Sample size calculation for survival time data in cluster randomized trials using NIATx 200 as an example

Antje Jahn-Eimermacher* Jim Robinson† Andrew Quanbeck‡
Dennis McCarty§

Abstract
In several trials with a time to event endpoint, the interventions to be compared are randomized not to individuals but to groups of individuals (clusters), which might be patients of the same hospital. One example is the NIATx 200 trial, a cluster randomized trial investigating methods of disseminating quality improvement to addiction treatment centers in the U.S.. One of the primary endpoints to be compared between the different interventions is the time patients have to wait for their first treatment. Members of the same cluster tend to be more similar than members of different clusters causing intra-cluster correlation. Correlation affects the power of a trial and thus has to be considered when planning the sample size. We illustrate how to plan the sample size for clustered time to event data using the NIATx 200 trial as an example.

Key Words: sample size; cluster randomized; survival; time to event

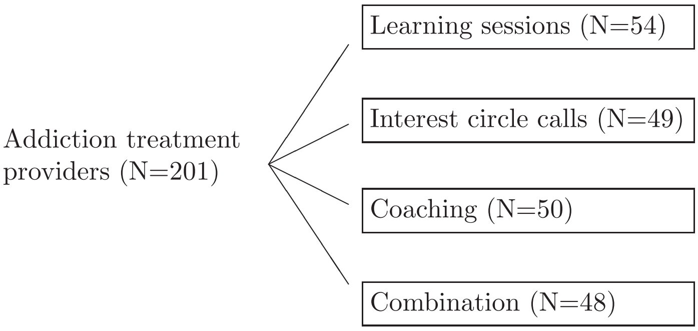
1. Introduction

In several trials with a time to event endpoint, the interventions to be compared are randomized not to individuals but to groups of individuals (clusters), which might be patients of the same hospital or practionier or children of the same school [2]. Examples are interventions involving training of health care professionals or being implemented at a hospital level. The NIATx 200 trial [9] is a cluster randomized trial investigating methods of disseminating quality improvement to addiction treatment centers in the U.S.. One of the primary endpoints to be compared between the different interventions is the time patients have to wait for their first treatment. In cluster randomized trials observations within the same cluster tend to be more similar than observations of different clusters. This correlation is most probably caused by unobserved or unobservable covariates which affect the outcome and are shared by members of the same cluster. Examples are a common social structure, the same standard of medical care or a similar lifestyle within a cluster. Extensions of the Cox proportional hazards regression allowing for clustering are well established [11]. Correlation reduces the statistical information in the data and thus the effective sample size. Therefore the clustered design also has to be considered when planning the sample size of a trial to ensure an adequate power to detect intervention effects. Methods for sample size calculation in cluster randomized trials are well established if a continous, binary or person years rate outcome is of primary interest[4, 2]. Recently a sample size formula has been proposed for time to event data as the primary outcome [6]. In the present publication, this formula will be illustrated on the NIATx 200 trial.

[*] Institute of Medical Biostatistics, Epidemiology and Informatics, Medical Center of the Johannes Gutenberg-University, Langenbeckstr.1, 55131 Mainz, Germany
[†] University of Wisconsin-Madison, Madison, WI 53706
[‡] University of Wisconsin-Madison, Madison, WI 53706
[§] Oregon Health & Science University, Portland, Oregon 97239

2. NIATx 200

Drug and alcohol treatment programs often have long delays between first contact to the treatment provider and first treatment appointment, which has been shown to decrease the probability of starting and completing treatment [5]. Research on addiction treatment has produced effective methods to improve the waiting time (days to treatment from first contact), but is disseminating them slowly. A large cluster randomized trial (NIATx 200) has been implemented to evaluate interventions for disseminating quality improvement to addiction treatment centers in the U.S. [9]. Four interventions (interest circle calls including monthly teleconferences, coaching including an initial site visit, face-to-face learning sessions and the combination of all) are randomly and balanced allocated to the treatment centers.


