Research Article

A Bayesian Design for Phase II Clinical Trials with Delayed Responses Based on Multiple Imputation

Chunyan Cai[a], Suyu Liu[b] and Ying Yuan[b]*

Interim monitoring is routinely conducted in phase II clinical trials to terminate the trial early if the experimental treatment is futile. Interim monitoring requires that patients' responses be ascertained shortly after the initiation of treatment so that the outcomes are known by the time the interim decision must be made. However, in some cases, response outcomes require a long time to be assessed, which causes difficulties for interim monitoring. To address this issue, we propose a Bayesian trial design to allow for continuously monitoring phase II clinical trials in the presence of delayed responses. We treat the delayed responses as missing data and handle them using a multiple imputation approach. Extensive simulations show that the proposed design yields desirable operating characteristics under various settings and dramatically reduces the trial duration.

Keywords: Continuous monitoring; Delayed responses; Multiple imputation.

1. Introduction

The primary objective of phase II clinical trials is to test the preliminary efficacy of experimental agents and decide whether the agents are sufficiently promising to be sent to phase III trials. To avoid assigning a large number of patients to inferior treatments and to increase the design efficiency, interim monitoring is routinely conducted in phase II trials to allow for the possibility of early termination if the experimental treatment is futile.

Numerous phase II trial designs have been proposed. Gehan [1] proposed a two-stage phase II design for cancer research in which the trial is terminated if there are no favorable responses observed in the first stage. Fleming [2] developed a multiple-stage testing procedure that allows for early termination and also preserves the simplicity of the single-stage procedure. Simon [3] presented optimal two-stage designs that minimize the expected or maximum sample size. As extensions of Simon's design, Green and Dahlberg [4] proposed a two-stage design for multicenter trials when the attained sample size is not the planned one, and Chen [5] proposed a three-stage design. Other extensions of Simon's design include those of Ensign *et al.* [6], Hanfelt *et al.* [7], Jung *et al.* [8], Shuster [9], and Lin and Shih [10]. Under the Bayesian framework, Tan and Machin [11] proposed a Bayesian two-stage design in which the parameters are calibrated based on

[a] *Division of Clinical and Translational Sciences, Department of Internal Medicine, Medical School; Biostatistics/Epidemiology/Research Design Core, Center for Clinical and Translational Sciences, The University of Texas Health Science Center at Houston, Houston, TX 77030,U.S.A.* [b] *Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A.*

* *Correspondence to: Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A. E-mail: yyuan@mdanderson.org*

the posterior probability approach. That design was extended by Sambucini [12] to take into account the uncertainty of future data.

In addition to these two and multiple stage designs, phase II designs with continuous monitoring have been proposed to improve the efficiency of interim monitoring. Thall and Simon [13] developed practical Bayesian guidelines for the design and analysis of phase II clinical trials with continuous monitoring. Thall *et al.* [14] proposed a Bayesian single-arm phase II design with sequential monitoring for multiple outcomes using a Dirichlet-multinomial model. Heitjan [15] proposed a flexible Bayesian phase II design using "persuasion probability", which allows for termination at any interim analysis as long as the persuasion probability exceeds its critical value. Lee and Liu [16] developed a flexible phase II design with continuous monitoring based on Bayesian predictive probability. Johnson and Cook [17] derived a class of Bayesian designs based on formal hypothesis tests using non-local alternative prior densities with continuous monitoring. Wathen *et al.* [18] proposed a flexible Bayesian single-arm phase II design with subgroup-specific early-stopping rules, which allows the decision of trial termination to differ within each subgroup. Zohar *et al.* [19] provided an excellent tutorial on how to conduct Bayesian phase II single-arm clinical trials with binary outcomes.

A major practical impediment when implementing phase II clinical trial designs, particularly with continuous monitoring, is that the responses must be observed early enough to apply the stopping rules. However, in practice, the efficacy response may take a relatively long time to be observed, with respect to the accrual rate. As an example, a single-arm phase II clinical trial recently initiated at MD Anderson Cancer Center investigated the efficacy of a combination of everolimus with a novel kinase inhibitor in patients with glioblastoma. The assessment of the response to the treatment (i.e., partial and complete response) requires 3 months. The lowest acceptable response rate for this trial was 40%; that is, if the response rate of the experimental treatment is below that value, we should terminate the trial for futility. The accrual rate was about 2 patients per month. The difficulty of conducting futility monitoring for that trial is that the response takes a relatively long follow-up to assess; thus, at each monitoring time point, the response outcome may not be observable for some patients.

One possible solution to this practical dilemma is to suspend the accrual and wait until the data in the trial mature before enrolling the next new patient in order to always have complete data for the interim monitoring. However, this complete-data method is typically infeasible in practice because repeatedly suspending patient accrual is not practical and often leads to unacceptably long trials. Another approach is to conduct interim analyses based on the observed data from patients who have responded to the treatment or/and completed the follow-up so that suspension of patient accrual is not needed [14, 15, 18]. Unfortunately, this observed-data approach is biased because the response outcome (i.e., response or no response) is more likely to be observed for patients who will respond to the treatment than for those who will not, i.e., the observed data comprise a biased sample of the patients [20]. In order to obtain unbiased inference, we need to take into account the partially followed patients (i.e., the patients who have not completed their follow-up assessments and have not yet responded to the treatment). By treating the treatment response as a time-to-event outcome, Follman and Albert [21] proposed an interim monitoring method based on a discrete-time survival model; and Zhao *et al.* [22] developed a Bayesian decision theoretic two-stage phase II design. Cheung and Thall [23] kept the treatment response in its conventional form as a binary outcome and weighted the binomial likelihood with patients' follow-up times to account for partially followed patients.

