Journal of Biostatistics and Biometric Applications
Volume 3 | Issue 2
ISSN: 2455-765X

Research article Open Access

Development of New Method for the Prediction of Clinical Trial Results Using Compressive Sensing of Artificial Intelligence

Miyagi Y[*1,2], Fujiwara K[3], Oda T[4], Miyake T[4], and Coleman RL[5]

[1]Department of Gynecology, Miyake Ofuku Clinic, Okayama City, Japan
[2]Department of Artificial Intelligence, Medical Data Labo, Okayama City, Japan
[3]Department of Gynecologic Oncology, Saitama Medical University International Medical Center, Hidaka City, Japan
[4]Department of Obstetrics and Gynecology, Miyake Clinic, Okayama City, Japan
[5]Department of Gynecologic Oncology and Reproductive Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[*]**Corresponding author:** Miyagi Y, Department of Gynecology, Miyake Ofuku Clinic, 393-1 Ofuku, Minami ward, Okayama City, 701-0204, Japan, Department of Artificial Intelligence, Medical Data Labo, Okayama City, 703-8267, Japan, Fax: +81-86-281-7575, Tel: +81-86-281-2020, E-mail: ymiyagi@mac.com

Citation: Miyagi Y, Fujiwara K, Oda T , Miyake T, Coleman RL (2018) Development of New Method for the Prediction of Clinical Trial Results Using Compressive Sensing of Artificial Intelligence. J Biostat Biometric App 3(2): 202

Summary

We propose a novel strategy to predict unknown data from a smaller sample size by using compressive sensing which is one of the information technology used in artificial intelligence. In this study, we applied the compressive sensing to an empirical sample of clinical trials comprising of two groups. Each group required the time data of 25 patients but the data of 12 more patients had to be acquired to satisfy alpha and beta errors. The unknown data were thus predicted using compressive sensing method. The calculation was repeated 1,000 times for each, and 1 million (=1,000^2) log-rank tests between the two groups were carried out following, which a histogram of the P-values was obtained. The information on the distribution of predicted data obtained by repeated calculations reflects the probability of statistical significance. The method could thus increase efficiency by saving on human resources, financial resources, and time. As such, further studies examining the feasibility of compressive sensing in clinical trials are warranted.

Keywords: Clinical trial; Artificial intelligence; Compressive sensing; Computer science; Study design

Introduction

Well-designed clinical trials are essential to proper and accurate inference through clear hypothesis testing. The end-result may lead to clinical practice changes when superior or unexpected outcomes are identified. The outcomes measures of such clinical trials are defined by pre-investigation statistical plans and procedures. One critically important factor in clinical trials is sample size and is determined by calculations according to α and β errors. When data are insufficient or ambiguous in a clinical trial, more data has to be acquired. However, if the lacking data can be reliably predicted, the clinical trial can proceed with increased efficiency while saving on human and financial resources and time. Specifically, if reliable methodology could be developed to better predict a likely future outcome, better decisions regarding further expansion vs triage could be made at an earlier times point in investigation, such after acquisition of phase II endpoints.

There are some methods to estimate unknown complete data from known incomplete data. The method of full-information maximum likelihood would show the expectation values of the complete data that are under exponential distribution [1]. The expectation–maximization algorithm is a method to find maximum-likelihood estimates for model parameters when data are incomplete. This method does not always imply convergence [2]. Local principal components from a large-scale data set with missing values could be extracted [3]. The last observation carried forward is also a method of handling missing data. It is simply to impute, or fill in, values based on existing data, but it may underestimate the variances of the estimated result [4,5]. Baseline observation carried forward is a method to handle missing data from early treatment discontinuation. It may also underestimate the variances of the estimated result [6,7].

Compressive sensing [8] is one of the information technologies employed to estimate unknown complete data from known incomplete data in computer science and information engineering and is widely used in many fields, such as in astronomy [9,10]. In medicine, MRI [11,12] and CT imaging [13] have been investigated using this information technology. However, the compressive sensing has not been applied to clinical study design yet. Here, we applied compressive sensing with L1 type regularization, a summation of absolute values of vector elements, which is often used in an artificial intelligence (AI) of computer science [8]. In this study, we introduced the compressive sensing with an original modification, which could obtain an outcome by predicting unknown data including the observed data, to clinical trials. We demonstrate an empirical case to propose the feasibility of using this technology for clinical trials.

