Introduction

Over the past years, immunotherapy has brought a paradigm shift in the treatment of advanced cancer patients. Nowadays, 26 immunotherapies have gained the approval from regulatory agencies and proofs of benefit have been reported in at least seventeen cancer types (1). In particular, twenty-five indications for six cytotoxic T-lymphocyte antigen-4 (CTLA-4) or programmed death-1 (PD-1) and its ligand PD-L1 inhibitors, have granted Food and Drug Administration (FDA) approval for metastatic solid tumors from March 2011 to August 2017 (2) (Table 1). In addition, several combinatorial treatment strategies are currently being tested. Nearly
1,502 clinical trials are evaluating PD-1/PD-L1 inhibitors in cancer patients, of these 1,105 are combination studies of anti PD-1/PD-L1 agents with other immunotherapies, targeted therapies, chemotherapies or radiotherapies (1). Interestingly, the rapidity of clinical trial enrollment and regulatory agencies accelerated approval has left many unsolved issues to explore in the next wave of immuno-oncology trials. Specifically, relevant unanswered questions concern the optimal study design, endpoints and statistical methods for evaluating immunotherapeutic drugs, the appropriate radiological assessment of antitumor responses, the development of predictive biomarkers and

<table>
<thead>
<tr>
<th>Drug</th>
<th>Indications</th>
<th>Line</th>
<th>Primary endpoint</th>
<th>FDA approval</th>
<th>EMA approval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ipilimumab</td>
<td>Melanoma</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>Any in combination with nivolumab</td>
<td></td>
<td>PFS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td>Nivolumab</td>
<td>Melanoma</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>NSCLC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>RCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>SCCHN</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>UCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>MSI-H CRC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td>Pembrolizumab</td>
<td>Melanoma</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>NSCLC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>PFS</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>1&lt;sup&gt;st&lt;/sup&gt; in combination with chemotherapy</td>
<td></td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>SCCHN</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>UCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>MSI-H any histology</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Gastric/GEJ</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td>Atezolizumab</td>
<td>NSCLC&lt;sup&gt;a&lt;/sup&gt;</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>OS</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td></td>
<td>UCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
<tr>
<td>Durvalumab</td>
<td>UCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td>Avelumab</td>
<td>UCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>–</td>
</tr>
<tr>
<td>Merkel carcinoma</td>
<td>UCC</td>
<td>≥2&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>ORR</td>
<td>Approved</td>
<td>Approved</td>
</tr>
</tbody>
</table>

<sup>a</sup>, after progression to sorafenib; <sup>b</sup>, in patients with PD-L1 expression ≥1%; <sup>c</sup>, in EGFR and ALK wild type patients with PD-L1 expression ≥50%; <sup>d</sup>, platinum-based chemotherapy; <sup>e</sup>, both squamous and non-squamous histology; NSCLC, non-small cell lung cancer; RCC, renal cell carcinoma; UCC, urothelial carcinoma; CRC, colorectal cancer; MSI-H, microsatellite instability high; SCCHN, squamous cell carcinoma of the head and neck; GEJ, gastroesophageal junction; ORR, overall response rate; PFS, progression-free survival; OS, overall survival; FDA, Food and Drug Administration; EMA, European Medical Agency.
the harmonization of the assays to test these biomarkers in large patient populations. Most of these issues are related to the intrinsic mechanism of action and kinetic of immune checkpoint inhibitors (ICI). Differently from chemotherapy and targeted agents, ICI induce a continuum of biological events that starts early with immune system activation and that procrastinates until the ideal obtainment of a (sometimes) delayed clinical benefit. This peculiar feature should be carefully considered when designing clinical trials with ICI and innovative study methodologies should be applied to appropriately assess the delayed effect of immunotherapeutic agents in terms of responses and survival benefit.

This review will explore the major methodological issues and challenges regarding endpoints, statistical methods, predictive biomarkers assessment and clinical trials design for ICI, focusing in particular on non-small cell lung cancer (NSCLC) patients.

**Methodology and endpoints in clinical trials with ICI**

In the immune-oncology era, traditional endpoints of randomized clinical trials, such as objective response rate (ORR) according to Response Evaluation Criteria in Solid Tumors (RECIST 1.1) (3), median progression-free survival (PFS) and overall survival (OS) have been extensively questioned. In fact, although they are appropriate for assessing the activity of agents able to induce rapid control of tumor growth, such as targeted or cytotoxic therapies, they may be less suitable for treatments, such as immune checkpoint blockade, where tumor control may develop over time. In particular, unconventional response patterns as pseudoprogression (4) or dissociated responses (5), have been recently characterized in different tumor types treated with ICI and they are generally not adequately identified using traditional RECIST 1.1. Similarly, median PFS based on RECIST 1.1 potentially underestimate the activity of ICI in patients with prolonged stable disease or unconventional responses, and median OS or hazard ratios (HR) are largely suboptimal to capture the key attributes of immunotherapeutic agents, such as delayed clinical effect and long-term survival (6). For these reasons, alternative response evaluation criteria, namely irRC (7), irRECIST (8) and iRECIST (9) and innovative statistical models, such as milestone analysis (10,11), weighted log-rank test (12), restrictive mean survival time (13) and Weibull distribution (14) are currently under development in order to assess delayed effect and prolonged survival of ICI.

