Improved Designs and Analyses of Safety and Efficacy in Immuno-Oncology

Bo Huang* Pei-Fen Kuan†

Abstract

Immuno-oncology has emerged as a new prominence in oncology. Common immunotherapy approaches include cancer vaccine, effector cell therapy and T-cell stimulating antibodies. Checkpoint inhibitors such as CTLA-4 and PD-1 antagonists have shown promising results in multiple indications in solid tumors and hematology. However, the mechanisms of action of these novel drugs pose unique statistical challenges in the accurate evaluation of clinical safety and efficacy, including late-onset toxicity, dose optimization, evaluation of combination agents, pseudoprogression, delayed and lasting clinical activity. Traditional statistical methods may not be most accurate or efficient. There is high unmet need to develop the most suitable statistical methodologies to efficiently develop cancer immunotherapies. In this paper, we summarize these issues and discuss alternative methods to meet the challenges in the clinical development of these novel agents for both safety and efficacy endpoints. For safety evaluation, we propose using the time-to-event model-based design to handle late toxicity, a simple three-step procedure for dose optimization, and flexible rule-based or model-based designs for combination agents. For efficacy evaluation, we propose alternative endpoints/designs/tests including optimal designs for time-specific probability endpoint, restricted mean survival time, generalized pairwise comparison, immune-related response criteria and weighted log-rank test. Benefits and limitations of these methods are considered and some recommendations are proposed for applied researchers to implement these methods in clinical practice.

Key Words: Immunotherapy, Safety, Efficacy, Delayed Effect, Time-to-Event, Optimization

1. Introduction

Immuno-oncology is a fast-growing and exciting research area in oncology. Immunotherapies work by harnessing the immune system to induce anti-tumor response. Immuno-oncology has been in the limelight over the past several years thanks to the breakthrough of checkpoint inhibitors including CTLA-4 and PD-1 antagonists that have demonstrated dramatic and durable activity in multiple indications of solid tumors and hematology.

Cancer cells can bind to T-cells and turn off their ability to detect and kill tumor cells. Immunotherapies can block tumor cells from deactivating T-cells. Depending on the mechanism of actions, there are a number of approaches to cancer immunotherapies, including cancer vaccine, effector cell therapy and T-cell stimulating therapy (2015 PMDA guidance). Unlike childhood vaccine aiming at preventing diseases, cancer vaccines are aimed at treating cancer on patients who have it. The idea is to prompt the immune system to attack the disease by presenting it with some piece of the cancer. Effector (adoptive) cell therapies work by removing immune cells from the body, altering them genetically to fight cancer, multiplied then transferred back to the human body to boost their anti-cancer effect.

Most success has come from checkpoint inhibitors which are T-cell stimulating therapies. Checkpoint inhibitors are inhibitory antagonists that block inhibitory receptors such as CTLA-4 and PD-1. By releasing the brake on the immune system, they can boost the immune system and have so far led to high therapeutic benefits in many patients. Contrary to this approach, stimulatory immunotherapies are agonist antibodies against immune-stimulating molecules such as 4-1BB, OX-40 and GITR, which are inductively expressed mainly on activated T cells and serve as receptors transmitting stimulatory immune signals.

*Pfizer Inc., 445 Eastern Point Rd, Groton, CT 06340

†Department of Applied Math and Statistics, Stony Brook University, Stony Brook, NY 11794

Despite the tremendous success in clinical trials and the result of remarkable and durable responses observed in multiple cancer indications, the unique mechanisms of action of these novel drugs pose unique statistical challenges in the accurate evaluation of clinical safety and efficacy, including late-onset toxicity, dose selection, pseudoprogression, delayed and lasting clinical activity. Traditional statistical methods may not be most accurate or efficient.

In this paper, we discuss these issues and propose alternative methods to meet these challenges in the clinical development of cancer immunotherapies. Section 2 describes the statistical challenges and considerations in safety studies of immunotherapies, followed by efficacy challenges and considerations in Section 3. Section 4 concludes with a discussion.

2. Statistical Challenges and Considerations in Safety Studies of Immunotherapies

The main statistical challenges in safety trial of cancer immunotherapies are management of toxicity, dose optimization and evaluation of immunotherapy combinations.

Unlike chemotherapies that attach tumor cells directly, immunotherapies target the immune system to elicit effect on cancer cells. The indirect effect could lead to late or cumulative effect on both safety and efficacy outcomes. For safety evaluation, one concern is late-onset toxicity beyond the traditional first-cycle DLT observation window. Traditional dose finding designs making dose escalation and de-escalation decisions based upon Cycle 1 of treatment may fail to adequately assess the safety profile of the experimental drug.

Due to the life-threatening nature of cancer, a higher degree of drug toxicity (often referred as dose limiting toxicity [DLT]) is generally considered acceptable. It is commonly assumed higher dose will lead to higher toxicity. However, for drug exposure and efficacy, the same assumption may not hold for cancer immunotherapies, especially for stimulatory antibodies. It is possible that efficacy could go up first and then go down due to overstimulation of the immune system. As a result, the optimal dose may not be the maximum tolerated dose (MTD) because higher dose may have higher toxicity but less clinical activity.

It is also appealing to combine two or more immunotherapy agents to elicit synergistic effect to maximize the immune activity, creating a so called Immunotherapy "cocktail". However, complexity of the design of a phase I trial increases exponentially with the number of different drugs included in the combination strategy.

