SPORE Workshop on Statistical Design for Translational Research

July 16 – 9 a.m.-12 noon

Co-chairs: Steven Piantadosi, Johns Hopkins University
Donald Berry, M.D. Anderson Cancer Center
George Klee, Mayo Clinic
Garnet Anderson, Fred Hutchinson Cancer Research Center

Introduction

Jorge Gomez, M.D., Ph.D., Chief, Organ Systems Branch, NCI

Dr. Gomez thanked the co-chairs. This workshop is a result of issues raised about how we are sponsoring some clinical studies in the SPORE program: clinical trials, biomarker, and population basis study prototypes. It is a special problem, because the SPORE program is multi-disciplinary. People from many areas bring lab discoveries into the clinic and work with various types of biomarkers. This has been done relatively efficiently, but the methods we use do not allow us to have concerted efforts or receive input. Some studies are well-designed, but some are not. Many biological endpoints are used in clinical studies and interventions, with no good methodology. Biomarkers have proliferated, but many data sets are not even reproducible results. Sometimes 20 or 30 biomarkers are used in a study that lacks the power to test even one. We should to be more careful how we use our resources, and we need advice from experts.

An additional issue is the lack of reference materials. Some reagents are developed in the lab, and many papers are published, but many assays have not been developed carefully and may not test what they purport to test.

We want to bring some level of awareness, and to develop, revise, and publish the methods and methodologies. Some complain that statisticians are a different breed and speak a different language. But we need to have appropriate statistical design and want to advise biologists and medical oncologists so that they have appropriate statistical input and their methods test what they are supposed to be testing. The goal is to do better science. We are forming a group to go over these issues, to help us educate and select how these studies should be done. Dr. Piantadosi and Dr. Berry have been asked to help in the area of clinical trials, and Dr. Klee and Dr. Anderson in the area of biomarkers.

There are other issues regarding correlative science, such as differentiating pre-validation from validation. I want this group to plan for a larger workshop we will sponsor in the fall. There will be a series of workshops to help the community develop better methods of validation for clinical studies with an impact on human cancers. Planning for the fall meetings will include identifying other experts to include in these meetings. There could be a subsequent meeting, if necessary, next year. We are committed to health, and need the help of this group to improve how clinical science is being done.

Some New Statistical Design Perspectives for Effective Translational Research

Steven Piantadosi, M.D., Ph.D., Johns Hopkins

The perspective presented here may be new to some of my colleagues. First, let’s enumerate the obstacles to effective translational research:

- Use of a stereotypical developmental paradigm that has been around since the 1960s; this is applied to a “pre-developmental” process, so . . .
- Terminology and jargon inhibit description and study design based on first principles;
- Lack of methods development, heterogeneity in translational research, and weak institutional support.

Generating evidence in translational trials should involve arguing points from first principles. *Passive observation* may yield some lucky accidents. *Active observation* (experiments) can employ theory in their design or in interpretation. “Theory” can be defined as a body of established, consistent knowledge, or a speculation based on inconclusive evidence. If we subordinate theory to evidence, we have structure but not its full power. We want true experiments in which theory is used *a priori*.

First principle of translational research: *Established biological knowledge, well-integrated into a good study design, minimizes the size required for an experiment.* A “true” experiment, built on theory, can be smaller than an empirical experiment based on the same ideas.
Identical evidence can lead to different conclusions, depending on *a priori* theory. For instance, a person’s assessment of obtaining four “heads” in four coin flips will change if they know there is a 50:50 chance that a two-headed coin is being used.

Definition of a *Translational trial* (from Piantadosi, S. (2005) *Clinical Trials* 2:182-192.):
A clinical trial where the primary outcome: 1) is a biological measurement (target) derived from a well established paradigm of disease, and 2) represents an irrefutable signal regarding the intended therapeutic effect. The design and purposes of the trial are to guide further experiments in the laboratory or clinic, inform treatment modifications, and validate the target, but not necessarily to provide reliable evidence regarding clinical outcomes.

Notice that the purpose of translational trials is *not* to assess clinical outcomes, but to *inform subsequent experimentation by reducing uncertainty*. The treatment may go back into the lab for modification, and the process may be repeated. A single experimental design class may not be able to satisfy the diverse objectives of translational research.

