Journal Title
XX(X):2–29

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Modelling and sample size reestimation for longitudinal count data with incomplete follow up

Thomas Asendorf[1], Robin Henderson[2], Heinz Schmidli[3] and Tim Friede[1]

Abstract

We consider modelling and inference as well as sample size estimation and reestimation for clinical trials with longitudinal count data as outcomes. Our approach is general but is rooted in design and analysis of multiple sclerosis trials where lesion counts obtained by magnetic resonance imaging are important endpoints. We adopt a binomial thinning model that allows for correlated counts with marginal Poisson or negative binomial distributions. Methods for sample size planning and blinded sample size reestimation for randomized controlled clinical trials with such outcomes are developed. The models and approaches are applicable to data with incomplete observations. A simulation study is conducted to assess the effectiveness of sample size estimation and blinded sample size reestimation methods. Sample sizes attained through these procedures are shown to maintain the desired study power without inflating the type I error. Data from a recent trial in patients with secondary progressive multiple sclerosis illustrate the modelling approach.

Keywords

adaptive design, lesion counts, sample size reestimation, negative binomial, discrete autoregressive process, time dependent

1 Introduction

Many biological experiments and clinical trials use count data as primary outcomes, for example the number of events within a certain time frame. Our particular interest is in the context of multiple sclerosis (MS) clinical trials, where the number of brain lesions assessed at regular intervals by magnetic resonance imaging (MRI) form an important measure of disease severity.[1] The response of interest is thus longitudinal count data, and because counts tend to be low, the highly discrete nature of the response needs to be taken into account properly, in design and analysis. We consider both in this article.

A variety of analysis methods are available for longitudinal count data.[2] Examples include early consideration of integer time series as Markov chains,[3–5] later alternative transition or other parametric models,[6–9] and general classes of model-free estimating equations[10,11] or random effect and latent process approaches.[12–15] Given that count data are commonly assumed to follow possibly overdispersed Poisson or negative binomial distributions, and that a fully specified model is needed for sample size estimation, our preference is for a parametric approach which accounts for within-person dependence but maintains negative binomial marginal distributions. While this can be true for certain frailty models[12,14] a binomial thinning construction formally introduced by Steutel and van Harn[16] can lead to an intuitive mechanistic interpretation for lesion count data, as we will explain later. Hence the approach we take is an adaptation of the binomial thinning developments of McKenzie,[17] and Al-Osh and Alzaid.[18] This will be described in Section 3.

Turning to design, an important step in the planning of clinical trials is the determination of an appropriate sample size for primary, secondary and sometimes also key outcomes. However, when calculating the sample size required to achieve a certain power, assumptions on the distribution of the outcome have to be made. Such assumptions are of course subject to uncertainty, which might be amplified if the

[1]University Medical Center Göttingen, Department of Medical Statistics, Göttingen, Germany
[2]University of Newcastle, School of Mathematics and Statistics, Newcastle Upon Tyne, United Kingdom
[3]Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

Corresponding author:
Thomas Asendorf, University Medical Center Göttingen, Department of Medical Statistics, Humboldtallee 32, 37073 Göttingen, Germany.
Email: thomas.asendorf@med.uni-goettingen.de

Prepared using sagej.cls

proposed trial has different characteristics from the historical trials used to inform sample size calculation, such as different eligibility criteria perhaps, or variations in length of follow up. Adaptive study designs using an internal pilot study (IPS) to reestimate the sample size, based on the distribution of outcomes observed in interim data, can reduce uncertainty[19] and have been encouraged by regulatory authorities.[20,21] In comparison to final analyses, analyses on interim data are often more challenging. Firstly, unblinded analyses at interim bear the risk of introducing bias to the study. To avoid these unwanted influences on the study outcome, an unblinding of the interim data is not recommended and should be avoided whenever possible.[22] Secondly, at interim many patients recruited into the study might only have provided incomplete follow up. Therefore, any sample size reestimation procedure should be applicable to blinded data and allow data with incomplete follow up. In this paper, we focus on developing procedures for sample size estimation and blinded sample size reestimation (BSSR) in randomized controlled two-arm trials, by using the proposed modelling approach to address the challenge of incomplete follow up times. Statistical inference is then performed on the ratio of group specific rates.

The paper is organised as follows. In Section 2 we give some background on MS studies and lesion count outcomes and we describe two trials which we use to calibrate our approach. Section 3 continues with a description of the proposed binomial thinning model as well as a formal derivation of a Wald-type test statistic for inference on the ratio of experimental and control group rates. In Section 4 we derive sample size estimation and blinded sample size reestimation procedures and in Section 5 we demonstrate these procedures in retrospect on a recent MS trial. Simulation results to investigate the accuracy of the proposed procedures are described in Section 6 and the paper is concluded in Section 7 with discussion of possibilities for further extending the modelling approach.

2 Lesion counts and MS-trials

Lesion counts from MRI play an essential role in the diagnosis of MS.[23] They are used as supportive secondary outcomes in phase III trials but are established as primary outcome measures in phase II trials, which tend to be shorter and smaller in terms of number of patients enrolled, and rely on short-term indicators of disease activity.[24] Lesion counts seem to fulfill the requirements of such a sensitive indicator.[25–27] In MRI, different techniques have been developed to visualize lesions. Two classic measures are

T1-hypointense lesions, also referred to as 'black holes' and T2-hyperintense lesions, also referred to as 'white spots'. These lesion types are inter-related yet different. T1-hypointense lesions appear as dark areas on an image and may arise from severe chronic demyelination, inflammation, permanent axonal damage and gliosis through multiple sclerosis. They may also represent areas of edema or swelling, and can be temporary and disappear in subsequent scans. T2-hyperintense lesions in contrast can show the total number of lesions as they are a marker of past injury, which only rarely disappear completely.[28] Both imaging endpoints have been considered as endpoints in numerous MS studies.[29–34] T2 lesions have been shown to be a reasonable surrogate for the number of relapses, both for relapsing-remitting MS (RRMS) and secondary progressive MS (SPMS).[35,36]

There are two quite different modelling strategies for lesion count data in MS. One relies on assuming the *marginal* distributions of lesion counts at observation times to be negative binomial,[37] whereas the second relies on an assumption that the *sum* of lesion counts over a certain time span is negative binomial.[38] Both modelling possibilities seem adequate when investigating real data, but when the observations are dependent, in theory only one can be true, since the sum of dependent negative binomial distributed variables generally does not follow a negative binomial distribution. The second, accumulation, approach has advantages in leading to simpler analyses, but the first is more flexible in modelling and in handling data with incomplete follow up. In the following, we will thus use a model which allows for marginal distributions of lesion counts to be from a negative binomial distribution.

For illustration and motivation of the techniques to come, we will consider two very different MS studies in which counts of T2 lesions have been used as endpoints. The first is a study conducted by Tubridy et al.,[39] which investigated whether the T2 outcome distribution depends upon the type of MS. It listed newly active lesions at monthly MRI scans for 31 RRMS patients and 26 SPMS patients from natural history studies or placebo arms of therapeutical trials, with six months follow up per patient. The second example we consider is a randomized controlled trial by Chataway et al.[29] The study aimed at demonstrating an effect of high-dose simvastatin on brain atrophy and disability in SPMS. For this purpose, 140 patients were randomly assigned (1:1) to receive either 80 mg of simvastatin or placebo. The trial was planned to go over 4 years with a recruitment phase of 2 years. Each of the 140 patients was followed up for 25 months. The primary outcome was the annualised rate of whole-brain atrophy measured from serial volumetric MRI. However, a secondary outcome, to be considered here, was new and enlarging T2

lesions, monitored at approximately annual MRI scans (baseline, month 12 and month 25).

3 Proposed model and Wald-type test

3.1 Modelling autoregressive negative binomial observations

We propose a statistical model which relies on a binomial thinning operation, developed by Steutel and van Harn,[16] to maintain integer based values within a time series. Suppose we observe n_C patients in the control group and n_E patients in the experiment group over a total of T time points. We will denote the observed counts of patient j of group i at time point t by $X_{ij}^{(t)}$ where $i = E, C$, $j = 1, \ldots, n_i$ and $t = 1, \ldots, T$. To model these observations in such a way, that they are dependent over time points and marginally negative binomial distributed, we reparametrise a negative binomial integrated value autoregressive process of first order (NB-INAR(1)), first proposed in McKenzie.[17] In the following, we specify the reparametrised model explicitly and show its most important properties.

