Review Article

Challenges of evaluating immunotherapy efficacy in solid tumors

Rilan Bai, Wenqian Li, Nawen Du, Jiuwei Cui

Cancer Center, the First Hospital of Jilin University, Jilin 130021, China

Correspondence to: Jiuwei Cui. Cancer Center, the First Hospital of Jilin University, Jilin 130021, China. Email: cuijw@jlu.edu.cn.

Abstract

Immunotherapy is one of the most promising treatments for multiple tumor types. The significant clinical benefits and durable responses of immunotherapy have led to the emergence of various immune-related clinical response patterns that extend beyond those achieved with cytotoxic agents. Various studies investigated the efficacy of immunotherapy, including the effect on tumor size, long-term survival benefits, and the ability to overcome the particularly challenging survival curves tailing phenomenon. The current immune-related methods guidelines, such as immune-related Response Criteria (irRC), immune-related Response Evaluation Criteria in Solid Tumors (irRECIST), immune Response Evaluation Criteria in Solid Tumors (iRECIST), and immune-modified Response Evaluation Criteria in Solid Tumors (imRECIST), could be well-adapted to identify the heterogeneity of responses that appear in patients receiving immunotherapy, such as pseudoprogression (PsPD) and hyperprogressive disease (HPD), and to some extent to overcome the limitation of evaluating the efficacy of immunotherapy on tumor size by imaging. Additionally, a second type of evaluation method was proposed based on survival, which includes milestone analysis and restricted mean survival time. Currently, milestone analysis is a complementary tool to summarize and interpret trial results along with more conventional measures of survival and other less established metrics. A golden standard evaluation method to distinguish the efficacy of immunotherapy may improve the process of imaging and aid survival-based efficacy evaluation in patients with solid tumors.

Keywords: Neoplasms; immunotherapy; pseudoprogression; response evaluation criteria; milestone analysis

Submitted Apr 29, 2019. Accepted for publication Sep 04, 2019.
doi: 10.21147/j.issn.1000-9604.2019.06.02
View this article at: https://doi.org/10.21147/j.issn.1000-9604.2019.06.02

Introduction

Traditionally, the most direct method of evaluating treatment efficacy involves the imaging of changes in the tumor size. However, in recent years, extensive researches have been conducted on immunotherapy which appears to be a promising treatment with significant clinical benefits in multiple cancer types (1,2). In consideration of their peculiar mechanism, immunotherapies can determine atypical response patterns or be called immune-related clinical response patterns that extend beyond those of cytotoxic agents, such as pseudoprogression (PsPD) (3), delayed responses (4), hyperprogressive disease (HPD) (5,6), etc. The patterns mentioned above had made clinicians and diagnostic radiologists increasingly face a great challenge in evaluating the clinical efficacy of these novel treatments using imaging accurately, particularly when the tumor burden increases from the initial or new lesions appear in imaging. Thus, it seems that it may not be possible to completely and accurately assess the efficacy of immunotherapy. In addition to the challenges associated with the evaluation of tumor size, recent developments have been made in assessing the objective response rate (ORR), long-term survival benefits, and overcoming the tailing phenomenon of the survival curves difficulty. Additionally, assessing the traditional reasonable endpoint in clinical trials, that is, overall survival (OS), has become difficult. This suggests that current techniques cannot accurately quantify the proportion of patients who are "cured" and thus these methods are likely suboptimal to determine the accurate effects of immunotherapy. Therefore, the emergence of these problems and challenges in evaluating the efficacy of immunotherapy for treating solid tumors prompt the development of novel efficacy evaluation and survival analysis methods based on imaging evidence, survival data and exploratory studies.

Immune-related evaluation criteria for tumor efficacy based on imaging evidence

Immune-related Response Criteria (irRC)

