Biostatistics Department Technical Report

BST2009-001

Modelling Lesion Counts Data in Multiple Sclerosis
When Patients have been Selected for Baseline Activity

Charity J. Morgan, PhD
Inmaculada B. Aban, PhD
Gary R. Cutter, PhD

May 2009

Department of Biostatistics
School of Public Health
University of Alabama at Birmingham
cjmorgan@uab.edu

Modelling Lesion Counts Data in Multiple Sclerosis When Patients have been Selected for Baseline Activity

Charity J. Morgan, Inmaculada B. Aban, Gary R. Cutter

Department of Biostatistics, University of Alabama at Birmingham

May 21, 2009

Abstract

The number of new gadolinium-enhancing lesions discovered via magnetic resonance imaging (MRI) is a well-established outcome for multiple sclerosis studies and the negative binomial distribution has proven to be a valuable tool for the analysis of this type of data. Due to the high cost of MRI scans, many investigators select participants for the presence of one or more lesions on a baseline scan. While this selection procedure has been shown to improve the power of subsequent inferences, the effect of screening for baseline activity on parameter estimation and interval coverage has not yet been examined. We performed computer simulations to investigate the influence of the screening process on inferences about treatment effects in two independent samples. We demonstrate here that, while screening for baseline activity improves point estimation, it also decreases interval coverage and inflates the Type I error rate.

1 Introduction

The number of new gadolinium-enhancing lesions discovered via magnetic resonance imaging (MRI) is a well-established outcome variable for multiple sclerosis (MS) studies [1, 2]. Unfortunately, MRI is a relatively expensive technique, making the always important task of accurately calculating power and estimating sample size particularly crucial. Sormani *et al.* [3] note that the utilization of an appropriate parametric distribution for the number of lesions would improve sample size and power calculations. The negative binomial distribution (NB) has been shown to provide a good fit to this type of data [3, 4, 5], and has already begun to be used in sample size calculations [6, 7]. A particularly informative result is the confirmation that traditional nonparametric resampling methods (see for example Nauta *et al.* [8]) have the potential to significantly overestimate the power of a study [3].

For many MS studies, investigators pre-screen subjects for baseline MRI activity (i.e. the presence of one or more lesions, see for example Comi *et al.* [9]), the intuition being that subjects who show no activity on a baseline scan are less likely to show activity in future scans. If this reasoning is indeed correct, filtering out these subjects reduces the natural variation present in the sample and thus should increase the power of inferences made using the remaining subjects. Sormani *et al.* [3] investigated the result of this screening on sample size estimation and found that screening for baseline activity did in fact increase power. However, the effect of this screening on parameter estimates and interval coverage are unclear and previously unexamined. In this paper, we investigate the performance of the negative binomial distribution when patients have been selected for activity on a baseline scan.

2 The negative binomial model for lesion count

The negative binomial distribution belongs to a family of distributions, known as the mixed Poisson distributions, which model count data as Poisson random variables with random means. Let y_i be the number of new gadolinium-enhancing lesions for patient i. Conditional on a latent mean λ_i, y_i follows a Poisson distribution. We use the notation $y_i|\lambda_i \sim Poisson(\lambda_i)$. If λ_i follows a Gamma distribution with shape parameter θ and rate parameter θ/μ, denoted Gamma(θ, θ/μ), then the marginal distribution of y_i is a negative binomial:

$$\begin{aligned} f(y) &= \int_{\lambda=0}^{\infty} f(y|\lambda) f(\lambda) d\lambda \\ &= \int_{\lambda=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \frac{\theta^\theta}{\mu^\theta \Gamma(\theta)} \lambda^{\theta-1} e^{-\frac{\theta}{\mu}\lambda} \\ &= \frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)} \frac{\theta^\theta \mu^y}{(\theta+\mu)^{(\theta+y)}} \end{aligned} \tag{1}$$

We use the notation $y_i \sim NB(\mu, \theta)$. While we note that λ may be distributed as any of a wide range of distributions, including for example the inverse-Gaussian distribution, none provides as clean a representation and as nice computational properties as the Gamma distribution.

Aban *et al.* [10] extended this model to allow for the parameterization of a treatment effect, and developed tests and confidence intervals for use in MS trials that were based on the negative binomial likelihood. These procedures assume that a treatment affects only the location parameter μ of the negative binomial distribution and not the shape parameter θ. Suppose we apply a treatment to patient i such that the average number of gadolinium-enhancing lesions with treatment becomes $\gamma\lambda_i$. We refer to γ as the *treatment effect* and note that y_i is still marginally distributed as a negative binomial:

$$
\begin{aligned}
y_i|\lambda_i &\sim Poisson(\gamma\lambda_i) \\
f(y) &= \int_{\lambda=0}^{\infty} f(y|\lambda)f(\lambda)d\lambda \\
&= \int_{\lambda=0}^{\infty} \frac{(\gamma\lambda)^y e^{-\gamma\lambda}}{y!} \frac{\theta^\theta}{\mu^\theta\Gamma(\theta)} \lambda^{\theta-1} e^{-\frac{\theta}{\mu}\lambda} \\
&= \frac{\Gamma(\theta+y)}{\Gamma(\theta)\Gamma(y+1)} \frac{\theta^\theta(\gamma\mu)^y}{(\theta+\gamma\mu)^{(\theta+y)}} \\
y_i &\sim NB(\gamma\mu, \theta)
\end{aligned} \tag{2}
$$