Materials and Methods

We developed a compressive sensing algorithm so that the larger number of unknown data, vector X, which includes the smaller number of known data, vector Y, may be predicted. The L1 type regularization was added to enable allowance of no strict sparseness of data [14-16].

Suppose m patients are required to fit a clinical trial, and only n patients (n<m) are enrolled. The data of the remaining patients who should be enrolled will be predicted as follows;

$$\arg\min \frac{1}{2n} \| Y - AX \|^2 + \lambda \sum_{j=1}^{m} \left| x_j \right|$$

where X=t(x1,..., xm), Y=t(y1,..., yn), $m, n \in \mathbb{N}$, m>n, $\forall x \geq 0$, $\forall y \geq 0$, $\forall S \in N(0,1), A \in S_{n \times m}$ and $\forall y \in X$. Suppose m>n, one already knows n patients' data, vector Y, and one needs to know m patients' data, vector X. Let vector X include vector Y as a novel idea. The elements in n×m matrix A consist of random real numbers of standard normal distribution. Then let the formula above calculate by varying matrix A and vector X. Then one obtains vector X as a predicted answer. But this is merely one candidate of the total possible results. Therefore, repeated calculations are required to obtain a list of m-dimensional vectors as the final format of this study.

In this study, we present an empirical sample of phase II of surviving analysis (Figure 1). Because the data are time in the sample, our program lets the values of all data be greater than or equal to zero. The sample is supposed to be an incomplete situation of a two-arm survival clinical trial. Suppose that a simulated 25 patients have been enrolled in group A and B as known data, respectively. In group A, uncensored and censored time data were {0.969, 0.655, 0.463, 0.330, 0.225, 0.154, 0.112, 0.076, 0.053, 0.038, 0.026, 0.018, 0.012, 0.009, 0.006, 0.004, 0.002, 0.002, 0.001} and {0.212, 0.427, 0.622, 0.847, 1.056, 1.265}, respectively. In group B, uncensored and censored time data were {0.969, 0.504, 0.273, 0.147, 0.080, 0.042, 0.023, 0.012, 0.006, 0.003, 0.002}, {0.093, 0.177, 0.271, 0.362, 0.447, 0.557, 0.619, 0.733, 0.801, 0.901, 1.01, 1.10, 1.160, 1.249}, respectively. The unit of time is arbitrary. A computer program generated randomly those 25 time data of which the maximum value is less than 1.350 in order to mimic conventional actual data. Sample size calculation with initial survival rates of the two groups revealed that 12 more patients were needed for each group to satisfy the prespecified alpha and beta error, 0.05 and 0.2, respectively. In group A, the number of uncensored data and censored known data are 19 and 6, respectively. Then the number of required censored data, which is proportional to the number of the known data, is 3 (=Round (6-(25+12)*6/25)), then the number of required censored data is 9 (=12-3). In group B, the