A phase II clinical trial of ipilimumab for the treatment of advanced melanoma, conducted by Wolchok *et al*. (7) demonstrated a special example of early PsPD, in which the patient experienced disease progression at the first assessment (month 3) after ipilimumab administration but the lesions regressed a month later after therapy continuation. Complete response was achieved at 2 years. Based on this result, Wolkok *et al*. (7) comprehensively tested the feasibility of immune-related evaluation criteria and confirmed the existence of PsPD. Thus, in 2009, Wolchok *et al*. (7) formally proposed the irRC based on the World Health Organization (WHO) criteria and explained in detail the proposed new concepts, guiding principles, and clinical applications. These criteria differ from the WHO criteria in the introduction of tumor burden, the restatement of measurable new lesions (≥5 mm × 5 mm; ≥ 10 visceral target lesions and 5 skin target lesions; ≥5 lesions per organ), and the concept of calculating new lesions into the total tumor burden (8). For the appearance of new lesions, the irRC consider that progressive disease (PD) would not be assigned as long as the total tumor burden does not increase by >25%, and patients with a stable clinical status would be recommended treatment continuation and reassessed after at least 4 weeks. As the first immune-related efficacy evaluation guidelines, the irRC have made new regulations in the definition and division of new lesions and PD. Subsequent clinical trials have also confirmed their unique superiority.

Through comparing irRC with Response Evaluation Criteria in Solid Tumors v1.1 (RECIST v1.1), Kim *et al*. (9) reveal atypical patterns of immunology in patients with non-small-cell lung cancer (NSCLC). The results showed that two patients (4.9%) with existing PsPD were assessed with PD by RECIST v1.1, but non-PD by irRC. Similarly, Hodi *et al*. (10) conducted a phase Ib study on patients with advanced melanoma receiving pembrolizumab to compare the two sets of criteria. Of the 655 patients in their study cohort, 327 had ≥28 weeks of imaging follow-up, of whom 24 (7.4%) had unusual responses, including 15 (4.6%) with early PsPD and 9 (2.8%) with delayed response. Moreover, the trial compared OS among three groups of patients; namely, those with non-PD as per both sets of criteria (group I), those with PD as per RECIST v1.1 but non-PD as per irRC (group II), and those with PD as per both sets of criteria (group III). The median OS has not yet been reached in group I [95% confidence interval (95% CI): 25.9 months to not reached], whereas it was 22.5 months in group II (95% CI: 16.5 months to not reached) and 8.4 months in group III (95% CI: 6.6−9.9 months). The 2-year OS rates were 77.6%, 37.5% and 17.3%, respectively in the three groups. These trials above showed that the traditional guidelines underestimate the benefits of immunotherapy and ignore the information of continuous treatment in patients with initial PD, whereas the irRC could actually assess the efficacy of immunotherapy and avoid the too-early termination of treatment, thereby preventing the lost opportunity of effective therapy for patients with a potential response.

Immune-related RECIST (irRECIST)

Although the irRC were developed to try to evaluate the tumor response to immunotherapy as adequately as possible, they are based on bidimensional measurements, not the unidimensional measurements defined by the RECIST v1.1 guidelines that have been widely used in solid tumors. Therefore, in 2013, Nishino *et al.* (11) conducted a study to compare the response assessments between unidimensional irRC and bidimensional irRC measurements in patients with advanced melanoma receiving ipilimumab. The interobserver variability of unidimensional *vs.* bidimensional measurements was assessed in 25 randomly selected patients, with results showing the percentage changes of measurements at follow-up, the immune best overall response (iBOR), and the time to progression as being highly concordant between the two sets of criteria. However, the unidimensional measurements (95% CI: −16.1%, 5.8%) were more repeatable than the bidimensional measurements (95% CI: −31.3%, 19.7%). Therefore, the use of unidimensional irRC was proposed for assessments of the response to immunotherapy in solid tumors, given the simplicity, higher repeatability, and higher concordance with the bidimensional irRC. Moreover, other studies had confirmed that the bidimensional measurements would exaggerate the actual degree of tumor change in size to some extent (especially when the lesion changes are actually very small) (12), resulting in many patients being mistakenly assessed with PD. Because of such cases, the irRC have not been widely applied ever since their proposal, promoting the updating process for new guidelines.

On the bases of the irRC and the study by Nishino *et al.* (11), researchers first proposed the irRECIST at the 2014 European Society for Medical Oncology (13). The guidelines extended the unidimensional measurements of RECISTv1.1 and several concepts of irRC; for example, new lesions should be added into the original tumor burden; non-target and new lesions also have a reference value when assessing PD; cases initially assessed as PD should be reevaluated after at least 4 weeks, and so on. Disappointingly, however, although the irRECIST guidelines have been applied in clinical trials ever since their proposal, satisfactory results have not been observed.