In this article, we propose a new missing-data approach to handle the issue of delayed response. We naturally treat unobserved delayed response outcomes as missing data, while keeping the observed (binary) response outcomes intact. We impute the unobserved responses using the multiple imputation approach based on a flexible piece-wise exponential model. Unlike the methods of Follman and Albert [21] and Zhao *et al.* [22], our approach treats the response as a binary outcome, which is more consistent with conventional phase II trials. In addition, as our method keeps the observed response data intact, it is more robust to model misspecification (e.g., the piece-wise exponential model). That is, the model misspecification only affects the imputed data, and more importantly, such effect attenuates along the trial and eventually disappears because when the trial moves on, more and more patients and eventually all patients' outcomes are observed.

Compared to the method of Cheung and Thall [23], for which the weighted likelihood may not always correspond to the actual data likelihood, our method is a fully likelihood-based approach and thus more efficient. Simulation studies show that the proposed method possesses desirable operating characteristics under various practical scenarios, and allows for continuous monitoring without prolonging the duration of the trial.

The remainder of this article is organized as follows. In Section 2, we introduce the Bayesian continuous monitoring rule and the multiple imputation approach to handle the delayed response. In Section 3, we examine the operating characteristics of the proposed design through extensive simulation studies and sensitivity analyses. We conclude with a brief discussion in Section 4.

2. Methods

2.1. Bayesian continuous monitoring

Consider a single-arm phase II trial in which patients enter the trial sequentially and are followed for a fixed period of time T after the treatment to assess their responses to an experimental treatment, e.g., partial or complete response. We assume that the treatment response y is binary, with $y = 1$ if the response is observed within the assessment window $[0, T]$, and $y = 0$ otherwise. That is, y follows a Bernoulli distribution,

$$y \sim Bernoulli(\pi),$$

where π is the treatment response rate. Depending on the nature of the disease and the response, the pre-specified assessment window T varies from weeks to months.

Suppose that at an interim monitoring time, n patients have been enrolled into the trial and their responses $\boldsymbol{y} = \{y_i, i = 1, \cdots, n\}$ have been fully observed, then the likelihood function of $\boldsymbol{y}$ is given by

$$L(\boldsymbol{y}|\pi) = \prod_{i=1}^{n} \pi^{y_i}\{1-\pi\}^{1-y_i}.$$

If we assign a conjugate beta prior to π with shape parameters ζ and ξ, i.e., $\pi \sim Beta(\zeta, \xi)$, the posterior distribution of π is

$$f(\pi|\boldsymbol{y}) = Beta(\zeta + \sum_{i=1}^{n} y_i,\ \xi + n - \sum_{i=1}^{n} y_i).$$

Let ϕ denote a lower bound of the acceptable response rate prespecified by physicians, and ψ denote a prespecified cutoff. Our Bayesian futility monitoring can be described as follows:

> At any time during the trial, if $\Pr(\pi < \phi|\boldsymbol{y}) > \psi$, we terminate the trial for futility; otherwise, we continue the accrual until we reach the maximum sample size.

In practice, to improve the reliability of the design, we typically apply the above continuous monitoring rule only after a certain number of patients, say n_0, have been treated and have completed their response assessments. The value of ψ can be chosen and calibrated by simulation studies to obtain desirable operating characteristics.

This continuous monitoring requires that treatment responses are quickly evaluable such that at each of the interim monitoring times, the responses of the enrolled patients are fully observable and available to inform the decision of stopping for futility. This requirement, however, is not satisfied when the response are delayed because, in this case, at the moment of interim monitoring, some patients might not have finished their evaluations and thus their responses are not yet observable.

2.2. *Accommodating delayed response using multiple imputation*

We propose to handle the delayed response using multiple imputation, which provides a systematic way to impute the unobserved responses and meanwhile account for the sampling uncertainty due to the missing values [24]. In the following subsection, we first define the missing data caused by delayed responses at the interim monitoring time and then describe the procedures to impute the missing data.

2.2.1. Missing data Let t_i denote the time to response for patient i. At the moment of interim monitoring, let r_i ($0 \leq r_i \leq T$) denote the actual follow-up time for patient i, and $m_i(r_i)$ denote the missing data indicator for y_i, with $m_i(r_i) = 1$ indicating the missingness of y_i. It follows that

$$m_i(r_i) = \begin{cases} 1 & \text{if} \quad t_i > r_i \text{ and } r_i < T, \\ 0 & \text{if} \quad t_i \leq r_i \text{ or } r_i = T. \end{cases} \tag{1}$$

That is, the response outcome y_i is missing (i.e., $m_i(r_i) = 1$) if the patient has not yet responded to the treatment ($t_i > r_i$) and has not been fully followed up to T ($r_i < T$). The response outcome y_i is observed (i.e., $m_i(r_i) = 0$) if the patient has either responded to the treatment ($t_i \leq r_i$) or completed the entire follow-up ($r_i = T$) without experiencing the response. For notational simplicity, we suppress the augmentation of r_i in $m_i(r_i)$ and let $\boldsymbol{m} = (m_1, \ldots, m_n)$. We partition $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$, where $\boldsymbol{y}_{obs}$ and $\boldsymbol{y}_{mis}$ denote the observed and unobserved response values, respectively.

A key point is that the missing y_i's are nonignorable in the sense that at any follow-up time, a patient who will not respond to the treatment by the end of follow-up (i.e., for whom it will turn out that $y_i = 0$) is more likely to have y_i missing than a patient who will respond to the treatment (i.e., for whom $y_i = 1$) [20]. Formally, $\Pr(m_i = 1|y_i = 0) > \Pr(m_i = 1|y_i = 1)$, which by Bayes' rule implies that

$$\frac{\pi_j}{1 - \pi_j} > \frac{\Pr(y_i = 1|m_i = 1)}{\Pr(y_i = 0|m_i = 1)}.$$

That is, the odds of response are smaller if y_i is missing, so the missingness indicator m_i contains information about the future value of y_i. Therefore, the simple approach of evaluating the stopping rule $\Pr(\pi < \phi|\boldsymbol{y}_{\text{obs}}) > \psi$ based on the observed response values $\boldsymbol{y}_{\text{obs}}$ leads to biased estimates and poor operating characteristics, as we show in our simulation study.