Moreover, the conventional oncology specific frameworks, traditionally used to estimate the value of cancer drugs, should also be modified, taking in account the new concept of durable clinical benefit. The American Society of Clinical Oncology (ASCO) has recently published an update of the original framework, incorporating the evaluation of long-term survival. Specifically, bonus points were awarded if the experimental regimen resulted in at least a 50% relative improvement in the percentage of patients alive at a time point corresponding to twice the median OS or PFS point for the control regimen and if at least 20% of patients receiving the control regimen were alive at this time (15). This novel ASCO framework could properly assess the clinical benefit of immunotherapies and recently it was used to review FDA approvals for ICI (2). Interestingly, only 3 out of 23 indications examined gained the long tail bonus points, namely second line ipilimumab and first line nivolumab for metastatic melanoma, second line nivolumab for squamous advanced NSCLC. Considering that 9 out of 23 approvals achieved the 50% improvement in patients alive in the experimental regimen compared with the standard treatment but did not receive bonus points because less than 20% of patients were alive in the control arm, definitive conclusions on where setting the bar to define a significant survival improvement with new immune-oncology agents are difficult to be drawn (16).

**Objective overall response rate**

Although it is a common belief that survival represents the main endpoint for regulatory agency approvals, 15 (60%) out of 25 FDA approvals for ICI were based on ORR as primary endpoint (2). Interestingly, in some patients treated with ICI, initial disease progression assessed by conventional tumor response criteria, such as WHO criteria (17) or RECIST 1.1 (3), may be followed by prolonged clinical stabilization or partial/complete responses. This phenomenon defined as pseudoprogression is caused by T-cell tumor infiltration as a result of immune activation and was described both with anti-PD-1/PD-L1 and anti-CTLA-4 agents in advanced melanoma patients (18,19) and with PD-1/PD-L1 inhibitors in advanced renal cell carcinoma (RCC) (20) and NSCLC patients (5,21,22). The emerging of pseudoprogression and dissociated responses to ICI brought to the development in 2009 of immune related response criteria (irRC) (7). The key differences compared to RECIST criteria were the...
introduction of bidimensional measurements (sum of products of the two largest perpendicular diameters), the inclusion of new lesions [usually classified as progressive disease (PD) according to RECIST 1.1] in the total tumor burden and the requirement of confirmation of PD on two consecutive scans at least 4 weeks apart. Subsequently, unidimensional irRC (irRECIST), which used the longest diameter measurements as in RECIST, demonstrated high concordance compared to bidimensional irRC, bypassing the methodological issues linked to the use of bidimensional measurements (8). Finally, the RECIST working group has recently developed a guideline for the use of modified RECIST (named iRECIST) in order to establish a common framework for the management of data from clinical trials with ICI (9). As irRECIST, iRECIST introduced the concept of immune unconfirmed PD (iUPD) which consents to reset the bar if RECIST progression is followed at the next assessment by tumor shrinkage. Basically, the main difference between irRECIST and iRECIST regards the new lesions, which are incorporated into the sum of target lesions in irRECIST while in iRECIST are recorded separately. However, high concordance has been recently reported between irRECIST and iRECIST in a retrospective study including advanced NSCLC patients treated with anti PD-1/PD-L1 agents. Interestingly, for only ~4% of NSCLC patients there was a mismatch between irRECIST and iRECIST, where iRECIST interpretation as iUPD led to unnecessary continuation of immunotherapy (5). To date, few clinical trials have used irRC/irRECIST as secondary response endpoints (23-25) and none has used iRECIST as response criteria to define their endpoints. Therefore, the regulatory agencies continue to base the approvals of new ICI on RECIST 1.1 defined outcomes. In the future, the integration and validation of irRECIST/iRECIST in clinical trials will be of paramount importance in order to provide to immunologists a practical and reliable tool to face the dilemma about whether and when continue immunotherapy beyond progression.

Another emerging challenge for immunotherapy trials is represented by the evaluation of accelerated tumor growth under ICI, a phenomenon known as hyperprogression (HPD) and recently described in 9% of advanced cancer patients (26), in 29% of head and neck cancers (27) and in 14% of NSCLC patients treated with ICI (28). Although each study used different methodologies to assess HPD, all of them highlighted the importance of measuring tumor growth speed on consecutive computed topography (CT) scans, before the start and during immunotherapy treatment. Retrospective evaluation of HPD in published randomized studies is actually difficult because the CT scans data before immunotherapy start are usually not captured. Therefore, a prospective assessment of HPD in adequately designed clinical trials, which collect CT scans before and during ICI and adopt innovative radiologic tools to quantify tumor kinetics and dynamics over time, will provide a confirmatory evidence regarding this rapid and atypical phenomenon. Finally, the use of ORR as a surrogate endpoint for OS in trials with ICI remains an unsolved question. A meta regression analysis of seventeen randomized trials testing ICI showed a weak but statistically significant correlation between the treatment effect on the ORR and the treatment effect on survival outcomes (i.e., OS and PFS) and suggested that the activity of ICI in terms of ORR explain ~50% of the effects detected in survival (29). Conversely, a systematic review of ten clinical trials evaluating PD-1/PD-L1 inhibitors in advanced NSCLC failed to show a significant correlation between response and survival (30). Considering that ICI activity potentially leads to prolonged disease stabilization and/or unconventional responses, it is likely that disease control rate (DCR), including both responses and tumor stabilization for at least 6 months of treatment (clinical benefit), may be a more clinically relevant surrogate endpoint for survival compared to ORR. The potential future validation of ORR or clinical benefit as surrogate endpoints for survival may consent an earlier analysis of trial data, allowing less expensive and prolonged studies and, most of all, rapidly addressing progressive patients towards other treatments.