The first human trial uses biological outcome(s) known with certainty relatively soon following treatment. The design requires an explicit definition of failure, and *either success or failure will be informative* for designing future experiments. These studies are often conducted by small groups with limited resources. We look for large effects on the target, although these may not be large clinical effects. Dr. Piantadosi observed that the inferences people draw from investigational studies are not driven by the formal hypothesis tests. The decision rules and even the next experiment(s) are protocol specified.

There are no major statistical issues with nesting a translational study within a developmental clinical trial because they are almost always smaller than the developmental trial; their outcomes tend to be simple and measured early; and the analysis is simple. Other potential issues are generally not directly germane to the methodology or the design.

Translational studies are *not* Phase I, dose-ranging, or dose-finding. Nor are they "pilot" or "proof of principle" trials, definitive tests of any clinical outcome, or sized, powered, or precise enough for full error control. These terms are used more to defray criticism of the rigor than because the terms are well-defined. The studies reduce uncertainty enough to guide the next step.

<u>Second principle of translational research</u>: *The primary goal of any translational clinical trial is to acquire sufficient information to guide the subsequent experiment(s).*
High precision and size are not necessarily fundamental design requirements, and may not be appropriate in the clinical setting. This is a departure from the usual statistical paradigm for clinical trials, but not from the use of statistical reasoning.
An example of such guidance: Data from engineering tests that preceded the Challenger disaster might have guided subsequent experiments on shuttle O-ring behavior at lower temperatures. A similar type of inference can often be drawn in translational studies with fairly murky evidence.

<u>Third principle of translational research</u>: *When maximal uncertainty exists, typical for the . . . treatment effect before human trials are done, a relatively small study coupled with established biology is a powerful device for gaining information.*
This situation is typical in pre-development, when you imagine the effects of a putative new therapy. Use entropy to measure information gained, relative to maximal uncertainty, and to suggest an appropriate size for a translational trial.

A typical statistical approach to inference is based on estimating a parameter from a model that contains a deterministic and a probabilistic part. *Parameter uncertainty* is the imprecision in the estimate of a model parameter, and determining model parameters with high precision requires a relatively large amount of data. This is true for both Bayesian and frequentist approaches. A second level of uncertainty, *outcome uncertainty*, is more relevant in clinical research and depends on the probability model being correct. With a dichotomy, the probability (p) of a given outcome is 0.5; but there is still maximal uncertainty regarding the effect of the treatment on any given subject. Outcome uncertainty may be more important than parameter uncertainty in pre-development, and can be partially resolved in pre-developmental studies even without resolving what "p" is.

Shannon (1948) identified *entropy* as the only appropriate measure of overall uncertainty. Entropy is primarily sensitive to outcome uncertainty and only weakly dependent on parameter uncertainty, so it may parallel the clinical and basic scientists' instinctive interpretation of clinical studies. Entropy can be used to assess the variance and bias in sampling distributions as a function of sample size, and see if there comes a point where a larger experiment yields diminishing returns. In one example, using the Gibbs distribution, as "n" increases, the true sampling distribution yields more information until reaching n = ~35. After this, a larger "n" yields no additional information (i.e. increase in entropy). It seems paradoxical that entropy increases as sample size increases. However, the entropy measure is biased in small samples, which do not capture all the uncertainty in the system. As sample size increases, the entropy approaches a limit – the bias going to zero. Variance also decreases as sample size increases.

Looking at a plot of entropy bias vs. variance, it matters very little what distribution is taken as the “true” state of nature. In terms of overall uncertainty (i.e. entropy), a sample size can be chosen that removes much of the variation and bias in information gained. Even a sample size as small as 15-20 subjects could be used to categorize therapies as very strong, some benefit, no change, somewhat negative, or going the wrong way.

Investigators should be aware that translational studies nearly always yield only weak empirical evidence, which is more likely to mislead us than strong evidence is. The biological paradigm, including target validity, can also be in error. These risks are worsened in the presence of a weak study methodology, so we need to minimize these risks.

Key points:

- The old cytotoxic drug development paradigms do not fit the needs of translational trials
- The essential purpose in a translational trial is to guide subsequent experiments
- Biological knowledge augmented by even a small well-designed translational study can reduce overall uncertainty and steer the next experiments
- Partial control of bias and variance of entropy can be used to select an appropriate sample size for a translational trial.

