DOI: 10.1093/jnci/djq310
Advance Access publication on September 8, 2010.

REVIEW

Improved Endpoints for Cancer Immunotherapy Trials

Axel Hoos, Alexander M. M. Eggermont, Sylvia Janetzki, F. Stephen Hodi, Ramy Ibrahim, Aparna Anderson, Rachel Humphrey, Brent Blumenstein, Lloyd Old, Jedd Wolchok

Manuscript received December 9, 2009; revised July 21, 2010; accepted July 21, 2010.

Correspondence to: Axel Hoos, MD, PhD, Bristol-Myers Squibb, Global Clinical Research, Oncology, Wallingford, CT 06492 (e-mail: axel.hoos@bms.com).

Unlike chemotherapy, which acts directly on the tumor, cancer immunotherapies exert their effects on the immune system and demonstrate new kinetics that involve building a cellular immune response, followed by changes in tumor burden or patient survival. Thus, adequate design and evaluation of some immunotherapy clinical trials require a new development paradigm that includes reconsideration of established endpoints. Between 2004 and 2009, several initiatives facilitated by the Cancer Immunotherapy Consortium of the Cancer Research Institute and partner organizations systematically evaluated an immunotherapy-focused clinical development paradigm and created the principles for redefining trial endpoints. On this basis, a body of clinical and laboratory data was generated that supports three novel endpoint recommendations. First, cellular immune response assays generate highly variable results. Assay harmonization in multicenter trials may minimize variability and help to establish cellular immune response as a reproducible biomarker, thus allowing investigation of its relationship with clinical outcomes. Second, immunotherapy may induce novel patterns of antitumor response not captured by Response Evaluation Criteria in Solid Tumors or World Health Organization criteria. New immune-related response criteria were defined to more comprehensively capture all response patterns. Third, delayed separation of Kaplan–Meier curves in randomized immunotherapy trials can affect results. Altered statistical models describing hazard ratios as a function of time and recognizing differences before and after separation of curves may allow improved planning of phase III trials. These recommendations may improve our tools for cancer immunotherapy trials and may offer a more realistic and useful model for clinical investigation.

J Natl Cancer Inst 2010;102:1388–1397

Despite important advances in the understanding of the immune response in cancer, clinical investigations of cancer immunotherapies have had marginal success. Most immunotherapeutic agents did not show sufficient activity in early trials, and whereas some were advanced to phase III investigation, most failed in randomized comparisons (1). Likely contributing factors include ineffective or marginally effective agents, an incomplete understanding of human tumor immunology, and the absence of adequate tools for development, including criteria for refined trial endpoints (1–3).

In recent years, immunologic science has evolved, and new mechanisms for targeted immunotherapies have been discovered (eg, immune checkpoint modulation such as anti-cytotoxic T lymphocyte–associated protein 4 [CTLA-4] or anti-programmed cell death-1 antibodies) (4–7). Adjusting the clinical development paradigm from chemotherapy to immunotherapy requires addressing the unique characteristics of immunotherapeutic agents in clinical trials to provide more adequate tools for evaluation (8), including the adjustment of clinical trial endpoints (9).

The continuum of biological events following administration of immunotherapy to a cancer patient can be divided into three steps: 1) immune activation and T-cell proliferation starting early after first administration, 2) clinically measurable antitumor effects mediated by activated immune cells over weeks to months, and 3) potential delayed effect on patient survival several months after first administration compared with agents not requiring immune activation (Table 1). Each biological event can be measured as a clinical trial endpoint and encompasses specific challenges.

The need to rethink clinical development of novel anticancer agents was recognized for solid tumors and hematologic malignancies (8–11), in particular for prostate cancer (12–14), melanoma (15–17), or lymphoma (18,19). From 2004 to 2009, several initiatives systematically evaluated an immunotherapy-focused clinical development paradigm, which supported the redefinition of biological outcome measures and clinical endpoints (2,8,9,20–22). These initiatives were planned and facilitated by the Cancer Immunotherapy Consortium of the Cancer Research Institute (CIC-CRI; formerly Cancer Vaccine Consortium) and conducted in collaboration with the Association for Cancer Immunotherapy (C-IMT) in Europe or the International Society for Biological Therapy of Cancer in the United States. Workshops or expert panels summarized community knowledge and defined challenges as a platform for improvements in trial endpoints. This knowledge was then used to systematically generate large datasets or analyze existing datasets with the aim of improving conventional endpoints. Several CIC-CRI and C-IMT proficiency panels identified sources of variability of T cell–response assays and defined harmonization criteria to control variability without limiting individual laboratory protocols. A large clinical trial program (n = 487) conducted by Bristol-Myers Squibb and Medarex investigated antitumor response and survival outcomes to propose new response criteria, as well

Table 1. Challenges and recommendations for assessment of cancer immunotherapy*

Immunotherapy start	Immune cell activation and proliferation	Effect on tumor	Effect on survival
Day 1	Days to weeks	Weeks to months	Several months
Endpoint	Cellular immune response	Antitumor response	Survival
Challenges	Complex assays exist Results are highly variable and not reproducible across trials Assay procedures are not harmonized	Conventional and novel response patterns are observed The translation of the immune response into an antitumor response takes time No systematic criteria to capture new response patterns exist	Translation of immune and antitumor response into a survival effect takes time Proportional hazards assumptions are not applicable Conventional statistical models do not account for nonproportional hazards and delayed separation of curves
Recommendations	Harmonized assay use through SOPs that accompany individual assay protocols	Identify relevant response patterns Use systematic criteria (irRC) to reproducibly capture new patterns	Employ statistical models that account for the delayed effect Carefully consider use of early interim and futility analyses

* irRC = immune-related response criteria; SOP = standard operating procedure.

as statistical methods more suitable for the analysis of survival kinetics.

Three novel endpoint considerations for immunotherapy investigation, which need to undergo prospective evaluation in clinical studies (9), were formulated as follows. First, minimize variability of T-cell assay results through assay harmonization to establish cellular immune response as a reproducible biomarker and subsequently investigate its relationship with clinical outcomes. Second, capture the spectrum of clinical patterns of antitumor response for immunotherapeutic agents through novel immune-related response criteria (irRC), which are adapted from Response Evaluation Criteria in Solid Tumors (RECIST) and World Health Organization (WHO) criteria. Third, use different statistical methods for trial design and analysis of survival outcomes in the presence of delayed separation of Kaplan–Meier survival curves.

The core aspects of these recommendations were discussed at a Workshop of the United States Food and Drug Administration in 2007 and were included in a recently issued guidance document "Clinical Considerations for Therapeutic Cancer Vaccines," thus providing developmental guidance to developers of immunotherapies (23). Taken together, these recommendations represent an expansion of methodology for cancer immunotherapy trials and may contribute to improved clinical investigations.

