Dose-finding methods: Moving away from the 3+3 to include richer outcomes Running

Xavier Paoletti, Damien Drubay, Laurence Collette

Title: Dose-finding methods: Moving away from the 3+3 to include richer outcomes

Running Title: Dose finding methods needs richer outcomes

Xavier Paoletti*[1], Damien Drubay[1], Laurence Collette[2]

1 – Biostatistic and Epidemiology department, Gustave Roussy Cancer Center; Université Paris 11 Saclay and INSERM U1018 CESP OncoStat, Villejuif, France
2 – European Organization for Research and Treatment of Cancer Headquarters, Brussels, Belgium

All correspondence to Xavier Paoletti, Biostatistics and epidemiology Dept & INSERM CESP OncoStat, Gustave Roussy, 114 rue Ed Vaillant, Villejuif Cedex, France

Abstract: The most commonly used method for dose-finding, the 3+3, has poor performances. New adaptive designs are more efficient. Nevertheless, they have reached a maximum performance level, and further improvement requires either larger sample sizes, or outcomes measures richer than the simplistic severe toxicity measured at cycle 1.

Keywords: Dose finding; outcome; adaptive; Repeated data.

Potential conflicts of interest: All declare they have no conflict of interest.

Funding: XP was partly supported by the ITN-Marie Curie IDEAS project; XP and DD were partly supported by the French NCI (INCA-Optidose program); LC was partly supported by FOCA (Fonds cancer) from Belgium.

Body

In this issue of Clinical Cancer Research, Yan et al. (1) propose a new adaptive dose-finding method, termed Keyboard as an alternative to the 3+3 design. The performance of the most commonly used dose-finding method, the 3+3, is disappointing: The chance of finding the correct dose in a phase I trial is no more than 40% (1). How many active, potentially promising agents are dropped due to incorrect dose selection during development? Few researches address this question (2), but stakeholders acknowledge the limitations of early phase trials (3). The statistical community unanimously agrees that the 3+3 is not an efficient method, raising ethical concerns according to NIH guidelines (4) and alternatives should be implemented.

Keyboard combines simplicity, performance and flexibility. Simplicity for two reasons: (i) it is based on a natural definition of the maximum tolerated dose (MTD), i.e., the dose at which the risk of dose-limiting toxicity (DLT) is within a predefined range (typically 20–35%), and (ii) the decision rules to (de)escalate are driven by the accumulated observations, and the observed proportion of DLT at a given dose level. The more patients enrolled at a specific dose, the more confident we can be that this dose is (close to) the MTD, and the more likely the same dose is recommended for the next patients.

However, simplicity is not an objective *per se* and the key point is performance. The Keyboard design is rooted in good and fruitful statistical concepts and properties. It belongs to the semi-parametric continual reassessment class of methods (semi-parametric CRM) (5), that has been shown to provide good operating characteristics and valid asymptotic properties, whilst being quite flexible. In their simulation study (1) the Keyboard design outperformed the 3+3, but was slightly inferior to the CRM (6). Importantly, these performances were quite close to the maximum achievable performances, given the sample size and the scenario. Figure 1 shows the distribution of the recommendations obtained with the optimal benchmark. This benchmark uses a so-called complete information, that is estimable only in a simulation framework, as if each patient could be treated at all the doses independently (7). This optimal method serves as a reference.

An important advantage of Keyboard over the 3+3 design is its flexibility. As with other model-based methods, all collected data contributes to the MTD assessment, and not only data from the last six patients.
This is crucial for monitoring patient tolerance throughout the trial, including during the dose expansion cohorts. These large cohorts of patients with selected tumor types are aimed to explore preliminary signs of activity. However, they too must be monitored for toxicity, as shown repeatedly by Iasonnos et al. (8). In cases of excessive toxicity rate, the investigated dose should be modified during the trial. Following the recommendations of the DLT-TARGETT (9), the expansion cohort should serve to refine the toxicity assessment. The possibility given by Keyboard to update the estimated risk of DLT and its confidence interval, based on all accumulated data, provides an additional tool to fulfil these requirements. The analysis combining data from the dose-escalation and the dose-expansion cohorts could increase the chance of selecting the correct dose (8).

To implement Keyboard into practice, we need easy-to-use software with two purposes: Firstly, given previous observations, it should indicate the next dose level to allocate (see Table 2 and Figure 2 in (1)). Secondly, and perhaps even more importantly, it should be able to run simulations similar to those presented in Figures 3 to 5 (1). Indeed, protocols that used adaptive designs require more preparatory work. There is no equivalent to the “null hypothesis, clinical targeted difference and type I error” used to design phase II or III trials. Each team preparing a phase I trial must simulate possible trials under various scenarios, specific to the trial context. This preparatory step is performed by the statistician in collaboration with the investigator in order to calibrate and tailor the design, and thereby to obtain the best possible operating characteristics. The scenarios of the simulations are designed to reflect the expected relationship between the dose and the risk of DLT, for the agent under investigation. For example, the scenario mimicking a trial of a monoclonal antibody will be different from that of a kinase inhibitor, since the former probably has a much larger therapeutic index than the latter. Similarly, to prepare a pediatric trial, we would choose scenarios based on the data collected in adults, as agents are usually tested in adults before being investigated in children (10). The examples from Figure 4 of Yan et al (1), highlight that there is no single method that is the most effective across all scenarios. This process is lengthy, but may reconcile the investigators with statistically oriented dose-finding designs.

