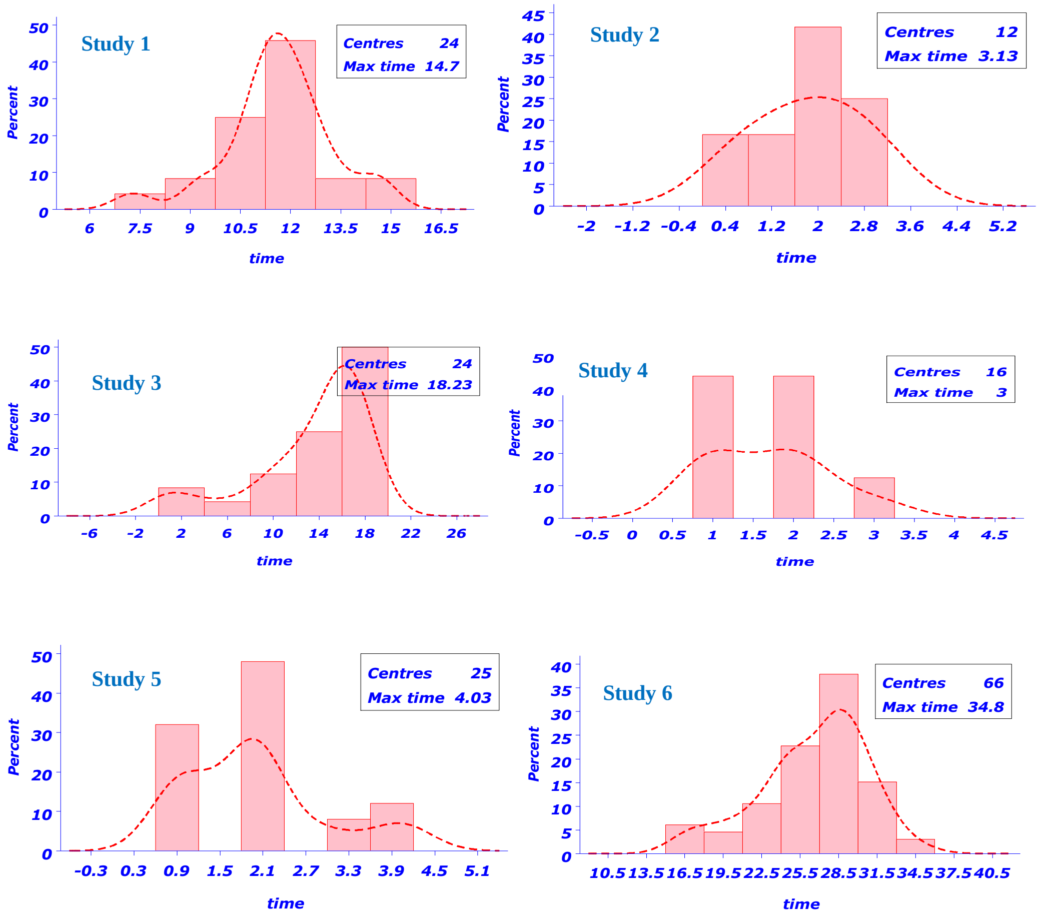
The initial aim is to model the patient recruitment process for each completed study. It is, in fact, to find a feasible distribution for the recruitment data in order to be able to assess how many patients arrive into the clinical centres during a maximum recruitment time in a real on-going trial.

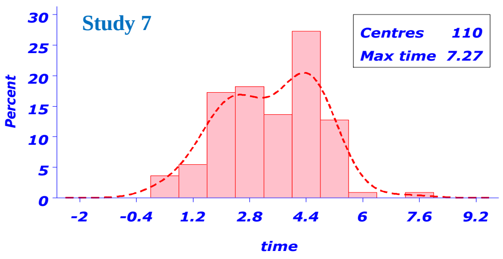
2.3. Time domain

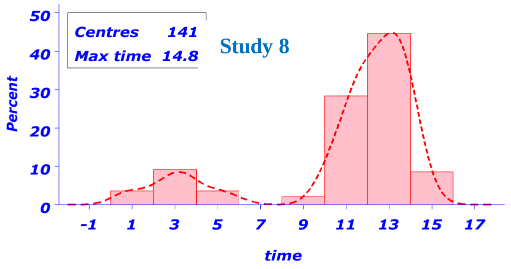
Analysing the accrual time (length of time to recruit patients), however, appears to be more important in clinical trials. This is due to the fact that in clinical trials the number of patients to be recruited is usually arranged in advance. Therefore, the companies know how many patients and centres are needed for a clinical trial. As a result, with a pre-arranged sample size and number of clinical centres, predicting the required time to recruit patients becomes essential. That is, a critical question when designing studies on how long will the studies take to recruit a pre-defined number of patients given a pre-specified number of centres. Hence, the length of time to recruit patients was also to be modelled for forecasting the maximum recruitment time. Nevertheless, there were some problems with the time data that had to be addressed for the prediction purpose.

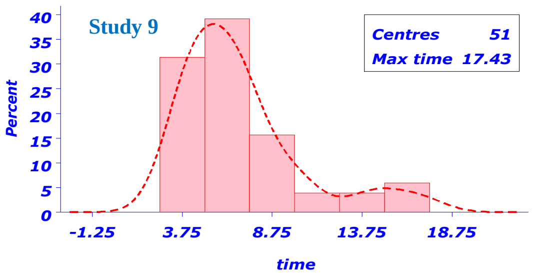
The first issue was inadequacy of information about the individual patients' arrival dates to the clinical centres. Only the total recruitment period for each centre was available in most trials. It meant that there was very little information about the waiting time between patient arrivals in clinical centres. It, then, made the analysis of the waiting time between patients more challenging. The possible solution was simulation, which is discussed in chapter 6.

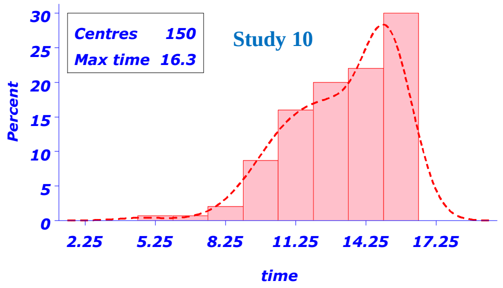
The second issue was the lack of compatibility between the distribution of the frequency domain and the distribution of the time domain. The density of the recruitment time data was not similar to what was expected in Anisimov's theory [3]. This problem is evidenced from the histograms and kernel density curves of the time data (figure 2.2). Overall, the arrival times in the trials do not follow a similar distribution. The general characteristics of the accrual time are illustrated in figures 2.2. The issue is expanded in chapter 3 and an analytical solution has been provided as well.

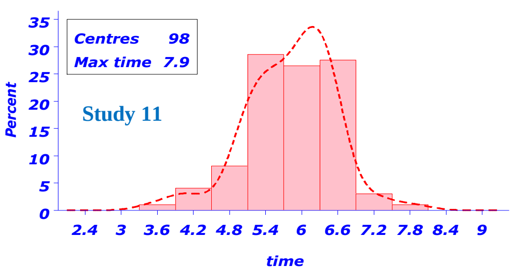


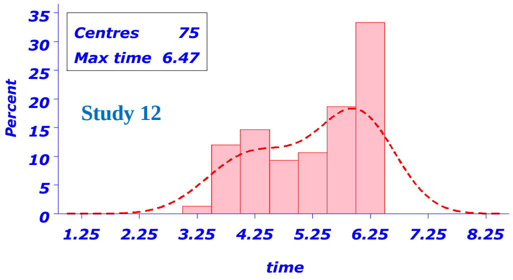

Study 7
Centres 110
Max time 7.27
Percent
0
5
10
15
20
25
30
-2
-0.4
1.2
2.8
4.4
6
7.6
9.2
time

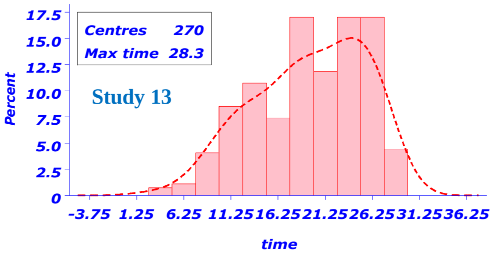

Centres 141
Max time 14.8
Study 8
Percent
0
10
20
30
40
50
-1
1
3
5
7
9
11
13
15
17
time

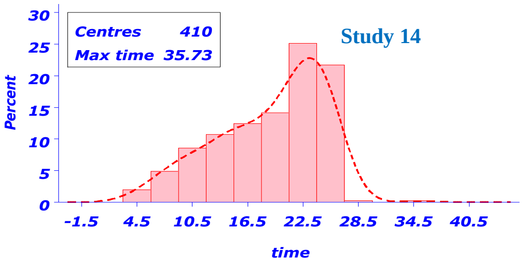

Study 9
Centres 51
Max time 17.43
Percent
0
5
10
15
20
25
30
35
40
-1.25
3.75
8.75
13.75
18.75
time

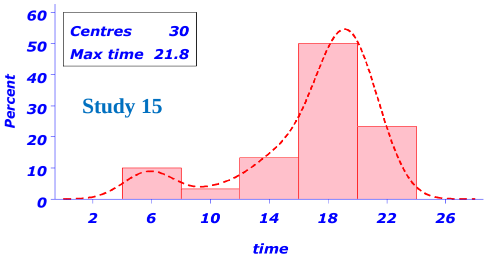

Centres 150
Max time 16.3
Study 10
Percent
0
5
10
15
20
25
30
2.25
5.25
8.25
11.25
14.25
17.25
time

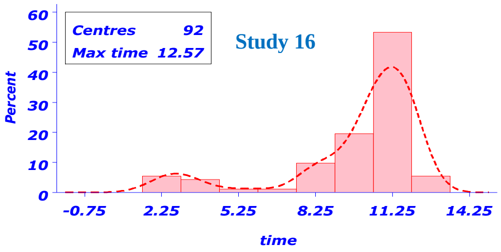

Centres 98
Max time 7.9
Study 11
Percent
0
5
10
15
20
25
30
35
2.4
3
3.6
4.2
4.8
5.4
6
6.6
7.2
7.8
8.4
9
time


Centres 75
Max time 6.47
Study 12
Percent
0
5
10
15
20
25
30
35
1.25
2.25
3.25
4.25
5.25
6.25
7.25
8.25
time


Centres 270
Max time 28.3
Study 13
Percent
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
-3.75
1.25
6.25
11.25
16.25
21.25
26.25
31.25
36.25
time


Centres 410
Max time 35.73
Study 14
Percent
0
5
10
15
20
25
30
-1.5
4.5
10.5
16.5
22.5
28.5
34.5
40.5
time


Centres 30
Max time 21.8
Study 15
Percent
0
10
20
30
40
50
60
2
6
10
14
18
22
26
time


Centres 92
Max time 12.57
Study 16
Percent
0
10
20
30
40
50
60
-0.75
2.25
5.25
8.25
11.25
14.25
time

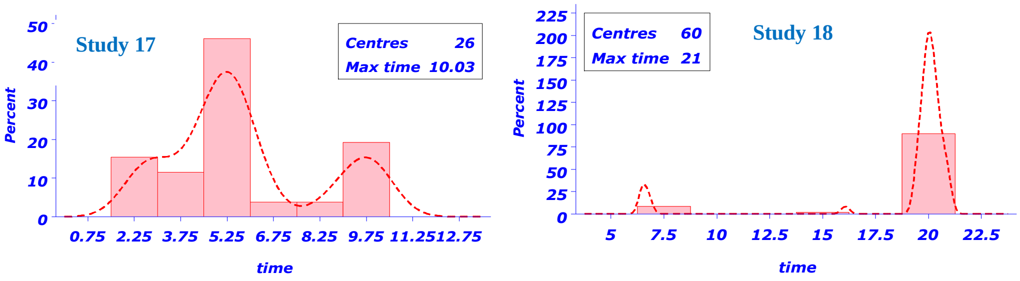


Figure 2.2: Probability histograms of patients' accrual times (months) in centres and estimated Kernel density curves

The units of the time domain were set to months throughout the analysis. Specifying the units of time in the time domain should be compatible with the units of time in the frequency domain. Rescaling the units from one domain to the other one would end up getting very different values in estimating the mean and variance of the distribution of the time domain and results in different parameters. Prof. Stephen Senn has highlighted this aspect in his notes and explained that it is the BETA parameter that is affected not the ALPHA parameter. (See appendix 2)

Chapter 3
Modelling the Recruitment Data

3.1. Modelling the frequency domain

Exploring the distributions of the variables in a data set is a fundamental step in data analysis. As discussed in the literature review chapter, the promising statistical model for the patient arrivals to clinical centres was the Poisson–Gamma mixture distribution.

The advantage of the Poisson–Gamma mixed model is that it lets the rate of patient arrivals, for a given centre; vary as a random realisation from a Gamma distribution. Then, patient arrivals at a centre are assumed to follow the Poisson distribution with the average rate given by the Gamma variable.

In the case that the mean parameter varies in the population but follows a Gamma distribution, the Poisson process can be suitably replaced by the Negative Binomial distribution. That is why the Negative Binomial distribution is more flexible in modelling the count data than the Poisson model. In contrast to the Poisson distribution in which the mean and variance are identical, in a Negative Binomial distribution the mean is smaller than the variance.

In this section, the recruitment data are modelled in the frequency domain assuming that they follow the Negative Binomial distribution. Then two methods of estimating the parameters are discussed and compared. Finally the goodness of fit test is developed to assess the suitability of the fitted model to the frequency data. However, it should be emphasised that only the frequency data have been modelled and the recruitment time has not been taken into account yet.

A further model, taking into account the recruitment time, will be discussed in chapter 5.

3.1.1. Distribution of the recruitment frequency data

If it is assumed that there are N clinical centres and each centre recruits $n_i, (i = 1, ..., N)$ patients with the recruitment rate λ_i; then the distribution of x_i, patient arrivals in centre i, follows the Poisson distribution with mean λ_i and probability density function

$$p(x_i \mid \lambda_i) = \frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!}, 0 \leq \lambda_i \leq \infty \tag{3.1}$$

If, however, λ_i is Gamma distributed with shape parameter α and scale parameter β and

$$f(\lambda_i) = \frac{\lambda_i^{\alpha-1} e^{\frac{\lambda_i}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, \alpha > 0, \beta > 0 \tag{3.2}$$

Then by multiplying the Poisson and Gamma function together and integrating out the unknown lambda [12], the outcome would be a Negative Binomial distribution with parameters $r = \alpha$ and $p = \frac{1}{1+\beta}$ [12,13]

$$P(x_i;\alpha,\beta)=\frac{\Gamma(x_i+\alpha)}{\Gamma(\alpha)\Gamma(x_i+1)}\left(\frac{\beta}{\beta+1}\right)^{x_i}\left(\frac{1}{\beta+1}\right)^{\alpha}=\begin{pmatrix}x_i+\alpha-1\\ x_i\end{pmatrix}\left(\frac{\beta}{\beta+1}\right)^{x_i}\left(\frac{1}{\beta+1}\right)^{\alpha}$$

$$i=1,\dots,N \tag{3.3}$$

In another word:

If $x \sim poisson(\lambda)$ & $\lambda \sim Gammd(\alpha,\beta) \Rightarrow x \sim NB\left(r=\alpha, p=\frac{1}{1+\beta}\right)$

With
$$\begin{aligned}&\Rightarrow \mu=\alpha\beta\\ &\sigma^2=\alpha\beta^2=\mu+\frac{1}{\alpha}\mu^2 \Rightarrow \sigma^2>\mu\end{aligned} \tag{3.4}$$

If, however, it is supposed that all the clinical centres initiate simultaneously and each centre recruits n_i patients during a fixed time period t, the total number of patients recruited up to time t follows the Negative Binomial distribution with parameters αN and $\frac{t}{\beta}$ [3] and the probability density:

$$P\left(x_i;\alpha N,\frac{t}{\beta}\right)=\frac{\Gamma(x_i+\alpha N)}{\Gamma(\alpha N)\Gamma(x_i+1)}\left(\frac{\frac{t}{\beta}}{\frac{t}{\beta}+1}\right)^{x_i}\left(\frac{1}{\frac{t}{\beta}+1}\right)^{\alpha N}=\begin{pmatrix}x_i+\alpha N-1\\ x_i\end{pmatrix}\left(\frac{\frac{t}{\beta}}{\frac{t}{\beta}+1}\right)^{x_i}\left(\frac{1}{\frac{t}{\beta}+1}\right)^{\alpha N}\Rightarrow$$

$$p\left(x_i;\alpha N,\frac{t}{\beta}\right)=\frac{\Gamma(x_i+\alpha N)}{\Gamma(\alpha N)\Gamma(x_i+1)}\left(\frac{t}{t+\beta}\right)^{x_i}\left(\frac{\beta}{t+\beta}\right)^{\alpha N}\quad i=1,\dots,N \tag{3.5}$$

Since all the centres are activated at the same time the variation in initiation date is not included in (3.5). Nevertheless, the recruitment time from trial to trial may vary due to different nature and structure of the studies. This aspect is considered in modelling the frequency domain in equation (3.5) as it follows the Negative Binomial distribution.

3.1.2. Parameter estimation

Before analysing the data, we need to assess the values of the parameters and translate this information into a sort of prior distribution for the parameters. The marginal distribution of the frequency data was modelled as a Poisson–Gamma mixture, which led to a Negative Binomial distribution. But, in order to use this model the value of the parameters should be known. Since the parameters of the Negative Binomial distribution were unknown, the Maximum Likelihood (ML) estimation and Method of Moments (MM) were applied in R as well as SAS® to estimate the parameters.

3.1.2.1. Maximum likelihood estimation

The log-likelihood function of a Negative Binomial distribution with the probability function (3.3) is

$$l(x_i;\alpha,\beta)=\sum_{i=1}^{N}\log\left(\Gamma\left(x_i+\alpha\right)\right)-N\log\left(\Gamma\left(\alpha\right)\right)-\sum_{i=1}^{N}\log\left(\Gamma\left(x_i+1\right)\right)-N\alpha\log\left(\beta+1\right)+\sum_{i=1}^{N}x_i\log\left(\frac{\beta}{\beta+1}\right) \tag{3.6}$$

And the log-likelihood function from the equation (3.5) is

$$l\left(x_i;\alpha N,\frac{t}{\beta}\right)=\sum_{i=1}^{N}\Gamma\left(x_i+\alpha N\right)-N\log\Gamma\left(\alpha N\right)-\sum_{i=1}^{N}\log\Gamma\left(x_i+1\right)+\sum_{i=1}^{N}x_i\left(\frac{t}{t+\beta}\right)+\alpha N^2\log\left(\frac{\beta}{t+\beta}\right) \tag{3.7}$$

The parameters were estimated by writing programs in R (see appendix 1) and also applying two different procedures in SAS®.

In R, the ‘function’ statement was applied to construct the Negative Binomial log-likelihood function then the ‘optim’ statement estimated the maximum likelihood of the parameters. To proceed with this estimation process, R requires initial values of the parameters. Hence, from the trial data the minimum numbers of patients in the completed trials were set as starting points for α ; and the starting point for β were derived with the help of method of moments from (3.4) as $\frac{E(x)}{\alpha}$. 'fitdistr' in the MASS package can also estimate the MLE of the parameters.

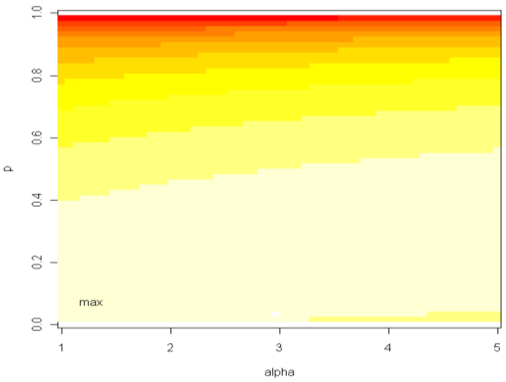
A straight-forward way to solve the estimation problem in SAS® is applying the Negative Binomial null model in the GENMOD procedure. The parameters to be estimated in PROC GENMOD are $\frac{1}{\alpha}$ and μ . In order to optimise a function, the NLP procedure in SAS® is a classic solution.

The estimated parameters were identical from applying either the log-likelihood function (3.6) or (3.7). This is because the maximum accrual times were entered as a known part of the function in (3.7). In fact, in this step the Negative Binomial distribution was only fitted to the numbers of patients recruited in clinical centres. It was, however, the case that the factor of accrual time had not been entered in the modelling. The outcome has been summarised in table 3.1.

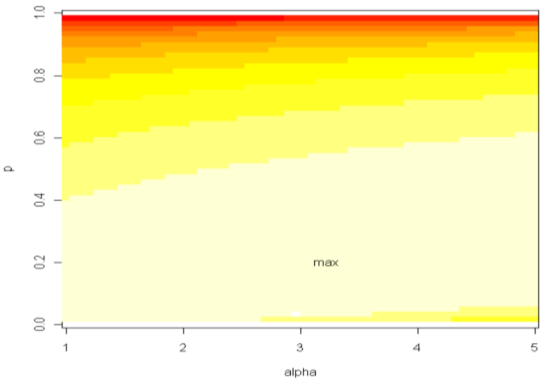
	Maximum Likelihood Estimation of frequency data							
Study	ALPHA	SE-alpha	BETA	SE-beta	P= 1/(1+beta)	mu	SE-mu	sd of data(NB)
1	1.273	0.361	12.599	4.286	0.074	16.042	3.015	18.334
2	3.220	1.641	3.934	2.126	0.203	12.667	2.282	7.905
3	2.392	0.717	14.125	4.655	0.066	33.791	4.615	22.608
4	1.280	0.534	3.907	1.896	0.204	5.000	1.238	4.953
5	2.745	0.977	3.556	1.356	0.220	9.760	1.334	6.668
6	1.435	0.270	8.402	1.826	0.106	12.061	1.311	10.649
7	1.417	0.218	7.225	1.271	0.122	10.236	0.875	9.176
8	1.416	0.198	6.216	0.991	0.139	8.801	0.671	7.970
9	2.478	0.540	7.334	1.744	0.120	18.176	1.723	12.308
10	1.314	0.162	10.146	1.462	0.090	13.333	0.995	12.191
11	1.739	0.252	18.449	3.046	0.051	30.816	2.523	24.979
12	3.988	1.043	1.826	0.495	0.354	7.280	0.524	4.536
13	0.995	0.083	16.245	1.699	0.058	16.159	1.016	16.693
14	1.227	0.095	6.505	0.592	0.133	7.985	0.382	7.742
15	2.191	1.006	1.567	0.761	0.390	3.433	0.542	2.969
16	1.070	0.182	5.416	1.096	0.156	5.794	0.636	6.097
17	0.816	0.214	25.873	8.868	0.037	21.115	4.672	23.821
18	2.130	0.401	13.270	2.777	0.070	28.267	2.593	20.084

Table 3.1: Estimated parameters from fitting Negative Binomial to the frequency data

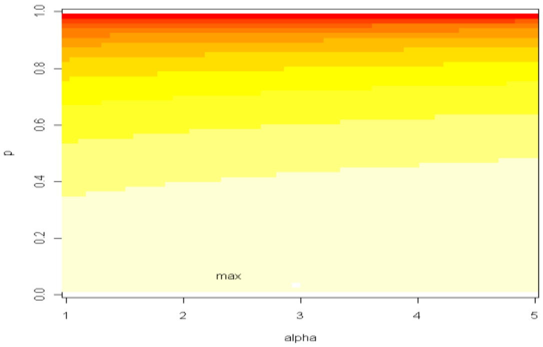
The contour plots of the Maximum likelihood functions with the maximum values of α and the probability values p, in which $p = \frac{1}{1+\beta}$, were provided (see appendix 1). The purpose of drawing the contour plots was to have a graphical view of the maximum likelihood point of both parameters in the function (figures 3.1). The 3D panels of the maximum likelihood estimation are also available (see appendix 1).



