Statistical Tutorial for Using Bayesian Optimal Interval (BOIN) Design for Phase I Oncology Trials

Ying Yuan* and Suyu Liu
Department of Biostatistics
The University of Texas MD Anderson Cancer Center
Houston, Texas 77030, U.S.A.
**email: yyuan@mdanderson.org*

July 22, 2015

1 Introduction

The Bayesian optimal interval (BOIN) design is a novel phase I trial design for finding the maximum tolerated dose (MTD). The BOIN design is motivated by top priority and concern of clinicians, which is to effectively treat patients and minimize the chance of exposing them to subtherapeutic or overly toxic doses. The most important advantage of the BOIN design is that it is easy to implement in practice and also has superior operating characteristics. The BOIN design is algorithm-based and thus can be implemented in a way similar to the traditional "3+3" design. The performance of the BOIN design is comparable to the continual reassessment method (CRM, one of the best performed model-based phase I trial designs) in terms of selecting the MTD, but has a lower risk of assigning patients to subtherapeutic or overly toxic doses (i.e., better patient ethics).

The idea behind the BOIN design is straightforward. The conduct of phase I trials can be essentially viewed as a sequence of decision-making steps of dose assignment for patients who are sequentially enrolled into the trial. At each moment of decision making, based on the observed data, we take one of three actions: escalate, deescalate or retain the current dose. Under the standard assumption that efficacy monotonically increases with toxicity (for cytotoxic agents), an ideal trial design would escalate the dose when the current dose is below the MTD in order to avoid treating a patient at a subtherapeutic dose level; deescalate the dose when the current dose is above the MTD in order to avoid exposing a patient to an overly toxic dose; and retain the same dose level when the current dose is equal (or close) to the MTD. However, such an ideal design is not available in practice because we do not know whether the current dose is actually below, above or equal (or close) to the MTD, and need to infer that information and make decisions based on the data collected from patients who have been enrolled and treated in the trial. Due to the randomness of the observed data and small

sample sizes of phase I trials, the decisions of dose assignment we make are often incorrect, e.g., we may erroneously escalate (or deescalate) the dose when it is actually higher (or lower) than the MTD, which results in overly aggressive (or conservative) dose assignments and treating excessive numbers of patients at dose levels above (or below) the MTD. From a practical and ethical viewpoint, it is highly desirable to minimize these decision errors, such that the actual design behaves as closely as possible to the ideal (error-free) design. The BOIN design is proposed to achieve this goal.

The BOIN design possesses sound theoretical properties. It is *long-term memory coherent* in the sense that the probability of dose escalation (or deescalation) is zero when the observed toxicity rate $\hat{p}_j$ at the current dose is higher (or lower) than the target toxicity rate ϕ. It has a convergence property similar to that of the continual reassessment method (CRM), and converges almost surely, at a $\sqrt{n}$ rate, to exclusive allocations of the target dose. The numerical study evaluating the design's performance shows that the BOIN design has superior operating characteristics that are comparable to or better than those of the CRM, but which are much easier to implement and has a substantially lower risk of assigning patients to subtherapeutic or overly toxic doses.

The BOIN design is very simple to implement in practice. In this design, dose transition is defined by the relative location of the observed toxicity rate (i.e., the number of patients who experienced toxicity divided by the total number of patients treated) at the current dose with respect to a prespecified toxicity tolerance interval. If the observed toxicity rate is located within the interval, we retain the current dose; if the observed toxicity rate is greater than the upper boundary of the interval, we deescalate the dose; and if the observed toxicity rate is smaller than the lower boundary of the interval, we escalate the dose.

In other words, to use the BOIN design, we need to specify only the interval (or dose escalation/deescalation) boundaries at the trial design phase, as they are the only design parameters that control dose escalation/deescalation. During the trial conduct, no additional software software is needed, and clinicians can simply count the number of patients who experience toxicity and compare the observed toxicity rate with the prespecified dose escalation/deescalation boundaries to determine dose assignment until the trial is completed. In the section that follows, we describe how to obtain the dose escalation/deescalation boundaries using the R package we have provided.

2 Software

The R package "BOIN" is freely available from CRAN. It contains three functions for implementing the proposed optimal designs:

- `get.boundary(···)`; This function is used to generate escalation and deescalation boundaries for the optimal interval design;

- `select.mtd(···)`; This function is used to select the MTD at the end of the trial based on isotonically transformed estimates;
- `get.oc(···)`; This function is used to generate operating characteristics for the proposed trial designs.