With the rapid development of immune checkpoint inhibitors (ICIs), the evaluation of immune efficacy has once again encountered challenges, bringing with them a lot of questions, such as whether all new lesions need to be measured, how the first time point of evaluation and the iBOR should be established, what the detailed principles of evaluation after the appearance of new lesions should be, etc. All these questions prompted the need for new immune-related criteria that would be more in line with clinical practice.

Immune Response Evaluation Criteria in Solid Tumors (iRECIST)

In early 2017, the RECIST Working Group formally proposed the iRECIST (14) to provide a consensus guide and standardized data management, collection, and analysis criteria for clinical trials of immunotherapy.

The guide proposes new terminology for efficacy evaluation, and has the prefix “i” (for “immune”) to differentiate the responses from those assigned by RECISTv1.1.; for example, “immune” complete response (iCR), “immune” partial response (iPR), “immune” stable disease (iSD), “immune” unconfirmed progressive disease (iUPD), and “immune” confirmed progressive disease (iCPD). The most prominent change is the aspect of resetting the bar if progression as per RECISTv1.1 is followed by tumor shrinkage at the next evaluation, and the introduction of the two key concepts iUPD and iCPD. That is, the RECISTv1.1-defined progression is first temporarily regarded as iUPD, and then evaluation for treatment continuation or discontinuation is made according to the clinical condition of the patient and the type and pathologic stage of the tumor. Finally, the revaluation is performed again after 4−6 weeks to confirm the iCPD status. Notably, under this mode, iSD, iPR, or iCR can reappear after iUPD, whereupon the bar would be reset, iUPD is again assigned, and iCPD is confirmed at the next evaluation. That is, as long as iCPD is not confirmed, cyclical evaluations are needed and the causes of unconfirmation should be recorded.

Each time point response using iRECIST guidelines requires a comprehensive analysis based on the assessment of target lesions, non-target lesions, and new lesions. For new lesions, many concepts of evaluation are unique with iRECIST. Measurable new lesions can be categorized as target lesions (≤5 lesions, ≤2 lesions per organ; which should not be included in the baseline of initial target lesions) or non-measurable lesions (with which the other measurable lesions are marked as new non-target lesions). There are two conditions for confirming iUPD on the basis of new lesions; that is, observing a further increase in the new target-lesions size on the basis of iUPD (sum of measures increase ≥5 mm) or any unambiguous increase of new non-target lesions. For target lesions, PD is first defined by RECISTv1.1 as iUPD, and is confirmed (after 4−8 weeks) as iCPD with a further increase in the sum of measures to at least 5 mm. The assessment of non-target lesions at each time point follows the similar guidelines. Generally, the confirmation of iUPD requires a further increase in the size or number of the lesion categories in which progression was identified for the first time, or an unambiguous progression in lesions that had not met progression as per RECISTv1.1 before, or the appearance of new lesions.

The efficacy evaluation of iRECIST can be categorized into the following three results: 1) confirmed iCPD, observing an increase of non-target/target lesions in the sum of the longest diameters of >5 mm, an increase in new non-target lesions, or the appearance of other new lesions; 2) assigned iCR, iPR, or iSD, where the status is reset if the lesions decrease to the corresponding RECISTv1.1-defined measurements; 3) and if no change is recorded, then the time-point response is still iUPD. In addition, the iRECIST guidelines clarify the iBOR, which is to record the best time point response from the beginning of the study treatment to the end. As long as iUPD is not confirmed, the best response is regarded as iBOR, and once iCPD is detected, it should be regarded as iBOR. The cyclical evaluation model innovatively proposed by the iRECIST could capture the emergence of atypical responses (e.g., PsPD and delayed response) in the era of immunotherapy. Recently, a clinical study conducted by Tazdait *et al*. (15) confirmed as much. It compared the discordance among the RECISTv1.1, irRECIST, and iRECIST guidelines in patients with advanced NSCLC, using ICIs, and it showed that 13% (20/160) of the patients had an unconventional response (5% PsPD and 8% dissociated responses). According to the survival analyses, RECISTv1.1 underestimated the benefits of immune checkpoint inhibitors in 11% (13/120) of the patients with PD. However, irRECIST and iRECIST could distinguish those unusual responses, with a 3.8% discrepancy rate.

