DOSE-FINDING DESIGNS FOR PHASE I CLINICAL TRIALS IN ONCOLOGY AND USE OF SELECTIVE PHENOTYPYING TO INCREASE POWER OF GENETIC ASSOCIATION STUDIES

Yunfei Wang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics of the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:

Anastasia Ivanova

Ethan M Lange

Yun Li

Matthew C Foster

Anna Snavely

ABSTRACT

YUNFEI WANG: DOSE-FINDING DESIGNS FOR PHASE I CLINICAL TRIALS IN ONCOLOGY AND USE OF SELECTIVE PHENOTYPYING TO INCREASE POWER OF GENETIC ASSOCIATION STUDIES
(Under the direction of Anastasia Ivanova & Ethan M. Lange)

The goal of phase I clinical trials in oncology is to find a dose for a study that has acceptable toxicity or adverse effect associated with a pre-specified probability in patients experiencing DLT (dose limiting toxicity) for a drug. We propose a dose-finding design for Phase I oncology trials where each new patient is assigned to the dose most likely to be the target dose given observed data. The only assumption is that the dose-toxicity curve is non-decreasing. This method is especially beneficial when it is desirable to enroll patients into a study as soon as they present for the trial. To prevent assignments to doses with limited toxicity information in fast accruing trials we propose assigning temporary fractional toxicities to patients still in follow-up.

The goal of a Phase I clinical trial in oncology is to find a dose with acceptable dose limiting toxicity rate. Often when a cytostatic drug is investigated or when the maximum tolerated dose is defined using a toxicity score, the main endpoint in a Phase I trial is continuous. We propose a new method to use in a dose-finding trial with continuous endpoints. The new method performs on par with other methods and provides more flexibility in assigning patients to doses in the course of the trial when the rate of accrual is fast relative to the follow-up time.

Blood-based biomarkers and other quantitative measures can provide valuable insights into disease etiology and are often used as intermediate outcomes for identifying risk factors associated with disease. Genome-wide association studies (GWAS) between quantitative traits

and single-nucleotide polymorphisms (SNPs) are routinely performed on large samples from population-based cohorts. Replication studies are an important step in controlling the type I error rate of reported GWAS findings. Many potential replication cohorts have existing genome-wide SNP data but have not yet measured the quantitative trait of interest. Measuring these traits can be expensive and time consuming, which can deter studies from pursuing replication. Given the expense and time of measuring these quantitative traits on large samples, it would be desirable to identify a subset of subjects that could be phenotyped to optimize statistical power under fixed sample size constraints. We describe an approach of utilizing existing genotype data to identify an optimal subset of samples to be phenotyped and included in a genetic replication study. Specifically, we have developed a simulated annealing-based algorithm to optimally select samples to be phenotyped conditional on a list of candidate SNPs and available genotype data for those SNPs under a fixed sample size constraint. We demonstrate the increase in power of our approach relative to random sampling using simulations and a real replication study for C-reactive protein levels. Our approach is flexible enough to allow constraints on available genotype counts and differential weighting of SNPs in the power calculations.

To my father and the memory of my mother

ACKNOWLEDGMENTS

I would like to thank Dr. Anastasia Ivanova, my advisor, for her continuous encouragement and assistance during the research on Phase I clinical trial study. I'm grateful to Dr. Ethan M. Lange, my co-advisor and supervisor, for his guidance in the research on genetics study and his generous support for my graduate study all these years. I'm thankful to Dr. Yun Li, for her one year support and her very poignant suggestions on the genetics paper. I would like to thank Drs. Matthew C. Foster and Anna Snavely, who kindly found time for my exams despite their busy schedules and provided useful comments on the clinical trial papers. I am also appreciated that Dr. Matthew C. Foster and the JHS group kindly let me use their data in my research.

I'm indebted to my wife, Liping Song, who takes care of the family though she has a full time job, and thank her for her love, support and sacrifice. I would like to thank my children, Yu, David and Henry, for the joy they have brought to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

LITERATURE REVIEW 1: DOSE-FINDING DESIGNS IN PHASE I CLINICAL TRIALS IN ONCOLOGY

1.1 Introduction

The primary goal for Phase I oncology trials is to find the maximally tolerable dose (MTD), the dose with certain probability of dose limiting toxicity (DLT). One important assumption for dose-finding is that the toxicity for a new intervention increases monotonically with doses in Phase I clinical trials. The toxicity associated with a dose is so called dose limiting toxicity (DLT) which is often defined as treatment related non-hematological toxicity of grade 3 or higher, or treatment related hematological toxicity of grade 4 or higher, where toxicity is measured on a scale from 0 to 5.

In Phase I oncology clinical trials, there are some issues that need be taken into considerations: ethics, delayed onset of DLT, and continuous outcome.

1) The ethical issue is the reason that we seldom use healthy volunteers and that we should neither put many patients in those doses far above MTD, nor put many patients in those doses far below MTD when the patients consent to the study wishing to find cure by taking a new treatment.

2) Traditional designs are usually used to find MTD in the first cycle, however, delayed onset of toxicity can be observed beyond the first cycle. In traditional designs, admitting new patients into a trial requires that the current patients have completed follow-up to determine the

dose assigned to the new patients. For a traditional design, it takes a long time to finish when there is delayed onset of toxicity. Therefore traditional designs are not practical in clinical practice since doctors will not let patients waiting for treatment.

3) Another purpose of dose-finding in Phase I trials is to provide an efficacy dose of cytostatic agents for the subsequent Phase II studies though to find a MTD is a primary goal for cytotoxic agents in Phase I trials. Continuous outcomes, such as measurement of target inhibition or pharmacokinetic endpoints such as plasma drug concentrations, are usually collected in such trials. The hypothesis that toxicity increases with increased dose is not necessary true for cytostatic agents. Toxicity may happen beyond the dose to yield sufficient efficacy. New drugs may incur only moderate toxicity with no DLT being observed. Therefore, there is a need to turn the moderate toxicity into continuous scores and to find the dose targeting continuous outcome or scores.

Our goal is to develop new approaches for dose-finding design in Phase I clinical trial for both binary and continuous outcomes with better operating characteristics without or with delayed onset toxicity.

1.2 Designs or methods to solve the mentioned issues

Designs for Phase I oncology trials can be categorized into two groups: non-parametric designs and parametric designs. In the following literature review section, we will describe these designs or methods in details.

1.2.1 Non-parametric designs

In non-parametric designs, no prior assumption of dose-toxicity curve is stipulated with no parameters being estimated for dose escalation. Instead, dose allocation depends on the calculated statistics such as toxicity rate estimates or t statistics according to random walk rules or up-and-down rules. The principle of these rules is that the dose for the future cohort of patient(s) is decided based on the statistics of the observed data from the previous cohort of treated patient(s). Traditional 3+3 design (Storer, 1989) which is a special form of A+B design (Lin and Shih, 2001), biased coin design (BCD) (Derman, 1957, Durham and Flournoy, 1994), group up-and-down design (Wetherill, 1963), moving average up-and-down design (Ivanova et al., 2003), group up-and-down design (Gezmu and Flournoy, 2006) are examples of non-parametric design for dose finding trials. Some of these designs, such as BCD or group up-and-down designs are Markov processes, and therefore their limiting distributions are easy to compute.

Many non-parametric designs use isotonic regression to estimate dose toxicity curve. The only assumption is that the DLT rate is nondecreasing function of the dose. These designs include isotonic design (Leung and Wang, 2001), and the Cumulative Cohort Design (Ivanova et al., 2007). The estimated toxicity rate at current dose is used to decide whether the next cohort is treated at current dose or one dose above or below. The rule for dose allocation is as follows:

i. If $\hat{q}_j \le \Gamma - \Delta$, then treat the next cohort at d_{j+1}.

ii. If $\hat{q}_j \ge \Gamma + \Delta$, then treat the next cohort at d_{j-1}.

iii. Otherwise, treat the next cohort at d_j.

The advantage of many non-parametric designs is their simplicity. It is easy to understand and to implement with no need of special software. The drawback is that many of these designs

do not use all available data in the decision rule and therefore converge slowly to the target dose or doses near target.

1.2.2 Parametric designs for dose-finding trials

In parametric designs, the probability of DLT can be given by a multi-parameter model:

$$p_i = F\left(\frac{d_i - \alpha}{\beta}\right)$$

where α and β is unknown shift and scale parameters (Rosenberger and Haines, 2002). The quantile of F is $\mu = \alpha + \beta F^{-1}(\Gamma)$ where Γ is a specified probability of toxicity and μ can be MTD. Two parameters are specified in the probability function and used in designs such as escalation with overdose control (EWOC) (Babb et al., 1998). Besides the two-parameter models, one parameter model has been also used in designs such as continual reassessment method (CRM) (O'Quigley et al., 1990).

In parametric designs, parameters are estimated using maximum likelihood or Bayesian methods. Parametric designs can be realized using maximum likelihood technique in frequentist form (O'Quigley and Shen, 1996), however, most of them belong to a class of Bayesian designs using Bayesian theorem (O'Quigley et al., 1990). A typical Bayesian decision design has four features: (a) a data model, (b) a prior distribution for parameters, (c) loss function, and (d) a set of actions (Whitehead, 2006). Decision of dose allocation is based on the posterior distribution of parameter(s) according to Bayesian theorem. To be specific, in one parameter model, the next patient is treated at the dose with the toxicity probability close to target p with parameter $\hat{\varphi}$ being calculated based on all the outcomes. The next patient will be treated at $d_{j+1} = \arg\min |F(d_k, \hat{\varphi}_j) - \hat{p}|$ (Cheung, 2010).

The CRM is a method to find MTD with certain toxicity rate for binary outcome. The one parameter working model is $P\left(d_j, a\right) = b_j^a$ where d_j is the dose, and j=1,…, K (the dose level), and a is a parameter. In a Bayesian form of the CRM, the prior for a is specified, for example, $f(a)=exp(-a)$. The posterior mean of parameter a is calculated using all the outcomes. Dose assignment is based on mean posterior probability that is closest to the target toxicity rate**.** The CRM can be used to reduce the number of patients to receive lower doses which may far below MTD as well and to achieve a more accurate estimation of MTD.

EWOC is a modified CRM in essence with safety measure to restrict a pre-specified proportion of patients to expose to a dose far above MTD. EWOC assigns doses based on the posterior probability that is overdosing. The loss function used in EWOC design is

$$l_a\left(x,\gamma\right) = \begin{cases} \alpha\left(\gamma - x\right) & \text{if } x \le \gamma \text{ (undose)} \\ \left(1-\alpha\right)\left(x-\gamma\right) & \text{if } x > \gamma \text{ (over dose)} \end{cases}$$

where γ is the parameter and $\gamma \in [X_{min}, X_{max}]$.

Neuenschwander et al. (2008) proposed a Bayesian method which can be used to solve an overdose issue too, and defined a rule to assign dose based on maximum posterior probability within an interval. The loss function in the approach is

$$L\left(\theta, d\right) = \begin{cases} l_1 = 1 & \text{if } \pi_\theta(d) \in (0, 0.2] \quad \text{(under-dosing)} \\ l_2 = 0 & \text{if } \pi_\theta(d) \in (0.2, 0.35] \quad \text{(Targeted toxicity)} \\ l_3 = 1 & \text{if } \pi_\theta(d) \in (0.35, 0.6] \quad \text{(Excessive toxicity)} \\ l_4 = 2 & \text{if } \pi_\theta(d) \in (0.6, 1] \quad \text{(Unacceptable toxicity)} \end{cases}$$

Ji et al. (2007, 2010) described a rule to assign doses based on the maximum posterior probability within an interval. A set of loss functions (Ji et al., 2010) were defined as following,

$$L(D, p_i) = \begin{cases} N_D, & \text{if } p_i - p_T < -\varepsilon_1 \\ K_D, & \text{if } -\varepsilon_1 \le p_i - p_T < \varepsilon_2 \\ 0, & \text{if } p_i - p_T > \varepsilon_2 \end{cases}$$

$$L(S, p_i) = \begin{cases} N_S, & \text{if } p_i - p_T < -\varepsilon_1 \\ 0, & \text{if } -\varepsilon_1 \le p_i - p_T < \varepsilon_2 \\ M_S, & \text{if } p_i - p_T > \varepsilon_2 \end{cases}$$

$$L(E, p_i) = \begin{cases} 0, & \text{if } p_i - p_T < -\varepsilon_1 \\ K_E, & \text{if } -\varepsilon_1 \le p_i - p_T < \varepsilon_2 \\ M_E, & \text{if } p_i - p_T > \varepsilon_2 \end{cases}$$

where $L(D, p_i)$ is the loss function for de-escalating, $L(S, p_i)$ is for staying, and $L(E, p_i)$ is for escalating.

The advantage of parametric designs over non-parametric designs is that all the available information is used to determine the dose for the next patient. On the other hand, parameter designs usually are harder to understand and more difficult to use compared to non-parametric designs.

1.2.3 Dose-finding for time to event outcome or delayed onset toxicity in Phase I study

In designs described previously, the outcomes of all previous patients should be known before the next patient can be assigned. In many Phase I oncology trials, the follow-up for toxicity is rather long. For example, in Lineberger Comprehensive Center trial of a nucleoside analog to treat acute myeloid leukemia, patients were observed for DLT for 35 days after the start of treatment. The follow-up for toxicity in a Phase I trial of perillyl alcohol at the University of Wisconsin Medical School was 4 weeks (Cheung and Chappell, 2000). Yuan (2009) describe a Phase I trial where patients were followed for toxicity for 3 months. Therefore, dose-finding

designs using continuous enrollment are needed to solve the problem of mandate trial suspension in traditional designs.

Several designs have been proposed to incorporate long follow-up time for toxicity in a Phase I trial. Two approaches have been used in these designs.

Cheung and Chappell (2000) assumed that time to toxicity is distributed uniformly in $(0, T)$ given that a DLT was observed in the follow-up window of (0, T). They proposed a modified CRM by introducing weight into the likelihood. In TITE-CRM, the dose assignment will converge to the MTD as in CRM (Cheung and Chappell, 2002). But due to rapid dose escalation, this method may end up with more patients in high toxic doses. Modifications for the TITE-CRM have been proposed by Braun et al. (2003, 2006) with the latter describing a generalization of TITE-CRM for early- and late-onset toxicities

The other approach is based on a Bayesian survival analysis methodology. This approach requires using Gibbs sampler or Markov Chain Monte Carlo (MCMC) algorithm for sampling from probability distributions through constructing a Markov chain to find the desired equilibrium distribution.

Kuo and Smith (1992) detailed how the Gibbs Sampler was used in Bayesian computation in survival models, and Rosner described the Gibbs Sampler in monitoring failure-time endpoints (Rosner, 2005). Yuan (2009) proposed a dose-finding method by jointly modeling toxicity and efficacy as time-to-event outcomes. In their study, they used the Cox proportional hazards model to model both toxicity and efficacy jointly with Weibull baseline hazards; and simulated the posterior distribution through Gibbs sampler assuming non-informative uniform vague priors. Braun et al. (2005) defined hazard function as a function of three parameters and total hazard of toxicity which depends on patient's time on study, number

of administrations and the times of treatment; assumed that the general beta distribution priors for the parameters and that the hazard increases to the maximum and decrease to zero linearly, and calculated the posterior through MCMC simulation to determine the maximum tolerated schedule rather than MTD.

Often times it takes long to observe toxicity compared to accrual. In some trials, when pace of disease is so rapid that a patient may succumb to the disease while waiting, it is desirable to enroll a patient into the trial and start treating as soon as possible. None of the existing time to event dose methods is suitable in this set up. The method of Bekele et al. (2008) solves the problem of fast accrual by prescribing when a trial should be halted as toxicity rate at the current dose becomes uncertain. As noted in Cheung and Chappell (2010) rapid accrual, especially in the beginning of the trial, may lead to assigning patients to doses with high toxicity rate when TITE-CRM is used. A rule frequently used is to assign at least 3 patients to a dose and wait for all 3 to be fully followed before any more patients can be assigned to that or higher doses. Then, if continuous enrollment is desirable, while the first 3 patients are being followed at the current dose, patients are assigned to a lower dose. In dose-finding designs (Wetherill, 1963, Ivanova et al., 2007) , the next assignment is determined based on the knowledge of the current dose. When assignments are being made to lower doses as well as the "current" dose, it is no longer clear what dose is the current dose. We describe an easy to use and understand dose-finding method where the dose-assignment strategy is not tied to a current dose and the next assignment is made to the dose that is most likely to be the MTD.

1.2.4 Dose-finding for continuous outcome

Dose limiting toxicity (yes or no) is the primary endpoint in most oncology phase I trials of cytotoxic agents. If a cytostatic agent is being investigated, toxicity is usually not a limiting factor and a continuous biomarker endpoint is often used as a primary endpoint, for example a measure of target inhibition or pharmacokinetic endpoints such as plasma drug concentrations that correlate with biological activity (Le Tourneau et al., 2009), or percentage inhibition of an enzyme (Plummer et al., 2008). Continuous endpoint also arises when multiple toxicity events in different body systems and multiple toxicity grades are combined into a single score (Bekele and Thall, 2004, Ezzalfani et al., 2012, Chen et al., 2010). A number of scores have been proposed recently, for example, total toxicity burden (TTB) (Bekele and Thall, 2004), total toxicity profile (TTP) (Ezzalfani et al., 2012), equivalent toxicity score (ETS) (Chen et al., 2010), and average toxicity score (ATS) (Bekele et al., 2010). These scores are generated through combining information from various toxicity grades, grades 1 through 5, into a single number with the goal of better reflecting toxicity burden on a patient compared to the binary outcome of dose limiting toxicity (DLT) that is usually defined as treatment related non-hematological toxicity of grade 3 or higher or treatment related hematological toxicity of grade 4 or higher.

Many methods in phase I clinical trials are developed for dose-finding with binary outcome (DLT, yes or no). Designs for continuous outcome have been proposed through controlling dose escalation via a binary assessment of efficacy (Mandrekar et al., 2007, Mandrekar et al., 2009, Hunsberger et al., 2005). Several methods work with continuous endpoint directly (Ivanova and Kim, 2009, Eichhorn and Zacks, 1973). Dose-finding designs for

toxicity score include *t*-statistics design (Ivanova and Kim, 2009), extended isotonic design (Chen et al., 2010), Quasi-CRM (Yuan et al., 2007) , Quasi-Likelihood-CRM (Ezzalfani et al., 2012) and other designs (Lee et al., 2010).

All the methods mentioned earlier require outcome of a patient being observed quickly. In many trials, time to observe the outcome is relatively long compared to the accrual rate. In some of such trials, for example in trials where urgent treatment is needed, it is desirable to assign a dose to a patient as soon as the patient enrolls in the trial. The proposed design allows making the best possible assignment for each incoming patient using all information available.

1.3 Other related methods or rules

In this section we review several methods or rules we have seen in phase I clinical trial protocols.

1.3.1 Start-up rule

It is important to avoid escalation that is too rapid or too slow (Cheung, 2005). With that goal in mind, the group size s in the start-up should be chosen according to the target toxicity level Γ. Ivanova et al. (Ivanova et al., 2003) suggested choosing group size according to the following formula $s = \lfloor \log(0.5)/\log(1-\Gamma) \rfloor$. For example, if $\Gamma = 0.5$, the start-up with $s = 1$ is used; if $\Gamma = 0.3$, the start-up with $s = 2$ is used; if $\Gamma = 0.2$, the start-up with $s = 3$ is used.

1.3.2 Isotonic regression

In clinical trials, the efficacy or toxicity of a drug is usually assumed to be a non-decreasing function of dose.

Several algorithms have been developed to compute maximum likelihood estimates under monotonic restriction. The most widely used one is pool adjacent violator algorithm (PAVA) (Barlow et al., 1972) to obtain isotonic estimates. The maximum likelihood estimates for $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_K)'$, can be obtained from unrestricted maximum likelihood estimates, $(\overline{y}_1, ..., \overline{y}_K)'$ for $j = 1, 2, ..., K$ using max-min formula

$$\hat{\mu}_j = \min_{t \in U_j} \max_{s \in L_j} \left(\frac{\sum_{h=s}^{t} n_h \overline{y}_h}{\sum_{h=s}^{t} n_h} \right)$$

where $L_j = \{1, ..., j\}$ and $U_j = \{j, ..., K\}$

PAVA yields restricted MLEs for normal and binomial outcomes when unrestricted MLEs are not correlated. Other methods includes simple averaging techniques (Mukarjee, 1988), quadratic optimization (Best and Chakravarti, 1990).

CHAPTER 2

LITERATURE REVIEW 2: THE USE OF SELECTIVE PHENOTYPING TO INCREASE THE POWER IN GENETICS ASSOCIATION STUDIES

2.1 Introduction

Blood-based biomarkers and other quantitative measures can provide valuable insights into disease etiology and are often used as intermediate outcomes for identifying risk factors associated with disease. Genome-wide association studies (GWAS) between quantitative traits and single-nucleotide polymorphisms (SNPs) are routinely performed on large samples from population-based cohorts. Replication studies are an important step in controlling the type I error rate of reported GWAS findings. Many potential replication cohorts have existing genome-wide SNP data but have not yet measured the quantitative trait of interest. Measuring these traits can be expensive and time consuming, which can deter studies from pursuing replication. Given the expense and time of measuring these quantitative traits on large samples, it would be desirable to identify a subset of subjects that could be phenotyped to optimize statistical power under fixed sample size constraints. The goal of this study is to find an approach of utilizing existing genotype data to identify an optimal subset of samples to be phenotyped and included in a genetic replication study.

2.2 Genetic association studies

Association studies have been widely used for most common diseases such as cancer, cardiovascular disease, and diabetes as an alternative strategy to pedigree-based linkage analysis studies which have been successfully applied to monogenic Mendelian diseases (Jimenez-Sanchez et al., 2001, Hirschhorn and Daly, 2005). For more complex diseases, the risk alleles are usually more probabilistic, or less penetrant, in increasing the chance of disease and many of the causal variants have relatively high frequency in the population. Linkage analysis is not well powered for searching for these common variants. Association studies represent a good strategy to uncover these variants (Hirschhorn and Daly, 2005, Risch and Merikangas, 1996, Rich, 2000, Cardon and Bell, 2001, Tabor et al., 2002, Carlson et al., 2004). While association analyses can be based on genotype transmission (from parent to offspring) in pedigrees, an approach designed to control for confounded association results due to population stratification (Hirschhorn and Daly, 2005, Schork et al., 2001), most genetic association studies are based on unrelated individuals. Genetic association studies are widely applied to both quantitative and qualitative (case/control) traits. According to scales of variants or single nucleate polymorphisms (SNPs), Balding (2006) described four types of association studies: candidate polymorphism (1 SNP), candidate gene (5-50 SNPs in a gene), fine-mapping (multiple SNPs in perhaps 5-50 genes), and genome-wide (> 300,000 well-chosen SNPs throughout the genome). The recent development of next generation sequencing technologies has now lead to a new class of "gene-burden" rare-variant association tests, which aggregate information from a defined set (usually some collection of variants in the same gene) of multiple variants into a single predictor (Li and Leal, 2008).

2.3 Genome-wide association studies (GWAS)

Genome-wide association studies (GWAS) that analyze the DNA variation throughout genome have become a powerful tool for investigating the role of common human genetic variation in common diseases (Bush and Moore, 2012). The goal of GWAS is to identify the variants which may be either directly causal or in linkage disequilibrium with the causal variant(s) (Hirschhorn and Daly, 2005). In GWAS, the association tests can be performed on a single locus with a single SNP, by far the most common application, or on multiple loci simultaneously with a combination of multiple SNPs (Bush et al., 2009). In these tests, logistic regression models or generalized linear regression models can be fitted for case and control studies or for quantitative trait designs, respectively, adjusted for covariates such as age, gender and principal components to control for population stratification. If family data or repeated measures are included then generalized estimating equations or linear mixed models are typically employed. Since usually over half a million SNPs are tested, a Bonferroni correction approach or a false discovery rate (FDR) procedure is used to control the rate of false positive results (Hochberg and Beniamini, 1990, van den Oord, 2008). Other procedures, such as permutation tests, are also occasionally used. It is now standard for top journals to require confirmation of new findings from a GWAS using a replication study from an additional independent sample (Chanock et al., 2007), which should repeat the design setting as close as possible and should be well-powered (Zollner and Pritchard, 2007).

2.4. Cost constraints of genotyping

Large scale genetic association studies such as GWAS can be prohibitively expensive due to the substantial sample sizes required to detect modest genetic effects. Historically, these costs

have largely been dominated by the high cost of genotyping. The cost for genotyping a single SNP was 0.50 $, later dropped to 0.01$ (Wang et al., 2006). The average cost of genotyping one sample was initially thousands of dollars per subject, but today the cost can be as low as a couple hundred dollars or less, depending on the genetic marker panel used. Today, GWAS data are routinely available on tens of thousands of subjects from deeply phenotyped cohorts. Though the expense of phenotyping samples can also be high, since GWAS are routinely performed in large "deeply-phenotyped" community based cohorts with existing phenotype data collected from earlier epidemiological studies, the major cost constraint, to date, has been the expense of genotyping (hundreds dollar per subjects) even though the price is dropping over time.

2.5 Statistical power

Statistical power is an important consideration for any experiment. Typically these calculations are made before initiating the experiment, usually to determine an appropriate sample size to achieve a desired probabilistic threshold for success in rejecting the null hypothesis in favor of the alternative when the null hypothesis is false. On occasion, *post-hoc* power is calculated after completion of the experiment if the experiment failed to reject the null hypothesis. The latter scenario is useful in helping the investigator to assess upper bounds on the possible deviation between the true and null values for a parameter of interest. Statistical power is determined by three factors: effect size, sampling error and the statistical significance threshold used to reject the null in favor of the alternative (Lipsey and Hurley, 1990). Effect size is the average difference in values between observations from different groups. Power increases with increasing effect size. Effect size is typically unknown and specified by the user. The choice is sometimes informed by results from previous experiments or by bounds based on meaningful

clinical differences. The second factor is sampling error. The larger the sample variability of the test statistic the harder it is to reject the null hypothesis. Power can be improved by decreasing sampling variability of the test statistic. Increasing sample size is the primary method for reducing the sampling variation, but who is sampled can also play an important role in reducing the variation. Finally, the statistical significance threshold, i.e. the critical value for the rejection region, is an important decision in both controlling power and the type I error. Typically, greater concern is applied to controlling the type I error of the experiment. The more stringent the applied significance threshold the greater the reduction in power will be.

2.6 Extremes of phenotype for selective genotyping

One mechanism that has been used to improve power when the extent of genotyping is constrained by cost is the utilization of an extremes-of-phenotype design (Darvasi and Soller, 1992). For a fixed sample size, it is often more powerful to genotype subjects at the opposite extreme ends of the phenotype distribution rather than choosing a random sample. Figure 2.1 demonstrates how we can improve power using extreme phenotype. Assuming that the phenotype in a study with two groups A and B, each normally distributed but with different means, shown in the upper part of Figure 2.1, we can fit a regression line using all the phenotype data from a random sample shown in the left hand of the lower part of Figure 2.1 and fit a regression line using a subset of subjects with extreme phenotypes from the same sample shown in the right hand of the lower part of Figure 2.1. The slope of the regression line using an extreme phenotype sample is often steeper than when using the entire sample. Therefore, we create a biased sample that tends to inflate effect size estimates under the alternative model (no

difference under null model) and hence typically increases statistical power over similar sized random samples.

2.7 Big data era

Genome-wide genotype data have been generated on hundreds of thousands of subjects. GWAS studies are systematically being performed on existing well-phenotyped biomarkers in large cohort studies (McManus, 2009). 10'000s samples from GWAS have been deposited in dbGAP --- a database (www.ncbi.nlm.nih.gov/gap) of genotypes and phenotypes--- developed to store and distribute genotype and phenotype for association studies and methodology development from interested outside researchers (Mailman et al, 2007). Because many large cohorts have already been genotyped through GWAS, cost constraints due to genotyping will be less of a concern in the near future.

2.8 New biomarkers

Novel biomarkers are being developed for cardiovascular disease and other disease endpoints (Vasan, 2006). The best biomarkers are ones that predict disease such as lipids, inflammatory markers (e.g. CRP and IL6), glucose/insulin levels, etc. The rationale for using biomarkers in association studies is: 1) Biomarkers may be more proximal to genetic factors that play an important role in disease endpoints. 2) Disease endpoints can have very heterogeneous etiology and it can be difficult to obtain sufficient numbers of cases to have good power. Biomarkers can substitute for clinical endpoints and be used in prognosis of a disease or in predicting the state of a disease or clinical benefit (Bhatt et al., 2010). 3) Discovery of genes or DNA sequence variation for biomarkers can become good targets for drug development or be

used in in personal medicine to provide health care tailed to the individual patient (Kumar and Sarin, 2009).

2.9 Cost constraints of phenotyping

Though many biomarkers are pre-existing in large cohort studies or community based studies, not all studies are phenotyped equally. Moreover, it is often difficult to obtain sufficient funding to cover the cost of measuring new biomarkers. If there exist *a priori* set of genetic markers of interest that have already been genotyped, we can select a subset of subjects to phenotype that maximizes power. Such a scenario would frequently occur for replication studies, where the phenotype of interest has yet to be measured in the replication sample. Another possible scenario would be a multistage genetic association study for a new, unmeasured, biomarker where a subset of random subjects are phenotyped in stage 1 and a targeted set of subjects, based on their genotype data, are phenotyped in stage 2 for variants showing the strongest evidence of association in stage 1.

2.10 Extremes of Genotype for selective phenotyping

One mechanism to increase power when phenotyping is constrained by cost is to use an extreme of genotype sampling strategy. The optimal strategy depends on the underlying genetic mechanism, but for an additive model this would likely mean preferentially excluding heterozygotes in favor of the less common homozygotes. Figure 2.2 illustrates the rationale how power can be improved using genotype extremes for a marker that is associated with the phenotype of interest. When we fit linear regression models, we often construct Wald or *t* test statistics to test our hypotheses, which include the beta estimate in the numerator and the

standard error of the beta estimate in the denominator. For a fixed sample size, selecting extremes of genotype should have little effect, on average, on the numerator (beyond stochastic variation) but would typically decrease the variation of the beta estimate (smaller denominator). Hence, using genotype extremes will increase our power to reject the null hypothesis in favor of the alternative, when true (Sen et al., 2009). For one marker contributing to a phenotype, it is easy to optimize the statistical power by simply choosing an equal number of the two homozygotes. However, for multiple markers contributing to a phenotype, it is a considerably more difficult problem, as there are many tradeoffs when selecting subjects. For example, a subject might have the more desirable less common homozygous genotype for one marker but have the less desirable heterozygous genotype at several of the other markers of interest. Selective phenotyping is about how to select subjects that optimizes power over a distribution of markers that are each putatively associated with a phenotype of interest.

2.11 Previous approaches in selective phenotyping

Existing methods have mainly been described for quantitative trait mapping (QTL), mostly in mouse crosses, where the goal is to optimize ability to position a putative causal locus in a region of established linkage. One such method is to maximize the recombination events in selected subjects. A subject with maximum recombinant across all the markers is selected first, and so on until the number of subjects is equal to the sample size (Jannink, 2005, Xu et al., 2005). The other method is SPARE (Selective Phenotyping Approach by Reduction Entropy) by Gagneur et al (2011), a Bayesian mapping approach that is designed to simultaneously optimize detection and localization of a QTL using pedigree data. In this method, a genotype-phenotype relation model is specified, and the cost function is defined as expected Shannon entropy

measuring the uncertainty about a single putative causative QTL. Selection strategy includes alternates of a selection step and a phenotyping step until the desired number of subjects is selected.

More relevant approaches to our question of interest, finding associated variants in large-scale genetic association studies of unrelated subjects, focus on balancing the distribution of alleles among subjects across the selected set of variants. In this category, scores like measure of similarity are used to find the subset of subjects with genotypes as dissimilar as possible (Sen et al., 2009, Jin et al., 2004). The simplest strategy is to, iteratively, select subjects to phenotype who have the largest number of homozygous genotypes across all variants (starting with the person with the most homozygous genotypes), and repeat until the predefined sample size is reached (Sen et al., 2009). This approach is not ideal because it treats all markers equivalently, regardless of the underlying alternative models for each, and doesn't factor in the uneven allele frequencies across the markers. For example, an experimenter might want to maximize power, simultaneously, for three markers. Assume two of the markers have common allele frequencies and would have excellent power even if a random sample were selected. Assume the third marker has considerably lower minor allele frequency and hence it would be critical to carefully select the genotypic extremes. Assigning equal weight to all markers when selecting subjects to phenotypes would be detrimental to the overall power of the experiment as some subjects with the less common homozygous genotype at the third marker would be passed over in favor of the subjects homozygous at both of the common markers. This is a simple example describing the largest limitation of existing methods. Namely, existing methods do not directly address the power functions of the individual variants when performing sample selection. The ideal method

should be able to directly assess the impact of sample selection on power for all markers simultaneously.

Figure 2.1 Mechanism to increase power using extreme phenotype. Select a subset of samples with extreme phenotype measures from the entire sample. Note, removing subjects from the middle of the phenotype distribution results in steeper regression lines than observed for the complete sample.

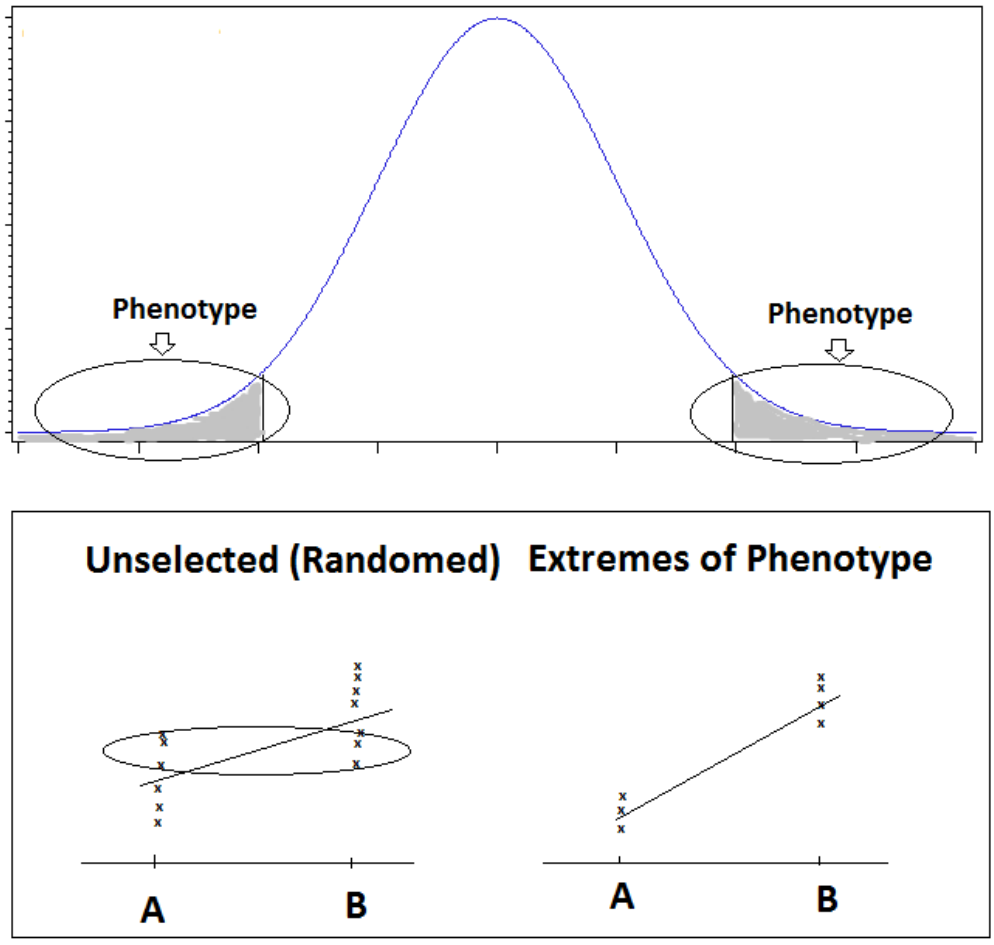
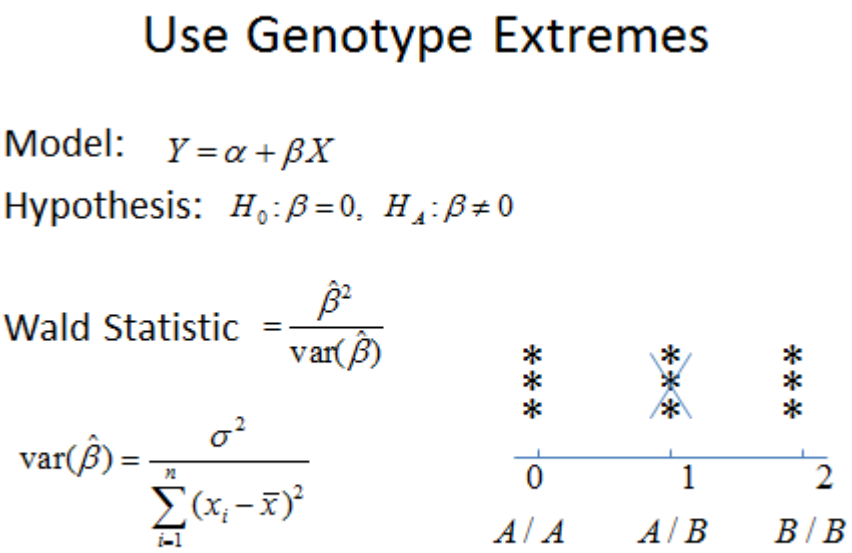


Figure 2.2 Mechanism to increase power using extreme phenotype to use genotype extremes



CHAPTER 3

THE RAPID ENROLLMENT DESIGN FOR PHASE I CLINICIAL TRIALS

3.1 Introduction

Dose-finding trials in oncology are conducted to learn about the dose-toxicity relationship of a drug and to estimate the maximum tolerated dose (MTD). There is a long history of oncology dose-finding methods for estimating a dose with a certain mean response when outcome is binary (Wetherill, 1963; O'Quigley et al., 1990; Babb et al., 1998). Often it takes a long time to observe toxicity compared to accrual rate and a number of methods have been developed for dose-finding with delayed outcome (Cheung and Chappell, 2000); (Ivanova et al., 2005; Bekele et al., 2008). In some trials, when pace of disease is so rapid that a patient may succumb to the disease while waiting, it is desirable to enroll a patient into the trial and start treating as soon as possible. None of the existing dose-finding methods developed for delayed toxicity is suitable for a trial with rapid enrollment where a patient should be assigned to a dose as soon as they are enrolled. The method of Bekele et al. (2008) solves the problem of fast accrual by prescribing when a trial should be halted as toxicity rate at the current dose becomes uncertain. In TITE-CRM (Cheung and Chappell, 2000) the accrual is not halted and a new patient is assigned to the dose closest to the MTD given current data including partial data from patients still in follow-up. As noted in Cheung and Chappell (2000) rapid accrual, especially in

the beginning of the trial, may lead to assigning patients to doses with high toxicity rate when TITE-CRM is used.

In many dose-finding designs (Wetherill, 1963; Yuan and Chappell, 2004; Ivanova et al., 2007; Ivanova and Kim, 2009; Ji et al., 2010) the next assignment is determined based on the data at the current dose. We propose assigning patients at lower doses while toxicity profile at higher doses is uncertain. When assignments are being made to lower doses as well as the higher dose, it is no longer clear what dose is the current dose and therefore none of the above cited rules can be used for assignment. In this paper we describe an intuitive dose-finding method where the dose-assignment strategy is not tied to a current dose and the next assignment is made to the dose that is most likely to be the MTD. We refer to the new method as the Rapid Enrollment Design (RED). We describe the dose-assignment design in Section 3.2. In Section 3.3 we propose how to mitigate the uncertainty from patients still in follow-up when a new assignment is made. We give an example in Section 3.4. We compare designs in Section 3.5 and discuss the findings in Section 3.6.

3.2 Dose-finding method

Let $D = \{d_1,...,d_K\}$ denote the set of ordered dose levels selected for a trial. Let T be the length of follow-up for toxicity. A subject's response at d_k has Bernoulli distribution with parameter q_k, where $(q_1,\ldots,q_K)$ is the vector of true toxicity rates at the K doses. We assume that $q_1,\ldots,q_K$ are non-decreasing, $q_1 \le ... \le q_K$. Observations from different subjects are independent. Only one observation per subject is taken. The goal is to find dose $d_m \in D$ such that $q_m = \Gamma$. If there is no such dose, the goal is to find the dose d_m with mean response closest to Γ. We refer to Γ as the target DLT rate and d_m as the target dose.

First we consider the case where T is short compared to accrual rate and therefore toxicity responses from all patients are known when a new patient is enrolled. In the proposed design, each new patient is assigned to the dose with the highest probability of being the target dose given data available so far. Let $\boldsymbol{n} = (n_1,...,n_K)$ be the number of subjects assigned to each dose and $\boldsymbol{m} = (m_1,...,m_K)$ be the number of DLTs observed at each dose. Then the DLT rates at each dose can be estimated by simple proportions $\hat{q}_j = m_j / n_j$, j = 1, …, K. Since $\hat{q}_1,...,\hat{q}_K$ might not be monotone we will compute isotonic estimates $(\hat{q}_1^*,...,\hat{q}_K^*)$ of DLT rates using the pool adjacent violator algorithm (PAVA) (Barlow et al., 1972). PAVA is described as follows. First we set $\hat{q}_j^* = \hat{q}_j$, j =1, …, K. If $\hat{q}_1 \le ... \le \hat{q}_K$, then $\hat{q}_1^* \le ... \le \hat{q}_K^*$ and nothing else is needed. Otherwise, find the smallest k such that $m_k / n_k > m_{k+1} / n_{k+1}$ and replace these two estimates with $\hat{q}_k^* = \hat{q}_{k+1}^* = (m_{k-1} + m_k) / (n_{k-1} + n_k)$. We also denote the estimate based on pooling data from doses d_k and d_{k+1} by $\hat{q}_{k,k+1}$ with "dose" $d_{k,k+1}$ and weight $n_{k,k+1} = n_k + n_{k+1}$. Look again for a dose for where the working isotonic estimate is greater than the working estimate at the next higher "dose", and repeat the averaging and concatenation process until isotonicity is actually obtained, that is, until $\hat{q}_1^* \le \cdots \le \hat{q}_K^*$. We note that the isotonic estimates are equivalent to the maximum likelihood estimates assuming $q_1 \le \cdots \le q_K$. In the dose-assignment design described in the next paragraph when estimates $(\hat{q}_k^*,...,\hat{q}_{k+m}^*)$, $\hat{q}_k^* = ... = \hat{q}_{k+m}^*$, are computed based on pooled data $\hat{q}_k^* = ... = \hat{q}_{k+j}^* = (m_k + ... + m_{k+j}) / (n_k + ... + n_{k+j})$, we use the highest dose on the estimated plateau, d_{k+j}, to represent the pooled doses if $\hat{q}_k^* \le \Gamma$ and the lowest dose on the estimated plateau, d_k, if $\hat{q}_k^* > \Gamma$. To facilitate the Bayesian decision rule described below we assign the

average number of DLTs $(m_k + ... + m_{k+j}) / (j+1)$ and the average sample size $(n_k + ... + n_{k+j}) / (j+1)$ of the pooled doses to that dose.

To compute the probability of each dose to be the MTD we use Bayesian computations and assume Beta(α,β) prior on q_j, j = 1, 2,..., K, the posterior distribution of q_j conditional on outcome data is

$$q_j \mid \boldsymbol{m}, \boldsymbol{n} \sim Beta(\alpha + m_j, \beta + n_j - m_j), j = 1, 2, \ldots, K.$$

The prior Beta(α,β) reflects the belief that there exists data from $\alpha + \beta$ patients, α patients with DLTs and β patients without a DLT. We use $\alpha = 0.3$ and $\beta = 0.01$. This prior slows the initial escalation because it reflects the prior DLT rate of $\alpha / (\alpha + \beta) = 0.3 / 0.31 = 0.97$, however it gets overruled by data very soon since the effective sample size in the prior is only $\alpha + \beta = 0.31$ of a patient.

Let $\{d_1, ..., d_k\}$ be the set of doses with at least one patient assigned. The design is based on the following rules

1)Initial escalation: do not escalate a dose unless at least s patients are assigned to current highest dose, k. The number s depends on the target rate Γ, s = 3 is often used if Γ = 0.25.

2)If $\hat{q}_k^* < \Gamma$, the next patient is assigned to d_{k+1} (or d_K if $k = K$).

3)If $\hat{q}_k^* \geq \Gamma$, if there is a dose d_j such that $\hat{q}_j^* = \Gamma$, the next patient is assigned to d_j. Otherwise, let j, $j \leq k-1$, be such that $\hat{q}_j^* < \Gamma$ and $\hat{q}_{j+1}^* > \Gamma$. Let $\pi_j = \Pr\{\Gamma - \varepsilon < q_j < \Gamma - \varepsilon\}$ and $\pi_{j+1} = \Pr\{\Gamma - \varepsilon < q_{j+1} < \Gamma - \varepsilon\}$. The next patient is assigned to the dose corresponding to the higher, π_j or π_{j+1}.

4)Do not assign patients to a dose where $\Pr\{q_j > \Gamma\} > 0.95$, j = 1, 2,..., k. If $\Pr\{q_1 > \Gamma\} > 0.95$, the trial is stopped because the lowest dose is too toxic.

5)At the end of the trial the dose that would have been recommended for the next patient is selected as the estimated MTD. No dose is selected if the lowest dose is deemed too toxic by the stopping rule.

In the rules above, ε, $\varepsilon \le \min(\Gamma, 1-\Gamma)$, is a design parameter. We recommend $\varepsilon = 0.05$. Robustness of this choice is discussed in Section 3.5.

Decision rules for two candidate doses, one with estimated DLT rate lower than Γ and one with rate higher than Γ, are shown in Table 3.1 (for $\Gamma = 0.2$), Table 3.2 (for $\Gamma = 0.25$) and Table 3.3 (for $\Gamma = 0.3$). The first column of each of these tables contains data yielding DLT proportion less than Γ, and the second column DLT proportion higher than Γ. The decision rule for each pair of the first and second column data is in column 3. For example, if $\Gamma = 0.2$, 3 patients were enrolled at dose 1 with no DLTs, and 6 patients were enrolled at dose 2 with 2 DLTs. The decision rule for these data is to assign the higher dose (line 9, Table 3.1). This is because the data at the two doses are (0/3, 2/6) yielding $\pi_1 = 0.090$ and $\pi_2 = 0.181$.

To implement the RED we developed web-based software available at http://www.unc.edu/wang484/red/red.php. The input is the number of DLTs at each dose, $\boldsymbol{m} = (m_1,...,m_k)$, and the number of subjects at each dose $\boldsymbol{n} = (n_1,...,n_K)$. The program identifies two candidate doses j and j + 1 based on isotonic estimates, and computes π_j, π_{j+1}. It also computes the probabilities that the DLT rates at the two doses are higher than Γ needed for the safety rule.

3.3 Mitigating uncertainty from patients still in follow-up

In many Phase I oncology trials follow-up time is long compared to accrual rate. Often 3 patients are assigned to a previously untried dose. These 3 patients need to complete follow-up at the dose before more patients can be assigned to that or higher dose level. This however, does not fully resolve uncertainty about safety of future assignments. For example, if one out of the three patients had a DLT at that dose, can we assign, say, 6 more patients to that dose at once or is it too risky? We propose a simple way to mitigate this risk. A patient with a DLT who has been in the follow-up for time u and therefore has completed a fraction u/T of the total follow-up with $1-u/T$ of the follow-up still remaining is counted as a patient with $1-u/T$ of a DLT. The total DLT count at a dose is the number of actual DLTs, m, plus the sum of $1-u_i/T$, where the sum is over all the patients assigned at that dose that are still in follow-up. The denominator is the total number of patients at that dose irrespective of their follow-up time. These data are used to determine the next assignment according to the rules in Section 3.2.

For example, 3 patients have completed follow up at dose 1 with one DLT and several patients are available to enroll. If $\Gamma = 0.20$, with 1 DTL out of 3, $\Pr\{q_1 > \Gamma\} = 0.75$, therefore since this probability is less than 0.95 at least one more patient can be enrolled. If one patient is enrolled, to mitigate potential DLT outcome from this patient, the data are augmented with 1 DLT at d_1 yielding 2/4 DLT at d_1 and $\Pr\{q_1 > \Gamma\} = 0.93$. Since this probability is less than 0.95 we can enroll one more patient. After augmenting 1/3 with 2/2, for two newly enrolled patients with $u = 0$ follow-up each, the DTL data at d_1 are 3/5 yielding $\Pr\{q_1 > \Gamma\} = 0.98$, therefore we cannot assign more patients to d_1 at this point. After the two newly enrolled patients have been followed for, say, half the total follow-up time without a DLT, the data at d_1 are 1/3 + 0.5/1 +

0.5/1 = 2/5 yielding $\Pr\{q_1 > \Gamma\} = 0.87$ and more patients can be enrolled. Patients are assigned to d_1 until the estimated DLT rate at d_1 becomes lower than $\Gamma = 0.20$, at which point patients are assigned to d_2.

In another example, there are 3 patients enrolled at d_1 with no DLTs, and 3 patients enrolled at d_2 with 1 DLT observed in these 3 patients. The data at the two doses are (0/3, 1/3), corresponding to $\pi_1 = 0.09$ and $\pi_2 = 0.15$, therefore the next patient, patient seven, is assigned to d_2. If patient eight is available at the time when patient seven is assigned, the augmented data are (0/3, 2/4), corresponding to $\pi_1 = 0.09$ and $\pi_2 = 0.08$, therefore patient eight should be assigned to dose 1. If, instead, patient eight arrives when patient seven has completed half of his follow-up without a DLT, the augmented data are (0/3, 1.5/4) yielding $\pi_1 = 0.09$ and $\pi_2 = 0.14$, therefore patient 8 is assigned to d_2.

3.4 Example

Dose-finding trials in acute leukemia typically require long follow-up for toxicity. This is because it is difficult to distinguish undue hematologic drug toxicity from the bone marrow effects of the disease itself. Often this requires follow-up for toxicity from an individual cycle of therapy that lasts 4-6 weeks rather than what is typical in solid tumors (3-4 weeks). In addition it is desirable to offer continuous enrollment in acute leukemia trials, since the pace of these leukemias is so rapid that the patient may succumb to the disease while waiting. The proposed strategy was implemented in a dose-finding study of a new derivative of thalidomide for older adults with acute myeloid leukemia. Since the trial is ongoing, we used data from a recently completed Phase I trial (Foster et al., 2012) to illustrate the method. The trial investigated clofarabine in combination with gemtuzumab ozogamicin in relapsed or refractory acute myeloid

leukemia patients. The DLT was defined as grade 3 or greater treatment-related toxicity lasting greater than 2 weeks or delay in hematologic recovery beyond 35 days from initiation of induction and not related to persistent or recurrent leukemia. Therefore the maximum observation period for toxicity was 5 weeks from the start of therapy, and the goal of the trial was to find a dose with the DLT rate of 0.26. The trial used the time-to-event CCD method from Ivanova et al. (2007) with an ad-hoc modification that allowed rapid enrollment with immediate assignment. We illustrate how the proposed dose-assignment algorithm would have worked if used in the gemtuzumab trial. We use patient enrollment times, their DLT outcomes and the time when a DLT has occurred. There were three patients who progressed or died between days 32 and 35 without a DLT, and therefore these patients were permanently censored for DLT before $T = 35$. In this illustration these patients are counted as patients with full follow-up of 35 days and no DLT. Dose assignments for the first 18 patients are presented in Table 3.4. When patient 19 was enrolled, all 13 patients assigned to dose 1 have completed their follow-up and 5 DLTs were observed. The posterior probability that the DLT rate at d_1 exceeds 0.26 is 0.85, and therefore according to our method the next assignment should be to d_1. In the actual acute myeloid leukemia trial, after seeing these data, the investigators decided to be conservative and to enroll patients 19 and 20 at the lower dose, dose -1.

3.5 Comparisons with other dose-finding methods

3.5.1 Comparison with mTPI and the *t*-statistic designs in trials with a short follow-up time

First we considered the case where DLT outcome is known right away. We compared the RED with the modified toxicity probability interval (mTPI) method (Ji et al., 2010), and the *t*-statistics design (Ivanova and Kim, 2009).

The mTPI is based on computing Bayesian posterior probabilities of the DLT rate being in certain intervals. Let d_j be the current dose, that is, the dose the last patient was assigned to. Calculate three probabilities $E = \Pr\{0 < q_j < \Gamma - \varepsilon_1\} / (\Gamma - \varepsilon_1)$, $S = \Pr\{\Gamma - \varepsilon_1 < q_j < \Gamma + \varepsilon_2\} / (\varepsilon_2 + \varepsilon_1)$ and $D = \Pr\{\Gamma + \varepsilon_2 < q_j < 1\} / (1 - \Gamma - \varepsilon_2)$. The next patient is assigned to d_{j+1} if E is the largest, to d_j if S is the largest, and to d_{j-1} if D is the largest. If $\Pr\{q_j > \Gamma\} > 0.95$, patients are assigned to lower doses. If $\Pr\{q_1 > \Gamma\} > 0.95$ the trial is stopped. The estimated MTD is the dose with the estimated DLT rate closest to Γ.

The *t*-statistics method is a dose-finding design in which the *t*-statistic, *T*, to test the hypothesis that the mean at the current dose is equal to the target is computed at each step. The next patient is assigned to the current dose if $-\Delta < T < \Delta$, otherwise the dose is reduced or increased depending on the sign of *T*. Here Δ is a design parameter and Δ = 1 is recommended. In the *t*-statistic design to estimate the target dose after the trial for the new design, we first obtained the isotonic estimates of DLT rates. The dose with the estimated DLT rate closest to Γ is the estimated MTD. If there were two or more such doses, the highest dose with the estimated DLT rate below Γ is chosen. If all the estimated rates at these doses were higher than Γ, the lowest of these doses is chosen.

We used all 10 scenarios from Ji et al. (2010) to compare designs. Results for mTPI for scenarios 1, 5, 7-10 were reproduced from Ji et al. (2010) and results for scenarios 2-4 and 6 were simulated using the program provided by Yaun Ji. Simulations were performed in R and comparison is made based on 4000 simulation runs for each scenario. The target DLT rate was $\Gamma = 0.25$. In all designs patients are assigned in cohorts of 3 and response is observed immediately. Note that all three designs can be used with any number of patients per cohort, assignment in cohorts of 3 was chosen to match the simulations in Ji et al. (2010). In the RED we

set $\alpha = 0.3$, $\beta = 0.01$ and $\varepsilon = 0.05$. Larger value of α is chosen to slow down escalation when there is not much information available at the dose. In mTPI design the prior parameters are $\alpha = 1$, $\beta = 1$ as in Ji et al. (2010). Since Ji et al. (2010), used α and β different from ours we calibrated our safety rule so that the two designs make the same safety decision for the same data. The re-calibrated safety rule for our design is that patients are not assigned to a dose if $\Pr\{q_j > \Gamma\} > 0.96$.