2.2.2. Imputation model Because the missing data are nonignorable, in order to impute the missing y_i's without bias, we need to model the missing data mechanism, which involves the time to response t_i as defined in (1). We specify a flexible correlated piecewise exponential model for the time to response. Specifically, we partition the follow-up period $[0, T]$ into a finite number of J disjoint intervals $[0, d_1), [d_1, d_2), \cdots, [d_{J-1}, d_J = T]$ and assume a constant hazard λ_j in the jth interval with $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_J)$. Based on numerical studies and the fact that the sample size of phase II trials is typically small or moderate, we recommend to set $4 \leq J \leq 8$ to keep the model parsimonious. In practice, we can start with $J = 4$ and use simulation studies to determine whether an increase of the value of J is needed to further improve the performance of the design. Our experience is that $J = 6$ is adequate for most practical applications. We define the observed time $x_i = min(r_i, t_i)$ and δ_{ij} as the indicator of the ith subject experiencing the response in the jth interval. The likelihood function of $\boldsymbol{x} = (x_1, \ldots, x_n)$ for n subjects is given by

$$L(\boldsymbol{x}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_j)^{\delta_{ij}} \exp(-\lambda_j e_{ij}),$$

where $e_{ij} = d_j - d_{j-1}$ if $x_i > d_j$; $e_{ij} = x_i - d_{j-1}$ if $x_i \in [d_{j-1}, d_j)$; and otherwise $e_{ij} = 0$.

To borrow information across the partition interval, we assume that λ_i follows a discrete-time martingale process [25, 26],

$$\lambda_j|\lambda_1,\cdots,\lambda_{j-1} \sim Gamma(c_j, \frac{c_j}{\lambda_{j-1}}), j = 1,\cdots,J$$

where $Gamma(\xi,\eta)$ represents a gamma distribution with a shape parameter ξ and a scale parameter η, and λ_0 and c_j are prespecified hyperparameters. This prior centers the hazard of an interval at that of the previous interval, i.e., $E(\lambda_j|\lambda_1,\cdots,\lambda_{j-1}) = \lambda_{j-1}$, thereby introducing correlations between the λ_i's in adjacent intervals and improving the smoothness of the estimates. The value of c_j controls the correlation and the smoothness of λ_j. If $c_j = 0$, λ_j is independent of λ_{j-1}, while if $c_j \to \infty$, $\lambda_j = \lambda_{j-1}$. Therefore, the posterior distribution of $\boldsymbol{\lambda}$ is given by

$$f(\boldsymbol{\lambda}|\boldsymbol{x}) \propto L(\boldsymbol{x}|\boldsymbol{\lambda}) \prod_{j=1}^{J} f(\lambda_j|\lambda_1,\cdots,\lambda_{j-1}), \tag{2}$$

which can be sampled using the Gibbs sampler.

2.2.3. Imputation procedure To carry out the multiple imputation, we draw the missing value of $y_i \in \boldsymbol{y}_{\text{mis}}$ from its posterior predictive distribution $f(y_i|\mathcal{D})$, where $\mathcal{D} = (\boldsymbol{y}_{\text{obs}}, \boldsymbol{m}, \boldsymbol{x})$. Because $f(y_i|\mathcal{D}) = \int f(y_i|\mathcal{D},\boldsymbol{\lambda}) f(\boldsymbol{\lambda}|\mathcal{D})\,\mathrm{d}\boldsymbol{\lambda}$, the multiple imputation of the missing value of y_i can be done in two steps:

1. Sample $\boldsymbol{\lambda}$ from its posterior distribution given by (2).
2. Conditional on $\boldsymbol{\lambda}$ sampled in step 1, draw the missing value of $y_i \in \boldsymbol{y}_{\text{mis}}$ from $f(y_i|\mathcal{D},\boldsymbol{\lambda})$, which is given by

$$f(y_i|\mathcal{D},\boldsymbol{\lambda}) = Bernoulli(\omega),$$

with

$$\begin{aligned} \omega &= \Pr(y_i = 1|m_i = 1,\boldsymbol{\lambda}) \\ &= \Pr(t_i < T|t_i > x_i,\boldsymbol{\lambda}) \\ &= \frac{F(T) - F(x_i)}{1 - F(x_i)}, \end{aligned}$$

where $F(\cdot)$ is a cumulative distribution function given by $F(s) = 1 - \sum_{j=1}^{J} \exp(-\lambda_j e_j)$, and $e_j = d_j - d_{j-1}$ if $s > d_j$; $e_j = s - d_{j-1}$ if $s \in [d_{j-1}, d_j)$; and otherwise $e_j = 0$.

We repeat steps 1 and 2 K times to obtain K sets of imputed values for $\boldsymbol{y}_{\text{mis}}$, denoted as $\boldsymbol{y}_{\text{mis}}^{(1)},\ldots,\boldsymbol{y}_{\text{mis}}^{(K)}$. Based on the K imputed datasets, our Bayesian continuous monitoring for delayed responses is given as follows: at any time during the trial, if $\Pr(\pi < \phi|\mathcal{D}) > \psi$, we terminate the trial for futility; otherwise, we continue the accrual until we reach the maximum sample size. The posterior probability $\Pr(\pi < \phi|\mathcal{D})$ is given by

$$\Pr(\pi < \phi|\mathcal{D}) = \frac{1}{K}\sum_{k=1}^{K} \Pr(\pi < \phi|\boldsymbol{y}^{(k)}),$$

where $\boldsymbol{y}^{(k)} = \{\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}^{(k)}\}$ denote the filled-in complete datasets based on the kth imputation.