Observations from the NB-INAR(1) model are given by

$$X_{ij}^{(t)} = \sum_{k=1}^{X_{ij}^{(t-1)}} B_{ijk}^{(t)}(U_{ij}^{(t)}) + W_{ij}^{(t)} \quad \text{for } t = 2, \ldots, T,\ j = 1, \ldots, n_i,\ i = E, C, \qquad (1)$$

where the random variables observed at the first time point $X_{ij}^{(1)}$ are defined as negative binomial random variables with mean λ_i and shape parameter η, i.e. $X_{ij}^{(1)} \overset{i.i.d.}{\sim} NB(\lambda_i, \eta)$. The desired marginal negative binomial property and correlation between observations $X_{ij}^{(t)}$ are attained through mixture distributions using the random variables $B_{ijk}^{(t)}(p)$, Bernoulli distributed random variables with parameter p, i.e. $B_{ijk}^{(t)}(p) \overset{i.i.d.}{\sim} Bernoulli(p)$, the random variables $U_{ij}^{(t)}$, beta distributed random variables with parameters $a\eta$ and $(1-a)\eta$, i.e. $U_{ij}^{(t)} \overset{i.i.d.}{\sim} Beta\,(a\eta, (1-a)\eta)$, and an independent random variable $W_{ij}^{(t)} \overset{i.i.d.}{\sim} NB\,((1-a)\lambda_i, (1-a)\eta)$, which is added to the random sum. It can be shown[17] that combining the random variables in this way the observations marginally follow a negative binomial distribution, i.e. $X_{ij}^{(t)} \sim NB(\lambda_i, \eta)$ for $j = 1, \ldots, n_i, i = E, C$ and $t = 1, \ldots, T$, and the covariance between observations at time points t and $t+k$ is $Cov(X_{ij}^{(t)}, X_{ij}^{(t+k)}) = a^k \cdot (\lambda_i + \lambda_i^2/\eta)$. For completeness we provide details and proofs in Appendices A and B.

The model is thus parametrised through λ_E, $\lambda_C \in (0, \infty)$, $a \in [0, 1]$ and $\eta \in (0, \infty)$. The lesion rates of the experiment group and control group are given through λ_E and λ_C respectively. The parameter a is a correlation parameter, which influences the degree to which two time points are dependent. At $a = 0$ observations are independent, while at $a = 1$ observations are fully dependent over time. The NB-INAR(1) model is further parametrised by a shape parameter η, which influences the amount of overdispersion observed in the marginal observations. Lower values of η indicate a higher degree of overdispersion. The assumption of a common shape parameter is frequently made, for example in standard statistical software (e.g. SAS, R). However, we point out that generally it would also be possible to define the model with group-specific nuisance parameters.

In contrast to alternative models which also allow for dependent count data with negative binomial marginals, e.g. serially correlated gamma frailty models,[12,14] the model does not include subject-specific covariates or time trends. However, it has the significant advantage that the full likelihood can be computed, marginal observations can be fully dependent, and closed-form method-of-moment estimators are available. It also has a mechanistic interpretation which would be appropriate for T1 counts in MS trials, as follows. A random proportion of lesions observed at time point t continues to time point $t + 1$, symbolized through the random sum in (1), together with independently occurring new lesions given through $W_{ij}^{(t)}$. For T2 lesions this mechanistic interpretation does not hold, as T2 lesions rarely disappear. Nevertheless, having correlated counts and marginal negative binomial distribution are appropriate properties for modeling T2 lesion counts.

3.2 Fitting the proposed model to data from the example studies

To give an indication of how data from the NB-INAR(1) model compares to real data from our examples, we considered MRI lesion counts from the previously mentioned MS studies.[29,39] We compare the calculated mean and shape parameter implied by the model with observed marginal means and shape parameters attained by negative binomial regression. Pointwise 95% confidence intervals for marginal estimates are also provided. Further, the mean correlation of two consecutive time points is compared to the correlation imposed by the NB-INAR(1) model. Figure 1 shows the observed data and compares the fitted model to observed mean and shape parameter.

The NB-INAR(1) model appears to capture the marginal mean and shape parameter well for the data from Tubridy et al.[39] and Chataway et al.[29] alike. Model means are very close to the empirically observed means. Furthermore, the observed means do not

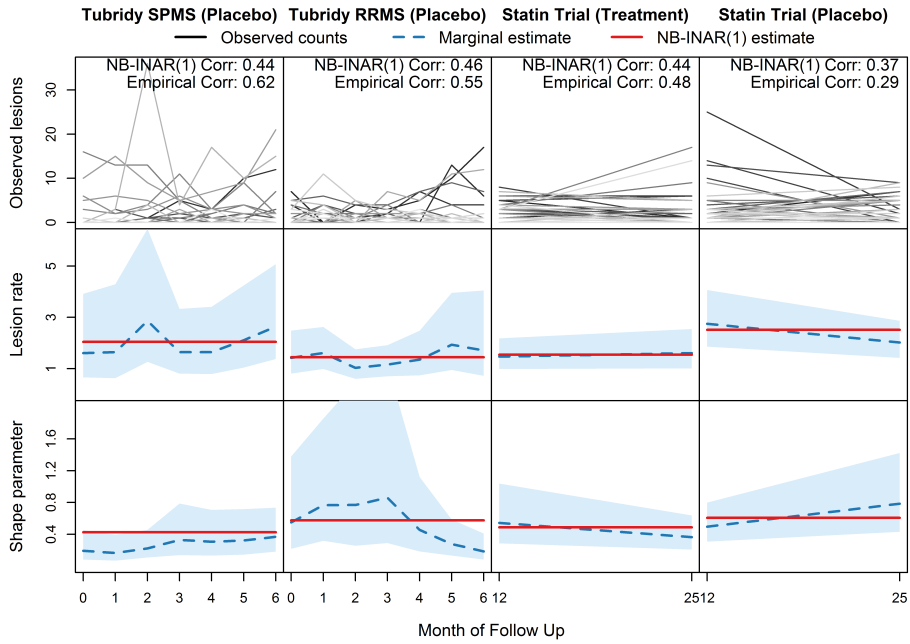

Figure 1. T2 lesion counts attained from Tubridy et al.[39] (first and second from left) and Chataway et al.[29] (third and fourth from left). The data from Tubridy et al.[39] contains new T2 lesion counts from placebo groups or patients natural history, separated according to SPMS and RRMS patients. The data was gathered over 6 months with monthly MRI scans. Data from Chataway et al.[29] contains new and enhancing T2 lesion counts from yearly MRI scans. In this case, new and enhancing lesion counts were measured 12 and 25 months after trial start.

seem to show a trend, so the assumption of constant rates seems plausible. The shape parameter observed in the data from Tubridy et al.[39] is not described as well, as it is rather heterogeneous across time points, while the NB-INAR(1) model assumes a constant shape parameter. Empirical pairwise correlations and model correlation are also different for the data from Tubridy et al.[39] This may partially be due to the relatively low sample size available. Both correlation and shape parameter are captured better for the data from Chataway et al.[29]

3.3 Derivation of a Wald-type test

We consider the NB-INAR(1) model introduced in Section 3.1, and denote the ratio of group rates by $\theta = \lambda_E/\lambda_C$. In this section we use moment estimators to develop a closed-form Wald-type statistic for testing the null hypothesis $H_0 : \theta = 1$ against the one-sided alternative $H_1 : \theta < 1$. The closed-form is especially useful when deriving a sample size formula, and would not have been attainable through a maximum likelihood approach using the observed Fisher information. The methods presented can be generalised to testing two sided hypotheses as well as absolute differences of the treatment rates.

Moment estimators for lesion rates λ_i, $i = E, C$ are given by

$$\widehat{\lambda}_i = \frac{1}{T \cdot n_i} \sum_{j=1}^{n_i} \sum_{t=1}^{T} X_{ij}^{(t)} \text{ for } i = E, C.$$

By the central limit theorem we have $\widehat{\lambda}_i \overset{\text{approx}}{\sim} N(\lambda_i, 1/n_i \cdot (\lambda_i + \lambda_i^2/\eta)\rho)$ where

$$\rho = \frac{1}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} a^{|t-s|} \tag{2}$$

is referred to as the dependency parameter. Applying the delta method yields $\log(\widehat{\lambda}_i) \overset{\text{approx}}{\sim} N(\log(\lambda_i), (\lambda_i^{-1} + \eta^{-1})\rho/n_i)$. The asymptotic distribution of $\widehat{\lambda}_i$ can be used to attain a pivotal quantity W

$$W = \frac{\log(\widehat{\lambda}_E/\widehat{\lambda}_C)}{\sqrt{\rho \cdot \left(\frac{1}{n_E}(\frac{1}{\lambda_E} + \frac{1}{\eta}) + \frac{1}{n_C}(\frac{1}{\lambda_C} + \frac{1}{\eta})\right)}} \overset{\text{approx}}{\sim} N(\mu, 1),$$

where the mean μ is given by

$$\mu = \frac{\log(\lambda_E/\lambda_C)}{\sqrt{\rho \cdot \left(\frac{1}{n_E}(\frac{1}{\lambda_E} + \frac{1}{\eta}) + \frac{1}{n_C}(\frac{1}{\lambda_C} + \frac{1}{\eta})\right)}}.$$