Addiction treatment providers (N=201)
Learning sessions (N=54)
Interest circle calls (N=49)
Coaching (N=50)
Combination (N=48)

The main outcome variables of NIATx 200 are patients' waiting time, the clinics' annual number of new patients and the average continuation rate. In the present manuscript we will focus on the pairwise comparison between interventions with respect to the patients' waiting time. Patients in centers being allocated to an ineffective intervention can be expected to wait an average time of 19.5 days for their first treatment corresponding to baseline data of a pilot study [8]. For sample size calculation we consider that a reduction of the mean waiting time by 15% to an average time of 16.6 days is considered as clinically meaningful. For an exponential distribution of the waiting time this equates to hazard rates of $1/19.5$ and $1/16.6$ in a pairwise comparison of two intervention groups, respectively, with the hazard ratio being $\frac{19.5}{16.6} = 1.17$.
The sample size of the NIATx 200 trial originally has been derived for a regression on cluster level analysing the averaged waiting times per cluster. We will illustrate our sample size formula for the pairwise comparison on the individual level, i.e. for analysing individual waiting times in an appropriate survival model.

3. Sample size determination

3.1 Notations

Assume we have a balanced trial design with N clusters per group each of size K. Subjects are recruited uniformly over an accrual period of size a and each subject is followed for an additional follow-up of length B. With i indicating clusters and k indicating the observations within clusters we define T_{ik} as the time to event, C_{ik} as the independent censoring time and $Y_{ik} = \mathbf{I}_{\{T_{ik} \leq C_{ik}\}}$ as the event indicator of subject k in cluster i. The hazard rate of subject i in cluster k for experiencing an event is denoted by λ_{ik}. Assume there is a single binary variable of interest which will be randomly and balanced allocated to whole

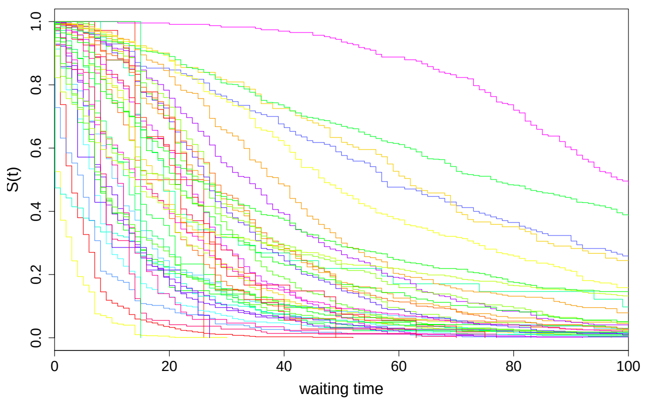


Figure 1: Kaplan-Meier estimates of the survival functions per cluster in group ”Learning sessions”. Only patients with a first treatment appointment are considered.

clusters, not to the individuals within clusters. For the NIATx 200 trial example this refers to a pairwise comparison of two interventions I_A and I_B.

3.2 Sample size calculation

A naive approach for sample size calculation ignoring the clustered design would be:

A.1 Assume a proportional hazards model

$$\lambda_{ik}(t) = \lambda_0(t)e^{\beta' W_i}$$

with β being the regression coefficient of interest and W_i being the regression covariate of interest in cluster i. Without loss of generality $W_i = 0$, if cluster i was randomized to intervention I_A and $W_i = 1$, if cluster i was randomized to intervention I_B. Assume the null hypothesis $H_0 = \{\beta = 0\}$ is to be tested at a significance level of α.

A.2 Apply Schoenfeld’s [10] sample size formula to calculate the number of clusters per group, N_0, required for a power of $1 - \gamma$ under an expected log hazard ratio of β_1 and an overall censoring probability of $P(C)$.

$$N_0 = \frac{2\,(z_{\alpha/2} + z_\gamma)^2}{\beta_1^2\,(1 - P(C))} / K$$

This approach is based on a common baseline hazard in all the centers. However, in practice a cluster heterogeneity in the baseline hazards might be more realistic, which causes within-cluster correlation and thus reduces the effective sample size of the trial. An example gives the NIATx 200 trial, where the randomized centers show a large heterogeneity in their patients waiting time even if randomized to the same intervention as can be observed from the estimated survival functions per center (Figure 1). In some centers more than 80% of the treated patients have their first appointment within 10 days, whereas in other centers this refers to less than 10% of the patients.

