Contemporary Clinical Trials Communications 22 (2021) 100694

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: http://www.elsevier.com/locate/conctc

The importance of loss functions: A note on the evolution of the Toxicity Probability Interval design

Jonathan Siegel

Bayer US LLC, United States

ARTICLE INFO

Keywords:
Dose escalation
Probability toxicity interval
Loss function
Clinical trial design
Statistical ethics

ABSTRACT

The Toxicity Probability Interval Design by Ji et al. (2007), which was subsequently modified by the mTPI (Ji et al., 2010), proposed a more efficient approach to early-phase dose-finding than conventional designs like 3 + 3. Subsequent authors reported issues with the method, finding that it tends to stay at a dose level when clinical intuition would suggest the toxicity level warrants decrease. Several iterations of refinement proceeded in an effort to address these issues, including the mTPI-2 and the keyboard method, as well as alternative approaches such as the BOIN. This author suggests the reason for these safety issues involves the underlying loss function. The TPI and mTPI used the identify function defined over wide intervals. As explained in this paper, this function and its domain can be problematic as a model of patients' loss experience. Later refinements moved the loss function closer to one more consistent with clinical intuition, and this explains their improved safety performance. Greater attention to quality as defined by fitness for use, including early evaluation of patient-experience and clinical-intuition implications of proposed loss functions, may improve future design efforts.

1. Introduction

This paper presents a critical historical look at a sequence of designs. While criticism has had its skeptics [1], insight into past approaches permits us to improve.

Ji et al. [2] developed the Toxicity Probability Interval (TPI) method for dose finding. The underlying insight was that a small early-phase trial cannot meaningfully detect small differences in toxicity, so it might be an efficient approach to regard an interval around the target toxicity as having equivalent toxicity. They reported that their method was more efficient than the conventional 3 + 3 design [2]. The method was quickly replaced with the calibration-free mTPI method [3], with consistent pre-specified probability intervals independent of observed results across all doses evaluated.

Subsequent authors, including Guo et al. [4], Yan et al. [5], and Zhou et al. [6], reported issues with the method. It tends to stay at a dose level when there is greater toxicity observed than clinical intuition might suggest is appropriate [4–7]. Several iterations refining the method to address these issues have been proposed, including the mTPI-2 [4], the keyboard method [5], and Pan et al.'s extension of the keyboard design to combination trials [8].

As explained below, I suggest the reason for these findings is that the mTPI method is based on a loss function which is problematic in describing the empirical losses patients experience when the toxicity level over- or under-shoots the target. This misspecification of the loss function results in mis-estimation.

2. Findings

The TPI method [2] divides the interval [0, 1] into 3 sub-intervals which are assumed to be of equivalent toxicity, and uses a beta-binomial model. The method makes an escalation, de-escalation, or stay decision at each dose level. In the original TPI method, not discussed in detail in this paper, the width of the middle interval was based on the posterior standard deviation σ_i of the toxicity found at the *i*th dose level evaluated with middle interval $\left[p_T - K_1\sigma_i,\ p_T + K_2\sigma_I\right]$, with K_1 and K_2 small positive pre-defined constants [2].

The original TPI was quickly superseded by the mTPI [3]. Under the mTPI, the middle interval is fixed a priori and is the same over all dose levels evaluated. It is assumed to be a small interval $[T - \varepsilon, T + \varepsilon]$ around the target toxicity; the lower interval is $[0,\ T - \varepsilon[$ and the upper interval is $]T + \varepsilon,\ 1]$. The mTPI selects with the highest score based on the score function

☆ This paper was produced during employment at Bayer, Inc., whose support the author gratefully acknowledges. All opinions expressed herein are the author's alone.

E-mail address: jonathan.siegel@bayer.com.

https://doi.org/10.1016/j.conctc.2020.100694
Received 28 June 2020; Received in revised form 14 September 2020; Accepted 16 December 2020
Available online 21 December 2020

$$S\left(I_I\right) = P\left(I_i\right)/M\left(I_i\right),\ i \in \{e, s, d\}$$

where $P\left(I_i\right)$ is the posterior probability of toxicity being in interval I_i, and $M\left(I_I\right)$ is the length of interval I_i. If the maximum score corresponds to the lower (e), middle (s), or upper (d) intervals, respectively, the algorithm escalates, stays, or de-escalates [3].

Guo et al. [4] showed that the mTPI score function under the beta-binomial model optimizes an identity loss function defined over the three intervals, loss 0 if the selected interval is the one in which the targeted toxicity probability lies, and loss 1 otherwise. Guo et al. also showed that selecting the interval with the highest score function results in the minimum loss, interpretable as a penalty for making wrong dose-finding decisions, under this loss function [4].