To attain a Wald-type test statistic Z, for testing the stated null hypothesis $H_0 : \theta = 1$, we further require estimators for the dependency parameter ρ and shape parameter η. Using the properties of the model, we can estimate ρ through the mean of all entries of the empirical correlation matrix. Since the means of the experiment and control group may be unequal, the correlation matrix is pooled accordingly. The resulting estimator $\widehat{\rho}$

is

$$\widehat{\rho} = \frac{1}{(n_1+n_2-2)T^2} \sum_{i=E,C} \sum_{t=1}^{T} \sum_{s=1}^{T} \frac{\sum_{k=1}^{n_i} \left(X_{ik}^{(t)} - \overline{X}_{i\cdot}^{(t)}\right)\left(X_{ik}^{(s)} - \overline{X}_{i\cdot}^{(s)}\right)}{\sqrt{\sum_{k=1}^{n_i} \left(X_{ik}^{(t)} - \overline{X}_{i\cdot}^{(t)}\right)^2 \sum_{k=1}^{n_i} \left(X_{ik}^{(s)} - \overline{X}_{i\cdot}^{(s)}\right)^2}}.$$

Rather than estimating λ_i and η separately and plugging the estimators into the denominator, the complete expression $1/\lambda_i + 1/\eta$ is estimated by

$$\widehat{\frac{1}{\lambda_i} + \frac{1}{\eta}} = \frac{1}{(n_i-1)T\widehat{\lambda_i}^2} \sum_{t=1}^{T} \sum_{k=1}^{n_i} \left(X_{ik}^{(t)} - X_{i\cdot}^{(t)}\right)^2.$$

Estimating the complete term, rather than plugging in separate estimates for λ_i and η, avoids the problem of negative shape parameter estimates when using moment estimators on underdispersed data. The case of no dispersion or underdispersion, however, is not typical for lesion counts of MS trials.[37]

By replacing all unknown parameters of the pivotal quantity W with their corresponding estimators, we attain a Wald-type test statistic which, by means of Slutsky's theorem, can be shown to be asymptotically normal distributed. The resulting statistic Z, for testing the null hypothesis $H_0 : \theta = 1$, is

$$Z = \frac{\log(\widehat{\lambda}_E/\widehat{\lambda}_C)}{\sqrt{\widehat{\rho} \cdot \left(\frac{1}{n_E}\widehat{(\frac{1}{\lambda_E} + \frac{1}{\eta})} + \frac{1}{n_C}\widehat{(\frac{1}{\lambda_C} + \frac{1}{\eta})}\right)}} \overset{\text{approx}}{\sim} N(0,1). \tag{3}$$

3.4 Numerical example

To demonstrate the use of the derived Wald-Test we consider the previously mentioned example data from Chataway et al.[29] Of the originally planned 140 patients, for 10 patients only baseline scans were provided. These patients were therefore discarded from the analyses presented in Chataway et al.[29] A further 8 patients had missing data on the 12 month MRI scan and 6 patients had no data on the 25 month MRI scan. We will discard the patients who did not have data on the 12 month MRI scan but consider the case of including the 6 patients with incomplete follow up. Therefore we consider two cases for inference.

1. Complete follow up only, with 55 patients in the placebo group and 61 patients in the control group.

2. Incomplete follow up, with 60 patients in the placebo group and 62 patients in the treatment group.

To assess for rate differences between control and treatment group, Chataway et al. performed a quasi-Poisson regression on the cumulative number of new lesions with adjustment for patient-specific characteristics. As the NB-INAR(1) model does not allow for covariates, for comparison purposes we repeat the quasi-Poisson regression[40] on the cumulative lesions counts without covariates, but including the number of scans as an offset variable in the case of incomplete follow up. The results are compared in Table 1 with results attained through the derived Wald-test and also with a test based on negative binomial regression[41] on the cumulative number of new lesions.

Table 1. Comparison of inference methods for testing the effect size log rate ratio (logRR) with standard error (SE) and rate ratio (RR) with 95% confidence interval (CI). Analysis are performed separately on complete and incomplete follow up. All p-values are one sided.

Follow up	Method	logRR (SE)	RR (95% CI)	Z	p-value
Complete	Quasi-Poisson	-0.39 (0.26)	0.68 (0.41-1.11)	-1.53	0.0639
	Negative Binomial	-0.39 (0.26)	0.68 (0.41-1.12)	-1.51	0.066
	NB-INAR(1) Wald-Test	-0.39 (0.26)	0.68 (0.41-1.12)	-1.51	0.065
Incomplete	Quasi-Poisson	-0.45 (0.26)	0.64 (0.38-1.07)	-1.71	0.0446
	Negative Binomial	-0.48 (0.25)	0.62 (0.37-1.01)	-1.91	0.0283
	NB-INAR(1) Wald-Test	-0.45 (0.25)	0.64 (0.39-1.05)	-1.78	0.0378

Table 1 displays the (log) rate ratio of mean relapse rates in the treatment group and the control group with standard errors for the log rate ratio, 95% confidence intervals for the rate ratios, Z-statistics and p-values (assuming normal approximation of the Z-statistic). Overall, the results appear to be similar across the methods. Especially for data with complete follow up, all three methods give almost equal results. The main difference occurs when incomplete follow up is considered for inference. The negative binomial regression with offset variable estimates a slightly larger effect size than Quasi-Poisson with offset and the NB-INAR(1) Wald-Test, resulting in a lower p-value.

4 Sample size formula and blinded sample size reestimation

4.1 Sample size: pre-trial calculation

Next we consider the derivation of a sample size formula given test size α, target power $1-\beta$ and specified clinically important effect size θ^*. Having the test statistic Z (3) in closed form means this is a relatively straightforward calculation. Given the null hypothesis is true, Z approximately follows a standard normal distribution and we simply need to solve $\Phi(z_\alpha - \mu^*) = 1-\beta$, where $\Phi(.)$ is the N(0,1) distribution function, z_α is the appropriate critical value and μ^* is the mean of Z at θ^*. Letting $k = n_E/n_C$ this gives

$$n_C = \frac{(z_\beta + z_\alpha)^2 \cdot \rho}{\log(\theta^*)^2} \left(\frac{(1+k\theta^*)^2}{(1+k)k\theta^*\overline{\lambda}} + \frac{1}{\eta}\left(1+\frac{1}{k}\right)\right), \tag{4}$$

where $\overline{\lambda} = (k\lambda_E + \lambda_C)/(k+1)$ is the overall lesion rate. In the special case of independent observations, i.e. $a = 0$, the formula simplifies with $\rho = 1/T$. Therefore, adding an observation time point to the experiment is equivalent to recruiting further $n_C + n_E$ patients. In the case of fully dependent observations, i.e. $a = 1$, adding further time points to the experiment does not improve the power and therefore does not influence the sample size. This is reflected through $\rho = 1$ and the formula reduces to the simpler case of a two group comparison with one time point, considered in Friede and Schmidli.[42]

To assess the accuracy of the proposed sample size formula, we ran a small simulation study to compare actual power with target for a range of values of the nuisance parameters $\overline{\lambda}$, a and η. We took a significance level of $\alpha = 0.025$, 80% power, number of time points $T = 7$, and sample size allocation parameter $k = 1$. We assumed the true effect size matched the specified clinically relevant effect size and took $\theta^* = \theta = 0.8$. The nuisance parameter ρ was implied by a and T through (2). All settings were simulated a total of $100,000$ times. The calculated sample sizes and achieved power are displayed in Table 2. In all cases the achieved power is within simulation noise (standard deviation 0.13%) of the target. Sample sizes tend to be smaller when counts are higher or less variable, larger when there is strong within-patient dependence.

While the relevant effect θ^* is chosen based on clinical considerations, the overall lesion rate $\overline{\lambda}$, the correlation parameter a, as well as the shape parameter η are unknown nuisance parameters and have to be estimated from previous trials or guessed. The need for precise nuisance parameter estimates is even stronger in situations where the sample size is sensitive to small changes of these. To study the influence of changes in nuisance

Table 2. Sample size calculations and corresponding simulated power for selected settings.

