Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2017 November ; 66(5): 1015–1030. doi:10.1111/rssc.12195.

Phase I Designs that Allow for Uncertainty in the Attribution of Adverse Events

Alexia Iasonos¹ and John O'Quigley²

¹Department of Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA

²Université Pierre et Marie Curie-Paris VI, Paris, France

Abstract

In determining dose limiting toxicities in Phase I studies, it is necessary to attribute adverse events (AE) to being drug related or not. Such determination is subjective and may introduce bias. In this paper, we develop methods for removing or at least diminishing the impact of this bias on the estimation of the maximum tolerated dose (MTD). The approach we suggest takes into account the subjectivity in the attribution of AE by using model-based dose escalation designs. The results show that gains can be achieved in terms of accuracy by recovering information lost to biases. These biases are a result of ignoring the errors in toxicity attribution.

Keywords

clinical trials; Phase I trials; dose limiting toxicity; dose finding algorithms; continual reassessment method; sequential monitoring

1 Introduction

The objective of a Phase I trial is to locate the maximum tolerated dose (MTD) at which the rate of drug related adverse events can be considered acceptable. Tracking and reporting the occurrence of adverse events is described in detail in most protocols (Sivendran et al 2014). The purpose is to be able to accurately quantify the impact of the investigational drug's dosage on the rate of observed adverse events. However, in oncology trials, separating out effects believed to be likely caused by the investigational drug from non drug related toxicities is a difficult process (Penel et al, 2011). Attribution can be inaccurate due to the potential presence of other competing factors, such as concomitant chemotherapy or medications, cumulative toxicities from previous treatment, disease progression, and/or various co-morbidities. Attribution is also challenging with targeted therapy where non-hematologic adverse events pre-dominate and common toxicities, such as fatigue, are frequent and possibly a result of the patients' underlying disease (Ratain, 2015). A retrospective review of attribution rates in randomized Phase III trials found that 50% of adverse events (AE) were reported as drug related on the placebo arm. Among those AEs

Supplemental Material:

Supplemental Material A includes the proofs for Section 4; Supplemental Material B includes supplemental Figures 1,2,3, and supplemental Tables 1 and 2.

reported more than once for the same patient, 36% changed in attribution over time (Hillman et al, 2010).

Attribution of AE is typically measured on a 5-tier system (not related, unlikely, possibly, probably, definitely; Table 1) but almost all Phase I designs rely on dose limiting toxicities (DLTs) which are assumed to be drug-related (binary: yes, no) and attribution is assumed to be observed without error. The statistical literature has examined the effect of different types and grades of toxicity, however current designs do not attempt to correct for uncertainty or error in toxicity attribution. Ordinal responses that represent the toxicity grade on a 0–5 scale (Van Meter et al 2011) or a toxicity burden that summarizes different types and grades of toxicities (Lee, Cheng, Cheung 2011) have been proposed but these do not address the problem of attribution error. Our proposed approach is different whereby investigators are no longer being asked to make a clear choice on the presence of drug-related AE or not, among serious AE. Instead, when the true attribution is uncertain, we propose that clinicians assign a score or a range of scores representing the probability of drug related toxicity in a given case. Based on all of the observed information, the investigator may provide an interval within which it can be assumed lies the probability of the observed AE being drug related. A single point might be considered a particular case of an interval and here the related work of Yuan, Chappell and Bailey (2007) is particularly relevant.

Instead of grouping uncertain categories into a binary response, which is the current practice, the proposed design allows for a range of scores that captures the uncertainty in assessment of attribution. A score in the scale from 0 to 1 reflects the uncertainty in attribution. Table 1 shows an example of how scores could correspond to the current attribution levels. For example, if we know that neutropenia is specific to a drug, then the score of serious neutropenia should be close to 1, counting as DLT. By contrast, if liver failure is suspected but not known for certain to be associated with a drug, an instance of liver failure might receive a score of 0.60. Patients who experience dermatologic toxicities could get a range of scores. For example, low scores (<0.30) would correspond to an AE being unlikely, mid-range scores ($0.30 < \text{probability} < 0.60$) to possibly, and high scores (probability > 0.65) to “probably” or “definitely” related (Table 1).

1.1 Motivation

Sherman et al (2011) suggest that there might have been over-reporting of serious adverse events (SAE), since there were no specific guidelines for drug causality (FDA, 1995). For this reason, the FDA issued an updated regulation that clarifies the definitions of adverse events (FDA, 2010). The updated FDA guidelines provide examples of the types of evidence that suggest causal relationship for investigators reporting a suspected adverse reaction. Determining causality of AE remains a difficult problem in Phase I trials, particularly when investigators need to judge on the basis of an isolated incident observed in a first in man study at the setting of advanced disease. There are clinical scenarios where it is beyond the clinician's ability to determine with high accuracy if a given toxicity is attributable to the study drug, especially early on in a trial and without access to the entire safety database (Sherman et al 2011). Eaton et al (2016) attempted to quantify this problem by estimating

the dose-toxicity curve using data from Phase I trials and Sharma and Ratain (2016) concluded that “a probability estimation would better support dose selection”.

2 Sources and types of error

In this section we provide a structure to the intuitive notions of the previous paragraph. The structure is centered around a framework in which the clinicians' uncertainty is expressed explicitly. Both the model and clinicians' uncertainty assessments can be updated over time in order to reflect the accumulated information on the safety profile of the investigational drug. For example, early on in a trial, when limited information is known about the experimental agent, the probability of drug related toxicity might appear higher than later on when more information has accumulated.