The results of the baseline scan contain information about a patient's unobserved mean, and we can use this baseline value to update our knowledge of this latent parameter. Let $y_i^{(0)}$ be the number of lesions present on patient i's baseline scan. From Bayes theorem,

$$
\begin{aligned}
y_i^{(0)}|\lambda_i &\sim Poisson(\lambda_i) \\
f(\lambda_i|y_i^{(0)}) &\propto f(y_i^{(0)}|\lambda_i)f(\lambda_i) \\
&\propto \frac{\lambda^{y_i^{(0)}} e^{-\lambda}}{y_i^{(0)}!} \frac{\theta^\theta}{\mu^\theta\Gamma(\theta)} \lambda_i^{\theta-1} e^{-\frac{\theta}{\mu}\lambda} \\
&\propto \lambda_i^{y_i^{(0)}+\theta-1} e^{-\frac{\theta+\mu}{\mu}\lambda} \\
\therefore \lambda_i|y_i^{(0)} &\sim Gamma(y_i^{(0)}+\theta, \frac{\theta+\mu}{\theta})
\end{aligned} \tag{3}
$$

Now assume that any individual exhibiting zero lesions on the baseline scan is removed from the sample. Conditional on the fact that a subject passed the screening and remains in the sample, that subject's latent mean is no longer distributed as a Gamma random variable.

$$
\begin{aligned}
f(\lambda_i|y_i^{(0)} > 0) &= \frac{P(y_i^{(0)} > 0|\lambda_i)f(\lambda_i)}{\int_0^\infty P(y_i^{(0)} > 0|\lambda)f(\lambda)d\lambda} \\
&= \frac{(\mu+\theta)^\theta}{(\mu+\theta)^\theta - \theta^\theta} \frac{\theta^\theta}{\mu^\theta\Gamma(\theta)} \lambda_i^{\theta-1} \left(e^{-\frac{\theta}{\mu}\lambda_i} - e^{-\frac{\theta+\mu}{\mu}\lambda_i}\right)
\end{aligned} \tag{4}
$$

Consequently, the number of gadolinium-enhancing lesions is no longer marginally distributed as a negative binomial. The true distribution is much less easier to manipulate than the negative binomial distribution and thus, despite this screening, it is tempting to use the negative binomial distribution to perform sample size calculations and estimate treatment effects.

3 Methods

We implemented a procedure similar to that of Sormani *et al.* [3] and performed a series of simulations to investigate the effect of screening for baseline activity on these calculations. For a sample of size $2n$ and treatment effect γ, we repeated the following procedure 10,000 times.

1. Create a pool of potential subjects by drawing their latent means from the Gamma distribution

$$\lambda_k^* \sim Gamma(\theta, \theta/\mu),\ k = 1, \ldots, N. \tag{5}$$

   (N should be sufficiently large that at least $2n$ of the potential subjects will exhibit baseline activity. We used N = 2000. We used the values of μ and θ that Aban *et al.* [10] estimated from actual data, $\hat{\mu} = 1.646$, $\hat{\theta} = .256$).

2. Simulate the results of the baseline scan:

$$b_k^* \sim Poisson(\lambda_k^*),\ k = 1, \ldots, N. \tag{6}$$

3. Out of the subjects such that $b_k^* > 0$, randomly select n and assign them to the control group and label their latent means as $\lambda_1, \ldots, \lambda_n$. Randomly assign another n subjects to the treatment group and denote their latent means as $\lambda_{n+1}, \ldots, \lambda_{2n}$.

4. Simulate the experimental results:

$$\begin{aligned} y_i &\sim Poisson(\lambda_i), i = 1, \ldots, n \\ y_i &\sim Poisson(\gamma\lambda_i), i = n+1, \ldots, 2n \end{aligned} \tag{7}$$

Aban *et al.* [10] note that the skewed distribution of the maximum likelihood estimate of the treatment effect, $\hat{\gamma}$, resulted in inferences that were not invariant to the labeling of treatment and control groups and recommended using the $log(\hat{\gamma})$ for inference. We follow this recommendation and use the following generalized linear model to estimate the treatment effect:

$$\begin{aligned} y_i &\sim NB(\eta_i, \theta), \\ log(\eta_i) &= \beta_0 + \beta_1 x_i, i = 1, \ldots, 2n \end{aligned} \tag{8}$$

where

$$x_i = \begin{cases} 1 & \text{if } i \geq n+1 \\ 0 & \text{if } i \leq n \end{cases} \tag{9}$$

Note that with this model $y_i \sim NB(e^{\beta_0}, \theta)$ for a patient in the control group and $y_i \sim NB(e^{\beta_0}e^{\beta_1}, \theta)$ for a patient in the treatment group. Thus e^{β_0} and e^{β_1} correspond to μ and γ in our original parametrization. We perform all inferences using β_0 and β_1 instead of μ and γ; for ease in interpretation, we report results in terms of γ, rather than β_1, where appropriate. This approach has the

added benefit of allowing us to take advantage of the wide range of software already available for fitting generalized linear models. For each replication, we fit this model to the data via maximum likelihood using the 'glm.nb' routine already implemented in SPlus [11].

