Statistical Challenges in Oncology Drug Development

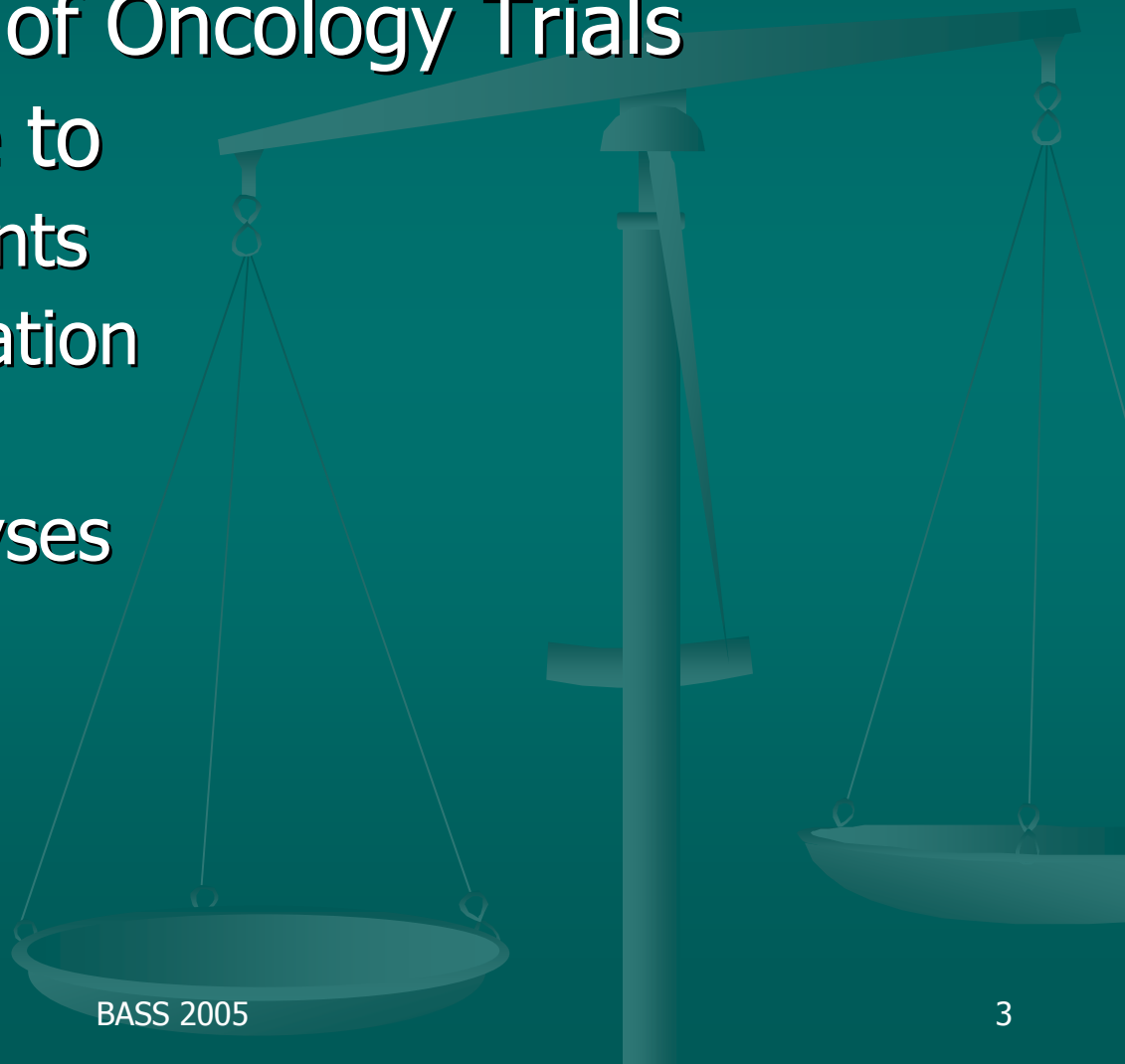
R. Sridhara, Ph.D.
Team Leader (Oncology Drug Products)
Office of Biostatistics
CDER, FDA

Disclaimer

This talk is not an official FDA guidance or policy statement. No official support or endorsement by the FDA is intended or should be inferred.

Outline

- Unique Aspects of Oncology Trials
- Challenges due to
  - Multiple Endpoints
  - Endpoint Evaluation
  - Missing Data
  - Subgroup Analyses
  - Multiple Looks
- Summary
- References



Uniqueness

- Open-label
- Single study
- Multi-center, multi-national, co-operative group studies
- Non-randomized studies
- Life threatening disease
  - Change of treatment during follow-up