There are two types of errors that might occur in toxicity attribution: 1) when an investigator incorrectly attributes a SAE as non drug related, when in fact it is related to the experimental drug (Type A error) and 2) when an investigator attributes toxicity to an experimental drug when in fact it is due to other causes (Type B error). Zohar and O'Quigley (2009) assessed the effect of errors on the Continual Reassessment Method and the 3+3 design in a simulation study. They showed model based designs to be significantly more robust than the standard 3+3 design. Both types of error result in biases (Iasonos, Gounder et al, 2012). Type A error potentially results in overdosing patients, while Type B error increases the risk of recommending a sub-therapeutic dose for further studies. Let us introduce notation and define these quantities. We assume the trial consists of k ordered dose levels, $d_1, d_2, \dots, d_k$, and a total of n patients. The visited dose level for patient j is denoted as X_j , and the binary true drug-related toxicity outcome is denoted as Y_j , where $Y_j = 1$ indicates a DLT for patient j , and 0 indicates absence of a DLT. We assume Y_j is unobserved, and it is a latent variable for which we want to make inferences based on the observed data Z_j which denotes the classification of a drug related toxicity by the investigator. As in any Phase I trial, we define the MTD as the dose with an acceptable rate of DLTs, thus the quantity of interest is $\Pr(Y = 1|d_i)$, $i = 1, \dots, k$. If Y_j were known then the errors of Type A ($1 - \lambda_1(d_i)$) and Type B ($\lambda_2(d_i)$) can be observed and are written:

$$\lambda_1(d_i) = 1 - \Pr(Z_j = 0|Y_j = 1, d_i), \quad \lambda_2(d_i) = \Pr(Z_j = 1|Y_j = 0, d_i) \quad (1)$$

The observed rate of DLTs depends on the error rates and the true DLT rate, i.e.,

$$\Pr(Z_j = 1|x_j) = \lambda_1(x_j)\Pr(Y_j = 1|x_j) + \lambda_2(x_j)\{1 - \Pr(Y_j = 1|x_j)\}. \quad (2)$$

In principle, it is no more difficult conceptually dealing with both sources of error in any model. In practice this may involve additional assumptions and, in the case of estimation, some loss of precision due to the inclusion of extra unknowns. In our experience the main worry is through Type B errors, i.e., modifying the dose escalation as a result of counting certain toxicities as DLTs that, in fact, are not related to the treatment. Type A errors are

likely to be much smaller and, in many cases, may be considered to be negligible (Eaton, Iasonos et al 2016).

3 Simple models for the error rates

We denote the probability of a true toxicity at $X_j = x_j$ by $R(x_j)$ and the probability of an observed or recorded toxicity as $P(x_j)$ where,

$$\begin{aligned} R(x_j) &= \Pr(Y_j=1|X_j=x_j) \approx \psi(x_j, a), \quad x_j \in \{d_1, d_2, \dots, d_k\} \\ P(x_j) &= \Pr(Z_j=1|X_j=x_j) \approx \psi(x_j, b), \quad x_j \in \{d_1, d_2, \dots, d_k\} \end{aligned}$$

where $\psi(d_i, a)$ is a simple working model for the dose toxicity relation, often chosen to take the simple form, $\psi(d_i, a) = \beta_i^{\exp(a)}$, where $a \in (-\infty, \infty)$ is the unknown parameter, and β_i are the standardized units representing the discrete dose levels d_i , $\beta_i \in (0, 1)$. Since drugs are assumed to be more toxic at higher dose levels, $\psi(d_i, a)$ is assumed to be an increasing function of d_i . We would like these two quantities $\Pr(Y_j=1|d_i, Z_j=1)$, and, $\Pr(Y_j=0|d_i, Z_j=0)$, to be high and, at least as an approximation, we can relate these to the error rates via,

$$\Pr(Y_j=1|Z_j=1, X_j=x_j) = \lambda_1(x_j) \psi^{-1}(x_j, b) \psi(x_j, a) \quad (3)$$

$$\Pr(Y_j=0|Z_j=0, X_j=x_j) = \{1 - \lambda_2(x_j)\} \{1 - \psi(x_j, b)\}^{-1} [1 - \psi(x_j, a)] \quad (4)$$

We can relate the different rates of events through,

$$P(x_j) = \lambda_1(x_j) R(x_j) + \lambda_2(x_j) \{1 - R(x_j)\} \approx \lambda_1(x_j) \psi(x_j, a) + \lambda_2(x_j) \{1 - \psi(x_j, a)\} \quad (5)$$

As a consequence we “lose” one of our parameters, b , reducing the dimension of the problem. At the same time this would require additional structure for the risks R and P . The parameter estimate $\hat{a}$ is given by $\sum_{j=1}^n U_j(\hat{a}) = 0$ where;

$$U_j(a) = \frac{Z_j \{\lambda_1(x_j) - \lambda_2(x_j)\} \psi'(x_j, a)}{\lambda_1(x_j) \psi(x_j, a) + \lambda_2(x_j) \{1 - \psi(x_j, a)\}} - \frac{(1 - Z_j) \{\lambda_1(x_j) - \lambda_2(x_j)\} \psi'(x_j, a)}{1 - \lambda_1(x_j) \psi(x_j, a) - \lambda_2(x_j) \{1 - \psi(x_j, a)\}}$$

From earlier or similar studies we may have knowledge of the error rates. Moreover, the uncertainty in this knowledge might be framed within a Bayesian structure as illustrated in Section 3.2.

3.1 Dose escalation design and model based corrections

The difficulty is in finding useful and sufficiently reliable input for the error rates. Information from related studies may help and, indeed, as the study progresses, it is likely that we will refine any knowledge we have on the rates of errors at earlier points of the study. We limit our study here to the cases in which any true drug related toxicity will be signaled as a DLT ($\lambda_1 = 1$). The only kind of error we consider is when an adverse event is observed, deemed to be related to the drug but, in fact, is not related to the drug; $\lambda_2(d_j) = \Pr(Z_j = 1 | Y_j = 0, d_j)$. We would like for $1 - \lambda_2(d_j) = \Pr(Z_j = 0 | Y_j = 0, d_j)$ to be high, as close to one as possible, which would mean lack of errors.