Measuring Immune Response: Reduction of Variability Through Assay Harmonization

Activation of the cellular arm of the immune system is seen as the first biological event after administration of immunotherapy, and consequently, measurement of such a response (mostly T-cell response) is of great interest. A variety of bioassays exist for immune monitoring, including the enzyme-linked immunosorbent spot (ELISPOT) assay and cytometry-based tests such as intracellular cytokine staining, HLA-peptide multimer staining, and the carboxyfluorescein succinimidyl ester assay (24–27). Even though the fundamentals of these assays have been well established, a plethora of different laboratory protocols is used (24,27,28).

Substantial variability in results among laboratories exists in multicenter trials, which hampers data reproducibility and prevents meaningful comparisons among studies (28). For example, in the first CIC-CRI ELISPOT proficiency panel, 36 laboratories used the ELISPOT assay to analyze the same donor peripheral blood mononuclear cell sample for a T-cell immune response, resulting in spot counts ranging from nondetectable to strongly positive (Figure 1). The challenge is the absence of a quality control measure for T cell–based assays that can be used as a gold standard (29), which has prevented the field from establishing T-cell response as a biomarker for immunotherapy trials and conducting reliable investigations of its relationship with clinical outcomes (30).

Under the auspices of CIC-CRI and C-IMT, two large international immune monitoring proficiency panel programs were initiated in 2005 (20–22). Their goals were to provide an external quality assurance process for laboratories conducting immune monitoring in clinical trials and to harmonize assay performance. Harmonization of clinical trial conduct and data collection as established through the International Conference on Harmonization for Good Clinical Practice had previously proven to be a highly successful model for improving clinical development procedures (31).

Proficiency panels are quality control experiments in which the same patient samples are tested across multiple laboratories, using their respective protocols, and the results are centrally evaluated (32). The CIC-CRI proficiency panel program is led by a scientific steering committee and involves a central laboratory for accrual of peripheral blood mononuclear cells, shipping logistics, and independent central data analysis. Peripheral blood mononuclear cell samples and antigens, pretested and defined for their response status, are sent to all participating laboratories, where they are tested for response under the local conditions. Test results and

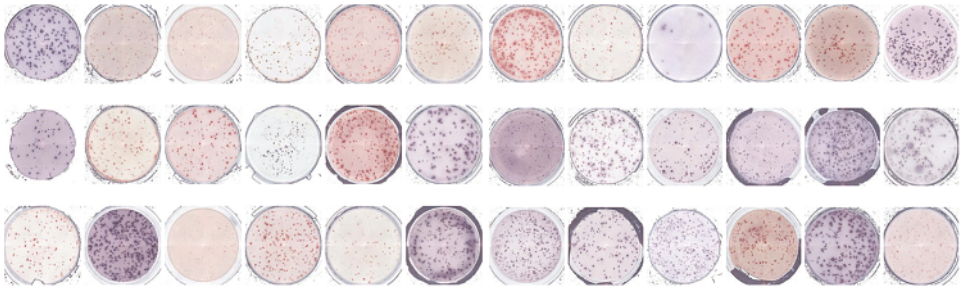

Figure 1. High variability of results for the enzyme-linked immunosorbent spot (ELISPOT) immune response assay. Identical peripheral blood mononuclear cell samples from the same patient were sent to 36 different laboratories experienced with ELISPOT methodology. The image shows the spot count results in microtiter plates in which each well represents the result of one laboratory. Some wells show high numbers of spots, whereas others are low or negative. Each spot in this assay represents a single T-cell capable of reacting against a defined target antigen. These results reflect the outcome of the first ELISPOT proficiency panel, which identified sources of variability among laboratories.

protocol details are reported back for central analysis. Since 2005, more than 80 laboratories from 14 countries have participated, encompassing the academic, nonprofit, biotech, and pharmaceutical sectors, and the United States Department of Defense (20,22).

The ELISPOT panel is the longest running program and provides the most mature results. The objective of the first ELISPOT panel was to identify sources of variability among assay procedures. The second panel adjusted these sources of variability while keeping the respective laboratory protocols intact. This adjustment led to a substantial reduction in variability: The percentage of participants unable to detect all responders (six responders among eight samples) was reduced from 47% to 14%, and the percentage of participants unable to detect 50% or more of all responders ("outliers") dropped from 11% to 0% (Figure 2). The combined panel results led to initial ELISPOT harmonization guidelines (Table 2) (20), which synchronize key variables across laboratories and substantially influence assay outcome but do not impose standardization of assays on individual laboratories. The introduction of these guidelines is central to this harmonization to provide general assistance for the conduct of the individual assay protocol in the context of standard operating procedures (eg, exclusion of apoptotic cells, use of pretested serum for background reduction, and quality control during the computerized spot evaluation) (Table 2) (20). Confirmatory findings from a parallel experiment were published by the C-IMT proficiency panels (21). As the use of serum is a crucial variable for ELISPOT assays, a separate ELISPOT proficiency panel focused on serum use and showed that serum-free medium for incubation of cells can be as effective as qualified serum-supplemented medium, thus addressing this crucial variable for assay protocols as part of the harmonization guidelines (33).

The first round of the HLA-peptide multimer-staining panel of CIC-CRI allowed the formulation of initial harmonization guidelines (22), which will likely reduce assay variability. These recommendations are 1) the use of more than two colors for staining, 2)

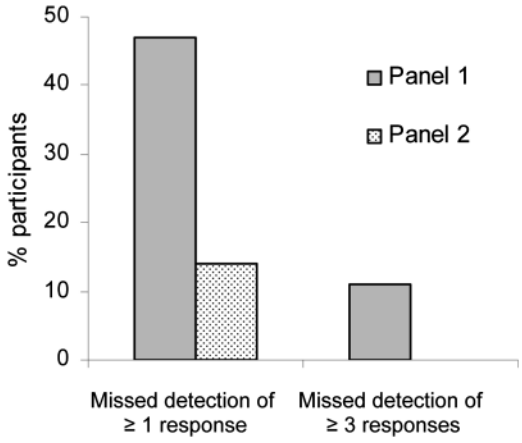

Figure 2. Effect of assay harmonization on data variability of the enzyme-linked immunosorbent spot (ELISPOT) assay. In the Cancer Immunotherapy Consortium of the Cancer Research Institute ELISPOT proficiency panel, participating laboratories reported the response status from eight different donor–antigen combinations. Grey bars represent the first panel round and stippled bars the second panel round. In the first panel round, 47% of panelists missed detection of at least one response correctly, and 11% of panelists failed to detect at least three responses correctly (characterized as an "outlier" because of high variability). Based on the first panel results, harmonization criteria were given to panelists, and the testing was repeated in the second panel (stippled bars). ELISPOT performance improved, with only 14% of panelists missing at least one responder and zero outliers.