For each replication, we calculated the estimated treatment effect, $\hat{\gamma} = e^{\hat{\beta}_1}$, and recorded whether the 95% confidence interval for γ contained the true value of γ, where the confidence interval for γ was calculated by creating a Wald confidence interval for β_1 and exponentiating.

For each value of n and γ, we counted the number of replications for which the null hypothesis $H_0 : \beta_1 = 0$ was rejected at a significance level of $\alpha = .05$. We performed this procedure for $\gamma =$ 0.4, 0.5, and 0.6 and n $= 20, 30, \ldots, 100$ as well as n = 150 and 200. In order to estimate the Type I error rate, we also performed this procedure for these values of n and $\gamma = 1$; we then calculated the proportion of replications in which the null hypothesis of no treatment effect was erroneously rejected.

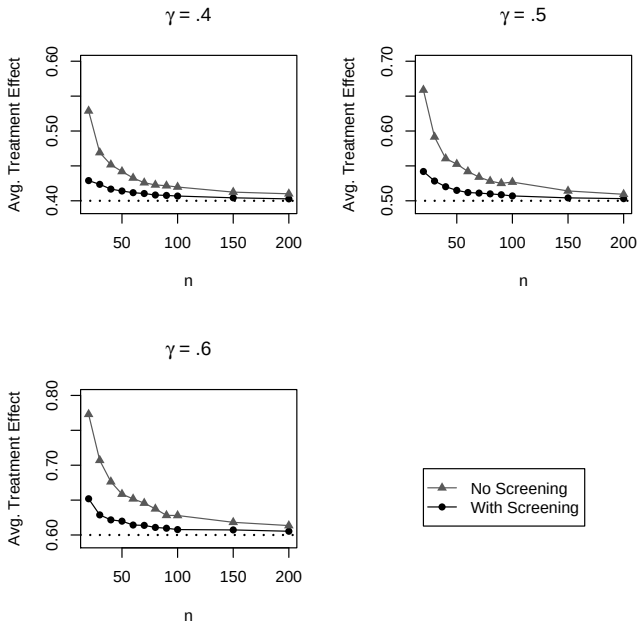
For the purposes of comparison, we repeated this procedure without screening for baseline activity. That is, in Step 3, $2n$ subjects were selected from the entire pool of potential subjects, rather than just those with positive baseline values.

4 Simulation results

4.1 Bias

For each value of γ and n, we averaged the estimated treatment effect, $\hat{\gamma}$, over the 10,000 replications. Figure 1 shows the average estimated treatment effect plotted against sample size. We see that $\hat{\gamma}$ is positively biased. This bias is present both when patients are screened for baseline activity and when no screening is performed, although the bias is less pronounced in the presence

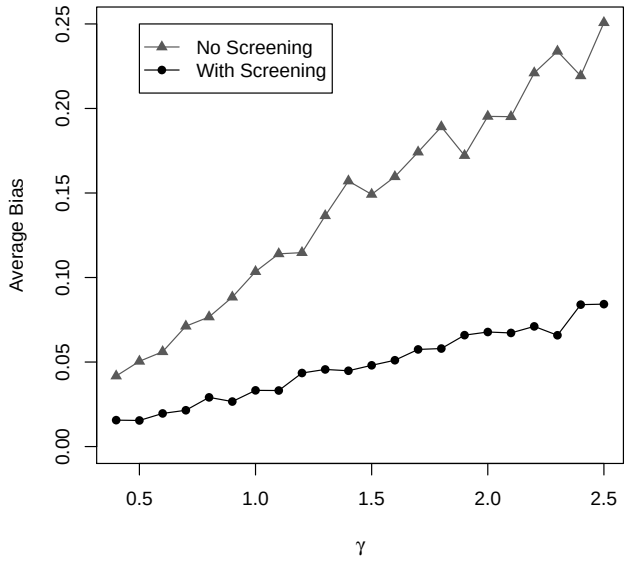
Figure 1: *Average estimated treatment effect. Dotted lines indicate the true treatment effect. Averaging across replications, we find that, regardless of whether patients are screened for baseline activity, $\hat{\gamma}$ shows positive bias. This bias decreases as the sample size increases and screening for baseline activity appears to reduce the bias.*



of screening. For both procedures, increasing the sample size reduces the bias in the estimated treatment effect.

Figure 2 displays the average bias of $\hat{\gamma}$ for $n = 50$ and varying values of γ. We find that, on average, the bias increases with γ. This instability is likely a consequence of the fact that our estimate of the treatment effect is obtained by exponentiating β_1; while $\hat{\beta}_1$ and $\hat{\gamma}$ are both asymptotically unbiased, for small n, it can be shown that the bias of $\hat{\gamma}$ is approximately a linear function of of γ:

Figure 2: *Average bias for the estimated treatment effect. For small sample size (here, $n = 50$), $\hat{\gamma}$ is positively biased. This bias increases linearly with γ. This instability in the estimation of γ is more pronounced when subjects have not been screened for baseline activity.*



$$
\begin{aligned}
E(\hat{\gamma}) &= E\left(e^{\hat{\beta}_1}\right) \\
&\approx e^{\beta_1} + e^{\beta_1} E\left(\hat{\beta}_1 - \beta_1\right) \\
&\approx \gamma + \gamma Bias\left(\hat{\beta}_1\right) \\
Bias(\hat{\gamma}) &\approx \gamma \cdot Bias\left(\hat{\beta}_1\right)
\end{aligned}
\tag{10}
$$

Screening for baseline activity provides some stabilization in the estimation of γ.