In the presence of errors, we do not observe y_j , but instead we observe z_j . We then have, $R(x_j) = P(Y_j = 1 | x_j) = P(Y_j = 1 | Z_j = 1, x_j) P(Z_j = 1 | x_j)$, assuming $P(Y_j = 1 | Z_j = 0) = 0$. We now introduce specific scores, s_j that are provided by the clinicians and can be interpreted as their best assessment of the conditional probability $\Pr(Y_j = 1 | Z_j = 1, X_j = x_j)$. Our approach relies on clinicians providing data on s_j which is equivalent to asking clinicians how probable is it that an observed toxicity for patient j is truly drug related. Although likely to be only roughly approximated early in the study we can anticipate this to improve through learning. We might consider two opposing situations: the first where the investigator is overconservative and assigns scores, or intervals of scores, that are too high on average. The second is where they are too low on average. The first situation will lead to a biased estimate of the MTD in which the expected probability of toxicity at this MTD will be lower than what is anticipated. The converse holds for the situation in which the investigator is anti-conservative - the bias is then in the opposite direction. Our goal is to obtain an unbiased estimate of the MTD and this is possible by requiring restrictions on the distributions allowed for the scores whether presented as points or intervals. We describe this in Sections 4 and 5.

We observe the dose levels x_j and the outcomes; either $z_j = 0$, in which case we will take it that the unobserved $y_j = 0$; or $z_j = 1$, in which case the clinician is asked to provide further data in the form of s_j , where $0 < s_j < 1$, is the clinician's assessment of the probability of the toxicity being drug related. Our observations now look like a collection of pairs, (x_j, s_j) where for each patient x_j is the dose level and s_j is the probability assigned by the clinician to represent the uncertainty in toxicity attribution for patient j . Including this in the likelihood and simplifying it we obtain an expression that parallels that of Yuan, Chappell and Bailey (2005), specifically for n patients we obtain an estimating equation of the form,

$$\sum_{j=1}^n \mathcal{U}_j(a) = 0 \text{ where,}$$

$$\mathcal{U}_j(a) = z_j \left\{ s_j \frac{\psi'}{\psi}(x_j, a) + (1 - s_j) \frac{-\psi'}{1 - \psi}(x_j, a) \right\} + (1 - z_j) \left\{ \frac{-\psi'}{1 - \psi}(x_j, a) \right\} \quad (6)$$

Note that, in Equation 6, we could make a notational economy by allowing s_j to assume the value zero whenever z_j takes on the value zero. The z_j would then disappear from the above equation and the expression would appear simpler for computation. This is what we use:

$$u(s_j, x_j, a) = s_j \frac{\psi'}{\psi}(x_j, a) + (1 - s_j) \frac{-\psi'}{1 - \psi}(x_j, a). \quad (7)$$

However, from a conceptual viewpoint, as well as the viewpoint of large sample theory, it is advantageous to leave z_j in the above expression and maintain the restriction that $0 < s_j < 1$. Once the current estimate of $\hat{a}$ and $\hat{R}(d_i) = \psi(d_i, \hat{a})$ are calculated, the MTD is defined to be the dose $d_m \in \{d_1, \dots, d_k\}$, $1 \leq m \leq k$ such that, $d_m = \arg \min_{d_j} |\hat{R}(d_j) - \theta|$, $i = 1, \dots, k$, where $|\hat{R}(d_j) - \theta| = |\hat{R}(d_j) - \theta|$ and θ is an acceptable DLT rate.

We also provide a solution analytically of the estimating equation given in Equation 6 using an approximation and show that this approximation is very close to the actual solution we obtain with maximization. This along with available software (available from the first author) can obtain the next dose assignment in real time using the proposed approach. The second derivative of the log likelihood, i.e. derivative of $\mathcal{U}(a)$ in Equation 6, can provide a lower bound for the variance of $\hat{a}_j$. For the power model $\psi(d_i, a) = \beta_i^{\exp(a)}$ this expression is given by:

$$\mathcal{U}'_j(a) = \frac{\log \psi(x_j, a) [(s_j - \psi(x_j, a))(1 - \psi(x_j, a)) - (1 - s_j)\psi(x_j, a) \log \psi(x_j, a)]}{(1 - \psi(x_j, a))^2} \quad (8)$$

Using an initial value of $a_j^{(0)}$ by fitting the model to the mean response, which here is the average score along with the mean dose level used, we can solve for $\hat{a}_j$ by using the recursion as follows. Let $a_j^{(0)} = \{\log(\sum_{\ell=1}^{j-1} s_\ell / (j-1))\} / \{\log(\sum_{\ell=1}^{j-1} \beta_\ell / (j-1))\}$ be the starting value (O'Quigley and Shen, 1996). Then,

$$\hat{a}_j^{(m)} = \hat{a}_j^{(m-1)} - \sum_{\ell=1}^{j-1} \mathcal{U}_\ell(\hat{a}_j^{(m-1)}) v(\hat{a}_j^{(m-1)}) \quad (9)$$

where $\hat{a}_j^{(m)}$ is the m^{th} iteration for $\hat{a}_j$; and $v(a)$ is approximated by $-1 / \sum_{\ell=1}^{j-1} \mathcal{U}'_\ell(a)$. The solution of a given by Equation 9 for $m = 1$ is close ($\pm 10^{-3}$) to the solution we get if we optimize the likelihood given by Equation 6 showing that this simple estimator is as good as the estimator we would get via optimization. The solution of $\hat{a}_j$ given a set of data from j patients that includes visited dose levels and corresponding probabilities of DLT (scores) can provide investigators the assigned dose for a newly enrolled patient seamlessly in practice.