Table 2. Initial harmonization guidelines for the enzyme-linked immunosorbent spot (ELISPOT) immune response assay*

A	Use only pretested and optimized serum allowing for low background : high signal ratio
B	Establish laboratory SOP for ELISPOT testing procedures, including:
B1	Counting method for apoptotic cells for determining adequate cell dilution for plating
B2	Overnight rest of cells before plating and incubation
C	Establish SOP for plate reading, including:
C1	Human auditing during reading process
C2	Adequate adjustments for technical artifacts
D	Only allow trained personnel, certified per laboratory SOP, to conduct assays

* Harmonization guidelines can be used by each individual laboratory performing an immune response assay in the context of Standard Operating Procedures (SOPs) and without adopting a standard assay protocol. Through general steps such as use of pretested serum [or serum-free media (33)], exclusion of apoptotic cells from the analysis, human auditing of the computerized assay read out procedure, and training of operators on the laboratory SOPs, quality of assays can be substantially improved. Courtesy of Janetzki et al. (20).

the collection of more than 100 000 CD8 T cells, and 3) the use of a background control sample to set appropriate analytical gates (data points of interest) (22). The second CIC-CRI panel round focused on confirmation of these guidelines, as well as assessment of the influence of specific protocol steps on final results.

The first intracellular cytokine staining panel was conducted under highly standardized conditions. However, the results demonstrated that the use of standardized reagents and protocols alone does not lead to the desired performance among panelists. The ongoing second panel allows free choice of reagents and protocols but requires compliance with evolving minimum guidelines based on the results of the first panel. Results are expected later in 2010.

The clinical trial application of immune assays was addressed by an international working group, the Cancer Vaccine Clinical Trial Working Group (8), which proposed that immunoassays in clinical trials should be performed at least at three different time points throughout any study—one baseline and two follow-up time points. Assays should be established, reproducible, and technically (not clinically) validated in the respective laboratories (according to the harmonization criteria for respective assays). At least two assays should be used in parallel to demonstrate the same findings (eg, ELISPOT and HLA-peptide multimer staining), which is particularly important if developmental decisions (eg, moving a new agent into more advanced trials) would be based on immune response data from such assays. Moreover, the cutoff values for an immune response should be identified prospectively to define the change necessary for a response and to define the proportion of study patients needed for a positive study outcome.

In addition, a separate initiative is under way to provide a publication framework for the results from immune monitoring trials. This project, named MIATA (minimal information about T-cell assays), is currently undergoing a public consultation process to obtain community feedback (28). Overall, only when the issue of current assay variability is adequately addressed can cellular immune response become a reliable parameter in clinical trials and may be more reliably studied for its relationship with clinical outcomes.

Measurement of Antitumor Response: irRC

For decades, investigators have relied on modified WHO criteria (34) or, more recently, RECIST (35,36) to assess clinical activity of anticancer agents. These standard criteria were designed to capture effects of cytotoxic agents and depend on tumor shrinkage to demonstrate activity. However, the response patterns seen with immunotherapeutic agents extend beyond those of cytotoxic agents and can manifest, for example, after a period of "stable disease" in which there is no tumor shrinkage or after initial tumor burden increase or appearance of new lesions (eg, tumor infiltrating lymphocytes) (15,37–40).

This potential delayed detection of clinical activity on radiographic assessment may reflect the dynamics of the immune system—the time required for T-cell expansion followed by infiltration of the tumor—and a subsequent measurable antitumor effect. For example, some reported clinical experiences with cancer vaccines (37–39) demonstrated that patients with stable disease or progressive disease may have subsequent tumor regression, whereas others may show initial mixed responses, with regression in some lesions while other lesions remain stable, progress, or first appear. In these studies, patients with measurable tumor burden decrease had more responses that did not fit currently accepted response criteria than conventional responses (eg, 5/6 and 4/5 unconventional out of all recorded responses) (37–39). Such patterns have been noted by many investigators; however, they were inconsistently included in publications or were not systematically captured because of the absence of suitable response criteria (8), which, in turn, did not allow for their clinical significance to be adequately studied (8). It has become evident that RECIST or WHO criteria may not offer a complete description of the response to immunotherapeutic agents, and either adjusted or new criteria are needed (8).

Similar to the immune response assays discussed above, the Cancer Vaccine Clinical Trial Working Group addressed the topic of clinical endpoints and potential delayed detection of clinical activity in a series of workshops. The main conclusions were that the appearance of measurable clinical activity at the tumor site may take longer for immunotherapies than for cytotoxic therapies; responses to immunotherapies may occur after conventional progressive disease (tumor burden increase); discontinuation of immune therapy may not be appropriate in some cases, unless progressive disease is confirmed (as is usually done for RECIST-based response); there should be allowance for "clinically insignificant" progressive disease (eg, small new lesions in the presence of other responsive lesions); and durable stable disease may represent clinical benefit. Such elements might be included in new antitumor response criteria that adapt standard criteria to the unique characteristics of immunotherapy (8).

Building on these recommendations, a novel set of response criteria based on WHO and RECIST were evaluated in a series of large multinational studies (17,41,42), including 487 patients with advanced melanoma participating in the Bristol-Myers Squibb and Medarex clinical trial program for ipilimumab, a fully human monoclonal antibody that blocks CTLA-4. In these studies, four distinct response patterns were detected: immediate response, durable stable disease, response after tumor burden increase, and response in the presence of new lesions (Figure 3). The first two patterns are conventional (Figure 3, A and B), whereas the latter two (Figure 3, C and E) are novel and specifically recognized with immunotherapeutic agents. Photographs of a case study of the first novel tumor response pattern (Figure 3, D) show that tumor burden initially increases (day 84) and then decreases (day 112) to a complete response (day 503). Importantly, all patterns (conventional and new) seem to be associated with favorable survival compared with patients with progressive disease by WHO criteria (43,44). To create a process, which systematically captures all observed response patterns, irRC were proposed (43–45).