The differences between iRECIST and the traditional standards include the assessment of new lesions, confirmation of PD, and consideration of the patient's clinical status. The RECISTv1.1 guidelines define the appearance of new lesions as PD directly, without measurement, whereas the iRECIST definition initially regards these as iUPD and then could confirm them as iCPD as long as the conditions are met again. The iRECIST guidelines propose to comprehensively consider the clinical status of the patients before all decisions regarding therapy continuation are made. Furthermore, the guidelines recommend a shorter period (4–6 weeks) before the next imaging assessment, to ensure that patients still have access to salvage therapies. However, a longer time window might be rational if PsPD is well caught in the tumor type, such as the CTLA4 inhibitor for melanomas, especially if there are no effective salvage therapies obtainable (e.g., for BRAF wild-type melanomas). If the researchers or patients believe it is appropriate to continue the treatment when iCPD is confirmed, then the continuation of data collection is recommended to further clarify the tumor growth dynamics with immunotherapy, and disease assessments should be continued until other therapies are started.

The introduction of iRECIST was a leap forward in the era of immunotherapy, being a consensus guideline for a standardized data management and analysis system for ongoing clinical trials. However, the problems that extend from iRECIST cannot be ignored. For example, many patients have been excluded from research owing to PD, resulting in many clinical problems (including HPD) being laid aside as a result of insufficient data. In addition, quantification of the differences in evaluation results between RECISTv1.1 and iRECIST should provide more valuable recommendations for revising iRECIST and addressing more clinical issues in future.

Immune-modified Response Evaluation Criteria in Solid Tumors (imRECIST)

With the widespread development of new immune agents, there is an urgent need to adapt unconventional responses by carrying out tumor assessments, and thereby require a complementary assessment of efficacy benefits with substitute endpoints [i.e., progression free survival (PFS), ORR] that usually mature well before OS. Massive data up until now have proven that these response measures often underrate survival benefits when the RECISTv1.1 guidelines are employed (16-18). Work to settle the inconsistency of unidimensional measurement based on the RECISTv1.1 framework began with the development of the irRC and has been extended with the imRECIST (19).

In brief, the imRECIST guidelines allow multiple reviews and BOR occurrence after PD assessment in imaging with patients continuing treatment. New lesions are added to the total tumor burden, as well as the sum of the target lesions when they are measurable. Lesions beyond measurement are not added to the PD evaluation. Moreover, PD is not defined by progression in non-target lesions. In the analysis of PFS defined by imRECIST (imPFS), imPD or death is regarded as a key event, except the situation that the time-point response is an SD, PR, or CR defined by imRECIST at the next scan (4 weeks later) after first regarded as imPD. In fact, as early as 2016, Mazieres *et al*. (20) compared the RECISTv1.1 and imRECIST guidelines in phase II POPLAR trial, which studied the continuing treatment of 144 patients with traditional progression after atezolizumab therapy. The results showed that compared with the measures made according to RECISTv1.1, the PFS increased by 1.5 months, OS increased slightly by 2%, the disease control rate increased by 13%, and the assessment of PD decreased by 16% when evaluated by imRECIST, which laid the foundation for the proposition of imRECIST.

Treatment beyond progression

PsPD is described as an increased lesion size or the visualization of new lesions, which might be followed by a durable response. Although well described, it could be challenging to differentiate transient PsPD from true progression. Therefore, we recommend clinical studies in which treatment beyond progression (TBP) as per

RECISTv1.1 (i.e., iUPD) would only permit patients with stable clinical status to continue on therapy until the next scan (≥4 weeks), which confirmed the opportunity for potential effective salvage therapy for patients with a non-response assessment. In a study that conducted TBP in 121 patients with advanced NSCLC after ICIs, the results showed that 10 (8.3%) patients had an additional tumor decrease by more than 35% (35%−100%, median value was 58%), and the best responses were 4 cases of PR, 2 of SD, and 4 of PD. At least 5 of the 10 patients responded for at least 6 months, and 3 patients for up to 1 year, and the median duration of response has not yet been reached (95% CI: 1.3 months to not reached) (21). The OAK study group presented the results of TBP after atezolizumab therapy for advanced NSCLC during the 2017 annual meeting of the American Society of Clinical Oncology (ASCO), reporting that 7% of 168 patients showed a decrease in target lesions and 49% had stable target lesions. The OS of these patients was 12.7 (95% CI: 9.0−14.9) months. In addition, the safety risk of continued therapy did not increase compared with chemotherapy, which showed that the patients had tolerated well to treatment and the benefit-to-risk ratio was high (18).