3.2 Dose escalation design based on intervals of uncertainty

Now let us assume that a more realistic representation is that an investigator provides an interval of uncertainty as opposed to a point. In this case, the uncertainty in attribution can be expressed as scores ranging from 0.70 – 0.90 for patient 1 for example, 0.50 – 0.70 for

patient 2 and so on. Both the interval and probability distribution are patient specific. We now solve the following equation:

$$\mathcal{U}_j^*(a) = \sum_{j=1}^n \left\{ z_j \int u(s_j, x_j, a) f_j(s) ds + (1 - z_j) \frac{-\psi'}{1 - \psi}(x_j, a) \right\} \quad (10)$$

where $u(s_j, x_j, a)$ is given in (7); $f_j(p)$ represents the patient-specific score distribution and p is the parameter vector that represents that distribution. For example, in the particular case where we assume each patient's scores comes from a Beta distribution with different shape and scale parameters for each patient say $B(g_j, h_j)$ where g_j, h_j are the shape and scale parameters respectively, then $f_j(s) = B(g_j, h_j)$ for patient j . To illustrate how the algorithm works suppose we use the simple one parameter model $\psi(x, a) = \beta_i^{\exp(a)}$ then Equation 10 becomes:

$$\mathcal{U}_j^*(a) = \sum_{j=1}^n \frac{\exp(a) \log(\beta_i)}{1 - \beta_i^{\exp(a)}} \left[z_j \int_{s_j^-}^{s_j^+} \frac{s}{c_j} B(g_j, h_j) ds - \beta_i^{\exp(a)} \right] \quad (11)$$

where $[s_j^-, s_j^+]$ is the range of scores specific to patient j , $B(g_j, h_j)$ is the distribution corresponding to this interval, and c_j is a normalizing constant.

4 Large Sample Theory

The following theoretical results ensure that the unknown parameter that models the dose-toxicity curve, a under certain conditions (shown in the Appendix), converges to the true population value almost surely. We build upon the work of Shen and O'Quigley (1996) to prove Theorem 1.

Lemma 1—Assume that conditions M_1, M_2, M_4 and M_5 are satisfied. Then S is an open and convex set, where $S = \{a : |\psi(x_0, a) - \theta| < |\psi(x_i, a) - \theta|, \text{ for all } x_i \neq x_0\}$, where x_0 is the target dose level.

Proof: See Shen and O'Quigley (1996).

Lemma 2—Define

$$I_{n2}(a) = \frac{1}{n} \sum [s_i - R(x_i)] \left[\frac{\psi'}{\psi}(x_i, t_{k0}) + \frac{-\psi'}{1 - \psi}(x_i, t_{k0}) \right]$$

For fixed t_{k0} , $\{nI_{n2} : n \geq 1\}$ forms a martingale and the terms in the summation in $I_{n2}(a)$ are bounded. $I_{n2}(a)$ converges to 0 almost surely.

Proof: See Supplemental A.

Lemma 3—Denote a_0 the solution to $u(s, x, a_0) = 0$. The solution of $u(s, x, a_0) = 0$ exists and it is unique for $\psi = \beta_i^{\exp(a)}$.

Proof: See Supplemental A.

Theorem 1—Assume conditions M_1 – M_7 are satisfied. For n large enough, let $\hat{a}_n$ be the maximum likelihood estimate of the parameter a , and $x(n+1)$ the recommended dose level for the next patient. Then almost surely, $\hat{a}_n \rightarrow a_0$ and $x(n+1) \rightarrow x_0$.

Proof: See Supplemental A.

These conditions are important in that they are sufficient to obtain useful large sample results such as consistency. Whether they are necessary, in other words can some of them be relaxed, is something that can be studied further. Most of these conditions provide a simple formulation of generally accepted hypotheses used in this context, for example M_1 , M_3 , M_4 , M_5 and M_6 . M_2 is a required condition under the theory of estimating equations and can be verified given any specific model. M_7 is a strong assumption that would only ever hold approximately, therefore it is important to investigate behavior under significant departures from the assumption. We do this in the next Section.

5 Simulation Studies

In order to assess how the proposed method performs under various error rates, we simulated hypothetical trials under various true dose toxicity curves and various parametric configurations and contrasted these with the ones obtained from conventional approaches where the error is ignored. Specifically, we simulated binary outcomes from a Bernoulli distribution with probability $R(d_i) = P(Y=1|d_i)$ denoting the true DLT probability. The outcome Y is unobserved in practice but here it will facilitate comparisons with the observed outcome Z . We compare the proposed method with regular CRM in the presence and absence of attribution errors (O'Quigley, Pepe, Fisher, 1990). We assume that investigators may erroneously flag adverse events as DLTs when in fact these are not drug-related. We evaluated the performance of various designs under various values of Type B error denoted as $\lambda_2 = \Pr(Z=1|Y=0) = 0.05, 0.10, 0.15$.

First, we assume for simplicity that $\lambda_1 = 1$ (i.e. $\Pr(Z=0|Y=1) = 0$), $\lambda_2 = 0$ in order to reflect the clinical scenario where true DLTs are not missed. Scores are assigned only for patients with $Z_j = 1$. Scores were generated from Uniform or Beta distribution with various means $\mu = 0.5, 0.7, 0.75, 0.85$ and variance parameters ($\sigma = 0.1, 0.2$). All simulation were performed using R (R version 3.1.1; <https://www.r-project.org/>). We evaluated scenarios where scores were sampled from the same distribution regardless of the dose level, i.e. we assume scores are constant across dose. These results are described in Section 5.1. We also evaluated scenarios where the distribution's mean varied across levels, with higher dose levels having a higher mean score to reflect the clinical scenario where higher doses result in more drug induced toxicity. These results are described in Section 5.2. For all scenarios, the following methods were compared:

1. proposed design using individualized uncertainty as described in Section 3.1

2. proposed design using intervals of uncertainty per patient allowing score variability as described in Section 3.2.
3. original CRM using the observed DLT outcome Z . This scheme helps us quantify the effect of errors on the dose toxicity parameter estimation.
4. original CRM using the latent variable Y . This scheme is used as a theoretical benchmark and it is based on the unobserved true DLT outcome.