The irRC enhance characterization of immunotherapy response patterns through new features. In particular, they allow for the assessment of tumor burden as a continuous variable, which considers index lesions identified at baseline together with new lesions as they may occur after treatment start. Only measurable lesions are taken into consideration. Measures are taken bidimensionally

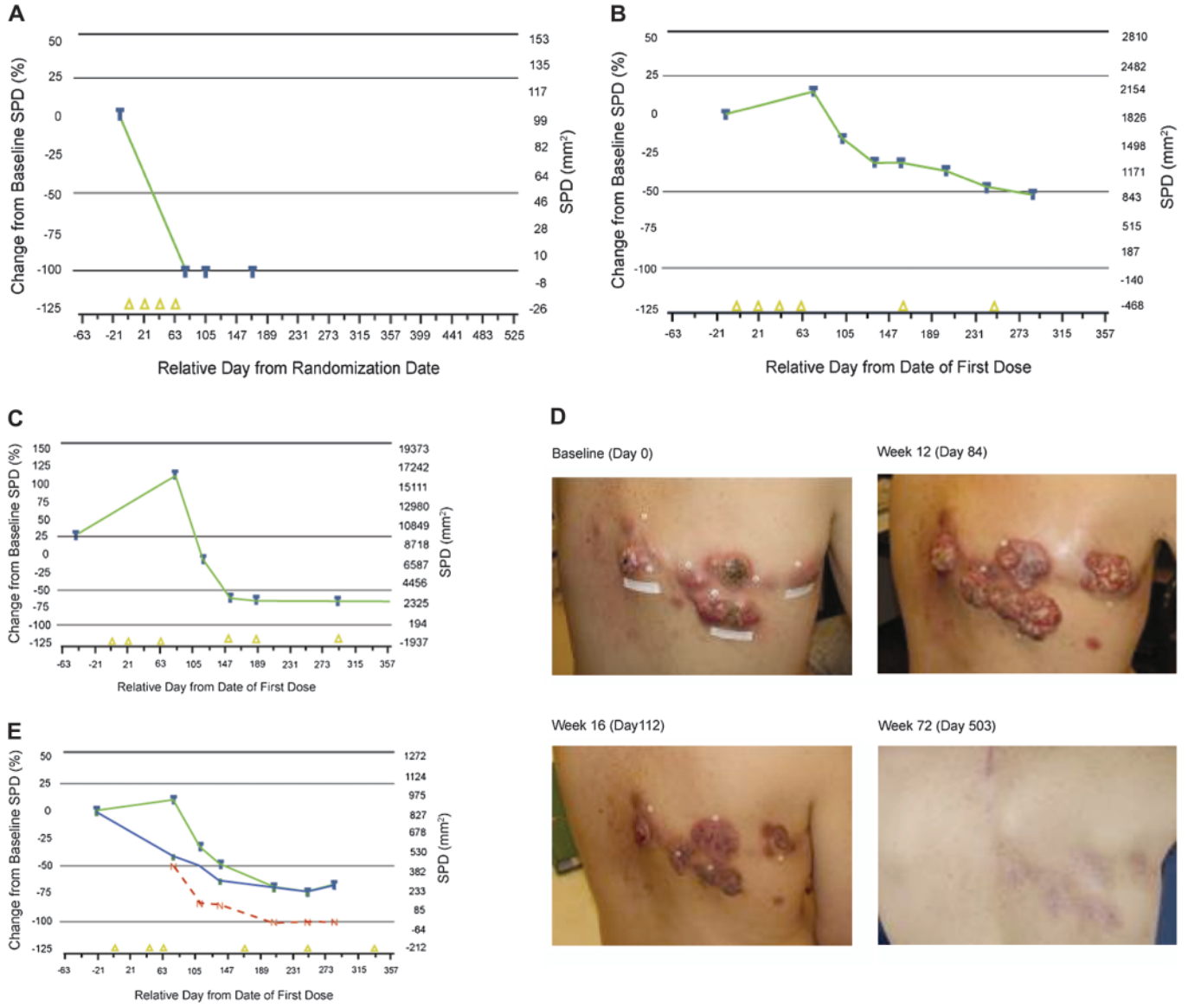

Figure 3. Clinical response patterns observed with anti-cytotoxic T lymphocyte–associated protein 4 immunotherapy (ipilimumab). Immunotherapy patterns of response depicted as a continuous variable of relative change of tumor burden (%) over time. Tumor burden is described through the sum of the perpendicular diameters (SPD) of all measurable lesions (baseline and new) at each time point. **A and B**) Conventional response patterns: (**A**) immediate response; (**B**) durable stable disease with possible slow decline in tumor burden. **C and E**) Novel immunotherapy response patterns: (**C**) increase in total tumor burden followed by response. (**D**) Clinical images corresponding to pattern (C): tumor burden on the skin at baseline **(day 0)** is increased at first follow-up **(day 84)** and subsequently declines **(day 112)** to a complete response (**day 503**) (courtesy of Dr K. Harmankaya). **E**) The second novel pattern shows a response in the presence of new lesions; existing lesions present at baseline (**blue**) and new lesions (**red**) are added to define the total tumor burden (**green**). Despite new lesions, the total tumor burden is still declining to a partial response. **Yellow triangles** indicate dosing with immunotherapy; **horizontal lines** indicate standard thresholds for response or progression. Modified after Wolchok et al. (43).

for each lesion. Measurability is defined as 5 × 5 mm or more on helical computer tomography scans. The sum of the perpendicular diameters (SPD) of index lesions at baseline is added to that of new lesions to calculate total tumor burden according to the following formula:

$$\text{Tumor Burden} = \text{SPD}_{\text{index lesions}} + \text{SPD}_{\text{new measurable lesions}}.$$

Thus, percentage changes in tumor burden between assessment time points describe the size and growth kinetics of total measurable tumor burden over time. Response categories under irRC are defined as immune-related complete response (irCR), immune-related partial response (irPR), immune-related stable disease (irSD), and immune-related progressive disease (irPD) using the same thresholds to distinguish between categories as defined under standard WHO criteria (Table 3). Decrease in total measurable tumor burden is assessed relative to the baseline tumor burden, that is, SPD of all index lesions at baseline. The response category irPD should be confirmed at two consecutive time points as already done for irPR or irCR. Overall, immune-related response based on two or more tumor assessments is derived as shown in Table 3.

Using irRC, the appearance of new lesions alone does not constitute irPD if they do not add to the tumor burden by at least

Table 3. Derivation of overall immune-related response in solid tumors*

Derivation of overall immune-related response for all assessed time points			
Measurable response	**Nonmeasurable response**		**Overall response**
Index and new measurable lesions (total measurable tumor burden)†	**Non-index lesions**	**New nonmeasurable lesions**	**Using irRC**
100% decrease	Absent	Absent	irCR‡
≥50% decrease	Any	Any	irPR‡
<50% decrease to <25% increase	Any	Any	irSD
≥25% increase	Any	Any	irPD‡

* After Wolchok et al. (43). irCR = immune-related compete response—complete disappearance of all index and new measurable lesions; irPR = immune-related partial response—decrease in tumor volume ≥50% relative to baseline; irSD = immune-related stable disease—not meeting criteria for irCR or irPR, in absence of irPD; irPD = immune-related progressive disease—increase in tumor volume ≥25% relative to nadir.