An approach for sample size calculation which takes the clustered design into account has recently been proposed [6]

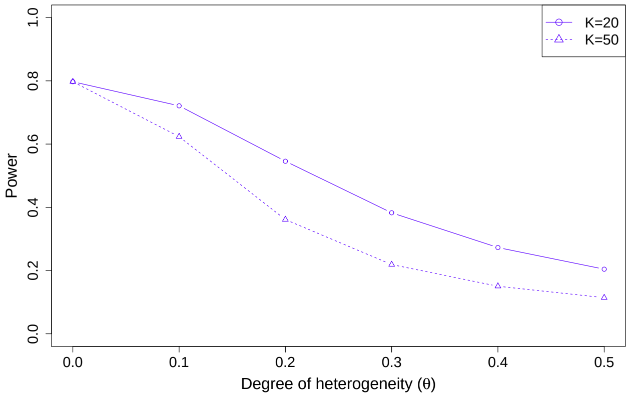


Figure 2: Power of a cluster randomized trial with $N_0 * K = 600$ subjects per group calculated by (A.2) to detect a reduction in mean waiting time from 19.5 days to 16.6 days corresponding to a HR=1.17 with anticipated 80% power, a significance level of 0.05 and no censoring.

B.1 Assume that cluster heterogeneity is caused by common unobserved or unobservable covariates affecting time to event and add a corresponding random term to the Cox model shared by the members of the same cluster (shared frailty)

$$\lambda_{ik}(t) \quad = \quad \lambda_0(t) Z_i e^{\beta' W_i}$$

$(Z_i)_{i=1\ldots2N}$ are independent identically distributed random variables with mean 1 and variance θ^2, which act multiplicatively on the marginal baseline hazard λ_0 and represent the unobserved or unobservable covariates defined on clusters.

B.2 Apply an adjusted Schoenfeld's formula

$$N \quad = \quad N_0 + (z_{\alpha/2} + z_\gamma)^2 \, \theta^2 \, \frac{1 + \exp(\beta_1)^2}{(1 - \exp(\beta_1))^2}$$

with N_0 derived by (A.2).

To demonstrate the importance of the correction term in formula (B.2) we calculate the power of a trial with sample size planned under a misspecified proportional hazards assumption using formula (A.2) in the presence of cluster heterogeneity. The power is calculated by inverting formula (B.2). For a fixed total sample size $N_0 * K$, which reaches the anticipated power of 80% in the homogeneous case ($\theta = 0$), the power decreases with increasing heterogeneity (Figure 2). Cluster heterogeneity more affects the power if there are only some independent clusters of large size than if there are more independent clusters of smaller size.

Figure 3 illustrates the sample size according to the adjusted Schoenfeld's formula (B.2) for different cluster sizes, K, and different assumptions on the hazard ratio. The impact of cluster heterogeneity on sample size increases with decreasing hazard ratio as can be observed from the slope of the sample size curves.

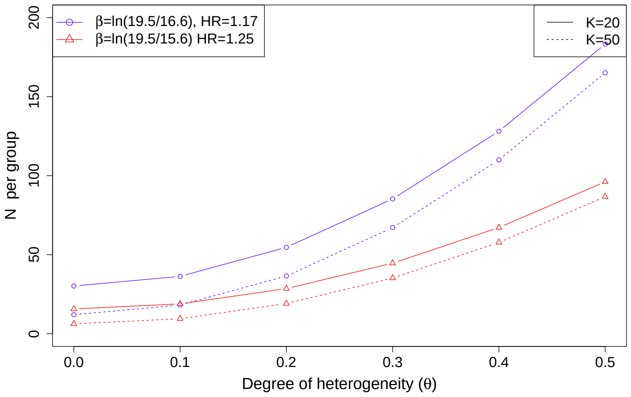


Figure 3: NIATx200: Required number of clusters per group according to (B.2) for a pairwise comparison to detect a reduction in mean waiting time from 19.5 days to 16.6 and 15.6 days corresponding to a HR=1.17 and 1.25, respectively, with 80% power, a significance level of 0.05 and no censoring.

3.3 Specification of θ

Usually, in the design phase of a cluster randomized trial, not much is known about the degree of heterogeneity. Data of a comparable cluster randomized trial or a pilot study can be used to estimate θ, if available. Additionally, characteristics of the frailty distribution might also help to find a reasonable assumption on θ for the sample size calculation. The most common distributional assumptions for the frailty variable are the gamma and log-normal distribution, the former mainly for mathematical convenience. The log-normal frailty distribution $\mathcal{LN}(-\sigma^2/2, \sigma^2)$ is modeling a normally distributed random term acting linear on the predictor, which fits well to generalized linear models [12]. A plot of the hazard rate distribution for different θ may help to illustrate the degree of heterogeneity in hazard rates caused by θ. This is exemplified for the NIATx trial in Figure (4):

For $\theta = 0.3$ and a constant marginal baseline hazard of $\lambda_0 = 19.5^{-1}$, 95% of cluster specific baseline hazards can be expected to lie within $[35.3^{-1}, 11.7^{-1}]$ corresponding to within-cluster waiting times from 11.7 days to 35.5 days. For $\theta = 0.45$, this range is enlarged to $[9, 51.7]$ days.
Although the lognormal distribution seems to be a reasonable distributional assumption, real data will always show deviations from this assumption. Due to the large cluster sizes in the NIATx 200 trial it is possible to fit a Cox proportional hazards model with intervention and addiction treatment center as fixed covariates, thus replacing the random frailty term in (B.1) by a fixed covariate:

$$\lambda_{ij} \quad = \quad \lambda_0(t)\exp(\beta' W_i + \delta_i) \tag{1}$$

Using these results we can estimate the frailty distribution by the Kernel density estimate of $(\exp(\hat{\delta}_i))_{i=1...2N}$, which is plotted in Figure 5. As we consider pairwise comparisons only, we used only clusters within two intervention groups (Interest circle calls and Coaching) for fitting model (1).

Except for some outliers with hazard ratio estimates of more than 5, the estimated frailty distribution in NIATx data comes close to a lognormal distribution.

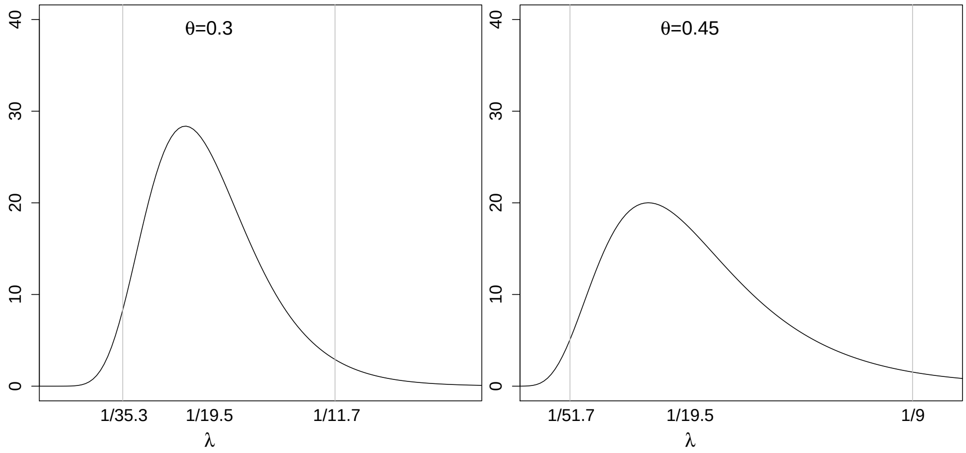


Figure 4: Density of hazard rates λ under a lognormal distributed and mean 1 frailty, $Z_i \sim \mathcal{LN}(-\frac{\sigma^2}{2}, \sigma^2)$ with $\sigma \approx \theta$ and a marginal hazard rate $\lambda_0 = 1/19.5$ (mean waiting time = 19.5 days). Vertical lines at $\exp(-\sigma^2/2 \pm 1.96\sigma)\lambda_0$ give a range around λ_0 where 95% of cluster specific baseline hazards λ can be expected to lie within.