2.1 Time-to-Event Continual Reassessment Method for Late-Onset Toxicity

It is desirable to consider a DLT or AE evaluation window of multiple cycles to account for delayed toxicities due to late and cumulative effect of immunotherapies. However, dose finding trials are sequential in nature, and a long DLT window will lead to long observation time and patients are susceptible to early drop-out (which usually requires replacement). Furthermore, trial accrual is subject to opening and closing which may pose additional risk to the success of the study.

To address this challenge, a time-to-event continual reassessment method (TITE-CRM) (Cheung and Chappell, 2000) should be considered as a potential dose finding design in safety trials. On top of the conventional CRM method (O'Quigley et al., 1990), TITE-CRM utilizes a time-to-event approach by employing a weight function (conditional probability of having a DLT during follow-up given it does occur within the DLT window) in the weighted binomial likelihood function, and can be open to accrual continually, while making timely dosing decision based on data from all treated patients. The weight (from 0 to 1) is an increasing function of the follow-up time of the patient, and if a DLT occurs, the weight becomes 1. Patients with incomplete follow-up will be incorporated into the statis-

tical model without holding enrollment and thus this method results in a faster enrollment and shorter trial. Variations of TITE-CRM have been developed. Mauguen et al. (2011) presented a hybrid design (TITE-EWOC) by introducing the time-to-event approach in the Escalation with Overdose Control (EWOC)method (Babb et al., 1998).Yuan and Yin (2011) proposed an expectation-maximization (EM) CRM approach to handling late-onset toxicity. Wages et al. (2013) extended the TITE-CRM design in the presence of partial ordering for a drug combination trial. Liu et al. (2013) proposed a data augmentation design (DA-CRM) for delayed toxicity by treating the unobserved toxicities as missing data. Huang and Kuan (2014) proposed an adaptive weight function from patients' cyclical safety data to describe the cycle-toxicity pattern by using a Multinomial-Dirichlet conjugate.

Figure 1 illustrates the potential time saving from the TITE-CRM method by simulations when the DLT evaluation time is 2 treatment cycles of 8 weeks, compared to the conventional CRM design and the standard 3+3 design. It is assumed patients are available when enrollment is open. For simplicity, the probability of DLT for patient i is linear up to patient number 16 with DLT rate capped at 25% afterwards. It is also assumed an early drop-out rate of 15% and the time delay due to early drop-out (which applies to both the 3+3 design and conventional CRM) follows a normal distribution with mean of 3 weeks and standard deviation of 1 week: $N(3, 1)$. For practical consideration, a short enrollment pause is assumed for TITE-CRM when a DLT is observed. Three scenarios of TITE-CRM are presented when enrollment pause follows a normal distribution with expected pause due to DLT of 1, 2 or 3 weeks, and standard deviation of one-third of the expected pause: $N(k, \frac{k}{3}),\ \ k = 1, 2, 3$. Two versions of CRM with altered cohort sizes are compared: CRM (1) and CRM (3) with cohort size of 1 and 3 respectively. The shaded areas are the 95% confidence limits. The simulation results demonstrate the apparent advantage of using a time-to-event endpoint to shorten the trial duration when the DLT evaluation window is long or when there is a non-negligible rate of loss-to-follow-up.

The TITE-CRM design using a cyclical adaptive weight function (Huang and Kuan, 2014) has been successfully implemented in some immunotherapy studies, including 4 ongoing/completed trials sponsored by Pfizer. To facilitate decision making in a transparent and systematic manner, it is recommended that a dose escalation steering committee be established for each study, with a charter written to document the trial conduct process and operational procedures. Details and additional considerations on practical implementation of the TITE-CRM design and other novel dose finding designs are discussed in a recent paper (Huang et al., 2016).

2.2 Statistical Considerations for Dose Optimization

Another challenge in the evaluation of immunotherapies is the selection of dose for subsequent investigation of clinical activity, or dose optimization. The current practice in oncology trials is to first estimate the MTD, and then further test it in an expansion cohort. However, depending on the mechanism of the drug, the MTD may not be the most efficacious dose, but definitely at least as toxic as the lower doses. In addition, the MTD may not be able to be identified for some agents (e.g. PD-1 or PD-L1 checkpoint inhibitors) due to their benign safety profile. Therefore, some work need to be conducted for dose optimization as opposed to following the traditional MTD paradigm in dose escalation safety trials.

There are novel methods (Braun, 2002; Thall and Cook, 2004; Fox et al., 2002; Piantadosi and Liu 1996; Mandrekar et al., 2007; Polly and Cheung, 2008) developed looking at composite endpoints such as safety plus efficacy (or biomarker/PK endpoint) as a way to prospectively select the optimal dose that is safe and efficacious. However, despite the great