5.1 Simulation studies with dose independent scores

Under the assumption that we do not know how far we are experimenting from the final MTD, having scores independent of dose is realistic. Scores are constant across dose level but vary per patient. The dose escalation algorithm follows the steps outlined in Section 3.1 and once we have solved the estimating equation for the unknown parameter and obtained the model estimated DLT rates we proceed like in any dose escalation design by assigning the next patient at the dose closest to a target rate. A target rate of $\theta = 0.2$ was used, sample size was fixed at $N = 25$ for testing $k = 6$ levels, and an initial vector of DLT probabilities known as a skeleton was specified as $(0.05, 0.1, 0.2, 0.3, 0.4, 0.7)$. We simulated 1000 trials with patients' scores, s_j , generated from a Uniform distribution with a mean score of $\mu = 0.5$, $0.7, 0.75, 0.85, \pm 0, 0.1, 0.2$, or generated from a Beta distribution with shape and scale parameters that corresponded to mean score of μ and σ .

Figure 1 shows a scenario with true DLT probabilities given by $0.01, 0.05, 0.07, 0.11, 0.20, 0.50$ (scenario 1) at each dose level respectively and scores are uniformly distributed from $U(0.55, 0.95)$. The results show that the percentage of trials in which we select the true MTD, level 5, increases from 43% (scheme 3) to 64% (scheme 1) compared to the dose we would have selected in the presence of attribution errors ($\lambda_2 = 0.05$). Note that scheme 4 denotes the theoretical benchmark under the assumption of no errors and it is not a scheme that can be selected to be used in practice. Scheme 2 is discussed in detail in section 5.3. We evaluated performance in various scenarios varying the location of the MTD and the true DLT rates. The true MTD is dose level d_5 in scenario 1 and is d_3 in scenario 2. The true DLT probabilities for scenario 2 are $0.07, 0.11, 0.23, 0.43, 0.84, 0.98$. Supplemental Figure 1 shows the results for scenario 2 when scores are uniformly distributed from $U(0.55, 0.95)$. We repeated all simulations with various levels of variability around each score $\pm 0, 0.10, 0.20$ and whether the patient specific score was centered around the mean or the score had some additional variability around its mean did not change the results (refer to Tables 2, 3 and Supplemental Tables 1, 2).

5.2 Simulation studies with dose dependent scores

There are clinical scenarios where investigators know from pre-clinical data that the trial starts at a very low dose thus treating patients far from the MTD. In this simulation setting, we allow investigators to assign different scores for each patient taking dose level into account. The scores vary depending on the dose the patient was treated at, with lower scores being given to lower doses, and higher scores be given to higher dose levels. The assigned scores were generated from Uniform distribution with $\mu = 0.10, 0.20, 0.30, 0.40, 0.50, 0.70$ at each dose level respectively, ± 0.20 (± 0.10 for d_1). The true scores depend on the true

underlying error rate, λ_2 , which is not known in practice (assuming $\lambda_1 = 0$) and true DLT probabilities, $R(x_j)$. For scenario 1, the true dose dependent scores are 0.17, 0.51, 0.60, 0.71, 0.83, 0.95 when $\lambda_2 = 0.05$; 0.09, 0.34, 0.43, 0.55, 0.71, 0.91 when $\lambda_2 = 10$; and 0.06, 0.26, 0.33, 0.45, 0.62, 0.87 when $\lambda_2 = 0.15$, for $d_1, \dots, d_6$ respectively. Figure 2 shows that the bias introduced by errors can be corrected by using the proposed approach. When investigators falsely attribute AE to the drug, i.e. $\lambda_2 > 0$, the MTD will be underestimated as shown by the bias, leading to greater recommendation of lower dose levels (scheme 3). The proposed method recommends the right MTD 86% of the time (scheme 1), as opposed to 42% under the original CRM (scheme 3) when the error rate is $\lambda_2 = 0.05$. The improvement here is greater compared to Figure 1, since giving low scores to low dose levels and higher scores to higher dose levels is a sensible approach given that clinicians learn about the drug's safety profile as the trial progresses. Scheme 4 denotes a method which is unrealistic in practice since it uses the unknown latent variable Y . The method that uses the latent variable however, provides a theoretical benchmark that helps quantify how well the proposed method performs. Supplemental Figures 2 and 3 show the results for scenario 2 when assigned scores increase as the dose increases and accuracy in finding the MTD is improved compared to ignoring errors.

5.3 Simulation studies with interval scores

In this section, we assume that investigators provide us with intervals of uncertainty as opposed to a single summary value. Scores were generated from a Beta distribution with shape and scale parameter that correspond to $\mu = 0.5, 0.7, 0.85$ and $\sigma = 0.04, 0.1, 0.15$. In order to simulate a range of scores, for each patient we first draw a random variable $s_j \sim B(g, h)$ that corresponds to a given μ, σ as above. We then find a corresponding Beta distribution denoted as $B(g_j, h_j)$, specific to patient j , such that the parameters g_j, h_j satisfy

$g_j = (s_j^2 - s_j^3 - s_j \sigma^2) / \sigma^2$, $h_j = g_j(1 - s_j) / s_j$. This ensures that the scores are well calibrated around the mean but we allow the interval to be narrow or wider (via σ^2). Once the solution to this estimating equation is found (Equation 11), $\hat{a}$, we proceed as above by assigning the next patient to the dose closest to θ . Tables 2 and 3 shows the comparison of the methods in terms of the percentage of trials that select each dose level using a point versus an interval under various error rates ($\lambda_2 = 0.05, 0.10, 0.15$) and $\lambda_2 = 0.1$ for two true dose toxicity scenarios. The two approaches of assigning an "uncertainty interval" as opposed to assigning a point score performed similarly in terms of accurately finding the MTD and while the point score approach is simpler, the interval approach allows investigators to more easily express their uncertainty in AE attribution. The results in Figure 1 and Figure 2 also show that the interval and point scores result in comparable performance.