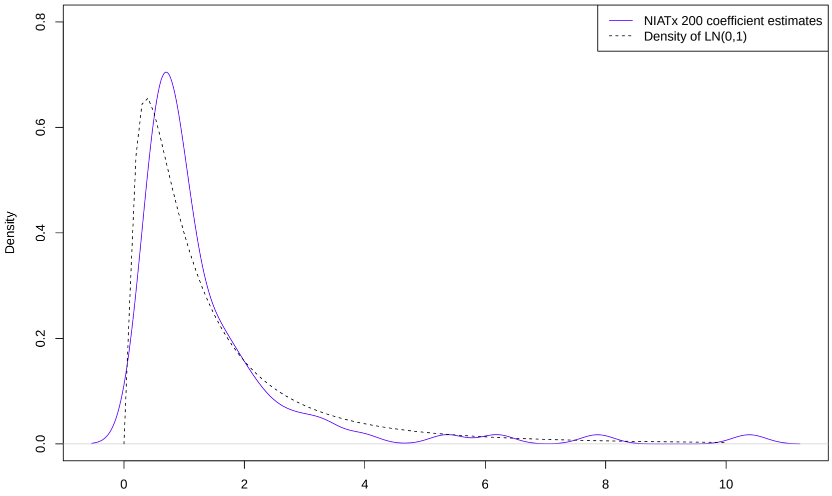


Figure 5: Kernel density estimate of the estimated cluster effects $\exp(\hat{\delta}_i)$

4. Discussion

We illustrated, how to realize the sample size determination in cluster randomized trials with a time to event endpoint applying the methods proposed by Jahn-Eimermacher *et.al.* [6]. The parameter θ, reflecting the degree of cluster heterogeneity, usually will be unknown in the planning phase of a trial. A sensitivity analysis using different values of θ should be performed to see how a misspecification of θ in the sample size determination will affect the power. In situations where power is substantially affected by a misspecified θ, but basically nothing is known about θ, it might be worth to consider sample size re-estimation procedures. Using a midtrial-estimate of nuisance parameters to adjust the sample size is well known from randomized clinical trials [3] and recently has been proposed for cluster randomized trials [1, 7]. Further research is required to evaluate the efficiency and validity of an internal pilot trial design.

The sample size formula is based on model assumptions which might not be justified in some applications. Deviations from the assumption of a constant marginal baseline hazard and of equal cluster size have been evaluated by the authors [6]. A further source of model misspecification might be the frailty distribution, which will conventionally be assumed to be a gamma or lognormal distribution. For the NIATx 200 trial we could demonstrate that a lognormal distribution would in fact be reasonable. However, this might look different for other applications and a deeper knowledge how frailty distribution misspecification affects the power of a trial, and thus the required sample size, would be valuable.

Acknowledgements

The authors would like to thank Harald Binder for constructive comments improving the publication. This research was supported by the Deutsche Forschungsgemeinschaft (DFG) grant number JA 1821. An award from the National Institute on Drug Abuse supported the NIATx 200 study (R01 DA 020832).

References

[1] MJ Campbell, A Donner, and N Klar. Developments in cluster randomized trials and *Statistics in Medicine*. *Statistics in Medicine*, 26:2–19, 2007.

[2] A Donner and N Klar. *Design and analysis of cluster randomization trials in health research*. John Wiley & Sons, 2000.

[3] T Friede and M Kieser. Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal*, 48:537–555, 2006.

[4] RJ Hayes and S Bennett. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, 28:319–326, 1999.

[5] KA Hoffman, JH Ford, CJ Tillotson, D Choi, and D McCarty. Days to treatment and early retention among patiens in treatment for alcohol and drug disorders. *Addictive Behaviors*, 36:643–647, 2011.

[6] A Jahn-Eimermacher, K Ingel, and A Schneider. Sample size in cluster randomized trials with a time to event as the primary endpoint. *Statistics in Medicine*, [Epub ahead of print], 2012.

[7] S Lake, E Kammann, N Klar, and R Betensky. Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, 21:1337–1350, 2002.

[8] D McCarty, DH Gustafson, JP Wisdom, J Ford, D Choi, T Molfenter, V Capoccia, and F Cotter. The network for the improvement of addiction treatment (niatx): Enhancing access and retention. *Drug and Alcohol Dependence*, 88:138–145, 2007.

[9] AR Quanbeck, DH Gustafson, JH Ford, A Pulvermacher, MT French, KJ McConnell, and D McCarty. Disseminating quality improvement: study protocol for a large cluster-randomized trial. *Implementation Science*, 6:44, 2011.

[10] D Schoenfeld. Sample size formula for the proportional-hazards regression model. *Biometrics*, 39:499–503, 1983.

[11] TM Therneau and PM Grambsch. *Modeling survival data*. Springer, New York, 2000.

[12] A Wienke. *Frailty models in survival analysis*. Chapman & Hall/CRC, Boca Raton, 2010.