**PFS**

Before the coming of ICI in the cancer treatment scenario, PFS has traditionally been considered a reasonable endpoint for new drug approval in a series of solid tumor, including lung cancer. Unlike OS, PFS is not influenced by post-progression therapies and it can provide an earlier assessment of efficacy and a direct measure of treatment effect, avoiding bias related to crossover (31). In many scenarios (32), and in particular in locally advanced lung cancer, a significant correlation between PFS and OS has been demonstrated (33). As a general rule, being OS the sum of PFS and survival post progression (SPP), the longer is the SPP, the lower is the chance that PFS and OS correlate (34). Regarding oncogene addicted NSCLC patients treated with targeted agents, the ratio between the HR for PFS and the
Table 2 HR for PFS and OS and HR PFS/OS rate for the main randomized phase II and III clinical trials of single agent ICI in advanced NSCLC patients

<table>
<thead>
<tr>
<th>Trial</th>
<th>Phase</th>
<th>Line</th>
<th>ICI</th>
<th>Treatment arms</th>
<th>Patients (n)</th>
<th>mOS (month)</th>
<th>HR OS</th>
<th>mPFS (month)</th>
<th>HR PFS</th>
<th>HR PFS/OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checkmate 017 (37)</td>
<td>III</td>
<td>≥2nd</td>
<td>Nivolumab</td>
<td>Nivolumab 3mg/kg q2 weeks; docetaxel</td>
<td>135; 137</td>
<td>9.2; 6</td>
<td>0.59</td>
<td>3.5; 2.8</td>
<td>0.62</td>
<td>1.05</td>
</tr>
<tr>
<td>Checkmate 057 (38)</td>
<td>III</td>
<td>≥2nd</td>
<td>Nivolumab</td>
<td>Nivolumab 3mg/kg q2 weeks; docetaxel</td>
<td>292; 290</td>
<td>12.2; 9.4</td>
<td>0.73</td>
<td>2.3; 4.2</td>
<td>0.92</td>
<td>1.26</td>
</tr>
<tr>
<td>Checkmate 026 (39)</td>
<td>III</td>
<td>1st</td>
<td>Nivolumab</td>
<td>Nivolumab 3mg/kg q2 weeks; platinum-based CT</td>
<td>271; 270</td>
<td>14.4; 13.2</td>
<td>1.02</td>
<td>4.2; 5.9</td>
<td>1.15</td>
<td>1.12</td>
</tr>
<tr>
<td>KEYNOTE 010 (25)</td>
<td>II/III</td>
<td>≥2nd</td>
<td>Pembrolizumab</td>
<td>Pembrolizumab 10 mg/kg q3 weeks; pembrolizumab 2 mg/kg q3 weeks; docetaxel</td>
<td>346; 345; 343</td>
<td>12.7; 10.4; 8.5</td>
<td>0.71*</td>
<td>4; 3.9; 4</td>
<td>0.88*</td>
<td>1.23</td>
</tr>
<tr>
<td>KEYNOTE 024 (40)</td>
<td>III</td>
<td>1st</td>
<td>Pembrolizumab</td>
<td>Pembrolizumab; platinum-based CT</td>
<td>154; 151</td>
<td>30; 14.2</td>
<td>0.63</td>
<td>10.3; 6</td>
<td>0.50</td>
<td>0.79</td>
</tr>
<tr>
<td>POPLAR (41)</td>
<td>II</td>
<td>≥2nd</td>
<td>Atezolizumab</td>
<td>Atezolizumab 1,200 mg q3 weeks; docetaxel</td>
<td>144; 143</td>
<td>12.6; 9.7</td>
<td>0.77</td>
<td>2.7; 3</td>
<td>0.98</td>
<td>1.27</td>
</tr>
<tr>
<td>OAK (42)</td>
<td>III</td>
<td>≥2nd</td>
<td>Atezolizumab</td>
<td>Atezolizumab 1,200 mg q3 weeks; docetaxel</td>
<td>425; 425</td>
<td>13.8; 9.6</td>
<td>0.73</td>
<td>2.8; 4</td>
<td>0.95</td>
<td>1.31</td>
</tr>
</tbody>
</table>

*, comparison pembrolizumab 2 mg/kg q3 weeks versus docetaxel; ICI, immune checkpoint inhibitors; q, every; mOS, median overall survival; HR, hazard ratio; mPFS, median progression free survival; CT, chemotherapy.