Although several published studies have achieved a certain degree of benefits from TBP, it is yet ambiguous whether tumor shrinkage after TBP is attributed to the delayed effects of immunotherapy. In addition, the risks of continued treatment (i.e., immunologically relevant adverse events) should be adequately assessed to balance the risks and benefits. Finally, there is a need to conduct more randomized controlled trials to further explore the biomarkers and to clarify the characteristics of the populations who really benefited from TBP.

Brief summary of response evaluation in imaging

Table 1 summarizes the tumor response assessments obtained according to the irRECIST, iRECIST, and imRECIST guidelines.

Multiple studies have shown that the incidences of PsPD are only 7%−10%. Chiou *et al.* (3) studied the incidence of PsPD in different solid tumors, showing 6.6% (31/471) for melanomas, 1.5% (1/65) for bladder cancers, and 1.8% (3/168) for renal cell carcinomas. In 2017, researchers reported that the overall incidence of HPD was 9%, being higher in elderly patients of over 65 years old (19%) (5). The current immune-related methods, such as irRC, irRECIST, iRECIST and imRECIST, could be well-adapted to the identification of patients with a response or PsPD, making a large proportion of patients no longer continue to receive an ineffective or even harmful treatment. Since pre-treatment data were not integrated

Table 1 Comparison of response categories between irRECIST, iRECIST and imRECIST criteria

Variables	irRECIST	iRECIST	imRECIST
PD	irPD • Increase ≥20% (≥5 mm) in TMTB compared with nadir or progression of non-target lesions or new lesions	iUPD • Increase ≥20% of SLD compared with nadir (≥5 mm) or progression of non-target lesions or new lesions	• Increase ≥20% (≥5 mm) in SLD compared with nadir • Determined only on the basis of measurable diseases
New lesions	• LD will be added to the total measured tumor burden of all target lesions at baseline • Do not correspond to a formal progression	• Do not correspond to a formal progression; New lesions are not incorporated in tumor burden	• New lesions do not categorically define PD • Measurable new lesions are incorporated in the total tumor burden
Confirmed PD	≥4 weeks after the first irPD assessment: • New unequivocal progression or worsened progression from initial PD visit • Appearance of another new lesion	≥4 weeks after the first iUPD assessment; iCPD: • Increased size of target or non-target lesions • Increase in the sum of new target lesions >5 mm • Progression of new non-target lesions • Appearance of another new lesion	≥4 weeks after the first PD assessment: • If the evaluation is non-PD, update to non-PD

irRECIST, immune-related Response Evaluation Criteria in Solid Tumors; iRECIST, immune Response Evaluation Criteria in Solid Tumors; imRECIST, immune-modified Response Evaluation Criteria in Solid Tumors; PD, progressive disease; irPD, immune-related progressive disease; TMTB, total measured tumor burden; LD, the longest diameter; iUPD, immune unconfirmed progressive disease; iCPD, immune confirmed progressive disease; SLD, sum of the longest diameter.

into an accurate calculation of tumor dynamics, current evaluation patterns might fail to monitor HPD during an early stage of therapy. Moreover, current methods are not able to distinct the true progression, PsPD or HPD, at early stage or within 8 weeks from starting treatment. The use of pre-treatment imaging and estimation of tumor growth rate (TGR) before and after treatment could identify patients with HPD at least at the time of the first disease assessment, considering that these patients are less likely to respond to treatment, another potentially effective treatment should be provided immediately. Regrettably, several challenges currently remain for radiologists and oncologists to perform tumor dynamics testing and TGR calculation in routine clinical practice.

Innovative efficacy evaluation models based on survival

Milestone survival as an intermediate endpoint

Delayed benefits of novel drugs lead to an extended survival in a quiet small patient population. In this scenario, following a traditional non-proportional risk model with reasonable endpoint such as OS is no longer applicable, and replaceable survival endpoints and statistical methods should be explored. Innovative methods such as milestone analysis, restricted mean survival time (RMST), and parametric models (i.e., Weibull distribution, weighted log rank test), should be used in clinical trials to fully quantify the proportion of "cured" patients reflected in the tails of the survival curves.