3. Numerical studies

3.1. Simulation study

We conducted extensive simulation studies to evaluate the operating characteristics of the proposed method. We assumed a maximum sample size of $N = 50$ and that patients were enrolled according to a Poisson process, with a rate of two patients per month. We set the assessment period for response at $T = 3$ months and took $J = 6$ equally spaced partitions for the piece-wise exponential model. We assigned vague gamma priors to the λ_j's, with $c_1 = \ldots = c_J = 0.01$, and set λ_0 such that the response probability at the end of the 3-month follow-up was equal to ϕ, the lower bound of the acceptable response rate, (i.e., the response rate of the experimental agent just satisfied the lowest requirement for efficacy *a priori*). The shape parameters ζ and ξ for the prior distribution of π were set at 0.1 and 0.2, respectively. The prespecified cutoff ψ in the futility monitoring rule was set at 0.95. As shown in Table 1, we considered two values for the lower bound, where $\phi = 0.3$ and 0.4, each paired with various true response rates π. Under each of these parameter configurations, we generated the time to response from a Weibull distribution, for which the scale and shape parameters were chosen such that (1) the response rate at the end of the assessment was equal to the value of π, and (2) 90% or 70% of responses occurred in the latter half of the assessment window $(T/2, T)$. The second condition was used to generate different degrees of delayed response. For each of the simulation settings, we simulated 1,000 trials and conducted continuous monitoring after $n_0 = 5$ patients had completed the assessment period.

We compared the proposed multiple imputation (MI) approach to three alternative methods: the naive method, observed-data (OD) method, and weighted (WT) method proposed by Cheung and Thall [23]. Specifically, for the naive method, at each of the monitoring times, we simply used the current status of each patient's response as y_i, regardless of whether the patient had completed the follow-up or not. This naive method can be biased because the patients who have not yet responded to the treatment may respond at a later time during the follow-up. In contrast, the OD method conducts the monitoring based on the data from the patients whose response outcomes have been observed. Note that because patients can respond to the treatment any time within the follow-up window (0, T], the set of patients whose response outcomes have been observed are not equivalent to the set of patients who have completed the follow-up. For the WT method, following Cheung and Thall [23], each patient's likelihood (i.e., a Bernoulli distribution) was weighted by r_i/T_i. For the purpose of comparison, we also implemented the complete-data (CD) method, which carries out the interim monitoring using complete data by suspending accrual and waiting for each patient's response outcome to mature before enrolling the next patient. As we described previously, the CD method is not feasible in practice; but it provides an optimal bound and a "gold standard" to evaluate the performance of other designs as it is based on fully observed data (i.e., no unobserved response).

Table 1 shows the simulation results, including the average percentage of early termination, average sample size, and the trial duration in months. To evaluate the performance of the different methods, we used the CD method as the gold standard and calculated the differences in the average percentage of early termination (and sample size) between the CD method and each of the other methods, the naive, OD, WT and proposed MI methods. Smaller differences indicate a better performance.We displayed the results in Figures 1 and 2, where the line closer to the zero horizontal line represents a better performance. In general, compared to the CD method, the naive method was overly stringent and tended to terminate the trial when the response rate was actually acceptable (i.e., higher than the lower bound ϕ). In contrast, the OD method was excessively liberal and was less likely to terminate the trial when the response rate was not acceptable (i.e., lower than ϕ). The proposed MI method performed best, yielding termination probabilities very similar to those of the CD method; that is, both designs terminated the trial with high (or low) probabilities when the response rate was lower (or higher) than ϕ. The WT method performed better than the naive method and OD method, but not as well as the MI method, especially when 90% of events occurred at the later half period of the assessment window. Compared to the CD method, the advantage of the MI method is that its trial duration was much shorter because of its ability to support continuous accrual. In most cases, the duration of the trial under the MI method was about a quarter of that under the CD method. For

example, when 90% of events occurred in the latter half of the assessment period $(T/2, T)$ and $\phi = 0.4$, if the true event rate $\pi = 0.3$, the CD method terminated 71.5% of the trials for futility, with an average sample size of 26.0 patients and a trial duration of 72.5 months. Compared to the CD method, the OD method terminated a lower percentage of trials for futility (64.3%), while the WT method over-terminated the trials (76.3%). Our proposed MI approach performed similarly to the CD method, with a termination rate of 70.3% for futility and a sample size of 27.6, but had a much shorter trial duration (17.1 versus 72.5 months). Although the naive method terminated the trial for futility with a high percentage of 89.1%, it also terminated the trial with a higher percentage when the true response rate was high. For example, when the true response rate was 0.6, the naive method terminated the trial 13.0% of the time, whereas the termination rates of the other four methods were all lower than 4%. A similar pattern of results was observed under $\phi = 0.3$ and 70% of responses occurring within $(T/2, T)$.

3.2. Sensitivity analyses

To evaluate the robustness of the proposed MI approach, we examined its performance when the time to response was generated from a log-logistic distribution. The results (see Table 2) were very similar to those obtained when the time to response was generated from the Weibull distribution (i.e., Table 1). The differences in the percentage of early termination for futility were typically less than 3%, suggesting that our method is not sensitive to the distribution of the time to response. Again, the proposed MI approach yielded average termination rates and sample sizes similar to those of the CD method, but had much shorter trial durations.