$\overline{\lambda}$	a (ρ)	η	Calculated Sample Size ($n_C = n_E$)	Simulated Power
1.0	0.3 (0.24)	0.8	171	79.8 %
1.0	0.3 (0.24)	1.6	124	79.9 %
1.0	0.6 (0.42)	0.8	301	80.1 %
1.0	0.6 (0.42)	1.6	218	80.0 %
2.5	0.3 (0.24)	0.8	125	79.7 %
2.5	0.3 (0.24)	1.6	78	80.2 %
2.5	0.6 (0.42)	0.8	221	80.2 %
2.5	0.6 (0.42)	1.6	137	80.1 %

parameters on the sample size formula, we interpret n_C as a function of the nuisance parameters and consider the derivatives with respect to the nuisance parameters. The derivatives of $n_C(\overline{\lambda}, \eta, a)$ with respect to the overall rate $\overline{\lambda}$ and shape parameter η are

$$\frac{\partial n_C(\overline{\lambda}, \eta, a)}{\partial \overline{\lambda}} = -\frac{1}{\overline{\lambda}^2} \cdot \frac{(z_\beta + z_\alpha)^2 \rho (1 + k\theta^*)^2}{log(\theta^*)^2 (1+k) k \theta^*}$$

and

$$\frac{\partial n_C(\overline{\lambda}, \eta, a)}{\partial \eta} = -\frac{1}{\eta^2} \cdot \frac{(z_\beta + z_\alpha)^2 \rho (1 + \frac{1}{k})}{log(\theta^*)^2}.$$

These reveal, that the sample size formula is especially sensitive to changes of the overall rate $\overline{\lambda}$ and shape parameter η, when these are near zero. The higher the parameter values are, the lower the impact of small changes to the parameters on the calculate sample size. The derivative of $n_C(\overline{\lambda}, \eta, a)$ with respect to the correlation coefficient a is

$$\frac{\partial n_C(\overline{\lambda}, \eta, a)}{\partial a} = \frac{(T-1)a^{T+1} - (T+1)a^T + a(T+1) - T + 1}{(a-1)^3} \cdot \frac{(z_\beta + z_\alpha)^2}{\log(\theta^*)^2} \left(\frac{(1+k\theta^*)^2}{(1+k)k\theta^* \overline{\lambda}} + \frac{1}{\eta} \left(1 + \frac{1}{k} \right) \right).$$

Analyzing the influence of the correlation coefficient is not as straightforward. To gain some insights on the influence of the correlation coefficient on the sample size formula, we consider the two cases of independent ($a = 0$) observations, which results in

$$\lim_{a \to 0} \frac{\partial n_C(\overline{\lambda}, \eta, a)}{\partial a} = (T-1) \cdot \frac{(z_\beta + z_\alpha)^2}{\log(\theta^*)^2} \left(\frac{(1+k\theta^*)^2}{(1+k)k\theta^* \overline{\lambda}} + \frac{1}{\eta} \left(1 + \frac{1}{k} \right) \right)$$

and fully dependent ($a = 1$) observations resulting in

$$\lim_{a \to 1} \frac{\partial n_C(\overline{\lambda}, \eta, a)}{\partial a} = \frac{1}{6} T(T^2 - 1) \cdot \frac{(z_\beta + z_\alpha)^2}{\log(\theta^*)^2} \left(\frac{(1 + k\theta^*)^2}{(1+k)k\theta^* \overline{\lambda}} + \frac{1}{\eta} \left(1 + \frac{1}{k} \right) \right).$$

From the two special cases, we conclude that the sample size is especially sensitive to changes of the correlation parameter when it is near 1 and the number of time points is high. It is in these situations, in which the sample size is sensitive to the nuisance parameters, that a reestimation of the nuisance parameters can be highly beneficial for attaining a reliable sample size estimate.

4.2 Sample size: blinded reestimation

For blinded sample size reestimation within the NB-INAR(1) model, we require estimators for the nuisance parameters $\overline{\lambda}$, a and η applicable on blinded data. Although moment estimators can be derived, we observed that maximum likelihood estimators are more efficient and imply a smaller standard deviation of the resulting sample size estimate. This observation is supported by general asymptotic theory and by other literature in the context of blinded parameter estimation.[43]

Let $x_j^{(t)}$ denote the observations from individual $j = 1, \ldots, n_E + n_C$ at time point $t = 1, \ldots, T$. Then the likelihood function for blinded estimation is

$$L\left(x_1^{(1)}, \ldots, x_{n_E+n_C}^{(T)}\right) = \prod_{j=1}^{n_E+n_C} \left(f_{X_j^{(1)}}(x_j^{(1)}) \cdot \prod_{t=1}^{T-1} f_{X_j^{(t+1)}|X_j^{(t)}}(x_j^{(t+1)}) \right), \qquad (5)$$

where $f_{X_j^{(1)}}(x_j^{(1)})$ is the marginal probability function of the first count and the subsequent conditional probabilities $f_{X_j^{(t+1)}|X_j^{(t)}}(x_j^{(t+1)})$ exploit the Markov structure of the model (1). Letting $\mathrm{P_{NB}}(x, \mu, \phi)$ denote the probability function of a negative binomial random variable with mean μ and shape parameter ϕ, the first observation has a mixture distribution

$$f_{X_j^{(1)}}(x_j^{(1)}) = \frac{1}{1+k} \cdot \left[k \cdot \mathrm{P_{NB}}\left(x_j^{(1)}, \frac{\overline{\lambda}(1+k)}{k + 1/\theta}, \eta \right) + \mathrm{P_{NB}}\left(x_j^{(1)}, \frac{\overline{\lambda}(1+k)}{1+k\theta}, \eta \right) \right].$$

Later observations have conditional probability functions

$$f_{X_j^{(t+1)}|X_j^{(t)}}(x_j^{(t+1)}) = \frac{1}{1+k} \sum_{y=0}^{min(x_j^{(t+1)}, x_j^{(t)})} \binom{x_j^{(t)}}{y} \frac{B(a\eta + y, (1-a)\eta + x_j^{(t)} - y)}{B(a\eta, (1-a)\eta)} \cdot$$
$$\left[k\mathrm{P}_{\mathrm{NB}}\left(x_j^{(t+1)} - y, \frac{\overline{\lambda}(1+k)}{k+1/\theta}, (1-a)\eta\right) + \mathrm{P}_{\mathrm{NB}}\left(x_j^{(t+1)} - y, \frac{\overline{\lambda}(1+k)}{1+k\theta}, (1-a)\eta\right)\right]$$

where $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x,y)$ with $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Numerically maximising the likelihood constructed from (5) and the above expressions leads to estimates of $\overline{\lambda}$, η and a which can be substituted into the sample size formula (4) to attain a blinded sample size reestimate. Unlike other approaches which estimate nuisance parameters from blinded data for blinded sample size reestimation, e.g. Friede and Schmidli,[42] observations are not assumed to be from a common negative binomial distribution, but from a mixture distribution between two negative binomial distributions with unequal means.

4.3 Extension to incomplete observations

When performing blinded sample size reestimation within a running trial, the experiment supervisor will almost always be confronted with incomplete observations at an interim analysis. This is frequently due to the underlying recruitment plans for such clinical trials. Figure 2 shows a possible recruitment scheme and its consequences for a sample size review.

If a sample size review is to be performed at 12 months, we see that numerous patients still have incomplete follow up. One advantage the proposed method offers over using a summary statistic as in Friede and Schmidli,[42] is that such incomplete observations can be considered for the blinded estimation of nuisance parameters. Data on these patients can significantly improve the estimation of nuisance parameters and should therefore not be ignored. Fortunately for our model incomplete observations can be incorporated within the likelihood function, in a blinded fashion, simply by allowing subject-specific final timepoints T_j for $j = 1, \ldots, n_E + n_C$ opposed to a fixed follow up time T as in (5).

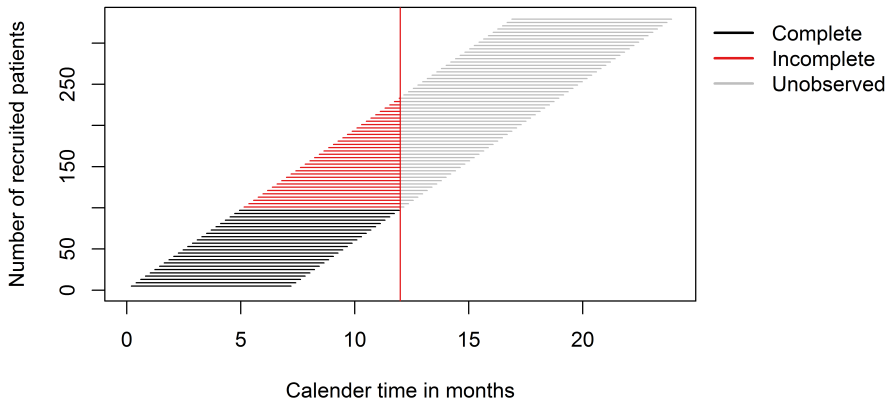

Figure 2. Possible recruitment scheme for a 24-month clinical trial, leading to incomplete observations at sample size review. Each line corresponds to one patient and the length of the line to the time in which the patient is examined. Thus, each patient is examined for a total of 7 months.

5 Example

To illustrate the initial sample size estimation and benefits of reestimation procedures, we retrospectively consider the previously mentioned simvastatin trial.[29]

5.1 Sample size: pre-trial calculation

To use the sample size formula (4) we need the design values and some knowledge of the nuisance parameters. In this example, design values are the number of time points $T = 2$, desired power $1 - \beta = 0.8$, type I error $\alpha = 0.025$ and sample size allocation factor $k = 1$. Further, the relevant effect which we would like to detect is $\theta^* = 0.6$. To estimate the nuisance parameters we need to find comparable studies from which to extract information. The simvastatin trial was planned around 2005, so the sample size estimates relied on publications before this year. However, especially in MRI studies before 2005, many different methods of counting new lesions can be found. For example, some authors count only new lesions[44] while others count new and enhancing lesions[45] or new and enlarging lesions.[46] Furthermore, there are different methods of adjusting lesions for baseline and further covariates, rendering comparisons between studies revolving around SPMS difficult if not impossible. For our illustration, estimates were based on Cohen et al.[46] This study included 385 SPMS patients in a 2-year placebo-controlled trial, with, as a secondary outcome, newly active or enlarging lesions, counted on a yearly basis. After some calculation from the summary information provided, the annual new active

lesion rate was estimated to be in the region of $\lambda_C = 1$. Then, assuming a treatment effect of $\theta = 0.6$ and $k = 1$, the estimated overall rate is $\overline{\lambda} = 0.8$. Further, a shape parameter estimate of $\eta = 0.58$ seems appropriate. Unfortunately the correlation parameter cannot be estimated from the data provided by Cohen et al.:[46] we will simply use an estimate of $a = 0.4$ (with $T = 2$, this corresponds to $\rho = 0.7$). Applying the sample size formula (4) then leads to $n_C = n_E = 129$ for our hypothetical design. Recall that the real design[29] had $n_C = n_E = 70$.