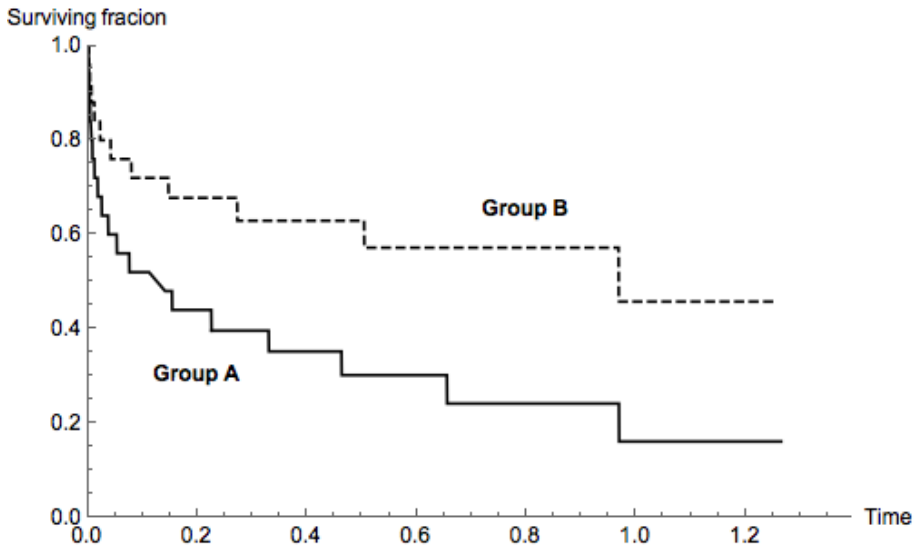

Figure 1: A sample of Kaplan-Meier survival curves of a clinical trial. There are two groups, each of which is consisted of 25 patients. The survival fractions are not significantly different (p=0.052) by log-rank test. But sample size is too small to fit. Twelve additional patients are necessary to satisfy both α error =0.05 and β error =0.20

number of uncensored data and censored known data are 11 and 14, respectively. Then the number of required censored data and uncensored data are 5 and 7, respectively. There is no significant difference between the two groups for survival at the initial situation. The time data of 12 patients in a group were predicted from known data of 25 patients with the condition that the maximum value of 12 patients was less than twice of the maximum value of 25 patients. The censored and uncensored data were calculated independently and repeated 1,000 times. Then 1,000 patterns of survival data from 37 patients in a group were obtained.

This procedure was then carried out for another group. One million (=$1{,}000^2$) log-rank tests between the two groups were performed and the histogram of the one million P-values of log-rank test were obtained. Then the probability of $P<0.05$ was obtained if the empirical study would be completely finished.

The differences between the known data and the corresponding elements in the predicted unknown data were investigated. The all of values of the known data were randomly set as between greater than or equal to 0 and less than or equal to 1. When the number of the known data and that of the adding data varied from 3 to 15 and from 1 to 18, respectively, the mean and the standard deviation of the variances of the differences were calculated by 10 times for each combination.

Results

A list of sets of predicted data was obtained from smaller number of samples by compressive sensing in the empirical sample.

The 1,000 of surviving fraction curves for each group are shown (Figure 2). Then the histogram of P-values of one million of log-rank tests between the two groups showed the probability of less than 0.05 was obtained as 98.8% (Figure 3). The median and average of P-values was 0.009 and 0.012, respectively. The first and the third quartile was 0.005 and 0.015, respectively.

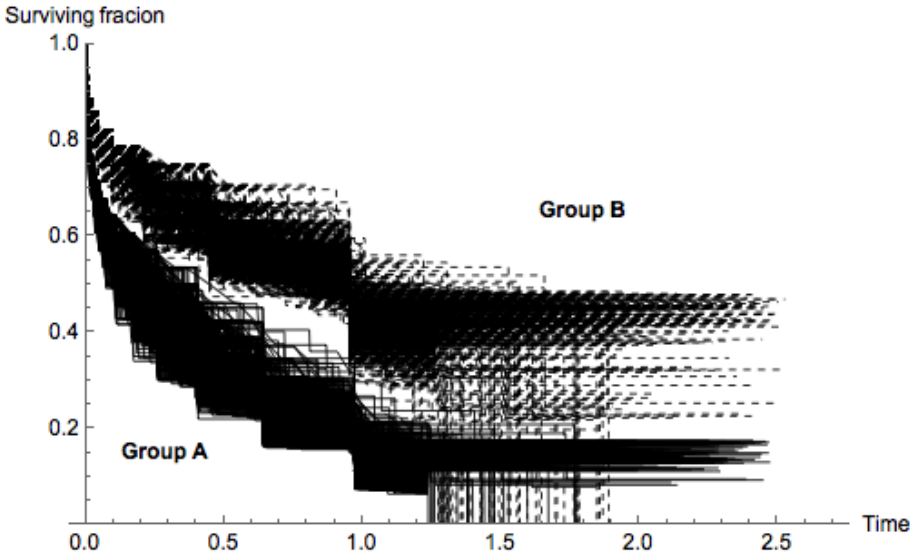

Figure 2: Kaplan-Meier survival curves generated from predicted data. The total 37 patients' data consisted of known 25 patients' data and predicted 12 patients' data were obtained by compressive sensing. The prediction was repeated 1,000 times for each group. Then the 1,000 of surviving fraction curves for each group are shown