In addition, we conducted a sensitivity analysis in terms of the number of partitions used in the piecewise exponential model, the accrual rate, and the prior distribution of λ_j. We considered J=8 and 12 partitions (in the piecewise exponential model), faster accrual rate of $\rho = 3$ and 4 patients per month, and prior hyperparameters $c_1 = \cdots = c_J$ =0.05 and 0.1. Across different settings, the results (see Table 3) were generally similar to those displayed in Table 1, where $J = 6, \rho = 2$ and $c_1 = \cdots = c_J$ =0.01. For example, when $\phi = 0.4$ and $\pi = 0.3$, with data simulated from Weibull distributions, the proposed method terminated the trial for futility $71.5\%, 72.2\%$ and 69.0% of the time when $J = 6, 8, 12$, respectively, and 71.5%, 67.3% and 70.0% of the time when $\rho = 2, 3, 4$, respectively. These results suggest that the proposed method is robust to these model parameters.

4. Conclusions

We have proposed a Bayesian phase II single-arm design with continuous monitoring to estimate the efficacy of the experimental drug. We incorporate a Bayesian stopping rule to terminate the trial early for futility and avoid assigning an unacceptable number of patients to inefficacious treatments. We handle the missing responses using the multiple imputation approach by modeling the time to response data using a piece-wise exponential model. Extensive simulations show that the proposed design yields desirable operating characteristics under various scenarios. The proposed design dramatically reduces the trial duration.

In this article, we focus on the trials with binary outcomes. Our proposed multiple imputation approach can be extended to the trials with ordinal outcomes. According to Response Evaluation Criteria in Solid Tumors, objective response in solid tumors can be classified into four ordered categories: complete response, partial response, stable disease and progressive disease. We can use the proportional-odds cumulative logit model [27] for the ordinal response. To handle and impute the missing ordinal data, we can device time-to-event models for each response category and link them with copula [28] to accommodate within-patient correlation.

Acknowledgement

The authors would like to thank associated editor and two reviewers for insightful and constructive comments that substantially improved the paper. Yuan's research was partially supported by Award Number R01 CA154591 and P50 CA098258 from the National Cancer Institute. Cai's research was supported by the National Institutes of Health's Clinical and Translational Science Award grant (UL1 TR000371), awarded to the University of Texas Health Science Center at Houston in 2012 by the National Center for Clinical and Translational Sciences. Article content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu.

References

1. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* 1961; **13**:346–353.
2. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**:143–151.
3. Simon RM. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
4. Green SJ, Dahlberg S. Planned versus attained design in phase II clinical trials. *Statistics in Medicine* 1992; **11**:853–862.
5. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 1997; **16**:2701–2711.
6. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. *Statistics in Medicine* 1994; **13**:1727–1736.
7. Hanfelt JJ, Slack RS, Gehan EA. A modification of Simon's optimal design for phase II trials when the criterion is median sample size. *Controlled Clinical Trials* 1999; **20**:555–566.
8. Jung SH, Carey M, Kim KM. Graphical search for two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 2001; **22**:367–372.
9. Shuster J. Optimal two-stage designs for single arm phase II cancer trials. *Journal of Biopharmaceutical Statistics* 2002; **12**:39–51.
10. Lin Y, Shih WJ. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics* 2004; **60**:482–490.
11. Tan SB, Machin D. Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine* 2002; **21**:1991–2012.
12. Sambucini V. A Bayesian predictive two-stage design for phase II clinical trials. *Statistics in Medicine* 2008; **27**:1199–1224.
13. Thall PF, Simon RM. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994; **50**:337–349.
14. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine* 1995; **14**:357–379.
15. Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine* 1997; **16**:1791–1802.
16. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clinical Trials* 2008; **5**:93–106.
17. Johnson VE, Cook JD. Bayesian design of single-arm phase II clinical trials with continuous monitoring. *Clinical Trials* 2009; **6**:217–226.
18. Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. *Statistics in Medicine* 2008; **27**:2802–2815.
19. Zohar S, Teramukai S, Zhou Y. Bayesian design and conduct of phase II single-arm clinical trials with binary outcomes: A tutorial. *Contemporary Clinical Trials* 2008; **29**:608–616.
20. Yuan Y, Yin G. Robust EM continual reassessment method in oncology dose finding. *Journal of the American Statistical Association* 2011; **106**:818–831.
21. Follmann DA, Albert PS. Bayesian monitoring of event rates with censored data. *Biometrics* 1999; **55**:603–607.
22. Zhao L, Taylor JM, Schuetze SM. Bayesian decision theoretic two-stage design in phase II clinical trials with survival endpoint. *Statistics in Medicine* 2012; **31**:1804–1820.
23. Cheung YK, Thall PF. Monitoring the rates of composite events with censored data in phase II clinical trials. *Biometrics* 2002; **58**:89–97.
24. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.
25. Arjas E, Gasbarra D. Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statistica Sinica* 1994; **4**:505–524.
26. Aslanidou H, Dey DK, Sinha D. Bayesian analysis of multivariate survival data using Monte Carlo methods. *Canadian Journal of Statistics* 1998; **26**:38–48.
27. Agresti A. *Categorical Data Analysis*. Wiley, 3rd edition, 2012.
28. Nelsen R. *An Introduction to Copulas*. New York: Springer, 1999.