Discussion

Q: In looking at pictures of a normal distribution, rather than entropy, by the time I got to n = 35, I wouldn’t see much change in the pictures. What is the advantage of working with entropy?
Piantadosi: I haven’t done that comparison. However, entropy is a function that is sensitive to the *uncertainty* implied by the *whole distribution*, not just one parameter. There are equal entropy states that are very different in terms of p values. For instance p = 0.9 or p = 0.1 have the same uncertainty, but look very different. If you were to axiomize uncertainty, you would end up with the entropy function. I’m just suggesting that this might be an appropriate tool.

Q: You mention what translational trials are not. Aren’t these reviewed in the SPORE program?
Piantadosi: In the SPORE program, the trials are regarded as some of the things that I say they are not. When people view translational studies the wrong way and use the wrong terminology, they limit their ability to see what is inappropriate design. On a recombinant DNA advisory committee, I’ve noticed that a lot of gene transfer studies being done are, in fact, translational studies. They are almost universally referred to as Phase I studies – people get a Phase I study from the local oncologist and plug their gene transfer design into that model. But there is no dose question; and the safety concerns are firmer than implied in a dose-finding experiment. It would be better to consider this a pre-development study: operating under a certain paradigm for a certain disease and how a system works to see if I can learn more about it.

Q: Have you communicated to IRBs?
Piantadosi: I am asked to comment on study design, and we’ve had a year-long discussion. We have five IRBs, and they wanted to come to coherent policy on pilot studies. I don’t even believe in pilot studies – the terminology means something very different than the way they are used. Also, the primary audience for the advisory study is the IRB, and the advice isn’t seen by many others. I can only hope the IRB will understand and ask the investigators to make the appropriate changes in the protocol.

Innovative Designs for Translational Research: Thinking outside the box

Donald A. Berry, Ph.D., M.D. Anderson Cancer Center

Thinking outside the box is dangerous: You may find yourself in quicksand, and think that others who think outside the box are thinking garbage. The goal is to learn the bad things we are doing, and correct them.

I want to give examples of things we are doing differently at M.D. Anderson. Many adaptive design issues are being addressed in pharmaceutical companies using Bayesian methods to build seamless phase trials that do something revolutionary in medical research. We look at the data, knowing that some things can lead us astray – as in life, if we are unsure where it is going to go, we go lightly.

The way we design trials based on power. Type I error is handy, but it could be better. We should ask what we are trying to do. If this is a pediatric cancer with 200 patients per year in the US, as opposed to cancers that have 200,000 patients, the trial size should take into consideration what we are trying to do – which is to treat as many patients as effectively as possible. There should be different standards in pediatric cancers. Predictive probabilities are essential in monitoring trials and are a critical aspect of experimental design. "We *must* ask where we are and whither we are tending." – Abraham Lincoln. When we ask where we are going, and the probability, it will lead us to something statistically significant.

For example, you may be interested in whether T or C is a better treatment. If 13 of 17 paired observations are successes for treatment T, you can still ask what would happen if you doubled the size of the trial. What is the probability of getting statistical significance at the end? Two histograms show best-fitting binomial vs. predictive probabilities: The graph of predictive probabilities has greater variability, and 88% probability of statistical significance; you would need at least 10 (T results) out of the next 17 to have significance. The binomial histogram has less variability, and 96% probability of statistical significance. This is a huge difference.

Another example is a trial for Herceptin in HER2+ breast cancer combined with two other therapies (FEC and Taxol). The trials involved 164 patients equally randomized between trials with Herceptin and trials without Herceptin. After the first 34 patients, the data monitoring committee showed that 67% of the 18 patients receiving Herceptin had responded vs. 25% of the 16 patients without Herceptin. The Bayesian probability of the outcome will still be positive (95%) after 164 patients. But we asked: Where are we going? How long will it take us to get there? What are the consequences if we continue? If we don't continue? Given the slow accrual to this experiment, we should present the information so people can use it. Some people complained about the small trial size, but it was an important piece, even using only 34 patients.

Predicting trial results requires that you:

- Simulate
- Model uncertainty
- Incorporate information (Bayesian-wise) on various outcomes
- Model relationships among early and late endpoints
- Consider alternative designs, including extraim analyses.

The true (predictive) power of a test is usually a good deal less than the traditional power.