Overall all three designs perform very similarly (Table 3.5). In scenario 7 the *t*-statistics design recommends the target dose (dose with the DLT rate of 0.05) in 0.46 of the trials compared to 0.82 and 0.83 for the other two designs. This is because a safety rule was implemented in RED and mTPI design but not in the *t*-statistic design. A frequentist safety rule for the *t*-statistic design can be constructed similarly to the Bayesian safety rule described in Section 3.3.

We investigated the robustness of the new design with respect to parameter ε. In the main simulation study (Table 3.5) we used ε = 0.05. We performed simulations with values of ε in the range (0.01, 0.1) and obtained very similar results (results are available from the authors).

3.5.2 Comparison with TITE-CRM when the follow-up for DLT is long

Cheung and Chappell (2000) proposed a time-to-event modification of the continual reassessment method (CRM), called TITE-CRM, for dose-finding trials where follow-up for DLT is long. Let x_i be the dose level received by subject i, $x_i \in D$, and $y_i = 1$ if the ith subject had a DLT and 0 otherwise. In the original CRM (O'Quigley et al., 1990), the calculation of posterior mean of θ at the time when (n+1)th subject enters the trial is based on the likelihood:

$$L_n(\theta) = \prod_{i=1}^{n} F(x_i, \theta)^{y_i} \left\{1 - F(x_i, \theta)\right\}^{1-y_i},$$

where $F(x_i, \theta) = b_{x_i}^{\theta}$ and $(b_1, ..., b_K)$ is a set of positive constants. In clinical trials that require long follow-up times, the toxicity rate at dose x_i is defined as the probability of observing toxicity at x_i during a time period of length *T* after initiation of therapy. Data for the *i*th subject, $i = 1, \ldots, n$, when (*n*+1)th subject is assigned to a treatment, consists of dose x_i, toxicity indicator y_i and the time u_i that has elapsed from the time of the *i*th subject's treatment assignment to the time of the (*n*+1)th subject's treatment assignment.

For TITE-CRM, Cheung and Chappell (2000) suggested the weighted likelihood

$$\tilde{L}_n(\theta) = \prod_{i=1}^{n} \left\{w_i F(x_i, \theta)\right\}^{y_i} \left\{1 - w_i F(x_i, \theta)\right\}^{1-y_i},$$

where w_i is the weight assigned to the *i*th observation prior to the entry of the (*n*+1)th subject. For example, setting $w_i = \min(u_i / T, 1)$ reflects an assumption that the density of time to toxicity is flat in (0,*T*).

We compared the performance of the TITE-CRM and the RED via simulations. In the RED the uncertainty from patients still in follow-up was mitigated as described in Section 3. Note that the approaches of handling patients still in follow up in the TITE-CRM and in the RED are very different. In the TITE-CRM, a patient who has been followed for a fraction *w* of full follow-up time *T* and have not had a DLT yet is contributing roughly as a fractional patient (*w* of a patient) without a DLT. On the other hand, in our approach such a patient is contributing as 1 patient with (1 – *w*) DLTs. As a result, if many patients are enrolled at once, the TITE-CRM escalates rapidly to higher doses and our method is conservative and escalates slowly.

Table 6 displays comparison of the two designs in a trial with 35-day follow up for DLT. Patients are assigned one at a time and enrolled at the rate of one patient per week. For patients with DLT, time to DLT was generated as uniform random variable on (0,35). We excluded scenario 8 since there is no MTD in this scenario. Due to aggressive escalation the TITE-CRM recommends the doses above the true MTD much more frequently than the RED. At the same time the recommendation of the correct dose is similar between the two designs. The RED recommends the correct MTD more often in scenarios where the MTD is among lower doses and the TITE-CRM recommends the correct MTD more often when the MTD is among higher doses. This is because the TITE-CRM escalates rapidly and the RED is conservative. Rapid escalation of the TITE-CRM to higher doses leads to observing 3 more DLTs on average in each trial compared to the RED.

The average length of a trial is 51 weeks. For comparison a trial that enrolls patients in cohorts of 3 at a time where each cohort is followed for 5 weeks before the next cohort is enrolled will have the length of approximately 70 weeks. The results of the latter are the same as those presented in Table 5. Comparison of Table 5 and 6 shows that the RED can provide good estimation of the MTD in trials with long follow-up for toxicity without exposing patients to doses with possibly high DLT rates.

3.6 Discussion

We proposed a new method to assign patients to doses in a Phase I oncology trial and a method to mitigate uncertainty from patients still in follow-up. Ji and Wang (2013) recently compared the mTPI and the 3+3 designs (Storer, 1989), to show that mTPI can identify the correct MTD better and also assigns fewer patients to the toxic doses above MTD. Our new

method performs similar to mTPI but also has flexibility to accommodate rapid enrollment in trials with long follow-up. We give tables with decision rules to implement the new design. A web-based program is available to identify a dose for the next assignment given trial data.

Table 3.1 Dose allocation decision based on the posterior probability at two candidate doses d_j and d_{j+1}. Target probability is $\Gamma = 0.20$. The observed toxicity rate at d_j is less than Γ and the observed toxicity rate at d_{j+1} is higher than Γ. Data at each dose is the number of DLTs over the number of patients assigned to the dose.

Observed data at d_j	Observed data at d_{j+1}	Decision
1/6:12, 2/11:12	1/3, 2/5, 3/8, 4/10, 4/11	Lower Dose
1/6:11, 2/11:12	2/6, 3/9, 4/12	Lower Dose
2/11:12	2/7, 3/10,1/4, 2/8:9, 3/11:12	Lower Dose
1/6:7	3/10	Lower Dose
1/6:8	2/7	Lower Dose
1/6:10	1/4	Lower Dose
0/1:12	1/3, 2/5, 3/8, 4/10:11	Higher Dose
0/1:12, 1/12	2/6, 3/9, 4/12	Higher Dose
0/1:12, 1/6:12	2/8:9, 3/11:12	Higher Dose
0/1:12, 1/11:12	1/4	Higher Dose
0/1:12, 1/9:12	2/7	Higher Dose
0/1:12, 1/8:12	3/10	Higher Dose

Table 3.2 Dose allocation decision based on the posterior probability at two candidate doses d_j and d_{j+1}. Target probability is $\Gamma = 0.25$. The observed toxicity rate at d_j is less than Γ and the observed toxicity rate at d_{j+1} is higher than Γ. Data at each dose is the number of DLTs over the number of patients assigned to the dose.

Observed data at d_j	Observed data at d_{j+1}	Decision
1/5:12, 2/9:12	1/2, 2/4, 3/6	Lower Dose
1/5:9, 2/9:12	1/3, 2/5, 3/7	Lower Dose
2/9:12	2/6, 2/7, 3/9, 4/11	Lower Dose
2/9:11	4/12	Lower Dose
2/9:11	4/12	Lower Dose
1/5:6, 2/9:12	3/8	Lower Dose
1/5:8, 2/9:12	4/10	Lower Dose
1/5:9, 2/9:12	5/12	Lower Dose
1/5:10, 2/9:12	4/9	Lower Dose
1/5:11, 2/9:12	5/11	Lower Dose
0/1:2, 1/5:12, 2/9:12	4/8	Lower Dose
0/1:3, 1/5:12, 2/9:12	5/10	Lower Dose
0/1:4, 1/5:12, 2/9:12	6/12	Lower Dose
0/1:6, 1/5:12, 2/9:12	6/11	Lower Dose
0/1:12	1/2, 2/4, 3/6	Higher Dose
0/1:12, 1/10:12	1/3, 2/5, 3/7, 5/12	Higher Dose
0/1:12, 1/5:12	2/6:7, 3/9, 4/11	Higher Dose
0/1:12, 1/5:12, 2/9:12	3/10:11	Higher Dose
0/7:12	6/11	Higher Dose
0/5:12	6/12	Higher Dose
0/4:12	5/10	Higher Dose
0/3:12	4/8	Higher Dose
0/1:12,1/12	5/11	Higher Dose
0/1:12,1/11:12	9/4	Higher Dose
0/1:12, 1/9:12	4/10	Higher Dose
0/1:12, 1/7:12	3/8	Higher Dose
0/1:12, 1/5:12, 2/12	4/12	Higher Dose

Table 3.3 Dose allocation decision based on the posterior probability at two candidate doses d_j and d_{j+1}. Target probability is $\Gamma = 0.30$. The observed toxicity rate at d_j is less than Γ and the observed toxicity rate at d_{j+1} is higher than Γ. Data at each dose is the number of DLTs over the number of patients assigned to the dose.

Observed data at d_j	Observed data at d_{j+1}	Decision
1/4:9, 2/7:12, 3/11:12	1/2, 6/12	Lower Dose
1/4:6, 2/7:12, 3/11:12	1/3, 5/11	Lower Dose
1/4:8, 2/7:12, 3/11:12	2/4, 3/6, 4/8, 5/10	Lower Dose
1/4:5, 2/7:12, 3/11:12	2/5, 4/9	Lower Dose
2/7:10, 3/11:12	2/6, 5/12	Lower Dose
3/11:12	3/9, 4/11	Lower Dose
1/5, 2/7:11, 3/11:12	3/7	Lower Dose
2/7:9, 3/11:12	3/8, 4/10	Lower Dose
0/1, 1/4:11, 2/7:12, 3/11:12	6/11	Lower Dose
0/1:12, 1/10:12	1/2, 6/12	Higher Dose
0/1:12, 1/7:12	1/3, 5/11	Higher Dose
0/1:12, 1/9:12	2/4, 3/6, 4/8, 5/10	Higher Dose
0/1:12, 1/6:12	2/5, 4/9	Higher Dose
0/1:12, 1/4:12, 2/11:12	2/6, 5/12	Higher Dose
0/1:12, 1/4:12, 2/7:12	3/9, 4/11	Higher Dose
0/1:12, 1/4:12, 2/12	3/7	Higher Dose
0/1:12, 1/4:12, 2/10:12	3/8, 4/10	Higher Dose
0/1:12, 1/4:12, 2/7:12, 3/11:12	4/12	Higher Dose
0/2:12, 1/12	6/11	Higher Dose

Table 3.4 Dose assignments for the first 18 patients in the gemtuzumab trial. The target DLT rate is 0.26. The length of follow-up is 35 days. DLTs were observed on patients 4, 9, 11, 12, 14, 16 and 17.

Pt	Day of enrollment	Data from patients with full follow-up at the time of enrollment		Additional temporary DLTs from patients still in follow-up at the time of enrollment		Estimated DLT rate		Posterior probability that DLT rate is in (0.21, 0.31)		Dose Assignment
		d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	
1	1	---	---	---	---	---	---	---	---	1
2	77	0/1	---	0/0	---	0.00	---	0.078	---	1
3	77	0/1	---	0/0	---	0.00	---	0.087	---	1
4	172	0/3	---	0/0	---	0.00	---	0.064	---	2
5	194	0/3	0/0	0/0	0.37/1	0.00	0.37	0.064	0.082	2
6	327	0/3	1/2	0/0	0/0	0.00	0.50	0.064	0.087	2
7	327	0/3	1/2	0/0	1/1	0.00	0.67	0.064	0.041	1
8	348	0/3	1/2	0.4/1	0.4/1	0.10	0.47	0.127	0.108	1
9	369	0/4	1/3	0.4/1	0/0	0.08	0.33	0.113	0.148	2
10	437	0/5	2/4	0/0	0/0	0.00	0.50	0.042	0.098	2
11	448	0/5	2/4	0/0	0.69/1	0.00	0.54	0.042	0.077	2
12	508	0/5	3/6	0/0	0/0	0.00	0.50	0.042	0.093	2
13	516	0/5	3/6	0/0	0.77/1	0.00	0.54	0.042	0.065*	1
14	565	0/6	4/7	0/0	0/0	0.00	0.57	0.033	0.047*	1
15	636	1/7	4/7	0/0	0/0	0.14	0.57	0.176	0.047*	1
16	671	1/8	4/7	0/0	0/0	0.13	0.57	0.159	0.047*	1
17	676	1/8	4/7	0.86/1	0/0	0.23	0.57	0.250	0.047*	1
18	801	3/10	4/7	0/0	0/0	0.30	0.57	0.270	0.047*	1
19	815	3/10	4/7	0.60/1	0/0	0.33	0.57	0.260	0.047*	1
20	850	3/12	4/7	0/0	0/0	0.25	0.57	0.307	0.047*	1

*The probability that the DLT rate exceeds 0.26 is higher than 0.95

Table 3.5 Comparison of the mTPI, the proposed RED method and the t-statistics design. The target DLT rate is 0.25 and the total sample size is n = 30. Proportion of times each dose is recommended as the target dose and the average number of subjects allocated to each dose. Numbers at the target dose are in bold.

		d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
Scenario 1		0.05	**0.25**	0.50	0.60	0.70	0.80	0.90	0.95
Recommendation	mTPI	0.14	**0.78**	0.08	0.00	0.00	0.00	0.00	0.00
	RED	0.14	**0.77**	0.09	0.00	0.00	0.00	0.00	0.00
	t-statistics	0.14	**0.78**	0.08	0.00	0.00	0.00	0.00	0.00
Allocation	mTPI	7.1	**18.3**	4.4	0.2	0.0	0.0	0.0	0.0
	RED	6.8	**17.5**	5.1	0.6	0.0	0.0	0.0	0.0
	t-statistics	7.5	**17.6**	4.5	0.3	0.0	0.0	0.0	0.0
Scenario 2		0.01	0.05	**0.25**	0.50	0.60	0.70	0.80	0.90
Recommendation	mTPI	0.00	0.14	**0.74**	0.11	0.01	0.00	0.00	0.00
	RED	0.00	0.15	**0.75**	0.10	0.00	0.00	0.00	0.00
	t-statistics	0.04	0.10	**0.78**	0.09	0.00	0.00	0.00	0.00
Allocation	mTPI	3.3	6.6	**15.7**	4.0	0.4	0.0	0.0	0.0
	RED	3.2	6.7	**15.0**	4.6	0.6	0.0	0.0	0.0
	t-statistics	3.2	7.0	**15.3**	4.1	0.3	0.0	0.0	0.0
Scenario 3		0.01	0.02	0.05	**0.25**	0.50	0.60	0.70	0.80
Recommendation	mTPI	0.00	0.00	0.13	**0.75**	0.11	0.01	0.00	0.00
	RED	0.00	0.00	0.15	**0.73**	0.11	0.01	0.00	0.00
	t-statistics	0.00	0.00	0.14	**0.76**	0.10	0.00	0.00	0.00
Allocation	mTPI	3.2	3.5	6.2	**13.1**	3.8	0.3	0.0	0.0
	RED	3.1	3.4	6.5	**12.5**	4.0	0.5	0.0	0.0
	t-statistics	3.2	3.5	6.3	**13.0**	3.6	0.3	0.0	0.0
Scenario 4		0.01	0.02	0.03	0.05	**0.25**	0.50	0.60	0.70
Recommendation	mTPI	0.00	0.00	0.00	0.14	**0.69**	0.15	0.02	0.00
	RED	0.00	0.00	0.00	0.15	**0.73**	0.10	0.01	0.00
	t-statistics	0.00	0.00	0.00	0.16	**0.71**	0.12	0.01	0.00
Allocation	mTPI	3.2	3.5	3.5	5.6	**10.7**	3.2	0.3	0.0
	RED	3.1	3.3	3.5	6.2	**10.0**	3.4	0.4	0.0
	t-statistics	3.2	3.5	3.7	5.8	**10.5**	3.0	0.2	0.0
Scenario 5		0.01	0.02	0.03	0.04	0.05	**0.25**	0.50	0.60
Recommendation	mTPI	0.00	0.00	0.00	0.02	0.16	**0.71**	0.10	0.01
	RED	0.00	0.00	0.00	0.01	0.24	**0.61**	0.12	0.01
	t-statistics	0.00	0.00	0.00	0.02	0.16	**0.69**	0.11	0.01
Allocation	mTPI	3.2	3.5	3.6	4.0	5.2	**8.1**	2.3	0.1
	RED	3.1	3.3	3.5	3.6	5.6	**7.7**	2.8	0.4
	t-statistics	3.2	3.5	3.7	4.0	5.1	**8.1**	2.2	0.2

Scenario 6		0.01	0.02	0.03	0.04	0.05	0.06	**0.25**	0.50
Recommendation	mTPI	0.00	0.00	0.00	0.02	0.04	0.28	**0.50**	0.15
	RED	0.00	0.00	0.00	0.01	0.04	0.25	**0.55**	0.14
	t-statistics	0.00	0.00	0.00	0.02	0.04	0.29	**0.51**	0.13
Allocation	mTPI	3.2	3.4	3.8	3.8	4.0	4.7	**5.3**	1.8
	RED	3.1	3.3	3.5	3.6	3.7	4.8	**5.6**	2.5
	t-statistics	3.2	3.5	3.7	4.0	4.0	4.6	**5.3**	1.7
Scenario 7		0.01	**0.05**	0.50	0.60	0.70	0.80	0.90	0.95
Recommendation	mTPI	0.00	**0.82**	0.17	0.00	0.00	0.00	0.00	0.00
	RED	0.00	**0.81**	0.19	0.01	0.00	0.00	0.00	0.00
	t-statistics	0.27	**0.46**	0.26	0.00	0.00	0.00	0.00	0.00
Allocation	mTPI	3.2	**15.9**	10.3	0.6	0.0	0.0	0.0	0.0
	RED	3.2	**15.9**	9.9	1.0	0.1	0.0	0.0	0.0
	t-statistics	3.2	**16.0**	10.1	0.7	0.0	0.0	0.0	0.0
Scenario 8		**0.40**	0.50	0.60	0.70	0.80	0.90	0.95	0.99
Recommendation	mTPI	**0.31**	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	RED	**0.34**	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	t-statistics	**0.32**	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Allocation	mTPI	**16.8**	2.0	0.2	0.0	0.0	0.0	0.0	0.0
	RED	**16.3**	2.8	0.3	0.0	0.0	0.0	0.0	0.0
	t-statistics	**16.9**	1.9	0.2	0.0	0.0	0.0	0.0	0.0
Scenario 9		0.15	**0.25**	0.35	0.45	0.55	0.65	0.75	0.85
Recommendation	mTPI	0.29	**0.45**	0.20	0.04	0.00	0.00	0.00	0.00
	RED	0.27	**0.45**	0.21	0.04	0.00	0.00	0.00	0.00
	t-statistics	0.30	**0.45**	0.20	0.04	0.00	0.00	0.00	0.00
Allocation	mTPI	12.4	**10.9**	5.0	1.1	0.1	0.0	0.0	0.0
	RED	10.7	**11.6**	5.6	1.5	0.3	0.0	0.0	0.0
	t-statistics	11.9	**11.7**	4.9	1.1	0.1	0.0	0.0	0.0
Scenario 10		0.05	0.15	**0.25**	0.35	0.45	0.55	0.65	0.75
Recommendation	mTPI	0.02	0.28	**0.42**	0.23	0.04	0.00	0.00	0.00
	RED	0.03	0.29	**0.43**	0.21	0.04	0.01	0.00	0.00
	t-statistics	0.03	0.30	**0.44**	0.19	0.03	0.00	0.00	0.00
Allocation	mTPI	4.9	10.2	**9.4**	4.5	0.9	0.1	0.0	0.0
	RED	4.7	9.6	**9.5**	4.6	1.3	0.2	0.0	0.0
	t-statistics	5.0	10.4	**9.7**	3.9	0.8	0.1	0.0	0.0