HR for OS is usually lower than 1 indicating that a larger benefit in PFS can translate in smaller advantages in OS. In the immunotherapy era this paradigm has been revolutionized (35). In fact, ICI usually induce a delayed clinical benefit that is not always adequately captured by PFS based on conventional RECIST 1.1, whereas it significantly improves OS. Therefore, in randomized trials testing ICI in advanced NSCLC patients, HR for PFS is generally lower than HR for OS and the ratio between them is higher than 1 (range, 1.05–1.31) with the exception of studies comparing pembrolizumab to platinum based chemotherapy in NSCLC patients. In the Checkmate 017 (37) study comparing pembrolizumab to chemotherapy, the HR for OS was 0.59 and the HR for PFS was 1.05, indicating a significant improvement in OS with pembrolizumab compared to chemotherapy.
was reported for the first time in clinical trials evaluating cancer vaccines, such as the phase III study comparing sipuleucel-T, an autologous active cellular immunotherapy, to placebo in advanced prostate cancer patients, where the effect on survival was not evident for the first 8 months of treatment (44). Similarly, also phase III trials of CTLA-4 or PD-1/PD-L1 inhibitors in advanced melanoma (45) and NSCLC (38,42) showed delayed separation of survival curves, or even a cross between them with an initial better survival outcome for chemotherapy compared to ICI, as observed in Checkmate 057, a phase III trial comparing nivolumab versus docetaxel in pretreated non-squamous NSCLC patients (38). Recently, an update of the phase I CA209-003 trial testing nivolumab in 129 previously-treated NSCLC patients showed that 5-year OS was 16% for squamous and 15% for non-squamous patients (46), however the 9.9 months of median OS did not adequately estimate the durable benefit demonstrated by the plateau in the tails of the survival curves. In Figure 1 is reported the hypothetical survival curve of a treatment (i.e., immunotherapy) that leads to long-term survival in a small proportion of patients (green line) compared to a standard therapy, potentially a cytotoxic agent, (red line) not associated with a prolonged survival benefit. Median OS, calculated as the time point after initiation of the treatment at which 50% of patients are still alive, clearly does not provide any information concerning the minor proportion of patients who occupies the tail of the curves (cure fraction). Therefore, median OS neither differentiates the proportion of patients alive or dead after 50% of patients have died nor reflects the survival time of the patients who are alive after the median OS is reached. In addition, the delayed clinical effect observed with ICI leads to the loss of statistical power if the trial is designed based on conventional proportional hazard model assumption (12). According to the proportional hazard model, HR is equal to 1 in the first part of the curves (early HR) and it becomes unequal to 1 after the separation of the curves (delayed HR). To demonstrate a statistically significant difference in OS, the delta between these two HRs should be high, in fact the HR after the separation of the curves must compensate the lack of separation during the first months of treatment (47) (Figure 1). However, the number of events required to have a large delta value should also increase and the study risks to definitively result as underpowered. In this regard, a recent report by the Institute for Clinical and Economic Review (ICER) highlighted the difficulty in using a proportional hazard model in studies evaluating ICI in advanced NSCLC patients (48). In particular, the ICER analysis stated that the existence of two populations in the immune-oncology arms of the trials, a majority who does not respond to ICI and has a high hazard for survival and a minority with sustained responses and low hazard for progression and mortality, makes difficult the use of proportional hazard models for survival analysis. Notably, survival curve statistic that optimally captures the benefit offered by a particular therapy can differ according to the class of drugs or the clinical context (35). As an example, traditional statistical methods (log rank and Cox model) and survival measures (median OS and HRs) can be usually applied for drugs that start to work early (the OS curve separate since the beginning) and continue to be more active compared to the control arm along the treatment is administered with the assumption that anything affecting the hazard does so by the same ratio at all times (Figure 2A). Median OS but not HR could be used for non-proportional risk models with absence of long-term survivors, as observed in trials evaluating targeted agents (Figure 2B). In fact, the initial large benefit driven by the target agent is entirely capture by the median survival, however with the emerging of resistance this difference disappears and the survival curves cross at a certain time, making the assumption of proportional hazards not applicable in this case. For drugs with delayed benefit, which lead to prolonged survival in a relatively small subset of patients, following a non-proportional risk model (Figure 2C), neither median OS, nor HR are appropriate and alternative statistical methods and survival measures should be reported.

Milestone survival analysis is a cross sectional assessment of OS at a pre-specified and clinically meaningful timepoint, using Kaplan Meier survival probabilities. Milestone...
analysis is usually conducted in a first cohort of randomized patients rather than in the whole population, and it provides long-term survival information in patients with sufficient follow-up, while the entire study continues with OS as primary endpoint. The main characteristic of milestone model is the requirement of a sufficient follow-up, long enough to allow robust estimation of the survival rates. In fact, the milestone analysis should not be conducted until all the patients have met the minimal follow-up time (11). Milestone analysis is based on the assumption deriving from cure rate models that the study population includes a distinct subset of patients who are “cured” and are represented by the tail in the Kaplan Meier curve (49). In addition, milestone outcome better reflects patients hopes and it is much more informative compared to median OS and HR because it answers to patients’ primary interest: the potential rate of cure with a specific treatment (10). The application of milestone in the immune-oncology trials could be useful to avoid wrong interpretations of survival data deriving from early interim analyses. As an example, an interim analysis of the phase III trial comparing first line tremelimumab to chemotherapy in advanced melanoma patients (50) showed no OS benefit, however, an extended follow-up revealed a potential separation of the curves, supporting the use of a milestone model to estimate the true survival benefit (51). Challenges with the milestone survival analysis are represented by the choice of the sample size cohort (in fact milestone does not account for the totality of the OS data), the selection of an optimal threshold for the type I error rate at the time of milestone analysis, the difficulty in maintaining study integrity and blinding prior to the final OS analysis and in assessing the post milestone treatment impact on survival (11). Besides that, the most important concern of the milestone analysis is the identification of a meaningful milestone time point. In fact, the survival analysis could be imprecise if the milestone timepoint is too early (too many events censored and no differences in survival) or too late (the set of patients at risk

Figure 2 Kaplan–Meier survival curves according to the class of drugs. (A) Drug starts to work early and continues to be more active compared to the control arm following the proportional hazard assumption, without long-term survival benefit; (B) treatment starts to work early, however resistance occurs with absence of prolonged survival benefit (non-proportional hazard model); (C) treatment starts to work late, however a long-term survival benefit is observed (non-proportional hazard model).
is small) (10). To avoid the latter condition, a milestone timepoint at which 10% to 20% of the patients in a Kaplan-Meier survival curve remains at risk has been previously proposed (52).