5.2 Sample size: blinded reestimation

To select an appropriate time point for an interim analysis, we require some knowledge of the underlying recruitment scheme. Figure 3 shows the recruitment scheme observed within the trial.

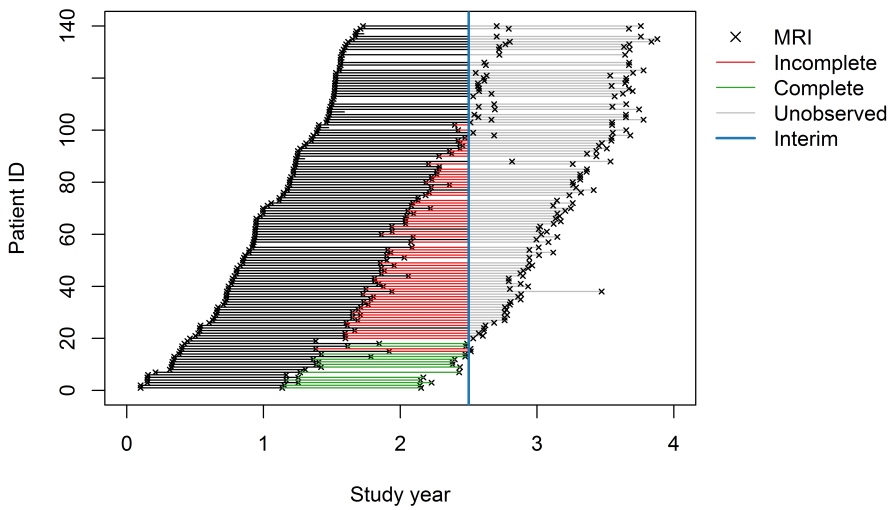

Figure 3. Recruitment scheme of the simvastatin trial.[29] The first MRI corresponds to the baseline count. The second MRI was conducted 12 months and the third MRI 25 months after baseline. A sample size reassessment 30 months after the start of the study would be possible by using the complete observations (green) and the incomplete observations (red).

From the recruitment scheme it seems that the earliest sensible time point for an interim analysis is after about 2.5 years into the study. At earlier time points there would be no patients with two consecutive observations after baseline, making it impossible to estimate the correlation parameter. Unfortunately, in this study, recruitment would have

already ended when performing the interim analysis. It would have been more desirable to perform a sample size reassessment after about 1.5 years, but doing so would have required MRI scans every 6 months. From this example we note that planning of an interim analysis for sample size reassessment requires taking into account the recruitment speed of patients as well as the frequency at which MRI scans are taken.

Assuming a sample size reassessment at 2.5 years leads to reestimated values of $n_C = n_E = 97$. The nuisance parameters at 2.5 years were estimated to be $\overline{\lambda} = 2.70$, $\eta = 0.56$ and $a = 0.42$. Thus, the planned sample size of 129 patients in each group is too high and the annual new active lesion rate and shape parameter were both initially underestimated. The difference between the marginal distribution assumed prior to the study and the marginal distribution estimated at the interim analysis is most noticeable in the probability of observing 0 new lesion counts. For the initially assumed distribution, the probability of observing 0 lesion counts is approximately 56%, much higher than the observed distribution at the interim analysis. At interim, the probability of observing 0 new lesions in a patient was approximately 34%.

6 Operating characteristics of the blinded sample size reestimation

We conducted a two-part simulation study to assess the effectiveness of the proposed BSSR procedure. In the first part, the fixed design is compared to the BSSR procedure in terms of significance level and power. In the second part, benefits of including patients with incomplete follow up for the sample size reestimation are examined. Throughout we took $T = 7$, one-sided test size $\alpha = 0.025$, target power $1 - \beta = 0.8$, and allocation factor $k = 1$. For each resulting setting $10,000$ simulation runs were conducted.

6.1 Significance level simulation of BSSR procedure vs. fixed sample size estimation

In a first simulation, we compare test sizes when 1) the a priori sample size was maintained, and 2) when sample size was adjusted following the BSSR procedure after half the (initial) target number of patients have been recruited. We assume full follow up for these patients: incomplete data will be considered later. The procedure can be summarized in four steps.

1. The initial sample size is calculated by plugging in *assumed* nuisance parameters $\overline{\lambda} = 2$, $a = 0.5$ and $\eta = 1$ and relevant effect size of $\theta^* = 0.8$ into the sample size formula (4). With these parameters, the estimated sample size of the control and treatment group are both given by $n_{est} = 165$.
2. Next $n_{est}/2$ observations are simulated for the control and treatment group respectively, corresponding to the first half of patients. The data is simulated under the null of no treatment effect ($\theta = 1$) and various selections of *true* nuisance parameters. We took an array of values for the true nuisance parameters: $a = 0.3, ..., 0.7$ in steps of 0.05, $\eta = 0.5, ..., 1.5$ in steps of 0.1 and $\overline{\lambda} = 1.5, ..., 2.5$ in steps of 0.1. Thus, in total 1089 different settings were considered. Because all a priori fixed sample sizes were calculated by using the same assumed nuisance parameters, these are misspecified in all but one of the settings.
3. A blinded sample size reestimation is performed on the simulated observations, to attain a reestimated sample size n_{reest}.
4. Finally, the attained significance level of the fixed design based on the initial $n_{est} = 165$ observations in each group is compared to the significance level of the BSSR design based on n_{reest} observations in each group. The number of observations in the BSSR design is allowed to be higher or smaller than n_{est}, but not smaller than the number of observations at the interim analysis ($n_{est}/2$).

Table 3 shows estimated test sizes under the null of no treatment effect for some of the simulated settings. In all cases the achieved test size is within simulation noise of the target 2.5%, and there is no evidence of difference in size between the two procedures: the mean values were 2.59% for fixed sample size and 2.58% for BSSR.

6.2 Power simulation of BSSR procedure vs. fixed sample size estimation

In further simulations to assess the test power, we again assume nuisance parameters $\overline{\lambda} = 2$, $a = 0.5$ and $\eta = 1$ when calculating the fixed sample size, but generate data under alternative true values. The procedure is as follows.

1. A sample size estimation is performed by plugging the assumed nuisance parameters and $\theta^* = 0.8$ into the sample size formula (4). Using the assumed nuisance parameters, the estimated sample size of the control and treatment group are both given by $n_{est} = 165$.

Table 3. Comparison of type I error for $\alpha = 0.025$ when using blinded sample size estimation (bold font) and using only the fixed design (normal font).

		$\overline{\lambda}$							
a	η	1.5		1.9		2.1		2.5	
0.30	0.50	**2.46**	2.38	**2.52**	2.53	**2.71**	2.78	**2.64**	2.77
	0.70	**2.45**	2.51	**2.63**	2.82	**2.70**	2.68	**2.82**	2.59
	0.90	**2.71**	2.38	**2.54**	2.37	**2.65**	2.87	**2.68**	2.53
	1.10	**2.67**	2.64	**2.57**	2.50	**2.57**	2.67	**2.70**	2.83
	1.30	**2.63**	2.44	**2.57**	2.30	**2.42**	2.31	**2.46**	2.39
	1.50	**2.56**	2.58	**2.44**	2.60	**2.59**	2.53	**2.87**	2.72
0.40	0.50	**2.73**	2.49	**2.83**	2.60	**2.60**	2.38	**2.56**	2.64
	0.70	**2.76**	2.92	**2.76**	2.79	**2.25**	2.44	**2.59**	2.70
	0.90	**2.46**	2.45	**2.58**	2.62	**2.46**	2.37	**2.59**	2.55
	1.10	**2.66**	2.65	**2.55**	2.55	**2.80**	2.62	**2.64**	2.58
	1.30	**2.54**	2.54	**2.68**	2.53	**2.67**	2.71	**2.48**	2.60
	1.50	**2.51**	2.52	**2.60**	2.67	**2.72**	2.75	**2.60**	2.64
0.50	0.50	**2.41**	2.34	**2.66**	2.50	**2.44**	2.27	**2.59**	2.78
	0.70	**2.68**	2.78	**2.43**	2.78	**2.71**	2.58	**2.57**	2.74
	0.90	**2.39**	2.49	**2.73**	2.71	**2.57**	2.48	**2.55**	2.47
	1.10	**2.39**	2.53	**2.37**	2.54	**2.77**	2.74	**2.42**	2.40
	1.30	**2.65**	2.60	**2.68**	2.54	**2.57**	2.68	**2.53**	2.45
	1.50	**2.60**	2.69	**2.49**	2.47	**2.71**	2.53	**2.69**	2.79
0.60	0.50	**2.65**	2.42	**2.46**	2.70	**2.51**	2.48	**2.37**	2.67
	0.70	**2.43**	2.63	**2.63**	2.57	**2.57**	2.89	**2.66**	2.99
	0.90	**2.64**	2.59	**2.73**	2.44	**2.58**	2.54	**2.71**	2.96
	1.10	**2.57**	2.62	**2.61**	2.82	**2.75**	2.90	**2.60**	2.67
	1.30	**2.34**	2.33	**2.79**	2.55	**2.66**	2.56	**2.28**	2.19
	1.50	**2.57**	2.68	**2.55**	2.49	**2.68**	2.73	**2.69**	2.47
0.70	0.50	**2.59**	2.75	**2.43**	2.65	**2.50**	2.65	**2.58**	2.86
	0.70	**2.70**	2.69	**2.76**	2.61	**2.43**	2.94	**2.28**	2.77
	0.90	**2.54**	2.45	**2.65**	2.34	**2.56**	2.53	**2.60**	2.87
	1.10	**3.08**	2.81	**2.51**	2.54	**2.37**	2.39	**2.72**	2.71
	1.30	**2.48**	2.44	**2.67**	2.54	**2.43**	2.42	**2.76**	2.66
	1.50	**2.82**	2.72	**2.80**	2.80	**2.35**	2.49	**2.57**	2.56