Table 3.6 Proportion of trials each dose was recommended as the MTD by the TITE-CRM and the RED. The length of follow-up for DLT is 35 days with enrollment rate of 1 patient per week. The target DLT rate is 0.25 and the total sample size is n = 30. Numbers at the target dose are in bold.

		d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
Scenario 1		0.05	**0.25**	0.50	0.60	0.70	0.80	0.90	0.95
Recommendation	TITE-CRM	0.11	**0.72**	0.16	0.00	0.00	0.00	0.00	0.00
	RED	0.11	**0.81**	0.08	0.00	0.00	0.00	0.00	0.00
Allocation	TITE-CRM	6.6	**14.5**	7.2	1.3	0.4	0.1	0.0	0.0
	RED	11.0	**15.5**	3.2	0.3	0.0	0.0	0.0	0.0
Scenario 2		0.01	0.05	**0.25**	0.50	0.60	0.70	0.80	0.90
Recommendation	TITE-CRM	0.00	0.09	**0.73**	0.17	0.01	0.00	0.00	0.00
	RED	0.00	0.13	**0.80**	0.07	0.00	0.00	0.00	0.00
Allocation	TITE-CRM	3.1	5.2	**13.3**	6.7	1.3	0.4	0.0	0.0
	RED	3.7	10.3	**13.0**	2.7	0.2	0.0	0.0	0.0
Scenario 3		0.01	0.02	0.05	**0.25**	0.50	0.60	0.70	0.80
Recommendation	TITE-CRM	0.00	0.00	0.12	**0.69**	0.18	0.01	0.00	0.00
	RED	0.00	0.01	0.18	**0.73**	0.08	0.00	0.00	0.00
Allocation	TITE-CRM	3.1	3.1	5.2	**11.1**	6.0	1.3	0.2	0.0
	RED	3.4	4.2	9.0	**10.7**	2.5	0.2	0.0	0.0
Scenario 4		0.01	0.02	0.03	0.05	**0.25**	0.50	0.60	0.70
Recommendation	TITE-CRM	0.00	0.00	0.00	0.15	**0.64**	0.20	0.01	0.00
	RED	0.00	0.01	0.05	0.21	**0.65**	0.09	0.00	0.00
Allocation	TITE-CRM	3.1	3.1	3.3	5.1	**9.5**	5.1	0.9	0.0
	RED	3.4	3.9	4.5	7.6	**8.4**	2.0	0.2	0.0
Scenario 5		0.01	0.02	0.03	0.04	0.05	**0.25**	0.50	0.60
Recommendation	TITE-CRM	0.00	0.00	0.00	0.01	0.16	**0.64**	0.18	0.01
	RED	0.00	0.00	0.05	0.12	0.19	**0.55**	0.08	0.00
Allocation	TITE-CRM	3.0	3.0	3.2	3.6	5.1	**8.2**	3.5	0.3
	RED	3.4	4.0	4.5	4.6	5.7	**6.0**	1.7	0.2
Scenario 6		0.01	0.02	0.03	0.04	0.05	0.06	**0.25**	0.50
Recommendation	TITE-CRM	0.00	0.00	0.00	0.01	0.04	0.23	**0.61**	0.12
	RED	0.00	0.01	0.05	0.12	0.12	0.21	**0.42**	0.08
Allocation	TITE-CRM	3.0	3.0	3.2	3.6	4.0	5.2	**6.3**	1.6
	RED	3.4	4.0	4.4	4.5	4.2	4.1	**4.1**	1.2
Scenario 7		0.01	**0.05**	0.50	0.60	0.70	0.80	0.90	0.95
Recommendation	TITE-CRM	0.00	**0.52**	0.47	0.01	0.00	0.00	0.00	0.00

	RED	0.00	**0.56**	0.44	0.00	0.00	0.00	0.00	0.00
Allocation	TITE-CRM	3.4	**11.5**	12.0	2.5	0.4	0.3	0.0	0.0
	RED	3.7	**18.5**	7.1	0.6	0.0	0.0	0.0	0.0
Scenario 9		0.15	**0.25**	0.35	0.45	0.55	0.65	0.75	0.85
Recommendation	TITE-CRM	0.19	**0.48**	0.29	0.04	0.00	0.00	0.00	0.00
	RED	0.36	**0.48**	0.14	0.02	0.00	0.00	0.00	0.00
Allocation	TITE-CRM	9.0	10.8	**7.3**	2.1	0.6	0.1	0.0	0.0
	RED	16.1	10.0	**3.2**	0.6	0.1	0.0	0.0	0.0
Scenario 10		0.05	0.15	**0.25**	0.35	0.45	0.55	0.65	0.75
Recommendation	TITE-CRM	0.01	0.17	**0.49**	0.28	0.05	0.00	0.00	0.00
	RED	0.03	0.41	**0.43**	0.12	0.01	0.00	0.00	0.00
Allocation	TITE-CRM	3.9	6.8	**10.6**	6.2	2.0	0.5	0.1	0.0
	RED	7.3	12.0	**7.7**	2.4	0.5	0.1	0.0	0.0

CHAPTER 4

DOSE-FINDING FOR CONTINUOUS OUTCOME IN PHASE I ONCOLOGY TRIALS

4.1 Introduction

Dose limiting toxicity (yes or no) is the primary endpoint in most oncology Phase I trials of cytotoxic agents. If a cytostatic agent is being investigated, toxicity is usually not a limiting factor. For example, out of 82 recent Phase I trials with cytostatic drugs reviewed by (Penel et al., 2011) dose limiting toxicity (DLT) was reached in 43 (52%) of the trials. Therefore, a continuous biomarker endpoint might be a better primary endpoint in a Phase I trial. Examples include a measure of target inhibition or pharmacokinetic endpoints such as plasma drug concentrations that correlate with biological activity (Le Tourneau et al., 2009), or percentage inhibition of an enzyme (Plummer et al., 2008).

Continuous endpoint also arises when multiple toxicity events in different body systems and multiple toxicity grades are combined into a single score (Bekele and Thall, 2004; Ezzalfani et al., 2012); Chen et al., 2010). Binary endpoint of DLT used in most of Phase I trials is defined as treatment related non-hematological toxicity of grade 3 or higher or treatment related hematological toxicity of grade 4 or higher. Wang et al. (2000) showed that the maximum tolerated dose (MTD) can be identified more accurately through incorporating the information on grade 3 and grade 4 toxicity instead of using DLT. The MTD is then defined as the dose with a certain mean toxicity score. Example of toxicity scores proposed recently include total toxicity burden (Bekele and Thall, 2004), equivalent toxicity score (Chen et al., 2010), average toxicity

score (ATS) (Bekele et al., 2010), and total toxicity profile (Ezzalfani et al., 2012). These scores are computed by combining information from various toxicity grades and various types of toxicity into a single number with the goal of better reflecting toxicity burden on a patient compared to the binary outcome of a DLT.

Dose-finding designs for continuous outcome have been proposed through controlling dose escalation via a dichotomized outcome (Mandrekar et al., 2007, Mandrekar et al., 2009, Hunsberger et al., 2005). Several methods work with the continuous endpoint directly (Ivanova and Kim, 2009, Eichhorn and Zacks, 1973). Dose-finding designs for toxicity score include quasi continual reassessment method (Yuan et al., 2007), extended isotonic design (Chen et al., 2010), the design in (Ezzalfani et al., 2012, Lee et al., 2010) and quasi-likelihood continual reassessment method of Ezzalfani et al. (2012). The *t*-statistics design (Ivanova and Kim, 2009) can also be used in trials with toxicity score as primary endpoint. All the methods mentioned earlier require outcome of a patient to be observed quickly. In many trials, time to observe the outcome is long compared to the accrual rate. In such trials when urgent treatment is needed, it is desirable to assign a dose to a patient as soon as the patient enrolls in the trial. The proposed design allows making the best possible assignment for each incoming patient using all information available.

In Section 4.2, we describe the proposed method. We give an example in Section 4.3. Simulation results are presented in Section 4.4 and discussion in Section 4.5.

4.2 Notation and methods

4.2.1 Probability model

Let $D = \{d_1, ..., d_K\}$ be the set of ordered dose levels selected for a trial. Let Y_j be the outcome at dose $d_j, j = 1, \ldots, K$, $Y_j \sim N(\mu_j, \sigma_j^2)$. We assume that both mean μ_j and variance σ_j^2

are unknown. The goal is to find a dose $d_m \in D$ with the mean response closest to the target response η. We assume that mean responses are non-decreasing with dose, $\mu_1 \le ... \le \mu_K$. The idea of the proposed design is to compute a Bayesian probability for each dose to be the target dose, that is, the probability for a dose to have the mean response close to the target η. The next assignment is made to the dose where this probability is the highest. We start with describing a Bayesian model for the data. We present the model for the case when outcome variances are assumed to be the same $\sigma_j^2 = \sigma^2, j = 1,\ldots, K$. The case when the variances are not the same is easier, as data from each dose is handled separately.

Ignoring the monotonicity, a conjugate prior density (Gelman et al., 1995) can be specified as

$$\mu_j \mid \sigma^2 \sim N(\mu_{0j}, \sigma^2 / k_{0j}),\ j = 1, 2,...,K, \text{ and } \sigma^2 \sim IG(\nu_0, \sigma_0^2),$$

where *IG* denotes inverse gamma distribution. Let n_j be the number of subjects assigned to d_j, $N = n_1 + \ldots + n_K$. The posterior of $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)'$ conditional on σ^2 and observed responses $\boldsymbol{y}$ is

$$\mu_j \mid \sigma^2, \boldsymbol{y} \sim N(M_j; V_j), j = 1, 2,..., K, \text{ and } \sigma^2 \mid \boldsymbol{y} \sim IG(\nu_n, \sigma_n), \qquad (1)$$

where $M_j = (k_{0j}\mu_{0j} + n_j \bar{y}_j)/(k_{0j} + n_j)$, $V_j = \sigma^2/(k_{0j} + n_j)$, $\nu_n = \nu_0 + N/2$, and

$$\sigma_n = \sigma_0 + \frac{1}{2}\sum_{j=1}^{K}\left\{(n_j - 1)s_j^2 + \frac{k_{0j}n_j}{k_{0j} + n_j}(\bar{y}_j - \mu_{0j})^2\right\}, \text{ with } (\bar{y}_1, ..., \bar{y}_K)' \text{ denoting the unrestricted}$$

maximum likelihood estimates of the mean response, and $(s_1^2,...,s_K^2)$ denoting the empirical variances.

To impose monotonic restriction on the posterior means $\mu_j \mid \sigma^2, \boldsymbol{y}$ we use the isotonic transformation approach of (Dunson and Neelon, 2003) and map unconstrained mean vector

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ from $R^K \to \Omega$ to obtain the posterior distribution for the restricted means. Here $\Omega \subset R^K$ is defined by a set of vectors such that $\mu_1 \leq \ldots \leq \mu_K$. We first compute the posterior distribution of unconstrained parameter vector $\boldsymbol{\mu}$, then obtain the draws from the posterior, and transform draws to the constrained draws from the posterior density for the constrained parameter vector, $\boldsymbol{\mu}^*$, using the pool adjacent violators algorithm (PAVA) (Barlow et al., 1972). From transformed draws, for each dose $d_j, j = 1,\ldots, K,$ we compute the following probabilities

$$\pi_j = P(\eta - \varepsilon < \mu_j < \eta + \varepsilon \mid \sigma^2, \boldsymbol{y}),$$
$$\rho_j = P(\mu_j > \eta \mid \sigma^2, \boldsymbol{y}),$$
$$\tau_j = P(\mu_j < \eta \mid \sigma^2, \boldsymbol{y})$$

The probability π_j shows how likely it is that dose d_j is the target dose. The probability ρ_j is the probability that the mean response at d_j exceeds η. This probability is used to stop the trial if the lowest dose is too toxic in trials where toxicity related outcome is the main endpoint. Probabilities ρ_j and τ_j are useful to decide whether a dose should be inserted as described in Section 4.2.2.

4.2.2 The Bayesian design for continuous outcomes (BDCO)

Subjects are assigned sequentially starting with the lowest dose. Similar to the start-up rule in toxicity dose-finding studies (Ivanova et al., 2003,Cheung, 2005), we recommend assigning at least three subjects to any untried dose before the dose can be escalated. Let k be the maximum dose with at least one subject assigned to it, and n_j be the number of subjects assigned to dose $d_j, j = 1,\ldots, k$. The next subject is assigned to the dose with the maximum probability π_j. That is, the next subject is assigned to the dose d_m that is most likely to be the target dose,

$m = \arg\max \pi_j$. If the primary endpoint is a safety endpoint, the dose for the next subject is chosen only among the doses such that $\rho_j < 0.95$.

Often, it takes several weeks to observe the outcome. As the first cohort of patients (often 3 patients) at the higher previously untried dose are being followed, patients can be assigned at the lower dose. The BCDO allows incorporation of all data to find the best assignment for the next patient. For example, if values of response much higher than the target are observed at the dose level right below the dose where initial cohort is being treated, lower doses will be chosen for the next patient, as the next patient is assigned to the dose with the highest value of π_j. Other dose-finding designs, for example, the *t*-statistic design (Ivanova and Kim, 2009) and extended isotonic design (Chen et al., 2010) can be cumbersome to use with delayed outcome and rapid enrollment since the assignments are tied to the "current" dose. In our example, the "current" dose is the dose where the initial cohort of 3 is being treated, therefore next assignment is determined based on responses from these 3 patients. In design with a parsimonious working model the next assignment should be to the dose prescribed by the design, otherwise convergence is not guaranteed. Therefore, when patients are assigned at lower dose levels while the initial cohort of 3 is being followed at the higher dose, information from these additional patients should not be used to determine the next assignment. If the primary endpoint is a safety endpoint, the trial is stopped earlier if $\rho_1 > 0.95$ since the lowest dose is unsafe.

Bayesian framework allows setting up a rule to insert a dose similarly to trials with binary outcome (Hu et al., 2013). If the posterior mean μ_1^* at the lowest dose is higher than the target, $\mu_1^* > \eta$, and $\rho_1 > 0.95$, a dose is inserted below d_1. If $\mu_K^* < \eta$ and $\tau_K > 0.95$ the dose is inserted above d_K. Otherwise, we find two adjacent doses such that $\mu_j^* < \eta < \mu_{j+1}^*$, then if

$\tau_j > 0.95$ and $\rho_{j+1} > 0.95$ a dose might be inserted between d_j and d_{j+1}. For example, in a trial with toxicity score endpoint and the target score of $\eta = 0.28$, 3 subjects were assigned to dose d_1, 12 subjects to d_2 and 7 subjects to d_3, yielding $\bar{y}_1 = 0.07$, $\bar{y}_2 = 0.18$, $\bar{y}_3 = 0.40$, $s_1^2 = 0.016$, $s_2^2 = 0.023$, and $s_3^2 = 0.006$. With prior parameters $\mu_{0j} = 0$, $k_{0j} = 0.1$, $\nu_0 = 0.05$ and $\sigma_0 = 0.0005$, and assuming equal variances, the order restricted posterior means at the three doses were $\hat{\mu}_1^* = 0.06$, $\hat{\mu}_2^* = 0.16$, and $\hat{\mu}_3^* = 0.39$. The probability that the mean at d_2 is below the target was $\tau_2 = 0.99$ and the probability that the mean at d_3 exceeds the target is $\rho_3 = 0.98$. Therefore one should consider inserting a dose between d_2 and d_3.