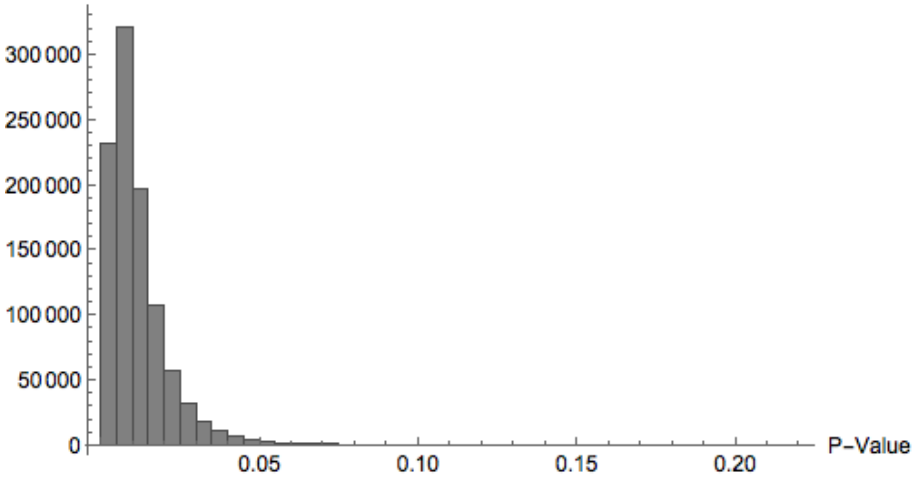

Figure 3: The histogram of the P-values of 1 million log-rank tests of the two groups. The probability of P-values less than 0.05 is 98.8% in this study. The median and average of P-values was 0.009 and 0.012, respectively. This may suggest that by using the compressive sensing, this clinical trial that requires 37 patients in each group with only enrolled 25 patients would result in significant difference of the two groups in terms of survival analysis

The variance profile of the differences between the known data and the unknown data is shown in Figure 4. The mean, standard deviation (SD), the maximum and the minimum number of the variances were 0.085, 0.091, 0.283 and 0.003, respectively. The left upper area, in which the number of the adding data is be greater than or equal to the number of the known data, is black which means the least variance. When the number of the adding data is the same as the number of the known data, the variance was 0.039±0.009 (Mean±SD).

Discussion

A list of the predicted data as a completed enrolled clinical trial was obtained from a smaller number of known data. It is emphasized that the predicted data included the known data. The outcome of a clinical trial can thus be predicted using compressive sensing. In the empirical case, the survival fractions were not significantly different initially. However, this would be significantly different at the probability of 98.8% if the clinical trial were to continue to reach the complete sample size. Findings from this study would thus likely be significant, and the investigators could proceed with designing the protocol for the next step, such as a phase III trial.

An unfavorable outcome predicted by this method would suggest for the investigators to abort the clinical trial immediately. The prediction can be repeatedly obtained at various points in the clinical trial when the information of a single patient is censored. The method using compressive sensing as well as the optimal two-stage design proposed by Simon [17,18] may be useful for clinical trials. Simon's method determines the adequate sample size by calculating the response probability of two groups with α and β errors, while our method used concrete data of every patient. Further studies to validate this method in Simon's two stage phase II trial is warranted and this would also be helpful in establishing futility benchmarks.

Our method using compressive sensing to predict unknown data including the observed data can be applied to various types of data including length, weight, volume and so on. Therefore, this method can also be widely applied to the development of drugs and other experiments, in addition to clinical trials. This method can be used not only for comparative studies but also non-comparative studies that handle quantity.

The results predicted by our method of compressive sensing does not present a confirmed outcome but rather a distribution of the probability of outcomes. Because the single outcome predicted by the method is only a candidate of unknown data, repeated calculations are required to obtain the outcome as a format of probability. When the probabilities of predicted outcomes of statistical significance are more than 95% among all predicted outcomes, the actual outcome with sufficient data of the clinical study with insufficient data would likely to be statistically significant. On the other hand, if the probabilities were not more than 95%, the actual outcome would likely be not significant. The probability of an outcome becomes more reliable when a statistical test is repeated. The log-rank tests in this study were repeated one million times. We could use less but the minimum is dependent on the mean (or median) survival times of the 2 groups [19,20].