Another novel survival tool currently emerging is the RMST or t-year mean survival time. RMST is a robust statistical procedure able to quantify treatment effect on survival, regardless of the model assumption. Visually, RMST is the area under the survival curve within a specific time window. As milestone analysis, RMST require that the follow-up duration is pre-specified and fixed (53-55). RMST has been recently applied to Checkmate 057 (38). In this study, the median OS in the nivolumab arm (12.2 months) does not adequately capture the long-term survival benefit estimated to be 18% at 3 years (56), similarly the HR of 0.73 cannot be used as a survival measure considering that the 2 survival curves were similar for the first 6 months of treatment and the proportional hazard assumption was not valid. With a follow-up duration of 24 months, the RMST is 13 months for nivolumab versus 11.3 months for docetaxel with a statistically significant difference of 1.7 months (95% CI, 0.4–3.1; P=0.01) in favor of nivolumab. Graphically this difference is represented by the area between the two Kaplan-Meier curves. In Checkmate 057, RMST of 13 months for the nivolumab arm means that NSCLC patients receiving nivolumab and followed for 24 months would survive for an average of 13 months (13).

The weighted log rank test is an additional statistical tool that can be used for non-proportional survival models with delayed clinical benefit and long-term survival. Basically, a weighted log rank test avoids loss of statistical power because it reduces the statistical weight of the early time period, during which survival curves might be similar (12,57). Weighted log ranks have been proposed as novel statistical methods for studies with ICI. However, the a priori definition of the weights is usually difficult because the point at which the survival curve diverge cannot be easily predicted at the start of the trial. Finally, besides weighted log rank, the Weibull distribution represents another parametric survival model, which could provide additional useful information in evaluating ICI effect on survival. In the Weibull model, the survival time depends on the shape parameter of survival curves and hazard is not a constant but a function of time (14,58). The Weibull model can fit well to the immune-oncology clinical trials because it takes into account the different shapes of survival curves and their variation during time. The Weibull model also allows the inclusion of covariates of survival times, useful to describe long tailed distributions.

Safety

Conventional safety analysis using the 3+3 dose escalation design and the first two cycles as a dose limiting toxicity (DLT) assessment period (59) might not accurately describe the safety profile of ICI. Remarkably, grade 3–4 immune related adverse events (irAE) with PD-1/PD-L1 inhibitors are infrequent, for example in pretreated NSCLC they range from 7% to 20% (25,37,38,42), and toxicities are observed usually late (from months to years) (60). Therefore, the phase II dose is often determined on the base of pharmacokinetic (PK)/pharmacodynamic profile or of the maximal administered dose rather than on the maximal tolerated dose (MTD). Future studies should introduce longer DLT periods (≥6 weeks) before escalate to higher doses in order to adequately select the phase II appropriate dose.

Regarding anti CTLA-4 agents, irAE can rapidly become life threatening (61), for this reason, in some protocols it is recommend to hold the anti-CTLA-4 until recovering from grade 2 toxicities. Although these irAE decrease drug exposure and definitively influence treatment dose intensity, they are not formally classified as DLT because they are not grade 3–4 toxicities. Finally, late irAE presenting several months after the last dose of ICI have been described (62), highlighting the importance to incorporate longer follow-up periods (up to 1 year) in clinical trials evaluating ICI in order to capture late post discontinuation toxicities and to characterize their effect on subsequent anticancer therapies. For these reasons, innovative phase I trial designs should better define the timing and the best way to assess DLT and MTD for ICI.

PD-L1 as predictive biomarker in NSCLC patients: challenges and methodological issues

A challenging unmet need for clinical trials evaluating ICI is the absence of reliable biomarkers of response to immunotherapeutic agents, able to identify before the treatment initiation which patients are more likely to experience clinical benefit. Although several biomarkers are currently being tested in different disease settings (63), most of the available data from clinical trials with PD-1/ PD-L1 inhibitors evaluated the predictive role of PD-L1 expression. Interestingly, across different tumor types, a relevant number of patients with PD-L1 positivity (40–50%) does not achieve objective response to anti PD-1/PD-L1...
therapies, in addition 15% of patients negative for PD-L1 expression, experience objective responses (64). However, a significant correlation between PD-L1 expression and ORR to anti PD-1/PD-L1 agents was reported by a sensitivity analysis of trials investigating PD-1/PD-L1 inhibitors in different cancer types (65) and by a meta-analysis in NSCLC patients (66).