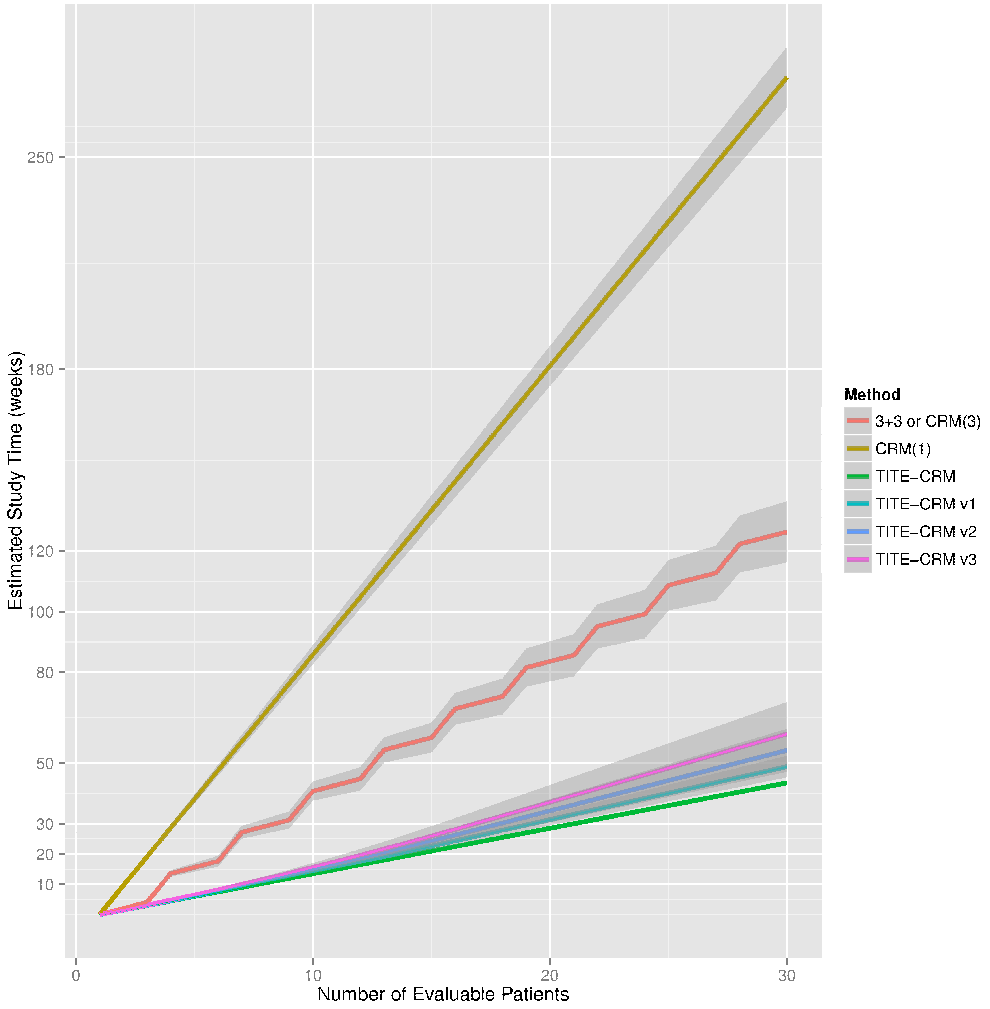

Figure 1: Simulated trial durations comparing the TITE-CRM method with the 3 + 3 design and conventional CRMmethods. CRM(1) and CRM(3) correspond to cohort size of 1 and 3 respectively.

efforts by researchers, the use of these novel methods in practice is very limited. There are several reasons behind this. First of all, efficacy requires a much longer follow-up than safety, making sequential dose escalation/de-escalation more difficult. Secondly, unlike safety, efficacy is sensitive to patient heterogeneity and the efficacy profile varies dramatically in different populations. In addition, response occurrence is usually scarce in phase I dose escalation trials to produce the data needed for model fitting. Furthermore, these designs are complex Bayesian model-based methods that are computationally intensive with many parameters, making practical implementation challenging.

Alternatively, we propose a practical and pragmatic approach for dose optimization in a 3-step procedure:

1. Implement a Bayesian model-based method to prospectively assess safety. The advantage of a model-based approach is the possibility of allowing flexible definition of the MTD according to the mechanism of action (e.g. instead of targeting 33% DLT rate as for cytotoxic drugs, one can target 20% or 15% DLT rate), or revise the DLT definition by including clinically relevant Grade 2 adverse events.

2. Conduct adequate retrospective analysis of PK, evaluate relevant PD biomarkers, investigate the dose-exposure relationship and assess early efficacy signal to have a good understanding of the totality of the data.

3. Evaluate efficacy of 1-3 doses in the expansion cohort in a homogeneous patient population, and randomize patients to the selected dose groups. One can use fixed randomization, or use response adaptive randomization to assign more patients to potentially more effective doses.

Compared to the designs that prospectively assess safety and efficacy in both dose escalation and final dose selection, this pragmatic approach may take longer and require more patients. However, it does not rely on complex statistical models (only parsimonious safety models) and can be implemented easily. It also provides reasonable confidence moving forward in further testing of the treatment in randomized controlled studies. Therefore, the design may be considered as an alternative approach to other novel methods.

2.3 Statistical Considerations for Combination of Immunotherapies

Clinical trials for combination of agents have become increasingly common in recent years, especially in the field of immuno-oncology, as it is believed that combining an immunotherapy with chemotherapies, targeted therapies or other IO agents could elicit synergistic effect and significantly boost efficacy. A combination of drugs can target cancer cells that have different drug susceptibilities, achieve a higher intensity of dose if the drugs have non-overlapping toxicities and reduce the likelihood of drug resistance (Dancey and Chen, 2006). However, complexity of the design of a phase I trial increases exponentially with the number of different drugs included in the combination strategy. When drugs in combination have different mechanisms of action or non-overlapping toxicities, the recommended dose for phase II for the combination may be close to the recommended dose of each drug given as a single agent. However, as the biological effects of the combination may be quite complex and the PK/PD drug-drug interaction between the agents is largely unknown, it is often difficult to administer at the recommended dose of each drug given as a single agent. Furthermore, unlike single-agent dose escalation where monotonicity is generally assumed to be true, in drug combination only partial ordering is known for the dose-toxicity relationship.