The variances of the differences between known and unknown data revealed the characteristic profile (Figure 4). The larger the number of adding data, the smaller the variance. The variance is quite small when the number of added data is larger than the number of the known data. When the number of the adding data is equal to the number of the known data, the variance was 0.039±0.009 (Mean±SD). Therefore, if the number of the adding data is equal to the number of the known data, the variance will be likely to 3.9% of the maximum value of the known data statistically. The size of the variance that is less than 5% could be acceptable clinically. That equality of the number of both the known data and the unknown data seems to be the boundary of the profile, and to be important to determine the proper adding number of the unknown data. In other words, the clinical data prediction with compressive sensing should be used so that the number of the added data might be greater than or equal to the number of the known data, which can be recognized as the left upper black area in Figure 4.

There are several nuances and issues to be resolved in future work. For instance, the specific conditions determined for the calculations are unique to each case. In this example of an empirical case, the maximum value of unknown data was defined as less than twice of the maximum value of known data. In addition, a normal distribution was applied to generate the matrix for survival data in this study. But this assumption has not been established and it might be better to select a different distribution such as the Weibull distribution, which is sometimes used as a parametric model of survival. Moreover, it is supposed that the uncensored event could follow the Weibull distribution and the censored event could occur by chance, which would imply that censored events may follow a normal distribution. Therefore, we may need to select the different distribution according to the feature of the data. Compressive sensing is able to reconstruct data especially when data are sparse. However, clinical data are not always sparse. The value of regularization would be determined not theoretically, but retrospectively. The fact that the investigator can define these conditions such as the maximum value of unknown data, the distribution model, the value of regularization and so on in every clinical trial implies that unpredictable extreme outcomes cannot be obtained. This remains the limitation of current AI. Currently, there is no standard method to complete lacking data in the clinical trial. Therefore, further validation to test the feasibility and applicability with real clinical trial should be conducted.