2. Next $n_{est}/2$ observations are simulated for the control and treatment group respectively, corresponding to the first half of patients, and with a true treatment effect of $\theta = 0.8$ and various selections of nuisance parameters.
3. A blinded sample size reestimation is performed on these observations, upon which the number of further simulated observations is increased according to the reestimated sample size n_{reest}.
4. Finally, the attained power of the fixed design based on the initial $n_{est} = 165$ observations in each group is compared to the power of the BSSR design based on n_{reest} observations in each group. The number of observations in the BSSR

design is allowed to be higher or smaller than n_{est}, but not smaller than the number of observations at the interim analysis, i.e. $n_{est}/2$.

Figure 4 shows estimated power when the nuisance parameters are varied one at a time. In all cases we see that a misspecification of the nuisance parameters will lead to over- or underpowered studies when the sample size is not adjusted properly. The loss in power can be dramatic, especially when the dependency or shape parameters are misspecified. A proper sample size review however, will lead to an appropriate sample size which maintains the desired power.

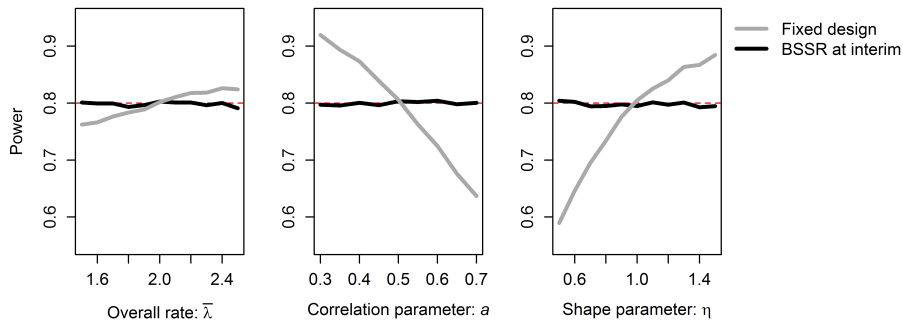

Figure 4. *Left*: The true overall rate is altered while keeping the true correlation parameter at $a = 0.5$ and shape parameter $\eta = 1$. *Center*: The true correlation parameter is altered while keeping the true overall lesion rate at $\overline{\lambda} = 2$ and shape parameter $\eta = 1$. *Right*: The true shape parameter is altered while keeping the true overall lesion rate at $\overline{\lambda} = 2$ and the true correlation parameter at $a = 0.5$.

6.3 Using observations with incomplete follow up

We now consider the effect of using observations from patients with incomplete follow up, using the same basic simulation scenario as in the previous section, but with fixed nuisance parameters $\overline{\lambda} = 2$, $a = 0.5$ and $\eta = 1$. The recruitment scheme is assumed to correspond to that in Figure 2, with 10 patients being recruited at each time point. In contrast to performing the sample size review after a specified number of patients have been fully observed, the sample size review is now performed at a specified time. Table 4 compares the sample size estimates when using BSSR with incomplete observations to BSSR without incomplete observations as the review time is varied.

Table 4. Mean and standard deviation of sample size estimates at different review time points.

	Incomplete Observations		Complete Observations	
Month of Review	Mean	SD	Mean	SD
8	164.9	15.9	164.4	23.6
10	165.1	13.1	165.0	16.5
12	165.3	11.5	165.2	13.7
14	165.2	10.4	165.1	11.8
16	165.3	9.4	165.3	10.5

The main benefit of including data from incomplete observations lies in a higher precision for the estimated sample size. In this example, taking into account incomplete observations enables a sample size review 2 months earlier, while maintaining the same precision.

7 Discussion

The presented method allows for sample size calculation, blinded sample size recalculation and inference within a reparametrised version of the NB-INAR(1) model. Within this model, the presented sample size calculations are accurate and the inference method maintains the nominal significance level. Adjustments to sample size within a running trial did not show any inflation of the significance level and may thus be considered in MS and other clinical trials with longitudinal count data. Especially in situations where the sample size estimates are sensitive to nuisance parameters and good estimates of nuisance parameters are difficult to attain for an initial sample size estimate, the blinded sample size estimation can be of great benefit. From Section 4.1 we observed, that misspecifications of the overall rate and shape parameter have a large influence when these are near zero. Further, misspecifications of the correlation parameter influence the sample size more when the number of time points is high and the correlation is near 1, than in the case of independent observations and less time points. By further incorporating incomplete observations, the precision of sample size reestimation can be markedly improved.

Although the methods work well within the NB-INAR(1) model, there is space for improvement outside of this model. For instance, a smoother estimation of the time series could be obtained by considering an autoregressive process of higher order, NB-INAR(p) with $p > 1$. For further details on how to generally extend the binomial thinning model to higher orders, we refer to McKenzie [17] and Alzaid. [47] Furthermore, our approach

does not allow for patient-specific covariates nor variation in lesion rates over time. While individual covariate information would not be available at pre-trial sample size estimation, considering covariates in inference has become popular among MS studies and population characteristics might therefore also be considered in sample size planning. Turning to time trends, in some trials it might be expected that the mean count changes systematically with time. If this is because of an initial stabilization period as treatment takes effect, then it may be sensible to use our model after a burn-in period to allow for levelling-off of trends, discarding the early time points. More generally it is possible to include time-variation in the model by allowing the parameters of the distribution of the innovation $W_{ij}^{(t)}$ in (1) to depend on time t. However, this way of incorporating a trend results in marginal distributions that are no longer negative binomial. Another possibility for introducing time variation while maintaining the property of marginal negative binomial observations, is given by gamma frailty models.[12,14] Here the conditional distribution of counts at time t for individual j, given time-varying frailty $Z_j^{(t)}$, is Poisson with mean $Z_j^{(t)}\mu_j^{(t)}$ for arbitrarily parametrised mean $\mu_j^{(t)}$. The frailties are constructed with marginal gamma distributions and between-time associations, leading to the desired characteristics of negative binomial marginals and within-patient serial correlation. Time trends and the effect of patient-specific covariates can be modelled through $\mu_j^{(t)}$. The model is limited, however, in the correlation structures that can be described and, importantly, has no closed expression for the likelihood. Overcoming these drawbacks and implementing sample size reestimation techniques within gamma frailty models may be an appropriate subject for future research.

Acknowledgements

We thank Jeremy Chataway (University College London), Richard Nicholas (Imperial College Healthcare NHS Trust) and Jennifer Nicholas (London School of Hygiene & Tropical Medicine) for providing us with the data of the MS-STAT trial.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [grant number FR 3070/1-1].

Appendix

A Proof of marginal negative binomial distribution

Assumption: For reasons of readability, we omit group and patient specific indexing within the following proof. We define following random variables:

- $U^{(t)} \overset{i.i.d.}{\sim} Beta(a\eta, (1-a)\eta)$ for $t = 2, ..., T$
- $B_k^{(t)}(p) \overset{i.i.d.}{\sim} Bernoulli(p)$ for $t = 2, ..., T$ and $k \in \mathbb{N}$
- $X^{(1)} \sim NB(\lambda, \eta)$
- $W^{(t)} \overset{i.i.d.}{\sim} NB(a\lambda, a\eta)$ for $t = 2, ..., T$
- $V^{(t)} = \sum_{k=0}^{X^{(t-1)}} B_k^{(t)}(U^{(t)})$ for $t = 2, ..., T$
- $X^{(t)} = V^{(t)} + W^{(t)}$ for $t = 2, ..., T$

Claim: Then it holds, that $X^{(t)} \sim NB(\lambda, \eta)$ for $t = 1, ..., T$.