A number of escalation strategies have been proposed in the literature and were described in Harrington et al. (2013). For dual-agent combination, when escalation occurs on both agents (no agent is fixed), the rule-based 3+3 or A+B design can be extended to the 3+3+3 or A+B+C design. Some flexible but more complex model-based designs were also published with model parameters to account for the inherent complexity of drug combination in the dose-toxicity relationship (Kramar, et al., 1999; Thall et al., 2003; Huang et al., 2007; Yin and Yuan, 2009; Wages et al., 2011, Mander and Sweeting, 2015). Choosing the suitable dose escalation strategy and the right doses remains a challenge in the development of combination therapies, and it should be determined by the best possible scientific and clinical practice rationale.

3. Statistical Challenges and Considerations in Efficacy Studies of Immunotherapies

There are a number of statistical challenges in efficacy studies of immunotherapies. First of all, unlike chemotherapies and targeted therapies, immunotherapy agents have an indirect effect on cancer cells. As a result, it is not uncommon to observe initial pseudoprogression followed by tumor regression and clinical activity. Response to treatment may also

demonstrate immune-related patterns that cannot be fully characterized by standard RECIST criteria. Secondly, the proportional hazard assumption used in power calculation of time-to-event endpoints is unlikely to hold for immunotherapies. Due to the delayed and durable anti-tumor effect on cancer cells, the survival curves may take a while to separate and the curve of the immunotherapy agent can have a very long tail. Therefore, the log-rank test may not be the most appropriate test and the hazard ratio (HR) may be difficult to interpret. Furthermore, there could be a weak or negative correlation between short-term surrogate endpoint of time-to-progression (TTP) and progression-free survival (PFS) with the long-term survival endpoint. Treatments may prolong survival but not imaging-based surrogate endpoint, as is the case in recent phase 3 trials of nivolumab in advanced nonsquamous non-small-cell lung cancer and renal cell carcinoma (Borghaei et al., 2015; Motzer et al., 2015).

In recent years, the oncology community are looking for alternative endpoints and statistical methods and moving beyond the HR to summarize treatment effect (e.g., Uno et al., 2014; Hoos et al., 2010; Buyse, 2010) Some of these methods related to cancer immunotherapies include using time-specific probability endpoint (landmark endpoint), using restricted mean survival time (RMST) to summarize treatment effect, implementing immune-related response criteria (irRC), extending the Wilcoxon MannWhitney statistic for generalized pairwise comparison, and performing weighted log-rank test for hypothesis testing.

3.1 Time-Specific Probability Endpoint

There are a number of benefits in using a time-specific survival probability endpoint (e.g. 12-month PFS, 2-year OS). We can potentially assess benefit risk early if the median time-to-event is long, an appealing feature particularly in phase 2 proof-of-concept (POC) studies. Unlike time-to-event endpoints, a time-specific probability endpoint is not event driven, thus operationally more predictable with respect to the timeline of interim and final analyses. When delayed treatment effect exists, selection of a landmark timepoint after the curves separate may provide greater statistical power. Unlike the HR, the interpretation of the time-specific probability endpoint is easy even when there is large departure from the proportional hazard assumption. Recent European Medicines Agency draft guidelines on the evaluation of anti-cancer medicinal products (EMA,2011) noted that progression would be observed at a slow rate for some conditions, so event rates at a specified fixed time might be appropriate.

One obvious limitation is that it does not capture the entire survival distribution as it only evaluates the probability of event occurrence at one landmark time, the selection of which may be arbitrary. Under the proportional hazard assumption, the hypothesis test based upon the time-specific probability has lower statistical power and requires a larger sample size compared to the log-rank test (Huang and Thomas, 2014) because the latter evaluates all the data up to the time when the maximum number of events is achieved. Since efficacy studies often include interim analysis for early stopping, one major limitation of the time-specific probability endpoint is that patients need to be followed for a fixed period (up to the landmark time) for endpoint availability, which is a significant operational challenge in multi-center multi-regional clinical trials.

3.1.1 Optimal Designs with Interim Analysis for Time-Specific Probability Endpoint

To address the statistical and operational challenge of the need for conducting interim analysis, statistical designs utilizing the Nelson-Aalen estimator (Nelson 1969) of the time-to-event distribution have been proposed (Lin et al., 1996; Case and Morgan, 2003; Huang,

Talukder and Thomas, 2010; Huang and Thomas, 2014).

Huang et al. (2010, 2014) propose a class of optimal designs for time-specific probability endpoint. The design can be set up in both the single-arm and two-arm randomized settings that allow 1 or 2 interim analyses. Optimality is defined as minimizing the expected sample size or expected the study duration under the null hypothesis. The variance term of the test statistic $Z(x;t)$ depends on the survival distribution which is assumed to be weibull. x is the landmark time and t is the study time. $Z(x;t)$ follows a stochastic Gaussian process and the joint distribution of $Z(x;t_1)$ and $Z(x;t_2)$ is asymptotically bivariate normal with correlation coefficient of $\rho = \sqrt{I(x;t_1)/I(x;t_2)}$ at $t_1 < t_2$, where $I(x;t)$ corresponds to the Fisher information at study time t. Early stopping for either efficacy or futility is built into the optimal design. Another nice feature of the design is that it assumes flexible accrual distribution (piecewise uniform) and allows a brief preplanned pause (≥ 0) in accrual before interim analyses to more efficiently use data from the accrued patients, and to allow focused data collection before the interim analyses. An R package *OptInterim* was developed (https://cran.r-project.org/) that creates the optimal designs and also simulates their properties to check asymptotic approximations and the robustness of the designs to differing conditions.