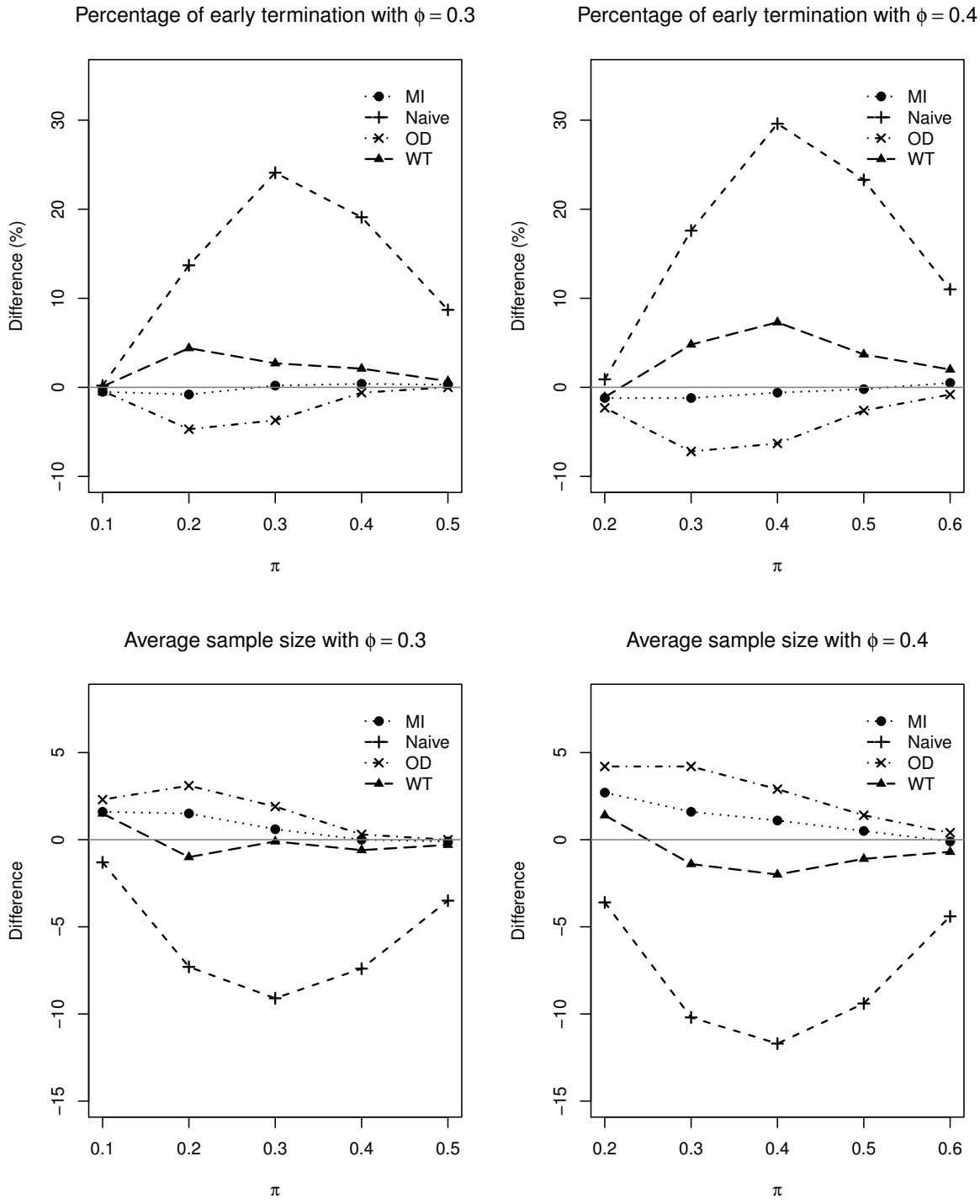

Figure 1. Differences in the average percentage of early termination and sample size between complete-data (CD) method and naive, observed-data (OD), weighted (WT), and the proposed multiple imputation (MI) methods with data generated from Weibull distributions and 90% of responses occurring in $(T/2, T)$