4.3 Example

We use data from Friedman et al. (1998) to illustrate the decision rules. The goal of the trial was to find a dose that produces undetectable AGT activity. Though the AGT outcome was continuous it was dichotomized at 10 fmol/mg to facilitate dose escalation in the trial. We use continuous outcome reported by Friedman et al. (1998). The trial accrued 30 patients; 3, 3, 13, and 11 patients to doses 40, 60, 80 and 100 mg/m^2, respectively. We set the goal of the trial as to finding the dose with AGT activity equal to 5 fmol/mg protein. To construct our example (Table 4.1) outcomes y_{31}, y_{32}, y_{33} were randomly selected from 13 outcomes reported at 80 mg/m^2, and we have randomly generated 3 outcomes at dose 120 mg/m^2. The hypothetical trial was stopped when the data at the 100 mg/m^2 dose-cohort were exhausted. Patients with undetectable AGT activity were assigned a value of 5 fmol/mg. Since the AGT activity was believed to be decreasing with dose, the dose was escalated if the mean AGT activity was higher than the target

5. Table 4.1 gives the averages and posterior means of AGT activity and decisions for each dose cohort.

4.4 Simulation study

For trials with continuous outcomes, we compared the new design with the *t*-statistic design ((Ivanova and Kim, 2009). In the *t*-statistics design, the *t*-statistic T for comparing the mean at the current dose with the target is computed. The next patient is assigned to the current dose if $-\Delta < T < \Delta$, otherwise the dose is reduced or increased depending on the sign of T. Here Δ is a design parameter and $\Delta = 1$ is recommended. We used the scenarios of true mean response as in (Ivanova and Kim, 2009) assuming the true variances are either equal or unequal across different doses. For each dose level d_j, j = 1, 2, 3, 4, 5, 6, the outcome Y_j has a normal distribution with mean $0.1j$ and variance 0.2. We also generated scenarios with unequal variances where Y_j has mean $0.1j$ and variance 0.1^2j^2. We varied the values of the target η, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, so that the target dose is dose d_1 in scenario 1, the target dose is d_2 in scenario 2 etc. The sample size was 36 for all scenarios. Prior parameters were $\mu_{0j} = 0$, $k_{0j} = 0.1$, $\nu_0 = 0.05$ and $\sigma_0 = 0.0005$. The BDCO parameter was set to $\varepsilon = 0.01$. The BDCO we simulated did not include the provision to insert doses. The BDCO was simulated when assuming equal variances and not making this assumption. The design performed slightly better when equal variances were assumed. Similarly both versions of the *t*-statistic design were simulated, under assumption of equal variances with and without this assumption. The *t*-statistic design performed slightly better when variances were not assumed to be equal. Therefore we only display results for the BDCO assuming equal variances and the *t*-statistic design assuming unequal variances. We obtain draws from the posterior density function described in Section 4.2.1 via Gibbs sampling algorithm and

then transform draws to the constrained draws from the posterior density of the constrained vector $\boldsymbol{\mu}^*$. This process was performed 1500 times with the first 500 iterations discarded as burn-in. Results were obtained by simulations with 5000 replicates.

Table 4.2 displays the proportion of runs where a dose was recommended as the target dose and the percentage of subjects allocated to each dose. As can be seen in Table 4.2, when the true variances are equal, the percent recommendations of correct dose are slightly higher for our new design than for the *t*-statistics design in five out of six scenarios. Whereas when the true variances are not equal, the percent recommendations of correct dose are about the same.

We also compared the performance of the new design in trials with toxicity score. We used total toxicity profile score proposed by Ezzalfani et al. (2012) that can be obtained by converting various toxicity grades in three different toxicity types, renal, neurological and hematologic, into a single score. We used all 8 scenarios from Ezzalfani et al. (2012). The total sample size was 36 and the target score was 0.28. The designs compared were the *t*-statistics method ((Ivanova and Kim, 2009) and the Quasi-Likelihood Continual Reassessment Method QLCRM (Ezzalfani et al., 2012). The new design performed as good as the *t*-statistics design (Table 4.3), and QLCRM performed better than the new design and the *t*-statistic in several scenarios. Recall that compared to other methods, the new design is more flexible as dose assignments are not tied to the current dose. The new method allows both stopping the trial early and inserting doses.

We also investigated the robustness of the choice of design parameter ε. We rerun simulations with parameter ε set to various values in ($\eta/100$, $\eta/10$) and obtained results very similar to results in Tables 4.2 and 4.3 (data are available from the authors).

4.5 Discussion

We proposed a new Bayesian dose-finding design for continuous outcomes that is based on computing, for each dose, the probability that the mean response is close to the target value. Somewhat similar approach is a dose-finding design by Ji et al. (2010) for binary outcomes which is based on computing quantities similar to π_j, τ_j and ρ_j. However, the choice of assignments in the design of Ji et al. (2010) is tied to the current dose and a different strategy is used to determine the assignment based on the three quantities. Our new design allows making the best possible assignment for each incoming patient using all information available. For example, consider a trial where patients are assigned to the lower doses while the initial cohort of subjects is being followed at the higher dose. Since assignments are not tied to a current dose, information from both subjects at lower doses as well as subjects at a higher dose is used to make each new assignment.

Table 4.1 Example of a trial with subjects assigned in cohorts of 3. The target AGT activity is 5 fmol/mg protein. Data were re-sampled from Friedman et al. (1998).

Dose	Data	$\bar{Y}_j,\ \mu_j^*$
Cohort 1, Dose 1	$(y_{11}, y_{12}, y_{13}) = (26.35, 42.00, 15.00)$	27.78, 26.89
Decision: Increase the dose, since $\mu_1^* > 5$		
Cohort 2, Dose 2	$(y_{21}, y_{22}, y_{23}) = (23.00, 13.50, 10.83)$	15.78, 15.20
Decision: Increase the dose, since $\mu_2^* > 5$		
Cohort 3, Dose 3	$(y_{31}, y_{32}, y_{33}) = (11.70, 9.03, 5.00)$	8.58, 8.12
Decision: Increase the dose, since $\mu_3^* > 5$		
Cohort 4, Dose 4	$(y_{41}, y_{42}, y_{43}) = (4.07, 5.00, 8.70)$	5.92, 5.08
Decision: Increase the dose, since $\mu_4^* > 5$		
Cohort 5, Dose 5	$(y_{51}, y_{52}, y_{53}) = (2.01, 3.00, 4.10)$	3.04, 2.56
While the first three subjects were followed at dose 5, three more subjects were enrolled at dose 4:		
Cohort 6, Dose 4	$(y_{44}, y_{45}, y_{46}) = (2.50, 4.07, 6.13)$	5.08, 5.17
Decision: Since $\mu_4^* > 5$ and $\mu_5^* < 5$, compute $\pi_4 = \Pr(\mu_4 > 5) = 0.040$ and $\pi_5 = \Pr(\mu_5 < 5) = 0.020$. Dose 4 has higher likelihood of being the target dose, therefore next assignment is to dose 4		
Cohort 7, Dose 4	$(y_{47}, y_{48}, y_{49}) = (3.60, 5.00, 5.00)$	4.90, 4.95
Decision: Since $\mu_3^* > 5$ and $\mu_4^* < 5$, compute $\pi_3 = 0.020$ and $\pi_4 = 0.050$. Dose 4 has higher likelihood of being the target dose, therefore next assignment is to dose 4		
	$(y_{410}, y_{411}) = (6.80, 6.60)$	5.22

Table 4.2 Proportion of trials where each dose was recommended as the target dose and proportion of subjects allocated to each dose by the t-statistics design and the new Bayesian design (BDCO). In case of unequal variances the variance of Y_j is $0.1^2 j^2,\ j \in \{1,2,3,4,5,6\}$, in case of equal variances the variance is 0.2. The target value in scenario k, k = 1,…, 6 is 0.1k. Results for the target dose are in bold.

	TRUE	Percent recommendation						Percent allocation					
	variance	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Scenario 1		0.1	0.2	0.3	0.4	0.5	0.6						
t-statistics	equal	**0.93**	0.07	0.00	0.00	0.00	0.00	**0.82**	0.17	0.01	0.00	0.00	0.00
BDCO	equal	**0.91**	0.10	0.00	0.00	0.00	0.00	**0.78**	0.21	0.02	0.00	0.00	0.00
t-statistics	unequal	**0.97**	0.03	0.00	0.00	0.00	0.00	**0.82**	0.17	0.01	0.00	0.00	0.00
BDCO	unequal	**0.97**	0.03	0.00	0.00	0.00	0.00	**0.85**	0.13	0.02	0.00	0.00	0.00
Scenario 2		0.1	0.2	0.3	0.4	0.5	0.6						
t-statistics	equal	0.09	**0.83**	0.08	0.00	0.00	0.00	0.28	**0.58**	0.14	0.01	0.00	0.00
BDCO	equal	0.07	**0.86**	0.08	0.00	0.00	0.00	0.19	**0.63**	0.16	0.01	0.00	0.00
t-statistics	unequal	0.06	**0.82**	0.11	0.01	0.00	0.00	0.19	**0.62**	0.18	0.02	0.00	0.00
BDCO	unequal	0.04	**0.84**	0.11	0.01	0.00	0.00	0.16	**0.63**	0.17	0.03	0.00	0.00
Scenario 3		0.1	0.2	0.3	0.4	0.5	0.6						
t-statistics	equal	0.00	0.09	**0.82**	0.10	0.00	0.00	0.10	0.26	**0.51**	0.12	0.01	0.00
BDCO	equal	0.00	0.07	**0.83**	0.09	0.00	0.00	0.09	0.19	**0.55**	0.16	0.01	0.00
t-statistics	unequal	0.00	0.14	**0.67**	0.17	0.02	0.00	0.09	0.26	**0.48**	0.15	0.02	0.00
BDCO	unequal	0.00	0.16	**0.65**	0.16	0.02	0.00	0.08	0.25	**0.45**	0.16	0.04	0.01
Scenario 4		0.1	0.2	0.3	0.4	0.5	0.6						
t-statistics	equal	0.00	0.00	0.10	**0.80**	0.11	0.00	0.09	0.10	0.25	**0.45**	0.11	0.01
BDCO	equal	0.00	0.00	0.08	**0.81**	0.11	0.00	0.08	0.09	0.18	**0.49**	0.15	0.02
t-statistics	unequal	0.00	0.00	0.25	**0.55**	0.17	0.03	0.08	0.11	0.32	**0.36**	0.11	0.02
BDCO	unequal	0.00	0.00	0.27	**0.50**	0.18	0.04	0.08	0.09	0.29	**0.33**	0.15	0.05
Scenario 5		0.1	0.2	0.3	0.4	0.5	0.6						
t-statistics	equal	0.00	0.00	0.00	0.12	**0.77**	0.11	0.08	0.09	0.10	0.24	**0.39**	0.10
BDCO	equal	0.00	0.00	0.00	0.09	**0.80**	0.11	0.08	0.08	0.09	0.17	**0.43**	0.15
t-statistics	unequal	0.00	0.00	0.02	0.36	**0.45**	0.18	0.08	0.09	0.16	0.32	**0.25**	0.10
BDCO	unequal	0.00	0.00	0.02	0.34	**0.45**	0.20	0.08	0.08	0.12	0.28	**0.27**	0.16
Scenario 6		0.1	0.2	0.3	0.4	0.5	0.6						
t-statistics	equal	0.00	0.00	0.00	0.00	0.14	**0.86**	0.08	0.08	0.09	0.11	0.23	**0.41**
BDCO	equal	0.00	0.00	0.00	0.00	0.10	**0.90**	0.08	0.08	0.08	0.09	0.16	**0.50**
t-statistics	unequal	0.00	0.00	0.00	0.07	0.41	**0.52**	0.08	0.08	0.10	0.20	0.29	**0.24**
BDCO	unequal	0.00	0.00	0.00	0.07	0.40	**0.54**	0.08	0.08	0.09	0.15	0.26	**0.33**

Table 4.3 Proportion of trials where each dose was recommended as the target dose and proportion of subjects allocated to each dose by the QLCRM, the t-statistics design and the new Bayesian design (BDCO). The target is 0.28. Results at the target dose are in bold.

	Percent recommendation						Percent allocation					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
Scenario 1	0.18	0.28	0.36	0.41	0.43	0.44						
QLCRM	0.03	**0.86**	0.11	0.00	0.00	0.00	0.19	**0.63**	0.17	0.01	0.00	0.00
t-statistics	0.06	**0.85**	0.10	0.00	0.00	0.00	0.26	**0.60**	0.14	0.00	0.00	0.00
BDCO	0.04	**0.86**	0.09	0.00	0.00	0.00	0.22	**0.60**	0.15	0.02	0.00	0.00
Scenario 2	0.10	0.20	0.31	0.39	0.44	0.48						
QLCRM	0.00	0.13	**0.85**	0.02	0.00	0.00	0.09	0.21	**0.61**	0.09	0.00	0.00
t-statistics	0.00	0.18	**0.80**	0.01	0.00	0.00	0.10	0.35	**0.50**	0.05	0.00	0.00
BDCO	0.00	0.22	**0.77**	0.01	0.00	0.00	0.10	0.36	**0.46**	0.07	0.00	0.00
Scenario 3	0.11	0.18	0.28	0.36	0.41	0.43						
QLCRM	0.00	0.03	**0.84**	0.13	0.00	0.00	0.09	0.15	**0.57**	0.18	0.01	0.00
t-statistics	0.00	0.06	**0.83**	0.11	0.00	0.00	0.10	0.25	**0.52**	0.12	0.00	0.00
BDCO	0.00	0.06	**0.83**	0.10	0.00	0.00	0.10	0.22	**0.52**	0.15	0.02	0.00
Scenario 4	0.05	0.12	0.27	0.37	0.40	0.46						
QLCRM	0.00	0.00	**0.83**	0.17	0.00	0.00	0.08	0.09	**0.51**	0.30	0.01	0.00
t-statistics	0.00	0.00	**0.93**	0.07	0.00	0.00	0.08	0.16	**0.60**	0.16	0.00	0.00
BDCO	0.00	0.00	**0.92**	0.07	0.00	0.00	0.08	0.17	**0.60**	0.13	0.01	0.00
Scenario 5	0.05	0.10	0.19	0.31	0.42	0.45						
QLCRM	0.00	0.00	0.09	**0.91**	0.01	0.00	0.08	0.08	0.15	**0.60**	0.08	0.00
t-statistics	0.00	0.00	0.15	**0.84**	0.00	0.00	0.08	0.09	0.32	**0.45**	0.04	0.00
BDCO	0.00	0.00	0.14	**0.86**	0.00	0.00	0.08	0.10	0.32	**0.44**	0.06	0.00
Scenario 6	0.05	0.11	0.18	0.28	0.36	0.41						
QLCRM	0.00	0.00	0.03	**0.81**	0.17	0.00	0.08	0.09	0.13	**0.51**	0.19	0.01
t-statistics	0.00	0.00	0.07	**0.82**	0.11	0.00	0.08	0.10	0.24	**0.46**	0.11	0.00
BDCO	0.00	0.00	0.08	**0.82**	0.10	0.00	0.08	0.10	0.22	**0.45**	0.14	0.02
Scenario 7	0.04	0.05	0.11	0.18	0.28	0.36						
QLCRM	0.00	0.00	0.00	0.03	**0.80**	0.18	0.08	0.08	0.08	0.12	**0.45**	0.18
t-statistics	0.00	0.00	0.00	0.08	**0.79**	0.13	0.08	0.09	0.10	0.23	**0.40**	0.10
BDCO	0.00	0.00	0.00	0.10	**0.80**	0.10	0.08	0.08	0.10	0.21	**0.39**	0.13
Scenario 8	0.05	0.11	0.14	0.19	0.25	0.35						
QLCRM	0.00	0.00	0.00	0.02	**0.83**	0.16	0.08	0.08	0.09	0.15	**0.48**	0.12
t-statistics	0.00	0.00	0.00	0.01	**0.78**	0.21	0.08	0.09	0.10	0.16	**0.40**	0.17
BDCO	0.00	0.00	0.00	0.01	**0.80**	0.19	0.08	0.08	0.09	0.13	**0.44**	0.18

CHAPTER 5

USE OF SELECTIVE PHENOTYPYING TO INCREASE POWER OF GENETIC ASSOCIATION STUDIES

5.1 Introduction

Replication studies play an important role in validating results from a GWAS. For many GWASs, the trait being studied is a quantitative measure that required a significant investment of time and money to collect. There are many potential replication cohorts that have existing genotype data on candidate SNPs under study but who have yet to measure the quantitative trait of interest. For these studies, the number of subjects that can be measured for the quantitative trait could be limited by time and cost constraints. The selection of the subset of subjects to phenotype would likely play a critical role in the success or failure of the replication study.

Historically, the expense of genotyping has been the major cost constraint for large-scale genetic association studies. One solution to remedying the high cost of genotyping, when studying quantitative traits, was to preferentially select subjects with extreme values of the phenotype of interest for genotyping (Darvasi and Soller, 1992). Today, many large cohort studies have existing genome-wide SNP data and the burden of expense has now shifted to unmeasured phenotypes.

Methods to maximize power using selective phenotyping have been developed for experimental animal and plant crosses. These methods can be classified into two categories. One category was developed to maximize the number of recombination events in selected subjects in order to have increased power to narrow the location of the unknown underlying causal locus

amongst tested genetic markers in a linkage map. The subject with the largest number of recombination events between all markers is selected first, and so on, until the number of subjects is equal to the final allowable sample size (Jannink, 2005, Xu et al., 2005). The other category of methods focuses on balancing the distribution of alleles for the selected sets of markers. In this category, measures of similarity are used to find the subset of subjects with genotypes as dissimilar as possible (Sen et al., 2009, Jin et al., 2004). A more recent Bayesian approach combines some of the features of both approaches (Gagneur et al., 2011). Though there are obvious differences between these methods, the rationale behind the methods is similar. Namely, the goal is to identify the subset of subjects who are as genetically diverse as possible with respect to distributions of recombination events and/or marker genotype data.

Determining an optimal selective phenotyping strategy for human genetic association studies of multiple variants is challenging. Assuming an additive genetic risk model, maximizing power for a single SNP associated with a quantitative phenotype is achieved by phenotyping an equal number of subjects who are homozygous for the major and minor alleles. Identifying the optimal selection strategy when considering multiple SNPs is much more difficult, as such a strategy has to consider a possible wide-range of allele frequencies and effect sizes across the SNPs being studied. The power for a subset of SNPs of interest might not be sensitive to sample selection while the power for the remaining SNPs could be very sensitive to sample selection. The selection of subjects for phenotyping that maximizes power across a set of SNPs is ultimately a complex optimization problem.

Simulated annealing (SA) is a computationally efficient optimization algorithm that finds the global minimum or maximum of complex functions containing many variables (Kirkpatrick et al., 1983, Coghlan et al., 2001). SA originated from the Metropolis's Algorithm (Metropolis et

al., 1953). In contrast to other optimization algorithms, SA does not require calculation of derivatives. The method is motivated by an analogy to solving the combinatorial optimization problem related to the physical annealing process of heating solids to a high temperature, so that the atoms will be able to rearrange freely, followed by a slow cooling process that results in a strong crystal in the minimum energy configuration.

We describe a method, based on SA, which identifies the optimal subset of subjects, based on their available genotype data at a set of SNPs of interest, to be phenotyped for inclusion in a genetic association study. The samples are chosen to either maximize average or minimum power across the selected SNPs. We demonstrate, through simulations and an empirical example, the improved power that can be achieved by using this approach compared to the alternative approach of selecting a random sample for phenotyping.