3.2 Restricted Mean Survival Time

The difference or ratio of restricted mean survival time (RMST) (Irwin, 1949; Zucker, 1998) is an alternative between-group statistical measurement of time-to-event endpoint to the commonly used HR. The RMST $\mu(\tau)$ is the mean survival time truncated by a specific time τ and is simply the area under the survival curve $S(t)$ from $t = 0$ to $t = \tau$.

$$\mu(\tau) = \int_0^\tau S(t)dt \tag{1}$$

where $S(t)$ can be estimated by the Kaplan-Meier method from the actual data. $\hat{\mu}(\tau) = \int_0^\tau \hat{S}(t)dt$ approximately follows a normal distribution with its variance term derived as in Klein and Moeschberger (2005).

The RMST depends on the selection of time τ, which need to be pre-specified to avoid the introduction of bias. Common selections include fixed landmark times of clinical relevance (e.g. x-year), minimum of the largest observed event time in each of the two groups (Trinquart et al., 2016), or minimum of the largest observed time (event or censoring) in each of the two groups.

RMST-based statistical measures do not rely on any model assumptions. Thus, when there is departure from the proportional hazard assumption, the interpretation is still straightforward. In contrast, the HR varies by time and the value derived from the Cox-PH model cannot be interpreted as the average HR across times. Furthermore, unlike median event-free time and time-specific probability endpoint, the RMST can capture the entire event-free distribution up to time τ as the area under the Kaplan-Meier curve. Importantly, both the difference and ratio in RMST provide a clinically meaningful summary of the group difference in a randomized study. Unlike the HR, the difference allows for quantifying the absolute survival difference and the magnitude of clinical benefit. The capability of dual presentation of both relative and absolute measures is an important benefit of using the RMST. The lack of absolute measure from the HR is a major limitation in the evaluation of benefit-risk profile of the experimental drug, particularly when the absolute measure of time to event is small.

One limitation of the RMST is that its value depends on the truncated time τ, which should be clinically meaningful and closer to the end of the study follow-up so that the

majority of survival outcomes will be covered by the time interval. Due to censoring and early events, the number of patients in the later part of the curve is small, resulting in increased variability of the curve shape by a small number of events. A curve based on the RMST over time RMST(t) ($t \in [\tau_1, \tau_2]$) as an alternative summary to the survival function can be considered. The RMST curve can be constructed for each arm and for the difference in RMST between the treatment arm and the control arm. It provides a temporal profile of RMST or difference of RMSTs for evaluating the benefit of the experimental treatment over the control treatment over time and overcomes the restriction of selecting a single truncated time τ. The time interval $[\tau_1, \tau_2]$ can be selected to reflect the window of clinical relevance. For example, τ_1 can be selected as the minimum of (median survival time for the experimental arm, median survival time for the control arm), and τ_2 can be selected as the minimum of (largest observed survival time for the experimental arm, largest observed survival time for the control arm). Zhao et al. (2016) proposed inference based on simultaneous confidence bands for a single RMST curve and also the difference between two RMST curves.

Trial design and sample size calculation solely based on the RMST is not easily attainable, and depends heavily on assumptions. The precise relationship between the amount (and pattern) of censoring and the variance term of RMST is complex with no closed form derivation. In a typical trial with censoring induced by staggered entry of patients, for example, it is unclear how to determine realistic within-group variances (Royston and Parmar, 2011). As a result, it is recommended to design the trial based on the log-rank test and perform the RMST analysis whether or not the proportional hazard assumption is met.

3.3 Immune-Related Response Criteria

Unlike other cancer therapies, immunotherapies may have an indirect effect on cancer cells because of the mechanism of action specific to cancer immunotherapies. As a result, it is not uncommon to observe initial pseudoprogression followed by tumor regression and clinical activity. Response to treatment may demonstrate immune-related patterns that cannot be fully characterized by standard RECIST criteria. First observations of this immune-related response patterns were identified in anti-CTLA4 agents ipilimumab and tremelimumab in melanoma (Wolchok et al., 2008; Healey et al., 2010). Patients were allowed to continue treatment after initial progression due to enlarged lesions or new lesions. The early progression could be enlargement of target lesions, appearance of new lesions, and progression on non-target lesions. Subsequent responses compared to baseline and shrinkage of new lesions were observed in some patients. Some responses were quite durable.

To better quantify the clinical activity observed in patients treated with cancer immunotherapies, immune-related response criteria (irRC) were proposed as criteria for tumor regression, in both unidimensional (irRECIST) and bidimensional (irWHO) measurements (Wolchok et al., 2009; Nishino et al., 2013). Immune-related objective response (irOR) and immune-related PFS (irPFS) can be determined using these criteria. It may be necessary to use irRC and other new criteria in some cases, which may provide higher correlation with OS than the RECIST/WHO criteria, as demonstrated in melanoma. However, more experiences with irRC in patients treated with next-generation immunotherapies and in different indications are needed to better understand the mechanics and whether such criteria offer additional clinical relevant data to assess the benefit-risk of these novel treatments.