Milestone survival analysis is a cross-sectional evaluation of OS data at a pre-setting and clinically meaningful time-point (the milestone), such as at 12 months, using Kaplan Meier survival probabilities (22). Milestone survival analysis is generally performed first in a randomized group rather than in the entire population, and it collects patients' information of long-term survival with extended follow-up, while for the rest of the cohort, OS is still used as the primary endpoint. It is worth noting that the chosen milestone may stand for a time point beyond which the researchers consider the treatment benefit likely remain stable, rather than representing long-term survival. The potential benefits of milestone analysis include its simplicity and the ability to capture benefits beyond the median of the Kaplan-Meier curves, and it may be suitable for calculating the non-proportionality of survival curves.

For example, an interim analysis of a phase III trial comparing an immunotherapy agent tremelimumab to chemotherapy as a first-line treatment for patients with advanced melanoma did not observe benefit in the OS. However, prolonged follow-up showed a possible separation of the curves, supporting that the milestone model was sensible used in clinical trials to assess the true survival benefits (23). Another phase III study demonstrated that if the milestone analysis had been conducted in the study design, the efficacy of the experimental treatment would have been determined less than one and a half years ahead of the actual time during the trial (24). The optimal time point for milestone analysis may depend on several factors: the patient population being studied, the disease background, the drugs being studied or their therapeutic category, and the magnitude of effect sought (superiority, or noninferiority, *et al.*). However, the use of the milestone rate as an endpoint in clinical study has several shortcomings, including the incapability to consider the OS curve and the effectiveness of the review prior to the milestone time point.

Restrictive mean survival time (RMST), weighted log-rank test and Weibull distribution

Another emerging novel survival tool is the RMST, also called the t-year mean survival time (25). RMST is the area under the survival curve within a definite timeframe, and also called the average survival time. The effect of the treatment effect between groups can be measured by the differential value or the ratio of RMST, which gives objective and easily understandable results. RMST has been conducted in Checkmate 057 (26) recently. In this study, the median OS in the immunotherapy agent nivolumab group (12.2 months) does not sufficiently describe the long-term survival benefit, which is estimated to be 18% at 3 years with the same hazard ratio (HR) of 0.73 (27). The two survival curves were similar at the first 6 months' follow-up, but when extended at 24 months, the RMST of nivolumab was 13 months *vs.* 11.3 months for docetaxel, with an obviously difference of 1.7 (95% CI: 0.4−3.1) months statistically (P=0.01). The results of this analysis suggest that patients receiving nivolumab would survive about 13 months (28). Thus, the RMST-based method could be used as a primary tool when designing and analyzing comparative studies. In addition, this method also helps clinicians to better interpret the HR clinical significance when the proportional risk model presumption is not satisfied.

The weighted log-rank test (29) and the Weibull distribution (30) represent other parametric survival models, which can be used to analyze non-proportional survival curves of treatments with drugs having delayed clinical and long-term survival benefits. The Weibull distribution in particular fits well to the clinical trials with immunotherapy as it considers the survival curves' different shapes and variation over time.

Summary and future directions

These evaluation methods are constantly evolving to improve their accuracy of efficacy evaluation in tumor treatment and identify patients who can actually benefit from immunotherapy. There are two evaluation methods available to assess the efficacy of immunotherapy, one is the measurement of tumor size; however, recently, it seems to be unable to represent the substantial changes seen in tumors after immunotherapy. In recent years, studies have demonstrated that new predictive markers of therapeutic efficacy in tumors, such as the levels of lactate dehydrogenase and interleukin-8 in peripheral blood and circulating tumor DNA (31,32), can assist in assessing the immune response patterns. Dynamic monitoring of the changes of those markers will facilitate the accurate evaluation of such responses in more patients. In addition, the current study demonstrates that the efficacy evaluation methods for immunotherapy cannot accurately reflect the survival benefits to patients due to its long-term survival benefits and the tailing phenomenon of survival curves. At present, the gold standard endpoint for clinical trials is still OS. Consequently, other evaluation methods based on survival have been proposed, such as milestone analysis and RMST. Although this novel method should not be used as a primary intermediate endpoint in a study, they can serve as a secondary endpoint in future research, particularly for those that use immunotherapy alone or in combinations with other treatments. Milestone analysis in particular could be used as an exploratory or supplementary tool to summarize or illustrate the results along with more conventional measures of survival, as it had been shown that the efficacy of an experimental treatment would have been determined less than one and a half years ahead of the actual time during the trial.