A trial design that demonstrates adaptive randomization was originally proposed as a three-armed trial of Troxacitabine in acute myeloid leukemia (AML). Seventy-five patients were to be randomized between treatment groups with Idarubicin and Ara-C; Troxacitabine and Idarubicin; and Troxacitabine and Ara-C (see Giles et al. *JCO*, 2003). Dr. Berry suggested looking at the data already available, but Dr. Giles wanted to randomize patients equally until one of the arms dropped, and then drop it completely. He would also drop alternative treatments if standard therapy was doing well. After 34 patients, the TI group was doing so poorly that it was dropped. IA had a 56% remission rate; the TA arm had only 27% remission rate; the TI arm had no complete remissions. The journal *Blood* said it was impossible to do anything with such a small number of patients. Another journal accepted it, saying the drug (T) was a dud with this particular disease, but the trial was wonderful.

Moving on to a phaseI/II cancer trial: The investigator wanted to test two drugs and see how one drug worked with another drug. Should they be used concurrently or sequentially? He also wanted to do a Phase II trial, but not a Phase I. Dr. Berry suggested doing Phase I and Phase II at the same time to address the dose of D, dose of A, and whether to use them concurrently or sequentially. Two grids show the combinations of four dosages of A and four dosages of D, administered either concurrently or A followed by D – a total of 32 possible assignments. The study started with the lowest doses of A and D (concurrently or sequentially), then moved to the next higher dose for A or for D. The PI's earlier data suggested that a lower dose was better, so he focused on lower doses and eventually found the dose that was inadmissible because of toxicity. At any given time, the study design had the possibility of:

- Expanding or contracting admissible doses depending on toxicity
- Randomizing to admissible doses, adapting to tumor response
- Expanding the dose-range while still focusing on lower doses.

Another topic to consider is modeling early endpoints using longitudinal markers. An example of this would be CA125 in ovarian cancer. The question is how well this measures effective treatment. It is not accepted by the FDA as a surrogate marker of treatment benefit, but Dr. Berry wanted to use it to adapt a trial to what is going on with a patient. One pet peeve with clinical trial development is doing Phase II on one endpoint and Phase III on another endpoint – we need to use the same endpoints, or have a primary endpoint and use all the information as well as the relationships between them. For instance, if the endpoint is survival, we need to look at progressions, performance status, etc. Use the available data from the trial and outside it to model the relationship over time with survival, depending on therapy. We should also:

- Do predictive distributions
- Use covariates
- Possibly seamless phases II and III.

Several graphs modeled CA125 data and predictive distributions of survival for one patient (of many). CA-125 dropped precipitously early in treatment, then started to come back after ~500 days. What did the low period mean? What did the return mean? Other patients showed a similar curve for CA-125 levels. We have some patients barely into the study and some who have died. As we move through time and the CA-125 levels drop, we make new predictions. The incremental survival time is less (patient lives longer, but the survival after assessment is less).

We compare treatment A and treatment B, keep track of the number of times treatment A is better than treatment B, etc. We can extend accrual, curtail accrual, modify, add, or drop treatment arms. There is a lot to look at.

Discussion

Q: What is the political process of convincing other investigators and groups even on simple things like research design – especially at institutions without a group as influential as yours?
Berry: Our IRB obviously buys into this. If you walk down the hall at M.D. Anderson and ask people what kind of statistics they do, they'll say, "We're Bayesians." The IRB is so interested that if an outside company comes in, they are told they have to do interim analyses. There is also a new dawn at the FDA, with the Critical Path initiative. A contact at the FDA said recently that they have to be adaptive, and underlined things that have to be pushed by the FDA. Their two focuses are using biomarkers and innovative trial design.
Last week was a hot week for adaptive designs: An article in the *Wall Street Journal* on Monday (Scott Gottlieb's presentation); *JAMA* is writing an article about adaptive designs; *Nature* has an article about it – and quotes von Eschenbach about the importance of this type of thing. There was also a Japanese pharmaceutical newsletter with a heading that said: "the coming Bayesian tsunami in clinical development" – all about adaptive design. Half the pharmaceutical companies are doing this type of thing. Some of the things I talked about are company-sponsored designs.

Q: Could you comment on the amount of time required of statisticians compared to using a more standard design? Also what proportion of the trials at M.D. Anderson use these types of designs?
Berry: We did a survey of our trials at M.D. Anderson, and it's about 20 percent and increasing. About half of our studies are pharmaceutical studies; they come in with a design and we review it. But many investigator-initiated studies take this approach (over 200 in the past few years). As to the first part of your question, the first study took about a year-and-a-half to develop vs. people who "need this next week." For statisticians who haven't done this sort of thing before, there is a learning curve, but it is fun and makes life worthwhile. The logistics are not small. The biggest thing is to get people who can run the study. You need data flow and a logistical set-up that will do the grinding. Someone said to me, "So there is a negative side of doing it." But the positive in terms of treating patients effectively, having a more accurate study and fewer patients used . . . the trade-off is absolutely clear. And the process of setting up an adaptive design is so revealing. The team has to say: "If these data were to accrue, this is what I would do." They understand the need for that. The data-monitoring committee will be monitoring to be sure the design is followed, so people must say what they would do in following the data. Even if you were to go through that process and then use the standard design off the shelf, it would still be worthwhile.