3.4 Generalized Pairwise Comparison

Buyse (2010) proposed a generalized pairwise comparison method, an extension of the U-statistic of the Wilcoxon-Mann-Whitney test for the comparison between two groups

of observations. The observations can be outcomes of any type (e.g. discrete, continuous, time to event). Let X_i be the outcome of the i_{th} subject in the treatment arm $(i = 1, \ldots, n)$, Y_j be the outcome of the j_{th} subject in the control arm $(j = 1, \ldots, m)$.

$$U_{ij} = \begin{cases} +1 & \text{if}(X_i, Y_j)\text{pair is favorable} \\ -1 & \text{if}(X_i, Y_j)\text{pair is unfavorable} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Let $U = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}$. Then U is the difference between the proportion of favorable pairs and the proportion of unfavorable pairs, or called the "net chance of a better outcome" Δ. The empirical distribution of test statistic U can be obtained by permuting the treatment labels in order to construct confidence intervals and obtain p-value.

The generalized pairwise comparison method offers an alternative approach to standard non-parametric tests for the two-sample problem. It naturally leads to a patient-relevant, general measure of treatment effect, and allows for testing of differences thought to be clinically relevant. For survival analysis of time-to-event endpoint, this method can be used to assess the benefit-risk of new cancer immunotherapies whether or not the assumption of proportional hazards is met, with Δ being an intuitive measure interpreted as the net chance of longer survival by certain month (Peron et al., 2016). One limitation of this method is the lack of closed analytical form of the test statistic distribution, making derivations of statistical significance and confidence intervals computationally intensive, especially in simulations.

3.5 Weighted Log-Rank Test

The Log-rank test is the most commonly used statistical test for comparing survival curves. It assigns constant 1 as the weight function in the test statistic for each event. Alternatively, weighted log-rank test assigns unequal weights to events, with the choice of weight function being the number of patients at risk, function of time, or a function of the survival distribution. The log-rank test is aimed at detecting a consistent difference between hazards in the two groups and is best placed to have optimum power when the proportional hazard assumption applies.

For immunotherapy agents, as it takes time for immune activation and building of an immune response, the survival curves of the two treatment arms may take a while to separate and the curve tail of the experimental arm may be long due to the durable responses experienced by some patients. The idea of using weighted log-rank test is that putting more weights on the curve tails (late events) may provide us higher statistical power.

Fleming and Harrington (1981) proposed a very general class of tests that includes, as a special case, the log-rank test. Their weight function is given by

$$W_{p,q}(t_i) = \hat{S}(t_{i-1})^p [1 - \hat{S}(t_{i-1})]^q, \;\; p \geq 0, q \geq 0 \quad (3)$$

where $\hat{S}(t_{i-1})$ is the Kaplan-Meier survival function at the previous death time. When $p = q = 0$, we have the log-rank test. When $p = 1$ and $q = 0$, we have a version of the Mann-Whitney-Wilcoxon test. When $q = 0$ and $p > 0$, these weights give the most weight to early departures between the hazard rates, whereas, when $p = 0$ and $q > 0$, these tests give most weight to departures which occur late in time (Table 1).

The benefits of using weighted log-rank test for late separation is that, it will yield potentially higher statistical power. It is also in the same non-parametric testing framework as log-rank test, with log-rank test as the special case. However, there are several limitations. One major concern is the introduction of bias with manipulation of the weight selection, which may face regulatory hurdles in the acceptance of the test result and interpretation of

Table 1: Fleming-Harrington Test.

Parameter values	Test statistic	Weighting summary
$p = 0, q = 0$	Log-rank test	Equal weighting
$p = 1, q = 0$	Wilcoxon test	Most weight on early deviations
$p > 0, q = 0$	Fleming-Harrington test	Most weight on early deviations
$p = 0, q > 0$	Fleming-Harrington test	Most weight on late deviations

Table 2: Statistical challenges and considerations in safety and efficacy evaluation of cancer immunotherapies.

	Challenges	Considerations
Safety	management of late/cumulative toxicity, MTD not identified,	TITE-CRM design with multiple treatment cycles revise the MTD definition and broaden the DLT criteria
	dose optimization	randomized dose-response analysis of safety, efficacy, PK and PD
	combination of agents	statistical designs that account for partial ordering
Efficacy	pseudoprogression, immune-related response, non-proportional hazard, weak/negative correlation between PFS and OS	Alternative methods: time-specific probability endpoint, restricted mean survival time, generalized pairwise comparison, irOR, irPFS, weighted log-rank test, co-primary endpoint of OS and PFS

the data. Using the weighted log-rank test is also a double-edged sword because it may lose power if the curve assumption is wrong (e.g. bigger early separation, smaller late separation). Another major limitation is the lack of corresponding measure of difference as opposed to using the HR under the proportional hazard assumption.

4. Discussion

The clinical development of cancer immunotherapies pose unique statistical challenges in both safety and efficacy evaluations. In this paper, we present some alternative methods and statistical considerations to address these challenges, as summarized in Table 2.

The main challenges in early safety studies include management of late-onset toxicity beyond 1st cycle of treatment, dose optimization and evaluation of combination agents. It is recommended that a Bayesian model-based time-to-event design using the continual reassessment method be implemented to capture late or cumulative toxicities without a significantly prolonged trial and to handle missing data due to patient drop-out. On the other hand, the MTD may not always be identifiable for novel immunotherapy agents, particularly T-cell stimulatory antibodies, and higher dose may not provide clinically improved activity to patients. A pragmatic three-step procedure is proposed to select the recommended dose for further investigation of the benefit-risk profile of the experimental treatment, by prospectively evaluating safety in dose escalation, retrospectively assessing efficacy, PK and PD data, and prospectively testing multiple doses via randomization. The proposed methods have been utilized in clinical practice and are accepted by health authorities.