Proof: For $t = 1$ the claim holds by definition. For $t = 2$ we first consider the distribution of $V^{(2)}$. It holds that:

$$
\begin{aligned}
&\mathbb{P}\left(V^{(2)} = y\right) = \mathbb{P}\left(\sum_{k=1}^{X^{(1)}} B_k^{(2)}(U^{(2)}) = y\right) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\left(X^{(1)} = n\right) \cdot \mathbb{P}\left(\sum_{k=1}^{n} B_k^{(2)}(U^{(2)}) = y\right) \\
&= \sum_{n=0}^{\infty} \mathbb{P}\left(X^{(1)} = n\right) \cdot \int_0^1 \mathbb{P}\left(\sum_{k=1}^{n} B_k^{(2)}(p) = y\right) \cdot \frac{p^{a\eta-1}(1-p)^{(1-a)\eta-1}}{B(a\eta, (1-a)\eta)} dp \\
&= \sum_{n=0}^{\infty} \mathbb{P}\left(X^{(1)} = n\right) \cdot \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} \cdot \frac{p^{a\eta-1}(1-p)^{(1-a)\eta-1}}{B(a\eta, (1-a)\eta)} dp \\
&= \sum_{n=0}^{\infty} \mathbb{P}\left(X^{(1)} = n\right) \binom{n}{y} B(a\eta, (1-a)\eta)^{-1} \cdot \int_0^1 p^{y+a\eta-1}(1-p)^{n-y+(1-a)\eta-1} dp \\
&= \sum_{n=0}^{\infty} \mathbb{P}\left(X^{(1)} = n\right) \binom{n}{y} \frac{B(y + a\eta, n - y + (1-a)\eta)}{B(a\eta, (1-a)\eta)} dp \\
&= \sum_{n=0}^{\infty} \left(\frac{\eta}{\eta+\lambda}\right)^{\eta} \frac{\Gamma(n+\eta)}{\Gamma(n+1)\Gamma(\eta)} \left(\frac{\lambda}{\eta+\lambda}\right)^{n}.
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\Gamma(n+1)\Gamma(y+a\eta)\Gamma(n-y+(1-a)\eta)\Gamma(\eta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(n+\eta)\Gamma(a\eta)\Gamma((1-a)\eta)} \\
&= \sum_{n=0}^{\infty} \left(\frac{\eta}{\eta+\lambda}\right)^{\eta} \frac{\Gamma(y+a\eta)}{\Gamma(y+1)\Gamma(a\eta)} \left(\frac{\lambda}{\eta+\lambda}\right)^{n} \frac{\Gamma(n-y+(1-a)\eta)}{\Gamma(n-y+1)\Gamma((1-a)\eta)} \\
&= \left(\frac{\eta}{\eta+\lambda}\right)^{\eta} \frac{\Gamma(y+a\eta)}{\Gamma(y+1)\Gamma(a\eta)} \sum_{m=0}^{\infty} \left(\frac{\lambda}{\eta+\lambda}\right)^{m+y} \frac{\Gamma(m+(1-a)\eta)}{\Gamma(m+1)\Gamma((1-a)\eta)} \\
&= \left(\frac{a\eta}{a\eta+a\lambda}\right)^{a\eta} \frac{\Gamma(y+a\eta)}{\Gamma(y+1)\Gamma(a\eta)} \left(\frac{a\lambda}{a\eta+a\lambda}\right)^{y} \cdot \\
&\qquad \underbrace{\sum_{m=0}^{\infty} \left(\frac{\lambda}{\eta+\lambda}\right)^{m} \frac{\Gamma(m+(1-a)\eta)}{\Gamma(m+1)\Gamma((1-a)\eta)} \left(\frac{(1-a)\eta}{(1-a)\eta+(1-a)\lambda}\right)^{(1-a)\eta}}_{\text{Is probability function of } NB((1-a)\lambda,(1-a)\eta) \text{ random variable}} \\
&= \left(\frac{a\eta}{a\eta+a\lambda}\right)^{a\eta} \frac{\Gamma(y+a\eta)}{\Gamma(y+1)\Gamma(a\eta)} \left(\frac{a\lambda}{a\eta+a\lambda}\right)^{y} \Rightarrow V^{(2)} \sim NB(a\lambda, a\eta)
\end{aligned}
$$

Calculating the characteristic function of $X^{(2)}$ shows:

$$
\begin{aligned}
\varphi_{X^{(2)}}(u) &= \varphi_{V^{(2)}}(u) \cdot \varphi_{W^{(2)}}(u) \\
&= \left(\frac{a\eta}{(1-e^{iu})a\lambda + a\eta}\right)^{a\eta} \cdot \left(\frac{(1-a)\eta}{(1-e^{iu})(1-a)\lambda + (1-a)\eta}\right)^{(1-a)\eta} \\
&= \left(\frac{\eta}{(1-e^{iu})\lambda + \eta}\right)^{\eta} \Rightarrow X^{(2)} \sim NB(\lambda, \eta)
\end{aligned}
$$

For $t = 3, ..., T$ the distribution of $X^{(t)} \sim NB(\lambda, \eta)$ clearly follows by induction over t.

B Proof of covariance between time points

Assumption: For reasons of readability, we omit group and patient specific indexing within the following proof. We define following random variables:

- $U^{(t)} \overset{i.i.d.}{\sim} Beta(a\eta, (1-a)\eta)$ for $t = 2, ..., T$
- $B_k^{(t)}(p) \overset{i.i.d.}{\sim} Bernoulli(p)$ for $t = 2, ..., T$ and $k \in \mathbb{N}$
- $X^{(1)} \sim NB(\lambda, \eta)$
- $W^{(t)} \overset{i.i.d.}{\sim} NB(a\lambda, a\eta)$ for $t = 2, ..., T$
- $V^{(t)} = \sum_{k=0}^{X^{(t-1)}} B_k^{(t)}(U^{(t)})$ for $t = 2, ..., T$
- $X^{(t)} = V^{(t)} + W^{(t)}$ for $t = 2, ..., T$

Claim: Then it holds, that $Cov(X_t, X_{t+s}) = a^s \cdot (\frac{\lambda^2}{\eta} + \lambda)$.

Proof:

$$
\begin{aligned}
&Cov\left(X^{(t)}, X^{(t+s)}\right) = E\left(X^{(t)} \cdot X^{(t+s)}\right) - E\left(X^{(t)}\right) \cdot E\left(X^{(t+s)}\right) \\
&= E\left(E\left(X^{(t)} \cdot X^{(t+s)} | X^{(t+s-1)}\right)\right) - E\left(X^{(t)}\right) \cdot E\left(E\left(X^{(t+s)} | X^{(t+s-1)}\right)\right) \\
&= E\left(E\left(X^{(t)} \cdot \sum_{k=1}^{X^{(t+s-1)}} B_k^{(t+s)}(U^{(t+s)}) + W^{(t+s)} | X^{(t+s-1)}\right)\right) \\
&- E\left(X^{(t)}\right) \cdot E\left(E\left(\sum_{k=1}^{X^{(t+s-1)}} B_k^{(t+s)}(U^{(t+s)}) + W^{(t+s)} | X^{(t+s-1)}\right)\right) \\
&= E\left(X^{(t)} \cdot E\left(\sum_{k=1}^{X^{(t+s-1)}} B_k^{(t+s)}(U^{(t+s)}) | X^{(t+s-1)}\right)\right) - \\
&\qquad E\left(X^{(t)}\right) \cdot E\left(E\left(\sum_{k=1}^{X^{(t+s-1)}} B_k^{(t+s)}(U^{(t+s)}) | X^{(t+s-1)}\right)\right) \\
&+ E\left(X^{(t)} \cdot W^{(t+s)}\right) - E\left(X^{(t)}\right) E\left(W^{(t+s)}\right) \\
&= E\left(X^{(t)} \cdot X^{(t+s-1)} \cdot E\left(B_k^{(t+s)}(U^{(t+s)})\right)\right) - \\
&\qquad E\left(X^{(t)}\right) \cdot E\left(X^{(t+s-1)} E\left(B_k^{(t+s)}(U^{(t)})\right)\right) \\
&\qquad + E\left(X^{(t)} \cdot W^{(t+s-1)}\right) - E\left(X^{(t)}\right) \cdot E\left(W^{(t+s-1)}\right) \\
&= E\left(B_k^{(t+s)}(U^{(t)})\right) \cdot \left(E\left(X^{(t)} X^{(t+s-1)}\right) - E\left(X^{(t)}\right) E\left(X^{(t+s-1)}\right)\right) \\
&= a \cdot Cov\left(X^{(t)}, X^{(t+s-1)}\right) \\
&\Rightarrow Cov\left(X^{(t)}, X^{(t+s)}\right) = a^s \cdot Cov\left(X^{(t)}, X^{(t)}\right) = a^s \cdot \left(\frac{\lambda^2}{\eta} + \lambda\right)
\end{aligned}
$$

References

1. European Medicines Agency (EMEA). Guideline on clinical investigation of medicinal products for the treatment of multiple sclerosis. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500185161.pdf, 2015. Accessed: 2017-01-25.