Q: Can you comment on the software availability?
Berry: Cytel has software; Berry Consultants (the main person is my son). You can go to M.D. Anderson on the Web site. You can't get it from SAS, but the idea is to make it all available eventually.

Assay Performance Specifications and Assay Validation Strategies

George G. Klee, M.D, Ph.D., Mayo Clinic

Goals of clinical tests: Assays can be designed to meet certain performance specifications, but only the interplay of performance specifications and validation can confirm if an assay meets the specifications. We need input from people running the trials about how an assay will be used. Several medical questions are addressed when using research or clinical assays:

- *Screening*: Should this patient be evaluated further?
- *Diagnosis*: Does this patient have a specific disease? There are very few diagnostic tests, but we still need criteria for looking at those.
- *Selection*: Which therapy will work best for this patient?
- *Monitoring*: Is this patient improving?

Medical *diagnosis* separates test results into discrete disease categories. A test value from a single patient must be compared with the distribution of values from patients with confirmed diagnoses. Reference data should be collected using the same assay under the same conditions and with consistent performance specifications and standards. *Monitoring* tracks a patient's multiple test values over time. Test values are compared with prior values and/or therapeutic ranges which should have been collected with the same assay, conditions, and calibrations.

The CLIA standards for how to validate a test were published in the Federal Register (CLIA Std. 793.123). Before reporting test results (and perhaps before doing clinical studies) a lab must verify or establish . . . the performance specifications for the following characteristics: accuracy, precision, analytical sensitivity, analytical specificity, linearity, reference range(s) and any other characteristic required for test performance.

There must be a minimal analytic validation for new test introduction:

- Complete written procedure guide – including performance validation data.
- Precision check – a minimum of 20 data points is recommended for each component; the use of pooled samples, several concentrations, and inter- and intra-assay precision are important.
- Linearity checks – if multiple specimen types (urine, serum, CSF, etc.) will be assayed, a linearity study should be carried out for each specimen type, with at least four concentration levels. The dilution agent can influence the test results; also, an assay validated for one specimen type may not have been validated for another application.
- Recovery checks – if you put something into a specimen, you should be able to get it back out. Specimens with low values can be spiked with known amounts of the standard and analyzed.
- Interference studies – studies with compounds that may be interfering or cross-reacting substances should be documented.
- Carry over studies – whether sample results are affected by preceding or succeeding samples.
- Comparison of methods – at least 100 comparisons distributed over the assay range is recommended. It would be useful if clinical trial data could be linked back to the research data, before putting an assay into clinical practice.
- Normal value studies – reference values.
- Clinical sensitivity and clinical specificity.

Traceable standards is another pet issue. The FDA does not require traceable standards for assays in the US, but the European Union does require this. European Directive 98/79/EC for in vitro

Diagnostic Medical Devices requires that: *The traceability of values assigned to calibrators and/or control materials must be assured through available reference measurement procedures and/or available reference materials of a higher order.*
The FDA is looking into this but only requires reproducibility and correlation. Traceability is a multiple-step process that involves a sequence of measurement procedures and calibrators. An International Standards (SI)-unit is defined, and the primary standard is posted and is available through the National Institute of Standards and Technology (NIST). With proteins, for example, it would be helpful to get some consensus as we move forward.
A Joint Committee for Traceability in Laboratory Medicine (JCTLM) was established in June, 2002, to support comparability, reliability and equivalence of measurement results in laboratory medicine. Their site posts reference materials and methods.

How do we know if an assay is good enough? We want precision and accuracy. What are the *tolerance limits for precision*? *Analytic precision* is limited by the biological variation within a person: If analytical standard deviation (SD) is less then one-quarter of the biological SD, the total SD increases by only 3%. The tolerance limits for *analytic bias* (accuracy) are less forgiving. Bias directly affects test values, and small analytic changes can produce major shifts in the frequency distributions of clinical test values. Shifting a normal curve by one SD unit has a large impact on the number of patients crossing a given decision point (i.e. positive or negative test results). For example, a +10% bias in a cholesterol assay increases the number of positives (above 200 mg/dL) by 27.8%.