In efficacy trials of cancer immunotherapies, selection of adequate statistical design, analysis method and hypothesis test is of paramount importance to assess the clinical activity of these novel agents. The traditional methods for the evaluation of chemotherapies and targeted therapies may not be most appropriate or efficient. Alternative methods related to cancer immunotherapies should be considered in suitable settings. Some of these methods include using time-specific probability endpoint, using restricted mean survival time to summarize treatment effect, implementing immune-related response criteria, extending the Wilcoxon-Mann-Whitney statistic for generalized pairwise comparison, and performing weighted log-rank test for hypothesis testing.

There are additional practical approaches to further de-risk in the design of randomized pivotal studies by taking into consideration the unique mechanisms and patterns of efficacy assessment in immuno-oncology. For instance, the timing of interim and final analyses and choice of futility boundaries should be determined with great caution to reduce the risk of inflating the probability of making a false negative decision. Simulations are also important and should be conducted routinely to evaluate the operating characteristics of the trial design and analysis method in various scenarios, including but not limited to large departure from the proportional hazard assumption. Piecewise exponential distribution, weibull distribution or cure rate model can be considered for the survival distribution in different scenarios.

With the ever increasing cost of conducting clinical trials and increasing discovery of molecular subtypes of cancer, efficient designs such as umbrella trials and basket trials are of particular interest as multiple drugs or drug combinations can be assessed in multiple cancer indications or histologies in the same trial. Specifically, a basket trial design facilitates a particular targeted therapeutic strategy (i.e., inhibition of an oncogenically mutated kinase) across multiple cancer types. Examples are NCI's Molecular Analysis for Therapy Choice (MATCH) and the Molecular Profiling based Assignment of Cancer Therapeutics (MPACT, NCT01827384) trials (Conley and Doroshow, 2014). The concept of a basket trial design is ideal in immuno-oncology as immunotherapies targeting the immune system are more likely to work in multiple tumor types. One can utilize a Bayesian approach with hierarchical models to borrow information across tumor types, or a frequentist approach with statistical rigor for a confirmatory study (Chen et al., 2016).

In summary, immuno-oncology is an exciting research area that provides both opportunities and challenges to statisticians. Novel methods have been and are being developed to meet the challenges and to better assess the benefit-risk profiles of cancer immunotherapies. More use of these methods in clinical practice is encouraged to gather more data and experience in order to be accepted by health authorities.

REFERENCES

Guidance Development Review Committee. (2015). Guidance on cancer immunotherapy development in early-phase clinical studies. *Cancer science*, 106(12), 1761.

Cheung, Y. K., and Chappell, R. (2000). Sequential designs for phase I clinical trials with late onset toxicities. *Biometrics*, 56, 11771182.

OQuigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics*, 46, 3348.

Mauguen, A., Le Deley, M. C., and Zohar, S. (2011). Dose-finding approach for dose escalation with overdose control considering incomplete observations. *Statistics in Medicine*, 30, 15841594.

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, 17, 11031120.

Yuan, Y., and Yin, G. (2011). Robust EMcontinual reassessment method in oncology dose finding. *Journal of the American Statistical Association*, 104 , 818831.

Wages, N. A., Conaway, M. R., and OQuigley, J. (2013). Using the time-to-event continual reassessment method in the presence of partial orders. *Statistics in medicine*, 32, 131141.

Huang, B., and Kuan, P. (2014). Time-to-event continual reassessment method incorporating treatment cycle information with application to an oncology phase I trial. *Biometrical Journal*, 6: 933-946.

Huang, B., and Bycott, P., Talukder, E. (2016). Novel Dose-Finding Designs and Considerations on Practical Implementations in Oncology Clinical Trials. *Journal of Biopharmaceutical Statistics* (to appear). DOI:10.1080/10543406.2016.1148715.

Braun, T. (2002). The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials*, 23(3): 240-256.

Thall, P., and Cook, J. (2004). Dose-Finding Based on EfficacyToxicity Trade-Offs. *Biometrics*, 60(3), 684-693.

Fox, E., Curt, G., and Balis, F. (2002). Clinical trial design for target-based therapy. *The oncologist*, 7(5): 401-409.

Piantadosi, S., and Liu, G. (1996). Improved designs for dose escalation studies using pharmacokinetic measurements. *Statistics in medicine*, 15(15): 1605-1618.

Mandrekar, S., Cui, Y., and Sargent, D. (2007). An adaptive phase I design for identifying a biologically optimal dose for dual agent drug combinations. *Statistics in medicine*, 26(11): 2317-2330.

Polley, M., and Cheung, Y. (2008). Two-Stage Designs for Dose-Finding Trials with a Biologic Endpoint Using Stepwise Tests. *Biometrics*, 64(1), 232-241.

Dancey, J., and Chen, H. (2006). Strategies for optimizing combinations of molecularly targeted anticancer agents. *Nature reviews Drug discovery*, 5(8): 649-659.

Harrington, J., Wheeler, G., Sweeting, M., Mander, A., and Jodrell, D. (2013). Adaptive designs for dual-agent phase I dose-escalation studies. *Nature Reviews Clinical Oncology*, 10(5): 277-288.

Kramar, A., Lebecq, A., and Candalh, E. (1999). Continual reassessment methods in phase I trials of the combination of two drugs in oncology. *Statistics in medicine*, 18(14): 1849-1864.