This combination of relative simplicity, good operating characteristics, and flexibility to tackle various practical situations, should motivate investigators and sponsors to move away from the 3+3 and to adopt alternative designs.
Will conducting future phase I trials with the Keyboard method guarantee the systematic selection of the correct dose? Unfortunately, not. As shown in Figure 1, the chance of selecting the correct dose with the benchmark, (i.e the highest achievable performance), is below 60% in most of the explored scenarios. Despite the good statistical properties of Keyboard, its performances are limited by the type of primary endpoint used in phase I trials: the DLT variable is binary and has irreducible binomial variability for a given (often limited) sample size. Continuous, ordinal or multiple endpoints would be much more informative.
One may then question the choice of the DLT as sole endpoint. The DLT-TARGETT database consists of individual patient data from 54 phase I trials of single-targeted agents (9). Of these, a total of 25% completely missed the primary objective, as they failed to identify the MTD and recommend a dose for phase II, due to lack of DLT. In the 2,084 treated patients, more than 24,000 graded toxicity, as measured by the NCI common toxicity criteria scale, were recorded as possibly related to the treatment. These toxicities occurred at any of the first six cycles of treatment. Yet, only 164 DLTs were reported, reflecting the huge shrinkage in the available information. A large fraction (50%) of the first grade 3 or 4 toxicity occurred after the DLT assessment period, and dose-intensity was reduced in more than 20% of all treatment cycles, raising the question of tolerability of the treatments administered in the long-run. To define the recommended dose, we need richer outcomes, possibly reflecting both toxicity and activity. Continuous endpoints have a better sensitivity to agent effect and a better discriminatory value to rank dose activity.

There is an urgent need for innovative designs that are both efficient and simple, and reduce the high failure rate of dose-finding (2). This “well-tempered” Keyboard can help synthetize complex information, and may stimulate statisticians and principal investigators to work together to provide a good design for a particular agent. Nevertheless, more efficient use of the multiple sources of data that are collected in most phase I trials, is needed. Currently PK, pharmacodynamics biomarkers, late-onset toxicity, and complex imaging are not formally integrated in the dose-finding analysis (11); this is a waste of resources.

Acknowledgments: The authors are indebted to all members of the EORTC-led DLT-TARGETT group.

Figure 1: Percentage of correct selection with the 3+3, the Continual Reassessment Method (CRM) and the optimal benchmark in the 10 scenarios presented by Yan et col (1), with a target rate of either 20% (left part) or 30% (right part).

Bibliography

1- Yan F, Mandrekar SJ, and Yuan Y Keyboard: A Novel Bayesian Toxicity Probability Interval Design for Phase I Clinical Trials. Clinical Cancer Research xxx-xxx
2- Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature Review Drug Discovery. Aug;3(8):711-5 (2004)

3- Wong KM, Capasso A, Eckhardt SG. The changing landscape of phase I trials in oncology. Nature Review Clinical Oncology Feb;13(2):106-17. (2016)
4- Emanuel EJ, Wendker D, Grady C. What makes Clinical Research Ethical? JAMA May 24/31; 283(20) (2000)
5- Clertant M, O'Quigley J. Semi-parametric dose finding methods. Journal of the Royal Statistical Society, Series B (Statistical Methodology) Feb; (2017)
6- O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. Biometrics Mar;46(1):33-48; (1990)
7- Paoletti X, O'Quigley J, Maccario J. Design efficiency in dose finding studies. Computational Statistics and Data Analysis.45(2):197–214 (2004)
8- Iasonos A, O'Quigley J. Design considerations for dose-expansion cohorts in phase I trials. Journal of Clinical Oncology Nov 1;31(31):4014-21 (2013)
9- Postel-Vinay S, Collette L, Paoletti X, Rizzo E, Massard C, Olmos D, Fowst C, Levy B, Mancini P, Lacombe D, Ivy P, Seymour L, Le Tourneau C, Siu LL, Kaye SB, Verweij J, Soria JC. Towards new methods for the determination of dose limiting toxicities and the assessment of the recommended dose for further studies of molecularly targeted agents--dose-Limiting Toxicity and Toxicity Assessment Recommendation Group for Early Trials of Targeted therapies, an European Organisation for Research and Treatment of Cancer-led study. European Journal of Cancer. Aug;50(12):2040-9 (2014)
10- Paoletti X, Geoerger B, Doz F, Baruchel A, Lokiec F, Le Tourneau C. A comparative analysis of paediatric dose-finding trials of molecularly targeted agent with adults' trials. Eur J Cancer. Jul;49(10):2392-402 (2013)
11- Postel-Vinay S, Soria JC. Phase I trials in oncology: a new era has started. Annals of Oncology Jan;26(1):7-9 (2015)

Figure 1: