ISSUE BRIEF

Conference on Clinical Cancer Research

SEPTEMBER 2008

ENGELBERG CENTER for Health Care Reform at BROOKINGS

AACR CENTENNIAL 1907 - 2007 American Association for Cancer Research

ASCO® American Society of Clinical Oncology

FRIENDS of CANCER RESEARCH

Inside cover
BLANK

CONTENTS

Left page after TOC
BLANK

PANEL 1
Data Submissions Standards and Evidence Requirements

Richard Schilsky, MD, University of Chicago Medical Center
Jeffrey Abrams, MD, National Cancer Institute
Janet Woodcock, MD, Food and Drug Administration
Gwen Fyfe, MD, Genentech
Robert Irwin, Marti Nelson Cancer Foundation

The need for data collection and reporting standards

The design of any clinical trial involves a variety of tradeoffs. One set of issues involves balancing the opportunity to collect large amounts of data on each subject with other important priorities, such as the ability to verify the quality of the data, the number of subjects that can be studied in the trial, the length and intensity of follow-up, and the cost of the trial. In some cases, the most important factor is randomizing a sufficient number of subjects with only a few data points per subject, the large simple trial – for example, an outcome study of an approved drug. In many other cases, such as trials intended to support initial regulatory approval, extensive amounts of data are often collected on each subject. More data are sometimes collected than is necessary to ensure that cancer treatments are safe and effective, increasing the cost and duration of clinical trials. A risk also exists that the magnitude of data collection may compromise overall data quality by creating an enormous burden on investigators and clinical study sites.

The U.S. Food and Drug Administration (FDA) has issued Guidance on data collection in registration trials for cancer therapies and general Guidance on how to conduct safety reviews.[1,2,3] However, these Guidances are not prescriptive and sponsors often err on the side of caution by collecting a great deal of data on each subject. Variability exists across development programs regarding the nature and extent of data collection. Some variability is expected, given the diversity of drugs, diseases, phases of clinical investigation, and ultimate intended use of the product. However, better agreement and understanding about the essential data elements that should be collected for various types of oncology trials intended for regulatory submission would be beneficial. It is appropriate to ask what should be collected, why, and whether greater efficiency can be achieved – not only in developing a standardized approach, but also in minimizing certain aspects of data collection without jeopardizing development of the database essential to evaluate the safety and efficacy of the agent. Common data collection and reporting standards can improve the efficiency of clinical cancer research.

In cancer, different kinds of primary approvals have led to confusion and inconsistency in the approach to safety data collection for a supplemental approval. What is required in supplemental approvals depends very much on how “well-characterized” the safety profile of the agent is, and on the intended use of the product. This confusion and subsequent lack of consistency in regulatory approach sometimes leads to detailed and prolonged negotiations between the FDA and the sponsor on a protocol-specific basis, as well as inconsistent approaches to data collection.

These issues further burden a clinical trials system that is already struggling. In contrast to other countries where 20 percent or even greater percentages of adult cancer patients may enroll in trials, less than five percent of U.S. adult patients participate. In part this deficiency is caused by the high costs and low reimbursement offered to sites participating in clinical trials, as well as the regulatory risks associated with participation in complicated trials that have detailed data collection requirements. Given the broad commercial availability of oncology agents following initial regulatory approval, clinical trial participation in the U.S. is increasingly undertaken only by trial sites that are

deeply committed to the assessment of investigational agents in the medical milieu of America's health care system.

The Brookings Conference on Clinical Cancer Research provides a unique opportunity to discuss data collection elements for therapeutic cancer trials. This paper will serve as background for Panel 1 by laying out the benefits of data collection standards, proposing a framework for data collection, and noting areas where further dialogue on standards is necessary.

Benefits of data standards

A variety of data elements and formats for collection exist. As studies of adverse event (AE) data collection have demonstrated, the method of data collection can profoundly influence how a trial's results are interpreted. One study found that rates of reported AEs are dependent upon how information is gathered; patients who received a checklist of adverse events to complete reported significantly more adverse events than those who were asked open-ended questions.[4] The study authors observed that comparing treatments for the same indication will be uninformative if studies of those treatments use different methods for collecting AE data.[5] Other variables that may impact reporting of adverse events include the frequency of follow-up visits and how trial paperwork is completed.[6]

Uniform data collection standards could improve the interpretation of trial results by delineating consistent standards for content and methodology. Consistency of approach is likely to improve site performance. Conversely, the variety of approaches currently in use may hinder good safety data collection. A recent review of 22 trials sponsored by the National Cancer Institute (NCI) discovered that AEs for some trials were reported differently in published articles versus the NCI's Clinical Data Update System. The authors noted that published reports inadequately described low-grade, high-grade, or recurring AEs.[7] Another study of safety data reporting in randomized drug trials found that none of the 192 trials reviewed exhibited "safety reporting [that] can be deemed satisfactory."[8]

Faster and less costly trials

Pharmaceutical company sponsors and academic investigators/sponsors often gather more data than are necessary. The extent to which the adverse event data collected during trials – especially trials that are intended to support a subsequent cancer indication following full approval and a well-characterized safety profile – are useful for gauging treatment risks is not currently known. Recent post hoc simulations based on data gathered during the Avastin Non-Small Cell Lung Cancer (AVAIL) trial demonstrated that toxicity data collection could be streamlined without significantly increasing the risk of missing an important adverse event. If toxicity data on Grade 1 and 2 events had been collected in a subset of 200 patients per arm rather than in every trial participant (n=~650), approximately 2,500 fewer adverse events would have been collected; extrapolations to estimate monitoring and transcribing efforts suggest that this could result in a time savings of at least 2,500 hours.[9] When researchers simulated a scenario in which Grade 3 and 4 events were collected in a subset of patients (instead of all patients) during a large trial, adverse events occurring at least five percent more frequently in the study drug arm were almost always observed in the smaller subset,

although those at least two percent more frequent were missed almost half the time.[10] This simulation includes serious adverse events reported by the parallel expedited adverse event regulatory reporting system for "serious" events, which ensures that all toxicities leading to death or hospitalization are captured. Whether such abbreviated data capture is viewed as adequate depends in part on the size of the pre-existing database and the degree to which regulators and health care professionals find that this new information is likely to provide novel insights.[11]

Impact on regulatory review and approval

A shared understanding between regulators and sponsors regarding the quality and quantity of data that are necessary to adequately weigh a treatment's risks and benefits will benefit all. Clinical trials could be designed and conducted more efficiently and the regulatory review process could be more uniform and rapid if a set of data collection and reporting standards were consistently applied to clinical trials conducted by industry, academia, and the NCI's Cooperative Groups. The challenge, however, is in balancing the need for efficiency in the conduct of trials with the need for developing adequate evidence to demonstrate safety and efficacy and for ensuring adequate labeling to inform clinical use. In addition, standards for individual data elements, as well as standard data collection approaches, need to be developed.

A proposed framework for data collection standards

Ideally, cancer clinical trial data collection standards would be based on what is already known about the treatment under investigation, the objectives of the study of the treatment, the study population and the intended use of the agent. Moreover, the standards would be flexible enough to deal with specific subsets and risk groups of interest.

When little is known about the therapeutic agent, as would usually be the case at the time of the primary (initial) cancer indication, the studies should include fairly extensive data collection.* This is congruent with the expectation of international regulatory bodies and general practice for pharmaceutical development. The FDA has approved cancer therapeutics with substantially less comprehensive data collection and in the absence of substantial prior existing data; however because of the great uncertainty associated with oncology drug development, this approach is risky and not recommended. Cancer therapeutics are particularly likely to fail in development compared to other therapies – one estimate shows that 60 percent of cancer drug development programs fail in the late clinical phase. On the chance that the cancer under study is incurable and the new drug offers substantial potential benefit, such as a survival advantage, patients and their physicians are generally willing to accept a greater degree of uncertainty about side effects. Too often, however, the efficacy is substantially less than anticipated. That, coupled with the inability to characterize the safety profile, may result in the need to conduct more trials.

A complete set of standards would identify a "core set" of data elements, along with recommendations about which elements to include in protocols and case report forms under various circumstances. Ultimately, such standards would be endorsed by regulatory agencies worldwide, so that evidence based on these standards could be globally applied.

All parties concur that a substantial amount of data reduction is not appropriate when little is known about the agent; usually this will be in the setting of the initial indication. Even in this setting, however, there is not universal agreement on what is essential, and a range of practices exist. Table 1 is intended to reflect what has been the range of data collection practices in the setting where the safety profile has not been established.

In contrast, for supplemental approvals or when substantial safety data are available, data reduction may be considered – for example, where the agent has previously received full approval (in contrast to accelerated approval) based on a randomized controlled trial and a substantial safety database of many hundreds of patients. Table 2 is intended to reflect areas where data reduction may be appropriate.

* Extensive data may already be available about the agent, such as from studies in other indications, or earlier studies in certain cancers that failed to demonstrate efficacy but generated substantial safety data. In such cases, it may be appropriate to reduce data collection in the new study intended to support an initial approval.

Most clinical studies intended to support a marketing claim will be the subject of an End-of-Phase II meeting with the FDA, which provides the opportune time to refine and modify the data collection elements as the situation warrants. The Brookings conference is the beginning of a process to develop a common approach to data collection for therapeutic oncology trials; identification of a common approach will, in turn, inform and improve the End-of-Phase II meetings.

The core set includes the following categories of data elements: eligibility, on-study form, medical history, lab findings, disease measurement, treatment, vital signs, non-protocol therapy, long-term follow-up, concomitant medications, and toxicity.