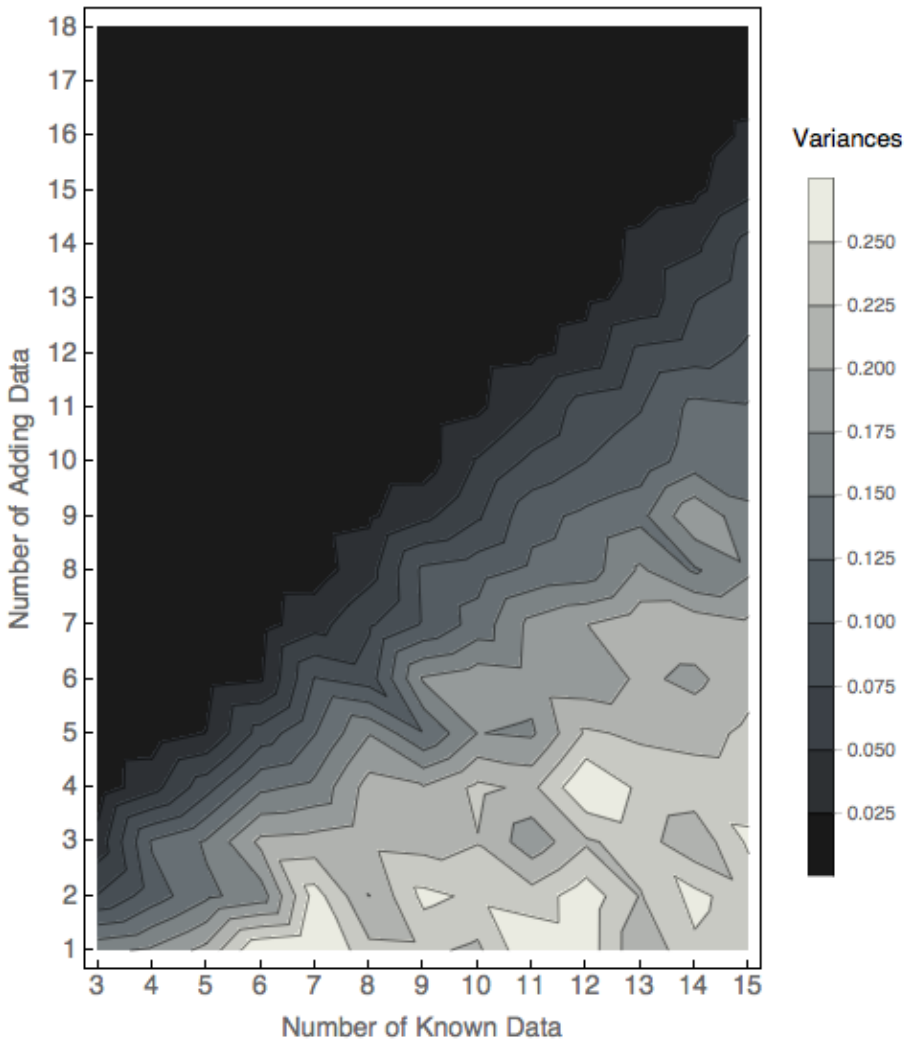

Figure 4: The density plot with contour of the variance that represents differences between the known data of value with a range between greater than or equal to 0 and less than or equal to 1 and the corresponding elements in the predicted unknown data is shown. When the number of the known data is from 3 to 15 and the number of the adding data is from 1 to 18, the variances were calculated by 10 times for each combination. It would be recommended that the number additional data should be greater than or equal to the number of the known data in order to reduce the variance, if possible

Conclusion

This method would be useful for clinical trials investigating rare diseases or if there is delay in the enrollment process. It would also be helpful in establishing futility benchmarks in phase II trials.

References

1. Lord FM, Novick MR (2008) Statistical theories of mental test scores, Information Age Publishing. Charlotte, United States.

2. Wu CJ (1983) On the convergence properties of the EM algorithm. Ann Stat 11: 95-103.

3. Honda K, Kanda A, Ichihashi H, Yamakawa A (2002) Extraction of local principal components from data with missing values. Syst Contr Inform 15: 663-72.

4. Salim A, Mackinnon A, Christensen H, Griffiths K (2008) Comparison of data analysis strategies for intent-to-treat analysis in pre-test–post-test designs with substantial dropout rates. Psychiatry Res 160: 335-45.

5. Molnar FJ, Hutton B, Fergusson D (2008) Does analysis using "last observation carried forward" introduce bias in dementia research? CMAJ 179: 751-3.

6. Liu-Seifert H, Zhang S, Deborah D'Souza D, Skljarevski V (2010) A closer look at the baseline-observation-carried-forward (BOCF). Patient Prefer Adherence 4: 11-6.

7. Kaiser KA, Affuso O, Beasley TM, Allison DB (2012) Getting carried away: a note showing baseline observation carried forward (BOCF) results can be calculated from published complete-cases results. Int J Obes (Lond) 36: 886-9.

8. Candés EJ, Romberg J, Toa T (2006) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inf Theory 52: 489-509.

9. Bobin J, Starck JL, Ottensamer R (2008) Compressed sensing in astronomy. IEEE J Sel Top Signal Process 2: 718-26.

10. Starck JL, Bobin J (2010) Astronomical data analysis and sparsity: from wavelets to compressed sensing. Proceedings of the IEEE 98: 1021-30.

11. Xi Y, Zhao J, Bennett JR, Stacy MR, Sinusas AJ, et al. (2016) Simultaneous CT-MRI Reconstruction for Constrained Imaging Geometries Using Structural Coupling and Compressive Sensing. IEEE Trans Biomed Eng 63: 1301-09.

12. Lustig M, Donoho D, Pauly JM (2007) Sparse MRI: The application of compressed sensing for rapid MR imaging. Magn Reson Med 58: 1182-95.

13. Chen GH, Tang J, Leng S (2008) Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly under sampled projection data sets. Med Phys 35: 660-63.

14. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 58: 267-88.

15. Chen S, Donoho DL, Saunders M (1998) Atomic decomposition by basis pursuit. SIAM REVIEW 43: 129-59.

16. Donoho D (2006) Compressed sensing. IEEE Trans Inf Theory 52: 1289-306.

17. Simon R (1989) Optimal two-stage designs for phase II clinical trials. Control Clin Trials 10: 1-10.

18. Jung SH, Lee T, Kim K, George SL (2004) Admissible two stage designs for phase II cancer clinical trials. Stat Med 23: 561-9.

19. Lakatos E (1988) Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics 44: 229-41.

20. Lakatos E, Lan KK (1992) A comparison of sample size methods for the logrank statistic. Stat Med 11: 179-91.