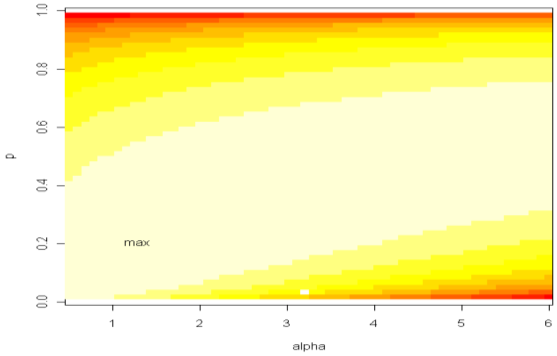
Study 1



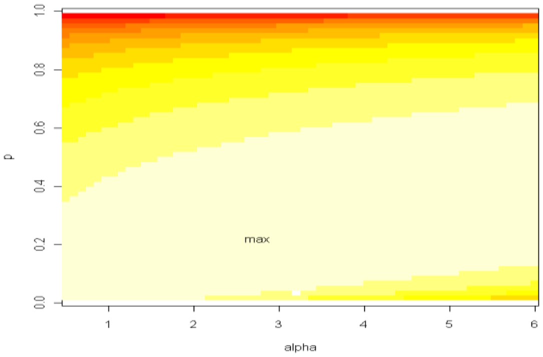
Study 2



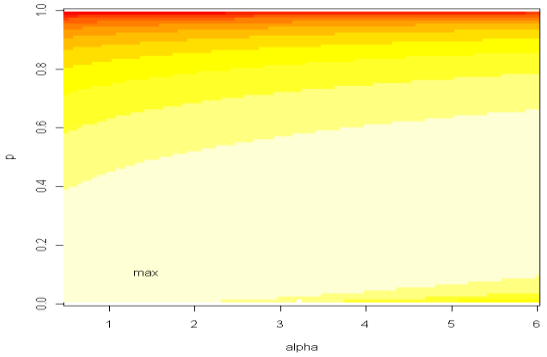
Study 3



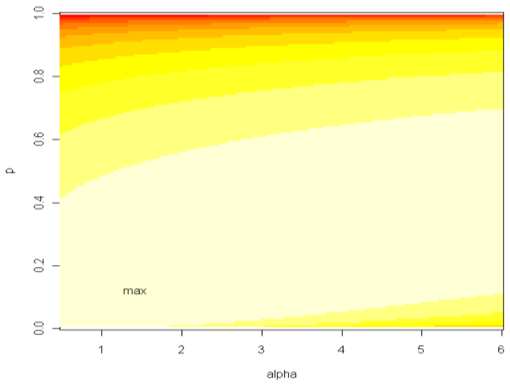
Study 4



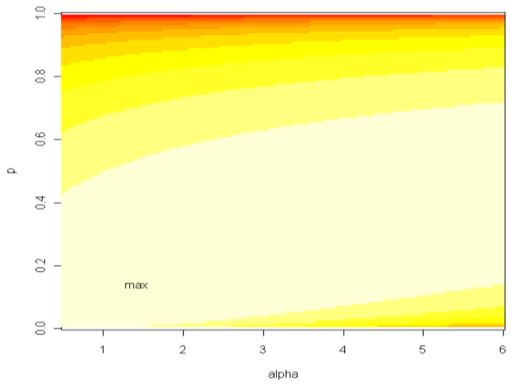
Study 5



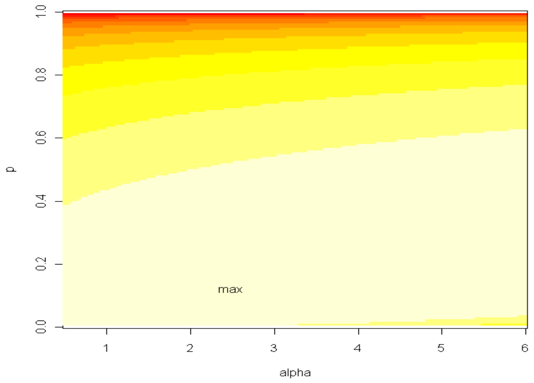
Study 6



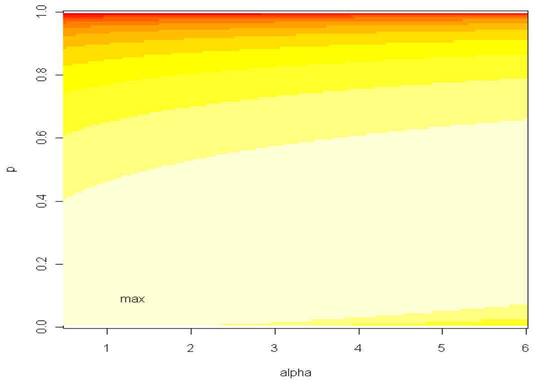
Study 7



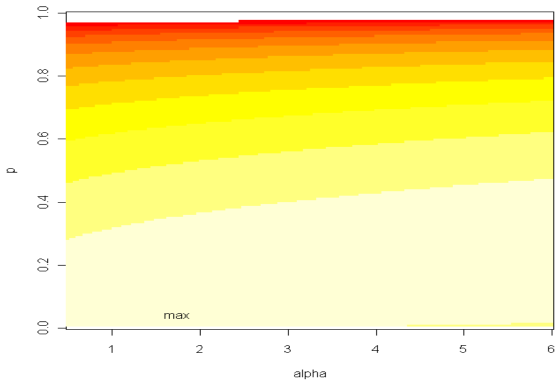
Study 8



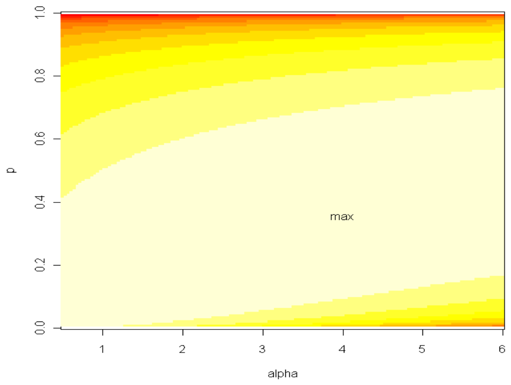
Study 9



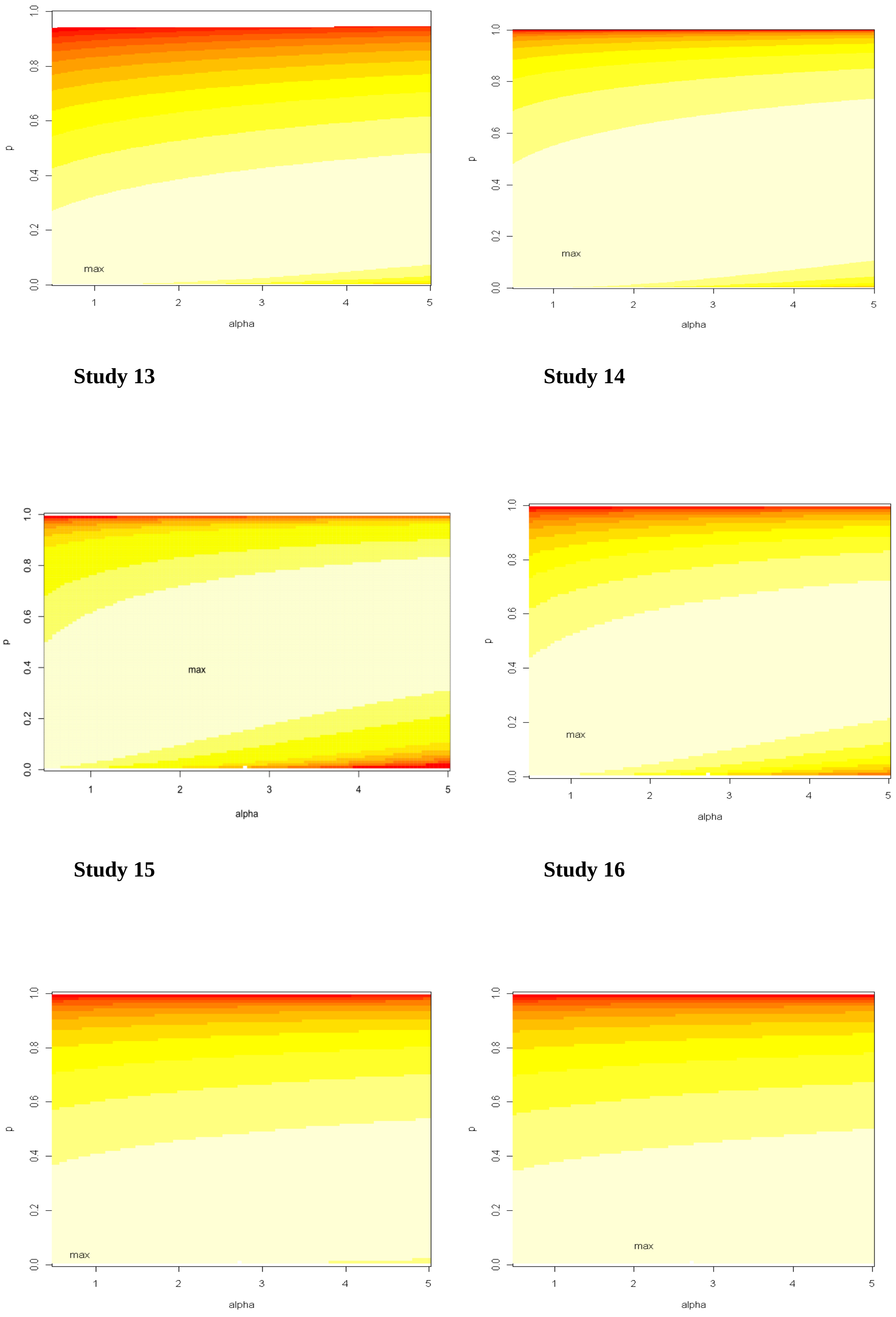
Study 10



Study 11



Study 12



Study 13

Study 14

Study 15

Study 16

Study 17

Study 18

Figure 3.1: Maximum Likelihood contour plots for parameters in the studies

3.1.2.2. Method of Moments

To capture the parameter estimation from the Method of Moments, the relationship between the parameters, mean and variance were considered as follows:

$$\alpha = \frac{(E(x))^2}{(\mathrm{var}(x))^2 - E(x)} \qquad \beta = \frac{\mathrm{E(x)}}{\alpha} \qquad \mathrm{p} = \frac{1}{1+\beta} \tag{3.8}$$

Table 3.2 shows the estimated parameter derived from Method of Moments.

	Method of moment		
Study	**ALPHA**	**BETA**	**p**
1	**0.799**	**20.080**	**0.047**
2	**2.415**	**5.244**	**0.160**
3	**2.608**	**12.957**	**0.072**
4	**0.822**	**6.080**	**0.141**
5	**2.059**	**4.741**	**0.174**
6	**1.454**	**8.294**	**0.108**
7	**1.505**	**6.803**	**0.128**
8	**1.216**	**7.236**	**0.121**
9	**1.949**	**9.326**	**0.097**
10	**1.176**	**11.333**	**0.081**
11	**1.384**	**23.179**	**0.041**
12	**2.651**	**2.746**	**0.267**
13	**0.413**	**39.133**	**0.025**
14	**0.827**	**9.651**	**0.094**
15	**1.428**	**2.404**	**0.294**
16	**0.532**	**10.886**	**0.084**
17	**0.886**	**23.832**	**0.040**
18	**1.826**	**15.478**	**0.061**

Table 3.2: Estimated parameters from Method of Moments

Figures 3.2 and 3.3 compare the Maximum likelihood method and Method of moments in estimating the values of the Negative Binomial parameters. The Maximum likelihood estimator produces slightly bigger values for the scale parameter α compared to the method of moments. However, most of the estimated β values are smaller in Maximum likelihood estimation method.

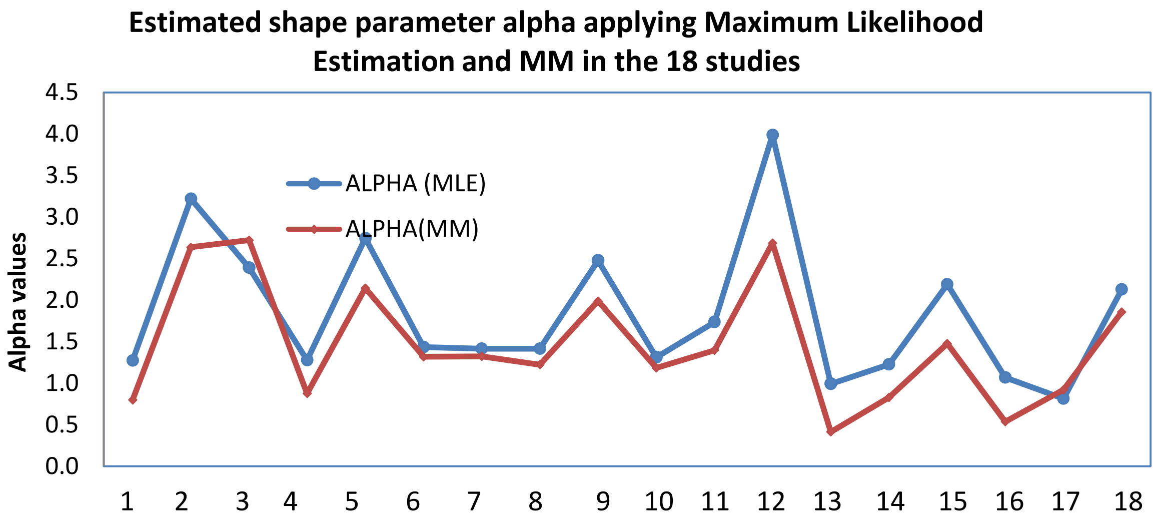


Figure 3.2: Comparing the Maximum Likelihood Estimation and Method of Moments in estimating shape parameter α by fitting Negative Binomial distribution to the frequency data

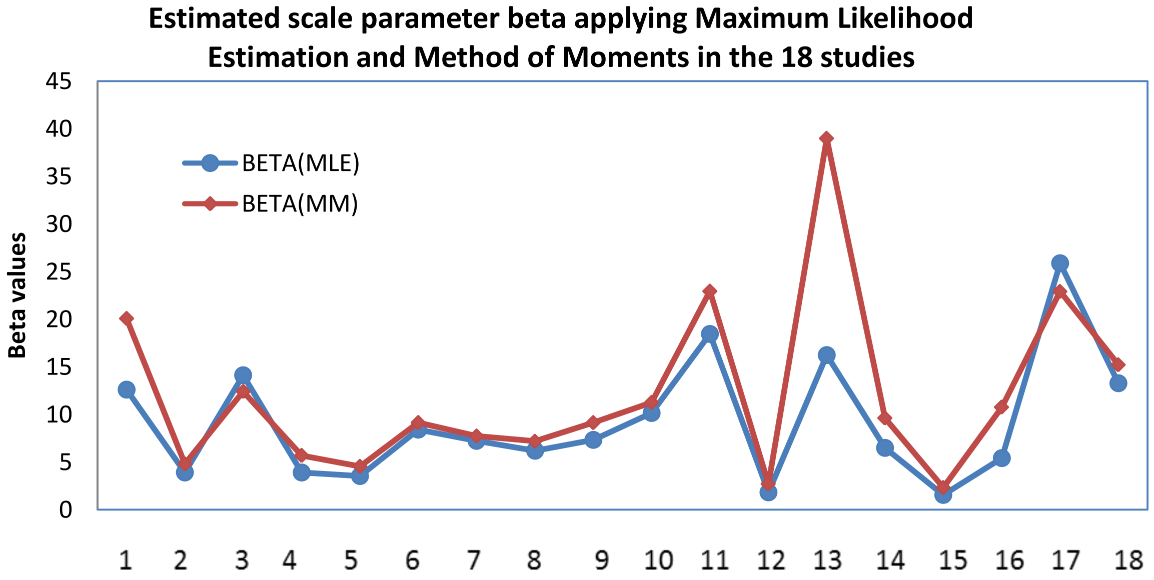


Figure 3.3: Comparing the Maximum Likelihood Estimation and Method of Moments in estimating scale parameter β by fitting the Negative Binomial distribution to the frequency data

3.1.3. Goodness of fit test (GOF)

The goodness of fit tests were run to make sure that the Negative binomial model fits well to the recruitment data. In fact, this was to measure the discrepancy between the recruitment data and the values expected for the Negative Binomial model. These measures were found by the Pearson's chi-square test in SAS.

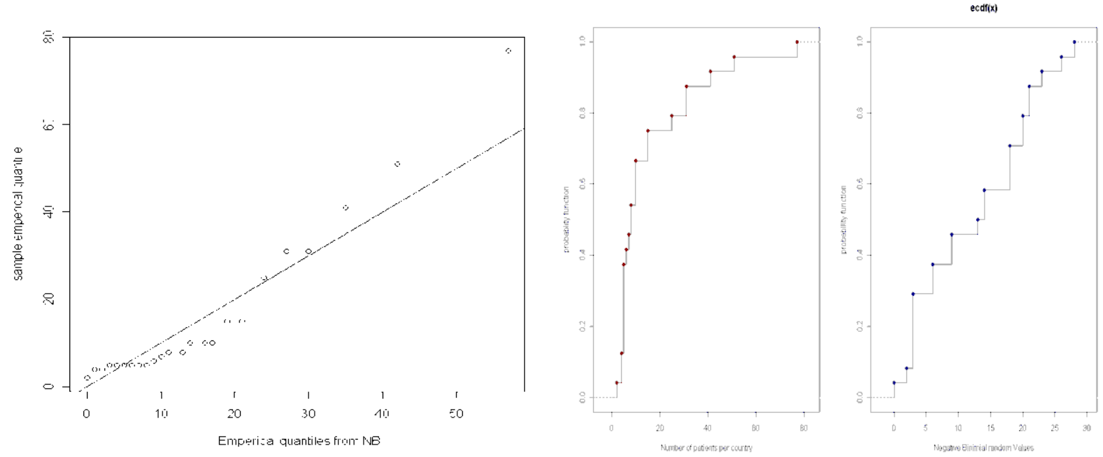
The null hypotheses were that the frequency data follow the Negative Binomial distribution. To test the hypotheses the Pearson chi-square values were compared to a chi-squared distribution. In some of the studies, the test was not significant (table 3.3), which shows that the Negative Binomial distribution has been a suitable fit for the frequency domain. Nevertheless in one or two trials the fit is clearly far from perfect and this suggests that future work could consider alternatives to the NB.

NB GOF Test in SAS				
Study	**Pearson Chi-Square value in GOF test**	**p-value**	**value in chi-sq table** $(\alpha = 0.01)$	**df**
1	**35.438**	**0.05**	**41.638**	**23**
2	**11.692**	**0.39**	**24.725**	**11**
3	**19.703**	**0.66**	**41.638**	**23**
4	**18.587**	**0.23**	**30.578**	**15**
5	**24.977**	**0.41**	**42.980**	**24**
6	**64.253**	**0.50**	**94.422**	**65**
7	**103.396**	**0.63**	**146.257**	**109**
8	**140.383**	**0.48**	**181.840**	**140**
9	**55.950**	**0.26**	**76.154**	**50**
10	**151.499**	**0.43**	**192.073**	**149**
11	**115.603**	**0.10**	**132.309**	**97**
12	**71.904**	**0.55**	**105.202**	**74**
13	**610.415**	**0.00**	**325.881**	**269**
14	**525.943**	**0.00**	**478.461**	**409**
15	**27.158**	**0.56**	**49.588**	**29**
16	**154.387**	**0.00**	**125.290**	**91**
17	**22.171**	**0.63**	**44.314**	**25**
18	**63.997**	**0.31**	**87.166**	**59**

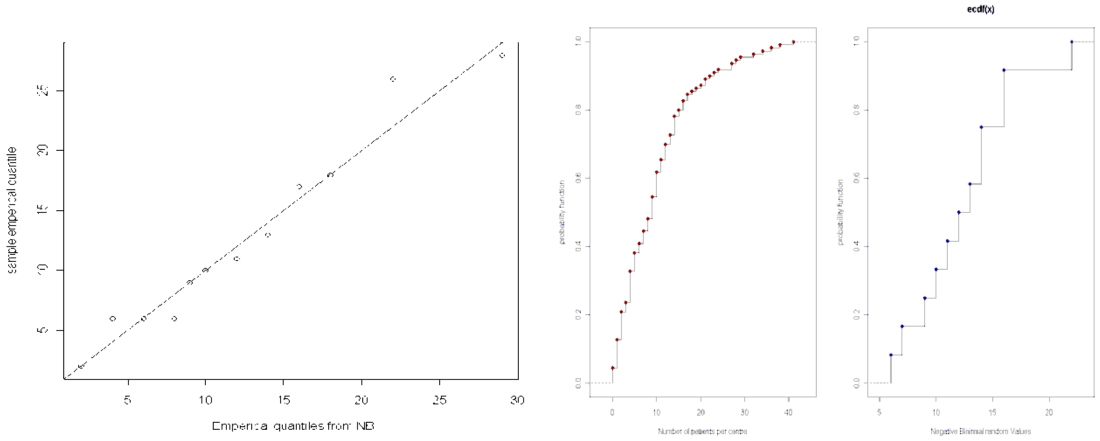
Table 3.3: Negative Binomial Goodness of Fit test to the recruitment data (Frequency domain)

Drawing the QQ-Plot is a popular graphical approach to assess the suitability of fitting the negative Binomial to the frequency data. This is the plot of the quartiles of the negative Binomial and the quartiles of the sorted recruitment data. The plot, however, should not be far too much away from the straight line (figures 3.4)

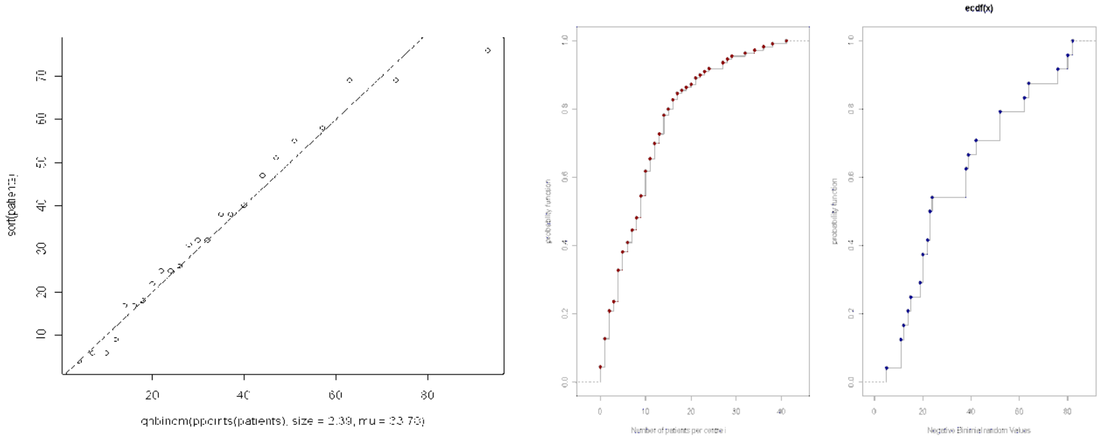
Another visual method was also applied to test the goodness of fit for the Negative Binomial to the data. The method was to compare the empirical cumulative density function (ecdf) of the frequency data and the random values of the Negative Binomial. The two cumulative functions should be similar for a good fit (figure 3.4). The graphical GOF test approve that the Negative Binomial is a good model for recruitment data in the frequency domain.



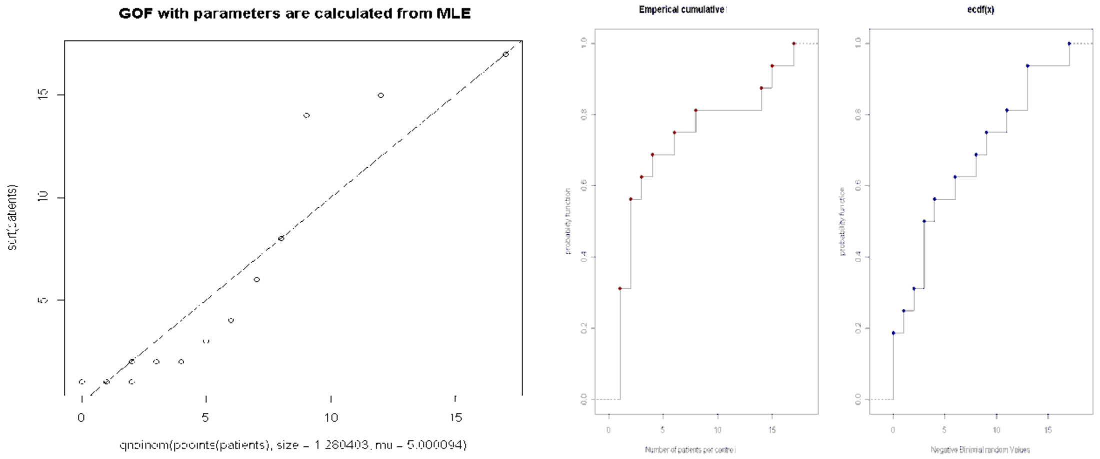
Study 1, QQ-plot (left) and the Empirical cumulative density functions (right)



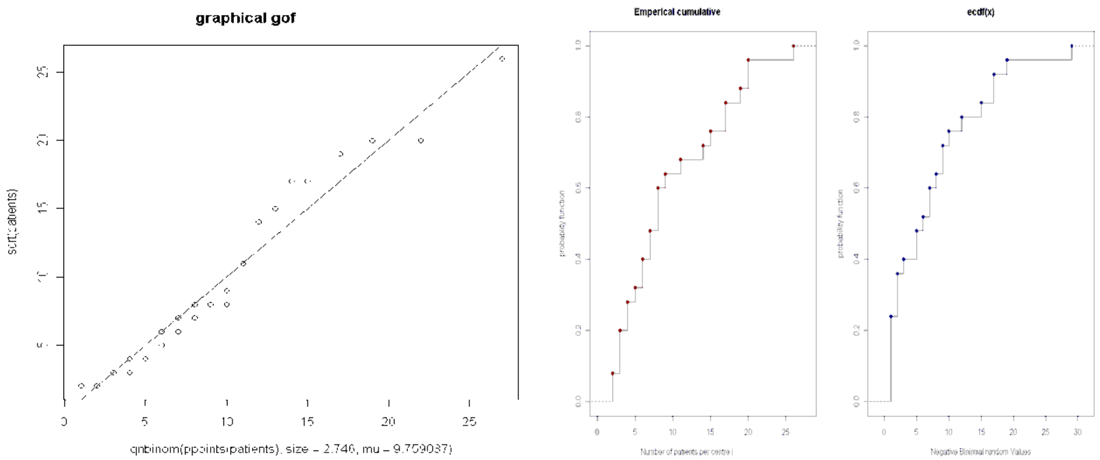
Study 2, QQ-plot (left) and the Empirical cumulative density functions (right)



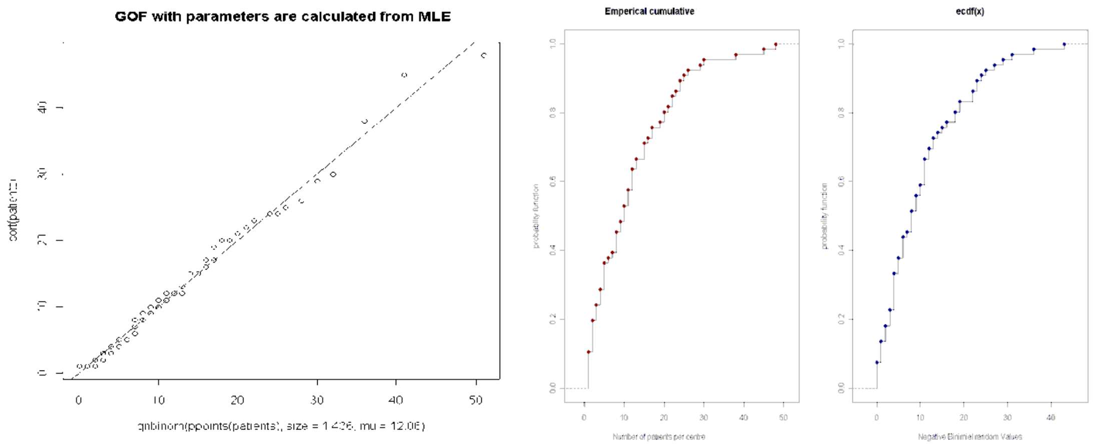
Study 3, QQ-plot (left) and the Empirical cumulative density functions (right)



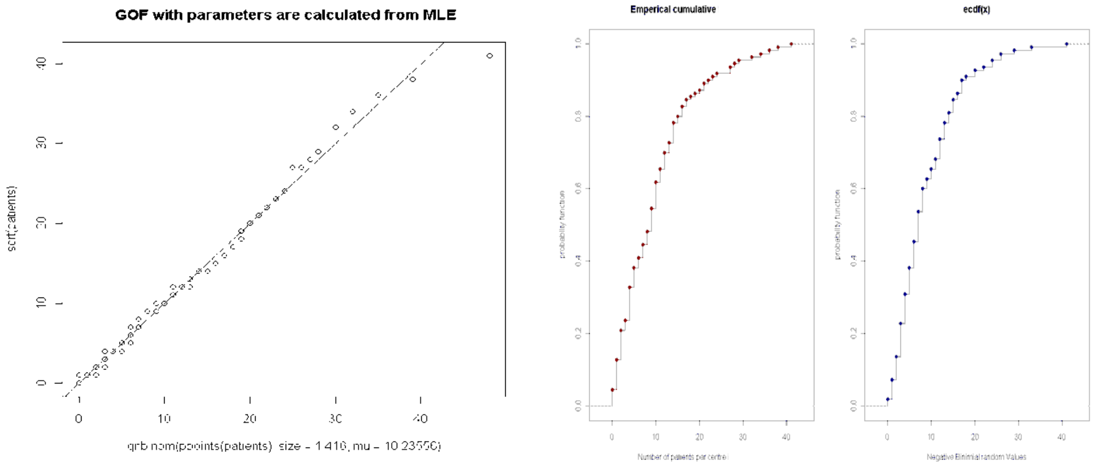
Study 4, QQ-plot (left) and the Empirical cumulative density functions (right)



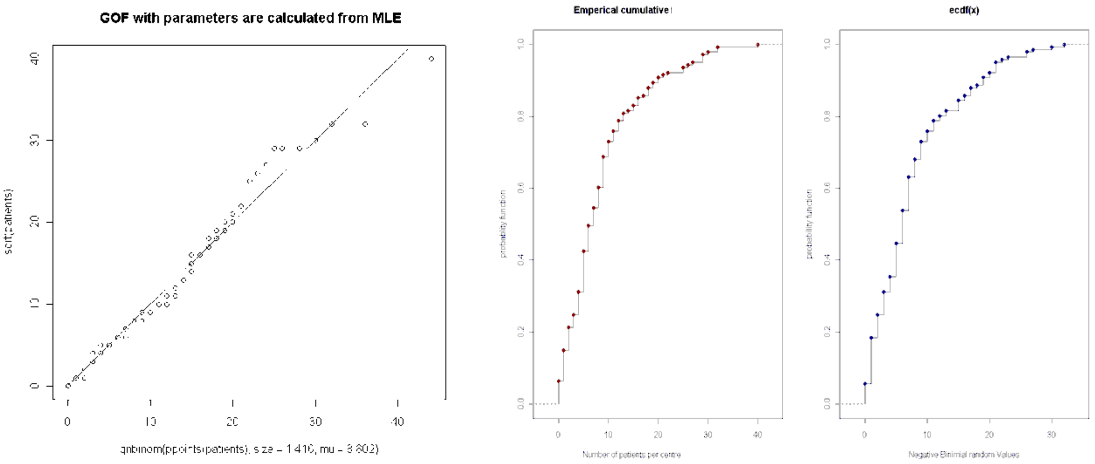
Study 5, QQ-plot (left) and the Empirical cumulative density functions (right)



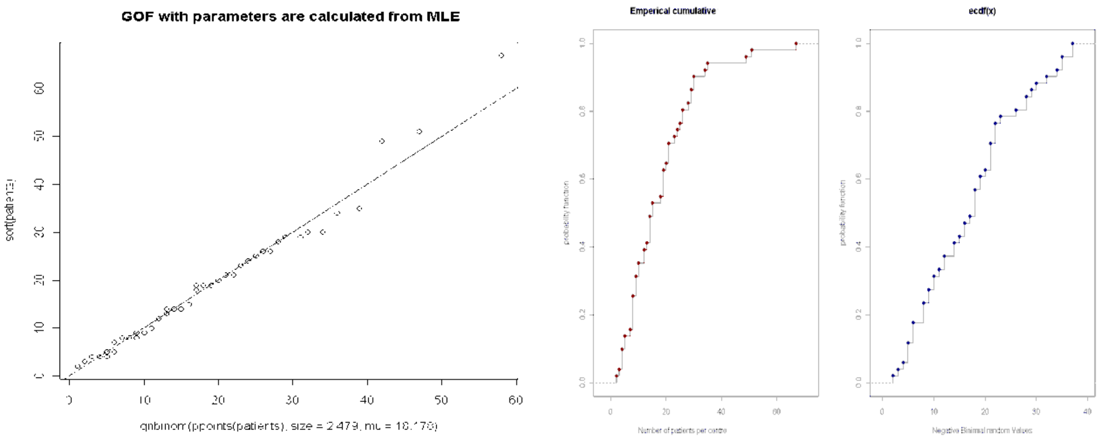
Study 6, QQ-plot (left) and the Empirical cumulative density functions (right)



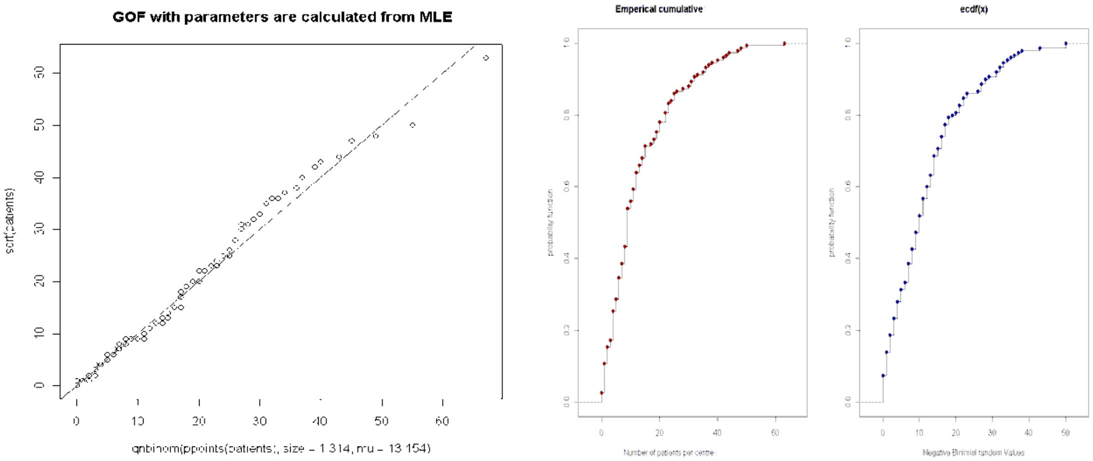
Study 7, QQ-plot (left) and the Empirical cumulative density functions (right)



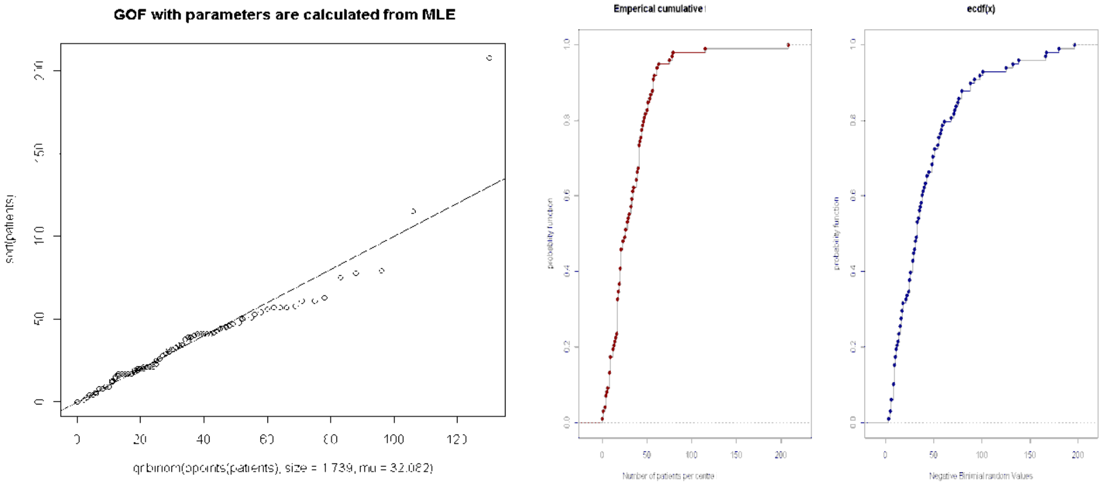
Study 8, QQ-plot (left) and the Empirical cumulative density functions (right)



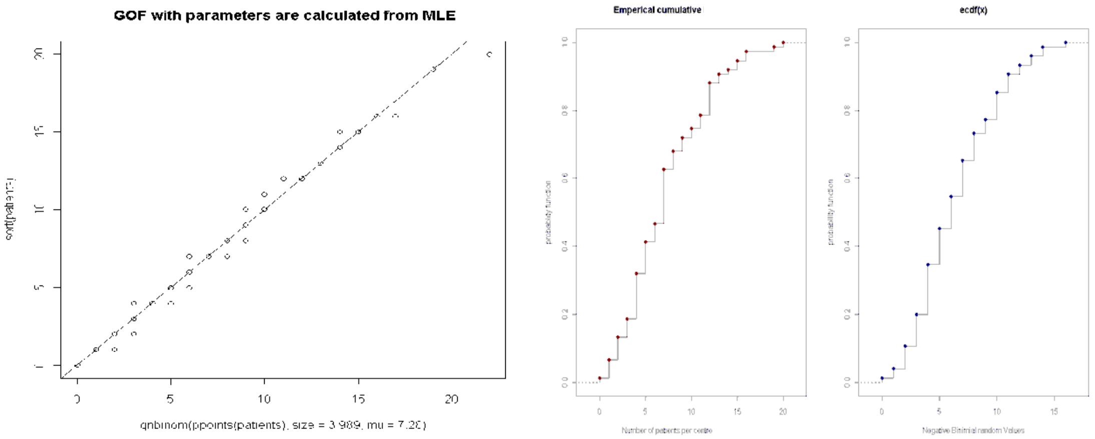
Study 9, QQ-plot (left) and the Empirical cumulative density functions (right)



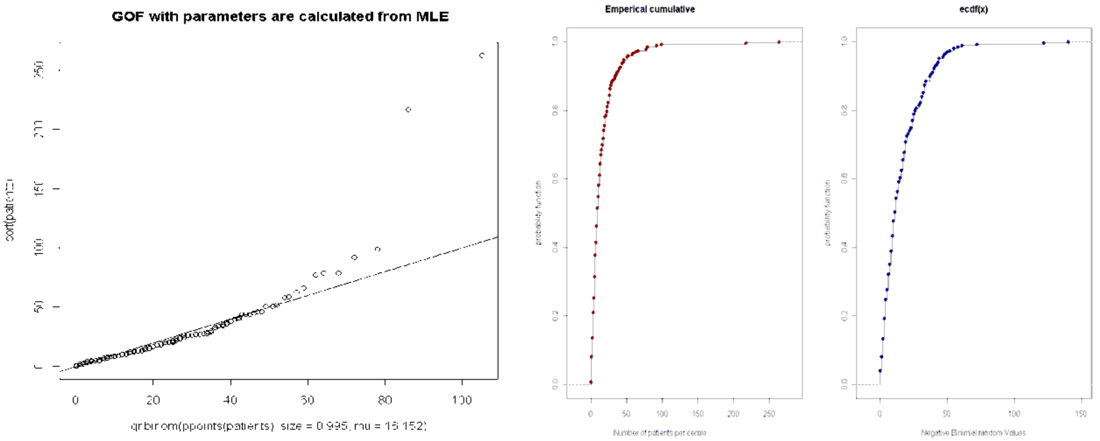
Study 10, QQ-plot (left) and the Empirical cumulative density functions (right)



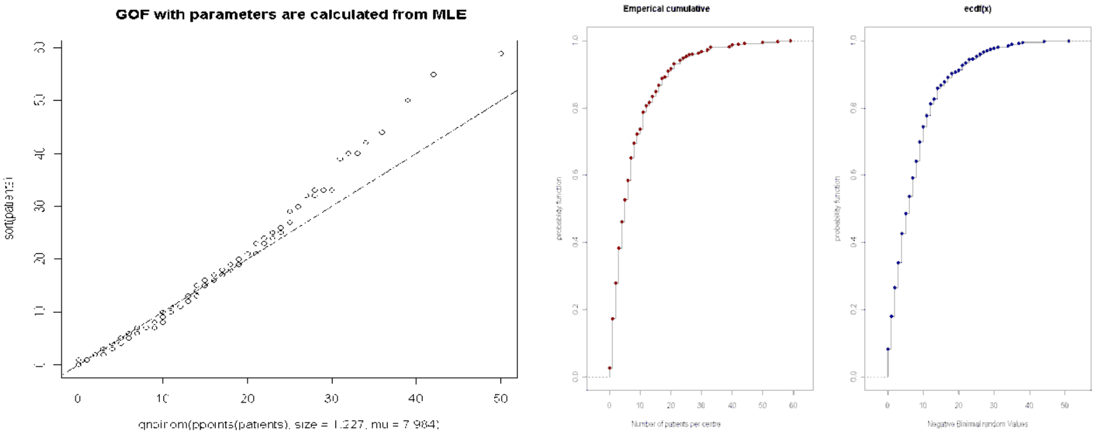
Study 11, QQ-plot (left) and the Empirical cumulative density functions (right)



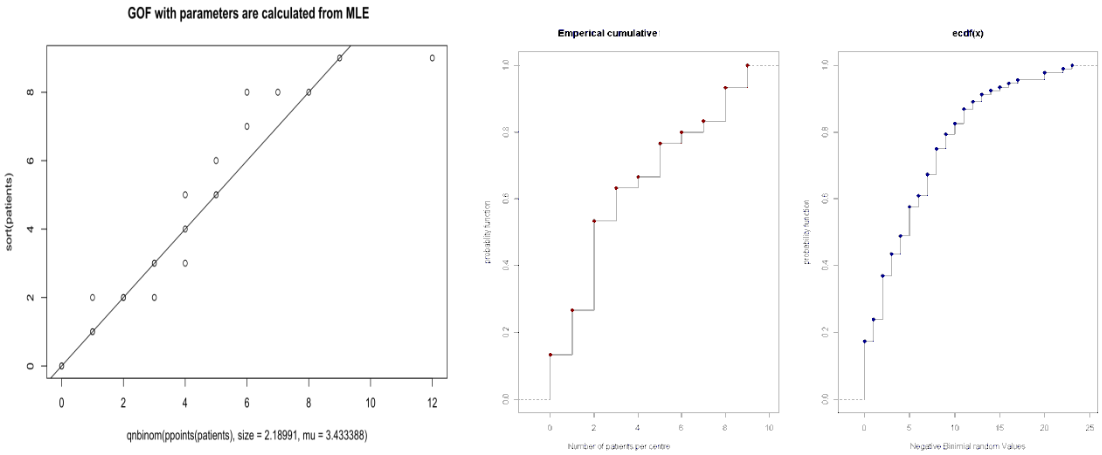
Study 12, QQ-plot (left) and the Empirical cumulative density functions (right)



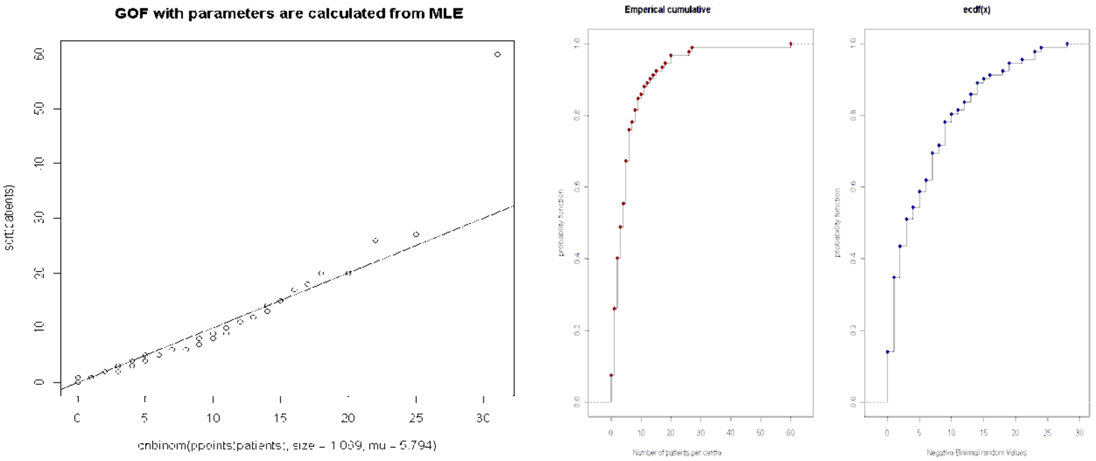
Study 13, QQ-plot (left) and the Empirical cumulative density functions (right)



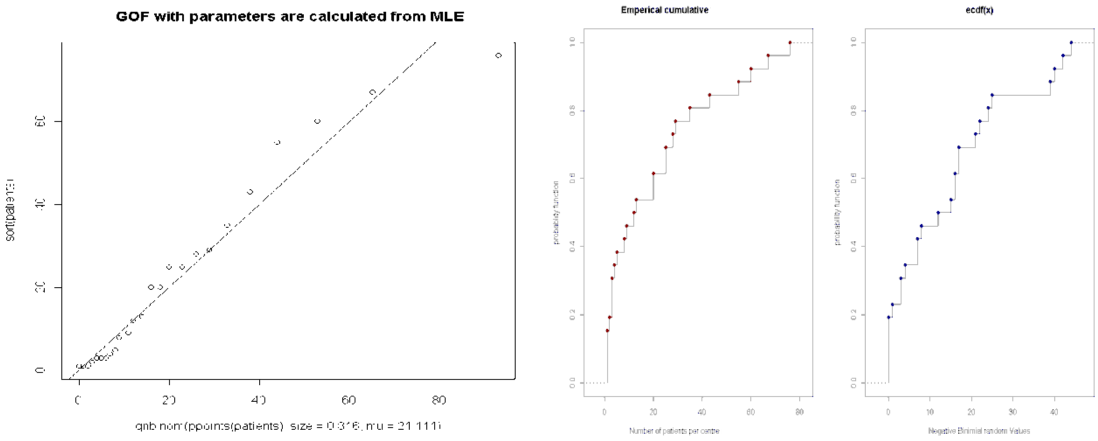
Study 14, QQ-plot (left) and the Empirical cumulative density functions (right)



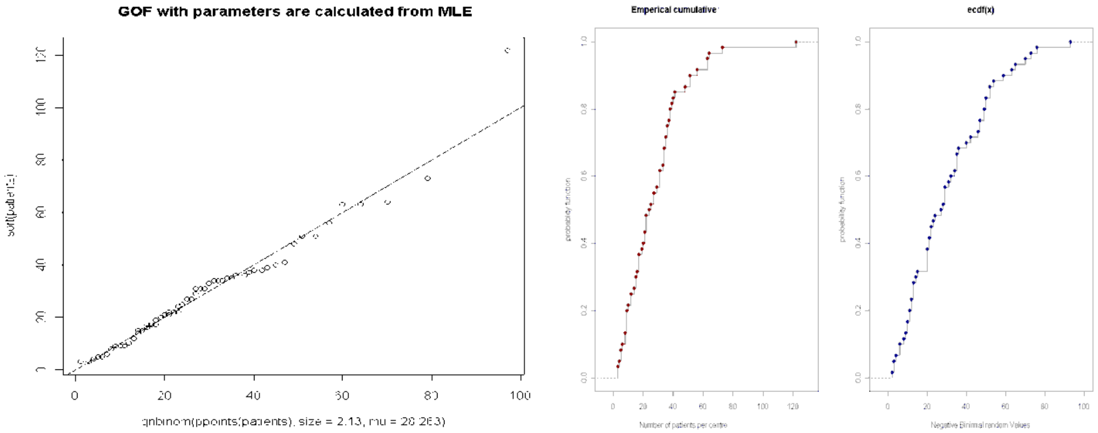
Study 15, QQ-plot (left) and the Empirical cumulative density functions (right)



Study 16, QQ-plot (left) and the Empirical cumulative density functions (right)



Study 17, QQ-plot (left) and the Empirical cumulative density functions (right)



Study 18, QQ-plot (left) and the Empirical cumulative density functions (right)

Figures 3.4: QQ-Plots and the Empirical cumulative density functions to test the suitability of the Negative Binomial to the frequency data.

3.1.4. Discussion

The unknown parameters of fitting the Negative Binomial to the frequency domain were estimated. Moreover, the tests show that the Negative Binomial is a suitable model for many trials of the recruitment frequency data. However, there are still two main issues to deal with.

First is finding a posterior distribution for the recruitment data. This requires an adequate knowledge of the true distribution of the parameters. However, there are only estimated parameters from the past studies available. Chapter 4 highlights the obstacles of modelling the parameters and brings some ideas to deal with the issue.

Second is the accrual time as an effective factor in patient recruitment analysis. The time should be considered as a factor that varies from centre to centre. Hence, in chapter 5 the accrual time was considered as an offset factor in analysing the recruitment frequency data. If the offset time variables are included in the model, the general Negative Binomial model remains unchanged although the parameters could be slightly different. In the current chapter, it was assumed that all the centres initiate simultaneously. In the case that centres initiate in different dates during the trial then the recruitment time would come across to another source of variation. This issue is also solved by adding the start time in simulating the recruitment period in chapter 5.

3.2. Modelling in time domain

Based on the theory that Anisimov & Federov [3] had in their paper, if $n_i(t)\left(i=1,...,N\right)$, which is the number of patients that are recruited in the clinical centre i up to time t, follow the Poisson process with parameter λ_i then the total number of patients recruited until time t in all the N clinical centres is $n(t)=\sum_{i=1}^{N} n_i(t)$ and follows the Poisson distribution with the overall rate parameter $\Lambda=\sum_{i=1}^{N}\lambda_i$.