Precision adds by the square root of the variances, while *accuracy* adds linearly.

- We can try to standardize the pre-analytic error components that influence *precision*: pre-analytic, instrument, reagent, measurement, biological.
- We can also try to standardize the pre-analytic error components that contribute to *bias (accuracy)*: collection, instability, calibration, pipetting, data reduction.

Finally, there are problems with the use of multi-parameter algorithms for test interpretation. We won't find a single biomarker for all diagnoses and patients at risk. New systems can measure 10 or 20 parameters at the same time, but statistical noise increases exponentially with the number of parameters. It is difficult to statistically classify patients without other clinical information and test results. As an example: Determining the risk of a neural tube defect when the gestational date may be off by +10 days (risk of 1 in 1600) or -10 days (risk of 1 in 75) illustrates this point. In addition, specimens collected in a research environment may be very different from those collected in a clinical setting. "Learning" systems with real-time feedback of medical outcomes may help.

In summary:

- Assay performance characteristics should be linked to clinical use of diagnosis and/or monitoring;
- Assay validation requires assessment of precision, linearity, recovery, and interference, with assigned performance limits;
- Assay traceability and harmonization are important;
- Multi-parameter test interpretation algorithms require extensive ongoing validation and robust assay performance characteristics.

Discussion

Q: You gave the example of a multi-parameter test. But in the tests we use, in a scenario with many genes, the variability of one may be compensated for by the variability of another, so the change won't be as strong . . .
Klee: I would disagree. In our multi-variant Gaussian model, there are about 10 parameters. When we saw a change in one set point, it was enough to interact with others. We don't take advantage of the linkage between genes. We should be able to say – if these factors are not consistent (we expect them to go the same direction) we need to look at the situation more closely. An error coming in from one can throw the whole model off.

Q: For those in translational research, can you comment on the challenges where we don't have a pure sample of the candidate biomarker or information from an earlier experimental stage?
Klee: I work in biomarker development. They bring me over-expressed genes and say, "We know the protein, now set up an assay for it." We can make recombinant proteins now. We can set up a vector to make the recombinant protein in a relatively pure form, post that, and assign some value to it. Also, with peptides, we can synthesize a peptide in pure form and assign a value to it to use as a reference.

Clinical Validation and Performance Assessment

Garnet Anderson, Ph.D., Fred Hutchinson Cancer Research Consortium

I want to describe some frustrations and ad hoc approaches that have been a problem in our biomarker research, and generate ideas for moving forward. Biomarkers have potential roles in cancer risk, early detection, diagnosis, prognosis, and prediction; other roles may include their use in therapeutic targeting, as intermediate endpoints, and surrogate endpoints.

One source of frustration is ambiguous terminology, typical of multi-disciplinary research. Talking across disciplines often tends to get confused because of differing uses of terms. The following terms (with suggested definitions) are often a source of confusion:

- Sensitivity – fraction of diseased individuals who test positive
- Specificity – fraction of non-diseased individuals who test negative
- Controls – individuals without designated outcome that serve as a comparison group
- Validation – a study to confirm initial findings using a new population sample
- Reproducibility – extent to which the findings apply to a new sample of individuals
- Prognostic marker – helps identify patients at different degrees of risk (without regard to treatment)
- Predictive marker – helps to identify patients that may respond to a particular therapy.

In the longer term, we want to accumulate the terms and multiple definitions, and agree on them.

Dr. Anderson presented a "motivating example" of early detection of ovarian cancer. Survival depends strongly on the stage at diagnosis. Women with the local stage can be cured by surgery alone, so early detection and treatment offer great promise for reducing morbidity and mortality. The natural history of cancer progression goes through initiation, detectable, symptomatic, and death. "Observed survival" occurs between the last two stages. With routine screening, cancer can be detected earlier, and when treatment is combined with screening, death can be deferred.