† Index and non-index lesions are selected at baseline. Index lesions are measurable (>5 × 5 mm), and non-index lesions are not measurable (<5 × 5 mm, ascites, bone lesions, etc.). Changes are assessed relative to baseline and include measurable lesions only (>5 × 5 mm).

‡ Assuming response and progression are confirmed by a second assessment at least 4 weeks apart.

25%. Patients with new lesions but an overall tumor burden decrease qualifying for partial response (≥50% decrease) or qualifying for stable disease (<50% decrease to >25% increase) are considered to have irPR or irSD, respectively (same percentage changes including new lesions) (Table 3). These new patterns are considered clinically meaningful because they appear to be associated with favorable survival (43,44). Importantly, early increase in the size of lesions, which may be attributable to the infiltration of lymphocytes, does not preclude an irCR, irPR, or irSD from being obtained at the next consecutive time point. If a patient is classified as having irPD, confirmation by a second scan in the absence of rapid clinical deterioration is required. Thus, the definition of confirmation of progression represents an increase in tumor burden of at least 25% compared with baseline at two consecutive time points at least 4 weeks apart. It is recommended that this confirmation be done at the discretion of the investigator in the context of the patient's tumor type, disease stage, and clinical status because awaiting a response after tumor burden increase may not be appropriate for patients with rapid symptomatic progression accompanied by a decline in performance status.

In summary, the development of irRC may provide a novel and long-needed tool for clinical trials to assess signs of activity of immunotherapies. Because data obtained with ipilimumab (43) suggest an association of immune-related response patterns with favorable patient survival, these criteria may identify important clinical patterns otherwise characterized as progressive disease by WHO criteria. Further prospective evaluation of irRC in immunotherapy trials is required to confirm their clinical utility. Complete details of the new irRC are described by Wolchok et al. (43).

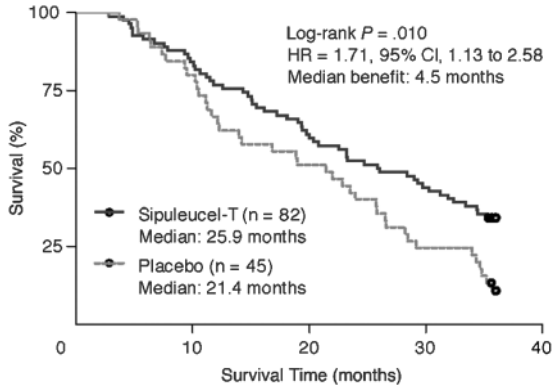

Figure 4. Delayed separation of survival curves of sipuleucel-T immunotherapy vs placebo in advanced prostate cancer, where the separation of Kaplan–Meier curves occurred after approximately 8 months after random assignment. HR = hazard ratio; CI = confidence interval. Courtesy of Small et al. (12). Reprinted with permission. Copyright 2008 American Society of Clinical Oncology. All rights reserved.

Survival Kinetics: Differences Between Immunotherapy and Conventional Therapies

In contrast to chemotherapy, for which an early clinical effect is possible, immunotherapies often demonstrate delayed clinical effects (12,14,17,46–49). For example, Figure 4 shows the overall survival outcome for a study of sipuleucel-T immunotherapy vs placebo in advanced prostate cancer (12), where the separation of Kaplan–Meier curves occurs after approximately 8 months after random assignment. In randomized trials, in which immunotherapies are compared with either placebo or inactive controls, Kaplan–Meier survival curves may be superimposable for a time before separation is observed. Generally, if there is such a delayed separation, the statistical power to differentiate the entire curves is reduced (1,9,50).

To better understand possible lessons from existing datasets of phase III immunotherapy trials, a workshop of the CIC-CRI in 2006 reviewed all publicly accessible data from such trials. This analysis suggested the possibility of a delayed separation of Kaplan–Meier curves for immunotherapeutic agents (mostly cancer vaccines) (1), including melacine in stage II melanoma (46), vitespen in advanced melanoma (47), vitespen in stage II/III renal cell carcinoma (48), sipuleucel-T (12) (Figure 4), PROSTVAC-VF (14) in advanced prostate cancer, and the anti-CTLA-4 antibody ipilimumab in advanced melanoma (17,49). A separation of curves in these trials was observed after 4–8 months or later following random assignment. However, such kinetics may differ among agents and disease settings, thus requiring individual investigation.

The survival analysis methods commonly used to design and analyze two-arm randomized clinical trials do not have a provision for data demonstrating a delay in the separation of Kaplan–Meier curves. These methods assume that the ratio of the arm-specific hazard rates (risks for patients to experience events) is a constant over time (proportional hazards assumption) and is necessary when using the standard formula for computing the number of events required for the final analysis of a trial (51). Furthermore, the log-rank test has optimal properties under proportional hazards, and Cox regression models assume proportional hazards. Under these

assumptions, the hypothesized hazard ratio (HR), describing the difference in survival between the study arms, applies immediately after random assignment, and the event rates at all times will have the same ratio to differentiate between the arms.

Delayed separation is a form of deviation from proportional hazards. Initially, the hazard rates are equal, and therefore, the hazard ratio is 1.0, and the Kaplan–Meier curve estimates will be indistinguishable (if an inactive comparator is used). Subsequently, the hazard rates become unequal and the curves will separate. For practical purposes, it is assumed for this discussion that the hazard rates after the delayed clinical effect eventually become proportional. Thus, the events occurring before the time when the hazard rates become unequal will not differentiate the study arms. As a consequence, computation of the required number of events for final analysis under proportional hazards assumptions will lead to a statistical power insufficient for a trial with a delayed separation. Depending on the timing of the delay, this loss of statistical power can be substantial (1,9,50).

Alternative methods should be considered to compute the required number of events for final analysis when delayed separation is expected (1,9). A variety of methods including simulation and numerical integration of a proposed theoretical hazard function can be used. Simulation is attractive because it may also account for other deviations from usual trial planning assumptions. However, any method used to compute the required number of events will require careful scrutiny and validation. Another critical element for the computation of the number of events is the timing (quantification) of the delayed effect. Ideally, the quantification used to compute the required events will come from previous randomized trials, possibly in the phase II setting (9).

The following example illustrates the practical implications of the delayed separation on study statistical power, number of events, and study duration. Figure 5 illustrates the mathematical form of a delayed separation, with a hazard ratio of 0.7 after the separation. The control arm has an exponential survival distribution with median survival of 18 months (red dashed curve). The form of the delayed separation is specified by a hazard ratio function (HR(t), solid grey line) that has value 1.0 for the first 3 months and then decreases linearly between 3 and 6 months to become 0.7 after 6 months. The experimental arm survival distribution (solid blue line) is the result of mathematically blending the control arm survival distribution function and the hazard ratio function. The survival distributions in Figure 5 can be used to illustrate the consequences of delayed separation. The following additional specifications are made: 1:1 random assignment, N = 600, accrual time 18 months, two-sided type I error probability .05, statistical power of 90%, and use of the log-rank statistic. Under proportional hazards (no delayed separation), 331 events are required for final analysis and are projected to occur at 2.85 years. With a delayed separation to 3 months as specified, 331 events would result in a statistical power estimate of only 62.7%. Using simulations that take into account delayed separation, it is found that 90% statistical power requires 510 events and would not be realized until 5.66 years.