4.2 Coverage

The true coverage of the 95% confidence intervals for γ are displayed in Figure 3. We see that coverage is slightly less than 95% for both the screened data and unscreened data, and that screening for baseline activity slightly reduces coverage. This result is not surprising since, as shown above, the data resulting from this type of screening is not distributed as negative binomial, and therefore confidence intervals based on the negative binomial distribution will not be entirely accurate. Nevertheless, the negative binomial model appears to be relatively robust to this misspecified distribution.

4.3 Standard Error

Figure 4 displays the average standard error of $\hat{\beta}_1$. Screening for baseline activity appears to reduce the variability in this estimate. We can approximate the standard error of $\hat{\gamma}$ using the delta method:

Figure 3: *True coverage of confidence interval for the treatment effect. All confidence intervals have a nominal coverage of 95%. For both types of procedures, coverage is slightly less than the nominal coverage. Screening for baseline activity reduces the true coverage, however we see a slight improvement in coverage for larger sample sizes.*

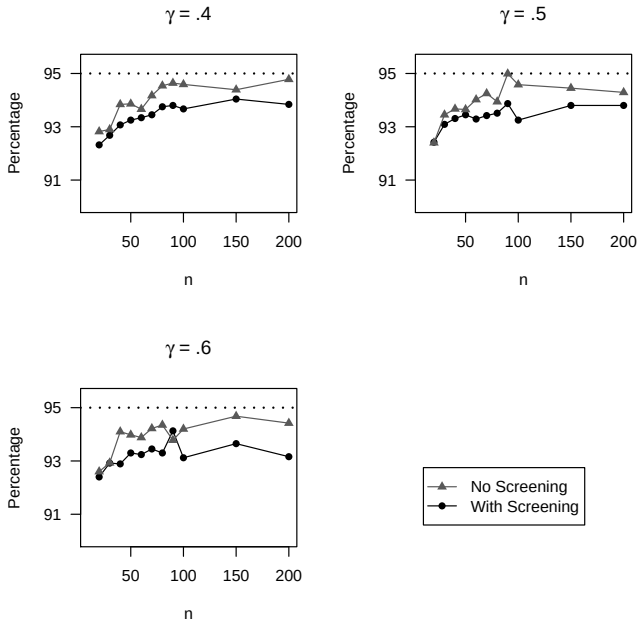
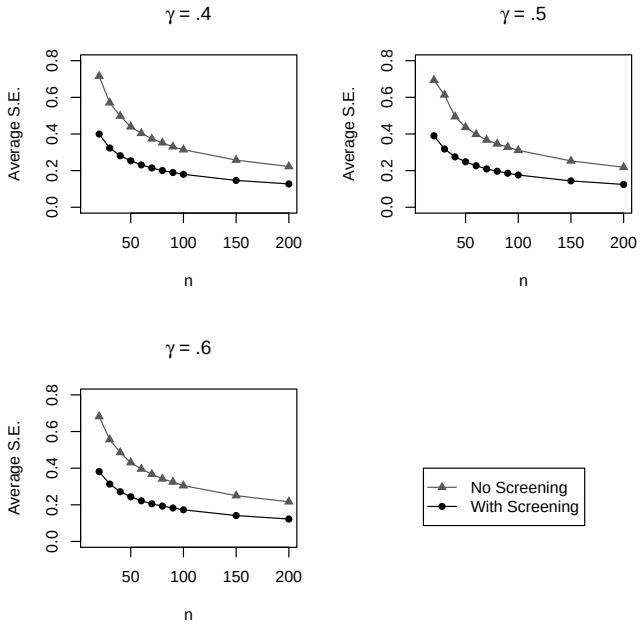


Figure 4: *Mean standard errors of the estimate of log of the treatment effect. Screening for baseline activity reduces the variability in this estimate.*



$$
\begin{aligned}
Var\left(\hat{\gamma}\right) &\approx \left(e^{E\left(\hat{\beta}_1\right)}\right)^2 \cdot Var\left(\hat{\beta}_1\right) \\
SE\left(\hat{\gamma}\right) &\approx e^{\hat{\beta}_1} \cdot SE\left(\hat{\beta}_1\right) \\
&\approx \hat{\gamma} SE\left(\hat{\beta}_1\right)
\end{aligned}
\tag{11}
$$

Figure 5 displays the average standard error of $\hat{\gamma}$. As with $\hat{\beta}_1$, screening for baseline activity reduces the variability in the estimate and the difference in variability between the two procedures decreases as sample size increases.