Despite multiple evaluation techniques for evaluating immunotherapy efficacy have been proposed recently, there are still several practical considerations to be made about the current research progress. First, the research on milestone analysis is not deep enough, so future investigation should evaluate the practical working characteristics of milestone survival analysis, including the most suitable size of the cohort and the time boundary for monitoring. Moreover, the efficiency conducted by milestone analysis and its function in the rest cohort on the final OS analysis should be compared with traditional methods. Second, due to the peculiar mechanism of immunotherapies, it is expected that clinical radiologists' professional skill requirements will be higher in the future. They will need to have a deep understanding of different classification criteria of immune response in particular in order to correctly identify the heterogeneity of clinical trials' outcomes. Furthermore, multiple imaging radiotracers are under investigation for giving a better interpretation of unconventional responses patterns of immunotherapy imaging, including several novel PET radiotracers (amino acids, nucleotides, *et al.*). Finally, recent studies have described patient-reported outcomes (PROs) as a new classification method of treatment outcomes. This method has attracted increased attention and advocacy from researchers, as it provides a unique indicator for clinical research and practice to study disease behavior and evaluate treatment efficacy from patients' perspective. However, most phase II studies still do not collect PROs information. As some studies showed that ORR appeared to be related with a recovery of symptoms and PROs, this new mean may still be an important subordinate endpoint in phase II studies, but further research is needed.

Finally, although the studies exploring the efficacy evaluation methods for immunotherapy have made great progress, no uniform standard evaluation method exists, which creats hurdles for further exploration. To enrich the data on survival-based efficacy evaluation, comparison of a number of clinical trials and accumulation of their quantitative data are required, which will provide new reference indicators for data collection and evaluation or analysis of clinical studies in the future. Nonetheless, with the progress of clinical research and the emergence of various efficacy evaluation methods, the present evaluation system is constantly improving and enriching itself, and will guide immunotherapy of solid tumor patients in the future.

Acknowledgements

This study was supported by the National Key Research and Development Program of China (No.

2016YFC1303804), the Key Laboratory Construction Project of Department of Science and Technology of Jilin Province (No. 20170622011JC), the Special Project of Development and Reform Commission in Jilin Province (No. 2017C022), and the Special Project of the National Health and Family Planning Commission of China (No. ZX-07-C2016004).

Footnote

Conflicts of Interest: These authors have no conflicts of interest to declare.

References

1. Song M, Chen X, Wang L, et al. Future of anti-PD-1/PD-L1 applications: Combinations with other therapeutic regimens. Chin J Cancer Res 2018;30: 157-72.
2. Li Y, Sun R. Tumor immunotherapy: New aspects of natural killer cells. Chin J Cancer Res 2018;30: 173-96.
3. Chiou VL, Burotto M. Pseudoprogression and immune-related response in solid tumors. J Clin Oncol 2015;33:3541-3.
4. Borcoman E, Nandikolla A, Long G, et al. Patterns of response and progression to immunotherapy. Am Soc Clin Oncol Educ Book 2018;38:169-78.
5. Champiat S, Dercle L, Ammari S, et al. Hyperprogressive disease is a new pattern of progression in cancer patients treated by anti-PD-1/PD-L1. Clin Cancer Res 2017;23:1920-8.
6. Wang Q, Gao J, Wu X. Pseudoprogression and hyperprogression after checkpoint blockade. Int Immunopharmacol 2018;58:125-35.
7. Wolchok JD, Hoos A, O'Day S, et al. Guidelines for the evaluation of immune therapy activity in solid tumors: immune-related response criteria. Clin Cancer Res 2009;15:7412-20.
8. Nishino M, Tirumani SH, Ramaiya NH, et al. Cancer immunotherapy and immune-related response assessment: the role of radiologists in the new arena of cancer treatment. Eur J Radiol 2015;84:1259-68.
9. Kim HK, Heo MH, Lee HS, et al. Comparison of RECIST to immune-related response criteria in patients with non-small cell lung cancer treated with immune-checkpoint inhibitors. Cancer Chemother Pharmacol 2017;80:591-8.
10. Hodi FS, Hwu WJ, Kefford R, et al. Evaluation of immune-related response criteria and RECIST v1.1 in patients with advanced melanoma treated with pembrolizumab. J Clin Oncol 2016;34:1510-7.
11. Nishino M, Jagannathan JP, Krajewski KM, et al. Personalized tumor response assessment in the era of molecular medicine: cancer-specific and therapy-specific response criteria to complement pitfalls of RECIST. AJR Am J Roentgenol 2012;198:737-45.
12. Nishino M, Giobbie-Hurder A, Gargano M, et al. Developing a common language for tumor response to immunotherapy: immune-related response criteria using unidimensional measurements. Clin Cancer Res 2013;19:3936-43.
13. Bohnsack O, Hoos A, Ludajic K. Adaptation of the immune related response criteria: iRRECIST. Ann Oncol 2014;25:iv361-72.
14. Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol 2017;18:e143-52.
15. Tazdait M, Mezquita L, Lahmar J, et al. Patterns of responses in metastatic NSCLC during PD-1 or PDL-1 inhibitor therapy: Comparison of RECIST 1.1, irRECIST and iRECIST criteria. Eur J Cancer 2017;88:38-47.
16. Kataoka Y, Hirano K, Narabayashi T, et al. Concordance between the response evaluation criteria in solid tumors version 1.1 and the immune-related response criteria in patients with non-small cell lung cancer treated with nivolumab: a multicenter retrospective cohort study. Cancer Chemother Pharmacol 2018;81:333-7.
17. Beer L, Hochmair M, Haug AR, et al. Comparison of RECIST, iRECIST, and PERCIST for the evaluation of response to PD-1/PD-L1 blockade therapy in patients with non-small cell lung cancer. Clin Nucl Med 2019;44:535-43.
18. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet 2017;389:255-65.
19. Hodi FS, Ballinger M, Lyons B, et al. Immune-modified response evaluation criteria in solid tumors (imRECIST): Refining guidelines to assess the clinical