The problem with this loss function is its assumption that losses within each interval – not just the narrow one around the target, but the much wider upper and lower intervals – will be identical. While losses from incorrect toxicity estimates within a small interval around the target can be reasonably assumed to be equivalent, the same is not the case for the outer two intervals. If the target toxicity rate was 25%, then toxicity rates in the interval [20%, 30%] might reasonably be considered similar. But the method assumes that 31% and 100%, and 19% and 0%, can all also be treated as essentially equal. I respectfully suggest that this is not a clinically reasonable model of empirical patient loss experience, resulting in the method optimizing an unreasonable model of patient losses, resulting in the method selecting clinically unreasonable estimates as optimal.

A more clinically reasonable loss function would tend to be approximately continuous, with smaller losses for mis-estimates at smaller distances from target and greater losses for mis-estimates at greater distances. Clinical intuition would suggest a smaller difference in toxicity from target generally results in less loss than a larger one. Differences can be ignored over a small interval. But it may be ethically problematic to ignore them over a large one.[1]

The problem can be further explained by examination of the score function. Although Ji et al. [3] characterized the mTPI's score function as a probability mass function, the score function might better be characterized as a probability intensity function, where the term intensity function is used with analogy to use of an intensity function in physics (e.g. energy intensity) or to indicators familiar to clinical trialists like dose intensity. Intensity functions in this sense constitute a quantity divided by a measure of a space, here the length of an interval. A general characteristic of intensity functions (in this sense) is that, as quotients, they can be prone to estimation errors due to overweighting when small intervals result in small denominators. For this reason, unduly narrow middle intervals can overweight the middle interval score. Accordingly, this intensity-function analogy explains the mechanism by which the mTPI can result in stay decisions when clinical intuition would suggest escalating or de-escalating. There is a tension between an ethical need to specify a middle interval sufficiently narrow for the assumption of toxicity equivalence within the interval to be ethically reasonable, and an operational need to specify a middle interval sufficiently wide that dividing by its length does not result in unduly overweighting the score function.

Improvements to the mTPI method have tended to move the assumed loss function further in the direction of a continuous one. For example, Guo et al.'s mTPI-2 method [4], incorporating additional intervals and intermediate loss levels, resulted in use of a multi-step loss function, becoming a closer approximation to continuous as more levels are added. Similarly, Yan, Mandrekar and Yuan's keyboard design [5,6], which creates a series of "keys" of width equal to the interval around the target, moved further in the direction of an approximately continuous loss function by imposing a further rule that the equivalence intervals (except at the extrema) must be the same width, thereby more closely approximating continuity, and avoiding denominator-related estimation problems in the score function. In addition, ensuring that all interior intervals must have width no wider than the target interval helped address the mTPI's original problem, positing unduly wide toxicity intervals as having ethically equivalent loss.

3. Interval boundary designs

The Bayesian Optimal Interval Design (BOIN) [9] is an evolution of the cumulative cohort design (CCD) [10]. Liu and Yuan defined both a local and a global design, but recommended the local design for use in practice [9]. The CCD and BOIN are classified by Ji and Yang as interval boundary designs, distinguishing them from the class of interval designs which include the TPI and its successors [11]. As Ji and Yang explain, a key difference between an interval design and an interval boundary design is that interval designs base decisions on the posterior probability that the toxicity rate p_i lies within an interval, while in interval boundary designs the endpoints of the middle interval serve as boundaries for assessing the point estimate $\widehat{p}_i$ of the toxicity rate. In addition, interval boundary designs, unlike interval designs, do not use a classical decision-theoretic framework involving a formal loss function as an essential element. The optimality suggested by BOIN is based on an error function that mimics the Type I error in hypothesis testing [11].

Under the local BOIN design, given a target toxicity probability φ, pre-specified lower and upper interval boundaries $\varphi_1, \varphi_2 \ni \varphi_1 < \varphi < \varphi_2$, and given prior probabilities $\pi_{0j}, \pi_{1j}, \pi_{2j}$ of hypotheses H0: $p_j = \varphi$, H1: $p_j = \varphi_1$, and H2: $p_j = \varphi_2$ respectively for toxicity level p_j of the jth dose level, and number of patients n_jn$_j$, the design selects boundariesλ_{1j} and λ_{2j} with $\varphi_1 < \lambda_{1j} < \varphi < \lambda_{2j} < \varphi_2$ minimizing the probability of making an incorrect escalation decision, using a binomial probability based error function.

$$\begin{aligned}\alpha\left(\lambda_1, \lambda_2\right) = \pi_{0j}\left\{Bin\left(n_j\lambda_{1j}; n_j, \phi\right) + 1\right. \\ \left. - Bin\left(n_j\lambda_{2j} - 1; n_j, \phi\right)\right\} \\ + \pi_{1j}\left\{1 - Bin\left(n_j\lambda_{1j}; n_j, \phi_1\right)\right\} \\ + \pi_{2j}Bin\left(n_j\lambda_{2j} - 1; n_j, \phi_2\right)\end{aligned}$$

For estimated toxicity $\widehat{p}_j$, the BOIN design escalates if $\widehat{p}_i < \lambda_{1j}$, de-escalates if $\widehat{p}_i > \lambda_{2j}$, and remains at current dose level if $\lambda_{1j} < \widehat{p}_j < \lambda_{2j}$ [9].

The analogy to sample size determination in hypothesis testing [9] provides a way to obtain an implicit loss function. As in a Bayesian hypothesis test, the implicit expected loss is a sum of weighted identity functions, and the score function is simply the probability of a correct decision $1 - \alpha\left(\lambda_1, \lambda_2\right)$. While the mTPI uses a true interval approach, the local BOIN design's interval boundary approach is analogous to a point hypothesis test, not an interval hypothesis test. In a point hypothesis test, rejection of H0 is merely evidence that a parameter is not at the null point value; it is not evidence that it lies at or beyond the alternative hypothesis point value.[2] Similarly, escalation or de-escalation decisions in the BOIN design do not require establishing that the toxicity

[1] Perhaps a physical analogy might help illustrate the relationship for nonstatistical readers. If one throws a ball in a U-shaped valley, it will bounce off and head towards the bottom. But in a square valley with vertical cliffs and level plateaus above the cliffs, once a ball is thrown over the cliff there is nothing to keep it going to infinity. If everything above the cliff has the same loss function, then balls landing close to the cliff will not be preferred over ones that roll farther away.

[2] In both frequentist and Bayesian point hypothesis testing, rejection of a point null hypothesis does not provide reliable evidence for acceptance of the stated point alternative hypothesis. The alternative hypothesis is merely the value that will result in rejection of the null hypothesis with a pre-determined reliability. A boundary is a point at which the likelihood of two hypotheses becomes equivalent, and hence must lie between the two. The same is true with the boundaries in the local BOIN design.

rate lies (respectively) below or above the interval. Rather, it is sufficient to establish that the toxicity rate lies below or above the target rate. Accordingly, as in a conventional hypothesis test, the boundaries must always lie inside the interval, between the null and the alternative (s). This, in turn, guarantees a principal comparative benefit of the BOIN design, that the null hypothesis (stay at dose level) can never be accepted when the estimated toxicity rate lies outside the interval boundaries. Because the BOIN only considers point alternatives within the posited interval around the target (including its endpoints), and simply does not consider alternatives outside it, it never violates the ethical principal that only toxicity rates within an appropriately small interval should be considered ethically equivalent. Because of the analogy to hypothesis testing, the score function is a simple probability, not a probability intensity as in the mTPI. For this reason, it is not subject to the principal problem with the mTPI's score function. Because narrowing the interval around the target does not increase the weight of the target interval's score, prespecifying an unduly narrow interval does not lead to overweighting the target toxicity score.

4. Discussion

Guo et al. explained the safety issues they observed in the mTPI method as “Ockham's razor is too sharp” [4]. In their view, based on Jeffries and Burger's characterization of Ockham's razor[3] [12], the problem involved a conflict between simplicity in statistical inference and clinical considerations regarding patient safety. From this perspective, the statistical profession's role is to introduce mathematically efficient and simple designs, which the mTPI design successfully did. It is clinicians who are responsible for applying clinical considerations in selecting a design that is fit for use, such as being safe for their patients. In this formulation fitness-for-use issues like safety are simply not the statistician's responsibility. If Ockham's razor is too sharp – if a simple and efficient design results in safety concerns – then that is the user's problem and not the designer's.

I respectfully propose that this may be too narrow a conception of the role and responsibility of the clinical statistician. Loss functions and their empirical and clinical reasonableness are matters within the statistician's purview. Statisticians are in a position to ascertain and ensure that models of patient loss are empirically reasonable and clinically appropriate, in collaboration with clinicians and patients, up-front, before design development gets very far. The statistician's responsibility arises in no small part because the statistician will often be the only member of a design team in a position to interpret, and hence assess the practical clinical implications of, mathematical assumptions and functions. To ensure this, an ability to cooperate closely with clinicians and an understanding of the clinical implications of designs – not just the mathematics but the meaning -- is an essential part of the needed skill set.