Consequently, $T(n,N)$, the length of time to recruit n patients in N clinical centres will follow a Gamma distribution with shape parameter n and scale parameter Λ with the probability density function:

$$p(T,n,\Lambda)=e^{-\Lambda T}\beta^{n}T^{n-1}\Gamma(n)^{-1} \tag{3.9}$$

in which $\Gamma(n)$ is the gamma function. Hence, the expected time to recruit n patients in N clinical centres is $\frac{n}{\Lambda}$ and $Var(T)=\frac{n}{\Lambda^2}$.

However, back to the assumption of the distribution of the frequency domain, λ_i is a random variable from a Gamma distribution with parameters α and β . Hence, the overall rate $\Lambda=\sum_{i=1}^{N}\lambda_i$ is also Gamma distributed with parameters $(\alpha N,\beta)$.

As a result, the total time to recruit patients has a mixture of two independent Gamma distributions and is

$$T(n,N)\sim Gamma(n,Gamma(\alpha N,\beta)) \tag{3.10}$$

The above process leads to the Pearson VI distribution with the probability density function

$$p(T,n,\alpha N,\beta)=\frac{1}{Beta(n,\alpha N)}\frac{T^{n-1}\beta^{\alpha N}}{(T+\beta)^{n+\alpha N}}\text{ , T>=0} \tag{3.11}$$

in which a $Beta(n,\alpha N)$ is beta function [3]. Therefore, the expected time to recruit n patients in N clinical centres will be $\frac{\beta n}{\alpha N-1}$ and the variance of the recruiting time is $\frac{\beta^2 n(n+\alpha N-1)}{(\alpha N-1)^2(\alpha N-2)}$, $\alpha N>2$.

3.2.1 Parameter estimation

The parameters of the Pearson VI distribution in table 3.4 were estimated by applying the Maximum Likelihood Estimation method in SAS and also R. The log-likelihood function of the Pearson distribution with the probability function (3.11) is:

$$l(t_i, n, \alpha N, \beta) = -N\log(beta(n, \alpha N)) + (n-1)\sum_{i=1}^{N}\log(t_i) + \alpha N^2\log(\beta) - (n+\alpha N)\sum_{i=1}^{N}\log(t_i+\beta) \tag{3.12}$$

	Maximum Likelihood Estimation of the Time domain				
Study	ALPHA	SE-alpha	BETA	SE-beta	E(T)[3]
1	2.260	0.753	1.610	0.540	11.648
2	0.196	0.004	0.019	0.004	2.159
3	0.046	0.012	0.009	0.003	73.130
4	0.396	0.146	0.113	0.044	1.689
5	0.192	0.054	0.032	0.009	2.019
6	0.534	0.097	1.142	0.209	26.525
7	0.024	0.003	0.006	0.001	3.995
8	0.011	0.001	0.009	0.001	18.568
9	0.121	0.003	0.039	0.002	6.945
10	0.174	0.020	0.167	0.020	13.335
11	0.571	0.083	0.102	0.015	5.820
12	0.373	0.064	0.260	0.045	5.274
13	0.024	0.002	0.025	0.002	20.240
14	0.014	0.001	0.029	0.002	19.720
15	0.229	0.064	0.996	0.291	17.501
16	0.045	0.006	0.063	0.010	10.698
17	0.233	0.064	0.054	0.015	5.849
18	0.133	0.024	0.081	0.081	19.594

Table 3.4: Estimated parameters from fitting the Pearson VI distribution to the time domain data

[3] E(T) is the expected length of time (months) to recruit patients in trials with the fixed number of patients based on the assumption that the time domain follow the Pearson VI distribution with the estimated parameters.

In SAS, the NLP procedure estimated the parameters with the maximum likelihood estimation method. The parameters were also estimated in the MASS package in R. The ‘function’ statement was applied for the Pearson VI distribution log-likelihood function then the ‘optim’ statement estimated the maximum likelihood of the parameters. The package ‘pearsonDS’ also resulted in the similar ML estimation of the parameters.

3.3. Time and frequency parameters correlation test

It was taken from the theory that the relationship between the Negative Binomial distribution and the Pearson VI distribution is similar to the relationship between the Poisson process and the Exponential distribution. It means that the parameters α and β in the Poisson-gamma mixture are the same as the parameters α and β in the Pearson VI distribution [3]. Hence, it was expected that in the completed trials the shape parameter α in the time domain, which was estimated from Pearson VI distribution, to be broadly similar to the shape parameter α estimated from Negative Binomial distribution in frequency domain. Also the scale parameters β are to be very similar in both recruitment domains. Even if the parameters were not similar due to different parameterisations, they were expected to be strongly related. However, due to difficulty of estimating the parameters from Pearson VI distribution and probably because of software issues, the results were not as they were expected. Lack of specific option in SAS procedures for Pearson VI distribution could reduce the reliability of estimating the parameters. The Pearson VI distribution was hardly a good fit for most of the time data among the 18 completed trials. More over the frequency data were well better presented than the time data.

Nevertheless, the analysis were carried out to draw a scatter plot and run statistical correlation test for any possible correlation coefficients between estimated parameters α in frequency and time domain. The scatter plot hardly showed any relationship between the α values (figure 3.5) estimated from Negative Binomial and Pearson VI distribution.

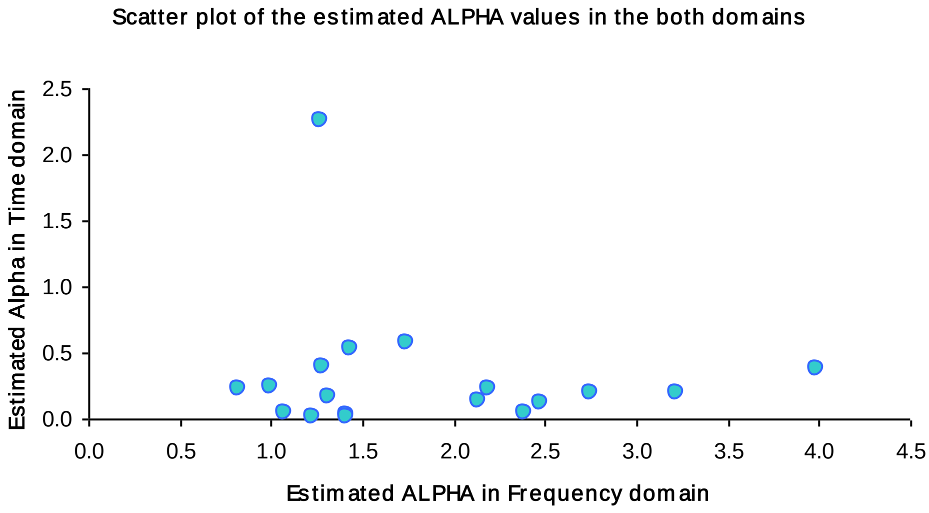


Figure 3.5: Scatter plot of the estimated ALPHA values in Frequency and time domain

Table 3.5 presents the SAS output of the correlation test between the α values in frequency domain, which were estimated from Negative Binomial distribution and the estimated α values from the Pearson VI distribution in time domain. Based on the calculated p-value (0.63) and the 95% confidence correlation limit (-0.5, 0.4) in the Pearson correlation coefficient test, the null hypotheses of no correlation between α values could not be rejected.

```
                         The CORR Procedure between alpha values

Variable      N       Mean      StdDev      Median      Min         Max
        alpha_freq          18     1.84033     0.85517     1.42600     0.81600     3.98800
        alpha_time          18     0.32178     0.51210     0.19400     0.01100     2.26000


                         Pearson Correlation Coefficients, N = 18
                               Correlation Estimate=-0.12
                                   P-value= 0.6316


              Pearson Correlation Statistics (Fisher's z Transformation)
  Variable             With variable          95% Confidence Limits       p Value for H0:Rho=0

alpha_freq                      alpha_time     -0.554176      0.369404                 0.6369
```

Table 3.5: SAS output of the correlation test between the ALPHA values in Frequency and Time domain

Similar analyses were run to test if the β values estimated in Frequency domain and time domain was correlated. The scatter plot hardly illustrates any relationship among β values either (figure 3.6).

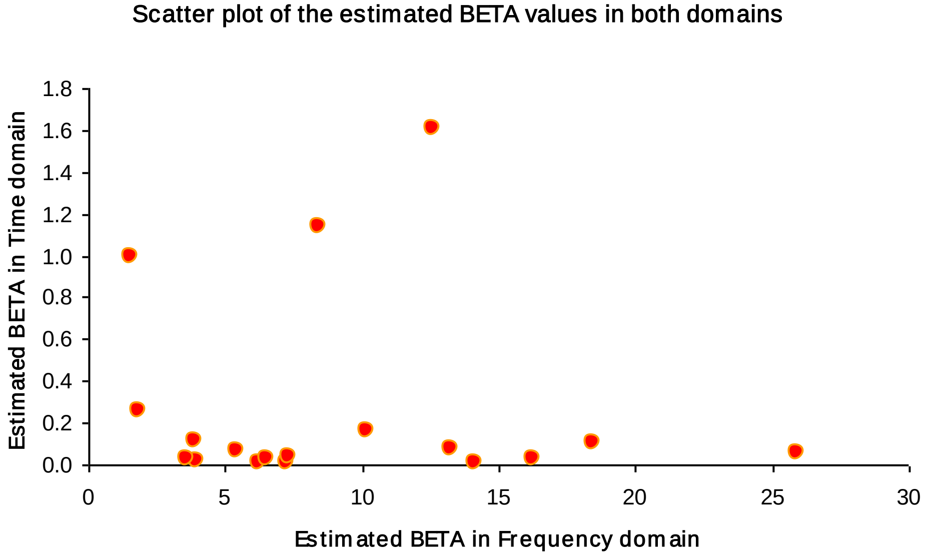
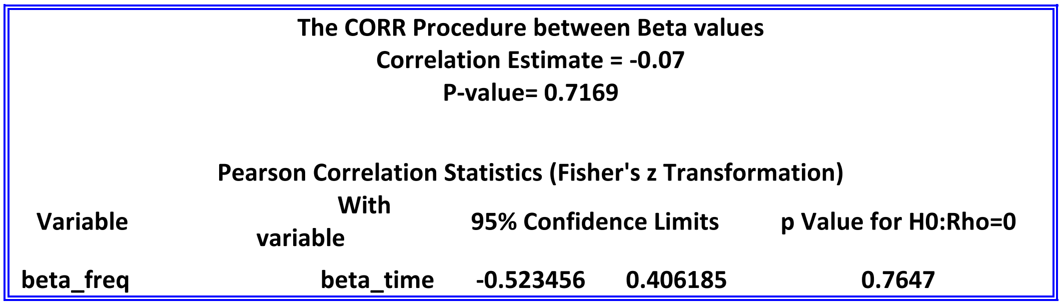


Figure 3.6: Scatter plot of the estimated BETA values in Frequency and time domain

The SAS out put of the correlation test between the estimated β values in the Frequency domain and the estimated β values in the time domain are displayed in table 3.6. In the CORR procedure the null hypotheses was not rejected within the 95% confidence correlation coefficient limit.



The CORR Procedure between Beta values
Correlation Estimate = -0.07
P-value= 0.7169

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With variable	95% Confidence Limits		p Value for H0:Rho=0
beta_freq	beta_time	-0.523456	0.406185	0.7647

Table 3.6: SAS output of the correlation test between the BETA values in Frequency and Time domain

3.4. A proposed solution to the time analysis problem

A logical and practical solution for the issue of analysing the recruitment time was to derive a time domain distribution directly from Poisson-gamma process. The application of the first principle of the relationship between frequency domain and time domain is that if the frequency domain follows Poisson-gamma mixture, the time to recruit patients will follow the Gamma-exponential mixture. The Gamma-exponential mixture is a special case of the Gamma-gamma mixture and it would end up with a similar expression to what Anisimov & Federov [3] had in their paper.

For the forecasting purpose, if the patient arrivals are simulated using the Poisson-gamma mixture, then the recruitment time are estimated by summing up the individual waiting times between patients until the last recruited patient in clinical centres.

In the proposed Bayesian based simulation method, a prior distribution had to be found for the parameters fitted from the Negative Binomial distribution. Then, to precede the simulation, the posterior mean and variance of the estimated parameters in the frequency domain were required as well. The next chapter expands the practical obstacles of modelling the parameters for the Bayesian prior distribution. Finally, in chapter 5 it has been tried to come up with a feasible modelling solution for all the analytical issues throughout the research.

Chapter 4

Independence of Model Parameters

The key aspect for Bayesian updating was to have a suitable prior distribution of the true value of the parameters. However, the uncertainty about the value of the parameters was a major issue in forecasting the recruitment. A fundamental assumption was that the parameters α and β had to be independent in the frequency domain as well as the time domain so that they can be modelled individually. Otherwise, fitting them to a bivariate prior distribution could be very complicated.

In this chapter, the independency of the estimated parameters in the recruitment frequency domain was tested. Then, regardless of the issues in modelling the time domain (chapter 3), the relationship between shape and scale parameters estimated from Pearson VI distribution has also been analysed.

4.1. Parameters estimated from Recruitment frequency domain

A scatter plot was very informative and told most of the story about the relationship between parameters. Figure 4.1, illustrates a simple scatter plot and reveals graphically the correlation between estimated parameters α and β in the recruitment frequency data.

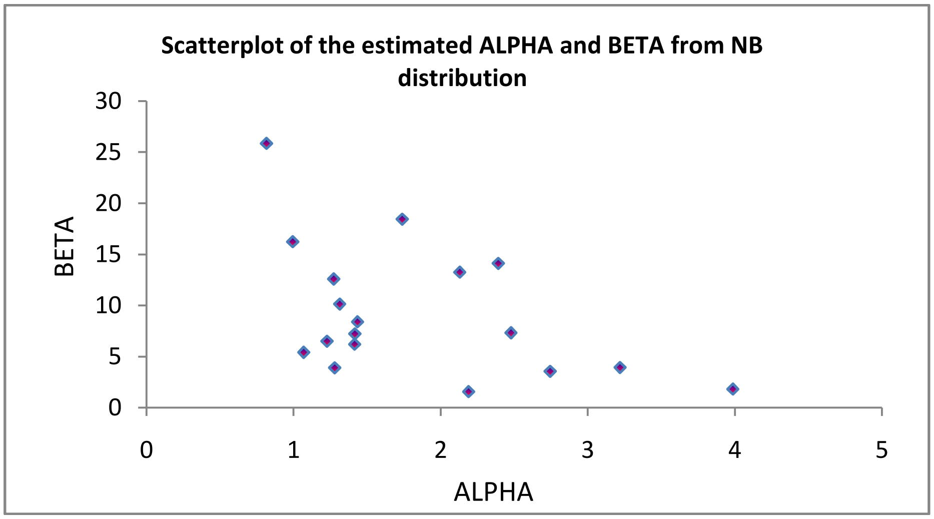


Figure 4.1: Scatter plot of estimated α and β from NB distribution in frequency domain

The scatter plot of α and β with their standard errors, however, gave a wider view of the relationship between parameters. If it is imagined that a horizontal whisker illustrates the standard error of α and a vertical line represents the standard error of β through each point in the scatter plot, it ends up having something like '+' around each point. As a result, the size of each point could indicate how reliable and informative the estimate was. The scatter plot of the shape parameter α and the scale parameter β, which were estimated by fitting the Negative Binomial distribution to the recruitment frequency domain, with one standard deviation has been visualised in figure 4.2. The green horizontal lines are the standard errors of the estimated parameter α and the blue vertical whiskers represent the standard errors of the scale parameter β in clinical trials. Obviously, the points with bigger standard errors are less informative and less reliable than others.

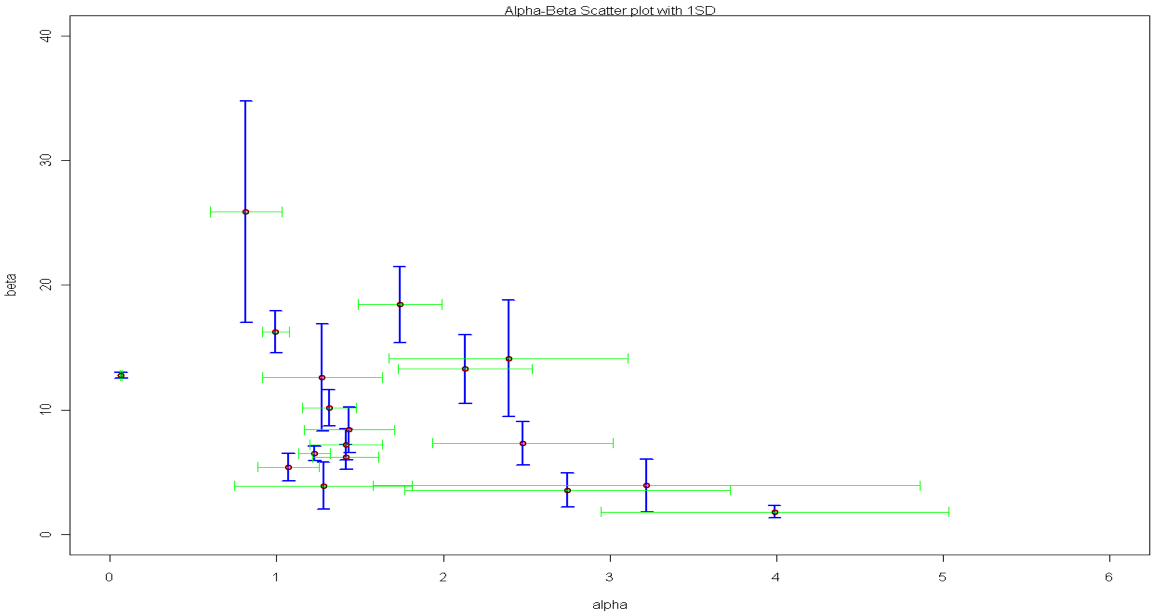


Figure 4.2: Scatter plot of α and β with 1*SE

The scatter plot in figure 5.2 illustrated only the between-trial correlations between the shape and the scale parameters. Overall, there were within trial correlation coefficients of the parameters which were negative (table 4.1 and 4.2) and a between trial coefficient that was also negative (figure 4.2 & table 4.3).

The within-trial correlation coefficients displayed in table 4.1 were calculated in SAS GENMOD procedure while estimating the parameters from fitting the Negative Binomial to the recruitment frequency data. This was, however, before taking the accrual time-spans into account. The table illustrates that there was quite a strong negative correlation coefficient between each pair of (α, β).

Study	ALPHA	BETA	corr-coefficient
1	1.273	12.599	-0.860
2	3.220	3.934	-0.943
3	2.392	14.125	-0.910
4	1.280	3.907	-0.860
5	2.745	3.556	-0.934
6	1.435	8.402	-0.866
7	1.417	7.225	-0.874
8	1.416	6.216	-0.878
9	2.478	7.334	-0.917
10	1.314	10.146	-0.855
11	1.739	18.449	-0.879
12	3.988	1.826	-0.964
13	0.995	16.245	-0.799
14	1.227	6.505	-0.850
15	2.191	1.567	-0.946
16	1.070	5.416	-0.840
17	0.816	25.873	-0.764
18	2.130	13.270	-0.899

Table 4.1: Estimated values of parameters from Negative Binomial and within trial correlation coefficient

It is also evidenced from the SAS output, which is displayed in table 4.2, that the mean of the within trial correlation coefficient in the frequency domain was not zero under the default 95% confidence limit.

```
                                  The SAS System
                             The UNIVARIATE Procedure
              Within-trial   correlation coefficient (alpha vs beta in each trial)
 N                  18                              Sum Weights            18
Mean               -0.88                            Sum Observations     -15.84
Std Deviation       0.051                           Variance                0.003
Skewness            0.40                            Kurtosis                0.24
Uncorrected SS  13.98                               Corrected SS            0.04
Coeff Variation  -5.85                              Std Error Mean          0.012
                              Tests for Location: Mu0=0
                Test                      -Statistic-               -----p Value------
            Student's t           t           -72.4724               Pr > |t|    <.0001
                Sign              M              -9                 Pr >= |M|   <.0001
            Signed Rank           S            -85.5                 Pr >= |S|   <.0001
```

Table 4.2: SAS output for the location test of the within trial correlation between parameters in frequency domain

The negative between trial correlations coefficient in the frequency domain has also been statistically documented in the SAS CORR procedure in table 4.3. The between trial correlation were calculated between the two vectors of the estimated parameters.

```
                         The SAS System
                       The CORR Procedure
             Between-trial correlation coefficient
               (Between the vectors of alpha and
                             beta)

                2  Variables:   ALPHA      BETA
Variable        N      Mean    Std Dev     Median    Minimum     Maximum
   ALPHA       18   1.84033    0.85517    1.42600    0.81600     3.98800
   BETA        18   9.25528    6.44033    7.27950    1.56700    25.87300
                        P_value = 0.0456
```

Table 4.3: SAS output of the between correlation coefficient of parameters among all the trials in the frequency domain

It was clear from the analysis that there were strong negative correlations between the parameters α and β that were estimated from Negative Binomial distribution in the recruitment frequency data. This implied that the parameters could not be modelled individually as a prior distribution for the Bayesian update. Therefore, the question raised here was whether a particular parameterisation could be found to make the estimates of α and β independent in some level. In this case, the new parameters had to be estimated and modelled for all the trials.

4.2. Parameter estimation from recruitment time domain

Despite the issue of incomparability of modelling the time domain from Pearson VI distribution (chapter 3), the correlation coefficients between shape and scale parameters were put under assessment. The scatter plots in figure 4.3 & 4.4 and also the statistical tests in table 4.5 revealed that the estimated parameters are highly correlated inside the trials. This time, however, the correlation coefficients between the shape and the scale parameters are strongly positive.

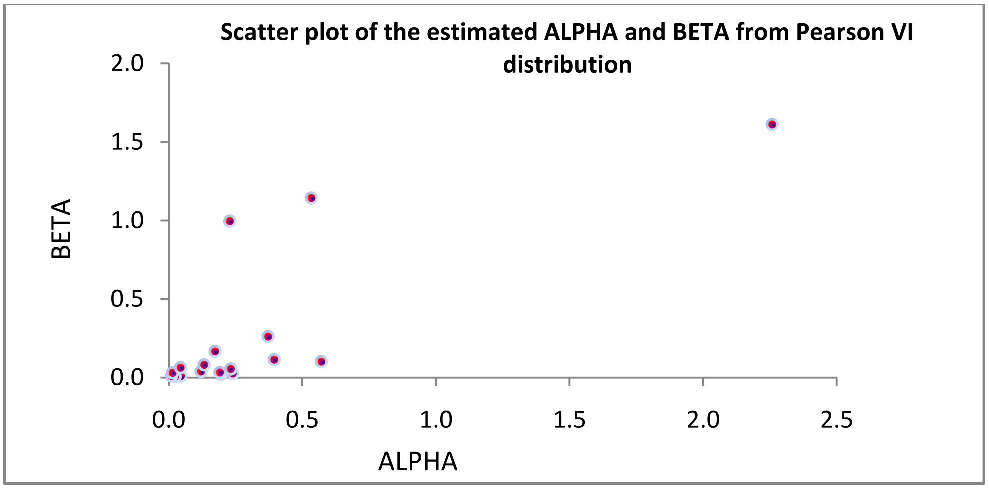


Figure 4.3: Scatter plot of Alpha and beta in time domain

Opposite the frequency domain, the estimated parameters in the time domain were less scattered and more informative. This was due to the smaller standard errors for each point (figure 4.4). Although the parameters estimated in the recruitment time data seemed to be more reliable, the quality of data provided in the frequency domain was more suitable for modelling patient recruitment. So, it is still aimed to analyse the recruitment time by moving from frequency domain to time domain.