2. Sutradhar B. An overview on regression models for discrete longitudinal responses. *Statistical Science* 2003; 18: 377–393.
3. Cox D and Miller H. The theory of stochastic processes. London: Methuen and Co. Ltd. X, 398 p. (1965)., 1965.
4. Jacobs P and Lewis P. Discrete time series generated by mixtures. I: Correlational and runs properties. *J R Stat Soc, Ser B* 1978; 40: 94–105.
5. Jacobs PA and Lewis PA. Discrete time series generated by mixtures II: Asymptotic properties. *J R Stat Soc, Ser B* 1978; 40: 222–228.
6. Grunwald G, Bruce S, Jiang L et al. A statistical model for under- or overdispersed clustered and longitudinal count data. *Biometrical Journal* 2011; 53: 578–594.
7. Altman RM and Petkau AJ. Application of hidden Markov models to multiple sclerosis lesion count data. *Stat Med* 2005; 24(15): 2335–2344.
8. Altman RM, Petkau AJ, Vrecko D et al. A longitudinal model for magnetic resonance imaging lesion count data in multiple sclerosis patients. *Stat Med* 2012; 31(5): 449–469.
9. Altman R, Petkau A, Vrecko D et al. MRI-based clinical trials in relapsing-remitting MS: new sample size calculations based on a longitudinal model. *Multiple Sclerosis Journal* 2012; 18(11): 1600–1608. DOI:10.1177/1352458512444326. URL http://dx.doi.org/10.1177/1352458512444326. PMID: 22495948, http://dx.doi.org/10.1177/1352458512444326.
10. Jowaheer V and Sutradhar B. Analysing longitudinal count data with overdispersion. *Biometrika* 2002; 89: 389–399.
11. Genest C and Neslehova J. A primer on copulas for count data. *Astin Bulletin* 2007; 37: 475–515.
12. Henderson R and Shimakura S. A serially correlated gamma frailty model for longitudinal count data. *Biometrika* 2003; 90(2): 355–366. DOI:10.1093/biomet/90.2.355.
13. Hall P, Mueller H and Yao F. Modelling sparse generalised longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society, Series B* 2008; 70: 703–723.
14. Fiocco M, Putter H and Van Houwelingen JC. A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics* 2009; 10(2): 245–57. DOI:10.1093/biostatistics/kxn031. URL http://dx.doi.org/10.1093/biostatistics/kxn031.
15. Canale A and Dunson D. Nonparametric bayes modelling of count processes. *Biometrika* 2013; 100: 801–816.
16. Steutel F and van Harn K. Discrete analogues of self-decomposability and stability. *Ann Probab* 1979; 7: 893–899. DOI:10.1214/aop/1176994950.

17. McKenzie E. Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv Appl Probab* 1986; 18: 679–705. DOI:10.2307/1427183.
18. Al-Osh M and Alzaid A. First-order integer-valued autoregressive (INAR(1)) process. *J Time Ser Anal* 1987; 8: 261–275. DOI:10.1111/j.1467-9892.1987.tb00438.x.
19. Wittes J and Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* 1990; 9(1-2): 65–72. DOI:10.1002/sim.4780090113. URL http://dx.doi.org/10.1002/sim.4780090113.
20. Food and Drug Administration (FDA). Guidance for industry-adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf, 2010. Accessed: 2016-12-23.
21. European Medicines Agency (EMEA). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf, 2007. Accessed: 2016-12-23.
22. Food and Drug Administration (FDA). Guidance for clinical trial sponsors. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf, 2006. Accessed: 2016-12-23.
23. McDonald WI, Compston A, Edan G et al. Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis. *Annals of Neurology* 2001; 50(1): 121–127. DOI:10.1002/ana.1032. URL http://dx.doi.org/10.1002/ana.1032.
24. Chataway J, Nicholas R, Todd S et al. A novel adaptive design strategy increases the efficiency of clinical trials in secondary progressive multiple sclerosis. *Mult Scler* 2011; 17(1): 81–88.
25. Simon JH. Contrast-enhanced MR imaging in the evaluation of treatment response and prediction of outcome in multiple sclerosis. *J Magn Reson Imaging* 1997; 7(1): 29–37.
26. Grossman RI and McGowan JC. Perspectives on multiple sclerosis. *American Journal of Neuroradiology* 1998; 19(7): 1251–65. URL http://www.ajnr.org/content/19/7/1251.short. http://www.ajnr.org/content/19/7/1251.full.pdf+html.
27. McFarland HF, Barkhof F, Antel J et al. The role of MRI as a surrogate outcome measure in multiple sclerosis. *Mult Scler* 2002; 8(1): 40–51.
28. Cohen J and Rudick R. *Multiple sclerosis therapeutics, fourth edition*. Cambridge University Press, 2011. ISBN 9781139023986. DOI:10.1017/CBO9781139023986.

29. Chataway J, Schuerer N, Alsanousi A et al. Effect of high-dose simvastatin on brain atrophy and disability in secondary progressive multiple sclerosis (MS-STAT): a randomised, placebo-controlled, phase 2 trial. *Lancet* 2014; 383(9936): 2213–2221.
30. Olsson T, Boster A, Fernandez O et al. Oral ponesimod in relapsing-remitting multiple sclerosis: a randomised phase II trial. *J Neurol Neurosurg Psychiatr* 2014; 85(11): 1198–1208.
31. Khan O, Rieckmann P, Boyko A et al. Three times weekly glatiramer acetate in relapsing-remitting multiple sclerosis. *Ann Neurol* 2013; 73(6): 705–713.
32. Giovannoni G, Comi G, Cook S et al. A placebo-controlled trial of oral cladribine for relapsing multiple sclerosis. *N Engl J Med* 2010; 362(5): 416–426.
33. Barkhof F, Hulst HE, Drulovic J et al. Ibudilast in relapsing-remitting multiple sclerosis: a neuroprotectant? *Neurology* 2010; 74(13): 1033–1040.
34. Kappos L, Radue EW, O'Connor P et al. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *N Engl J Med* 2010; 362(5): 387–401.
35. Sormani MP, Stubinski B, Cornelisse P et al. Magnetic resonance active lesions as individual-level surrogate for relapses in multiple sclerosis. *Mult Scler* 2011; 17(5): 541–549.
36. Pozzi L, Schmidli H and Ohlssen DI. A bayesian hierarchical surrogate outcome model for multiple sclerosis. *Pharmaceutical Statistics* 2016; 15(4): 341–348. DOI:10.1002/pst.1749. URL http://dx.doi.org/10.1002/pst.1749. Pst.1749.
37. Sormani MP, Bruzzi P, Miller DH et al. Modelling MRI enhancing lesion counts in multiple sclerosis using a negative binomial model: implications for clinical trials. *J Neurol Sci* 1999; 163(1): 74–80.
38. Friede T and Schmidli H. Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Stat Med* 2010; 29(10): 1145–1156.
39. Tubridy N, Ader HJ, Barkhof F et al. Exploratory treatment trials in multiple sclerosis using MRI: sample size calculations for relapsing-remitting and secondary progressive subgroups using placebo controlled parallel groups. *J Neurol Neurosurg Psychiatr* 1998; 64(1): 50–55.
40. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.
41. Venables WN and Ripley BD. *Modern Applied Statistics with S*. Fourth ed. New York: Springer, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.
42. Friede T and Schmidli H. Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Methods Inf Med* 2010; 49(6): 618–624.

43. Schneider S, Schmidli H and Friede T. Robustness of methods for blinded sample size re-estimation with overdispersed count data. *Statistics in Medicine* 2013; 32(21): 3623–3635. DOI:10.1002/sim.5800. URL http://dx.doi.org/10.1002/sim.5800.
44. The North American Study Group on Interferon beta-1b in Secondary Progressive MS. Interferon beta-1b in secondary progressive ms: Results from a 3-year controlled study. *Neurology* 2004; 63(10): 1788–1795. DOI:10.1212/01.WNL.0000146958.77317.3E. URL http://www.neurology.org/content/63/10/1788.abstract. http://www.neurology.org/content/63/10/1788.full.pdf+html.
45. Skurkovich S, Boiko A, Beliaeva I et al. Randomized study of antibodies to IFN-gamma and TNF-alpha in secondary progressive multiple sclerosis. *Mult Scler* 2001; 7(5): 277–284.
46. Cohen JA, Cutter GR, Fischer JS et al. Benefit of interferon beta-1a on MSFC progression in secondary progressive MS. *Neurology* 2002; 59(5): 679–687.
47. Alzaid A and Al-Osh M. An integer-valued p th-order autoregressive structure (INAR(p)) process. *J Appl Probab* 1990; 27(2): 314–324. DOI:10.2307/3214650.