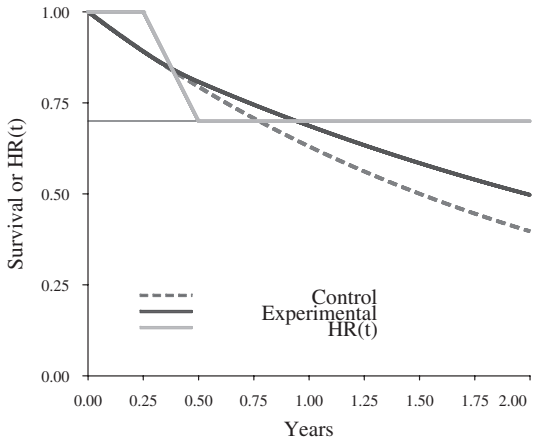

Figure 5. Mathematical illustration of a delayed separation of curves. Example of a two-arm study with an ultimate hazard ratio of 0.7. The control arm has an exponential survival distribution with median survival of 18 months (**red dashed curve**). The form of the delayed separation is specified by a hazard ratio function (HR(t), **solid gray line**) that has the value 1.0 for 3 months and then decreases linearly between 3 and 6 months to become 0.7 at 6 months and then remains constant. The experimental arm survival distribution (**solid blue line**) is the consequence of mathematically blending the control arm survival distribution function and the hazard ratio function.

It is therefore recommended that the computations of the required events be based on a plausible specification of the timing of the delayed separation, the desired statistical power, use of the log-rank statistic, and the understanding that the trial will be overpowered if delayed separation is not observed or is less than that specified. The log-rank test is recommended for study planning and final analysis to avoid deviation from conventional methods, to avoid prespecification of parameters describing the delay (52,53), and as a hedge against absence of a delay. The cost of using the log-rank statistic is some loss of optimal conditions if a delay is present. Because immunotherapeutic agents may differ regarding presence and timing of delayed separation, randomized phase II trials may improve the ability to plan for definitive phase III trials.

Further challenges include the planning of interim and futility analyses. These must be carefully considered because the presence of a delayed separation will inflate the chances of a negative early interim analysis and of concluding futility because of projected results without a delayed separation.

Conclusions

As our knowledge of immunologic and clinical science has evolved (4–7), we have begun to address the unique characteristics of immunotherapeutic agents in clinical trials and to use more appropriate endpoints (8,9,10,43). Immunotherapies demonstrate different kinetics compared with cytotoxic agents. Thus, they may induce cellular immune responses before influencing tumor burden or patient survival (8,9). For adequate investigation of immunotherapies in clinical trials, a new development paradigm is needed, including adjustment of established endpoints to address this biology. The challenges around adequate clinical endpoint use in immunotherapy trials were addressed through community-wide workshops paired with comprehensive laboratory and clinical programs providing large datasets (8,9,17,20–22,41,42).

The inability to use cellular (T-cell) immune response assays to define biomarkers and to investigate their relationship with clinical outcomes has its roots in highly variable and often nonreproducible assay results in multicenter trials (28,30). As was demonstrated

by several large international proficiency panels, the use of harmonized assays can reduce this variability (20–22) and may help to build a general framework for assay use in multicenter clinical trials similar to what the International Conference on Harmonization/Good Clinical Practice did for clinical protocols. Furthermore, harmonized assays may allow investigation of the relationship between immune responses and clinical outcomes. Correlation of immune response and clinical outcomes would certainly accelerate the learning process regarding the biological activity of immunotherapeutic agents in the clinic and help the early selection of promising candidates for advanced trials.

As observed in many clinical trials, immunotherapy can induce novel patterns of antitumor responses distinct from those of chemotherapy, which are consequently not captured by the WHO or RECIST criteria. In a 5-year collaborative effort between academia and industry (8,44), clinical observations were translated into new response criteria (43), which more comprehensively capture all observed response patterns. The irRC provide a systematic description for phenomena such as mixed responses, in which new lesions appear while existing lesions shrink or in which some lesions shrink while others grow. Through the new irRC system, immunotherapy patterns of response are described as tumor burden over time where tumor burden is the SPD of all measurable lesions (baseline and new). The irRC are generally based on the WHO and RECIST criteria and do not require a substantial departure from standard oncology practice. The novelty of the irRC lies in the measurement of new lesions, which are included in the overall tumor burden, allowing for it to be described as a continuous variable (before and after conventional progression) (9,44). This response analysis bears similarities to a concept described for the investigation of cytostatic agents such as sorafenib, in which tumor size was described as a continuous variable (54).

Histological evidence in cases in which biopsy material was available suggests that the appearance of new lesions or increase in the size of existing lesions may be the result of lymphocytic infiltration and may not represent true disease progression (40,55). In these cases, the irRC provide a means of accounting for delayed changes in tumor burden through confirmation of progression at subsequent time points. This new confirmation of progression is similar to the confirmation routinely done for response and may enable the detection of a change in kinetics when initial tumor burden increase through lymphocytic infiltration is followed by a lymphocyte-induced tumor response (delayed response). Importantly, initial observations in advanced melanoma with anti-CLTA-4 antibody indicate an association of immune-related responses with favorable survival, suggesting that these criteria may identify patients with previously unrecognized benefit (43,44). The clinical applicability of these criteria lies principally in the more comprehensive description of clinical signals of activity in early trials. Consequently, the irRC offer an additional tool for investigating immunotherapies and are undergoing prospective validation.

Considering the time of translation of immunologic responses into clinical activity, the survival of patients may not be affected until some months after treatment start compared with chemotherapy. A resulting delayed separation of Kaplan–Meier curves was observed in multiple randomized immunotherapy trials and was often not apparent until 4–8 months or more after random assignment (12,14,17,46–49). In patient populations with advanced cancers, a substantial number of events can occur in the time window before separation and thus lead to substantial loss of statistical power (1,9,50). Whereas it remains unclear whether past studies would have had a different outcome if delayed separation had been considered during trial planning, its consideration for future evaluations appears important. Knowledge of survival kinetics may also allow better planning of interim analyses because if they are performed too early, they may give misleading results and conclude futility prematurely. In addition, immunotherapeutic agents may differ in the timing of delayed separation and should be tested in randomized phase II trials to improve the planning for phase III trials. Use of modified statistical methods to characterize the hazard ratios before and after separation of survival curves allows improved planning of randomized trials in which a delayed separation of curves is expected.