In particular, in advanced lung cancer, the predictive role of PD-L1 did not clearly emerged from clinical trials evaluating nivolumab in pretreated NSCLC patients (37,38,46,67), although a post hoc analysis from Checkmate 057 suggested an improved benefit for patients with PD-L1 positive tumors at the threshold levels of 1%, 5% and 10% (38). Similarly, PD-L1 expression was not predictive of nivolumab benefit in first line setting (39,68). On the contrary, the phase II (KEYNOTE-010) (25) and III (KEYNOTE-024) (36) development of pembrolizumab in NSCLC was restricted to PD-L1 positive patients (threshold 1% for KEYNOTE-010 and 50% for KEYNOTE-024) due to the higher ORR (45% for PD-L1 ≥50%, 16.5% for PD-L1 in the range of 1–49% and 10.7% for PD-L1 <1%) observed in the expansion cohort of a phase I trial (KEYNOTE-001) (69). In the case of the anti-PD-L1 antibody atezolizumab, PD-L1 expression was evaluated on both tumor cells (TC) and immune cells (IC) (70). Although the survival benefit with atezolizumab appeared to correlate with PD-L1 expression on TC and IC in pretreated NSCLC patients (POPLAR trial) (41), this finding was not confirmed in the phase III OAK trial which showed a significant OS benefit in favor of atezolizumab compared to docetaxel regardless of PD-L1 expression (42). In first line setting, the development of atezolizumab followed a different strategy and two phase II trials, BIRCH (71) and FIR (72) tested atezolizumab only in PD-L1 positive (TC and IC) patients. Finally, early trials evaluating durvalumab and avelumab showed a higher response rate in patients with PD-L1 expression on TC ≥25% (ATLANTIC trial) (73) and ≥1% (Javelin solid tumor trial) (74), respectively.

A comprehensive characterization of the putative predictive role of PD-L1 or other biomarkers, such as tumor mutational burden (75,76), IFN-γ mRNA expression (77), tumor infiltrating lymphocytes (78,79), serum circulating factors (80), gut microbiota (81) across several tumor types (63) and in NSCLC patients (82), has been recently reported elsewhere and it goes far beyond the aim of this review. Therefore, we will provide an insight on some methodological and practical issues regarding PD-L1 assessment in advanced NSCLC such as concordance between PD-L1 immunohistochemistry (IHC) assays, variability of PD-L1 assessment on TC and IC and the impact of spatial and temporal tumor heterogeneity on PD-L1 expression. Considering that each PD-1/PD-L1 inhibitor has its own PD-L1 diagnostic test and that different level of PD-L1 expression have been evaluated for correlation with clinical outcome in trials in NSCLC patients, several harmonization studies have recently tried to reduce the high variability of the assays. The Blueprint PD-L1 IHC Assay Comparison Project reported high concordance for PD-L1 level detection between 28-8 (IHC test for nivolumab), SP263 (IHC test for nivolumab, pembrolizumab, durvalumab) and 22C3 (IHC test for pembrolizumab) assays, whereas lower PD-L1 expression was detected by SP142, the companion diagnostic test for atezolizumab (83). These findings suggest a potential risk of false negative results when the antibody SP142 is used to detect PD-L1 on tumor samples. On the bases of Blueprint study, PD-L1 expression was re-evaluated with the 22C3 IHC assay in 400 tumors from OAK trial. Surprisingly, atezolizumab was superior to docetaxel in all subgroups, including tumors with less than 1% PD-L1 expression on TC (84). However, these data are still a matter of debate because PD-L1 <1% was found in 55% of the tumors, whereas the expected rate of negative tumors is around 30%, making it likely that the PD-L1 <1% population included false-negative tumors.

The low performance of SP142 in detecting PD-L1 on TC was additionally confirmed in other 3 harmonization trials (85-87), and the high degree of concordance between 28-8, SP263 and 22C3 were consistent across several different studies (85-90). However, conflicting results emerged from recent analyses showing a higher expression of PD-L1 with SP142 test (91) and a lower expression with 22C3 (92) and 28-8 (91) antibodies.

Differently from PD-L1 expression on TC, PD-L1 IHC on IC is characterized by a greater variability and low interobserver concordance. These discordant results might be due to the co-existence of both cytoplasmic and membranous PD-L1 staining in IC (93) and to the lack of pre-specified criteria for assessment of PD-L1 staining on IC. Besides the variability among assays and between PD-L1 assessment on TC or IC, another critical issue is represented by the spatial and temporal heterogeneity of PD-L1 expression (94). In this regard, KEYNOTE-010 comparing pembrolizumab to docetaxel in PD-L1 ≥1% pretreated NSCLC patients (25), showed that the prevalence of PD-L1 levels ≥50% was similar in archival or rebiopsy
samples (~40–45%) and clinical outcomes in patients with PD-L1 ≥50% did not differ between archival and new samples (95). A superimposable result was reported in an exploratory analysis of the ATLANTIC trial (73), a phase II study of durvalumab in pretreated NSCLC patients, which showed high concordance between fresh biopsies acquired 3 months before treatment compared to older tumor samples (96). Similarly, spatial heterogeneity does not seem to strongly influence PD-L1 expression. In fact, a retrospective study (97) and an exploratory analysis of the ATLANTIC trial (96) showed good concordance for PD-L1 expression between primary tumor and metastatic samples. Finally, regarding PD-L1 intratumor heterogeneity, data are conflicting with some studies showing high concordance of PD-L1 staining between different samples of the same tumor site, tested with the same PD-L1 IHC assay (98), and other studies reporting discordant PD-L1 staining from matched specimens (99).

Characterizing inter and intratumor heterogeneity of PD-L1 expression and overcoming the hurdles of inter-assays variability and of discordant TC-IC stains represent important issues that need to be addressed in future clinical trials with ICI in cancer patients.

**Clinical trials design for ICI**

The traditional clinical trials designs have been widely reshaped by the advent of ICI, with changes concerning all the different phases of drug development.