The overriding principle is that less data may be collected when substantial safety data are already available; ordinarily, we consider this in the context of secondary versus primary indications, although occasionally substantial data are available at the time of the initial indication. Regardless of whether the indication is primary or secondary, when the safety profile of the agent has been well-characterized, the burden on investigators, human subjects, sponsors, regulators, and patients in need of the treatment under study can be lessened without compromising the adequate evaluation of risk and benefit. It could also be argued that, by focusing investigator efforts on critical safety variables, that more accurate and complete safety reporting of the most serious events will be facilitated.

For drugs without a well-characterized toxicity profile, most participants agree that collection of all Grade 3-4 AEs, at least a subset of Grade 1-2 AEs, and all concomitant medications is appropriate. For both primary and secondary indications, investigators would always collect data on deaths, serious adverse events, and adverse events leading to the discontinuation of treatment.

For secondary indications (i.e., when data are deemed sufficient to characterize the safety of the agent), some members of the panel suggest that investigators collect data on Grade 3-4 adverse events by cycle in either a subset of sites or another subset of patients. In addition, based on the known biology and safety profile of the study drug, targeted AEs could be collected in all patients. Similarly, targeted concomitant medications could be collected in a subset of sites or patients based on the pharmacology of the drug and its known safety profile.

The suggested new approaches for concomitant medications and toxicity might lead to concern that insufficient data will be available on which to assess safety and formulate risk/benefit assessments. If less data on concomitant medications were collected, investigators might overlook possible drug-drug interactions that they might otherwise catch. Likewise, collecting less data on toxicity during trials that study secondary indications could be argued to risk underestimating the rates of a treatment's adverse effects. This might be a particular concern when the populations for initial and supplemental indications differ greatly, such as a third-line setting in advanced cancer for initial and adjuvant setting for supplement. The benefit/risk analysis may differ, and assessment of lower-grade adverse events potentially might become more important. These issues will be raised during the panel discussion.

Minimizing collection of redundant or unnecessary data in these two categories when testing secondary indications could potentially make trials more efficient without posing additional risk to patients, especially in the context of adequate post-marketing surveillance. It would be important in this context to specify what designs would make post-marketing surveillance "adequate" to detect events that could have been missed using this type of data collection. When a treatment undergoes testing for secondary indications, data on adverse events – including complete sets of information on Grades 1 through 4 adverse events, deaths, drug discontinuations, dose decreases and delays, chemotherapy modifications, and serious adverse events – should already be available. Moreover, in many circumstances, the additional information gained from the full cohort of subjects may not be more informative than that based on a subset. Empirical studies and statistical simulations have

demonstrated that while increasing the number of subjects can narrow confidence intervals, the point estimate for AE rates are relatively stable under small samples. As a result, the amount of AE data gathered should be carefully considered, particularly in instances where a treatment is undergoing testing for a secondary indication.[12]

Beyond the framework outlined here, it may also be useful to establish an additional set of standards related to post-marketing evaluation. The conduct of post-approval trials according to appropriately designed standards might facilitate the more rapid approval of certain therapies. Such standards would aim to increase precision around safety and efficacy estimates – perhaps in specific subpopulations. Under the FDA Amendments Act of 2007 (FDAAA), the FDA can require post-marketing studies to further characterize the safety profile of an approved drug or biological product. Implementation of the FDA Sentinel Initiative to conduct population-level post-marketing surveillance will be an additional step toward ensuring that gains in efficiency are not offset by risk to patients.[13]

Electronic data reporting

These recommendations identify what investigators should collect, but do not address how the standard items should be collected. We support the collection of clinical trial data through electronic means using consistent definitions and formats across trials. The connection of data collection standards and electronic data reporting standards, as developed by the Clinical Data Standards Interchange Consortium (CDISC) and Health Level 7 (HL7), can further increase clinical trial efficiency and speed up regulatory review.

CDISC and HL7 are two principal groups that have taken on the task of creating clinical-research reporting standards. CDISC has created models to standardize the submission of data to regulators, and HL7 produces standards for clinical and administrative data. In order to promote interoperability among applications using CDISC or HL7 standards, the FDA, NCI, CDISC, and HL7 created the Biomedical Research Integrated Domain Group (BRIDG).[14] The BRIDG model, released in 2007, has been adopted by the NCI's cancer Biomedical Informatics Grid (caBIG) initiative.

CaBIG is an open-source, open-access, open-development, and federated platform that facilitates the collection and sharing of standardized data across participating sites, and shows great potential for increasing the pace of innovation in cancer treatment. CaBIG offers participants tools that aid clinical trials management, integrate and analyze various types of data, and assure secure data-sharing connections among the 56 NCI cancer centers and 16 community health centers in the network. Whenever possible, investigators should utilize the CaBIG platform and adhere to accepted standards for data reporting. Of particular interest will be the forthcoming proposed rule for *Electronic Submission of Data from Studies Evaluating Human Drugs and Biologics*, expected in September 2008.

Areas for panel discussion, with a focus on scenarios where an adequate safety database exists

- Data collection on CRFs for toxicity;
- Data capture on CRFs for concurrent medications;
- Types of studies in post-marketing that can provide robust safety information; and
- A request for industry to conduct retrospective and prospective evaluations of differing amounts of toxicity or concomitant medication data collection in their clinical trials.

Conclusion

Development of standards for data collection – both qualitative and quantitative – can improve the efficiency of cancer clinical research and development. Adoption and consistent implementation of such standards throughout the development process could help facilitate the rapid development of safe and effective cancer therapies.

Table 1. Standard practices for data collection when little prior data available†

DATA TYPE	SCOPE OF COLLECTION
Eligibility	Collect major inclusion/exclusion criteria (e.g. PS, disease or treatment characteristics) as individual yes/no boxes and remaining eligibility as a single yes/no on a case report form (CRF); do not collect source data (e.g. labs, scans).
On-Study Form	Collect all relevant patient and baseline characteristics.
Medical History	Collect targeted baseline medical history in checkbox format (e.g. diabetes, hypertension requiring treatment, history of myocardial infarction).
Physical Exam	Variable: (ranges from all to none)
Lab findings	Varies: from all routine laboratory values at baseline and during treatment to subset, and via central lab vs. site laboratory.
Disease Measurement	Collect all tumor assessment measurements at all time points.
Treatment	Collect actual dose and treatment date, or reason for modification, delay, hold, or discontinuation.
Vital Signs	Collect routine vital signs. Collect weight/height or body surface area (BSA) on initial Treatment page. If a change from the initial dose, a reason (weight change, toxicity, protocol specified, etc.) must be provided.
Non-Protocol Therapy (NPT)	Collect all NPT (but not doses), including start and stop date (month/year), until first progression. (Need to clearly define what therapies are included.)
Concomitant Medications	Collect all concomitant medications at baseline by name; practices vary from all start and stop dates, to by cycle.
Toxicity	Collect deaths, Grades 3-4 toxicity, serious AEs, AEs leading to discontinuation of treatment. For Grades 1-2, practices range from collection of all grades with start and stop dates, to all grades by cycle, to collection of Grades 1-2 in a subset.
Long-Term Follow-Up	First treatment initiated after disease progression; dose and duration of treatment not needed.

† This usually applies to studies supporting initial indications; the table shows a range of practices.

Table 2. Data collection for secondary indications or where substantial data exist‡

DATA TYPE	SCOPE OF COLLECTION
Eligibility	Collect major inclusion/exclusion criteria (e.g. PS, disease or treatment characteristics) as individual yes/no boxes and remaining eligibility as a single yes/no on a case report form (CRF); do not collect source data (e.g. labs, scans).
On-Study Form	Collect all relevant patient and baseline characteristics.
Medical History	Collect targeted baseline medical history in checkbox format (e.g. diabetes, hypertension requiring treatment, history of myocardial infarction).
Physical Exam	Do not record physical exam on CRF.
Lab findings	Do not collect routine laboratory values (except in the case when they are eligibility criteria or where certain targeted laboratory data are important) at baseline or during treatment except as adverse events. However, if there is a lab-related serious adverse event (SAE), the SAE should include whether the patient's initial value was normal, prior treatment values that were abnormal and related history.
Disease Measurement	Collect all tumor assessment measurements at all time points.
Treatment	Collect actual dose and treatment date, or reason for modification, delay, hold, or discontinuation.
Vital Signs	Do not collect routinely except where certain targeted vital signs are important. Collect weight/height or body surface area (BSA) on initial Treatment page. If there is a change from the initial dose, a reason (weight change, toxicity, protocol specified, etc.) must be provided.
Non-Protocol Therapy (NPT)	Collect all NPT (but not doses), including start and stop date (month/year), until first progression. (Need to clearly define what therapies are included.)
Concomitant Medications	Needs further discussion Current proposal: Collect targeted concomitant medications by specific name based on safety profile of drug. Collect at baseline and when a SAE occurs.
Toxicity	Needs further discussion Current proposal: Collect deaths, targeted AEs, serious AEs, AEs leading to the discontinuation of treatment; collect Grades 3-4 events by cycle at subset of sites (or patients).
Long-Term Follow-Up	First treatment initiated after disease progression; dose and duration of treatment not needed.

‡ Data collection refers to data that are specifically recorded on case report forms; it is expected that all patients will receive routine evaluations (physical examination including vital signs, laboratory evaluation, etc.) as per standard of medical care.

References

[1] U.S. Food and Drug Administration. Reviewer Guidance: Conducting a Clinical Safety Review of a New Product Application and Preparing a Report on the Review [online], <http://www.fda.gov/CDER/guidance/3580fnl.pdf> (2005).

[2] U.S. Food and Drug Administration. Guidance for Industry: Premarketing Risk Assessment [online], <http://www.fda.gov/CDER/guidance/6357fnl.pdf > (2005).

[3] U.S. Food and Drug Administration. Guidance for Industry: Cancer Drug and Biological Products - Clinical Data in Marketing Applications [online], <http://www.fda.gov/cder/guidance/4332fnl.htm> (2001).

[4] Bent, S., Padula. A. and Avins, A.L. Brief Communication: Better Ways to Question Patients about Adverse Medical Events: A Randomized, Controlled Trial. *Ann. Intern Med.* 144, 257-261 (2006).

[5] Bent, S. Padula, A. and Avins, A.L. 2006.

[6] Ioannidis, J.P.A., Mulrow, C.D., and Goodman, S.N. Adverse Events: The More You Search, the More You Find. *Ann. Intern. Med.* 144, 298-300 (2006).

[7] Scharf, O. and Colevas, A.D. Adverse Event Reporting in Publications Compared With Sponsor Database for Cancer Clinical Trials," *J. of Clin. Oncol* 24, 3933-3938 (2006).

[8] Ioannidis, J.P.A. and Lau, J. Completeness of Safety Reporting in Randomized Trials: An Evaluation of 7 Medical Areas. *J. Am. Med.* Assoc. 285, 437-443 (2001).

[9] Genentech, data on file.

[10] Genentech, data on file.

[11] Genentech, data on file.

[12] Mahoney M.R. et al. Dealing with a Deluge of Data. *J. of Clin. Oncol.* 23, 9275-9281 (2005).

[13] U.S. Food and Drug Administration. FDA Sentinel Initiative: A National Strategy for Monitoring Medical Product Safety [online], <http://www.fda.gov/oc/initiatives/advance/reports/report0508.html> (2008).

[14] Richesson, R. and Krischer, J. Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions. *J Am. Med. Inform. Assoc.* 14, 687-696 (2007).

PANEL 2
Improved Insights into Effects of Cancer Therapies

Raymond DuBois, MD, M.D. Anderson Cancer Center
Donald Berry, PhD, M.D. Anderson Cancer Center
Jim Doroshow, MD, FACP, National Cancer Institute
Paolo Paoletti, MD, Glaxo SmithKline
Richard Pazdur, MD, Food and Drug Administration
Nancy Roach, C3: Colorectal Cancer Coalition

The need for clarity on efficacy endpoints

In an effort to accelerate safe and effective cancer drug development and to decrease the time to drug approval, the oncology community has long sought endpoints other than overall survival (OS) to evaluate new agents. Measures of disease progression, health-related quality of life, patient-reported symptoms, and biomarkers have been proposed and tested in clinical studies, but consensus has not been reached on the role of these endpoints in determining the overall clinical benefit of a therapy. One auxiliary endpoint that has been the focus of particularly intense discussion is progression-free survival (PFS), which employs the RECIST criteria to determine the progression of cancer based on imaging.[1] PFS is the length of time during and after treatment in which a patient is living with a disease that does not worsen, according to the established criteria. RECIST defines progression as an increase of 20 percent in a single dimension on computed tomogram or magnetic resonance imaging. However, concern about potential biases in the measurement of disease progression by radiographic imaging has resulted in debate over the current use of PFS in clinical trials.

Two recent cancer drug approvals were based solely on evidence of PFS, underscoring the need to reach agreement on how this endpoint is defined and interpreted. In December 2005, the U.S. Food and Drug Administration (FDA) approved sorafenib for the treatment of advanced kidney cancer based on an increase in progression-free survival, despite the absence of a statistically significant benefit in overall survival. More recently, the FDA granted accelerated approval for bevacizumab in patients with breast cancer. Data showed that a combination of bevacizumab and paclitaxel nearly doubled median PFS compared with paclitaxel alone, but the secondary endpoint of overall survival in this trial did not reach statistical significance. A review of the data by the FDA will be required for the accelerated approval to be converted into a full approval by the end of 2008. If upcoming trials do not show a survival benefit, then the accelerated approval for breast cancer could be revoked or curtailed until more data are collected.

Lack of clarity around the appropriate use of PFS and other auxiliary endpoints can be a barrier to efficient clinical cancer research, as well as to the review and approval of cancer therapies. This paper will brief conference attendees on the complex issues surrounding measurement of treatment efficacy and propose a set of principles to guide the evaluation and use of auxiliary endpoints. Finally, these principles will be illustrated using PFS as a case study.

Issues around the use of auxiliary endpoints

The gold standard for clinical effectiveness of a given agent is an improvement in a defined endpoint in a randomized clinical trial.[2,3] In oncology, overall survival is most often seen as the best single agreed-upon endpoint. However, randomization rarely occurs in the Phase II setting, as is common in other areas of drug development. Instead, Phase II trials measure the rate of complete and partial

response to given agents prior to progression to randomized Phase III trials. Single-arm, historically-controlled Phase II trials are rarely employed outside of oncology. This issue is multifactorial, but this may contribute to the high failure rate of drugs proceeding from the Phase II to the Phase III setting in cancer drug development. Thus, there is a need for endpoints that can quickly detect drug efficacy or failure, in order to avoid unnecessary resource allocation to drugs that will ultimately fail to exhibit a patient benefit.

Overall survival is objectively measured and not prone to the potential investigator biases associated with endpoints that require clinical judgment. However, using overall survival as a primary endpoint significantly slows the rate of cancer drug development. As approved therapies have become increasingly effective at prolonging survival, so too have they prolonged the duration of trials designed to detect that endpoint. Delays in recruitment and follow-up ultimately serve to prolong the regulatory review and approval of newer agents that could provide needed options for cancer patients. Furthermore, overall survival is a crude instrument for measuring the effects of many targeted therapies, which may be designed to work in a subset of patients with specific molecular targets. Thus, a great need exists to define and validate alternative markers of effect.[4]

While the term "surrogate endpoint" has been more commonly used in the literature, "auxiliary endpoint" is a preferable term because the endpoints under investigation are not meant to supplant more conventional endpoints, but rather to be evaluated in conjunction with other endpoints. We define auxiliary endpoints to include the collection of endpoints – other than overall survival – 1used to infer the effects of cancer therapies from clinical trials. Auxiliary endpoints may be primary or secondary endpoints within a trial, and may include progression-free survival (time to progression), response rate, patient-reported outcomes (e.g., quality of life), and biomarkers (e.g., tumor size, circulating tumor cells, and tumor-specific markers). Clearly defining the strengths, limitations, and appropriate uses of auxiliary endpoints could accelerate the development, review, and approval of new treatments.

Principles for the evaluation and use of auxiliary endpoints

We propose three basic principles to consider when selecting auxiliary endpoints for a given trial. First, a strong biological rationale should support the potential auxiliary endpoint as a marker of treatment effectiveness. For example, biomarkers that predict variability in survival time may be preferred endpoint candidates. Second, the potential auxiliary endpoint should be shown to explain variability in treatment outcomes in terms of survival for treated patients.[5] Third, ideal auxiliary endpoints should accurately assess the efficacy of the drug being evaluated with minimal risk of subjectivity or bias. Where the possibility of bias exists, the trial design should compensate by seeking to minimize potential bias.[6]

The development of new drugs for AIDS patients closely follows this model of auxiliary endpoint development. CD4 count and viral load were validated as auxiliary endpoints in trials in the late 1980s and early 1990s, allowing for an explosion in the available therapies for AIDS patients. Applying this model to the oncology community will be more difficult. Since cancer involves many heterogeneous disease processes, many auxiliary endpoints will need to be developed according to these criteria.

Toward a rational and valid process for evaluating progression-free survival

Progression-free survival (PFS) is a desired endpoint in many settings, but it is not a surrogate for overall survival. Advantages of PFS as a primary endpoint include a more rapid clinical trial and the elimination of confounding effects when evaluating experimental therapies in diseases with existing effective therapies. For example, if a patient enters a clinical trial after four failed conventional therapies and later discontinues that trial due to progression, numerous other approved therapies

may be available. The patient could survive long after a given trial ceased accruing patients, and other therapies could contribute to his or her demise, minimizing the effect of the experimental therapy. Further, non-trivial improvements in PFS are considered a clinical benefit in some settings. Patients may see a benefit in a lack of progression in their tumor burden, irrespective of the benefits in overall survival.

Although PFS has many advantages, it is not without limitations. Unlike OS, the precise timing of PFS is not known. This leads to the potential for evaluation-time bias, which produces biased estimates of treatment effectiveness when the evaluation times for progression status differ by treatment arm.[7] Further, elements of subjectivity remain in spite of efforts, such as RECIST, to standardize the evaluation of progressive disease.[8] Indeed, a non-trivial number of discrepancies between radiologists evaluating the progression status of the same patients are to be expected. These discrepancies can come from multiple sources. At the start of trials, baseline lesions are usually defined, but occasionally, these lesions are altered or ignored in the course of a trial. Radiographic scans can be misplaced, leading to clinical judgments based on varied amounts of radiographic information. In addition, radiologists may have different interpretations of the available scans. In a recent trial, the discrepancy rate between two expert radiologists blinded to the treatment assignment was 34 percent. When these discrepancies are unrelated to treatment, they are a source of measurement variability, which results in attenuated estimates of effect sizes. Measurement variability reduces the power to detect a true difference but will not lead to invalid conclusions when the experimental therapy is truly ineffective. In other words, measurement variability alone will not result in ineffective therapies entering the oncology community. However, if the variability is large enough, it could preclude effective therapies from being revealed.

The most significant concern about discrepancies in assessment arises when progressive disease evaluations are influenced by an investigator's lack of objectivity about the therapies under study. The potential for evaluations to be influenced by knowledge of treatment assignment, combined with pre-existing views about their relative effectiveness, has led to the introduction of Blinded Independent Central Reviews (BICR) as a suggested means of validating efficacy in trials with PFS endpoints. However, the use of BICR is problematic and may lead to invalid analyses, as it does not always provide an unbiased estimate of a treatment's effectiveness. Specifically, BICR analyses for PFS are likely subject to the presence of informative censoring, which invalidates standard analyses. The methodology relies on the assumption that censoring is independent of factors associated with progression or survival.

Informative censoring arises in the following manner: Patients who progress by investigator assessment may not have the same assessed time of progression under the BICR. Once a patient has progressed according to the investigator, he or she will be taken "off protocol" and further follow-up is not likely. If the BICR does not determine that a patient has progressed by the time the patient is off protocol, the patient is censored for the purpose of analysis. This patient, however, is more likely to have progressed, as assessed by BICR, sooner than those remaining in the at-risk cohort. This violates the standard assumptions for censoring subjects and, as a result, survival-analysis estimates are biased. Further, although methods for modeling informative censoring exist, these methods cannot conclusively eliminate the potential effects of informative censoring. Dodd et al. provide a more detailed discussion of this issue with an example from a clinical trial.[9]

In a review of Phase III oncology trials published in the last five years that had BICRs as a component of assessment, no cases were shown to have substantial differences between analyses between the BICR and investigator assessments.[10] (See Tables 1 and 2.) The lack of differences is striking in light of the seemingly high discrepancy rates between BICR and local review, which range from 36 to 53 percent. However, these discrepancies are likely due in large part to random, rather than systematic, differences between the clinicians who evaluate the radiographic imaging scans. This variation in assessment between two independent reviewers is a

well-studied phenomenon in many therapeutic areas.[11] Further, there was no trend that would indicate that either BICR or local review resulted in a stronger treatment effect.

In conclusion, BICR does not necessarily provide a less biased estimate of a treatment's effectiveness than local review, and situations in which the BICR conclusions differ from those based on the investigators' assessments result in an ambiguous situation. The discrepancy may be caused by measurement variability, informative censoring, or true evaluation bias. Methods that effectively reduce evaluation bias where it is most likely to affect trial outcomes are needed. Four approaches are worthy of consideration.

Proposals for the auditing of PFS

Matter for clarification: No BICR when trials are double-blinded.

Blinding of treatment assignment would eliminate systematic bias in PFS evaluation related to knowledge of treatment assignment. Therefore, there should be no requirement for central review in double-blinded trials, except in the case where an extreme imbalance between treatment arms in the incidence of side effects could lead to a considerable level of unblinding. This level of imbalance would be characterized by the majority of patients in the treatment arm experiencing a particular side effect with a virtual absence of this same side effect in the control arm.

Case 1: An open-label superiority trial with an BICR-based audit of progression.

Detection of meaningful evaluation bias will be gauged via an audit of progression determinations. BICR could be performed in both arms of a trial on a subset of cases. A sample size for the audit would be specified in advance (for example, 10 percent of participants or a minimum number of cases). If bias is suspected, then the audit would expand to a larger proportion of cases. The goal of the audit would be to determine whether there is a meaningful difference in hazard ratios between the local review and BICR. It is recognized that, given the potential biases present with both BICR and local review, a discrepancy in assessments would make a conclusion about a treatment's efficacy more difficult.

Large effect sizes will likely be robust to small discrepancies between treatment arms, while smaller effect sizes will be quite sensitive to small discrepancies. Therefore, when effect sizes are large, relatively smaller audits may be necessary to detect the amount of bias needed to alter the trial conclusion substantively. However, in some cases, no audit, no matter its size, can rule out evaluation bias.

Because the goal of the audit is to detect actual bias, measurement variability should be controlled. Technologies that enable synchronization, allowing patients to be followed by the BICR as a trial is ongoing, are strongly encouraged.

It is recognized that data-driven analyses are necessary to develop the scientific justification for the selection of the recommended audit size. The Phamaceutical Researchers and Manufacturers of America (PhRMA) and the National Cancer Institute (NCI) have begun research projects to address this specific issue. The NCI will be collecting patient-level data from multiple large clinical trials with data from both central and local reviews to better inform the audit process. Since it not expected that these data sets will contain meaningful bias, such bias will be introduced into the data so that the auditing strategy can be tested. Simulation studies, based on an understanding of the trial data, will also inform recommendations. Clearly, an understanding of what is an important level of bias for a particular study given an observed effect size is needed.

Case 2: An open-label superiority trial with large effect size.

When treatment effect sizes are large enough, an audit is not necessary, since evaluation bias is not expected to be of a magnitude that would meaningfully impact the observed effect size. As part of this proposal, increased monitoring of the protocol-specified imaging procedures at the local site could be undertaken. It is expected that the investigator is the greatest potential source for bias in a PFS assessment. It should also be noted that a local radiologist is frequently unaware that patients are participating in a clinical trial, so a procedure that records the measurements or progression assessments of both the radiologist and the investigator is recommended. Whenever the investigator overrides the radiologist's determination, the reasons for this will be documented. When this occurs more frequently in one treatment arm and the reasons are not easy to verify objectively, concern about bias will arise.

Case 3: PFS evaluation at two time points with auditing at these evaluation times.

Evaluation of treatment effectiveness could be based on the proportion of patients whose cancer has progressed at two time points, rather than using an analysis based on a survival model. Two time points for imaging assessments would be determined prospectively, corresponding to the approximate median PFS and approximately twice median PFS of the control arm or conventional therapies. Summary statistics would include the proportion alive and progression-free at each time point. Progressions that have been documented prior to the designated imaging assessment time would be counted as an event for the rate of progression or death, and images would be audited at the two time points. For patients who progress prior to the designated scan times, the audit would be based on the scan that determined progressive disease.

This two-point approach reduces evaluation-time bias and results in a simpler trial design.[12] Since the approach limits the focus to the two imaging assessments, the issues of compliance, timing, rigor, and consistency are easier to maintain or verify. Further, central review of two time points should be easier to implement. While one might have concerns about a loss in power of the trial design as compared to a log-rank analysis, the loss in power with two time points is less than that from a single time point. Indeed, Freidlin et al. demonstrate that there is little risk in major power loss from this approach.[13] The trade-off for some loss in power, however, is decreased susceptibility to bias.

Conclusion

This paper has presented a proposal for auditing PFS in three different scenarios. Establishing such auditing procedures can help build confidence in PFS as an indicator of clinical benefit. The suggestions listed above also hint at a way forward for improving the reliability of information produced by other auxiliary endpoints. If the cancer research community can determine how to most effectively utilize auxiliary endpoints – without compromising the quality of safety and efficacy data – cancer patients will benefit greatly.

Table 1. Trials that have used retrospective blinded independent central reviews[a]

Disease	Trial	Sample size	Hazard ratio and 95% confidence interval per central review	Hazard ratio and 95% confidence interval per local review
Renal Cell Carcinoma	sorafenib vs. placebo[14], b	903	0.44 (0.35-0.55)	0.51 (0.43-0.60)
	sunitinib vs. interferon alpha[15]	750	0.42 (0.32-0.54)	0.42 (0.33-0.52)
Colorectal cancer	panitumumab plus best supportive care vs. best supportive care[16]	463	0.54 (0.44-0.66)	0.39 (0.32-0.48)
Breast Cancer	lapatinib plus capecitabine vs. capecitabine[17]	324	0.49 (0.34-0.71)	0.59 (0.42-0.84)
	bevacizumab plus capecitabine vs. capecitabine[18]	462	0.98 (0.77-1.25)	0.90 (0.72-1.12)
	ixabepilone plus capecitabine vs capecitabine[19]	752	Median PFS[c] 5.8(5.45-6.97) vs 4.2(3.81-4.50) months	Median PFS[c] 5.3 vs 3.8 months[d]
	bevacizumab plus paclitaxel vs paclitaxel[20, 21]	722	0.48 (0.39, 0.61)	0.42 (0.34,0.52)

Notes for Table 1

[a] We reviewed the literature and searched PubMed for studies in breast cancer, colorectal cancer, lung cancer and renal cell carcinoma. Search terms included, "progression free survival" or "time to progression," with filters of "randomized controlled trial" and "published in last five years." This revealed 209 manuscripts, of which only six reported having a central review of progression. The bevacizumab plus paclitaxel trial in breast cancer (last row) was included separately because it generated much discussion during an FDA Oncologic Drug Advisory Committee meeting on Dec. 5, 2007. All of these trials implemented a retrospective BICR. The panitumimab trial allowed cross-over at the time of locally determined progression amongst patients receiving the control treatment. As a result, patients for whom progression was not confirmed centrally continued to be evaluated centrally for progression.

[b] Double-blinded trial

[c] Hazard ratios not reported for local review.

[d] Difference statistically significant ($p<0.0011$). 95% CI for median PFS not reported.

Table 2. Discrepancy rates for three trials with central review

	Discrepancy Rate in Assignment of Progression/ Censoring Datea	Discrepancy Rate in Assignment of PFS Status	Per Central Review		Per Local Review	
			HR	95% CI	HR	95% CI
Lapatinib plus capecitabine v capecitabine[22]	Lapatinib plus capecitabine 87 of 163 = 53% Capecitabine, 69 of 161 = 43%	Lapatinib plus capecitabine, 40 of 163 = 25% Capecitabine, 40 of 161=25%	0.49	0.34 to 0.71	0.59	0.42 to 0.84
Bevacizumab plus paclitaxel v paclitaxel[23, 24]	Bevacizumab plus paclitaxel 118/330 = 35.7% Paclitaxel, 114/319 = 35.7%b, 25	Bevacizumab plus paclitaxel 90 of 368 = 24.5% Paclitaxel, 84 of 354 = 23.7%c	0.48	0.39 to 0.61	0.42	0.34 to 0.52
Bevacizumab plus capecitabine v capecitabine[25]	Bevacizumab plus capecitabine, 88/232=38%, Capecitabine, 99/230=43%27	Both arms combined: 105 of 462 = 23%	0.98	0.77 to 1.25	0.90[28]	0.72 to 1.12[29]

Notes for Table 2

[a] Computed as agreement in date of progression or date of censoring.

[b] Estimated amongst the 649 (of 722) patients for whom images were available for central review. An agreement was counted if dates were within 6 weeks of one another. This is in contrast to the lapatinib plus capecitabine and bevacizumab plus capecitabine trials, in which exact date was used for agreement.

[c] A discrepancy was counted if either status assignment differed or if no image was available for central review. As a result, a total of 722 (and not 649) patients were included.

References

[1] Therasse, P. et al. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *J. Natl. Cancer Inst.* 92, 205-216 (2000).

[2] Freidlin, B. et al. Proposal for the Use of Progression-Free Survival in Unblinded Randomized Trials. *J Clin Oncol.* 25, 2122-2126 (2007).

[3] Ratain, M.J. et al. Recommended changes to oncology clinical trial design: Revolution or evolution? *Eur. J. Cancer.* 44, 8-11 (2008).

[4] Schilsky, R.L. End Points in Cancer Clinical Trials and the Drug Approval Process. *Clin. Cancer Res.* 8, 935-938 (2002).

[5] Ellenberg, S.S. Surrogate end points in clinical trials. *BMJ*. 302, 63-64 (1991).

[6] Pazdur, R. Endpoints for Assessing Drug Activity in Clinical Trials. *Oncologist* 13 suppl 2, 19-21 (2008).

[7] Freidlin, B. et al., 2007.

[8] Therasse et al., 2000.

[9] Dodd, L.E., et al. Blinded Independent Central Review of Progression-Free Survival in Phase III Clinical Trials: Important Design Element or Unnecessary Expense? *J. Clin. Oncol.* 26, 3791-3796 (2008).

[10] Dodd, L.E. et al., 2008.

[11] Feinstein, A.R. A bibliography of publications on observer variability. *J. Chronic Disease.* 38, 619-632 (1985).

[12] Freidlin, B. et al., 2007.

[13] Friedlin, B. et al., 2007.

[14] Escudier, B. et al. Sorafenib in advanced clear-cell renal-cell carcinoma. *N. Engl. J. Med.* 356, 125-134 (2007).

[15] Motzer, R.J. et al. Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. *N. Engl. J. Med.* 356, 115-124 (2007).

[16] Van Cutsem, E. et al. Open-Label Phase III Trial of Panitumumab Plus Best Supportive Care Compared with Best Supportive Care Alone in Patients With Chemotherapy-Refractory Metastatic Colorectal Cancer. *J. Clin. Oncol.* 25, 1658-1664 (2007).

[17] Geyer, C.E. et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N. Engl. J. Med.* 355, 2733-2743 (2006).

[18] Miller, K.D. et al., Randomized Phase III Trial of Capecitabine Compared with Bevacizumab plus Capecitabine in Patients with Previously Treated Metastatic breast Cancer. *J. Clin. Oncol.* 23, 792-799 (2005).

[19] Thomas, E.S. et al. Ixabepilone Plus Capecitabine for Metastatic Breast Cancer Progressing after Anthracycline and Taxane Treatment. *J. Clin. Oncol.* 25, 5210-5217.

[20] U.S. Food and Drug Administration. *FDA Briefing Document: Oncology Drug Advisory Committee Meeting BLA STN 125085/91.018 Avastin® (bevacizumab).* (2007).

[21] Genentech, Oncology Drugs Advisory Committee Meeting: 5 December 2007.

[22] Geyer, C.E. et al., 2006.

[23] U.S. Food and Drug Administration, 2007.

[24] Genentech, Oncology Drugs Advisory Committee Meeting: 5 December 2007.

[25] Personal communication. Suman Bhattacharya, PhD, Bio-oncology, Genentech.

[26] Miller, K.D. et al., 2005.

[27] Personal communication, Suman Bhattacharya.

[28] Personal communication, Suman Bhattacharya.

[29] Personal communication, Suman Bhattacharya.

PANEL 3

Co-Development of Diagnostics and Therapeutics

Daniel Hayes, MD, University of Michigan
Steven Gutman, MD, MBA, Food and Drug Administration
Richard Frank, MD, PhD, GE Healthcare
Nancy Roach, C3: Colorectal Cancer Coalition
Richard Simon, DSc, National Cancer Institute
Ray Woosley, MD, PhD, Critical Path Institute

Personalized cancer therapy requires co-development

For decades, cancer therapies worked by non-selectively inhibiting rapidly dividing cells. The effectiveness of treatments was thus determined by how much toxicity could be tolerated and how well toxicity could be managed. Today, advances in molecular biology, genetics and imaging have enabled the identification of more specific disease targets and the development of therapies which act directly on those targets. Several examples of targeted cancer therapies include the following:

- Endocrine therapies (such as tamoxifen and the aromatase inhibitors);
- Anti-HER2 therapies for breast cancer (such as trastuzumab and lapatinib);
- Imatinib, the first drug to directly turn off the signal of a protein known to cause a cancer; and
- Anti-EGFR therapies for patients whose tumors overexpress the EGFR protein due to a specific gene mutation (such as cetuximab, gefitinib and erlotinib).

Determining whether individual cancer patients are likely to respond to targeted therapies – and thus more effective, efficient use of those therapies – is a key step in fulfilling the promise of personalized medicine. Unfortunately, the development of diagnostic tests to identify patients who will benefit from targeted therapies has typically lagged behind the development of the therapies themselves. A diagnostic test may be developed after a corresponding treatment has received regulatory approval (often aided by archived specimens collected during trials of the therapy); prior to the development of a corresponding treatment (in which case the diagnostic test could be used to measure efficacy of future therapies); or at the same time as the targeted therapy (co-development).

Ideally, therapies and their targets would be developed and approved in parallel, so that both are marketed at the same time. However, few co-development efforts have been successful to date, and the promise of personalized cancer therapy remains largely unfulfilled. This paper will serve as background to the Panel 3 discussion at the Brookings Conference on Clinical Cancer Research by (1) describing the current problems and barriers to development of diagnostics, and (2) identifying promising models for regulatory review of diagnostics in general and for co-development in particular.

Three barriers to effective co-development

Currently, there are three main impediments to the efficient development of diagnostics for tumor markers, which we will define as molecular or process-oriented assays beyond classic hematoxylin and eosin pathology or standard imaging, that indicate future behavior of a cancer – either independently of therapy (designated prognostic factors) or specifically related to the likelihood that a therapeutic strategy will work (designated predictive factors). Summarily, problems exist in the (1) translational research and product pipelines in diagnostics for tumor markers, (2) processes for regulatory evaluation and approval globally, and (3) inadequate reimbursement for innovative, highly effective diagnostics for tumor markers.

Problems in Translational Research and Pipelines

Translation of basic-science discoveries in the field of cancer genomics into products and therapies has been slow in recent years, which has concerned all parties invested in this research area. In order to address this problem, the National Institutes of Health launched the Roadmap Initiative, and the U.S. Food and Drug Administration (FDA) created the Critical Path Initiative (CPI). The CPI delineated the "pipeline problem" for both therapeutics and diagnostics in 2004, noting that the rate of development has declined for new drugs and diagnostics over the preceding several years despite an explosion in scientific discovery.[1] According to Phillips, et al., who conducted interviews with stakeholders from the diagnostics industry and regulatory agencies, addressing translational-research challenges in the area of biomarkers and diagnostics is essential. Specific scientific challenges include identification markers of abnormal cellular signaling pathways, identification of pre-treatment biomarkers that predict patient response to specific therapies, and development of *in vitro* assays and imaging diagnostics with sufficient sensitivity and specificity to be clinically useful. These barriers combine with the issues below to discourage venture capitalists and other investors from funding diagnostics companies' research, further contributing to an empty pipeline.[2]

Current Regulatory Challenges

The FDA faces the challenge of simultaneously addressing scientific rigor, practicality, and efficiency in the process of regulating co-developed technologies to use in risk assessment, screening for early detection, diagnosis, staging and prognosis for choice of therapeutic approach, and monitoring of treatment effect for individualization of regimen. Clear understanding of what data are required by FDA to demonstrate the benefits of using a particular biomarker test is essential to warrant pharmaceutical and device companies' investment in their clinical trials.

Currently, diagnostic and therapeutic products applications are reviewed in two distinct Centers at FDA, each with their own criteria for approval. Different regulatory statutes and standards at the Centers make co-development of tumor markers and drugs particularly challenging. Historically, predictive tests (tests that predict whether a patient will respond to a specific drug) and drugs have been developed separately. For example, the test that evaluates HER2 status in women was developed prior to the research that demonstrated that trastuzumab increased survival in women with HER2+ tumors. It is comparatively difficult to design a clinical-trials program that shows the safety and efficacy of a drug and demonstrates the functionality and clinical utility of a companion diagnostic. This difficulty leads to significant increases in both research costs and time to market. Adding to the complexity are tests that evaluate multiple markers simultaneously, associated labeling changes, and determination of the appropriateness of prospective clinical trials addressing the use of the marker versus *prospectively planned analyses* or retrospective studies of archived tumor samples. Given that diagnostics and therapeutics are equally essential for personalized cancer medicine, addressing these joint issues in their regulatory evaluation is critical.

The imperative is to use "qualified" biomarkers in research and "approved" diagnostics in clinical practice to promote the appropriate use of cancer therapies, resulting in improved patient outcomes, more efficient delivery of health care, and wider access to novel therapeutics and diagnostics.

Inadequate Reimbursement

Reimbursement for diagnostic products by private payers and Medicare does not provide adequate support for sustaining the development and use of new diagnostics that meet criteria for clinical utility. Payment for lab tests is largely based on the tests' incremental costs, rather than a broader determination of their value. The policies for fee determination and adjustment were enacted in the mid-1980s, and are outdated in light of the newfound importance of molecular diagnostics in targeted cancer treatments. Typically, reimbursement is set by a fee cap, known as the National Limitation Amount (NLA).[3] The NLA is calculated in two steps. First, the median fee paid for a specific

test by Medicare's regional carriers is determined. Some of these payment rates, though occasionally adjusted for inflation, are based on lab charges from 1983.[4] Then the median fee is reduced by a specified percentage; over the years, this percentage has decreased from 115 percent of the median fee for lab fees to only 74 percent.[5]

Fees for new tests are set according to mechanisms known as "gap-filling" or "cross-walking." While the cross-walking procedure applies to tests that resemble pre-existing technology, gap-filling is used to determine reimbursement for innovative tests. The gap-filling procedure gives regional Medicare carriers wide latitude in setting their own payments for a new test. CMS collects this information and uses it to establish an NLA for the test. This process can result in fees that are set below the cost of the test and which cannot be easily changed.[6]

The consequences of poor reimbursement include less investment in new diagnostic tests and the failure of some diagnostics companies. The recent bankruptcy of Immunicon – developer of the first quantitative assay for circulating tumor cells – and the subsequent acquisition of its assets by Johnson & Johnson have been interpreted by some observers in the cancer community as yet another sign that the current reimbursement environment cannot sustain the development and commercialization of diagnostics unless the costs of diagnostics can be subsidized by a corresponding treatment.

But large pharmaceutical companies have also been reluctant to engage in drug-diagnostic co-development. These firms may perceive that diagnostic development slows the drug development process while adding little value to research portfolios. Given current reimbursement policies described above, diagnostics are less profitable than treatments. Thus, diagnostics that result in targeted use of a comparatively well-reimbursed treatment can reduce not only revenue but also profit margins. Drug developers are also sensitive to the risk that an otherwise marketable treatment could be denied FDA approval if its corresponding assay is not approved.

In order to modernize reimbursement and thus the economic incentives for contemporary diagnostic technology, the clinical and economic value of these tests must be demonstrated and communicated to payers and patients in meaningful terms. Such evaluations must consider the cost offsets that come from reduced untargeted utilization of therapies based on the sensitivity and specificity of the test. New models of reimbursement for stand-alone diagnostics may need to differ from reimbursement for diagnostic-therapeutic combinations.

Three recommendations

Perhaps the most direct way to remove the barriers above is to develop a clear path for the co-development, co-review, and co-approval of therapeutic/diagnostic combinations. We recommend three specific lines of activity to accomplish that objective:

1) A clear pathway for development of diagnostics for tumor markers should be developed in consultation with the community external to the FDA, and the procedures and timeline for doing so should be clearly outlined in an FDA Guidance. Members of the community that should be engaged in this effort include relevant Device Advisory Panels – in particular, the Immunology and Hematology Panel responsible for providing input on tumor markers – professional societies, and the FDA-convened Oncologic Drugs Advisory Committee (ODAC), made up of extramural experts who assess data on cancer treatments and make non-binding recommendations on whether or not treatments being considered should be approved.

   External professional societies possess expertise in diagnostic development. For example, the American Society of Clinical Oncology has had a standing Tumor Markers Guidelines Committee for the last decade, and the National Cancer Center Network has a strong record of developing clinical guidelines that have included use of tumor markers. Likewise, the American Association

for Cancer Research has recently partnered with the FDA and the National Cancer Institute (NCI) to review the field of tumor marker development. These organizations could provide experience and expertise to develop a clear pathway for marker approval, and to generate a committee similar to ODAC to address Tumor Markers (see the third recommendation). However, it is also important to ensure that unique diagnostic and device issues are considered as well by including participation by relevant members of the Center for Devices and Radiological Health and in some, cases, Center for Biologics Evaluation and Research panels.

2) Tumor-marker clearance and approval should be based on demonstrated clinical benefit. However, when a marker is being co-developed with a therapy, this pathway should be approached in the most practical manner possible to avoid delay in patient access to a drug known to work.

3) An ODAC-like advisory committee should be developed for tumor marker clearance and approval in order to improve consistency and coordination with other oncology programs in the agency. Membership in this system should include an appropriate mixture of expertise including clinicians, trialists, laboratorians, statisticians and representatives of consumer and advocacy groups. The proper mix of expertise is critical to ensure good science and sound public policy.

Regulatory review of co-developed combinations

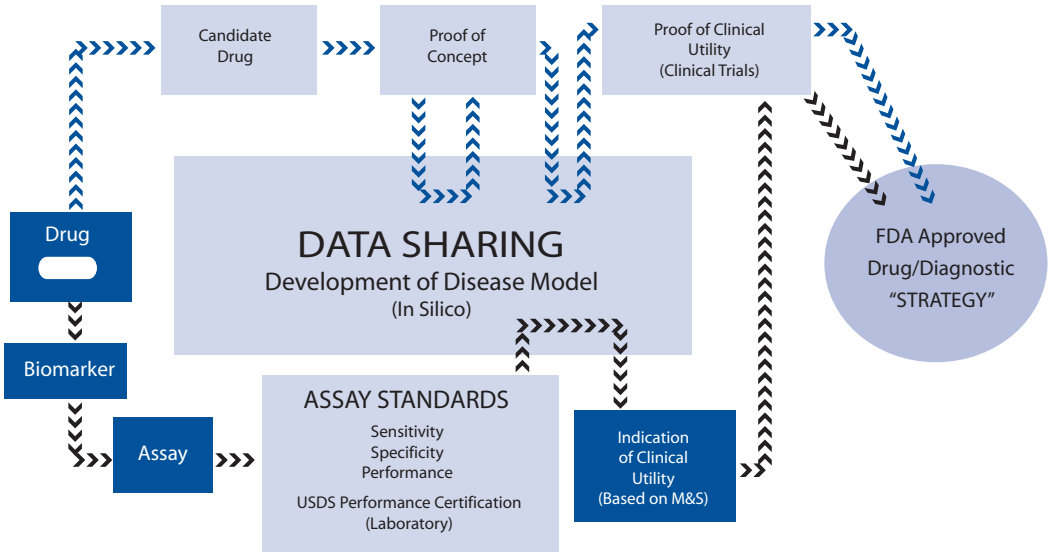
In April 2005, the FDA issued a concept paper, "Drug-diagnostic co-development," which outlined preliminary Agency thoughts on how to prospectively co-develop a drug or biological therapy and diagnostic test in a scientifically robust and efficient way.[7] Among the important issues discussed in that paper are:

- *Review procedure issues:* processes and procedures for submitting and reviewing a co-developed drug-test product
- *Analytical test validation:* the in vitro ability to accurately and reliably measure the analyte of interest, including analytical sensitivity and specificity, with focus on the laboratory component of drug/test development
- *Clinical test validation:* the ability of a test to detect or predict the associated disorder in patients, including clinical sensitivity and specificity, and/or other performance attributes of testing biological samples
- *Clinical test utility:* elements that should be considered when evaluating the patient risks and benefits in diagnosing or predicting efficacy or risk for an event (drug response, presence of a health condition)[8]

Figure 1, also presented in the concept paper, depicts a possible pathway for the development and regulation of a therapy and a corresponding assay. In this model, the regulatory process is coordinated so that the diagnostic and the therapy would, if approved, enter the market at the same time. Co-development remains on the Center for Drug Evaluation and Research (CDER) Guidance Agenda for 2008, and we recommend that it be prioritized and completed.

The clearance and approval of all diagnostic tests, whether or not co-developed, should be based on demonstrated clinical benefit. However, efforts to refine the regulatory process for diagnostics should also ensure that regulation of biomarkers does not become so burdensome as to discourage co-development and to render tumor-marker evaluation impractical. This can potentially be accomplished by defining different models for study designs addressing the clinical utility of biomarkers for *existing drugs* versus biomarkers that are paired with *new drugs*. Different guidelines for conducting prospective studies versus retrospective analyses of archived samples should also be introduced.

Figure 1. Drug-device co-development process: key steps during development[9]



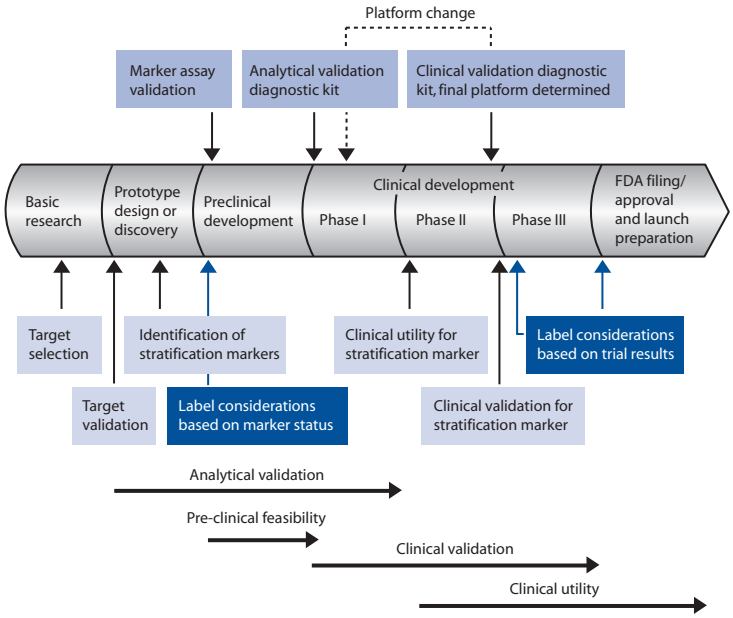
Finally, an advisory committee similar to the ODAC should be created at the FDA to address co-development of targeted therapies and companion diagnostics. The purpose of this committee would be to improve consistency and coordination across oncology programs at the Agency, allowing for more efficient approval of effective therapies.