In summary, the described recommendations for improved clinical endpoints have the potential to positively alter the clinical investigation of cancer immunotherapy.

Supplementary Data

Supplementary data can be found at http://www.jnci.oxfordjournals.org/.

References

1. Finke LH, Wentworth K, Blumenstein B, et al.. Lessons from randomized phase III studies with active cancer immunotherapies—outcomes from the 2006 meeting of the Cancer Vaccine Consortium (CVC). *Vaccine.* 2007; 25(suppl 2):B97–B109.
2. Copier J, Dalgleish AG, Britten CM, et al. Improving the efficacy of cancer immunotherapy. *Eur J Cancer.* 2009;45(8):1424–1431.
3. Rosenberg SA, Yang JC, Restifo NP. Cancer immunotherapy: moving beyond current vaccines. *Nat Med.* 2004;10(9):909–915.
4. Finn OJ. Cancer immunology. *N Engl J Med.* 2008;358(25):2704–2715.
5. Melero I, Hervas-Stubbs S, Glennie M, et al. Immunostimulatory monoclonal antibodies for cancer therapy. *Nat Rev Cancer.* 2007;7(2):95–106.
6. Peggs KS, Quezada SA, Korman AJ, et al. Principles and use of anti-CTLA4 antibody in human cancer immunotherapy. *Curr Opin Immunol.* 2006;18(2):206–213.
7. Ribas A, Butterfield LH, Glaspy JA, et al. Current developments in cancer vaccines and cellular immunotherapy. *J Clin Oncol.* 2003;21(12):2415–2432.
8. Hoos A, Parmiani G, Hege K, et al. A clinical development paradigm for cancer vaccines and related biologics. *J Immunother.* 2007;30(1):1–15.
9. Hoos A. Proposal of a clinical development paradigm for cancer immunotherapy: novel endpoints. In: Endpoints for Immunotherapy Studies: Design and Regulatory Implications, American Society of Clinical Oncology (ASCO) Annual Meeting; June 2, 2008; Chicago, IL.
10. Michaelis LC, Ratain MJ. Measuring response in a post-RECIST world: from black and white to shades of grey. *Nat Reviews Cancer.* 2006;6(5):409–414.
11. Bendandi M. Idiotype vaccines for lymphoma: proof-of-principles and clinical trial failures. *Nat Reviews Cancer.* 2009;9(9):675–681.
12. Small EJ, Schellhammer PF, Higano CS, et al. Placebo-controlled phase III trial of immunologic therapy with sipuleucel-T (APC8015) in patients with metastatic, asymptomatic hormone refractory prostate cancer. *J Clin Oncol.* 2006;24(19):3089–3094.
13. Small EJ, Fong L. Developing immunotherapy as legitimate therapy for patients with prostate cancer [published online ahead of print January 25, 2010]. *J Clin Oncol.* 2010;28(7):1085–1087.
14. Kantoff PW, Schuetz TJ, Blumenstein BA, et al. Overall survival analysis of a phase II randomized controlled trial of a poxviral-based PSA-targeted immunotherapy in metastatic castration-resistant prostate cancer [published online ahead of print January 25, 2010]. *J Clin Oncol.* 2010;28(7):1099–1105.