Regarding phase I trials, considering the low rate of grade 3–4 toxicities and the relatively absence of DLT for PD-1/PD-L1 inhibitors, alternative designs, such as modified toxicity probability interval design, have been recently developed (100,101). According to this model the proportion of targeted DLT can be less than 17% (in classical 3+3 dose escalation phase I trials the targeted proportion of DLT is 17–33% of patients). PK and pharmacodynamic properties of ICI have been explored only in a limited number of phase I trials (102,103). For example, regarding nivolumab, doses from 0.1 to 10 mg/kg demonstrated 64–70% PD-1 receptor occupancy on CD3+ T cells (103) and the initial FDA approved dose of 3 mg/Kg every 2 (q2) weeks was subsequently changed to a flat dosing of 240 mg q2 weeks based on population PK demonstrating comparability of safety and efficacy for most disease indications (104). A model based PK analysis in different cancer types reported that an alternative flat dosing of 480 mg q4 weeks resulted in similar exposure, efficacy and safety as the 3 mg/kg q2 weeks (105). A better knowledge and interpretation of PK data of phase I trials are of paramount importance considering that advanced melanoma (106), RCC (107) and NSCLC (22) patients may experience prolonged responses after treatment discontinuation and that responses may happen also after rechallenge with the same drug (108). In this regard, a phase III/IV trial (Checkmate 153) comparing continuous nivolumab to observation after 1 year of nivolumab in advanced NSCLC patients recently showed an improvement in PFS (not reached versus 10.3 months, HR =0.42; 95% CI, 0.25–0.71) in patients receiving continuous treatment (109). Despite these hypothesis-generating results, additional data and innovative phase I trials with a deeper insight in the pharmacological properties of ICI are urgently needed.

As previously reported for some targeted agents (110), phase I trial testing ICI had to face the issue of answering to multiple clinical questions in a shorter timeframe, with the final aim of reducing the development time from phase I to registration by regulatory authorities. Remarkably, adaptive and basket designs with biomarker enrichment strategies have led to approval of several ICI, revolutionizing the traditional drug development paradigm based on 3 or more steps (phase I, phase II and phase III). In adaptive design, modifications of the trial are prospectively planned, so that changes may take place while the study is ongoing. The main goal of adaptive design trials is to learn and address several hypotheses at one time in order to speed up the development of the compound (111). One example is KEYNOTE-001, a phase I trial which led to FDA approval of pembrolizumab both in advanced melanoma and NSCLC, in a timeframe <4 years (112). With 1,245 patients enrolled, KEYNOTE-001 is the largest phase I trial to date. Its adaptive design (at least 8 protocol amendments including, among others, modification of the primary endpoint from irRC to RECIST 1.1, addition or abandoning of specific cohorts, increasing sample size for certain cohorts), allowed the trial to simultaneously generate several efficacy data rather than starting different studies for each clinical question (112). Basket designs have been successfully adopted in clinical trial evaluating ICI. Basket trials test the effect of a drug on a single target in a variety of cancer types. In basket studies, the investigators can separately analyze the responses of patients by tumor types, and choose to expand or close patient cohorts according to the benefit of the experimental treatment (111). KEYNOTE-059 is an example of a successful and innovative phase II basket
trial, testing pembrolizumab in patients with high level or microsatellite instability (MSI-H) or deficiency in mismatch DNA repair (dMMR). In ~150 patients with MSI-H or dMMR, pembrolizumab showed an ORR of ~40% and responses were observed regardless of tumor histology, leading to the first histology-agnostic FDA approval of a cancer treatment in USA (1,113). In this regard, the histology-independent benefit of ICI clearly differentiates them from targeted agents. In fact, basket trials are not always reliable in oncogene addicted disease, considering that the simple molecular abnormality (i.e., BRAF mutation) does not imply the efficacy of specific inhibitors (i.e., vemurafenib) and that tumor response strongly depend on the disease context (114).