5.4 Sensitivity Analysis

In order to assess the sensitivity of the proposed approach to the score assignment, in this section we ran simulations where the assigned scores are systematically high or systematically low at all dose levels. As expected if scores are systematically very low, then it is as though we had seen less toxicity in practice, and since less toxicity will lead to higher levels, we then tend to overdose (Supplemental Figure 2 when there is no error). On the other hand, systematically high scores might lead to treating patients with subtherapeutic

dose levels, since we indicate potential DLTs when in fact they are not, shifting experimentation to lower levels (Figure 3). However, in the absence of systematic bias, we found no cases where the proposed method using scores performs worse than the current practice of possibly erroneously attributing AE to the drug. Generally across many scenarios examined the new method leads to considerable improvement in the accuracy of determining the MTD, as long as the scores are well calibrated (on average across patients the interval contains the true DLT rate). When scores are systematically biased, for example underestimating the DLT risk, the proposed approach will tend to select a higher dose as the MTD more likely compared to a lower dose especially if the true error rate approaches 0 ($\lambda_2 = 0$). Finally, although not common, it is possible that the proposed method performs better than the theoretical benchmark. The reasons are complex and relate to the fact that the input data are continuous instead of binary. This also relates to super-optimality (O'Quigley J, Paoletti X, Maccario J, 2002; O'Quigley J, 2006).

6 Discussion

An important aspect of Phase I trials in oncology is that the patient population has advanced disease, which makes it difficult for clinicians to distinguish between drug related toxicity and disease deterioration (Penel et al 2011). In the interest of patient safety, it is not uncommon for investigators to attribute an AE to the study drug when it is unclear whether the drug is in actuality the underlying cause (Mukherjee et al 2011). Errors in toxicity attribution is a topic that has been neglected in the statistical literature and current Phase I designs assume that the DLT outcome is measured without error (Pan, Zhu et al 2014). Our proposed design that allows patient specific scores for AE attribution provides a conceptual link between the clinical challenges of Phase I trials and the statistical complexity of recent model based dose escalation algorithms. Our work focused on the single agent setting but the proposed approach has the potential to be extended to the context of drug combinations (Wages 2016, Wheeler et al. 2016). For example, there are cases where one of the components of a combination is known to be associated with some particular AEs or type of toxicities such as dermatologic, while the other is not. These algorithms allow for dynamic learning as the trial progresses and the added contribution of the proposed design is to extend this dynamic learning to include knowledge from earlier studies together with the clinical investigators expertise. In some cases, the investigators may be close to certain on correct attribution. In other cases, they will be less certain and our purpose is to explicitly address this fact. The investigators' input tailored to the individual patients can enable more accurate dose finding and greater flexibility in escalation while still maintaining a rigorous adherence to protocol definitions.

Phase I designs should incorporate a level of uncertainty in toxicity attribution, since in the majority of these trials the agents mechanism of action is not fully understood. In the proposed framework, the protocol team, which includes the investigators, the disease management team, or the Data Safety Monitoring Board, can assign a score representing the likelihood of an event being drug related. Petroni et al. (2016) recently discussed the logistics and the timeline involved in reviewing safety data from Phase I trials which would apply here. The scores can describe numerically existing verbal descriptors such as possibly, probably or unlikely (Table 1). These descriptors will map onto separate scores or a group of

scores. This mechanism will allow for individual, patient-specific scores as well as dynamic scores. For example, knowledge based on late-onset toxicities observed in the post-DLT monitoring period on previously enrolled patients can be incorporated when assigning scores for new patients, and hence the dose escalation algorithm will reflect accumulating information on the drug's complete toxicity profile. Patients can be assigned a variety of scores: scores that vary by dose level or patient inclusion number so that they reflect accumulated information throughout the trial. The investigator may wish to err on the side of caution, especially early on in the trial. For example for the first 10 enrolled patients, the scores could be similar given the limited safety information available, while for the next 10 patients the scores can vary depending on the dose level the patient received and the accumulated safety data.

While Type A errors might be less common than type B, it could be the case that Type A errors reduce the bias brought on by any Type B errors made (Iasonos et al, 2012). The assumption that all true drug related toxicities show up seems reasonable and this enables us to consider the Type A error as negligible. There may be cases where this assumption would require closer scrutiny. Time dependent toxicities come under this heading where there is late onset of drug related toxicities. Multi-center trials may be another source of potential Type A error in which the relevant toxicity information is slow coming through and a DLT may be classified incorrectly at least for some part of the study. These are topics for future research. Another situation is when there is great heterogeneity among patients' disease status. An observed DLT for one patient may carry more weight than a similar DLT on a patient with more advanced disease. These clinical scenarios are challenging in terms of distinguishing drug related from disease related adverse events and this is a topic of ongoing work (Sharma and Ratain, 2016).

We showed that a method that uses a personalized score instead of a binary DLT outcome can lead us accurately to the MTD because if the scores are well calibrated, on average we expect patients with high probabilities of DLTs to correlate with an actual presence of DLT and similarly low scores to correlate with absence of DLT. For example, if the clinicians' scores have no systematic variability and agree with the presence of DLTs, then we expect the proposed method to be more accurate than using a binary outcome with misclassification errors. Systematic biases however, as expected, will lead to under estimation or overestimation of the MTD when bias here is defined as systematically assigning low scores to true DLT or very high scores to non DLT's throughout the trial. The scores can be inexact for every patient but the method's accuracy will not be unduly affected as long as the scores are well calibrated and not systematically biased. This novel conceptual framework incorporates the subjectivity involved in toxicity attribution into Phase I designs, such that the final recommended dose reflects this uncertainty. It also allows the investigators to strictly adhere to the protocol by recording all DLTs and, yet, still allow expert opinion to prevent AE that are most likely not drug related from seriously compromising the exercise of identifying the correct MTD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially funded by the National Cancer Institute (Grant Number R01 CA142859-04A1) and The Transnational and Integrative Medicine Research Fund at Memorial Sloan Kettering Cancer Center, NY.