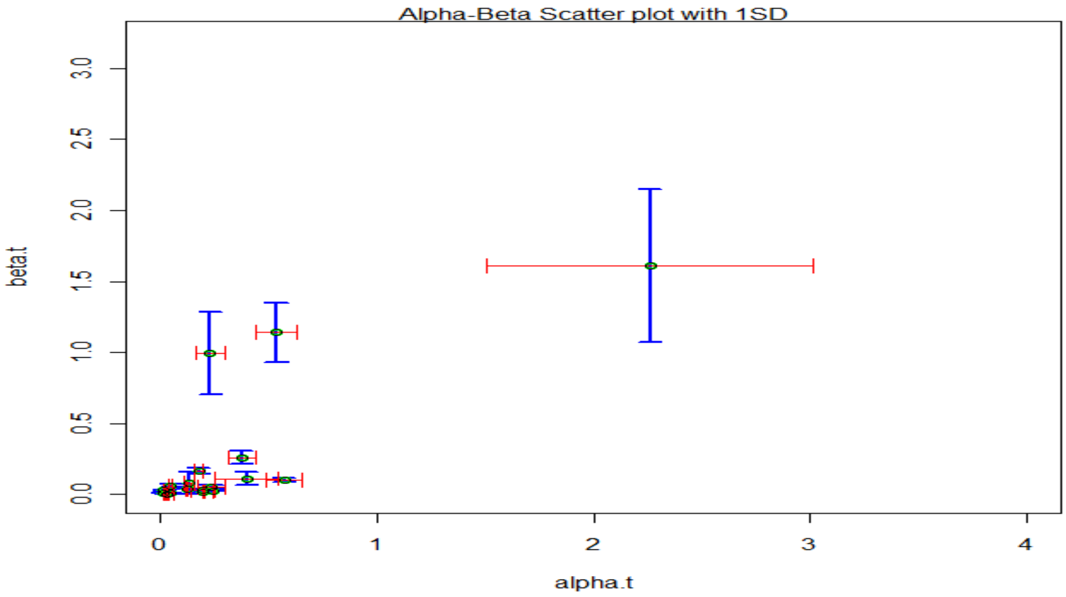


Figure 4.4: Scatter plot of ALPHA and BETA estimated from Pearson VI distribution with 1*Standard Deviation

The within trial correlation coefficients in table 4.4 were estimated in SAS NLP procedure after each Pearson VI parameter were estimated. The location test for the within trial correlation coefficient has been summarised in table 4.5, which shows a strong positive value of 0.87.

Study	ALPHA	BETA	corr-coefficient
1	2.260	1.610	0.996
2	0.196	0.019	0.146
3	0.046	0.009	0.798
4	0.396	0.113	0.964
5	0.192	0.032	0.950
6	0.534	1.142	0.993
7	0.024	0.006	0.909
8	0.011	0.009	0.853
9	0.121	0.039	0.420
10	0.174	0.167	0.991
11	0.571	0.102	0.996
12	0.373	0.260	0.992
13	0.240	0.025	0.962
14	0.014	0.029	0.958
15	0.229	0.996	0.969
16	0.045	0.063	0.942
17	0.233	0.054	0.960
18	0.133	0.081	0.969

Table 4.4: Estimated values of parameters from Pearson VI distribution and within trial correlation coefficient

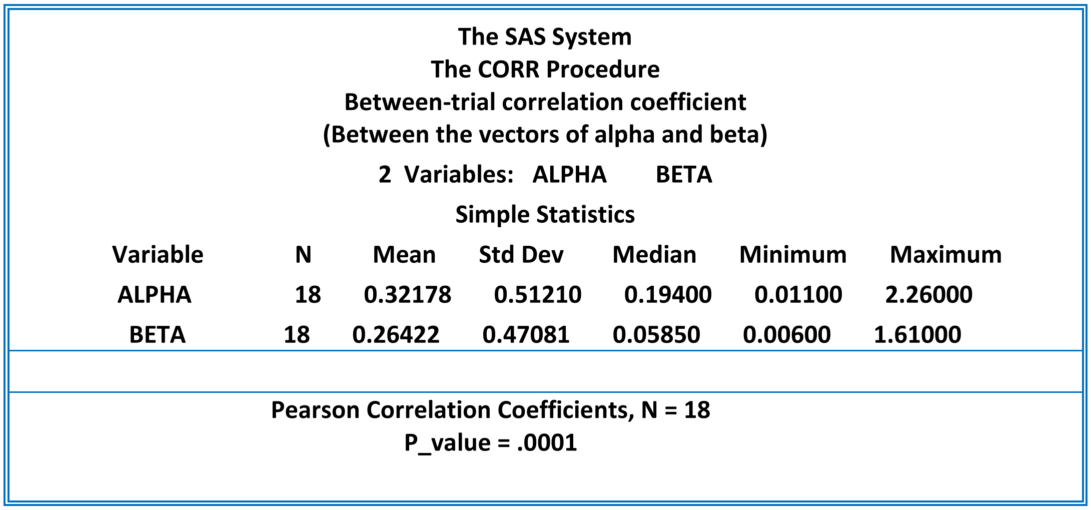
```
                         The SAS System
                    The UNIVARIATE Procedure
                Within-trial correlation coefficient
           (Between alpha and beta values in each trial)

                      Moments
   N                  18             Sum Weights          18
   Mean               0.88           Sum Observations     15.77
   Std Deviation      0.23           Variance          0.05
   Skewness          -2.68           Kurtosis          6.81
   Uncorrected SS     14.69          Corrected SS        0.87
   Coeff Variation    25.88          Std Error Mean      0.05

                 Tests for Location: Mu0=0
 Student's t                                       Pr > |t|    <.0001
 Signed Rank                                       Pr >= |S|   <.0001
```

Table 4.5: SAS output for the location test of the within trial correlation between parameters in time domain

The CORR procedure output, which is displayed in table 4.5, proved that there was also a high correlation-coefficient between the trials in time domain. The correlation of the between trial parameters was also positive.



```
                              The SAS System
                            The CORR Procedure
                    Between-trial correlation coefficient
                  (Between the vectors of alpha and beta)
                       2  Variables:   ALPHA      BETA
                            Simple Statistics
    Variable        N      Mean     Std Dev     Median     Minimum     Maximum
    ALPHA          18    0.32178    0.51210    0.19400    0.01100    2.26000
     BETA         18    0.26422    0.47081    0.05850    0.00600    1.61000


                 Pearson Correlation Coefficients, N = 18
                            P_value = .0001
```

Table 4.6: SAS output of the between correlation coefficient of parameters among all the trials in the time domain

4.3. Discussion

In the previous chapter it was decided to analyse the recruitment by moving smoothly from the recruitment frequency data to the time domain. This required the frequency parameters to be estimated and the distribution of the true parameters were to be found. Nevertheless, as it was highlighted formerly, the two Negative Binomial parameters had to be independent. In the next step, the issue was addressed by including the time spans into modelling the frequency domain. Simultaneously, two other parameters were estimated by fitting the Negative Binomial distribution to the recruitment frequency data. Therefore, the new independent parameters could be modelled individually to find a suitable prior distribution for the frequency data.

Chapter 5

Transformations and Predictions

5.1. Issues in modelling and solutions

The modelling and predicting process came across several issues, some of which were covered in previous chapters. In chapter 3, the shape parameters α and scale parameters β were estimated by fitting the Negative Binomial distribution to the recruitment frequency data only $(x \sim NB(\alpha, \beta))$. The frequency scale was later found to be broadly compatible when the time-spans of patient recruitment were included in the modelling. This update was applied by using 'offset' option in SAS modelling procedures. The offset variable was equalled to the patients' accrual time (the time unit was set to months throughout the research) in centres, which considered the time as a fixed factor in the null regression model.

The second problem appeared when the estimated parameters in the time domain (chapter 3) were far different from the theoretical expectations. Hence, the data that were provided were really only suitable for modelling the frequency domain, but the time domain was more important for practical forecasting. This issue was addressed by simulating the frequency data in the Poisson-gamma process and generating the waiting time between patients from the Gamma-exponential mixture.

The dependency of α and β especially in the frequency domain made the business of applying Bayesian methods more complex. Hence, it was not feasible to treat them as if they are independent for handling forecasting. Therefore, a transformation of parameters was desirable to make them independent [18]. To deal with the problem, two transformed parameters $k = \frac{1}{\alpha}$ and $\mu = \alpha\beta$, were considered for the estimation instead of (α, β).

To apply the Bayesian forecasting methods properly, further data of individual recruitment dates were required. The forecasting approach for waiting time between patients and also the length of time required to accrue patients in clinical centres were finally illuminated by applying the simulation method using a Poisson-gamma and Gamma- exponential mixture.

5.2. Parameter estimation

In addition to estimating two different parameters by fitting the Negative Binomial distribution (equation 3.3) to the recruitment frequency data, the time-spans have been included in the model by using 'offset' option in Genmod procedure in SAS. The new parameters were the dispersion parameter, which is $k = \frac{1}{\alpha}$ and the mean parameter which is in fact $\mu = \alpha\beta$; μ is the parameter representing the mean number of recruited patients per centre in the trial.

The GENMOD procedure in SAS® was applied to estimate the new parameters using the method of maximum likelihood. The results are displayed in table 5.1. The values of the estimated parameters are slightly different from the estimated parameters in table 3.1, which is due to including recruitment time into the Negative Binomial parameter estimation using ‘offset’ option.

Study	n	K = 1/ALPHA= dispersion	SE (K)	within trial variance (k)	mu	SE (mu)	within trial variance (mu)
1	385	0.789	0.238	0.057	13.658	3.026	9.155
2	152	0.274	0.145	0.021	11.838	2.138	4.569
3	811	0.406	0.122	0.015	30.960	4.522	20.451
4	80	0.673	0.295	0.087	4.065	1.109	1.229
5	244	0.351	0.127	0.016	9.088	1.304	1.701
6	796	0.687	0.130	0.017	8.679	1.296	1.680
7	1126	0.678	0.106	0.011	8.884	0.847	0.718
8	1241	1.000	0.135	0.018	8.802	0.671	0.451
9	927	0.395	0.087	0.007	16.187	1.700	2.889
10	2000	0.754	0.093	0.009	10.669	0.988	0.975
11	2936	0.574	0.083	0.007	30.304	2.521	6.353
12	546	0.250	0.065	0.004	5.634	0.523	0.273
13	4363	0.986	0.083	0.007	12.882	0.995	0.991
14	3274	0.781	0.061	0.004	4.851	0.369	0.136
15	103	0.528	0.227	0.051	0.903	0.569	0.324
16	533	0.894	0.154	0.024	3.367	0.610	0.372
17	549	1.226	0.321	0.103	19.475	4.675	21.859
18	1696	0.469	0.088	0.008	25.364	2.592	6.721

Table 5.1: Estimated transformed parameters from fitting Negative Binomial to the frequency data and including time-spans to the model

5.3. Correlation test for the transformed parameters in frequency domain

The transformation of the estimated parameters was beneficial only if the parameters end up being independent. In this case, it is possible to model them individually to find a prior distribution for the parameters. There were two approaches to look at independence. The first was combining the evidence from each trial as regards the within-study correlation test. The second was examining the pairs of parameter estimates from each trial to look at the between trial correlation. It is clear from table 5.2 that the within-study correlation coefficients between the estimated values of parameters (k, μ) are remarkably reduced compared to the correlation coefficient between α and β in table 4.1. Of course, including an offset might in any case produce some change but this is not the explanation here.

Study	n	1/ALPHA= k= dispersion	mu	corr-coefficient (k & mu)	cov (k & mu)
1	385	0.789	13.658	0.0002	0.00013
2	152	0.274	11.838	-0.0101	-0.00313
3	811	0.406	30.960	-0.0010	-0.00054
4	80	0.673	4.065	-0.0234	-0.00765
5	244	0.351	9.088	-0.0024	-0.00040
6	796	0.687	8.679	-0.0009	-0.00015
7	1126	0.678	8.884	-0.0029	-0.00026
8	1241	1.000	8.802	0.0000	0.00000
9	927	0.395	16.187	-0.0011	-0.00016
10	2000	0.754	10.669	-0.0005	-0.00005
11	2936	0.574	30.304	0.0000	-0.00001
12	546	0.250	5.634	-0.0001	0.00000
13	4363	0.986	12.882	-0.0009	-0.00008
14	3274	0.781	4.851	-0.0038	-0.00009
15	103	0.528	0.903	0.0306	0.00395
16	533	0.894	3.367	-0.0055	-0.00051
17	549	1.226	19.475	0.0000	0.00002
18	1696	0.469	25.364	0.0000	0.00000

Table 5.2: Table of the correlation-coefficient and covariance between estimated parameters *k* and *mu* from Negative Binomial with adding time-spans in the model

As regards the first matter, the statistical correlation test between parameters (k, μ) in trials suggested that it was possible to treat them as two independent parameters. Table 5.3 tests the 18 within-study correlation coefficients to see if there is any evidence that on average they are different from zero. The summary of results is displayed below.

```
                          The SAS System
                    The UNIVARIATE Procedure
                        Variable:  corr
     (Within-trial correlation coefficient between k and mu)
-------------------------------------------------------------------
                   Tests for Location: Mu0=0
                       p Value= 0.6057
```

Table 5.3: SAS output of the within trial correlation coefficient test between parameters k & mu

The second matter was also addressed by testing the independence between pairs of estimated parameters in trials. The Pearson correlation coefficient test (table 5.4) did not reject the independence of the parameters among the trials either, which meant that two issues have been solved simultaneously. Adding the time-spans (accrual time) into the Negative Binomial model tuned the estimations into more compatible results. At the same time, a transformation in parameters made them independent from each other.

```
                           The SAS System
                         The CORR Procedure
                   Between-trial correlation coefficient
                        2  Variables:    k       mu
                           Simple Statistics
Variable        N       Mean      Std Dev        Sum
     k          18      0.65       0.27          11.71
    mu          18     12.53       8.87          225.61
-------------------------------------------------------------------

-------------------------------------------------------------------
                           Pearson Correlation Coefficients
                                      P_value=  0.6604
```

Table 5.4: SAS output of the correlation test between parameters k & mu among the trials

5.4. Prior distribution of parameters

Based on the literature review, the most feasible marginal distribution for the recruitment frequency data was the Negative Binomial distribution. Then, to proceed to the Bayesian prediction method, the true values of the parameters were to be modelled. The true values are different from what would be estimated. The true parameters are what are observed if each trial had an infinitive number of centres with an infinitive number of patients. Because this is not the case, the variability of the observed parameters is bigger than the true parameters. In addition to that, the initial parameters α and β were previously found to be dependent. Hence, the transformed parameters k and μ were considered for the maximum likelihood estimation.

The parameters k and μ were modelled individually to see if they follow the Normal distribution. The Normal distribution although not ideal could be a suitable approximate prior distribution for the Negative Binomial parameters given an appropriate transformation of the parameters. If the Normal distribution was a good fit for parameters, then estimating the posterior values of the parameters would be easier. The result of testing how well the Normal distribution fits to the to the Negative Binomial parameters k and μ showed that it was a practical fit. The SAS univariate outputs are displayed in table 5.5 for the parameter k and table 5.6 for μ. All the three statistical GOF tests approved the assumptions of the Negative Binomial parameters following the Normal distribution.

```
                    The SAS System
              The UNIVARIATE Procedure
               Fitted Distribution for k

           Parameters for Normal Distribution
              Parameter   Symbol   Estimate
                Mean        Mu       0.651
               StdDev     Sigma     0.271

      Goodness-of-Fit Tests for Normal Distribution
                  Test              -----p Value-----
Kolmogorov-Smirnov   >0.150
Cramer-von Mises     >0.250
Anderson-Darling     >0.250
```

Table 5.5: SAS output of the normality test of the parameter k

```
                    The SAS System
              The UNIVARIATE Procedure
               Fitted Distribution for mu

           Parameters for Normal Distribution
              Parameter   Symbol   Estimate
                Mean        Mu       12.53
               StdDev     Sigma      8.87

      Goodness-of-Fit Tests for Normal Distribution
                  Test              -----p Value-----
Kolmogorov-Smirnov   >0.150
Cramer-von Mises     0.065
Anderson-Darling     0.051
```

Table 5.6: SAS output of the normality test of the parameter mu

The normality test was also run for Ln(k). It was to apply *Ln (k)* instead of *k* in the simulation process in order to avoid generating negative *k* values. The parameter *k* in the Negative Binomial distribution gets positive values only. However, the estimated values of the parameter *k* could get negative values in the process of generating the Normal distribution. For that reason, the parameter *Ln (k)* was initially applied. There was no such consideration for parameter μ since its mean was big enough compared to its standard deviation in each trial.

```
                    The SAS System
               The UNIVARIATE Procedure
              Fitted Distribution for Ln_k

          Parameters for Normal Distribution
             Parameter   Symbol   Estimate
                Mean       Mu      -0.51
                StdDev    Sigma     0.43

        Goodness-of-Fit Tests for Normal Distribution
              Test                -----p Value-----
 Kolmogorov-Smirnov                     >0.150
 Cramer-von Mises                       >0.250
 Anderson-Darling                       >0.250
```

Table 5.7: SAS output of the normality test of Ln (k)

Among all the 18 completed trials, one trial was chosen to illustrate the calculation of the posterior distribution. The prediction process relied on the posterior values of the selected trial. To calculate the posterior values of the trial it was necessary to estimate the variances between the 18 completed trials. In theory the relationship between the variances of parameters is:

Total variance = between trial variance + within trial variance (5.1)

The within trial variances for parameters *k* and μ (displayed in table 5.1 and table 5.8) were estimated in the SAS GENMOD procedure in Fitting the Negative Binomial distribution to the frequency data. The within trial variances for *Ln (k)*, however, were calculated using the equation:

$$\operatorname{var}(\ln(k)) \cong \frac{1}{k^2}\operatorname{var}(k) \tag{5.2}$$

The above is an approximate formula based on Taylor’s expansion of Ln(k) using the so-called ‘delta method’.

Study	mu	SE (mu)	within trial variance (mu)	Ln_k	Se (Ln_k)	within trial variance (Ln_k)
1	13.658	3.026	9.155	-0.237	0.302	0.091
2	11.838	2.138	4.569	-1.295	0.529	0.280
3	30.960	4.522	20.451	-0.902	0.301	0.091
4	4.065	1.109	1.229	-0.396	0.439	0.193
5	9.088	1.304	1.701	-1.048	0.361	0.130
6	8.679	1.296	1.680	-0.376	0.189	0.036
7	8.884	0.847	0.718	-0.389	0.156	0.024
8	8.802	0.671	0.451	0.000	0.134	0.018
9	16.187	1.700	2.889	-0.929	0.219	0.048
10	10.669	0.988	0.975	-0.282	0.123	0.015
11	30.304	2.521	6.353	-0.555	0.145	0.021
12	5.634	0.523	0.273	-1.388	0.262	0.069
13	12.882	0.995	0.991	-0.014	0.084	0.007
14	4.851	0.369	0.136	-0.247	0.078	0.006
15	0.903	0.569	0.324	-0.638	0.429	0.184
16	3.367	0.610	0.372	-0.112	0.172	0.030
17	19.475	4.675	21.859	0.204	0.262	0.069
18	25.364	2.592	6.721	-0.756	0.188	0.035

Table 5.8: Estimated parameters mu and Ln (k) with their within-trial standard errors and variances

The observed (total) variance of the parameters could be calculated while modelling the parameters as a normal distribution. From table 5.5 the total variance of parameter *k* is $(0.271)^2 = 0.073$, based on table 5.6 the total variance of parameter μ is $(8.873)^2 = 78.730$ and according to the table 5.7 the total variance of parameter *Ln (k)* is $(0.433)^2 = 0.187$.

5.5. Random Effect meta-analysis [19, 20]

The general purpose of meta-analysis is to estimate the true effect size of the parameters taken from a study under special assumptions and conditions. The meta-analysis combines the results of the completed trials and gives the effect size of the parameters as its output. Since the parameters are independent, the one dimensional random effect analysis would work efficiently. That is to say, a separate meta-analysis was applied to both *Ln (k)* and μ.

The random effect meta-analysis macro in SAS [19,20] takes the estimated values as well as the estimated within-trial standard deviations of the parameter in all trials and calculates the posterior estimation of the parameter based on the normal distribution. It also estimates the random effect variance of the parameter. The result of the random effect meta-analysis of parameters *k*, Ln *(k)* and μ are summarised in tables 5.9, 5.10 and 5.11 respectively. The posterior variances of the parameters, however, were calculated based on the assumption of normal distribution of parameters and applying the equation:

$$Posterior.\mathrm{var}iance = \frac{1}{\frac{1}{R.E.Variance} + \frac{1}{within.trial.\mathrm{var}iance}} \tag{5.3}$$

Meta-analysis, Der Simonian& Laird method				
Study	1/ALPHA= k= dispersion	within trial variance (k)	Posterior estimation of k	Posterior variance of k
1	0.789	0.057	0.707	0.028
2	0.274	0.021	0.372	0.015
3	0.406	0.015	0.453	0.012
4	0.673	0.087	0.645	0.034
5	0.351	0.016	0.413	0.012
6	0.687	0.017	0.673	0.013
7	0.678	0.011	0.669	0.009
8	1.000	0.018	0.908	0.014
9	0.395	0.007	0.423	0.007
10	0.754	0.009	0.737	0.007
11	0.574	0.007	0.580	0.006
12	0.250	0.004	0.277	0.004
13	0.986	0.007	0.946	0.006
14	0.781	0.004	0.772	0.003
15	0.528	0.051	0.576	0.027
16	0.894	0.024	0.813	0.017
17	1.226	0.103	0.835	0.036
18	0.469	0.008	0.489	0.007
Estimated mean of *k* from meta-analysis				0.627
R.E Variance of the parameter *k*				0.055

Table 5.9: SAS results of the random effects meta-analysis for parameter *k*

Meta-analysis, Der Simonian& Laird method				
Study	**Ln_k**	**within trial variance (Ln_k)**	**Posterior estimation of Ln_k**	**Posterior variance of Ln_k**
1	**-0.237**	**0.091**	**-0.342**	**0.044**
2	**-1.295**	**0.280**	**-0.642**	**0.066**
3	**-0.902**	**0.091**	**-0.665**	**0.044**
4	**-0.396**	**0.193**	**-0.427**	**0.059**
5	**-1.048**	**0.130**	**-0.682**	**0.052**
6	**-0.376**	**0.036**	**-0.395**	**0.025**
7	**-0.389**	**0.024**	**-0.400**	**0.019**
8	**0.000**	**0.018**	**-0.076**	**0.015**
9	**-0.929**	**0.048**	**-0.754**	**0.031**
10	**-0.282**	**0.015**	**-0.306**	**0.013**
11	**-0.555**	**0.021**	**-0.532**	**0.017**
12	**-1.388**	**0.069**	**-0.967**	**0.038**
13	**-0.014**	**0.007**	**-0.047**	**0.007**
14	**-0.247**	**0.006**	**-0.260**	**0.006**
15	**-0.638**	**0.184**	**-0.504**	**0.059**
16	**-0.112**	**0.030**	**-0.197**	**0.022**
17	**0.204**	**0.069**	**-0.082**	**0.038**
18	**-0.756**	**0.035**	**-0.664**	**0.025**
Estimated mean of Ln(k) from meta-analysis				**-0.441**
R.E Variance of parameter Ln(k)				**0.086**

Table 5.10: SAS results of the random effects meta-analysis for parameter Ln (k)

Meta- analysis , Der Simonian& Laird method				
Study	mu	within trial variance (mu)	Posterior estimation of mu	Posterior variance of mu
1	13.658	9.155	12.921	6.300
2	11.838	4.569	11.739	3.726
3	30.960	20.451	21.066	10.163
4	4.065	1.229	4.475	1.159
5	9.088	1.701	9.261	1.569
6	8.679	1.680	8.880	1.551
7	8.884	0.718	8.963	0.693
8	8.802	0.451	8.854	0.441
9	16.187	2.889	15.577	2.527
10	10.669	0.975	10.699	0.930
11	30.304	6.353	25.751	4.833
12	5.634	0.273	5.705	0.269
13	12.882	0.991	12.806	0.944
14	4.851	0.136	4.893	0.135
15	0.903	0.324	1.064	0.319
16	3.367	0.372	3.513	0.365
17	19.475	21.859	15.225	10.499
18	25.364	6.721	21.849	5.043
Estimated mean of mu from meta-analysis				11.291
R.E Variance of the parameter mu				20.202

Table 5.11: SAS results of the random effects meta-analysis for parameter mu

5.6. Posterior estimations for the selected trial

The Bayesian prediction method was applied on one completed trial. The chosen trial had special information that was essential for the forecasting process. The study included the details about the individual recruitment date as well as each centre's activation date. Therefore, the waiting time between patients in the whole trial was available. The trial had recruited 152 patients in 12 centres in just less than 4 months.

The trial, then, was monitored in four different time intervals. After each interval a new dataset was produced and the Negative Binomial parameters in the frequency domain were estimated for every new data set. The first dataset included the number of patients that had been recruited by the first month after the first centre had been activated. The second data sheet had the details about the patients in clinical centres just 2 months after the first site had been activated. In the third data set, there was the number of patients that were recruited by the third month of the start of the trial. Finally the last data set contained the total recruitment details of the trial by the time it had finished, when all the clinical centres had entered the trial. There were 4 data sets in total and the parameters were estimated by fitting the Negative Binomial distribution to the frequency domain.

The recruitment had not initiated simultaneously in all centres. In other word, the clinical centres had been activated in different dates. Consequently, in each data set, the recruitment times varied among centres and depend on the activation date. The clinical centres were added to the datasets one by one. Only the fourth data set included all the centres in the trial.

Table 5.12 illustrates the number of patients that had been recruited in the clinical centres as well as their accrual time in the four data sets.