A potential scientific approach to co-development

The principal scientific challenge in co-development is the absence of a proven effective therapy to demonstrate the utility of the diagnostic test, and the absence of a proven effective test to demonstrate the effectiveness of the treatment. Because our panel has not evaluated all the possible methods for co-development, we do not provide a consensus recommendation here. However, we describe below one such proposal which we hope will stimulate further discussion of alternatives and solutions.

The Critical Path Institute (C-Path) has proposed a co-development process that employs retrospective data and pre-competitive data-sharing to validate biomarkers by testing their ability to predict patient-level variability in disease outcomes (Figure 2).[10] After a disease target is identified, diagnostic development and treatment development should begin at the same time. To ensure that assays will perform consistently in clinical laboratories across the nation, the creation of an independent laboratory – similar to an Underwriters Laboratory or the U.S. Pharmacopeia – is needed to certify the performance of assays. Once the reliability of an assay's performance is established (certified), the assay's clinical value must be determined. Although it is generally assumed that a clinical trial will determine the assay's clinical value, it is not possible to assess this value if the assay does not correspond to a drug with proven efficacy. In this situation, the disease model can be used to conduct simulations of the possible outcomes of clinical trials with a hypothetical drug (e.g., an EGFR antibody) to test the potential reliability of the diagnostic test. This model would also incorporate the test's performance characteristics into the simulation.

Figure 2. A pathway for co-development using a quantitative disease model[11]



If the model predicts that the diagnostic test has a reasonable likelihood of accurately identifying a population responsive to the hypothetical drug, the test could then be deemed "qualified" for use in the development of a new drug with the same general characteristics of the hypothetical drug. If a clinical trial finds that the population identified by the diagnostic test has the desired clinical outcome when treated with the drug predicted to be effective, the data would be submitted as a "strategy" for approval by the FDA. Instead of a drug approval or a diagnostic approval, the strategy approved would assume that the drug would only be recommended for use when the diagnostic test predicts a beneficial response—the realization of truly personalized medicine. In this model, Phase III data could be utilized to seek FDA approval of both a therapy and its companion diagnostic test. Analogous efficiencies in regulatory processes should be considered for biomarkers, including imaging, which also quantify treatment effect and enable further individualization of the regimen in accordance with individual patients' responses to treatment.

References

1 U.S. Food and Drug Administration. Innovation or Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products [online], <http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html> (2004).

2 Phillips K.A., Van Bebber S.L., and Issa A.M. Diagnostics and biomarker development: priming the pipeline. *Nature Rev. Drug Discov.* 5, 463-469 (2006).

3 Medicare Payment Advisory Commission, Clinical Laboratory Services Payment System [online], <http://www.medpac.gov/documents/MedPAC_briefs_Payment_Basics_07_clinical_lab.pdf> (2007).

4 Medicare Payment Advisory Commission, 2007.

5 Ridge, J.R. et al. Reimbursement and coverage challenges associated with bringing emerging molecular diagnostics into the personalized medicine paradigm. *Personalized Medicine.* 3, 345-348 2006.

6 Steinwald B. Medicare Laboratory Payment Policy and Patient Access to Technology. Testimony March 1, 2001 before the Subcommittee on Health and the Subcommittee on Oversight and Investigations, Energy and Commerce Committee, U.S. House of Representatives [online], <http://www7.nationalacademies.org/ocga/testimony/Medicare_Laboratory_Payment_Policy.asp> (2001).

7 U.S. Food and Drug Administration. Drug-Diagnostic Co-Development Concept Paper [online], <http://www.fda.gov/cder/genomics/pharmacoconceptfn.pdf> (2005).

8 U.S. Food and Drug Administration, 2005.

9 Phillips, K.A., Van Bebber, S.L., and Issa, A.M, 2006.

10 Woosley, R. Drug-Diagnostic Co-Development: A New Paradigm [online], <http://c-path.org/Portals/0/Codevelopment.ppt> (2008).

11 Woosley, 2008.

EVENT SUMMARY

Conference on Clinical Cancer Research

Engelberg Center for Health Care Reform at Brookings
Friends of Cancer Research
American Association for Cancer Research
American Society of Clinical Oncology

September 26, 2008

The Engelberg Center for Health Care Reform, in collaboration with Friends of Cancer Research (Friends), the American Association for Cancer Research (AACR) and the American Society of Clinical Oncology (ASCO), hosted a conference to address key challenges facing clinical cancer research.

Four panels of experts came together at the event to discuss recommendations on the specific challenges of data submission standards and evidence requirements, efficacy endpoints, co-development of diagnostics and therapeutics, and a vision for the future of FDA. Issue briefs on the first three topics were developed prior to the meeting, and are available online.

Introduction

In his opening remarks, National Cancer Institute (NCI) Director Dr. John Niederhuber suggested that cancer research should move beyond organ-site based, single agent cancer clinical trials towards a new, molecule- and systems-based era in translational science. In order to achieve the full potential of this new scientific paradigm, researchers will need to adopt drug discovery methods that include mining large datasets and relating genetic information to new targets and pathways. When scientific changes create challenges for the conduct of research, NCI can act as a "safe harbor" where pressing issues can be resolved. The agency has already worked with the CEO Roundtable on Cancer (CRC) to develop common language for clinical trial contracts. The model clauses, which the Justice Department has stated it will not challenge, will reduce the start-up time for clinical trials.

Panel 1 – Data Submission Standards and Evidence Requirements

During this discussion, which was moderated by Dr. Richard Schilsky, associate dean for clinical research at the University of Chicago Medical Center, panelists proposed data collection standards for cancer clinical trials. The panel's recommendations differ depending on whether or not substantial data exist on the safety and efficacy of the treatment being studied. Several panelists noted that streamlined data collection for treatments about which much is already known will make trials faster, less costly, and less burdensome on volunteers. Simplifying data collection may also encourage physician and, in turn, patient participation in clinical trials.

Although the establishment of a standards-based minimum data set will increase the efficiency of trials, potential safety signals must be not overlooked. It was recommended that researchers rely on statistical simulations and post-hoc analyses of previously conducted clinical trials to create a decision tree to guide future data collection, providing sponsors with a mechanism to implement evidence-based data standards. The decision tree, which would take into account factors like whether the treatment was being studied for a new or a supplemental use, would guide sponsors to collect the optimal type and quantity of data on the safety and efficacy of a therapy.

Panel 2 – Improved Insights into Effects of Cancer Therapies

The second panel addressed the issue of auxiliary endpoints, a broad term for endpoints other than overall survival which can potentially be used to help learn about the benefits and risks of cancer therapies in clinical trials. Such endpoints can include progression free survival (PFS), cancer biomarkers, and patient-reported outcomes (e.g., quality of life). As defined by the panel, auxiliary endpoints are not meant to be surrogates for or supplant overall survival, but rather to be evaluated in conjunction with that endpoint.

The panelists focused on PFS because it is now widely used in trials, it has been accepted by the FDA as the basis for approval of some cancer therapies, and it exemplifies many of the challenges in using auxiliary endpoints. PFS, defined by the NCI as the "length of time during and after treatment in which a patient is living with a disease that does not worsen," has been increasingly used to demonstrate efficacy in clinical trials. Compared to overall survival, PFS can often be observed sooner, resulting in less time to show clinical benefit and faster clinical trials. However, since many cancer trials are open-label and progression is typically inferred from radiographic images, that endpoint may be subject to bias.

This concern has prompted debate about whether and to what degree auditing via Blinded Independent Central Review (BICR) is appropriate. As panelists pointed out, BICR is unnecessary in certain situations, but appropriate in others. For example, the FDA has already stated that BICR is unnecessary for double-blinded trials, except when an imbalance in side effects in the treatment arms causes a considerable level of unblinding. In an open-label superiority trial, BICR-based audits of a subset of the PFS endpoints is potentially beneficial; however, the audit would add little or no value if the observed effect size is large enough. Finally, the panelists proposed that an evaluation of PFS at two time points with an audit might reduce bias without much loss of statistical power.

Discussants concluded that researchers should conduct statistical simulations using datasets from previously completed trials to create further standards for what constitutes PFS. This investigation would help develop a more definitive link between PFS and progression, clarify when the use of BICR is appropriate, and yield guidelines on the percentage of cases that should be audited in trials using PFS. If appropriate data sets were compiled, NCI statisticians could complete the necessary analyses within six months, according to Dr. James Doroshow, director of NCI's Division of Cancer Treatment and Diagnosis.

Luncheon Keynote from FDA Commissioner Dr. Andrew von Eschenbach

In his remarks, Dr. von Eschenbach described important changes on the horizon at FDA. He said that the agency is going to upgrade its information technology systems and reassess salary packages and career paths for its most important scientists. The recently launched FDA Fellows program allows professional scientists to learn about the science and policies that underlie the FDA's regulatory decision-making. The FDA has also created the Beyond our Borders initiative, which will result in the opening of FDA offices in Asia, Europe, Latin America and the Middle East.

Panel 3 – Co-Development of Diagnostics and Therapeutics

Panelists discussed steps that should be taken to advance diagnostic test development and regulation, including co-development with cancer therapies. They noted that an unclear, inconsistent regulatory path to clinical acceptance has hampered the development of diagnostic tests with demonstrated clinical utility. Moderator Dr. Daniel Hayes, professor of Internal Medicine at the University of Michigan, noted that FDA approval does not guarantee that an assay is clinically useful, and the "Home Brew" rule means that an assay can be marketed without FDA approval.