Figure 5: *Mean standard errors of the estimate of the treatment effect. Screening for baseline activity reduces the variability in estimation.*

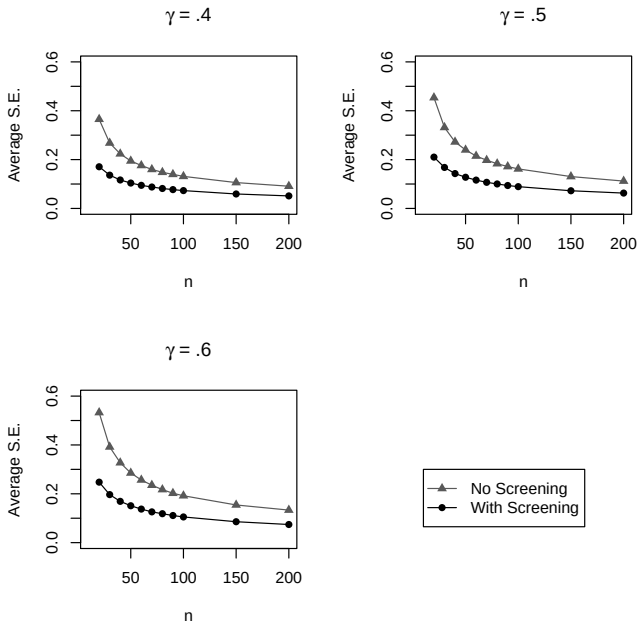
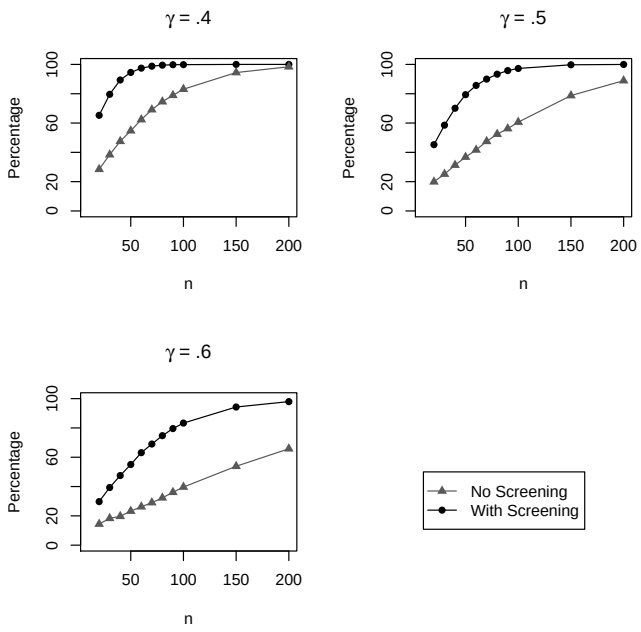


Figure 6: *Estimated power curves. Screening for baseline activity dramatically improves power.*



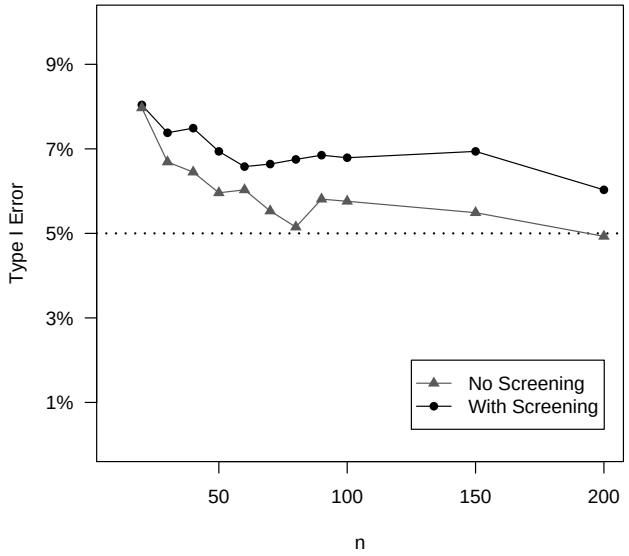
4.4 Power

Figure 6 shows the power curves for the screened and unscreened data. We find that screening for baseline activity results in a dramatic improvement in power. For example, for a true treatment effect of $\gamma = 0.4$, an estimated $2n = 180$ subjects are required to obtain approximately 80% power, if no screening is performed. In contrast, this level of power can be obtained with a sample size of only 60 if subjects are screened for baseline activity. These results align well with those reported by Sormani *et al.* [7].

4.5 Type I Error

The simulated Type I error rates are displayed in Figure 7. The probability of a Type I error is slightly greater than 5% regardless of whether subjects are screened for baseline activity. We see

Figure 7: *Simulated Type I errors. For both procedures, the Type I error rate is larger than the specified significance level of 5%. The probability of a Type I error decreases as sample size increases. We also find that screening for baseline activity increases the probability of a Type I error.*



that screening for baseline activity inflates the Type I error. For both types of procedures, the true Type I error decreases as sample size increases.

5 Application to real data

Screening for baseline MRI activity increases power because the results of the baseline scan contain information about each subject's unobserved mean number of lesions. Using these results to select subjects creates a more homogeneous sample, which makes any treatment effect more apparent. However, simply removing any subject exhibiting zero lesions at baseline uses this information in a rather crude way. It stands to reason that incorporating the results of the baseline

scan into the analysis should provide even stronger results.