As an example, suppose we want to conduct a phase I trial with $J = 5$ dose levels and a target toxicity rate of $\phi = 0.3$. The maximum sample size is 30 patients, and patients are treated in cohort sizes of 3. The BOIN design allows the cohort size to vary from one cohort to another.

2.1 Trial design

To design the trial, we only need to run the function `get.boundary(·)` to get the dose escalation and deescalation boundaries, which are all we need to run the trial. This function has eight arguments:

- `target` the target toxicity rate
- `ncohort` the total number of cohorts
- `cohortsize` the cohort size
- `n.earlystop` early stopping parameter. If the number of patients treated at the current dose reaches `n.earlystop`, early stop the trial and select the MTD based on the observe data. The default value of `n.earlystop = 100` essentially turns off this type of early stopping.
- `p.saf` the highest toxicity probability that is deemed subtherapeutic (i.e., below the MTD) such that dose escalation should be made. The default value of `p.saf` = `0.6` $\times$ `target`.
- `p.tox` the lowest toxicity probability that is deemed overly toxic such that dose deescalation is required. The default value of `p.tox = 1.4` $\times$ `target`.
- `design` to use the local optimal design (the default, `design = 1`) or the global optimal design (`design = 2`). We generally recommend the local optimal design.
- `cutoff.eli` the cutoff to eliminate the overly toxicity dose for safety. We recommend the default value `cutoff.eli = 0.95` for general use.
- `extrasafe` set `extrasafe = TRUE` to impose a more strict stopping rule.

- `offset` a small positive number (between 0 and 0.5) to control how strict the stopping rule is when `extrasafe = TRUE`. A larger value leads to a more strict stopping rule. The default value `offset = 0.05` generally works well.

- `print` to print out the boundary results.

In practice, we should avoid setting the values of `p.saf` and `p.tox` very close to `target`. This is because the small sample sizes of typical phase I trials prevent us from being able to discriminate the target toxicity rate from the rates close to it. For example, at the significance level of 0.1, there is only 7% power to distinguish 0.25 from 0.35 with a total of 30 patients given just two doses. In addition, in most clinical applications, the target toxicity rate is often a rough guess, and finding a dose level with toxicity rates reasonably close to the target rate selected by the investigator will still prove to be of interest to the investigator. Based on our experience with phase I oncology trials, and also the operating characteristics of the proposed design, we find that $\phi_1 \in [0.5\phi, 0.7\phi]$ and $\phi_2 \in [1.3\phi, 1.5\phi]$ are reasonable values for most clinical applications. As default values, we recommend $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$ (i.e., 40% deviation from the target) for general use. Although this looks like a large deviation from the target, under typical phase I sample sizes, 40% deviation from the target actually is a rather small difference and the power to detect such a difference is quite limited. For example, if $\phi = 0.25$, then $\phi_1 = 0.15$ and $\phi_2 = 0.35$. Given a sample size of 30 patients and only two doses, we only have 7% power to distinguish 0.25 from 0.35, and 9% to distinguish 0.25 from 0.15, based on Fisher's exact test with a significance level of 0.1.

Although the function allows the user to specify whether the trial will use the local BOIN design or global BOIN design (based on different optimization criteria) by setting the value of `design`, the default is the local BOIN design (i.e., `design = 1`) because our experience has shown that it has better operating characteristics and also easier to implement (see Liu and Yuan, 2015 for details).

The BOIN design has two built-in stopping rules: (1) stop the trial if the lowest dose is eliminated due to toxicity. In this case, no dose should be selected as the MTD; and (2) stop the trial and select the MTD if the number of patients treated at the current dose reaches `n.earlystop`. The first stopping rule is a safety rule to protect patients from the case that all doses are overly toxic.

The rationale for the second stopping rule is that when there are a large number (i.e., `n.earlystop`) of patients assigned to a dose, it means that the dose finding algorithm has approximately converged. Thus, we can early stop the trial and select the MTD to save the sample size and reduce the trial duration. The default value `n.earlystop = 100` essentially turns off this type of early stopping. Note that by setting `n.earlystop` at a value like 12 can potentially save sample size and finish the trial early, the trade-off is that it may affect the selection percentage and decrease the rate of safety stopping if the first dose is overly toxic.