15. Hamid O, Urba WJ, Yellin M, et al. Kinetics of response to ipilimumab (MDX-010) in patients with stage III/IV melanoma [abstract 8525]. *J Clin Oncol.* 2007;25(suppl):18s.
16. Wolchok JD, Ibrahim R, DePril V, et al. Antitumor response and new lesions in advanced melanoma patients on ipilimumab treatment [abstract 3020]. *J Clin Oncol.* 2008;26(suppl):19s.
17. Wolchok JD, Neyns B, Linette G, et al. Ipilimumab monotherapy in patients with pretreated advanced melanoma: a randomised, double-blind, multicentre, phase 2, dose-ranging study. *Lancet Oncol.* 2010;11(2): 155–164.
18. Inoges S, Rodriguez-Calvillo M, Zabalegui N, et al. Clinical benefit associated with idiotype vaccination in patients with follicular lymphoma. *J Natl Cancer Inst.* 2006;98(18):1292–1301.
19. Longo DL. Idiotype vaccination in follicular lymphoma: knocking on the doorway to cure. *J Natl Cancer Inst.* 2006;98(18):1263–1265.
20. Janetzki S, Panageas KS, Ben-Porat L, et al.; for the Elispot Proficiency Panel of the CVC Immune Assay Working Group 2007. Results and Harmonization Guidelines from two large-scale international Elispot proficiency panels conducted by the Cancer Vaccine Consortium (CVC/SVI). *Cancer Immunol Immunother.* 2008;57(3):303–315.
21. Britten CM, Gouttefangeas C, Schoenmaekers-Welters MJP, et al. The CIMT-Monitoring panel: a two-step approach to harmonize the enumeration of antigen-specific CD8+ T lymphocytes by structural and functional assays. *Cancer Immunol Immunother.* 2008;57(3):303–315.
22. Britten CM, Janetzki S, Ben-Porat L, et al.; for the HLA-peptide Multimer Proficiency Panel of the CVC-CRI Immune Assay Working Group. Harmonization guidelines for HLA-peptide multimer assays derived from results of a large scale international proficiency panel of the Cancer Vaccine Consortium (CVC). *Cancer Immunol Immunother.* 2009;58(10):1701–1713.
23. *Guidance for Industry: Clinical Considerations for Therapeutic Cancer Vaccines.* U.S. Department of Health and Human Services, Food and Drug Administration, Center for Biologics Evaluation and Research; 2009. http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/default.htm.
24. Maecker HT. The role of immune monitoring in evaluating cancer immunotherapy. In: Disis ML, eds. *Cancer Drug Discovery and Development: Immunotherapy of Cancer.*Totowa, NJ: Humana Press; 2005:59–72.
25. Hobeika AC, Morse MA, Osada T, et al. Enumerating antigen-specific T-cell responses in peripheral blood: a comparison of peptide MHC Tetramer, ELISpot, and intracellular cytokine analysis. *J Immunother.* 2005;28(1):63–72.
26. Walker EB, Disis ML. Monitoring immune responses in cancer patients receiving tumor vaccines. *Int Rev Immunol.* 2003;22(3–4):283–319.
27. Keilholz U, Weber J, Finke LH, et al. Immunologic monitoring of cancer vaccine therapy: results of a workshop sponsored by the Society for Biological Therapy. *J Immunother.* 2002;25(2):97–138.
28. Janetzki S, Britten CM, Kalos M, et al. MIATA—minimal information about T cell assays. *Immunity.* 2009;31(4):527–528.
29. Tuomela M, Stanescu I, Krohn K. Validation overview of bio-analytical methods. *Gene Ther.* 2005;12(suppl 1):S131–S138.
30. Britten CM, Janetzki S, van der Burg SH, et al. Toward the harmonization of immune monitoring in clinical trials: quo vadis? *Cancer Immunol Immunother.* 2008;57(3):285–288.
31. International Conference on Harmonization (ICH). *Guidance for Industry: Good Clinical Practice*. 1997;62(90):25691–25709E7.
32. NCCLS. *Evaluation of Matrix Effects, Approved Guidelines. NCCLS document EP14-A* [ISBN 1-56238- 434-1]. Wayne, PA: NCCLS.
33. Janetzki S, Price L, Britten CM, et al. Performance of serum-supplemented and serum-free media in IFNgamma Elispot Assays for human T cells. *Cancer Immunol Immunother.* 2010;59(4):609–618.
34. *WHO Handbook for Reporting Results of Cancer Treatment.*Geneva, Switzerland: World Health Organization Offset Publication No. 48; 1979.
35. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. *J Natl Cancer Inst.* 2000;92(3): 205–216.
36. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumors: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45(2):228–247.
37. Berd D, Sato T, Cohn H, et al. Treatment of metastatic melanoma with autologous, hapten-modified melanoma vaccine: regression of pulmonary metastases. *Int J Cancer.* 2001;94(4):531–539.
38. Kruit WHJ, van Ojik HH, Brichard VG, et al. Phase 1/2 study of subcutaneous and intradermal immunization with a recombinant MAGE-3 protein in patients with detectable metastatic melanoma. *Int J Cancer.* 2005;117(4):596–604.
39. van Baren N, Bonnet MC, Dréno B, et al. Tumoral and immunologic response after vaccination of melanoma patients with an ALVAC virus encoding MAGE antigens recognized by T cells. *J Clin Oncol.* 2005;23(35): 9008–9021.
40. Hodi FS, Butler M, Oble DA, et al. Immunologic and clinical effects of antibody blockade of cytotoxic T lymphocyte-associated antigen 4 in previously vaccinated cancer patients. *Proc Natl Acad Sci U S A.* 2008;105(8):3005–3010.
41. Weber J, Thompson JA, Hamid O, et al. A randomized, double-blind, placebo-controlled, phase II study comparing the tolerability and efficacy of ipilimumab administered with or without prophylactic budesonide in patients with unresectable stage III or IV melanoma. *Clin Cancer Res.* 2009;15(17):5591–5598.
42. O'Day S, Ibrahim R, DePril V, et al. Efficacy and safety of ipilimumab induction and maintenance dosing in patients with advanced melanoma who progressed on one or more prior therapies [abstract 9021]. *Proc Am Soc Clin Onc.* 2008;26(20 suppl).
43. 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(23):7412–7420.
44. Hodi FS, Hoos A, Ibrahim R, et al. Novel efficacy criteria for antitumor activity to immunotherapy using the example of ipilimumab, and anti-CTLA-4 monoclonal antibody [abstract 3008]. *J Clin Oncol.* 2008;26(suppl):19s.
45. Tuma RS. New response criteria proposed for immunotherapies. *J Natl Cancer Inst.* 2008;100(18):1280–1281.
46. Sosman JA, Unger JM, Liu PY, et al.; Southwest Oncology Group. Adjuvant immunotherapy of resected, intermediate-thickness, node-negative melanoma with an allogeneic tumor vaccine: impact of HLA class I antigen expression on outcome. *J Clin Oncol.* 2002;20(8):2067–2075.
47. Testori A, Richards J, Whitman E, et al.; for the C-100-21 Study Group. Comparison of autologous tumor-derived heat shock protein gp96-peptide complexes (vitespen) and physician's choice in a randomized phase 3 trial in patients with stage IV melanoma. *J Clin Oncol.* 2008;26(6):955–962.
48. Wood C, Srivastava P, Bukowski R, et al.; for the C-100-12 RCC Study Group. An adjuvant autologous therapeutic vaccine (HSPPC-96; vitespen) versus observation alone for patients at high risk of recurrence after nephrectomy for renal cell carcinoma: a multicentre, open-label, randomised phase III trial. *Lancet.* 2008;372(9633):145–154.
49. Hodi FS, O'Day S, McDermott DF, et al. Improved survival with ipilimumab in patients with metastatic melanoma [published online ahead of print June 14, 2010]. *N Engl J Med.* 2010doi:10.1056/NEJMoa1003466.
50. Fine GD. Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Inf J.* 2007;41:535–539.
51. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis.*New York, NY: John Wiley & Sons; 1991.
52. Lakatos E. Sample size based on the log-rank statistic in complex clinical trials. *Biometrics.* 1988;44(1):229–241.
53. Sposto R, Stablein D, Carter-Campbell S. A partially grouped logrank test. *Stat Med.* 1997;16(6):695–704.
54. Karrison TG, Maitland ML, Stadler WM, et al. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *J Natl Cancer Inst.* 2007;99(19):1455–1461.
55. Hodi FS, Oble DA, Drappatz J, et al. CTLA-4 blockade with ipilimumab induces significant clinical benefit in a female with melanoma metastases to the CNS. *Nat Clin Pract Oncol.* 2008;5(9):557–561.

Funding

Funding for scientific meetings, workshops, and the immune monitoring proficiency panels was provided by the Cancer Immunotherapy Consortium of the Cancer Research Institute, a nonprofit organization.

Notes

We thank all participants of the workshops and community-wide initiatives who contributed knowledge to this article. A. Hoos, R. Ibrahim, A. Anderson, and R. Humphrey are employees of Bristol-Myers Squibb. J. Wolchok is an ad hoc Advisory Board member (compensated) for Bristol-Myers Squibb. F. S. Hodi has received research support and served as a consultant for Bristol-Myers Squibb. Clinical trials with ipilimumab immunotherapy were sponsored by Bristol-Myers Squibb.

Affiliations of authors: Cancer Immunotherapy Consortium of the Cancer Research Institute (CIC-CRI; formerly Cancer Vaccine Consortium), New York, NY (AH, SJ, RI, LO, JW); Global Clinical Research Oncology, Bristol-Myers Squibb, Wallingford, CT (AH, RI, AA, RH); Department of Surgery, Erasmus University Medical Center, Rotterdam, the Netherlands (AMME); ZellNet Consulting, Fort Lee, NJ (SJ); Department of Medical Oncoology, Dana-Farber Cancer Institute, Boston, MA (FSH); TriArc Consulting, Washington, DC (BB); Ludwig Institute for Cancer Research, New York, NY (LO); Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY (JW).