We expand our previous model to include repeated measurements. Let y_{ij} be the number of gadolinium-enhancing lesions observed for patient i at time j. We apply the following generalized linear model with mixed effects:

$$\begin{aligned} y_{ij} &\sim Poisson(\eta_{ij}) \\ log(\eta_{ij}) &= \alpha_i + \beta x_{ij} \\ \alpha_i &\sim N(0, \sigma^2) \end{aligned} \tag{12}$$

where x_{ij} is a vector of covariates measured on patient i at time j. Note that this model allows us to include the results of a scan performed at baseline, as well as any other time points.

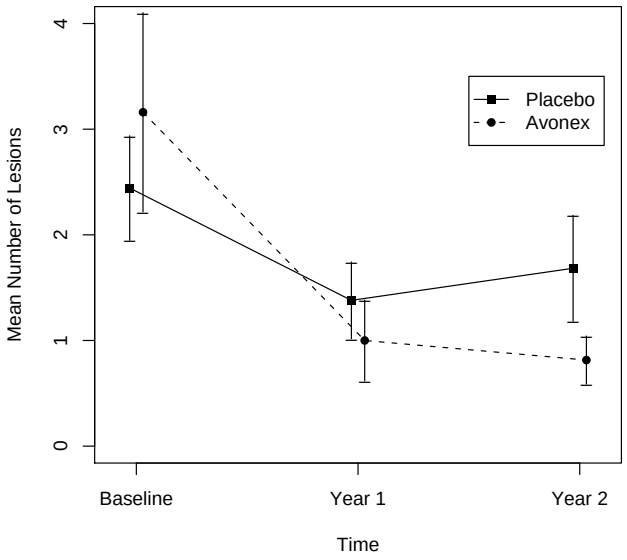
We apply this model to data from a study reported by Rudick *et al.* [12]. 172 patients were assigned to either the drug Avonex or to placebo and followed for two years. MRI was administered at baseline and at the end of each subsequent year. We consider here only those patients for whom measurements for all three time points are available (81 of 85 subjects in the Avonex group, 79 of 87 in the placebo group). Figure 8 shows the the change in average number of lesions over time for each group. Subjects in this study were not selected for baseline activity.

To test for a treatment effect, we consider the following model for the mean number of lesions:

$$\begin{aligned} log(\eta_{ij}) &= \alpha_i + \beta_0 + \beta_1 x_{ij}, \\ \text{where } x_{ij} &= \begin{cases} 1 & \text{if subject } i \text{ is receiving Avonex at time } j \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{13}$$

Under this model, the number of lesions for a subject i who is not receiving treatment at time j is distributed as $y_{ij} \sim Poisson(e^{\alpha_i + \beta_0})$, thus the expected number of lesions is $E(y_{ij}) = E(e^{\alpha_i + \beta_0}) =$

Figure 8: *Mean lesion count by time and treatment.*



$e^{\beta_0}E(e^{\alpha_i})$.[1] Let $\mu = e^{\beta_0}E(e^{\alpha_i})$. Similarly, if patient i is receiving treatment at time j, then $y_{ij} \sim Poisson(e^{\alpha_i+\beta_0+\beta_1})$ and $E(y_{ij}) = E(e^{\alpha_i+\beta_0+\beta_1}) = e^{\beta_1}e^{\beta_0}E(e^{\alpha_i}) = e^{\beta_1}\mu$. As before, we see that if $\gamma = e^{\beta_1}$, then we can use this model to test for a treatment effect by the testing the null hypothesis $H_0 : \beta_1 = 0$, which is equivalent to $H_0 : \gamma = 1$.

We also consider a modification of (12) that replaces the Poisson distribution used to model the number of lesions with a negative binomial with random mean:

$$\begin{aligned} y_{ij} &\sim NB(\eta_{ij}, \theta) \\ log(\eta_{ij}) &= \alpha_i + \beta x_{ij} \\ \alpha_i &\sim N(0, \sigma^2) \end{aligned} \tag{14}$$

[1] $\alpha_i \sim$ N(0, σ^2) and therefore e^{α_i} follows a lognormal distribution, with E(e^{α_i}) = $e^{\sigma^2/2}$.

Table 1: *Comparison of the Poisson-based and negative binomial-based models. We use the AIC to compare the fits of these two models. According to the AIC, the negative binomial model provides a better fit to these data.*

Distribution	Log-likelihood	AIC
Poisson	-813.05	1632.1
Negative Binomial	-704.15	1416.3

Note that as before, we assume that the treatment has no effect on the shape parameter θ. The interpretation of β_0 and β_1 is the same as in the Poisson-based model.