The value of `n.earlystop` should be calibrated by simulation to obtain desirable operating characteristics. In generally, we recommend `n.earlystop = 9` to `18`. Our experience is that this stopping rule is particularly useful when there is strong prior knowledge that the first dose is safe since a major side effect of using the stopping rule is that it decreases the rate of safety stopping when the first dose is actually overly toxic.

Although the BOIN design has a built-in safety stopping rule (i.e., the stopping rule (1) described above), for some applications, investigators may prefer a more strict stopping rule for extra safety when the lowest dose is possibly overly toxic. Setting `extrasafe = TRUE` imposes the following stronger stopping rule:

> stop the trial if (1) the number of patients treated the lowest dose ≥ 3, and (2) Pr(toxicity rate of the lowest dose $>$ `target` | data) $>$ `cutoff.eli - offset`.

Note that as the tradeoff, a more strict stopping rule will decrease the selection percentage of the MTD when the lowest dose actually is the true MTD. When use the option `extrasafe = TRUE`, we recommend the default value `offset = 0.05`, but users can calibrate the value of `offset` to obtain desired operating characteristics. In practice `offset` is rarely greater than 0.2.

Using the default values of `p.saf`, `p.tox`, `design` and `cutoff.eli` automatically provided by the function, we can design the aforementioned trial example by running `get.boundary(·)` as follows:

```
> get.boundary(target=0.3, ncohort=10, cohortsize=3)
Escalate dose if the observed toxicity rate at the current dose <= 0.2364907
Deescalate dose if the observed toxicity rate at the current dose >= 0.3585195

This is equivalent to the following decision boundaries

Number of patients treated     3 6 9 12 15 18 21 24 27 30
Escalate if # of DLT <=        0 1 2 2 3 4 4 5 6 7
Deescalate if # of DLT >=      2 3 4 5 6 7 8 9 10 11
Eliminate if # of DLT >=       3 4 5 7 8 9 10 11 12 14

A more completed version of the decision boundaries is given by

Number of patients treated     1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
20 21 22 23 24 25 26 27 28 29 30
Escalate if # of DLT <=        0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 4 5 5 5
5 6 6 6 6 7
```

```
Deescalate if # of DLT >=      1 1 2 2 2 3 3 3 4 4 4 5 5 6 6 6 7 7 7 8 8 8 9
9 9 10 10 11 11 11
Eliminate if # of DLT >=      NA NA 3 3 4 4 5 5 5 6 6 7 7 8 8 8 9 9 9 10 10
11 11 11 12 12 12 13 13 14

Default stopping rule:  stop the trial if the lowest dose is eliminated.
```

Remarks:

- The output presents the dose escalation/deescalation rule in two forms: based on the observed toxicity rate, and based on the observed dose-limiting toxicity (DLT). We recommend the latter because clinical researchers often find it easier to use.
- For convenience, two versions of decision boundaries are displayed: one is based on cohort size (3, 6, 9, ..., 30), the other is for all possible sample size (1, 2, 3, ..., 30). Although the design assumes a constant cohort size of 3, in practice the actual cohort size may vary along the trial due to various reasons. If that is the case, it is more appropriate to present the completed version of the decision boundaries in the protocol such that clinicians can make the decision of dose assignment, at any time of the trial, for any given number of patients who have been treated at the current dose level. Actually, this is one of the important advantage of the BOIN design, which allows the cohort size to vary from one cohort to another and to make the decision of dose escalation/deescalation at any time of the trial conduct.
- The elimination boundaries are used to prevent treating patients at overly toxic doses, based on the following Bayesian safety rule: if $\text{pr}(p_j > \phi | m_j, n_j) > 0.95$ and $n_j \geq 3$, dose levels j and higher are eliminated from the trial; and the trial is terminated if the first dose level is eliminated.