5.2 Methods

5.2.1 Statistical power calculation

We follow the approach described by Ambrosius et al. to calculate statistical power for a SNP associated with a quantitative trait (Ambrosius et al., 2004). Specifically, we utilized a linear regression model to study the association between the phenotype and SNP (i.e.,Model : $\mathbf{y}=X\beta + \varepsilon$, where $\mathbf{y}$ is the vector containing the dependent continuous variable for phenotype; X is the matrix of independent variables written as $X=[\mathbf{1}\ \mathbf{g}]$, $\mathbf{g}$ being the vector of independent variables for the genotype; β is the vector of regression coefficients written as $\beta=[\beta_0, \beta_1]^T$ such that β_0 is the parameter for the intercept and β_1 is the coefficient explaining the effect of genotype; ε is a vector of independent and normally distributed random variable with mean zero and variance σ^2). In genetic association models, we test the null hypothesis (H_0) that no

association exists between the phenotype and genotype (i.e., H_0: $\beta_1=0$). This hypothesis can be tested using an *F*-test. To calculate the power under the alternative hypothesis (H_1) that there is an association between phenotype and genotype (i.e., H_1: $\beta_1 \neq 0$), we need to specify the critical value (crit) used to assess significance and the value of the non-centrality parameter, φ. The critical value is calculated under the null hypothesis satisfying $Pr[F_{df1,fdf2} > F_{crit}]=\alpha$. Assuming an additive effects model for genotype on phenotype, we calculate power, $Pr[F_{df1, fdf2,\varphi} > F_{crit}]$, as follows (Ambrosius et al., 2004).

$$Pow = Pr[F_{df1,df2,\ \varphi} > F_{crit}],\ \text{where: } \varphi = \beta' Z' Z \beta / (2\sigma^2),$$

the non-centrality parameter for the F distribution

$$Z'Z = k_1\left(\frac{k_2 + 2k_3}{k}\right)^2 + k_2\left(\frac{k_1 - k_3}{k}\right)^2 + k_3\left(\frac{2k_1 + k_2}{k}\right)^2$$

$k=k_1+k_2+k_3$, $df_1=1, df_2=k-2$

k_1 is number of subjects with genotype AA;

k_2 is number of subjects with genotype AB;

k_3 is number of subjects with genotype BB;

σ^2 is the within group variance.

5.2.2 Simulated Annealing Algorithms

Our goal is to find the subsample, of fixed size, of subjects from the complete sample that provides the greatest overall *average* power across all SNPs of interest. Constructing and calculating average power for all possible subsamples is intractable. Simulated annealing, a Monte Carlo algorithm designed to mimic the heating and subsequent cooling of a substance into a solid state which minimizes the overall kinetic energy, is a widely used algorithm for minimization/maximization problems (Kirkpatrick et al., 1983, Borneman et al., 2001).

Below is the detail of simulated annealing algorithm used in our study:

Parameters include the function to be optimized (here average power), the initial temperature (t^0), the rate of cooling (α) and the convergence criteria (ε).

Let K be total number of available subjects for phenotyping and k ($\leq$ K) be the number of subjects that will be phenotyped. Let n be the number of candidate SNPs that will be tested for association. Let $S^i = \{s_j\}$, where $\{s_j\}$ is a set of k subjects, without replacement, from the complete sample S at iteration i. Let $S^{i+1} = S^i - s_a + s_b$, where subject a is randomly selected and removed from S^i and subject b is randomly selected from S/S^i and added to S^{i+1}. Let $Pow(S^i)$ and $Pow(S^{i+1})$ = average power across n SNPs for set of subjects S^i and S^{i+1}, respectively. To calculate average power, we calculate the power for each individual selected SNP, sum up the powers and divide by the total number of selected SNPs.

Iterative approach:

1) Start with S^0 and calculate $Pow(S^0)$ (i=0).

2) Create S^1 (i=1) by randomly dropping one subject out from S^0 and randomly adding one subject from S/S^0.

3) Calculate $Pow(S^1)$.

4) If $Pow(S^1) > Pow(S^0)$, choose in favor of sample S^1 as new "best" sample. Else, with probability = $\exp((Pow(S^1) - Pow(S^0))/t^0)$, existing sample S^1 is replaced by S^0. (i.e. always accept a favorable move and some times, with decreasing probability as the system cools, choose the less favorable sample.)

5) $t^1 = \alpha \times t^0$, where t^0 is initial starting temperature and α = user defined constant cooling rate.

6) Continue iteratively, creating sample S^{i+1} from S^i in similar fashion and calculating power for sample S^{i+1}.

7) If $Pow(S^{i+1}) > Pow(S^i)$, choose in favor of sample S^{i+1} as new "best" sample. Else, with probability = $\exp((Pow(S^{i+1}) - Pow(S^i))/t^i)$, existing sample S^{i+1} is replaced by S^i and carried forward to the next iteration.

8) $t^i = \alpha \times t^{i-1}$

9) Continue until $t^i < \varepsilon$, where ε is a user-defined threshold for convergence. Final sample at last iteration is chosen as optimal sample.

The algorithm was programmed in R (version 2.11.0; www.r-project.org).

5.2.3 Data simulation

We simulated two sample pools of size K=10,000, each with a different number of target SNPs (n=8 [minor allele frequency (MAF) = 0.05 (x2 SNPs), 0.1 (x2), 0.2, 0.3, 0.4, 0.5]; n=80 [MAF=0.05 (x20), 0.1 (x20), 0.2 (x10), 0.3 (x10), 0.4 (x10), 0.5 (x10)]). For each constructed SNP dataset, SNP genotype data were generated randomly, from their respective MAFs, assuming they were independent from other SNPs and followed Hardy-Weinberg equilibrium. For each sample design, we then calculated power under two scenarios: 1) assuming a constant mean effect (fixed β or change in mean per variant allele) and 2) assuming a constant proportion of total variation explained (fixed R^2) by each SNP. We assumed that $\sigma^2 = 1$ in all power calculations and used a Bonferroni significance threshold of α=0.00625 (0.05/8) and α=0.000625 (0.05/80) for the models with 8 and 80 SNPs, respectively. Only 2000 subjects were assumed to be available for phenotyping; genotype data were assumed to be complete and available on all subjects prior to phenotype sample selection. For each alternative model we

selected 100 random samples of 2000 subjects. The average power across all SNPs, and the individual power for each SNP, was calculated for each random sample. These same 100 random samples were then used as initial starting values in our simulated annealing algorithm to identify a sample of 2000 subjects that maximized the average power across all selected SNPs. The detailed information about the alternative models are shown in Table 5.1. The parameters used for the simulated annealing algorithm were as follows: t^0=400, α =0.9999, ε=0.000001.

5.2.4 Replication of C Reactive Protein (CRP) associations in the Jackson Heart Study (JHS)

To assess the utility of our approach in a "real data" replication study, we used genotype data from JHS to attempt to confirm a reported association from the Women's Health Initiative (WHI) between five independent genetic variants and CRP. Reiner et al. (Reiner et al., 2012) recently reported five genetic variants to be significantly associated with CRP levels in African Americans. JHS, a longitudinal population-based cohort from Jackson, Mississippi, represents the largest single-site, prospective, epidemiological investigation of cardiovascular disease in African Americans (Sempos et al., 1999). Genotype data, generated from the Affymetrix 6.0 Array, on 2983 JHS subjects with CRP data were available to be included in a replication study of the Reiner et al. findings. Three reported SNPs (rs1160985, rs6734238, and rs7748513) were directly available on JHS subjects from the Affymetrix array. Two proxy SNPs were used for the reported SNPs rs16827466 and rs797943. The selected proxy SNP for rs16827466 was rs12239267 (r^2=0.945 and D'=1.00). The selected proxy SNP for rs797943 was rs2393791 (r^2=0.558 and D'=1.00). We first attempted to replicate the previous findings using the entire JHS sample. CRP was log-transformed to meet distributional assumptions and extreme values of CRP>100, likely representing acute infections were removed. Linear regression models were

performed, regressing log-CRP on each SNP individually with covariate adjustment for BMI, sex, age, smoking status (yes/no) and the first four principal components used to control for population stratification.

We next considered the scenario where we could have only afforded to have phenotyped 500 subjects. For each SNP, we randomly sampled 100 different samples, without replacement, performed the same linear models as for the complete data and calculated, across these 100 random samples, the mean p-value, the standard deviation of the p-values, and the proportion of samples where the SNP tests (p-value < 0.05) would have rejected the null hypothesis and confirmed the previous results. We next used so same random samples as initial samples in our simulated annealing algorithm to identify the, *a priori*, optimal sample for phenotyping and repeated the linear model analyses performed for the random samples. The alternative models for each SNP used in the simulated annealing algorithm were defined by the observed effect size estimates in the original report.

5.3 Results

5.3.1 Simulations

Results for the simulation analyses are presented in Tables 5.2 and 5.3. For Model 1, we assumed all SNPs, regardless of MAF, explained the same proportion of variation of the quantitative phenotype (Table 5.2). As expected, for random samples mean power was near uniform across MAFs for both the 8 (average mean power~0.67 for Model 1A) and 80 (average mean power ~0.68 for Model 1B) SNP models. Gains in mean power were observed for all SNPs, regardless of MAF, with application of the simulated annealing algorithm. The largest gains were observed for the SNPs with lower MAF and bigger gains in mean power were

observed for the 8 SNP model compared to the 80 SNP model. For the models with constant main effect (Models 2A and 2B; Table 5.3), not surprisingly SNPs with lower MAF had less mean power than SNPs with higher MAF for the random samples. While this pattern remained true for samples selected from the simulated annealing algorithm, big gains in power were especially noted for SNPs with low MAF. Once again, bigger gains in power for the simulated annealing approach were observed for the model with 8 candidate SNPs compared to the model with 80 candidate SNPs.

In addition to higher mean power estimates when using our simulated annealing approach to selective phenotyping, we note that there is considerably more variability of the power estimates for random samples compared to the simulated annealing selected phenotype samples that used the same random samples as initial samples (Tables 5.2 and 5.3). As illustrated in Figure 5.1 for Model 1A, the mean power for the simulated annealing approach converged quickly to similar values irrespective of the initial starting sample.

5.3.2 JHS CRP replication study

The reported association with CRP in WHI was replicated ($p<0.05$) for four out of five SNPs using the complete sample in JHS (Table 5.4). Restricting the JHS sample size to random draws of size n=500 considerably reduced the evidence for replication in JHS. Focusing on the four SNPs with evidence for association in the complete sample, only rs12239267, a proxy for WHI SNP rs16827466, was consistently replicated in the random draws (mean p-value = 0.014; 96% of draws replicated the result) (Table 5.5). While rs1160985 was replicated in 65% of the random draws, the mean p-value was only 0.12. The remaining two SNPs consistently demonstrated little evidence for association. Applying our selective phenotyping approach

resulted in much higher probabilities to replicate the WHI findings. Specifically, 100% of the selective phenotype samples replicated the findings at rs12239267 and rs1160985, while 86% of the time the result at rs6734238 was also replicated. In addition, the mean p-values were substantially smaller for these SNPs using the selective phenotyping method compared to using random samples. Little evidence for association was observed for rs2393791, a proxy for WHI SNP rs797943, using either approach.

5.4 Discussion

Replication studies are important for controlling the type I error rate of GWAS discoveries. GWASs have identified many associations for quantitative measures and often these measures are missing in potential replication studies. We describe a method that provides investigators, who are concerned about the costs of measuring these quantitative outcomes, an approach to select an *a priori* optimal subsample of their data for phenotyping when conducting a replication study.

Simulated annealing is a popular optimization algorithm for complex functions. Here we describe the function as the average power across a selected set of SNPs. Simulated annealing has been used in solving optimization problems such as probe selection (Borneman et al., 2001) and optimal bit-pattern representation of amino acids (Coghlan et al., 2001). The algorithm itself is very flexible and other functions could be readily adapted. For example, an investigator might be interested in finding the subsample that maximizes the minimum power of a SNP among a set of SNPs. In addition, features such as constraints or weights could be easily included so that some SNPs, deemed more critical to the experiment, get more weight in the selection scheme. As with any simulated annealing design, care by the user should be applied when selecting variables

such as starting temperature, cooling rate and convergence criteria. As we applied in Figure 5.1, a plot of the current value of the power function across iteration is a useful guide to assess convergence properties.

We show through simulations and an empirical example the increase in power our approach can have relative to random sampling. It was not surprising that the biggest gains were observed for SNPs with lower MAF and that the gains were stronger for the smaller the set of SNPs being considered. Replication studies with a narrower set of SNPs to follow will gain more than studies with a more exhaustive set of variants to follow. For studies that want to initially focus on replication but then subsequently contribute to future discovery on a broader list of SNPs, the selection strategy should have little impact, relative to random sampling, on the power for SNPs not included among, and not in linkage disequilibrium with, the list of SNPs used for optimization. There is like little risk the method would harm power for non-selected SNPs. Like any power calculation, the method relies on user specified effect estimates, under the alternative hypothesis, for individual SNPs. Often these estimates can be obtained from the discovery study, though some caution is advised, especially for less common variants, that these initial effect estimates are often biased (too strong). This phenomenon has been described as "the winner's curse" (Bush and Moore, 2012). We have performed additional simulations (data not shown) that suggest the method is reasonably robust to model misspecification, either due to inflated effect estimates for a subset of selected SNPs or the inclusion of null SNPs.

Herein, we describe the method of optimal phenotype selection for replication studies, where the set of genotyped variants for follow up, and their associated effect size estimates, are relatively easy to define based on the preceding discovery study. The method should also be considered for candidate gene studies, where the list of SNPs to include is typically modest.

Under this scenario, the effect size estimates may not be available from previous studies; thus the choice of effect size estimates, similar to standard single SNP power analyses, are more arbitrary. Still, reasonably informed choices can be made based on plausibility (including information on heritability, MAF, etc.) and results from other studies. We are currently working on adapting the method for gene-based tests of less common variants (relatively straight forward for many "gene-burden" statistics) and multi-stage GWAS studies, where a random sample is selected in stage 1 and an informed, phenotype selected, sample is selected in stage 2. Finally, it is possible to extend the method to include imputed genotype data. Given the relative low expense of genotyping a small number of variants compared to measuring a new phenotype, many investigators will choose to directly genotype the SNPs in a replication study. Other burdens besides cost might limit the ability to perform additional genotyping and thus the consideration of using imputed genotype data is an important one.

Table 5.1 Four models used for simulations

MODEL 1A: Constant R^2 for 8 SNPs						
			SNP MAF			
	0.05(n=2)	0.1(n=2)	0.2(n=1)	0.3(n=1)	0.4(n=1)	0.5(n=1)
β	0.229	0.167	0.125	0.109	0.102	0.100
R^2	0.005	0.005	0.005	0.005	0.005	0.005
MODEL 1B: Constant R^2 for 80 SNPs						
			SNP MAF			
	0.05(n=20)	0.1(n=20)	0.2(n=10)	0.3(n=10)	0.4(n=10)	0.5(n=10)
β	0.281	0.204	0.153	0.134	0.125	0.123
R^2	0.0075	0.0075	0.0075	0.0075	0.0075	0.0075
MODEL 2A: Constant β for 8 SNPs						
			SNP MAF			
	0.05(n=2)	0.1(n=2)	0.2(n=1)	0.3(n=1)	0.4(n=1)	0.5(n=1)
β	0.125	0.125	0.125	0.125	0.125	0.125
R^2	0.0015	0.0028	0.005	0.0066	0.0075	0.0078
MODEL 2B: Constant β for 80 SNPs						
			SNP MAF			
	0.05(n=20)	0.1(n=20)	0.2(n=10)	0.3(n=10)	0.4(n=10)	0.5(n=10)
β	0.153	0.153	0.153	0.153	0.153	0.153
R^2	0.0022	0.0042	0.0075	0.0098	0.0113	0.0117

Table 5.2 Mean Power (SD) with constant R^2

Model 1A (8 SNPs, R^2=0.005, α=0.00625):						
Random Sample (100 draws of 2000/10,000 samples)						
			SNP MAF			
0.05(n=2)	0.1(n=2)	0.2(n=1)	0.3(n=1)	0.4(n=1)	0.5(n=1)	Avg(n=8)
0.679	0.663	0.656	0.667	0.666	0.667	0.67
(0.016)	(0.015)	(0.022)	(0.014)	(0.009)	(0.011)	(0.01)
Simulated Annealing Selected Sample (100 SA runs of 2000/10,000 samples)						
			SNP MAF			
0.05(n=2)	0.1(n=2)	0.2(n=1)	0.3(n=1)	0.4(n=1)	0.5(n=1)	Avg(n=8)
0.971	0.941	0.93	0.915	0.894	0.887	0.93
(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.00)
Model 1B (80 SNPs, R^2=0.0075, α=0.000625):						
Random Sample (100 draws of 2000/10,000 samples)						
			SNP MAF			
0.05(n=20)	0.1(n=20)	0.2(n=10)	0.3(n=10)	0.4(n=10)	0.5(n=10)	Avg(n=80)
0.674	0.668	0.679	0.678	0.678	0.679	0.68
(0.011)	(0.006)	(0.009)	(0.004)	(0.004)	(0.003)	(0.00)
Simulated Annealing Selected Sample (100 SA runs of 2000/10,000 samples)						
			SNP MAF			
0.05(n=20)	0.1(n=20)	0.2(n=10)	0.3(n=10)	0.4(n=10)	0.5(n=10)	Avg(n=80)
0.918	0.862	0.812	0.776	0.761	0.761	0.83
(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.00)

Table 5.3 Mean Power (SD) with constant β

Model 2A (8 SNPs, β=0.125, α=0.00625)						
Random Sample (100 draws of 2000/10,000 samples)						
			SNP MAF			
0.05(n=2)	0.1(n=2)	0.2(n=1)	0.3(n=1)	0.4(n=1)	0.5(n=1)	Avg(n=8)
0.155	0.351	0.655	0.811	0.876	0.889	0.53
(0.011)	(0.01)	(0.022)	(0.016)	(0.006)	(0.006)	(0.01)
Simulated Annealing Selected Sample						
			SNP MAF			
0.05(n=2)	0.1(n=2)	0.2(n=1)	0.3(n=1)	0.4(n=1)	0.5(n=1)	Avg(n=8)
0.519	0.698	0.929	0.953	0.959	0.961	0.78
(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.00)
Model 2B (80 SNPs, β=0.153, α=0.000625)						
Random Sample (100 draws of 2000/10,000 samples)						
			SNP MAF			
0.05(n=20)	0.1(n=20)	0.2(n=10)	0.3(n=10)	0.4(n=10)	0.5(n=10)	Avg(n=80)
0.094	0.301	0.673	0.851	0.912	0.926	0.52
(0.011)	(0.010)	(0.022)	(0.016)	(0.006)	(0.006)	(0.01)
Simulated Annealing Selected Sample						
			SNP MAF			
0.05(n=20)	0.1(n=20)	0.2(n=10)	0.3(n=10)	0.4(n=10)	0.5(n=10)	Avg(n=80)
0.208	0.535	0.808	0.899	0.932	0.945	0.63
(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.00)