		By the end of the 1st month (15/04/)		By the end of the 2nd month (15/05/)		By the end of the 3rd month (15/06/)		BY the end of the trial (15/07/)	
site	site activation date	patients	Accrual time-month	patients	Accrual time-month	patients	Accrual time-month	patients	Accrual time-month
1	15/03/	6	0.99	14	1.89	23	3.05	26	3.36
2	16/03/	2	0.89	5	1.79	6	2.52	6	2.52
3	16/03/	1	0.89	5	1.83	12	3.03	18	3.89
4	16/03/	7	0.99	18	1.90	25	3.03	28	3.79
5	30/05/					1	0.46	10	1.49
6	23/06/							11	0.70
7	30/05/							2	1.42
8	19/04/			4	0.80	5	1.44	6	2.37
9	19/04/			1	0.73	5	1.89	6	2.76
10	19/04/			2	0.79	5	1.19	9	2.86
11	19/04/			3	0.76	10	1.66	17	2.77
12	06/06/					4	0.29	13	1.27
Patientsrecruited		16		52		96		152	
Average recruitment per centre per month		4.00		3.25		3.20		3.25	

Table 5.12: The four recruitment data set made from the selected trial in four time intervals monitoring

In the next step the posterior mean values and the posterior variances of the parameters were calculated in all the four data sets derived from the trial.

The posterior estimations of the parameters were based on the assumption that the parameters follow the normal distribution. Equation (5.3) was applied to calculate the posterior variances of the parameters.

Estimated parameters and the posterior values				
	1st data set	2nd data set	3rd data set	4th data set
number of patients	**16**	**52**	**96**	**152**
parameter k	**0.187**	**0.463**	**0.423**	**0.290**
se (k)	**0.339**	**0.301**	**0.233**	**0.151**
within trial variance (k)	**0.115**	**0.090**	**0.054**	**0.023**
Ln_k	**-1.675**	**-0.770**	**-0.860**	**-1.239**
se (Ln_k)	**1.810**	**0.650**	**0.551**	**0.520**
within trial variance (Ln_k)	**3.278**	**0.422**	**0.303**	**0.271**
Posterior estimation (Ln_k)	**-0.473**	**-0.497**	**-0.534**	**-0.633**
Posterior variance (Ln_k)	**0.084**	**0.071**	**0.067**	**0.065**
parameter mu (estimated mean number of patients in centres after each interval)	**4.016**	**5.854**	**8.501**	**11.667**
Estimated mean number of patients per centre per month (mu/time interval)	**4.02**	**2.93**	**2.83**	**2.99**
se (mu)	**1.318**	**1.742**	**2.119**	**2.199**
within trial variance (mu)	**1.737**	**3.033**	**4.490**	**4.836**
posterior estimation (mu)	**4.592**	**6.564**	**9.009**	**11.595**
Posterior estimation of the average recruitment per centre per month (mu/time interval)	**4.59**	**3.28**	**3.00**	**2.97**
posterior variance (mu)	**1.599**	**2.637**	**3.674**	**3.902**

Table 5.13: Summary of the estimated parameters and the posterior values for the selected trial in four time intervals

The posterior mean of the parameters were calculated from substituting the values in the equations (5.4) and (5.5).

$$Posterior.mean(\mu)=\frac{\left(\frac{1}{R.E.\sigma^{2}(\hat{\mu})}.mean(\hat{\mu})\right)+\left(\frac{1}{within.trial.\sigma^{2}(\hat{\mu})}.\hat{\mu}\right)}{\left(\frac{1}{R.E.\sigma^{2}(\hat{\mu})}+\frac{1}{within.trial\,\sigma^{2}(\hat{\mu})}\right)} \qquad (5.4)$$

$$Posterior.mean\left(\ln(k)\right)=\frac{\left(\frac{1}{R.E.\sigma^{2}(\ln(\hat{k}))}.mean(\ln(\hat{k}))\right)+\left(\frac{1}{within.trial.\sigma^{2}\left(\ln(\hat{k})\right)}.\ln(\hat{k})\right)}{\left(\frac{1}{R.E.\sigma^{2}\left(\ln(\hat{k})\right)}+\frac{1}{within.trial\,\sigma^{2}\left(\ln(\hat{k})\right)}\right)} \qquad (5.5)$$

In calculating the posterior values, $R.E.\sigma^{2}(.)$ is the random effect variance of the parameters among the 18 completed trials and $mean(.)$ is the estimated mean of the parameters in the trials derived from meta-analysis. The $within.trial.\sigma^{2}(.)$, in the second part of the equations, is the within trial variance of the parameter, which was estimated from fitting the Negative Binomial distribution to the frequency data. The estimated parameters in the four time interval data sets and the within trial variances of the parameters are displayed in table (5.13).

5.7. Prediction of patient recruitment

The aim of the current simulation was to use the data from completed trials in the frequency domain and forecast the patient recruitment in the time domain in multi-centre clinical trials. In practice, predicting the accrual time for the pre-arranged sample size was more important than predicting the number of patients in the clinical trials. However, the data provided was more suitable for working on the frequency domain. This research, however, managed to simulate the patients accrual times (in month) in an on-going trial by using the frequency domain in completed trials.

Regarding the prediction approach, it was to look at the predictions that were estimated from the trial simulations and compare them with the real trial.

In section 5.7.1 we discuss the theoretical problem of predicting how long it would take to run a second 152 patient trial given the information available at month 1, 2, 3 and 4. For example, suppose we need to conduct a second similar trial for regulatory submission to satisfy the regulatory needs of replication. If the planning phase occurs whilst the initial study is going on, we investigate how long it will take to recruit 152 patients in a new study using the 1, 2 and 3 month data. The more interesting practical question of concern is how long it will take to complete the on-going trial which is discussed in section 5.7.2. That is, at month 1 how long will it take to recruit the additional 136 patients, at month 2 how long will it take to recruit the additional 100 patients etc.

The predictions were run 1000 times in R version 2.12.1. The simulations were initially based on the posterior information of the four datasets derived from the selected trial.

The selected trial had recruited 152 patients in 12 clinical centres. In order to make a similar trial, N=12 clinical centres were simulated for the trial. However, taking the safe side, 152 individuals were simulated for each centre, which makes a trial of n=12*152 patients. Although n=152 patients of the whole trial would be used only. It did not affect the predicting process since the n patients are considered by the order they arrived to the trial regardless of the centres.

5.7.1. Prediction in the frequency domain

Each centre had one pair of estimated parameters $(\ln(k), \mu)$ and it was supposed that they follow the Normal distribution. Hence, for each data set, the simulation approach was to generate N=12 random values of $\ln(k)$ and μ with the estimated posterior mean values and posterior variances displayed in table (5.13). Random values of *k* were derived by taking *exp(ln(k))* from the generated $\ln(k)$.

The next step for the four data sets was to generate n=152 random numbers (patients) from the Negative Binomial distribution with the parameters (k, μ) generated for each centre. Although, the total number of patients in the whole trial was 152, this amount was generated for each centre to be able to predict recruitment time for a second trial of the same size. So far the simulated trials of the four data sets have produced matrices of N=12 rows (centres) and n=152 frequencies, which represented the number of patients that have been recruited in clinical centres. Consequently, the overall mean of each matrix divided by their monitoring interval times gave the mean number of patients per centre per month in the simulated trial.

According to the data set collected by the end of 1st month of the trial (table 5.12), 4 patients in average were recruited in each centre during the first month of the trial. When the trial went on to its second month four new centres were activated as well as more patients were recruited. But the average number of patients reduced to 3.25 per centre per month. Another slight reduction of 0.05 in the average number of recruited patients appeared by the end of 3rd month. The average number of accrued patients per centre per month slightly recovered to reach to 3.62 by the end of the trial. Figure 5.1 illustrates the average recruitment trend per centre per month in the completed trial.

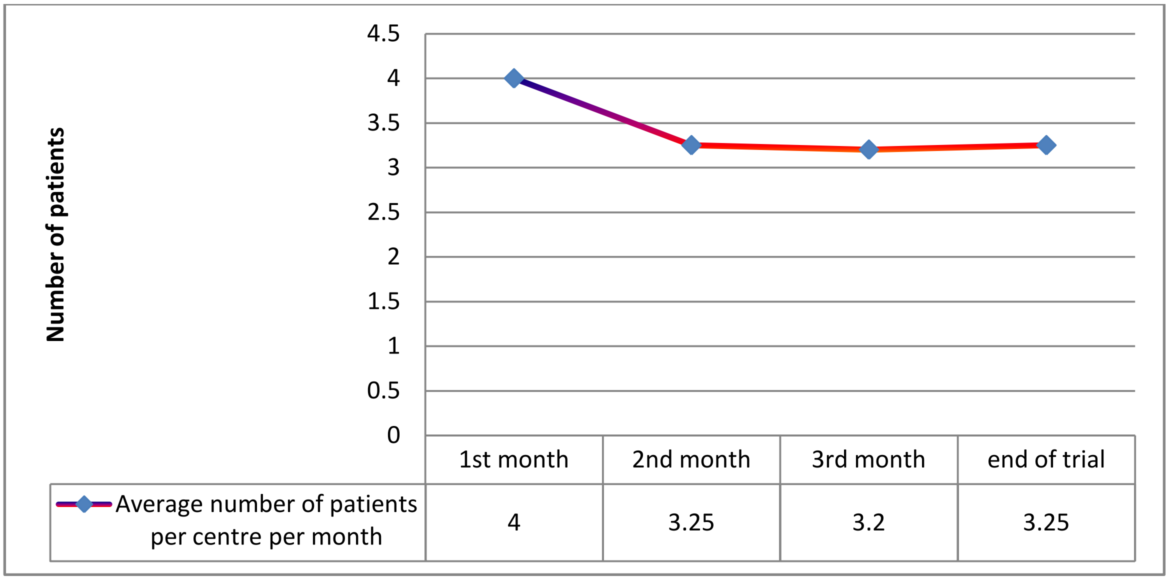


Figure 5.1: Line graph of the average number of recruited patients per centre per month in the completed trial througout the four monitoring intervals.

Figure 5.2 shows the histogram of the predicted average number of patients per centre per month based on the one-month information simulated trial. The histogram says that having had the recruitment data of the first month only, it is expected to recruit the average number of 4.09 to 5.13 (95% CI: 4.43-5.04) patients per centre per month throughout the trial. Although the value of the real trial is very close to the lower band of the histogram, the 95% confidence interval does not cover the values of the real trial. It means that the simulated trial has slightly over estimated the recruitment.

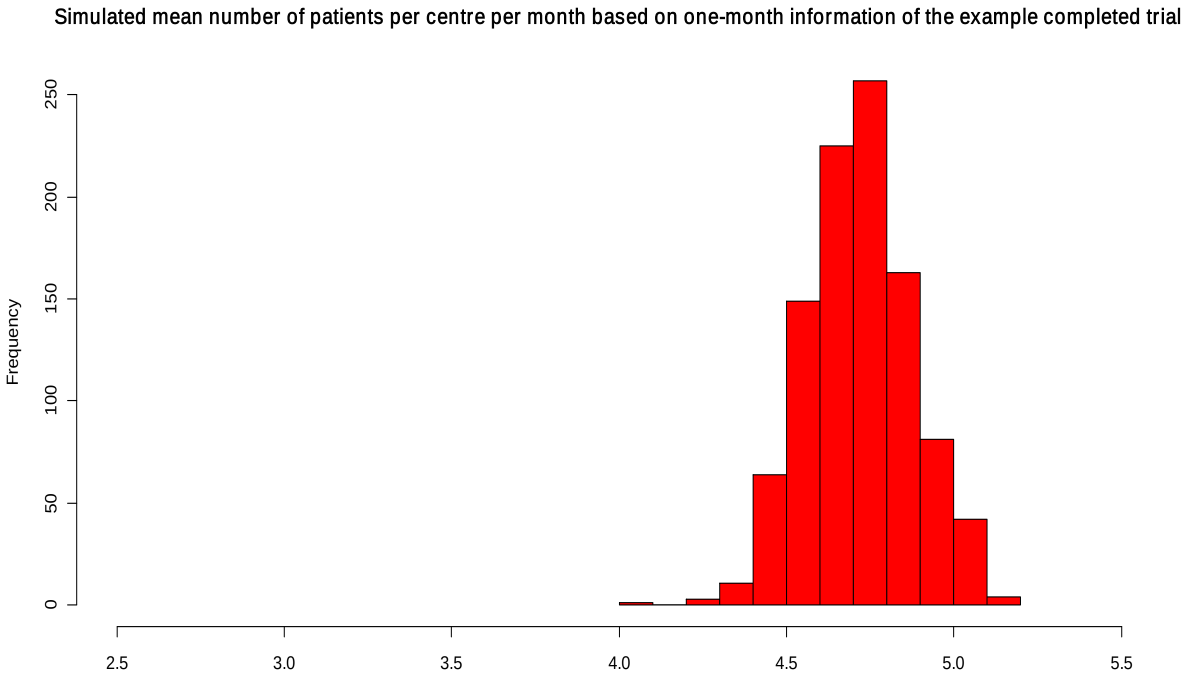


Figure 5.2: Average number of recruited patients per centre per month based on one month monitoring

The prediction on the frequency domain changed slightly after the trial finished the second month of recruiting patients. Based on the two months information in an on-going trial, the simulated average number of patients was a number between 3.02 and 3.73 (95% CI: 3.13-3.57) patients per centre per month (Figure 5.3). The prediction is perfectly matched to the values of the completed trial.

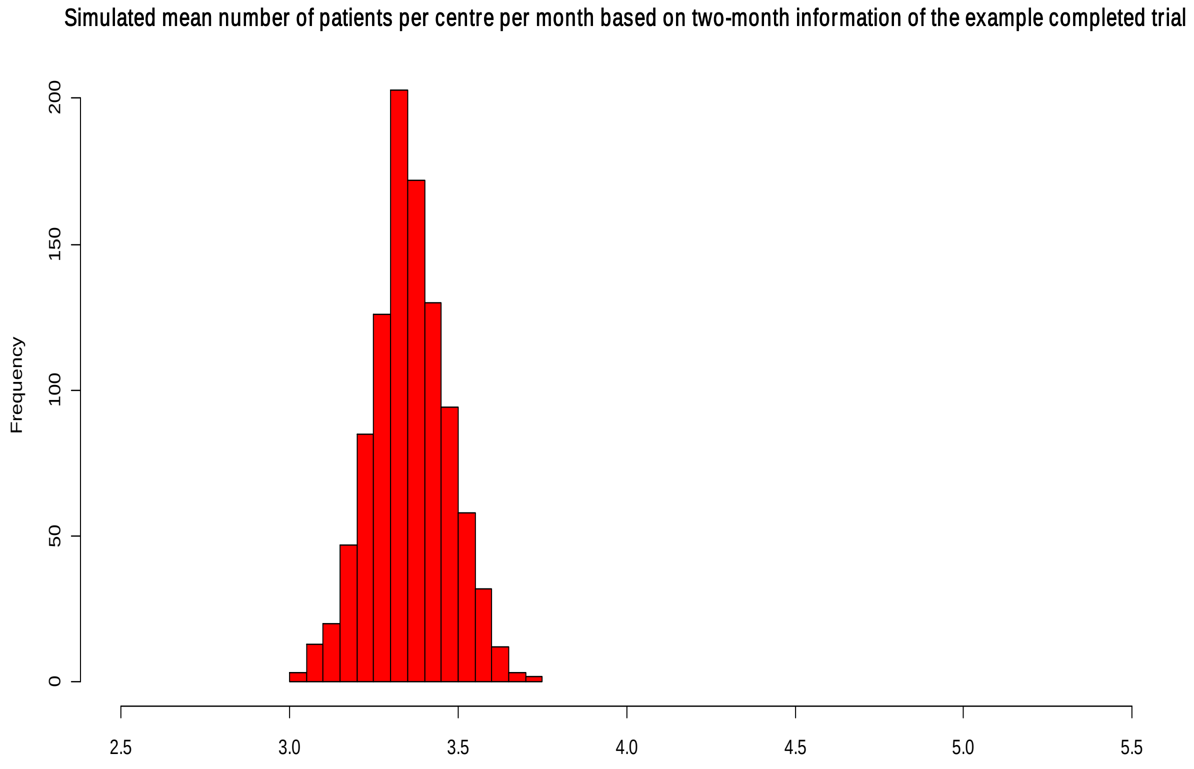


Figure 5.3: Average number of recruited patients per centre per month based on two months monitoring

The expected frequency of patients per centre per month is predicted to be between 2.71 and 3.60 (95% CI: 2.90-3.31) based on the simulated trial of the three-month data set in an ongoing trial (figure 5.4). The confidence interval included the average value of the real trial in the third data set (3.20) as well as the last data set (3.24).

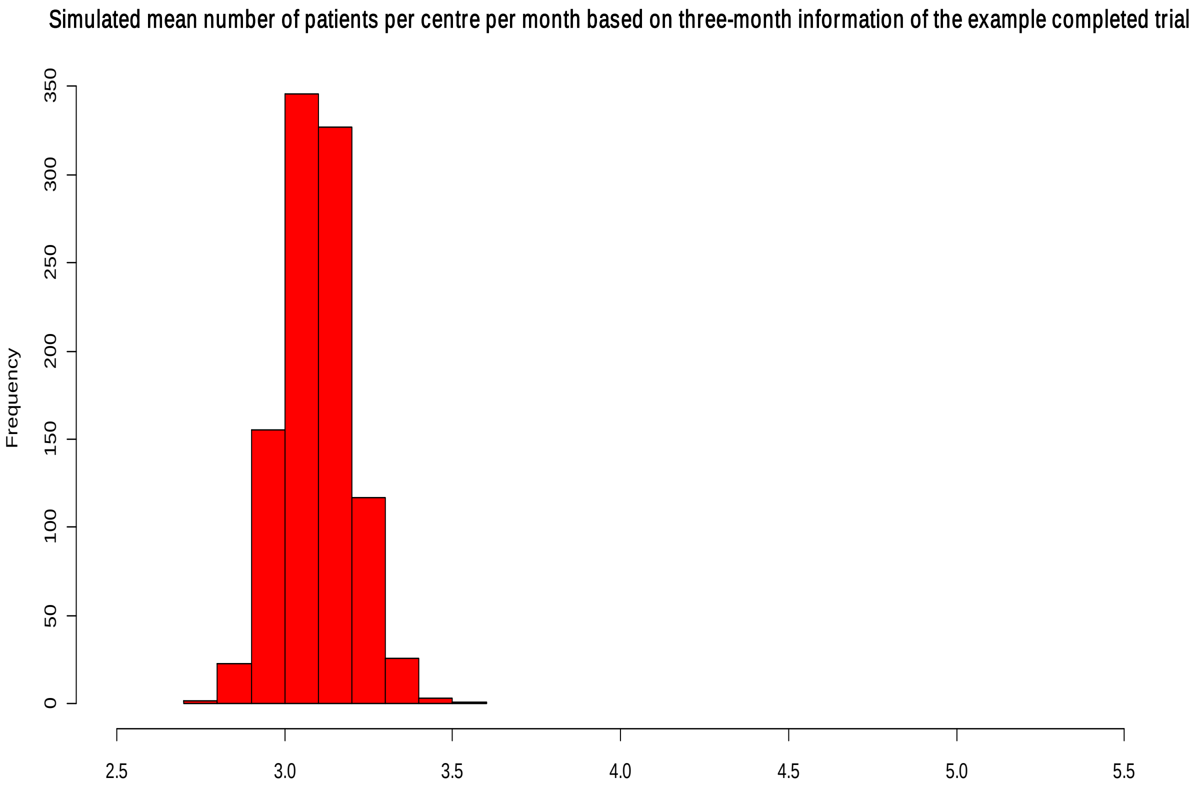


Figure 5.4: Average number of recruited patients per centre per month based on three months monitoring

By the end of the trial, the average number of patients per centre per month in the example trial were about 3.25. It was located inside the 95% confidence interval of the predicted average number of patients. In the simulated trial,derived after the trial ended, the expected range was between 2.60 and 3.39 (95% CI: 2.80-3.27). The histogram of the final simulation is displyed in figure 5.5.

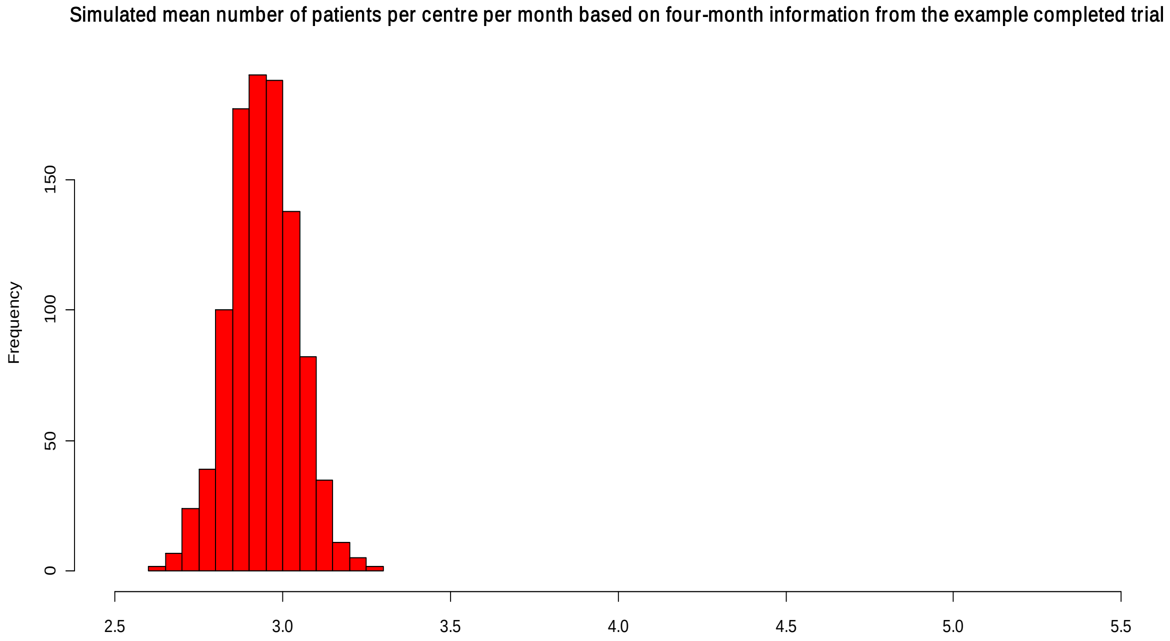
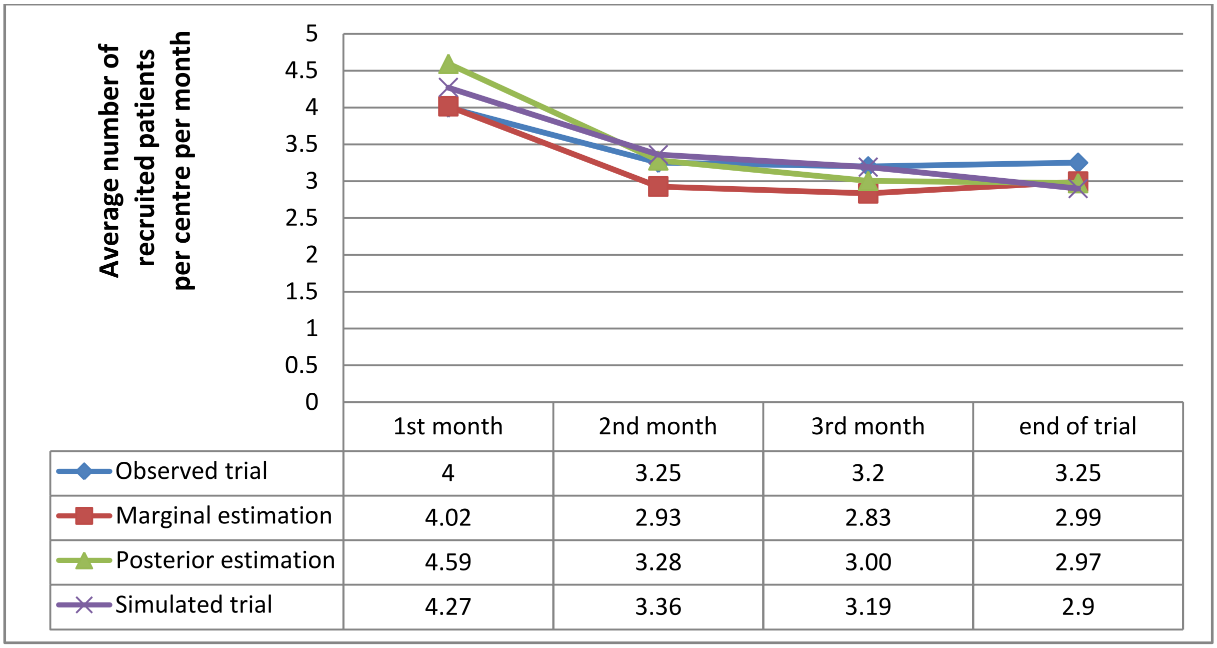


Figure 5.5: Average number of recruited patients per centre per month based on four months monitoring

Figure 5.6 illustrates a visual comparison between the average number of recruited patients per centre per month in the simulated trials and the real completed trial as well as the marginal and posterior estimation in different time intervals. The values of the observed data are taken from table 5.12 and the estimated values derived from table 5.13. The general trends of the mean number of patients' arrivals to the clinical centres are similar except for the last data, where the recruitment has a clear reduction. The average number of patients per centre per month is slightly underestimated in the last simulation.