If we set `extrasafe = TRUE` to turn on the `extrasafe` feature, the output will include the extra stopping boundaries as follows,

```
> get.boundary(target=0.3, ncohort=10, cohortsize=3, extrasafe=T)

   ......

In addition to the default stopping rule (i.e., stop the trial if the lowest dose
is eliminated), the following more strict stopping safety rule will be used for
extra safety:
   stop the trial if (1) the number of patients treated at the lowest dose >=
3 AND (2) Pr(the toxicity rate of the lowest dose > 0.3 | data) > 0.9, which
```

```
corresponds to the following stopping boundary:

Number of patients treated at the lowest dose    3 6 9 12 15 18 21 24 27 30
Stop the trial if # of DLT >=                     3 4 5 6 7 8 9 10 12 13
```

2.2 Obtain operating characteristics

For protocol preparation, it is often useful to obtain the operating characteristics of the design. The function `get.oc(·)` can be used for this purpose. This function shares the same set of arguments as the function `get.boundary`(·) described previously, with three additional arguments:

- `p.true` the true toxicity probabilities of the investigational dose levels.
- `startdose` the starting dose level for treating the first cohort of patients. The default value is `startdose = 1`, i.e., starting from the lowest dose.
- `ntrial` the number of trials to be simulated.

Using the same setting as above and assuming that the true toxicity scenario is `p.true` = $(0.05, 0.15, 0.30, 0.45, 0.6)$, we show below how to obtain the operating characteristics based on 1000 simulated trials.

```
> get.oc(target=0.3, p.true=c(0.05, 0.15, 0.3, 0.45, 0.6), ncohort=10, cohortsize=3,
ntrial=1000)

selection percentage at each dose level (%):  1.1 23.4 54.2 20.2 1.1
number of patients treated at each dose level:  4.2 9.3 11.0 4.9 0.7
number of toxicity observed at each dose level:  0.2 1.4 3.3 2.2 0.4
average number of toxicities:  7.4
average number of patients:  30.0
percentage of early stopping due to toxicity:  0.0%
risk of poor allocation:  17.9%
risk of high toxicity:  8.0%
```

Remarks: In the output, *the risk of poor allocation* is defined as the percentage of simulation runs in which the number of patients allocated to the MTD (say n_{MTD}) is less than that of a standard non-sequential design, which assigns the equal number of patients to each dose, i.e., $\Pr(n_{\text{MTD}} < n/J)$. *The risk of high toxicity* is defined as the percentage of simulation runs

in which the total number of toxicities is greater than $n\phi$. These risk measures are of great practical importance because they gauge the likelihood of a trial turning out to be a "bad" trial (e.g., performing worse than a standard non-sequential trial) under a specific design. In other words, they measure the reliability (or variation) of the design. This important aspect of trial design has been largely overlooked by the existing literature, which typically focuses only on the mean or average performance of a design.

2.3 Select the MTD when the trial is completed

When the trial is completed, based on the observed data, we can select the MTD using the function `select.mtd(···)`. This function has six arguments: `target`, `ntox`, `npts`, `cutoff.eli`, `extrasafe` and `offset`, where

- `ntox` the vector recording the number of patients experienced toxicity at each dose level.
- `npts` the vector recording the total number patients treated at each dose level.

Arguments `cutoff.eli`, `extrasafe` and `offset` are the same as (and should be consistent with) these in functions `get.boundary(·)` and `get.oc(·)`, with default values `cutoff.eli = 0.95`, `extrasafe = TRUE` and `offset = 0.05`. When the default values are used, there is no need to specify these arguments in `select.mtd(·)`. Assume that the number of patients treated at five doses is $n = (3, 3, 15, 9, 0)$ and the corresponding number of patients who experienced toxicity is $y = (0, 0, 4, 4, 0)$.

```
> n<-c(3, 3, 15, 9, 0)
> y<-c(0, 0, 4, 4, 0)
> select.mtd(target=0.3, ntox=y, npts=n)
The MTD is dose level 3

Dose     Posterior DLT          95%
Level       Estimate     Credible Interval     Pr(toxicity>0.3|data)
1             0.02        ( 0.00 , 0.20 )             0.01
2             0.02        ( 0.00 , 0.20 )             0.01
3             0.27        ( 0.09 , 0.51 )             0.36
4             0.45        ( 0.16 , 0.75 )             0.66
5             ----        ( ----------- )             ----
```

The result is that dose level 3 is selected as the MTD. Note that no estimate is provided for dose level 5 because that dose has never been used for treating patients (i.e., `n[5]=0`).

3 Contact

If have any questions or comments, please contact Ying Yuan at yyuan@mdanderson.org. You can also visit http://odin.mdacc.tmc.edu/~yyuan/index_code.html for more information.

References

Liu S. and Yuan, Y. (2015). Bayesian Optimal Interval (BOIN) Designs for Phase I Clinical Trials, *Journal of the Royal Statistical Society: Series C*, **64**, 507-523.