We used the SAS procedure PROC NLMIXED to fit these two models. The Akaike information criterion [AIC, 13] and log-likelihoods are displayed in Table 1. We use the AIC to determine the better-fitting model; The negative binomial-based model has an AIC (1416.3) closer to zero than the Poisson-based model (1632.1), so we select this distribution as providing a superior fit and report results specific to this model. The results for the negative binomial-based model are shown in Table 2. We find a significant treatment effect, $\hat{\gamma} = e^{\hat{\beta}_1} = 0.301$ ($p < .0001$). Exponentiating the confidence interval for β_1 yields a 95% confidence interval for γ : $(0.199,\ 0.455)$. The estimated mean (standard error) number of lesions for patients not receiving treatment is $\hat{\mu} = 2.571$ (0.553); the estimated mean (standard error) number of lesions for those receiving treatment is $\hat{\mu} \cdot \hat{\gamma} = 0.774$ (0.188).

This model assumes that the effect of the treatment is constant over time. It may however be the case that the effect of the treatment may change from Year 1 to Year 2. We can evaluate the reasonableness of this assumption by adding another covariate to the model. Let $x = (x_1, x_2)$,

Table 2: *Parameter estimates for the generalized linear model with mixed effects for the number of lesions. Distribution of lesion count assumed to be negative binomial with random mean. SAS estimates the parameters* θ *and* σ^2 *as* $\hat{\theta}$ *= 1.101 (0.206) and* $\hat{\sigma}^2$ *= 2.099 (0.417). The log-likelihood under this model is -704.15. This model assumes the effect of the treatment is constant over time.*

	Parameter	Estimate	Standard Error	p-value
Intercept	β_0	-0.105	0.163	.519
Treatment	β_1	-1.201	0.209	<.0001

where $x_{1ij} = 1$ if subject i is receiving Avonex at time j, and $x_{1ij} = 0$ otherwise. Let x_2 be an indicator for the second year of treatment. That is, $x_{2ij} = 1$ if subject i is receiving Avonex at time j and time j is Year 2, and $x_{2ij} = 0$ otherwise. The model for the mean number of lesions then becomes

$$log(\eta_{ij}) = \alpha_i + \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}, \tag{15}$$

Under this model, the expected number of lesions without treatment is still $e^{\beta_0} E(e^{\alpha_i}) = \mu$; the mean number of lesions for a person receiving treatment is $\gamma\mu$ for the first year and $e^{\beta_2}\gamma\mu$ for the second year. If we let $\kappa = e^{\beta_2}$, then κ measures the change in treatment effect from Year 1 to Year 2. A value of κ less than 1 indicates that the treatment becomes more effective in the second year, $\kappa > 1$ corresponds to a decrease in effectiveness, and $\kappa = 1$ indicates that there is no change in the effectiveness of the treatment after the first year.

The fit of this model is shown in Table 3. There is no significant change in the treatment effect

Table 3: *Parameter estimates for the model allowing a change in treatment effect from Year 1 to Year 2. SAS estimates the parameters θ and σ^2 as $\hat{\theta}$ = 1.101 (0.206) and $\hat{\sigma}^2$ = 2.099 (0.416). The log-likelihood under this model is -704.15 and thus the addition of another parameter does not result in a significant increase in the log-likelihood.*

	Parameter	Estimate	Standard Error	p-value
Intercept	β_0	-0.105	0.163	.519
Treatment	β_1	-1.176	0.256	<.0001
Year 2	β_2	-0.050	0.296	.867

from Year 1 to Year 2 ($\hat{\kappa}$ = .952, p = .867). Furthermore, the addition of β_2 to the model causes the AIC to increase to 1418.3 from 1416.3, and thus we select (14) as our final model.

Out of the 160 subjects for whom measurements at all time points were available, 72 (35 of the 79 subjects receiving the placebo and 37 of the 81 subjects receiving Avonex) exhibited zero lesions during the baseline scan. We illustrate the effect screening for baseline activity would have had on our conclusions by performing a hypothetical, post-hoc scan and conducting the above analyses on the remaining 88 (44 placebo and 44 Avonex) participants who exhibited at least one lesion at baseline.

We first compare the fits of the Poisson-based mixed effect generalized linear model (12) to its negative binomial-based counterpart (14), using the AIC. As before, the AIC identifies the negative binomial-based model (AIC = 1087.2) as providing a better fit than the Poisson-based model (AIC = 1287.2), and thus we use the negative binomial distribution to perform all further analyses.

The results for the negative binomial-based model when applied to the screened data are sum-

Table 4: *Comparison of results for mixed effect generalized linear model fit to the data both with and without baseline screening. These models assume the effect of the treatment is constant over time.*

	Full Data	Screened Data
$\hat{\gamma}$	0.301 (0.063)	0.287 (0.060)
95 % C.I. for γ	(.199, .455)	(.189, .434)
$\hat{\mu}$	2.571 (0.553)	4.053 (0.654)
$\hat{\theta}$	1.101 (0.206)	1.331 (0.244)
$\hat{\sigma}^2$	2.099 (0.417)	0.851 (0.202)
sample size	160	88

marized and presented alongside the results for the full data set in Table 4. We again find a significant treatment effect: $\hat{\gamma} = .287 (p < .0001)$. This value does not significantly differ from the effect size found when all subjects are considered (p = .815).

While screening for baseline activity does not significantly change our estimate of the treatment effect, we find a significant increase in $\hat{\mu}$ (p = .023), where the estimated mean number of lesions increases from 2.571 to 4.053 for patients not receiving treatment, and from 0.774 to 1.162 for patients receiving treatment. This discrepancy suggests that, although screening for baseline activity can accurately estimate the treatment effect with a reduced sample size, the estimated average lesion count is not generalizable to the general population.