	1st month	2nd month	3rd month	end of trial
◆ Observed trial	4	3.25	3.2	3.25
■ Marginal estimation	4.02	2.93	2.83	2.99
▲ Posterior estimation	4.59	3.28	3.00	2.97
× Simulated trial	4.27	3.36	3.19	2.9

Figure 5.6: Line graph of the mean number of patients arrived to the centres per month in the real trial, marginal estimation, posterior estimation and the simulated trial

5.7.2. Prediction in time domain

In this section we consider the more relevant question of how long it will take to complete the on-going trial. That is, at month 1 how long will it take to recruit the additional 136 patients, at month 2 how long will it take to recruit the additional 100 patients and at month 3 how long will it take to recruit 56 patients.

In this research, the recruitment frequency data were used for almost all the analysis. Therefore, the process of the simulation in the recruitment time required a movement from the frequency domain to the time domain. Due to the frequency domain of the recruitment data following the Poisson-gamma mixture, the time domain was assumed to have the Gamma-exponential mixture distribution. What the assumption implies is that the waiting times between patients follow the exponential distribution with parameter lambda, which itself is the Gamma distributed with parameters (α,β).

In order to simulate the patients' accrual time, waiting times between patients' arrival should be estimated first. Accrual time is the length of time that it would take to recruit a number of patients in clinical centres. But the waiting time between patients is the delay time between two patients' arrival into centres in a clinical trial. Therefore, by summing up the waiting time between patients we would get the accrual time for that number of patients.

To simulate the waiting time between patients, first, N=12 pairs of the parameters k and μ from the Normal distribution were generated. Then, for each pair, n=152 random numbers from the Gamma distribution with the shape parameter $\alpha = \frac{1}{k}$ and the scale parameter $\beta = k\mu$ were simulated. These numbers, which made a matrix of N=12 rows and n=152 columns, were in fact the values of the parameter lambda of the exponential distribution. So far, there were 152 lambda values generated for each pair of the parameters. Consequently, for each centre (row) the mean numbers of the lambda values were calculated to get only one representative parameter of lambda generated from the Gamma distribution for each centre. After this step the simulated matrix had N=12 rows (centres) and one column (average of lambda values generated from gamma distribution).

The lambda values that were generated from the Gamma distribution were supposed to be the parameters of the Exponential distribution. Hence, for each centre, n=152 waiting times between patients' arrival were simulated from the Exponential distribution using the lambda values as the distribution parameters. The result was again a matrix of N=12 rows and n=152 columns but in the new matrix the values represented waiting times between patients' arrivals.

Although the waiting times between patients' arrivals to the centres had been simulated, there was no clue about the patients' accrual time yet. To gain the answer, in the next step of the simulation process, the waiting times in centres were summed up to calculate the total recruiting time. This, however, was under the condition that all the centres initiated simultaneously. But, in the completed trial, different centres were activated in different dates. This meant that some centres started to recruit patients with some sorts of delay from the start of the trial. There was no doubt that the delay had to be considered in the simulation as well. To address the issue, the delay of the first site was set to zero. But the difference between the activation date of each centre and the first one was added to the matrix of the accrual time. The delay values were taken from the selected completed trial. Then, regardless of the centres, the whole data were sorted by their accrual time. As a result, a set of time data were generated, in which the ith value represented the length of time (in month) to recruit the ith patient in an on-going trial.

The process explained above was repeated for the four data sets gathered in the four time intervals of monitoring the completed trial. In fact, the four data sets of the trial were simulated separately to predict the accrual time based on the different recruiting data. Also, all the four sets of the simulation were run for 1000 times to be able to draw histograms of the accrual times.

In the selected trial, 16 patients were recruited during the first month of initiating the trial. But after the second month of the start of the trial the arrivals increased to three times as many as the first month. 36 more patients had arrived into the clinical centres to increase the total number of patients to 52 by the end of the second month. After the third month, the total quantity of the patients was 96, which meant that the trial had accrued 44 patients during the third month. By the end of the trial 152 patients had arrived into the clinical centres (table 5.12).

The line graph in figure 5.7 shows the sharp increasing trend of the recruitment process in the completed trial. Also, the pie chart in figure 5.8 reveals by how much the number of patients went up after each month.

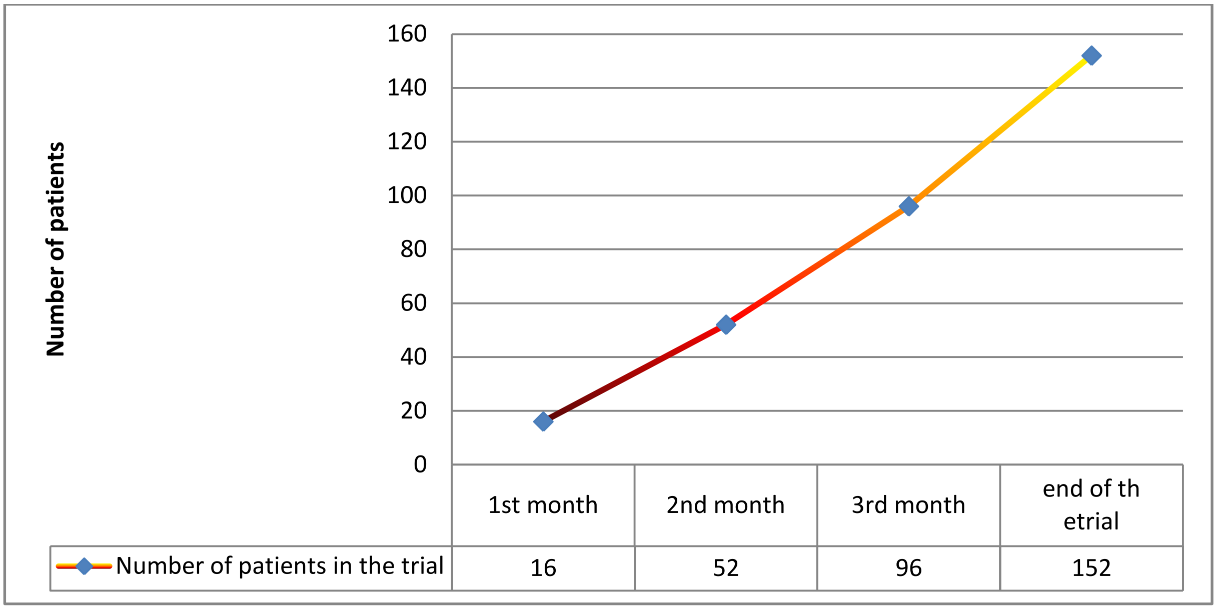


Figure 5.7: Line graph of the recruiting trend througout the trial

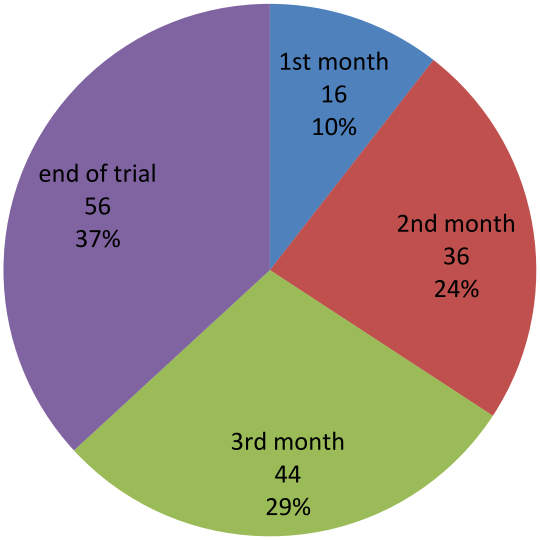


Figure 5.8: Pie chart of the share of each time interval in recruiting patients

An important consideration here is that the increasing rate of the patient recruitment would affect the accrual time. Therefore, it is expected to view a decreasing trend in the accrual time as the trial goes on. This research aimed to predict the accrual time for the remaining patients based on the information in an on-going trial, which was gained in the different follow up times.

Although the main interest was to predict the time to recruit the remaining patients to complete the on-going trial, which are illustrated in figures 5.12, 5.14 and 5.15, short term predictions were also considered (figures 5.10, 5.11 and 5.13). They were, in fact, comparisons between the simulated trial and the actual data. The histograms in figures 5.9, 5.10, 5.11 and 5.13 are mainly to check the compatibility of the simulated on-going trial and the actual trial. While the results in figures 5.12, 5.14 and 5.15 are the answers to the main question of the research. They illustrate the expected time to recruit the remaining individuals from the total sample size of 152 patients if 16, 52 and 96 patients had already arrived to the centres in the past 1st, 2nd and 3rd month of the trial, respectively.

The first trial was simulated based on the first month follow up, during which 16 patients only were recruited. Figure 5.9 is the simulated accrual time to recruit 16 patients. It is basically to check the comparability of the real trial and the simulated one. The length of time to recruit 16 patients estimated from 1000 simulation was stretched from 0.26 to 1.38 month (95% CI: 0.48-1.2). It represented the example trial very well since the one month accrual time in the trial was included in the predicted 95% confidence interval.

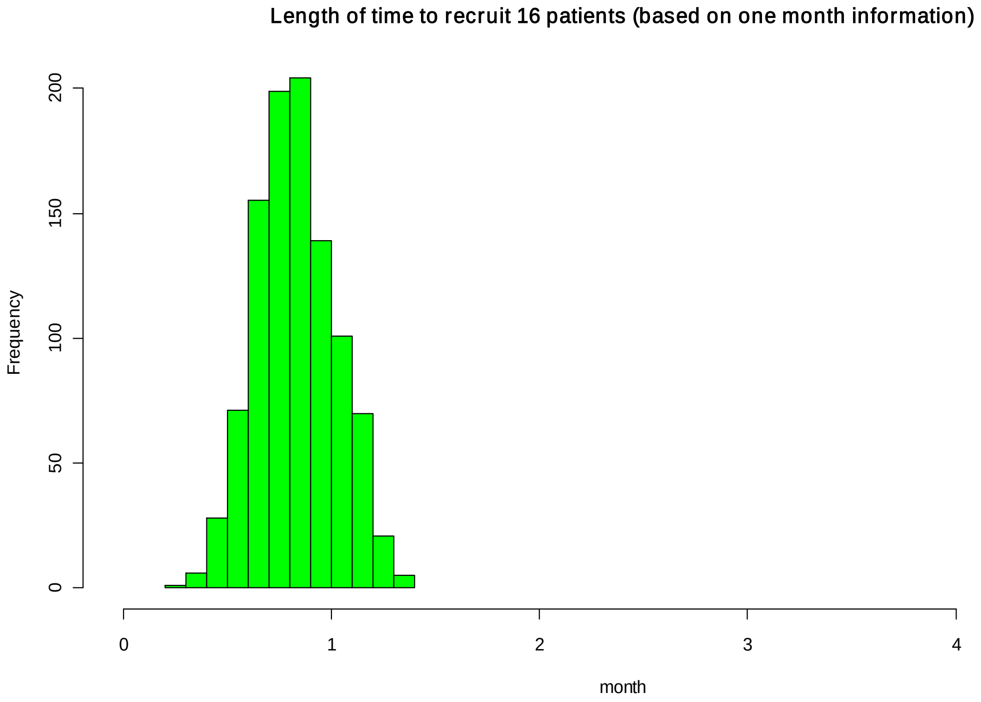


Figure 5.9: Histogram of the simulated accrual time for 16 patients based on one month follow up

The simulated trial, however, was mainly to predict the accrual time to recruit the remaining patients from the sample size. It had taken 1 month to recruit 36 more patients (52-16=36) in the completed trial after the 16 patients had already arrived into the clinical centres in the first month. The simulation predicted that it would be between 0.58 to 1.66 months (95% CI: 0.75-1.43, median=1.09 months) (figure 5.10).

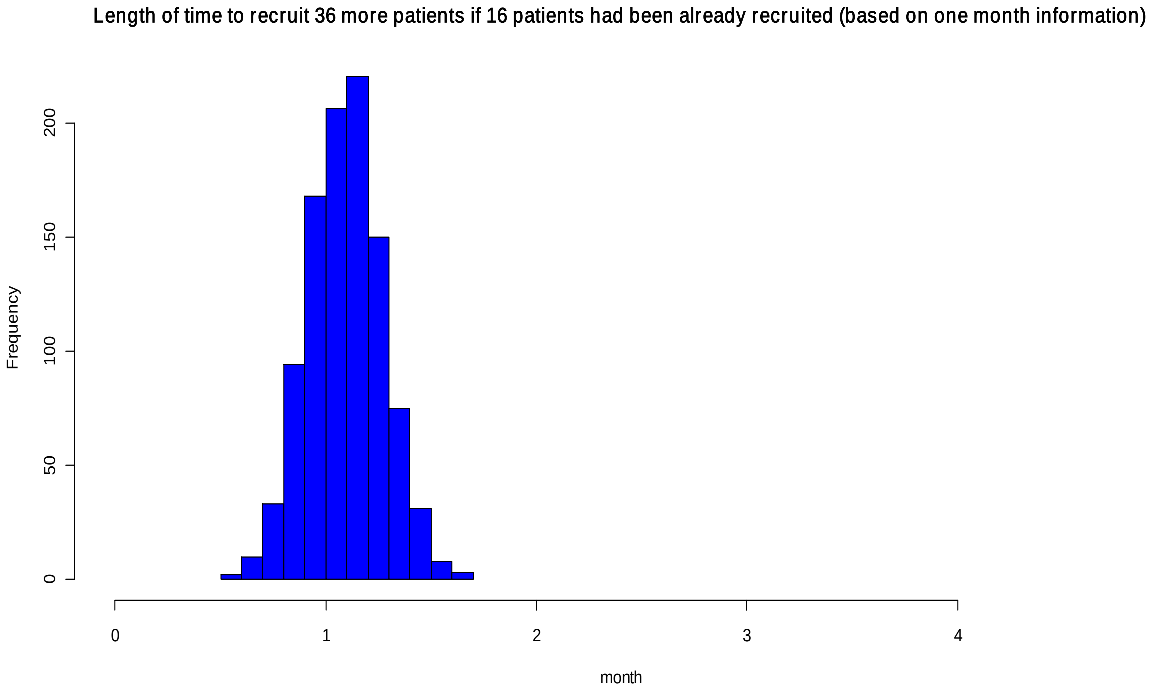


Figure 5.10: Histogram of the predicted accrual time of 36 patients based on the assumption that 16 patients had already arrived into the centres in the past month

Based on 1000 simulation, it was predicted to take between 1.51 to 2.92 months (95% CI: 1.73-2.47; median=2.10 months) for 80 more patients (96-16=80) to arrive the clinical centres if 16 patients had already done so in one month (figure 5.11). The value in the completed trial was 2 months, which is very similar to the predicted median and fell into the 95% confidence interval.

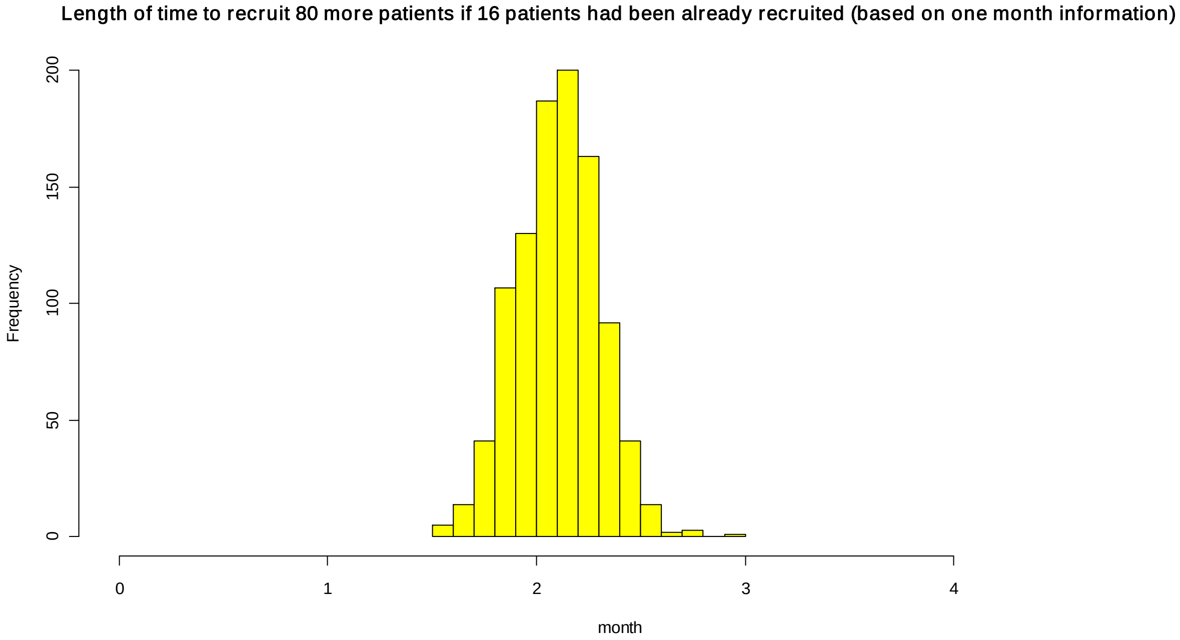


Figure 5.11: Histogram of the predicted accrual time of 80 patients based on the assumption that 16 patients had already arrived into the centres in the past month

If it was known that during the first month of the trial 16 patients had arrived into the centres, figure 5.12 shows the predicted accrual time for the remaining 136 patients out of the 152 sample size (152-16=136). The simulated trial, predicted an average of 3.11 months. In 95% of the trials the accrual time for the remaining 136 patients was between 2.67 and 3.60 months. In the real trial the total accrual time was about 3.90 months. Considering the arrivals of 16 patients in the first month, the accrual time for the remaining 136 patients had been about 2.90 months and it was contained in the confidence interval.

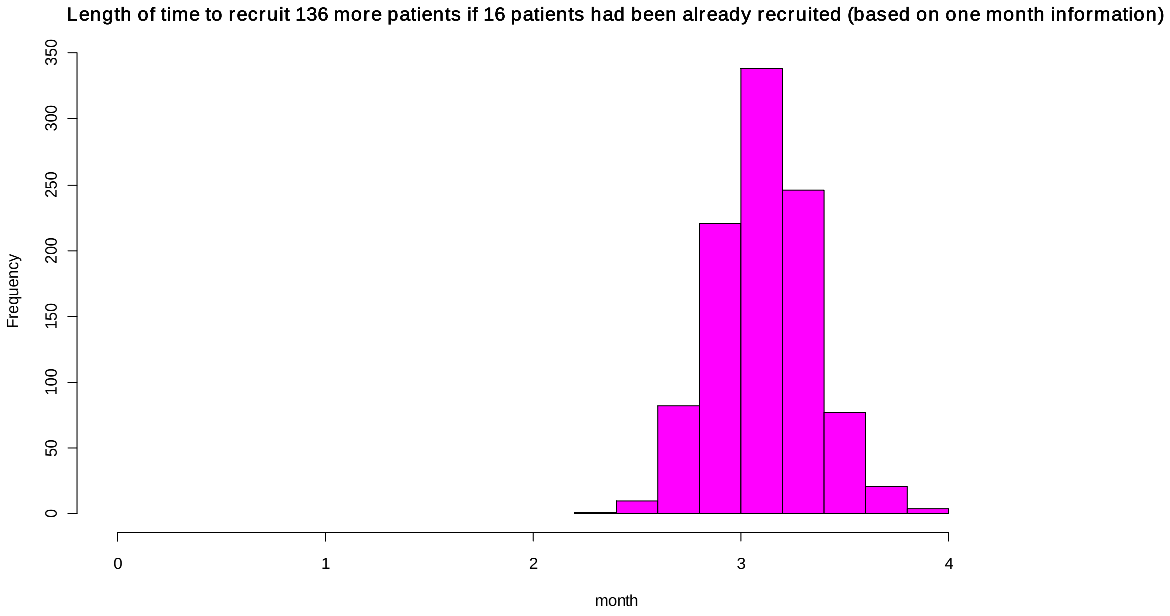


Figure 5.12: Histogram of the predicted accrual time of 136 patients based on the assumption that 16 patients had already arrived into the centres in the past month

The second trial was simulated based on two months follow up of the completed trial, during which 52 patients had arrived into clinical centres. The simulation was first to predict how long it would take to recruit another 44 patients (96-52=44). According to the simulation result, it was predicted to take 0.86 month in average, which is less than one month. However, the one month accrual time for the extra 44 patients in the real trial was included in the 95% confidence interval of the prediction (CI: 0.63-1.08 months) (figure 5.13).

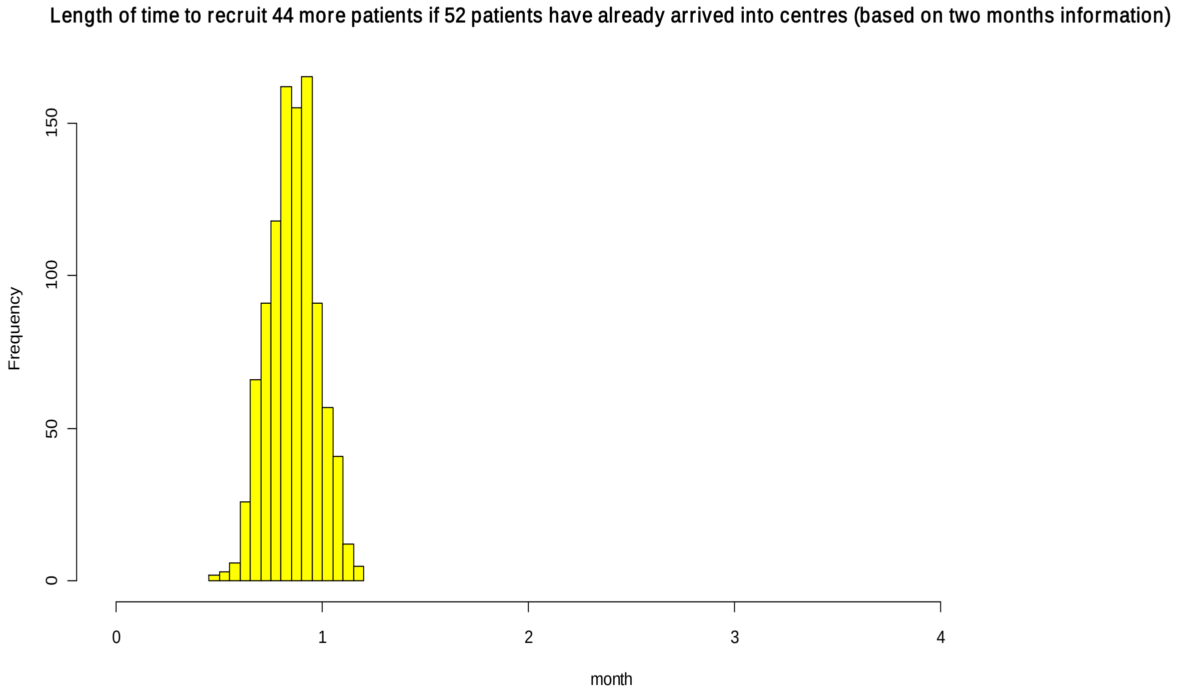


Figure 5.13: Histogram of the predicted accrual time of 44 patients based on the assumption that 52 patients had already arrived into the centres in the past two months

If it was to recruit 152 patients in a clinical trial, the question raised here could be how long it would take to recruit 100 individuals (152-52=100) if 52 had been recruited in the past two months. The 1000 simulations highlighted that the 95% confidence interval was expected to be between 1.40 and 1.96 months. The real accrual time for the remaining patients to arrive into centres was about 1.90 months which was very well covered by the predicted CI (figure 5.14).