Thall, P., Millikan, R., Mueller, P., and Lee, S. (2003). Dose-Finding with Two Agents in Phase I Oncology Trials. *Biometrics*, 59(3), 487-496.

Huang, X., Biswas, S., Oki, Y., Issa, J., and Berry, D. (2007). A parallel phase I/II clinical trial design for combination therapies. *Biometrics*, 63(2): 429-436.

Yin, G., and Yuan, Y. (2009). A latent contingency table approach to dose finding for combinations of two agents. *Biometrics*, 65(3): 866-875.

Wages, N., Conaway, M., and OQuigley, J. (2011). Continual reassessment method for partial ordering. *Biometrics*, 67:15551563.

Mander, A.P. and Sweeting, M.J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Statistics in medicine*, 34(8), pp.1261-1276.

Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D.R., Steins, M., Ready, N.E., Chow, L.Q., Vokes, E.E., Felip, E., Holgado, E. and Barlesi, F. (2015). Nivolumab versus docetaxel in advanced nonsquamous nonsmall-cell lung cancer. *New England Journal of Medicine*, 373(17), pp.1627-1639.

Motzer, R.J., Escudier, B., McDermott, D.F., George, S., Hammers, H.J., Srinivas, S., Tykodi, S.S., Sosman, J.A., Procopio, G., Plimack, E.R. and Castellano, D. (2015). Nivolumab versus everolimus in advanced renal-cell carcinoma. *New England Journal of Medicine*, 373(19), pp.1803-1813.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., ... Skali, H. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22), 2380-2385.

Hoos, A., Eggermont, A. M., Janetzki, S., Hodi, F. S., Ibrahim, R., Anderson, A., ... Wolchok, J. (2010). Improved endpoints for cancer immunotherapy trials. *Journal of the National Cancer Institute*.

Buyse, M. Generalized pairwise comparisons for prioritized outcomes in the two-sample problem. *Statist Med*, 29: 3245-57, 2010.

Huang, B., and Thomas, N. (2014). Optimal Designs with Interim Analyses for Randomized Studies with Long-Term Time-Specific Endpoints. *Statistics in Biopharmaceutical Research*, 6(2): 175-184.

Huang, B., Talukder, E., and Thomas, N. (2010). Optimal Two-Stage Phase II Designs with Long-Term Endpoints. *Statistics in Biopharmaceutical Research*, 51-61.

European Medicines Agency (2011), Guideline on the Evaluation of Anticancer Medicinal Products in Man. *EMA/CHMP/205/95/rev.4*.

Nelson, W. (1969). Hazard Plotting for Incomplete Failure Data. *Journal of Quality Technology*, 1, 2752.

Lin, D. Y., Shen, L., Ying, Z., and Breslow, N. E. (1996). Group Sequential Designs for Monitoring Survival Probabilities. *Biometrics*, 52, 10331042.

Case, L. D., and Morgan, T. M. (2003). Design of Phase II Cancer Trials Evaluating Survival Probabilities. *BMC Medical Research Methodology*, 3, 112.

Irwin, J. O. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene*, 47, 188189.

Zucker, D.M. (1998). Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93:702–709.

Klein, J.P. and Moeschberger, M.L. (2005). Survival analysis: techniques for censored and truncated data.

Springer Science and Business Media.

Royston, P. and Parmar, M.K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine*, 30(19), pp.2409-2421.

Trinquart, L., Jacot, J., Conner, S.C. and Porcher, R. (2016). Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, p.JCO642488.

Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M.A., Solomon, S.D., Trippa, L. and Wei, L.J. (2016). On the restricted mean survival time curve in survival analysis. *Biometrics*, 72, 215221.

Wolchok, J. D., Ibrahim, R., DePril, V.,... and Hamid, O. (2008). Antitumor response and new lesions in advanced melanoma patients on ipilimumab treatment. In *ASCO Annual Meeting Proceedings*, (Vol. 26, No. 15, p. 3020).

Healey D, Carlson P, Huang B, and Marshall, M. (2010). Clinical outcome of first-line melanoma patients who continue tremelimumab in spite of early disease progression. In *ASCO Annual Meeting Proceedings*, vol. 28, no. 15 p. 2574.

Wolchok, J. D., Hoos, A., O'Day, S., Weber, J. S., Hamid, O., Lebb, C., ... and Humphrey, R. (2009). Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. *CCR*, 15(23), 7412-7420.

Nishino, M., Giobbie-Hurder, A., Gargano, M., Suda, M., Ramaiya, N. H., and Hodi, F. S. (2013). Developing a common language for tumor response to immunotherapy: immune-related response criteria using unidimensional measurements. *CCR*, 19(14), 3936-3943.

Peron, J., Roy, P., Ozenne, B., Roche, L. and Buyse, M. (2016). The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials. *JAMA oncology*.

Fleming, T.R., and Harrington, D.P. (1981). A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics-Theory and Methods*, 763-794.

Conley, B.A., and Doroshow, J.H.(2014). Molecular analysis for therapy choice: NCI MATCH. *Semin Oncol*, 41:2979.

Chen, C., Li, N., Yuan, S., Antonijevic, Z., Kalamegham, R., Beckman, R.A. (2016). Statistical design and considerations of a Phase 3 basket trial for simultaneous investigation of multiple tumor types in one study. *Statistics in Biopharmaceutical Research*, DOI:10.1080/19466315.2016.1193044.