References

- Eaton A, Iasonos A, Gounder, et al. Toxicity Attribution in Phase I Trials: Evaluating the Effect of Dose on the Frequency of Related and Unrelated Toxicities. *Clin Cancer Res*. 2016; 22(3):553–9. [PubMed: 26324741]
- FDA. Guideline for industry: clinical safety data management: definitions and standards for expedited reporting. 1995. <http://www.fda.gov/downloads/Drugs/.../Guidances/ucm073087.pdf>
- FDA. Investigational new drug safety reporting requirements for human drug and biological products and safety reporting requirements for bioavailability and bioequivalence studies in humans. Final rule. *Fed Regist*. 2010; 75(188):59935–63. [PubMed: 20879180]
- Hillman SL, Mandrekar SJ, Bot B, et al. Evaluation of the value of attribution in the interpretation of adverse event data: a North Central Cancer Treatment Group and American College of Surgeons Oncology Group investigation. *Journal of Clinical Oncology*. 2010; 28(18):3002–7. [PubMed: 20479400]
- Iasonos A, Gounder M, Spriggs DR, et al. The impact of non-drug-related toxicities on the estimation of the maximum tolerated dose in phase I trials. *Clin Cancer Res*. 2012; 18(19):5179–87. [PubMed: 22825582]
- Lee SM, Cheng B, Cheung YK. Continual reassessment method with multiple toxicity constraints. *Biostatistics*. 2011; 12(2):386–98. [PubMed: 20876664]
- Mukherjee SD, Coombes ME, Levine M, Cosby J, Kowaleski B, Arnold A. A qualitative study evaluating causality attribution for serious adverse events during early phase oncology clinical trials. *Investigational New Drugs*. 2011; 29(5):1013–20. [PubMed: 20512397]
- O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*. 1990; 46:33–48. [PubMed: 2350571]
- O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. *Biometrics*. 1996; 52:673–84. [PubMed: 8672707]
- O'Quigley J, Paoletti X, Maccario J. Non-parametric optimal design in dose finding studies. *Biostatistics*. 2002; 3(1):51–56. [PubMed: 12933623]
- O'Quigley J. Theoretical study of the continual reassessment method. *Journal of Statistical Planning and Inference*. 2006; 136:1765–1780.
- Pan H, Zhu C, Zhang F, Yuan Y, Zhang S, et al. The Continual Reassessment Method for Multiple Toxicity Grades: A Bayesian Model Selection Approach. *PLoS ONE*. 2014; 9(5):e98147. [PubMed: 24875783]
- Penel N, Adenis A, Clisant S, Bonnetterre J. Nature and subjectivity of dose-limiting toxicities in contemporary phase 1 trials: comparison of cytotoxic versus non-cytotoxic drugs. *Investigational New Drugs*. 2011; 29(6):1414–9. [PubMed: 20617363]
- Petroni GR, Wages NA, Paux G, Dubois F. Implementation of adaptive methods in early-phase clinical trials. *Stat Med*. 2016 Epub ahead of print.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. <http://www.R-project.org/>
- Ratain MJ. Redefining the primary objective of phase I oncology trials. *Nat Rev Clin Oncol*. 2015; 12(3):126.
- Sharma MR, Ratain MJ. Taking a Measured Approach to Toxicity Data in Phase I Oncology Clinical Trials. *Clin Cancer Res*. 2016; 22(3):527–9. [PubMed: 26467385]
- Shen L, O'Quigley J. Consistency of continual reassessment method under model misspecification. *Biometrika*. 1996; 83(2):395–405.
- Sherman RB, et al. New FDA regulation to improve safety reporting in clinical trials. *N Engl J Med* 2011. 2011; 365(1):3–5.

- Sivendran S, Latif A, McBride RB, et al. Adverse event reporting in cancer clinical trial publications. *Journal of clinical Oncology*. 2014; 32(2):83–9. [PubMed: 24323037]
- Van Meter EM, Garrett-Mayer E, Bandyopadhyay D. Proportional odds model for dose-finding clinical trial designs with ordinal toxicity grading. *Stat Med*. 2011; 30(17):2070–80. [PubMed: 21344472]
- Wages NA. Identifying a maximum tolerated contour in two-dimensional dose finding. *Stat Med*. 2016 Epub ahead of print.
- Wheeler GM, Sweeting MJ, Mander AP, Lee SM, Cheung YK. Modelling semi-attributable toxicity in dual-agent phase I trials with non-concurrent drug administration. *Stat Med*. 2016 Epub ahead of print.
- Yuan Z, Chappell R, Bailey H. The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics*. 2007; 63(1):173–9. [PubMed: 17447942]
- Zohar S, O'Quigley J. Sensitivity of dose-finding studies to observation errors. *Contemp Clin Trials*. 2009; 30(6):523–30. [PubMed: 19580886]

Appendix

Conditions needed in Section 4

Assume the following conditions are satisfied. Namely,

M_1 : For each a , function $\psi(\cdot, a)$ is strictly increasing and it is continuous and strictly monotone in a in the same directions for all x .

M_2 : The function $u(s, x, a) = s \frac{\psi'}{\psi}(x, a) + (1-s) \frac{-\psi'}{1-\psi}(x, a)$, for each $0 < s < 1$ and each x , is continuous and strictly monotone in a .

M_3 : The parameter a belongs to finite interval $[A, B]$.

M_4 : The target dose level is x_0 , that is $R(x_0) = \theta$.

M_5 : The probabilities of toxicity at $x_1, \dots, x_k$ satisfy $0 < R_1 < \dots < R_k < 1$.

M_6 : For $i = 1, \dots, k$, $a_i \in S$ where $S = \{a : |\psi(x_0, a) - \theta| < |\psi(x_i, a) - \theta|, \text{ for all } x_i \neq x_0\}$.

M_7 : For every dose level i the expected value of assigned scores is $E(s_i) = R(x_i)$.

Figure 1.

Percentage of trials out of 1000 simulated trials (y axis) that select each dose level among 6 levels (x axis). True DLT probabilities are 0.01, 0.05, 0.07, 0.11, 0.20, 0.50 (scenario 1) at each dose level respectively and scores are dose independent but vary per patient (generated from Uniform with mean $\mu = 0.75$, ± 0.20 regardless of dose). Various panels show different error rates (0.15, 0.10, 0.05, 0). Scheme 1 (purple) uses a point score; Scheme 2 (green) uses interval scores; Scheme 3 (red) uses a binary DLT (yes/no); Benchmark (blue) uses the latent variable Y and assumes no errors exist.