In a recently published guidance, FDA highlights the key role of the enrichment trial design to identify specific subgroups of patients who would benefit from experimental treatments, encouraging physicians to widely adopt this strategy in clinical trials (115). Both KEYNOTE-001 and KEYNOTE-059 trials imply a biomarker based enrichment design. In the expansion cohort of KEYNOTE-001 in advanced NSCLC (69), the cut off selection for PD-L1 positivity and its validation provided the bases for phase II (25) and III (36) trials testing pembrolizumab in PD-L1 positive NSCLC patients. On these premises, the phase III trial comparing first line pembrolizumab to platinum based chemotherapy was specifically designed in EGFR/ALK wild type NSCLC patients with PD-L1 expression ≥50%, and the most updated results showed significant improvements in the ORR (45.5% vs. 29.8%), median PFS (10.3 vs. 6.0 months, HR =0.50; P=0.001) and median OS (30 vs. 14.2 months; HR =0.63; P=0.002) in favor of pembrolizumab (40). However, the reliability of biomarker enrichment strategies for ICI is still a matter of debate, considering that PD-L1 is neither a totally specific, nor a sensitive predictive biomarker, and that several others (such as tumor mutational burden or TIL) are currently being validated in clinical trials. For example, in the phase III OAK study, atezolizumab significantly improved OS compared to docetaxel (13.8 vs. 9.6 months; HR =0.73, P=0.0003) in pretreated NSCLC patients, regardless of PD-L1 expression on TC or IC (42), also when PD-L1 expression was evaluated with the 22C3 more sensitive diagnostic assay (84). Furthermore, the recently published Checkmate 026 failed to show a significant improvement in PFS (HR =1.15; 95% CI, 0.91–1.45; P=0.25) in advanced NSCLC patients with PD-L1 expression ≥5% (39). Of note, in patients with PD-L1 expression ≥50%, the lack of benefit for nivolumab persisted with an HR for progression or death of 1.07 (95% CI, 0.77–1.49). Overall, results from Checkmate 026 both in the whole population and for those tumors with strongly positive PD-L1 expression are inconsistent with first line nivolumab performance in a phase I trial (68). Besides Checkmate 026, another example of unsuccessful biomarkers enrichment strategy design is represented by MYSTIC trial comparing durvalumab vs. durvalumab + tremelimumab vs. platinum based chemotherapy in 1,092 treatment-naïve-EGFR/ALK wild-type NSCLC patients with PD-L1 expression ≥25%. The co-primary endpoints were OS and PFS. Results are not yet published, however the trial failed to show superiority in PFS of the combination durvalumab plus tremelimumab compared to platinum based chemotherapy in this PD-L1 enriched population (116). Although it is not possible to perform cross trial comparisons, the conflicting results between KEYNOTE-024 and Checkmate 026 or MYSTIC trials are difficult to attribute to differences in the pharmacologic and biologic properties among ICI, while discrepancies in patients selection, biomarker tests, and PD-L1 expression cut points could have contributed to these discordant findings (117). Results from a confirmatory phase III study of first line pembrolizumab vs. platinum based chemotherapy (KEYNOTE-042) in advanced NSCLC patients with PD-L1 expression ≥1% (118), will probably shed more light on the utility of enrichment design and on the performance of different biomarker thresholds for PD-L1 positivity.

Finally, due that multiple studies evaluating ICI have a low enrollment target (76 patients per trial on average for investigator initiated studies) (1), it is unrealistic that small single center trials will recruit enough patients to produce high quality results. Furthermore, the main pitfall accompanying the entering in the clinic of many different ICI will probably be the absence of direct comparisons between different compounds, tested in different clinical settings and in distinct patients’ populations.

Recently, FDA summarized examples of collaborative and novel trial designs that could allow more questions to be efficiently addressed in a single multicenter trial (119). A promising example is the LUNG MAP program using a common biomarker screening platform to classify molecular subgroups of patients and assign them to specific matched targeted therapies (120). Similarly for ICI, collaborative platforms, coordinated both by pharma companies and non-profit organizations, including studies with multiple arms and hundreds of active sites, will help to avoid excessive
data fragmentation and duplication and will provide the background for the development of high-quality designed clinical trials, where sharing of findings and resources can ultimately lead to accelerated scientific innovation.

**Conclusions**

Besides the paradigm shift in cancer treatment, the advent of ICI has also raised several questions regarding the most appropriate endpoints, statistical models, biomarker assessment methodologies and clinical trial designs. Specifically, a more extensive use of iRECIST to assess antitumor responses and the replacement of traditional statistical methods (log rank and Cox proportional hazard model) and survival measures (median OS and HR) with new models (such as milestone analysis or RMST) able to capture delayed survival benefit and long-term tails are key issues to address in the next wave of trials with immunotherapies. Regarding predictive biomarkers, in particular PD-L1 expression, the integration and harmonization of the existing assays are critical to reduce variability and provide a reliable test to identify responders or patients who should be early switched to different treatments. Furthermore, a single biomarker may not mirror the real systemic immunological landscape of the patient. Therefore, translating to ICI the idea of targeted therapies, for which one biomarker is usually enough to predictive the drug benefit, appears an unrealistic objective. Moreover, innovative study designs such as adaptive or basket and biomarker enriched clinical trials, which may address different hypotheses at one time, potentially identifying molecular subgroups of patients with increased benefit from ICI, represent a promising strategy to pursue. In conclusion, building large collaborative platforms of clinical trials and selecting the appropriate bars to assess the clear health benefit and value of ICI represent the major challenges for the future research in the immune-oncology field.

**Acknowledgements**

**Funding:** Dr. Roberto Ferrara was the recipient of the grant DUERTECC/EURONCO (Diplôme Universitaire Européen de Recherche Translationnelle Et Clinique en Cancérologie) for 2016–2017, and of the IASCL (International Association for the Study of Lung Cancer) Young Investigator Award for 2017–2018. Dr. Sara Pilotto was supported by a Fellowship Award of the International Association for the Study of Lung Cancer. Prof. Emilio Bria received honoraria or speakers’ fee from MSD, AstraZeneca, Celgene, Pfizer, Helsinn, Eli-Lilly, BMS, Novartis, Roche, and he received research support from A.I.R.C. (Associazione Italiana Ricerca sul Cancro, grants no. 14282 and no. 20583), I.A.S.L.C. (International Association for the Study of Lung Cancer), L.I.L.T. (Legaitaliana perla Lotta contro i Tumori), Fondazione Cariverona, Astra-Zeneca, Roche and Open Innovation.

**Footnote**

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

**References**

Ferrara et al. Methodological issues in immunotherapy trials


76. Gadgeel S, Kowanetz M, Zou V, et al. Clinical efficacy of Atezolizumab in PD-L1 subgroups defined by SP142 and 22C3 IHC assay in 2L+ NSCLC: results from the
randomized OAK study. Ann Oncol 2017;28:mdx380.001
104. FDA Opdivo prescribing information 2016. Available online: https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/125554s017s018lbl.pdf