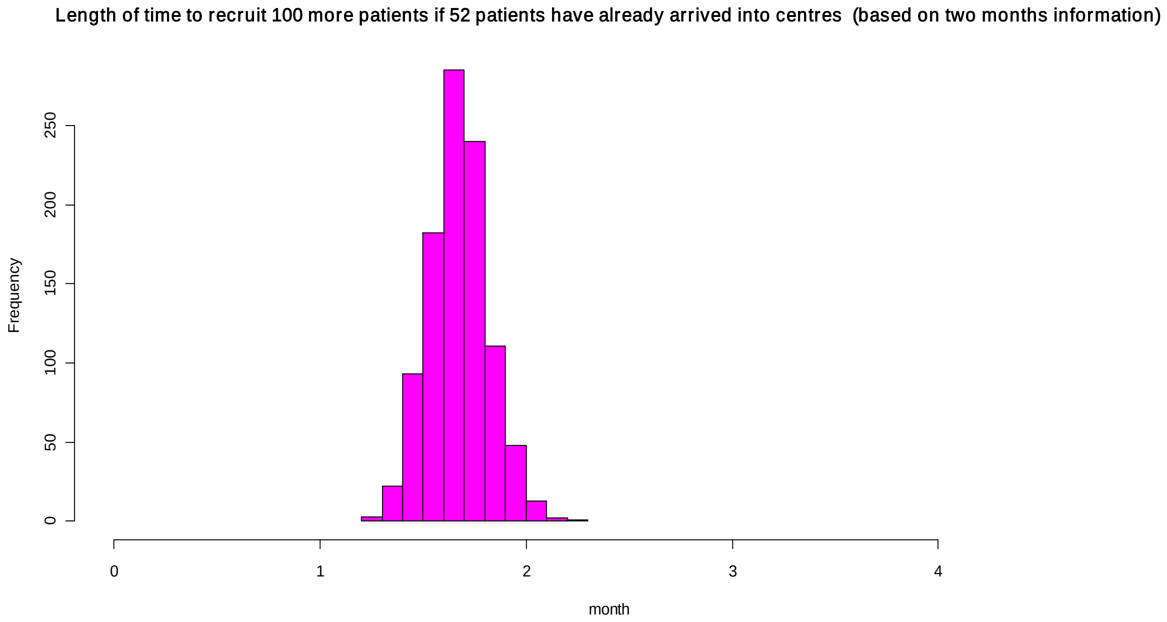


Figure 5.14: Histogram of the predicted accrual time of 100 patients based on the assumption that 52 patients had already arrived into the centres in the past two months

Finally, the simulated trial predicted the expected accrual time of 0.50 to 1.20(95% CI: 0.55 – 0.88) months to recruit the 56 remaining individuals from the sample size of 152 patients(152-96=56) in an on-going trial, in which 96 patients had been already recruited in three months. In the completed trial, the recruitment time for the remaining 56 patients was about 0.90 months which has been slightly underestimated in the last simulation (figure 5.15).

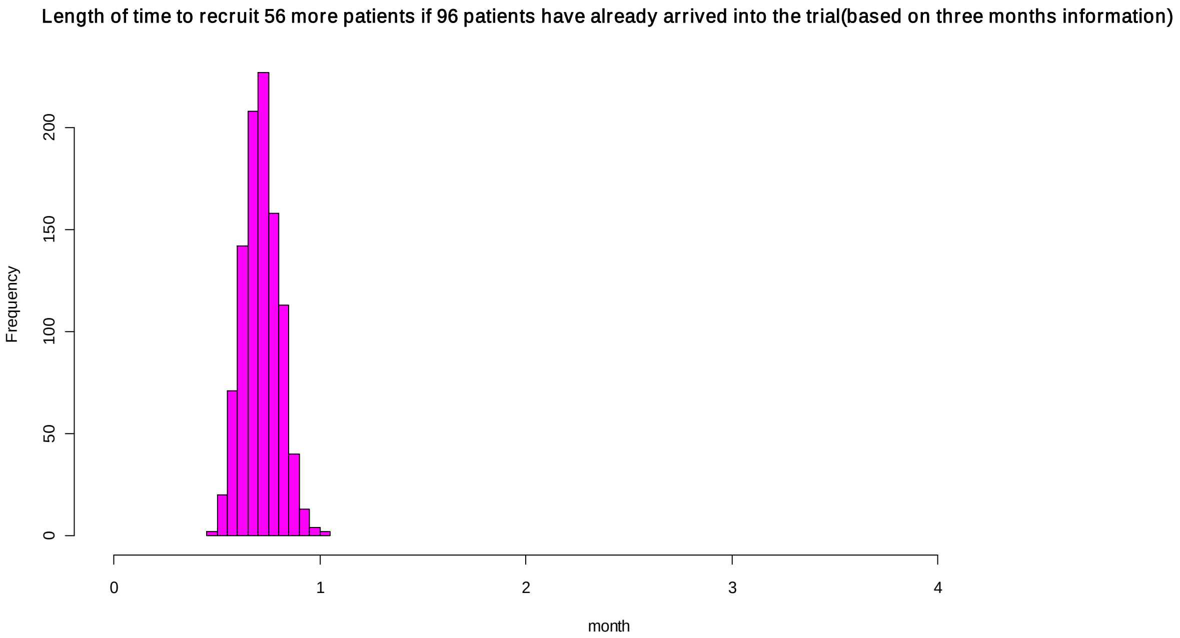


Figure 5.15: Histogram of the predicted accrual time of 56 patients based on the assumption that 96 patients had already arrived into the centres in the past three months

For the final view, table 5.14 shows how the predictions of the accrual time of the remaining patients changed based on the simulated on-going trials.

Time intervals	Patients to be recruited after each time interval	Mean of the predicted time (month) to recruit the remaining individuals	95% CI prediction for the time (month) to recruit the remaining patients	Actual time to predict the remaining patients
1st month	136	3.11	(2.67 - 3.60)	2.9
2nd month	100	1.6	(1.40 - 1.96)	1.9
3rd month	56	0.72	(0.56 - 0.88)	0.9

Table 5.14: Summary table of the predicted accrual times based on the 1st, 2nd and 3rd month data

Overall, the aim of the project was achieved in the way that it was successful in using the frequency domain to predict the time domain in an on-going multi-centre clinical trial. Moreover, almost all of the predictions covered the accrual time of the real trial.

Conclusion and discussion

In this research an on-going trial was simulated using the data from completed trials provided by ICON. The overall approach was to model the frequency domain from recruitment data and move smoothly to predict the time domain.

The frequency of patients' arrivals was modelled as if they followed the Poisson-gamma mixture distribution with the accrual times included in the model as the offset variables. Then, the waiting times between patients were simulated based on the assumption that they were Gamma-exponential distributed. The total recruitment time for a pre-fixed number of individuals were predicted by summing up the waiting times between the patients' arrivals.

The main interest was to predict the accrual time of the remaining individuals in an on-going trial. As an example, in the trial in which 152 patients were to be recruited by the end of the trial, the main interest was to predict the time to recruit the remaining 136 patients if 16 patients had already been taken into the clinical centres in the first month of the trial. However, short term predictions were applied as well to see whether the simulated trials provided similar time to the actual trials. To make it clearer, in the example above, the time to recruit 36 more patients was also predicted if 16 patients had already arrived by the end of the first month. Hence, the short term prediction was basically to see if the predicted recruiting time for 36 more patients is close to the actual recruiting time in the selected trial. Overall, three different on-going trials were simulated based on the information of the follow up intervals of a selected completed trial. Initially, the simulations predicted the expected number of recruitments per centre per month derived from the 1st, 2nd and 3rd month recruitment information. Then the expected length of time (month) to recruit the remaining number of patients and complete the trial was predicted based on the one, two and three months of data collection.

As regards the frequency domain, the comparison results between the simulated trials and the actual data were very reasonable. The average frequencies of the completed trials were well covered by the 95% confidence intervals of the expected values predicted from the simulated on-going trials. The exception, however, was the first simulation which slightly overestimated the average number of patients per centre per month. The average number of patients per centre per month is expected to remain unchanged throughout the trial. This value is remarkably more in the first month of the actual trial. That could be the reason of the over-estimation outcome. The line graph in figure 5.6 compares the expected number of individuals to be recruited in the simulated on-going trials and the actual completed trial in different time steps.

In the main forecasting area, the time domain, the predictions of the time to accrue the remaining individuals from the pre-arranged sample size were reasonable compared to the values in the actual trial. Although it may imply that the real accrual time of the remaining patients in the 3rd trial has been underestimated by the 3rd simulation, it is located in the upper end of the histogram.

In the first simulated trial, the question was that how long it would take to recruit the remaining 136 patients (152 patients to be recruited) if 16 patients had already arrived into the clinical centres during the first month of the trial. The recruitment time in the actual trial was about 3.90 month for the sample size of the 152 individuals and 16 patients were recruited by the end of the first month of the trial. Hence, the accrual time for the remaining 136 patients was 2.90 months. The relative simulated on-going trial, with 95% confidence interval, predicted that the recruitment time for the remaining 136 patients is between 2.67 and 3.60 month, which was a reasonable result.

In the second data set that was derived from the actual trial, 52 patients had arrived into the centres by the end of the second month. So, the time for the remaining 100 patients to enter the trial was 1.90 months. The second simulation (based on the two-month information) suggested 1.40-1.96 months in its 95% confidence interval. It is an acceptable result, although the value of the actual trial is moving towards the upper end of the CI.

By the end of the 3rd month, the trial had recruited 96 patients and took less than a month to take the remaining 56 individuals. However, in a 95% confidence interval, the simulated trial predicted a range of 0.56 to 0.88 months for 56 patients to be accrued in the clinical trial. As a discussion, it implies that it may slightly underestimate the accrual time if the model is applied. This is due to the observed data being at the upper end of the confidence interval.

Overall, the main purpose of the research was to predict the patients accrual time in an on-going clinical trial. It was in fact to forecast the recruitment time for the remaining patients in an on-going trial. Since the recruitment data were better presented in the frequency domain, the project had to manage to use the recruitment data from the completed trials in the frequency domain to forecast the recruitment in the time domain in an on-going trial. In future, we could improve the recruitment prediction by applying the statistical distribution to the time domain directly. In this case the Pearson VI distribution that was suggested by Anisimov et al would be a more reasonable fit.

Appendix 1

Parameters of Negative Binomial distribution were estimated applying maximum likelihood estimation method in R.

```
### *** Analysing  recruitment data by fitting Negative Binomial distribution to the
frequency domain with parameters r=alpha   ,    p=1/(1+beta)*** ###
# log-likelihood function #
log.like                <-function(par,x)
{
        a       <- log(gamma(x+par[1]))
        b       <- length(x)* log(gamma(par[1]))
        c       <- length(x)*par[1]*log(par[2]+1)
        d       <- x*log(par[2]/(par[2]+1))
        e       <- log(gamma(x+1))
log.like                <-sum(a)-b-c+sum(d)-sum(e)
}
# To optimize the function starting from lowest number of patient in centres for alpha(here is
2),  beta=mean/alpha (here is 8)#
estimation      <-optim(c(2,8),log.like, control=list(fnscale=-1) , hessian=TRUE, x=patients)
estimation

alpha   <-estimation$par[1]             #estimated parameter alpha#
beta    <-estimation$par[2]             #estimated parameter beta#
p       <-1/(1+beta)                    #scale parameter in NB distribution#
m.u     <- alpha*(1-p)/p                #mean number of values in NB #
variance        <-alpha*(1-p)/(p^2)             #variance of data in NB distribution#
s.d     <-  variance^.5                 #standard deviation#
```

```
#standard error of the parameter#
se.par   <- sqrt(diag(-solve(estimation$hessian)))

### **** confidence intervals for alpha and beta *** ###
b                <- c(1,0)        # vector to consider alpha only(the first parameter)#
lower.alpha      <-(t(b)%*%estimation$par)-(1.96*sqrt(t(b)%*%(-
solve(estimation$hessian))%*%b))
upper.alpha      <-(t(b)%*%estimation$par)+(1.96*sqrt(t(b)%*%(-
solve(estimation$hessian))%*%b))
# the vector to consider beta only  (the second parameter)#
b                <- c(0,1)
k                <-(-estimation$hessian)        # the sample information matrix#
lower.beta<-(t(b)%*%estimation$par)-(1.96*sqrt(t(b)%*%(-
solve(estimation$hessian))%*%b))                              #lower band for the CI#
upper.beta<-(t(b)%*%estimation$par)+(1.96*sqrt(t(b)%*%(-
solve(estimation$hessian))%*%b))                              #upper band for the CI#

# Estimating parameters using Maximum Likelihood Method in MASS package #
library(MASS)                                           # laoding package MASS #
fitdistr(patients , "negative binomial")        # fitting Negative binomial parameters #

### *** log-likelihood contour plot *** ###
ngrid    <-60                                           # Produce the number of grid lines#
alpha    <-seq(1,5,length=ngrid)                # location of grid line of parameter alpha #
P        <- seq( 0, 1, length=ngrid)            #location of grid line of  parameter p #
# Make a data frame and then matrix from alpha and p#
grid     <-as.matrix(expand.grid(alpha,p))
log.like                 <-rep(0, nrow(grid))
n                <-24                     #Number of patients in the centre#
for (i in 1:n)                            # to calculate the log-likelihood for the given data#
{
         log.like                         <-log.like+log(dnbinom(patients[i], grid[,1], grid[,2]))
         log.like[log.like        <-150] <-NA
```

```
        log.like                    <-matrix(log.like, nrow=ngrid)
}

contour (alpha, p, log.like)
#Draw the contour plot with x=alpha, y=p, z=log-likelihood#
title(xlab="alpha", ylab="p", main="Likelihood function")
text(1.273, .073, "max")

image (alpha, p, log.like)                  #Image of the contour plot#
text(1.273, .073, "max")                    #Label the MLE value#

### *** panel *** ###
library(rpanel)
log.like            <-function(theta, data)
# writing the likelihood function separately then adding all togetherfor simplicity#
{       par     <-theta
        x       <-data
        a       <- log(gamma(x+par[1]))
        b       <- length(x)* log(gamma(par[1]))
        c       <- length(x)*par[1]*log(par[2]+1)
        d       <- x*log(par[2]/(par[2]+1))
        e       <- log(gamma(x+1))
log.like<-sum(a)-b-c+sum(d)-sum(e)
invisible (sum(a)-b-c+sum(d)-sum(e))
}
# Draw the panel for log-likelihood estimated parameters#
rp.likelihood(log.like, patients, c(0.1, 2), c(2, 8) )
rp.likelihood("sum(dnbinom(data, theta[2], theta[1], log = TRUE))",   patients, c(0.1, 2), c(0.9, 4))

### *** Goodness of fit test *** ###
### *** H0: the data follow NB distribution  *** ##
library (vcd)
```

```
gf.NB  <-       goodfit(patients,type= "nbinomial", method= "MinChisq")
summary(gf.NB)

### *** estimating parameters applying method of moments *** ###
x.bar           <- mean (patients)          # expected value from the sample #
s2.bar          <-  var(patients)           # variance from the sample#
alpha.moment <- x.bar^2/s2.bar              # estimating alpha from the mean and variance #
beta.moment  <- x.bar/alpha.moment          # estimating beta from alpha and mean#
p.moment        <- 1/(1+beta.moment) # estimating probabity from its relationship with beta#


# *** qplot *** #
#to check if the number of patients recruited follow the negative binomial distribution#
plot(qnbinom(ppoints(patients), size=1.274, mu=16.044), sort(patients), xlab="
Empericalquantiles from NB", ylab="sample empericalquantile")
abline(0,1) #drawing a 45-degree reference line#

### *** emperical density functions *** ###
#producing random numbers from NB distribution with the given parameters#
y<-rnbinom(24,size=1.274,prob=.074)
#to draw the cumulative functions in one window#
Par(mfrow=c(1,2))
# to plot empirical cumulative distribution function for recruited patients #
plot.ecdf(patients, verticals = TRUE, pch=19, col="red",
xlab="Number of patients per country", ylab="probability function", main= "Empirical
cumulative step function, study 1")

#empirical cumulative step function plot for the random NB values#
plot.ecdf(y, verticals=TRUE, pch=19, col="darkblue",
xlab="Negative Binomial random Values", ylab="probability function")

### *** Histogram with fitted distribution *** ###
#histogram of the patients recruited#
```

```
hist(patients, freq=FALSE, xlim=c(0,80),
col="lightblue", border="red", plot=TRUE,
xlab="Number of patients recruited", ylab="Probability densities",
main="Histogram of recruited patients with the fitted NB in study1")


x<-seq(2,80,1)                              #produce a sequence of numbers from 2 to 80#
fn.nb<-dnbinom(x,size=1.274,prob=.074)      #calculate a function of negative binomial distribution with the given parameters#
lines(x,fn.nb)                              # Negative binomial curve for the given parameters #

### *** kernel density plot *** ###
d <- density(patients)                      # to estimate the frequency density of the data#
plot(d, main="Kernel Density of patients in study 1")
polygon(d, col="red", border="red")

### *** Analysing the number of patients by fitting Negative Binomial distribution with parameters r=alpha*N and p=t/beta *** ###
### *** N  isthe number of clinical centres and t is the total recruitment time  *** ###
#The log-likelihood function for NB distribution with parameters alpha*N and t/beta ##
logl<-function(par,y,N,t)
{
        a<-log(gamma(y+par[1]*N))
        b<-length(y)*log(gamma(par[1]*N))
        c<-log(gamma(y+1))
        d<-y*log((t/(t+par[2])))
        e<-length(y)*par[1]*N*log(par[2]/(t+par[2]))
        logl<-sum(a)-b-sum(c)+sum(d)+e
}
logl
```

```
est<-optim(c(2,5),logl, control=list(fnscale=-1),hessian=TRUE, y=patients,
N=length(country),t=max(time.month))# to estimate the parameter alpha=par[1] and
beta=par[2] by maximising the function #
est

alphat<- est$par[1]
betat<- est$par[2]
shape<- alphat * N
scale<- max(time.month)/betat
```

Appendix 2

Notes regarding the influence of units of time in modelling the Poisson process on the time or frequency domain.

Suppose we have a Poisson process with mean λ observed for time $T = 1$. Note, in practice, T is not unit-free it will have to have units attached. To take a concrete example, we might have $T = 1\, year$. The probability mass function for the number of events X is given by the Poisson distribution with

$$P\left(X = x \right) = \frac{\lambda^x e^{-\lambda}}{x!}.(1.1)$$

Note also that the λ that is used must be the λ that is consistent with T. This λ is a rate parameter and if the rate parameter is defined with respect to a different period of observation it must be adjusted accordingly so that when substituted in (1.1) it gives the correct probability. For example if we choose to define λ as a rate per month but observe the process for a year we must substitute 12λ for λ in (1.1). Note that the mean of the resulting Poisson is 12λ and its variance is also 12λ. This is because the resulting probability distribution does not involve any multiplication of X. Mere multiplication of an X that was observed for one month by 12 would no longer result in a Poisson distribution. Instead we would have a random variable with mean 12 times as large and variance $12^2 = 144$ times as large. On the other hand observing a Poisson process for 12 months and adding together the 12 independent random variables consisting of the numbers in each month would give a total that was still Poisson with expectation and variance of 12λ.

Let the inter-arrival time be Z. The mean inter-arrival time is $\mu = 1/\lambda$. Note that μ is also not unit free and must be measured in the same time units as Z. The probability density function for inter-arrival times is an exponential distribution with

$$f\left(Z = z\right) = \lambda e^{-\lambda z} dz = \frac{1}{\mu} e^{-\frac{z}{\mu}} dz. (1.2)$$

Now suppose that we choose to measure in new time units t. If we create a new random variable $Y = tZ$ then $y = tz, z = y/t, dz = dy/t$ and so we have

$$f\left(Y = y\right) = \frac{\lambda}{t} e^{-\frac{\lambda}{t} y} dy. (1.3)$$

To take a concrete example, we might decide to measure in days rather than years. In that case we would have $t = 365 days/year$. Clearly this new random variable is also an exponential distribution with mean $\mu^* = t/\lambda = t\mu$.

The gamma distribution which includes (1.2) as a special case is given by

$$g\left(W = w\right) = \frac{1}{\Gamma\left(\alpha\right)\beta^{\alpha}} e^{-\frac{w}{\beta}} w^{\alpha-1} dw. \qquad (1.4)$$

Clearly by setting $\alpha = 1$ in (1.4) and $W = Z, z = w$ we get (1.2). Specifically, if α takes on a positive integer value and we define

$$W = \sum_{i=1}^{\alpha} Z_i \qquad (1.5)$$

as the sum of α inter-arrival times then (1.4) can be used as the waiting time until α events occur. More generally, however, we can let α be some positive real number and apply a distribution like (1.4) as a mixing distribution for a Poisson process itself.

If, analogously to the case with the exponential distribution, we create a new random variable $V = tW$ then $v = tw, w = v/t, dw = dv/t$ and then substitute in (1.4) we get

$$g\left(V = v\right) = \frac{1}{\Gamma\left(\alpha\right)\beta^{\alpha}} e^{-\frac{v}{t\beta}} \left(\frac{v}{t}\right)^{\alpha-1} \frac{1}{t} dv = \frac{1}{\Gamma\left(\alpha\right)\left(t\beta\right)^{\alpha}} e^{-\frac{v}{t\beta}} v^{\alpha-1} dv \qquad (1.6)$$

Clearly this is a gamma distribution with parameters $\alpha, t\beta$ rather than α, β. Note also that (1.4) has mean $\alpha\beta$ and variance $\alpha\beta^2$. Whereas (1.6) has mean $\alpha t\beta$ and variance $\alpha t^2\beta^2$.

This is clearly appropriate since the random variable V is simply created by multiplying the random variable W by a constant t.

Stephen Senn, 7 June 2011

References

1. Senn Stephen, **Some controversies in planning and analysing multi-centre trials.** *Statistics in Medicine*1998, **17**: 1753-1765

2. Anisimov VV: **Recruitment modelling and predicting in clinical trials.** *Research Statistics Unit GlaxoSmithKline*

3. Anisimov VV, Fedorov VV: **Modelling, prediction and adaptive adjustment of recruitment in multicentre trials.** *Statistics in Medicine* 2007, **26**:4958-4975

4. Anisimov V, Federove V, **Design of multi-centre clinical trials with random enrolment,** in book *Advances in statistical methods for the health science. Applications to cancer and AIDS studies, genome sequence analysis and survival analysis.* Series: Statistics for industry and technology, 2006, Ch. 25, pp 393-406.

5. Anisimov VV, Fedorov VV: **Modelling of enrolment and estimation of parameters in multicentre trials.** *GlaxoSmithKline Pharmaceuticals* 2005.

6. Williford WO, Bingham SF, Weiss DG, Collins JF, Rains KT, Krol WF: **The "constant intake rate" assumption in interim recruitment goal methodology for multicentre clinical trials.** *Journal of Chronic Disease1987,* **40:** 297-307

7. Haidich AB, Ioannidis JP: **Determinants of patient recruitment in a multicentre clinical trials group: trends, seasonality and the effect of large studies.** *BMC Medical Research Methodology* 2001, 1:4

8. Carter RE, Sonne SC, and Brady KT: **Practical consideration for estimating clinical trial accrual periods: application to a multi-centre effectiveness study**. *BMC Medical Research Methodology* 2005, **5**:11

9. Carter RE: **Application of stochastic processes to participant recruitment in clinical trials.** *Controlled Clinical Trials* 2004, **25**:429-436

10. Abbas I, Rovira J, Casanovas j: **Clinical trial optimization: Monte Carlo simulation Markov model for planning clinical trials recruitment.** *Contemporary Clinical Trials* 2007*,* **28**: 220-231.

11. Gajewski BJ, Simon SD, and Carlson SE: **Predicting accrual in clinical trial with Bayesian posterior predictive distribution.** *Statistics in Medicine* 2008, **27**: 2328-2340.

12. Senn S. Inference 4 (notes), chapter 4.

13. Cook JD: **Notes in negative binomial distribution**, 2009 Oct 28. http://www.johndcook.com/negative binomial.pdf

14. Hastings N.A.J, Peacock J.B: **Statistical Distributions**. The Butterworth group:1975

15. Schoenberger JA: **Recruitment in the coronary drug project and the aspirin myocardial infarction study**. ClinPharmacolTher 1979, 25: 681-684.

16. Benedict GW: **LRC coronary prevention trial**. Baltimore.CtinPharmacolTher 1979, 25: 685-687.

17. Prout T: **Other examples of recruitment problems and solutions**. CIinPharmacolTher 1979, 25.

18. Douglas Gregory: **Time-cost optimization of complex clinical trials,** drug information journal, 2011 Vol 45: 345-356.

19. Stephen Senn, James Weir, Tsushung A. Hua, Conny Berlin, Michael Branson, EkkehardGlimm:**Creating a suite of macros for meta-analysis in SAS$^{®}$: A case study in collaboration, Statistics &amp.**Probability Letters, Volume 81, Issue 7, July 2011, Pages 842-851,ISSN 0167-7152, 10.1016/j.spl.2011.02.010.http://www.sciencedirect.com/science/article/pii/S0167715211000484

20. http://www.senns.demon.co.uk/SAS%20Macros/SASMacros.html

21. Wand M. P, Jones M C: **Kernel Smoothing.** Chapman & Hall, 1995.