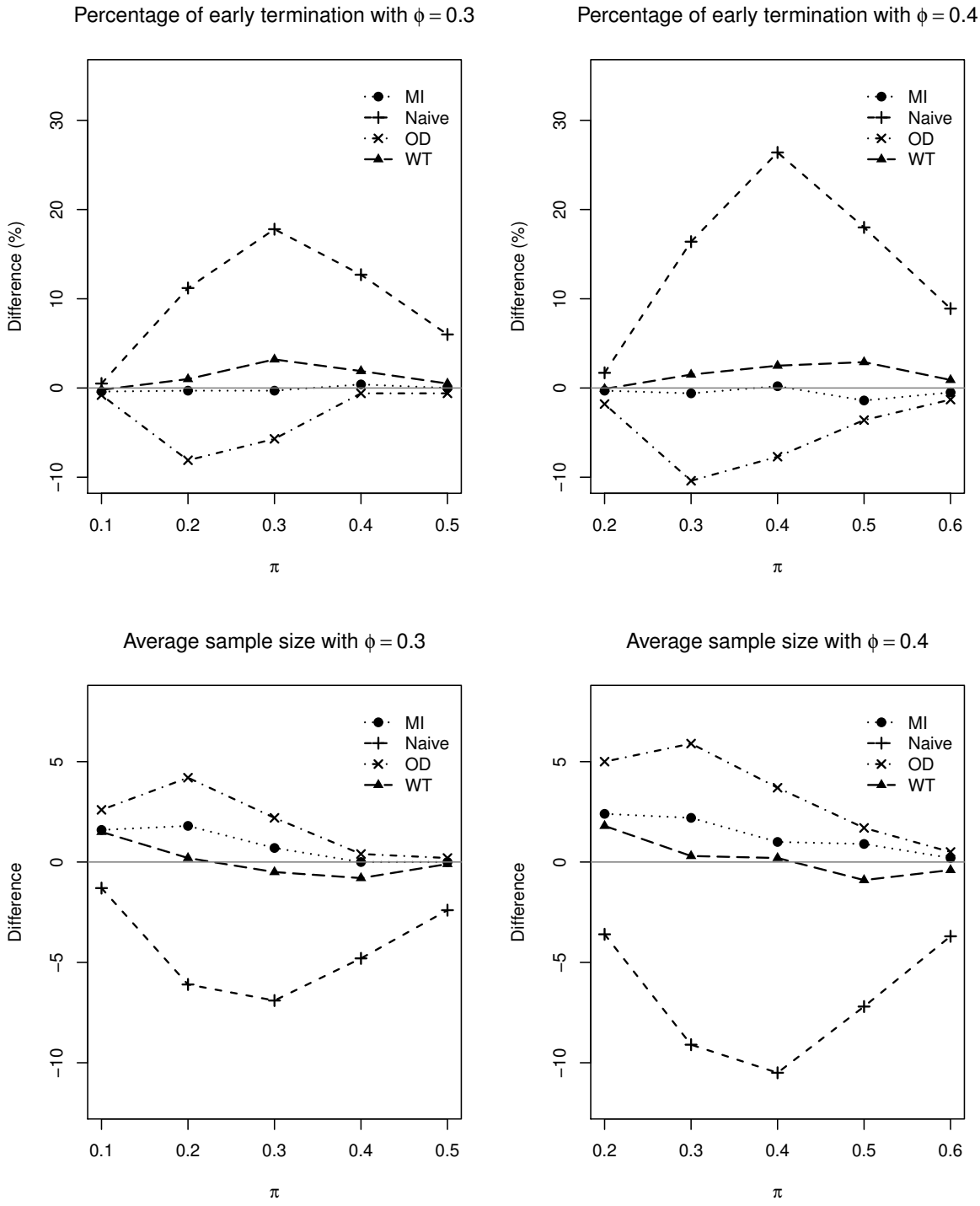

Figure 2. Differences in the average percentage of early termination and sample size between complete-data (CD) method and naive, observed-data (OD), weighted (WT), and the proposed multiple imputation (MI) methods with data generated from Weibull distributions and 70% of responses occurring in $(T/2, T)$

Table 1. Simulation results under naive, observed-data (OD), complete-data (CD), weighted (WT), and the proposed multiple imputation (MI) methods with data generated from Weibull distributions.

% of responses			% of early termination					Average sample size					Trial duration (months)				
in $(T/2, T)$	ϕ	π	Naive	OD	CD	MI	WT	Naive	OD	CD	MI	WT	Naive	OD	CD	MI	WT
90%	0.3	0.1	99.5	98.9	99.3	98.8	99.4	8.2	11.8	9.5	11.1	11.0	6.6	8.4	27.9	8.1	8.0
		0.2	89.0	70.6	75.3	74.5	79.7	16.6	27.0	23.9	25.4	22.9	11.1	16.8	68.5	15.9	14.5
		0.3	57.6	29.8	33.5	33.7	36.2	28.1	39.1	37.2	37.8	37.1	17.6	23.8	103.9	23.1	22.7
		0.4	29.4	9.7	10.3	10.7	12.4	38.4	46.1	45.8	45.8	45.2	23.5	27.9	124.4	27.7	27.3
		0.5	11.7	3.0	3.0	3.3	3.7	45.2	48.7	48.7	48.6	48.4	27.3	29.3	128.2	29.2	29.1
	0.4	0.2	99.6	96.4	98.7	97.5	97.6	8.8	16.6	12.4	15.1	13.8	6.9	10.9	35.4	10.1	9.4
		0.3	89.1	64.3	71.5	70.3	76.3	15.8	30.2	26.0	27.6	24.6	10.7	18.5	72.5	17.1	15.4
		0.4	59.7	23.8	30.1	29.5	37.4	27.3	41.9	39.0	40.1	37.0	17.2	25.4	105.8	24.3	22.5
		0.5	32.3	6.4	9.0	8.8	12.7	36.9	47.7	46.3	46.8	45.2	22.6	28.7	122.1	28.2	27.3
		0.6	13.0	1.2	2.0	2.5	4.0	44.7	49.5	49.1	49.0	48.4	27.0	29.7	125.7	29.5	29.1
70%	0.3	0.1	99.8	98.5	99.3	98.9	99.1	8.5	12.4	9.8	11.4	11.3	6.8	8.8	28.4	8.3	8.2
		0.2	87.4	68.1	76.2	75.9	77.2	17.1	27.4	23.2	25.0	23.4	11.4	17.0	64.5	15.6	14.8
		0.3	51.4	27.9	33.6	33.3	36.8	30.4	39.5	37.3	38.0	36.8	18.9	24.1	99.6	23.2	22.5
		0.4	22.0	8.7	9.3	9.7	11.2	41.3	46.5	46.1	46.1	45.3	25.1	28.1	117.9	27.9	27.4
		0.5	10.5	3.9	4.5	4.5	5.0	45.6	48.2	48.0	48.0	47.9	27.6	29.1	117.6	28.9	28.8
	0.4	0.2	99.4	95.9	97.7	97.4	97.6	9.5	18.1	13.1	15.5	14.9	7.3	11.7	36.4	10.4	10.0
		0.3	87.6	60.8	71.2	70.6	72.7	16.8	31.8	25.9	28.1	26.2	11.2	19.4	69.3	17.3	16.3
		0.4	54.4	20.3	28.0	28.2	30.5	28.9	43.1	39.4	40.4	39.6	18.1	26.1	100.6	24.5	24.0
		0.5	26.5	4.9	8.5	7.1	11.4	39.3	48.2	46.5	47.4	45.6	24.0	29.0	114.1	28.5	27.5
		0.6	11.2	1.0	2.3	1.8	3.2	45.4	49.6	49.1	49.3	48.7	27.5	29.8	114.8	29.6	29.3

Table 2. Sensitivity analysis with data generated from log-logistic distributions under naive, observed-data (OD), complete-data (CD), weighted (WT), and the proposed multiple imputation (MI) methods.