A diagram of the biomarker research pipeline (ca. 1998) shows 1000s of potential biomarkers entering the system by *discovery*, 100s of candidate biomarkers moving on to *translation*, 10s of them are validated and go through clinical testing to *assessment*, yielding a biomarker panel which, if *efficacious*, leads to a screening program. Margaret Pepe, at the Fred Hutchinson Cancer Research Center, developed a five-stage "phases of biomarker validation":

- Phase 1 – Preclinical exploratory: promising directions identified
- Phase 2 – Clinical assay and validation: clinical assay detects established disease
- Phase 3 – Retrospective longitudinal: biomarker detects disease early before it becomes clinical, and a "screen positive" rule is defined
- Phase 4 – Prospective screening: extent and characteristics of disease detected by the test and the false referral rate are identified
- Phase 5 – Cancer control: impact of screening on reducing burden of disease on population is quantified.

We have a way of thinking through the process of identifying a biomarker and getting it into clinical trials, but the pipeline is stopped up. The 2006 "biomarker research pipeline" has tens of thousands of potential biomarkers entering the *discovery* phase, 100s of markers entering *translation*, and a trickle going on to assessment and efficacy. What is the plug in the pipeline?

- Large number of candidates identified in phase I
- Lack of clinically relevant, commercially available assays for phase II
- Cost and time of developing assays for use in validation studies
- Design limitations of "validation" studies
- Funding/intellectual property issues.

The challenges of finding or developing assays for clinical validation studies include: discovery methods are often not suitable for validation studies; existing assays have limitations; developing a new assay may take a year or more – and it still may not work; there is no "gold standard" assay for comparison, nor a reference material against which to measure acceptable performance for a new assay of a novel marker.

Dr. Anderson described the use of CA-125 as a biomarker for breast cancer. The original test required 300 microliters of specimen, so she found a more time-consuming method that uses only 15 microliters. However, when broken down by case and control status, the correlation (*r*) in the screening controls falls to 0.64. That low a correlation *may* be acceptable, but she has no reference. The correlation estimate is affected by sample selection; there is no consensus on the threshold for accepting performance; and existing clinical assays have their own measurement properties – they may not be sensitive or specific for the targeted molecule; or may be too expensive, too time-consuming, or specimen greedy.

They have tried using the coefficient of variation (CV) to assess reproducibility of assays. This looked promising for CA125-RDI, but poor for HE4, and there is no reference material to work with. To address this problem, they created specimen pools of cases (N=50) or controls (N=9) to permit ongoing use for prospective assessment as laboratory controls, and to create a standardized curve for almost any biomarker of interest. The HE4 BioPlex assay showed a lot of variability for the various dilutions. If you leave out all the specimen preparation steps, assay variability is about 5%, but with all the preparation steps, variability is about 26%. Many assays are multi-step processes, and the variation (in specimen preparation, reagents, batches, and technicians) affects the studies. The only way to deal with this is boring, unpublishable studies – and these have been very enlightening.

Published coefficients of variation (CV) rarely indicate the sources of variability. They probably represent the assay performance alone, and so underestimate the variability we should expect. Although variability may be random, it can dilute the discriminatory performance of a biomarker assay. Biomarker performance is confounded with the assay performance – it is hard to know when to give up on a biomarker because of the assay.

Evaluating an assay for clinical validation studies should:

- Use correlations with a standard assay, if available, in a representative sample
- Have low overall CV
- Have good discriminatory power.

Failure to validate/discriminate may be due to general *assay* performance issues or lack of sensitivity/specificity to the targeted molecule. It may also be due to differences between discovery and validation study *specimen* collection, processing, or storage protocols (e.g., specimens from repositories at multiple sites; use of "least precious" most convenient samples available). Failure to validate/discriminate may be due to lack of *expression of the biomarker* in a new specimen type, or expression from other non-tumor tissues. Finally, *statistical or design issues* may confound validation studies.

Published validation studies can be difficult to evaluate because of differences in study populations, over-optimistic analyses or emphasis on statistical significance, poor reporting, or publication bias.

If we were starting over, our advice would be:

- Use the most clinically relevant specimens for discovery work
- If different specimen types are used for discovery and validation, do validation in the same patients' samples first
- Use the best statistical/epidemiological design feasible at each stage
- Develop/maintain/use standardized specimen sets to facilitate comparisons across markers
- Consider the use of specimen pools for discovery, for early biomarker validation, and for QA/normalization of novel markers – you need both methodologic efforts and empirical work to better understand the strengths and weaknesses of using pooled specimens
- Do validation studies that are publishable regardless of the results (publish all results)
- Use of biomarkers alone provides no benefit to subjects (biomarkers are not the ultimate goal).