To continue Guo et al.‘s [4] reference to Occam's Razor, if we include the specification of patient losses as part of the problem a useful clinical design seeks to solve, then Occam's Razor would select a design whose model of patient loss experience is both accurate and simple.

From this perspective, the TPI evolution began with a loss model which was indeed simple. The identify function is one of the simplest loss functions possible, and reducing the problem to selecting from three intervals similarly simplifies. But it was inaccurate in the sense of not appropriately modeling patients' loss experiences. In the historical evolution through the mTPI, mTPI-2, and Keyboard designs, each design iteration made a small improvement by adding new steps and constraints to the loss function, but did so at the expense of making the loss function more complex. As a result, the loss model moved from simple but inaccurate to more accurate but less simple. Perhaps further improvement is possible. Perhaps a method using a simpler description of the losses might be devisable that might result in an ethically appropriate but statistically more efficient approach.

The existing evolution might, however, be adequate to solve this particular problem. With enough intervals, a step loss function can be devised that is essentially indistinguishable from a continuous one for the small sample sizes appropriate to Phase 1 trials. Perhaps this is enough. In addition, interval boundary designs such as BOIN provide an alternative which, while still interval-based, avoids ethically unreasonable escalation and de-escalation decisions of the type this paper discusses.

For the future, including future design problems not yet posed, I would respectfully suggest that statisticians start off asking whether a proposed loss function represents a reasonable model of patient experience. It might be better to emphasize that loss functions have an empirical character. We can ask patients, either informally or through surveys, to describe their loss experience and their preferences and wishes to us. We can turn to clinicians for their clinical experience. It might be better to ask than to assume. While not having the same obligations as the physicians we work with, statisticians who devote their careers to the clinical environment and improving patient welfare should nonetheless also share their commitment to avoiding harm.

5. Conclusion

W. Edward Deming [13] is perhaps best known as an advocate of the position that quality (even of statisticians' work) ought to be judged by its fitness for use. Consistent with this view, he was also an early advocate of empirical loss functions. He argued against the tendency of statisticians in his day to treat loss functions as mathematical conveniences rather than as central to describing the problem to be solved. He argued for the importance of careful observation, empiricism, and subject-matter knowledge in evaluating loss experiences. This advice appears still relevant today.

References

[1] M. Twain, H.E. Smith, Autobiography of Mark Twain: Reader, s Edition, Univ. of California Press, 2012.

[2] Y. Ji, Y. Li, B.N. Bekele, Dose-finding in phase I clinical trials based on toxicity probability intervals, Clin. Trials 4 (2007) 235–244.

[3] Y. Ji, et al., A modified toxicity probability interval method for dose-finding trials, Clin. Trials 7 (2010) 653–663.

[4] Guo, et al., A Bayesian interval dose-finding design addressing Ockham's razor: mTPI-2, Contemp. Clin. Trials 58 (2017) 23–33.

[5] F. Yan, M.J. Sumithra, Y.Y. Keyboard, A novel bayesian toxicity probability interval design for phase I clinical trials, Clin. Canc. Res. 23 (2017) 3994–4003.

[6] H. Zhou, Y. Yuan, L. Nie, Accuracy, safety, and reliability of novel Phase I trial designs, Clin. Canc. Res. 24 (2018) 4357–4364.

[7] Y. Zhu, et al., Evaluating the effects of design parameters on the performance of Phase I trial designs, Contemporary Clinical Trials Communications 15 (2019), https://doi.org/10.1016/j.conctc.2019.100379.

[8] Pan, et al., Keyboard design for phase I drug-combination trials, Contemp. Clin. Trials 92 (2020), https://doi.org/10.1016/j.cct.2020.105972.

[9] S. Liu, Y. Yuan, Bayesian optimal interval designs for phase I clinical trials, J Royal Statist Soc Appl Statist 64 (2015) 507–523.

[10] A. Ivanova, N. Flournoy, Y. Chung, Cumulative cohort design for dose-finding, J. Stat. Plann. Inference 137 (2007) 2316–2327.

[11] Y. Ji, S. Yang, On the interval-based dose-finding designs, https://arxiv.org/pdf/1706.03277.pdf.

[12] W. Jeffries, J. Burger, Ockham's razor and bayesian analysis, Am. Sci. 80 (1992) 64–72.

[13] W.E. Deming, The New Economics for Industry, Government Education, 2nd Edition. MIT Press, 1994.

[3] Jeffries and Burger (1992) characterized Ockham's razor as "‘an explanation of the facts should be no more complicated than necessary,’ or ‘among competing hypotheses, favor the simplest one.’" [12].