Furthermore, public and private payers provide poor reimbursement for many diagnostic tests, slowing investment and innovation in potentially high-value tests.

Several measures that could be taken to create a clear pathway for the development of cancer assays were recommended. Panel members suggested that such a pathway be developed with input from FDA advisory panels, professional societies, and clinical investigators. They also emphasized that approvals for diagnostic tests should be based on demonstrated clinical benefit, but were open to various methods of measuring demonstrated benefit. Finally, panelists recommended that an advisory committee similar to the Oncologic Drugs Advisory Committee (ODAC) be created in order to facilitate the development of a coordinated process for tumor marker clearance. When diagnostic tests are co-developed with therapies, coordination of the regulatory process across FDA centers is especially important.

Dr. Ray Woosley, president and CEO of the Critical Path Institute, presented a method that could facilitate the development and approval of diagnostic and therapeutic companions through biomarker certification and the use of quantitative disease models. For example, if a disease model predicted that a diagnostic test was likely to correctly identify populations responsive to a hypothetical drug, then the FDA could qualify the test for use in developing a therapy that has the same characteristics as the hypothetical drug. Next, clinical trials would be conducted to determine the test's clinical utility in conjunction with the therapy. Ultimately, positive trial results of the diagnostic-therapy combination would be submitted to the FDA for approval as a "strategy" – that is, the drug would only be used when the assay predicts a beneficial response. This is not the only approach, therefore a formal dialogue is needed among professional societies, NCI, FDA, and diagnostics developers to identify and evaluate the range of alternatives.

Panel 4 – Vision for the Future of the FDA

The final panel outlined a vision for the FDA that incorporated the ideas discussed in the previous panel presentations, including panelists' own perspectives on the agency. Each noted the FDA's need for additional resources, and several mentioned that such resources could help the agency hire and retain additional scientific staff whose capabilities reflect the current state of biomedical science.

For example, current and future breakthroughs in cancer research are being built on the disciplines of systems-based biology, genomics, and nanotechnology. Dr. Robert Young, chancellor of Fox Chase Cancer Center, suggested that the FDA should create positions equivalent to Chief Science Officer and Chief Medical Officer and a board of external scientific advisors. Dr. Ellen Sigal, chairperson and founder of Friends of Cancer Research, also backed the idea that the FDA should solicit more input from outside advisors. Dr. Anna Barker, NCI's deputy director, highlighted the NCI-FDA Interagency Oncology Task Force's work on nanotechnology, standards for electronic data submission and biomarker qualification – and called for more collaboration between these influential organizations. Dr. Sigal recommended this task force be expanded to include external representation.

There was also consensus that the FDA's Centers need to work more closely with one another to advance regulatory science in light of rapidly advancing product development science. To ensure that new treatments are developed efficiently and maximally effective, more biomarkers and auxiliary endpoints need to be validated through consistent evaluation across public and private research programs. Dr. David Kessler, former FDA Commissioner, pointed out the importance of developing endpoints that can be measured quantitatively, rather than subjectively. David Epstein, president and CEO of Novartis Oncology, added that the issues linked to auxiliary endpoints need to be addressed soon in order to provide support for conducting clinical trials of treatments for rare diseases and better-defined subgroups of patients.

Dr. Kessler outlined the process by which the next FDA Commissioner could accelerate innovation in cancer therapies. He suggested the commissioner make a public commitment to finding more treatments that work for patients with cancer. To win the necessary public support for this effort, the FDA will need to concentrate on treatments for the most aggressive cancers. Finally, the FDA must develop a drug development roadmap for sponsors and regulators to follow. For cancer therapies, the roadmap should ensure that trials "get the science right," that accelerated approvals are granted where appropriate, and that models of success are identified for others to follow. A similar roadmap, Kessler said, was crucial to the rapid testing and approval of AIDS drugs in the early 1990s.

Next Steps

Dr. Mark McClellan, director of the Engelberg Center for Health Care Reform, summarized the conference discussion with a series of recommended next steps for making clinical cancer research more effective and efficient:

Data Submission Standards and Evidence Requirements

1. Building on the proposals from the first panel, efforts should be made to ensure that data are collected consistently and accurately, and that the data that are gathered are actually useful. To accomplish this goal, researchers should use data from completed trials to develop a decision tree that would help sponsors design optimal data collection protocols.

Improved Insights into Effects of Cancer Therapies

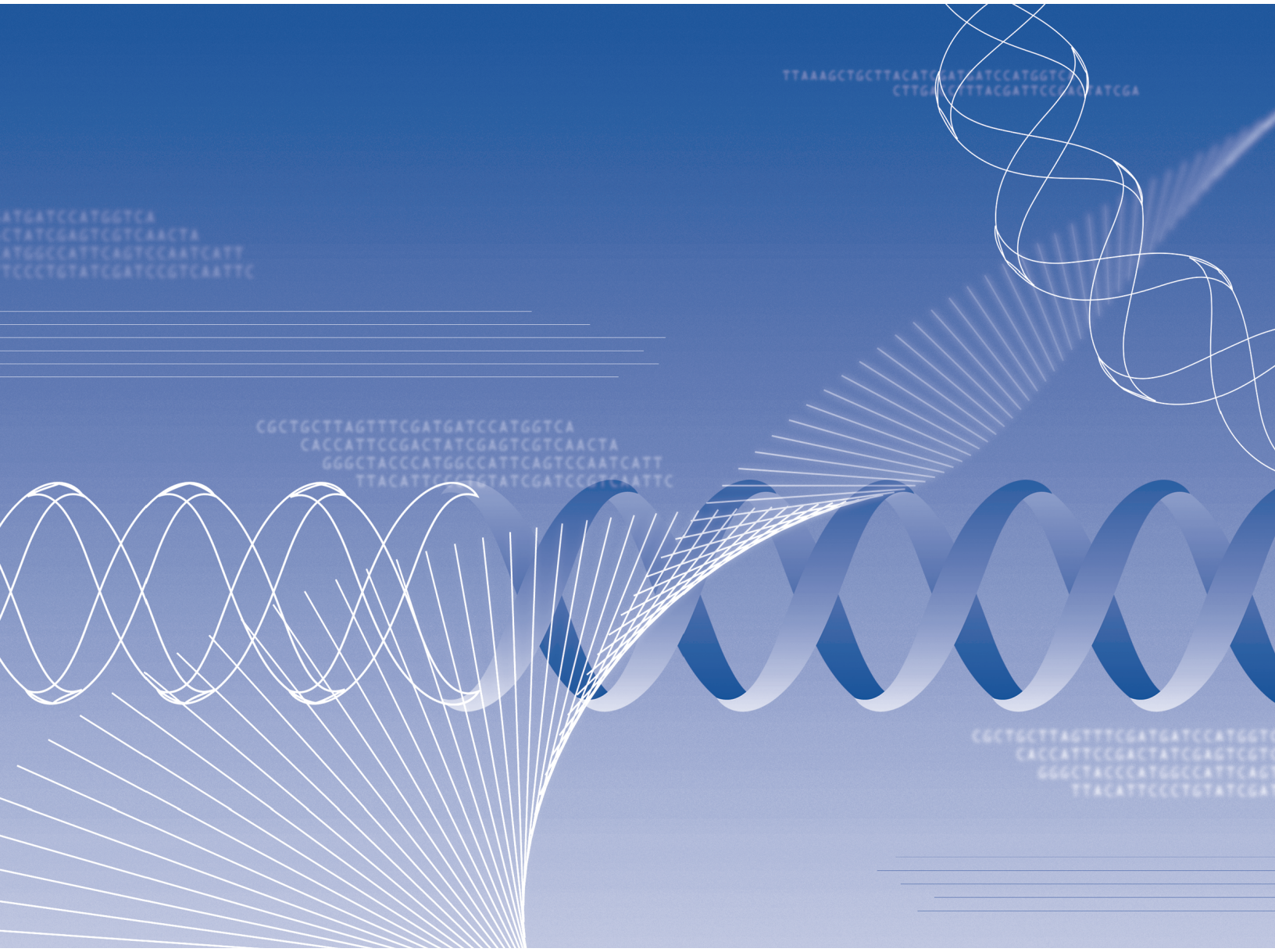
2. Researchers should perform simulations on data from a series of completed PFS trials involving BICR, to quantify the potential for bias. This analysis should yield evidence-based recommendations for when BICR will meaningfully reduce bias, and on what percentage of cases should be reviewed. The approach to improving auditing procedures for PFS suggests a method for increasing the consistency and quality of data produced in clinical trials utilizing other auxiliary endpoints: researchers should strive to make auxiliary endpoints standardized and quantifiable. Studies, analogous to those needed for PFS, should be carried out to validate other auxiliary endpoints.

Co-Development of Diagnostics and Therapeutics

3. The FDA, relevant FDA advisory panels, and professional societies should begin to discuss ways to clarify the development pathway for cancer diagnostics in general and those co-developed with targeted therapies in particular. This discussion should address methods for generating reliable evidence to show that diagnostics influence treatment decisions and impact health outcomes. Such evidence can influence reimbursement of diagnostic tests.

4. An ODAC-like committee could help improve consistency and intra-FDA coordination in tumor marker clearance and approval. An initial item of business for this committee and FDA staff should be to define a clearer pathway initial validation, further development, and approval for diagnostic tests that seem to have demonstrated clinical utility.

NOTES:



ENGELBERG CENTER for
Health Care Reform
at BROOKINGS

AACR CENTENNIAL 1907 - 2007
American Association for Cancer Research

ASCO®
American Society of Clinical Oncology

FRIENDS of CANCER RESEARCH