The goal is not the biomarkers, but to improve survival by developing a screening program based on serum biomarkers that can detect ovarian cancer at its most treatable stage. This can only be evaluated in a full-scale randomized trial.

Finally, consider the use of biomarkers as surrogate outcomes. The motivation is to reduce trial size and duration. Surrogate outcomes can only be established for a very specific treatment and outcome and evaluated in a full-scale trial with all variables available. The use of biomarkers as surrogate outcomes is erratic and limits our ability to look at overall risks and benefits of an intervention. As an intermediate outcome, the approach can help establish biologic plausibility.

Discussion

Q: Part of the blockage in the pipeline is education – not necessarily in statistics, but for clinical colleagues in the area of metrology. How can we, as a statistical community, go forward with a general approach?

Anderson: I have the luxury of being in an environment rife with statisticians. Dr. Gomez has asked us, as statisticians, to try to define criteria and educate our colleagues in this process, multiple agendas going toward the same target. Local institutions can try training opportunities.

Q: People are overwhelmed with the Phase process – how substances get pushed through the process.

Anderson: The Phase process is great if you focus on early detection, but we can shortcut the approval process with a different use of biomarkers.

Q: There have been changes in the natural history of the disease over time. Screening has its impact on early detection, and biomarkers may not be as effective now as they would have been five or ten years ago. How will we plan for designing appropriate models that will take into consideration the changing history of cancer?

Anderson: I work in ovarian cancer, and there is no screening in this population. With breast and prostate, incorporating the broad screening into the process is a challenge.

Berry: As statisticians, we try to look for the forest, but too frequently get burdened with the details. I am concerned about false positives and the diagnosis. The definition of a healthy person is someone who has not had a complete workup! If we had a biomarker that told us when that first cell becomes cancerous . . . What if a woman tests positive for breast cancer – do we recommend a mastectomy? The most important design consideration is: Why are we doing this?

Anderson: I agree. We shouldn't be doing this unless we are committed to full validation in follow-up trials. The community has hyped the value of early detection, but this has tremendous costs. If we start screening for all these diseases, the cost will be monumental.

Q: One thing in managing early disease is "expectant management." We can use pathologic and/or biomarker criteria to evaluate diseases over time. In prostate cancer, they are selecting patients with a low tumor burden and assessing them over time.

Berry: The problem is in understanding the early disease, and we don't do that.

Q: That is studying the mechanism of early disease, but we can apply biomarkers to move the process forward. Over-treatment is already happening. It's not due to biomarkers, but to early diagnosis. Part of what's happening at this meeting is trying not to make the problem worse.

Q: The problem in the world of ovarian cancer early detection is the "stopped-up" slide. How do we know which of the thousands of potential candidates will measure accurately in serum? Of the thousands of candidates, which ones do you take forward and invest a year or more developing an assay? I'm not sure there is a good answer.

Anderson: We're going to "charge" for it – bring statistical procedures and whatever biology we need to the table. We have two markers in validation studies now, and others in the pipeline.

Q: In terms of the SPOREs, how they are organized and the budgetary part, it's always cheaper and quicker to come up with a new biomarker. Doing the later stage development means a new biomarker project in the SPORE. Basic scientists don't get much recognition. The SPORE

program is organized to generate lots of partially-tested biomarkers with little ability or reward for doing the later work. This has happened repeatedly. It is driven from the bottom and the later part is just "cleaning up." Maybe SPOREs could be adjusted or somehow structure the funding.
Anderson: That's a good point, and the experience isn't limited to SPOREs. Biases in publication go toward the discovery process. It is hard to get negative information published.
Berry: One of the problems of our lives is that we are driven by publication. We get that, and then who cares what happens afterwards.

Q: In terms of statistical modeling – the adaptive and the entropy approach – we often hit a wall with respect to the accuracy. How do we design experiments where you achieve what is satisfactory statistically and clinically? When are we going to make decisions based on removing redundancy? When do we structure a design that assesses redundancy and looks at concordance? We want a marker that is 100% effective, but have we seen it? I've had things come across my desk claiming 99% specific and 99% sensitive, but when it is published, it turns out to be only 75%. How do we know when to say No and when to say Yes?
Berry: There is one interesting example of 99% specificity and sensitivity: dogs sniffing cancer. This has captured the imagination of lots of people. Both in lung cancer and in breast cancer.

Comment: Maybe we need more funding for methods research dealing with the statistical issues we're talking about; more new devices for measuring new biomarkers, and new technology.