benefit of cancer immunotherapy. J Clin Oncol 2018;36:850-8.

20. Mazieres J, Fehrenbacher L, Rittmeyer A, et al. Non-classical response measured by immune-modified RECIST and post-progression treatment effects of atezolizumab in 2L/3L NSCLC: Results from the randomized phase II study POPLAR. J Clin Oncol 2016;34:15_suppl.9032.
21. Finkel RS, Mercuri E, Darras BT, et al. Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. N Engl J Med 2017;377:1723-32.
22. Zhao S, Zhang Z, Zhang Y, et al. Progression-free survival and one-year milestone survival as surrogates for overall survival in previously treated advanced non-small cell lung cancer. Int J Cancer 2019;144: 2854-66.
23. Ribas A, Kefford R, Marshall MA, et al. Phase III randomized clinical trial comparing tremelimumab with standard-of-care chemotherapy in patients with advanced melanoma. J Clin Oncol 2013;31:616-22.
24. Robert C, Thomas L, Bondarenko I, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N Engl J Med 2011;364:2517-26.
25. Wang ZX, Wu HX, Xie L, et al. Correlation of milestone restricted mean survival time ratio with overall survival hazard ratio in randomized clinical trials of immune checkpoint inhibitors: A systematic review and meta-analysis. JAMA Netw Open 2019;2: e193433.
26. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627-39.
27. Haanen JBAG, Carbonnel F, Robert C, et al. Management of toxicities from immunotherapy: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv264-6.
28. Pak K, Uno H, Kim DH, et al. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. JAMA Oncol 2017;3:1692-6.
29. Liu S, Chu C, Rong A. Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. Pharm Stat 2018;17:541-4.
30. Ying GS, Heitjan DF. Weibull prediction of event times in clinical trials. Pharm Stat 2008;7:107-20.
31. Iijima Y, Hirotsu Y, Amemiya K, et al. Very early response of circulating tumour-derived DNA in plasma predicts efficacy of nivolumab treatment in patients with non-small cell lung cancer. Eur J Cancer 2017;86:349-57.
32. Sanmamed MF, Perez-Gracia JL, Schalper KA, et al. Changes in serum interleukin-8 (IL-8) levels reflect and predict response to anti-PD-1 treatment in melanoma and non-small-cell lung cancer patients. Ann Oncol 2017;28:1988-95.

Cite this article as: Bai R, Li W, Du N, Cui J. Challenges of evaluating immunotherapy efficacy in solid tumors. Chin J Cancer Res 2019;31(6):853-861. doi: 10.21147/j.issn.1000-9604.2019.06.02