% of responses			% of early termination					Average sample size					Trial duration (months)				
in $(T/2, T)$	ϕ	π	Naive	OD	CD	MI	WT	Naive	OD	CD	MI	WT	Naive	OD	CD	MI	WT
90%	0.3	0.1	99.9	98.9	99.6	98.8	99.8	8.5	12.7	10.2	12.0	10.6	6.8	8.9	29.8	8.6	7.8
		0.2	88.3	70.8	74.6	74.4	78.7	16.4	26.4	23.5	24.7	22.7	11.0	16.5	67.0	15.5	14.4
		0.3	52.9	26.6	29.4	30.5	37.7	29.7	39.7	38.4	38.5	36.3	18.6	24.2	106.8	23.5	22.2
		0.4	26.6	8.3	9.5	9.8	11.0	39.5	46.5	46.0	46.0	45.5	24.1	28.1	124.2	27.8	27.5
		0.5	11.9	3.0	3.3	3.2	4.1	45.2	48.7	48.5	48.6	48.2	27.3	29.3	127.1	29.2	29.0
	0.4	0.2	99.8	95.8	97.5	96.8	98.4	9.1	16.9	12.6	15.4	13.4	7.1	11.1	35.9	10.3	9.2
		0.3	88.9	61.9	69.7	68.2	77.3	15.9	31.0	26.6	28.1	24.1	10.7	19.0	74.0	17.4	15.1
		0.4	60.3	22.7	28.4	29.2	37.2	27.2	42.6	39.9	40.2	36.7	17.1	25.7	107.9	24.4	22.4
		0.5	31.1	7.2	9.9	9.8	14.9	37.7	47.3	46.0	46.4	44.3	23.1	28.5	120.4	28.0	26.8
		0.6	15.0	1.0	1.7	1.9	5.0	43.8	49.6	49.3	49.3	48.0	26.6	29.8	124.9	29.6	28.9
70%	0.3	0.1	100.0	98.9	99.4	99.2	99.4	8.1	11.6	9.2	10.7	11.1	6.6	8.3	26.8	7.9	8.1
		0.2	84.6	66.4	73.8	72.8	78.2	18.2	27.8	23.8	25.4	23.0	12.0	17.3	66.0	15.9	14.6
		0.3	45.4	22.2	27.6	26.9	35.7	32.9	41.6	39.6	40.2	36.9	20.3	25.3	105.4	24.5	22.5
		0.4	22.2	8.6	10.8	10.7	11.0	41.2	46.4	45.5	45.7	45.5	25.1	28.0	116.8	27.6	27.5
		0.5	8.4	2.6	2.9	2.9	3.5	46.6	48.9	48.7	48.3	48.5	28.1	29.4	118.9	29.3	29.2
	0.4	0.2	98.9	96.6	97.8	97.7	97.3	9.0	16.3	11.6	14.3	14.2	7.1	10.8	32.3	9.8	9.7
		0.3	87.8	62.3	73.4	71.9	73.6	16.6	31.0	24.9	27.2	25.8	11.1	19.0	66.5	16.8	16.1
		0.4	55.1	20.9	28.9	29.1	32.2	28.9	42.9	39.3	40.1	38.6	18.1	25.9	100.5	24.3	23.5
		0.5	28.5	6.8	10.8	9.5	9.9	38.4	47.3	45.5	46.3	46.1	23.5	28.5	111.1	27.9	27.8
		0.6	12.5	1.1	2.5	2.1	2.4	44.8	49.5	48.9	49.2	49.0	27.1	29.8	113.7	29.6	29.5

Table 3. Sensitivity analysis with different numbers of partitions, accrual rates and prior distributions of λ_j under the proposed MI approach with $\phi = 0.4$ and 90% of responses occurring within $(T/2, T)$

		Weibull			log-logistic		
Settings	π	% of termination	Average sample size	Trial duration (months)	% of termination	Average sample size	Trial duration (months)
J=8	0.2	96.7	15.9	10.6	97.2	15.0	10.1
	0.3	72.2	26.8	16.6	68.2	27.8	17.2
	0.4	29.7	39.7	24.1	28.7	40.3	24.4
	0.5	8.3	46.9	28.3	8.5	46.7	28.2
	0.6	2.5	49.0	29.5	2.5	49.0	29.5
J=12	0.2	96.2	15.6	10.4	96.6	15.5	10.4
	0.3	69.0	27.9	17.3	68.8	27.9	17.3
	0.4	31.7	39.2	23.8	28.1	40.8	24.7
	0.5	8.7	46.7	28.2	8.5	46.8	28.2
	0.6	2.6	49.0	29.4	3.1	48.7	29.3
ρ=3	0.2	97.7	16.4	8.2	96.7	16.1	8.1
	0.3	67.3	29.1	13.3	68.7	28.5	13.0
	0.4	32.8	43.2	17.7	29.8	40.6	18.1
	0.5	9.7	46.5	20.6	11.9	45.5	20.2
	0.6	3.3	48.7	21.5	3.1	48.8	21.6
ρ=4	0.2	97.1	16.8	7.0	96.9	17.2	7.2
	0.3	70.0	28.8	10.8	68.6	29.3	11.0
	0.4	32.4	40.0	14.6	30.8	40.4	14.8
	0.5	9.9	46.6	16.9	12.0	45.7	16.6
	0.6	3.9	48.5	17.5	2.7	49.0	17.7
c_j=0.05	0.2	97.3	14.2	9.7	97.3	14.3	9.8
	0.3	73.8	25.1	15.7	71.9	26.5	16.5
	0.4	34.5	37.9	23.1	34.2	37.7	23.0
	0.5	13.0	45.0	27.2	13.8	44.6	27.0
	0.6	4.7	48.1	29.0	4.4	48.2	29.0
c_j=0.1	0.2	97.0	13.2	9.2	97.8	12.9	9.0
	0.3	75.4	23.8	15.0	72.9	24.6	15.5
	0.4	36.8	36.3	22.3	39.1	35.6	21.9
	0.5	15.4	43.9	26.6	16.1	43.7	26.5
	0.6	5.4	47.8	28.8	6.3	47.5	28.6