Figure 2.

Percentage of trials out of 1000 simulated trials (y axis) that select each dose level among 6 levels (x axis). True DLT probabilities are 0.01, 0.05, 0.07, 0.11, 0.20, 0.50 (scenario 1) at each dose level respectively and scores are dose dependent (scores are generated from Uniform with mean $\mu = 0.10, 0.20, 0.30, 0.40, 0.50$ and 0.70 at each dose level respectively with ± 0.20 variability from patient to patient; except for d_1 which has ± 0.10 variability). Various panels show different error rates (0.15, 0.10, 0.05, 0). Scheme 1 (purple) uses a dose-dependent point score; Scheme 2 (green) uses dose-dependent interval scores; Scheme 3 (red) uses a binary DLT (yes/no); Benchmark (blue) uses the latent variable Y and assumes no errors exist.

Figure 3.

Percentage of trials (y axis) that select each dose level among 6 levels (x axis). True DLT probabilities are 0.01, 0.05, 0.07, 0.11, 0.20, 0.50 (scenario 1) at each dose level and scores vary depending on dose level but they are systematically overestimating DLT risk, i.e. scores induce bias by systematically assigning higher scores compared to true rates. Various panels show different error rates (0.15, 0.10, 0.05, 0). Scheme 1 (purple) uses a dose-dependent point score; Scheme 2 (green) uses dose-dependent interval scores; Scheme 3 (red) uses a binary DLT (yes/no); Benchmark (blue) uses the latent variable Y and assumes no errors exist.

Table 1

Illustrative example of scores for drug related toxicities 0: non DLT; 1: DLT.

Attribution	Hypothetical Score	Current designs: DLT yes/no
Unrelated	0	No (0)
Unlikely	0 to <0.30	No (0)
Possibly	0.30 to <0.65	Yes (1)
Probably	0.65 to <1	Yes (1)
Definitely	1	Yes (1)

Table 2

Percentage of trials selecting each dose level under four schemes for scenario 1 (true MTD is d_5^* , $R(d_i^*) = 0.01, 0.05, 0.07, 0.11, 0.2, 0.5, i = 1, \dots, 6$); μ denotes mean score; $\sigma = 0.10$. Scheme 1: point score; Scheme 2: interval scores; Scheme 3: CRM using Z ; Scheme 4: CRM using Y .

	$\mu = 0.85$						$\mu = 0.70$						$\mu = 0.50$					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
$\lambda_2=0.15$																		
Scheme 1	17	20	26	22	14	0	7	18	22	25	28	0	1	6	16	20	53	4
Scheme 2	16	20	25	23	16	0	7	17	21	27	28	0	1	6	16	20	54	4
Scheme 3	20	24	25	20	11	0	21	23	24	19	12	0	22	22	25	19	12	0
Scheme 4	0	0	6	26	67	2	0	0	6	26	66	1	0	1	6	26	65	2
$\lambda_2=0.1$																		
Scheme 1	4	14	22	27	32	1	2	9	16	23	48	0	0	2	7	15	67	9
Scheme 2	4	14	21	27	34	1	0	8	15	24	49	2	0	2	6	16	68	8
Scheme 3	8	15	25	29	22	0	8	17	22	29	23	0	8	15	23	28	25	0
Scheme 4	0	0	6	26	67	1	0	0	5	28	65	2	0	0	5	26	67	2
$\lambda_2=0.05$																		
Scheme 1	0	5	13	26	54	2	0	3	8	19	66	5	0	0	2	7	72	18
Scheme 2	1	4	12	25	56	2	0	3	7	17	68	6	0	0	2	7	72	18
Scheme 3	1	7	18	32	41	1	1	8	17	33	41	1	0	8	16	32	42	0
Scheme 4	0	0	0	6	67	2		0	5	26	67	1	0	0	6	23	69	1

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Percentage of trials selecting each dose level under four schemes for scenario 2 (true MTD is d_3 , $R(d_i) = 0.07, 0.11, 0.23, 0.43, 0.84, 0.98, i = 1, \dots, 6$); μ denotes mean score; $\sigma = 0.10$. Scheme 1: point score; Scheme 2: interval scores; Scheme 3: CRM using Z ; Scheme 4: CRM using Y .

$\mu = 0.85$													$\mu = 0.50$					
	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6	d_1	d_2	d_3	d_4	d_5	d_6
$\lambda_2=0.15$																		
Scheme 1	41	39	19	2	0	0	23	36	35	6	0	0	3	18	49	29	1	0
Scheme 2	39	37	22	2	0	0	22	36	36	6	0	0	3	18	50	28	0	0
Scheme 3	54	34	11	1	0	0	55	33	12	1	0	0	53	34	13	1	0	0
Scheme 4	4	31	56	10	0	0	4	31	56	9	0	0	4	32	54	10	0	0
$\lambda_2=0.1$																		
Scheme 1	23	41	32	4	0	0	10	32	47	12	0	0	1	9	47	43	1	0
Scheme 2	21	41	34	4	0	0	9	32	48	12	0	0	1	9	49	41	1	0
Scheme 3	33	43	23	1	0	0	34	39	25	2	0	0	34	42	23	2	0	0
Scheme 4	3	31	56	10	0	0	4	31	56	9	0	0	4	32	55	9	0	0
$\lambda_2=0.05$																		
Scheme 1	9	35	48	9	0	0	3	22	57	19	0	0	0	4	39	56	2	0
Scheme 2	8	34	48	10	0	0	2	20	58	20	0	0	0	4	38	56	2	0
Scheme 3	14	43	38	5	0	0	15	44	37	4	0	0	15	44	37	4	0	0
Scheme 4	3	33	55	9	0	0	4	32	56	8	0	0	4	34	54	8	0	0