We then refit the model allowing for a nonconstant treatment effect. As before there was no significant change in the treatment effect from Year 1 to Year 2 ($\hat{\kappa} = .871, p = .647$), and the addition of this term to the model causes the AIC to increase (from 1087.2 to 1089).

6 Discussion

Magnetic resonance imaging is a necessary but expensive tool for identifying brain lesions in multiple sclerosis patients. Pre-screening subjects for baseline activity and performing statistical analyses using a negative binomial model are two tools used by MS investigators to improve the power of their studies and consequently decrease the number of MRI scans needed to demonstrate treatment efficacy. We also note that the decision to screen or not to screen may depend not only on the cost of the trial but also on many other issues such as the duration of the trial and even the proposed mechanism by which the treatment operates.

While the screening process inflates Type I error and decreases interval coverage when the treatment effect is estimated via a negative binomial model, the effects are relatively minor, in-

dicating that the negative binomial distribution is robust in the face of screening. This result is crucial, as the availability of negative binomial-based generalized linear models not only facilitates the identification of treatment effects but also allows the detection of changes in those treatment effects over time. This aspect of MS therapy as well as other disease treatments is important for several reasons. Using these techniques we may isolate treatment lags and in future trials where head-to-head therapies are compared, we may be able to assess which drugs are superior based on longer term results. These later temporal effects of treatment fidelity are very important in treatments for chronic diseases such as multiple sclerosis, which may involve decades of therapy.

Furthermore, screening for baseline activity reduces the bias and variability in the estimation of treatment effects. Thus, while screening for baseline activity creates a trade-off between cost effectiveness and conservative inference that must be carefully considered, this approach, when combined with negative binomial-based modelling, is a promising tool for the analysis of MRI-monitored multiple sclerosis studies.

7 Acknowledgments

This research is based partially on work supported by Grant Number T32NS054584 from the National Institute of Neurological Disorders and Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health.

References

1. McFarland HF, Frank JA, Albert PS, Smith ME, Martin R, Harris JO, Patronas N, Maloni

H, McFarlin DE. Using gadolinium-enhanced magnetic resonance imaging lesions to monitor disease activity in multiple sclerosis. *Annals of Neurology* 1992; **32**(6):758–766.

2. Sormani MP, Filippi M. Statistical issues related to the use of MRI data in multiple sclerosis. *Journal of Neuroimaging* 2007; **17**(2):56S–59S.

3. Sormani MP, Bruzzi P, Miller DH, Gasperini C, Barkhof F, Filippi M. Modelling MRI enhancing lesion counts in multiple sclerosis using a negative binomial model: Implications for clinical trials. *Journal of the Neurological Sciences* 1999; **163**:74–80.

4. Sormani MP, Bruzzi P, Rovaris M, Barkhof F, Comi G, Miller DH, Cutter GR, Filippi M. Modelling new enhancing MRI lesion counts in multiple sclerosis. *Multiple Sclerosis* 2001; **7**:298–304.

5. van den Elskamp IJ, Knol DL, Uitdehaag BMJ, Barkhof F. The distribution of new enhancing lesion counts in multiple sclerosis: Further explorations. *Multiple Sclerosis* 2009; **15**:42–49.

6. Sormani MP, Molyneux PD, Gasperini C, Barkhof F, Yousry TA, Miller DH, Filippi M. Statistical power of MRI monitored trials in multiple sclerosis: New data and comparison with previous results. *Journal of Neurology, Neurosurgery, and Psychiatry* 1999; **66**:465–469.

7. Sormani MP, Miller DH, Comi G, Barkhof F, Rovaris M, Bruzzi P, Filippi M. Clinical trials of multiple sclerosis monitored with enhanced MRI: New sample size calculations based on large data sets. *Journal of Neurology, Neurosurgery, and Psychiatry* 2001; **70**:494–499.

8. Nauta JJP, Thompson AJ, Barkhof F, Miller DH. Magnetic resonance imaging in monitoring the treatment of multiple sclerosis patients: Statistical power of parallel-groups and crossover designs. *Journal of the Neurological Sciences* 1994; **122**:6–14.

9. Comi G, Filippi M, Wolinsky JS, European/Canadian Glatiramer Acetate Study Group. European/Canadian multicenter, double-blind, randomized, placebo-controlled study of the effects of glatiramer acetate on magnetic resonance imaging-measured disease activity and burden in patients with relapsing multiple sclerosis. *Annals of Neurology* 2001; **49**:290–297.

10. Aban IB, Cutter GR, Mavinga N. Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. *Computational Statistics and Data Analysis* 2009; **53**:820–833.

11. Venables WN, Ripley BD. *Modern Applied Statistics with S*. Springer, fourth edition 2002.

12. Rudick RA, Cutter G, Baier M, Fisher E, Dougherty D, Weinstock-Guttman B, Mass MK, Miller D, Simonian NA. Use of the multiple sclerosis functional composite to predict disability in relapsing MS. *Neurology* 2001; **56**:1324–1330.

13. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**(6):716–723.