Figure 5.1 Rapid Convergence of Simulated Annealing Algorithm Irrespective of Starting Sample

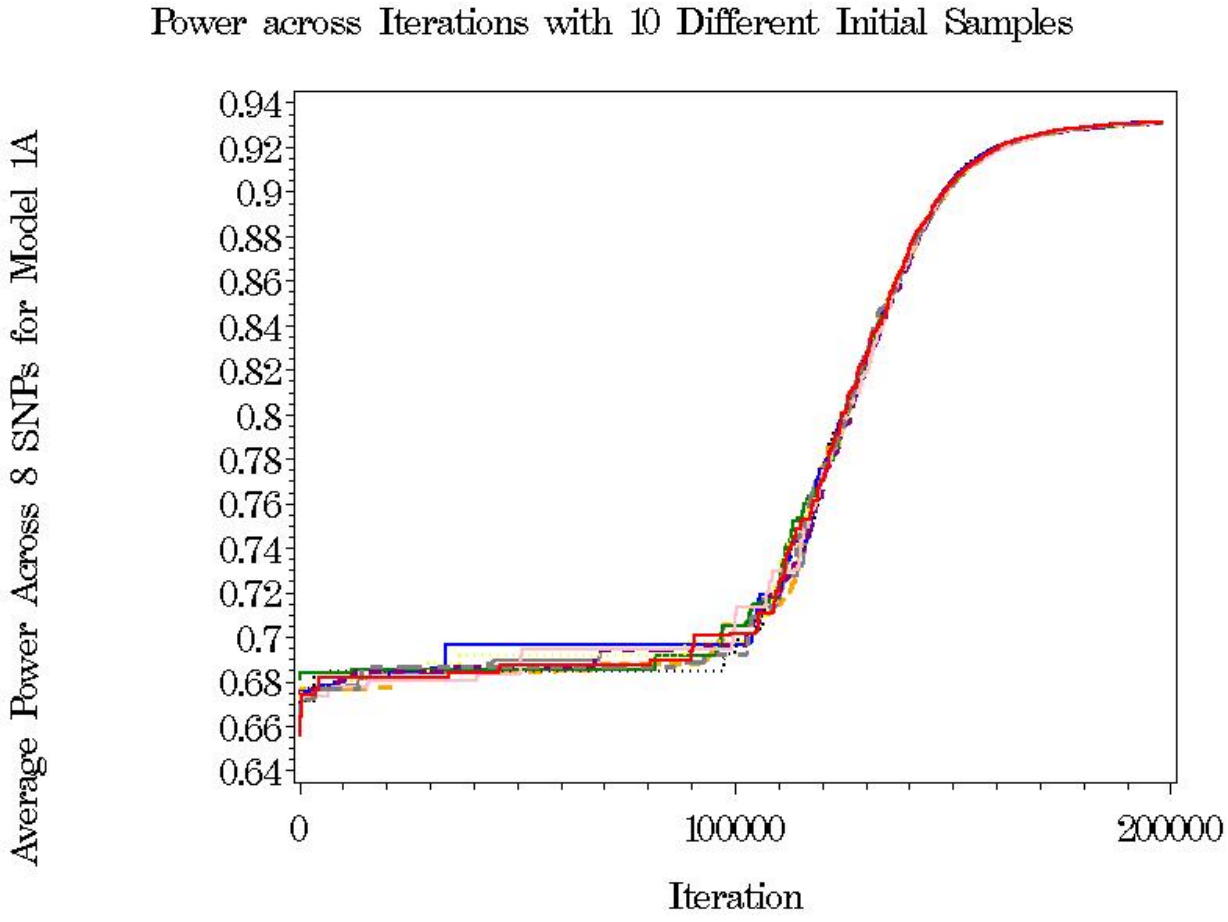


Table 5.4 Results for full sample (N=2983)

Variant	Chr	Position	MAF	Beta (β)	Se	p-value
rs1160985	19q13	50,095,252	0.37	-0.14	0.03	1.2×10^{-6}
rs12239267	1p13	159,670,928	0.04	0.35	0.04	2.4×10^{-22}
rs2393791	12p13	121,423,956	0.39	-0.058	0.03	0.039
rs6734238	2q13	113,557,501	0.47	0.072	0.03	0.0067
rs7748513	6p21	41,235,950	0.42	0.015	0.03	0.57

Table 5.5 Evidence for replication in JHS using random and selective phenotyping (100 samples, N=500)

Variant	Random sample		Simulated annealing sample	
	Mean p-value (std dev)	Proportion p<0.05	Mean p-value (std dev)	Proportion p<0.05
rs1160985	0.12 (0.18)	0.65	0.0081 (0.0068)	1.00
rs12239267	0.014 (0.066)	0.96	3.5×10^{-5} (7.0×10^{-5})	1.00
rs2393791	0.48 (0.3)	0.17	0.57 (0.27)	0.00
rs6734238	0.32 (0.28)	0.17	0.023 (0.026)	0.86

BIBLIOGRAPHY

Ambrosius, W. T., Lange, E. M., and Langefeld, C. D. (2004). Power for genetic association studies with random allele frequencies and genotype distributions. *Am. J. Hum. Genet.* **74**, 683-693.

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* **17**, 1103-1120.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781-791.

Barlow, R. E., Bartholomew, D. j., Bremner, J. M., and Brunk, H. D. (eds) (1972). *Statistical Inference under Order Restrictions*. London, New York: John Wiley & Sons.

Bekele, B. N., Ji, Y., Shen, Y., and Thall, P. F. (2008). Monitering late-onset toxicities in phase I trials using predicted risks. *Biostatistics* **9(3)**, 442-457.

Bekele, B. N., Li, Y., and Ji, Y. (2010). Risk-group-specific dose finding based on an average toxicity score. *Biometrics* **66**, 541-548.

Bekele, B. N. and Thall, P. F. (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association* **99**, 26-35.

Best, M. J. and Chakravarti, N. (1990). Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming* **47**, 425-439.

Bhatt, A. N., Mathur, R., Farooque, A., Verma, A., and Dwarakanath, B. S. (2010). Cancer biomarkers-current perspectives. *Indian J Med Res* **132**, 129-149.

Borneman, J., Chrobak, M., Vedova, G. D., Figueroa, A., and Jiang, T. (2001). Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics* **17(suppl1)**, S39-S48.

Braun, T. M. (2006). Generalizing the TITE-CRM to adapt for early- and late-onset toxicities. *Statistics in Medicine* **25**, 2071-2083.

Braun, T. M., Levine, J. E., and Ferrara, J. L. (2003). Determining a maximum tolerated cumulative dose: dose reassignment within the TITE-CRM. *Controlled Clinical Trials* **24**, 669-681.

Braun, T. M., Yuan, Z., and Thall, P. F. (2005). Determining a Maximum-Tolerated Schedule of a Cytotoxic Agent. *Biometrics* **61**, 335-343.

Burdakow, O., Grimwall, A., and Hussian, M. (2004). A generalized pav algorithm for monotonic regression in several variables. In *Compstat 2004 Symposium*761-767.

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput*, 368-379.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. 8(12):. *PLoS Comput Biol* **8(12)**, e1002822. doi-10.1371/journal.pcbi.1002822.

Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nat Rev Genet* **2**, 91-99.

Carlson, C. S., Eberie, M. A., Kruglyak, L., and Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446-452.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician* **46**, 167-174.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., and Hunter DJ. (2007). Replicating genotype-phenotype associations. *Nature* **447**, 655-660.

Chen, Z., Krailo, M. D., Azen, S. P., and Tighiouart, M. (2010). A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in Phase I clinical trials. *Contemporary Clinical Trials* **31**, 473-482.

Cheung, Y. K. (2005). Coherence principles in dose finding studies. *Biometrika* **92**, 863-873.

Cheung, Y. K. (2010). Stochastic approximation and modern model-based designs for dos-finding clinical trials. *Statistical Science* **25**, 191-201.

Cheung, Y. K. and Chappell, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56**, 1177-1182.

Cheung, Y. K. and Chappell, R. (2002). A simple technique to evaluate model sensitivity in the continual reassessment method. *Biometrics* **58**, 674.

Coghlan, A., Mac Donaill, D. A., and Buttimore, N. H. (2001). Representation of amino acids as five-bit or three-bit patterns for filtering protein databases. *Bioinformatics* **17(8)**, 676-685.

Darvasi, S. and Soller, M. (1992). Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor. Appl. Genet.* **85**, 353-359.

Derman, C. (1957). Nonparametric up and down experimentation. *Annals of Mathematical Statistics* **28**, 795-798.

Dunson, D. and Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics* **59**, 286-295.

Durham, S. D. and Flournoy, N. (1994). Random walks for quantile estimation. In *Statistical Decision Theory and Related Topics*, Berger, J. and Gupta, S. (eds), 467-476. New York: Springer-Verlag.

Eichhorn, B. H. and Zacks, S. (1973). Sequential search of an optimal dosage. *Journal of the American Statistical Association* **68**, 594-598.

Ezzalfani, M., Zohar, S., Qin, R., Mandrekar, S. J., and Le Deley, M. (2012). Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. *Statistics in Medicine* **32(16)**, 2728-2746.

Foster, M. C., Amin, C., Voorhees, P. M., van Deventer, H. W., Richards, K. L., Ivanova, A., Whitman, J., Chiu, W. M., Barr, N. D., and Shea, T. (2012). A phase I dose escalation study of clofarabine in combination with fractionated gemtuzumab ozogamicin in patients with refractory or relapsed acute myeloid leukemia. *Leukemia and Lymphoma* **53(7)**, 1331-1337.

Friedman, H.S., Kokkinakis, D.M., Pluda, J., Friedman, A.H., Cokgor, I., Haglung, M.M, Ashley, D. Rich, J. Dolan, M.E., Pegg, A., Moschel, R.C., Mclendon, R. E., Kerby, T., Herndon, J.E., B., D.D., and Schold, S. C. (1998). Phase I trial of O^6-benzylguanine for patients undergoing surgery for malignant glioma. *Journal of Clinical Oncology* **16**, 3570-3575.

Gagneur, J., Elze, M. C., and Tresch, A. (2011). Selective phenotyping, entropy reduction, and the mastermind game. *BMC Bioinformatics* **12**, 406.

Gelfand, A. E., Smith, A. F. M., and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association* **87**, 523-532.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (eds). *Bayesians Data Analysis*. London: Chapman and Hall.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721-741.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **57**, 97-109.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108.

Hochberg, Y. and Beniamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat Med* **9**, 811-818.

Hu, B., Bekele, B. N., and Ji, Y. (2013). Adaptive dose insertion in early phase clinical trials. *clinical trials* **10**, 216-224.

Hunsberger, S., Rubinstein, L. V., Dancey, J., and Korn, E. L. (2005). Dose escalation trial designs based on a molecularly targeted endpoint. *Statistics in medicine* **24**, 2171-2181.

Ivanova, A. and Kim, S. H. (2009). Dose finding for continuous and ordinal outcomes with a monotone objective function: a unified approach. *Biometrics* **65**, 307-315.

Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G., and Durham, S. D. (2003). Improved up-and-down designs for phase I trials. *Statistics in Medicine* **22**, 69-82.

Ivanova, A., Qaqish, B. F., and Schell, M. J. (2005). Continuous toxicity monitoring in phase I trials in oncology. *Biometrics* **61**, 540-545.

Ivanova, I., Flournoy, N., and Chung, Y. (2007). Cumulative cohort design for dose-finding. *Journal of Statistical Planning and Inference* **137**, 2316-2327.

Jannink, J. -. (2005). Selective phenotyping to accurately map quantitative trait loci. *Crop Science* **45(3)**, 901-908.

Ji, Y., Li, Y., and Bekele, B. N. (2007). Dose-finding in phase I clinical trials based on toxicity probability intervals. *Clinical Trials* **4**, 235-244.

Ji, Y., Liu, P., Li, Y., and Bekele, B. N. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical trials (London, England)* **7**, 653-663.

Ji, Y. and Wang, S. (2013). Modified toxicity probability interval design: a safer and more reliable method than the 3+3 design for practical phase I trials **31**, 1785-1791.

Jimenez-Sanchez, G., Childs, B., and Valle, D. (2001). Human Disease genes. *Nature* **409**, 853-855.

Jin, C., Lan, H., Attie, A. D., Churchill, G. A., Bulutuglo, D., and Yandel, B. S. (2004). Selective Phenotyping for Increased Efficiency in Genetic Mapping Studies. *Genet.* **168(4)**, 2258-2293.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, J. M. P. (1983). Optimization by simulated annealing. *Science* **220(4598)**, 671-680.

Kumar, M. and Sarin, S. K. (2009). Biomarkers of diseases in medicine. *Current trends in sciences*, 403-417.

Kuo, L. and Simith, A. F. M. (1992). Bayesian computations in survival models vi the gibbs sampler. *Survival Analysis: State of the Art*, 11-24.

Le Tourneau, C., Lee, J. J., and Siu, L. L. (2009). Dose escalation methods in phase I cancer clinical trials. *Journal of the National Cancer Institute* **101**, 708-720.

Lee, S. M., Cheng, B., and Cheung, Y. K. (2010). Continual reassessment method with multiple toxicity constraints. *Biostatistics* **12(2)**, 386-398.

Leung, D. H. and Wang, Y. G. (2001). Isotonic designs for Phase I trials. *Controlled Clinical Trials* **22**, 126-138.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311-321.

Lin, Y. and Shih, W. J. (2001). statistical properties of the modified algorithm-based designs for phase I cancer clinical trials. *Biostatistics* **2**, 203-215.

Lipsey, M. W. and Hurley, S. M. (1990). *Design sensitivity statistical power for applied experimental research*. London, UK: Sage publications.

Mandrekar, S. J., Cui, Y., and Sargent, D., J. (2007). An adaptive phase I design for identifying a biologically optimal dose for dual agent drug combinations. *Statistics in medicine* **26**, 2317-2330.

Mandrekar, S. J., Qin, R., and Sargent, D. J. (2009). Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: methods and challenges. *Statistics in Medicine* **29**, 1077-1083.

McManus, B. (2009). Trends in genomic biomarkers. *Heart Metab* **43**, 19-21.

Metropolis, N., Rosenberger, W. F., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1807-1091.

Mukarjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics* **16**, 741-750.

Neuenschwander, B., Branson, M., and Gsponer, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Statistics in Medicine* **27**, 2420-2439.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* **46**, 33-48.

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: a likelihood approach. *Biometrics* **52**, 673-684.

Penel, N., Adenis, A., Clisant, S., and Bonneterre, J. (2011). Nature and subjectivity of dose-limiting toxicities in contemporary phase 1 trials: comparison of cytotoxic versus non-cytotoxic drugs. *Invest New Drugs* **29**, 1414-1419.

Plummer, R., Jones, C., Middleton, M., Wilson, R., Evans, J., Olsen, A., Curtin, N., Boddy, A., McHugh, P., Newell, D., Harris, A., Johnson, P., Steinfeldt, H., Dewji, R., Wang, D., Robson, L., and Calvert, H. (2008). Phase I study of the poly(ADP-ribose) polymerase inhibitor, AG014699, in combination with temozolomide in patients with advanced solid tumors. *Clinical Cancer Research* **14**, 7917-7923.

Reiner, A. P., Beleza, S., Franceschini, N., Auer, P. L., Robinson, J. G., Kooperberg, C., Peters, U., and Tang, H. (2012). Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. *The Americal Journal of Human Genetics* **91**, 502-512.

Rich, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405**, 847-856.

Risch, N. and Merikangas, K. (1996). The future of genetics studies of complex human diseases. *Science* **273**, 1516-1517.

Rosenberger, W. F. and Haines, L. M. (2002). Competing designs for phase I clinical trials: a review. *Statistics in Medicine* **21**, 2757-2770.

Rosner, G. L. (2005). Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics* **61**, 239-245.

Schork, N. J., Fallin, D., Thiel, B., Xu, X., Broeckel, U., Jacob, H. J., and Cohen, D. (2001). The future of genetic case-control studies. *Adv Genet* **42**, 191-212.

Sempos, C. T., Bild, D. E., and Manolio, T. A. (1999). Overview of the Jackson Heart Study: a study of cardiovascular diseases in African American men and women. *Am J Med Sci* **317(3)**, 142-146.

Sen, S., Johannes, F., and Broman, K. W. (2009). Selective genotyping and phenotyping strategies in a complex trait context. *Genet.* **181**, 1613-2626.

Smith, A. F. M. (1993). Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society* **55**, 3-23.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45**, 925-937.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* **3**, 391-397.

van den Oord, E. J. (2008). Controlling false discoveries in genetic studies. *Am J Med Genet B Neuropsychiatr Genet* **147B**, 637-644.

Vasan, R. S. (2006). Biomarkers of cardiovascular disease: molecular basis and practical considerations. *Circulation* **113**, 2335-2362.

Wang, C., Chen, T. T., and Tyan, I. (2000). Designs for phase I cancer clinical trials with differentiation of graded toxicity. *Communications in Statistics-Theory* **29(5&6)**, 975-987.

Wang, H., Thomas, D. C., Peer, I., and Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genetic Epidemiology* **30**, 356-368.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society B* **25**, 1-48.

Whitehead, J. (2006). Using Bayesian decision theory in dose-escalation studies. In *Statistical Methods for Dose-finding Experiments*, Chevret, S. (ed), 149-171. Chichester, West Sussex, England ;Hoboken, NJ: John Wiley & Sons.

Xu, Z., Zou, F., and Vision, T. J. (2005). Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics* **170**, 401-408.

Yuan, Y. and Yin, G. (2009). Bayesian dose finding by jointly modelling toxicity and efficacy as time-to-event outcomes. *Appl. Statist.* **58**, 719-736.

Yuan, Z. and Chappell, R. (2004). Isotonic designs for phase I cancer clinical trials with multiple risk groups. *Clinical Trials* **1**, 499-508.

Yuan, Z., Chappell, R., and Bailey, H. (2007). The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics* **63**, 173-179.

Zollner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80**, 605-615.