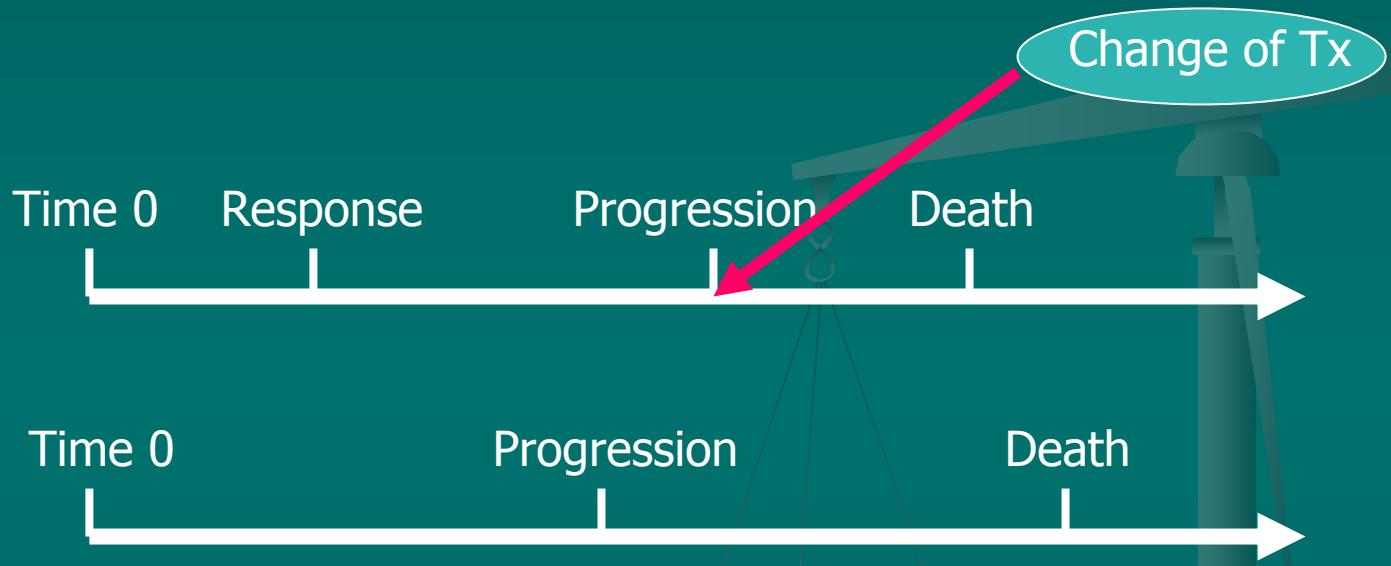
Evaluation of Efficacy

- Is the observed effect true? …….. Statistical question
- Is the magnitude meaningful? …….. Clinical question

Multiple Endpoints

- Solid tumors vs. Hematological malignancies
  - Tumor response rate, Time to progression, Progression-free survival, Time to recurrence, Disease-free survival, PRO, Overall survival
  - Complete remission, Duration of remission, Time to recurrence, Relapse-free survival, Patient reported outcome (PRO), Overall survival
- Primary vs. secondary endpoints

Sequential Endpoints



Censoring can happen at any time due to toxicity or drop out or change of therapy (transplant)

Challenges - 1

- Even though Survival may be the primary endpoint, other endpoints are evaluable early
- Patients receive other therapy after progression
- Possibility of Accelerated Approval (AA) based on early endpoints (RR or PFS)
  - AA based on "surrogate" likely to predict clinical benefit and not considered as clinical benefit
  - Difficulty to further follow patients for survival
  - Treatment cross-over – Estimated effect size ?
  - Efficacy based on interim analysis of "surrogate" endpoint ??
  - Interpretation of p-value ???

Example - 1[1]

- Velcade vs. high-dose Dexamethasone[1] in approximately 700 relapsed multiple myeloma patients, randomized 1:1
- Primary endpoint TTP, but OS (secondary endpoint) is the ultimate endpoint of interest
- Planned one interim analysis for TTP with OBF adjustment
- Protocol did not state total number of events for the OS final analysis

Example -1 (Contd.)

- Interim efficacy analysis of TTP with 50% progression events: p – value < 0.0001, HR ~ 0.55
- DSMB advised to stop the trial, patients (44%) crossed over to new treatment
- Only 20% of enrolled patients were dead, OS analysis with stratified log-rank also significant (?) – Is this real, how to estimate OS effect and report p-value?

Example - 2^2

Taxol for the adjuvant treatment of node positive breast cancer

- Independent DSMB
- Pre-specified plan to conduct 3 interim efficacy analysis with OBF α - spending
  - **25% info, α = 0.00005**
  - 50% info, α = 0.00304
  - 75% info, α = 0.01625
  - 100% info, α = 0.03070
- Trial stopped after first interim analysis (p = 0.0026 (DFS), p = 0.0076 (OS))

Example - 2

Trial Design: 3 × 2 factorial design

Comparison of AC vs. AC + T

	+ T	- T
A 60 mg/m^2 + C 600 mg/m^2	N = 515	N = 533
A 75 mg/m^2 + C 600 mg/m^2	N = 523	N = 517
A 90 mg/m^2 + C 600 mg/m^2	N = 513	N = 520

Example - 2

- Open-label, multi-center, randomized, one phase III study
- Stratified by number of histologically positive lymph nodes at surgery
- Primary endpoints DFS and OS
- Patients were first randomized to receive one of 3 doxorubicin doses and then re-randomized to receive taxol or no taxol.

Example - 2

- Per protocol sample size = 3000 patients (1800 recurrences) based on power to detect 25% decrease in HR for DFS, AC + T arm vs. AC
- Planned Interim analysis at 450, 900, 1350 recurrences, OBF adjustment. **First interim analysis with 25% info, α = 0.00005,** Final analysis α = 0.0307
- Accrual May 1994 – April 1997
- A total of 3170 patients were randomized from 530 centers.

Example – 2: DSMB action

- First interim conducted with 22% events in the AC arm and 18% events in the AC + T arm
- DFS: log-rank p = **0.0026**, Cox model, p = 0.0022
- OS: log-rank p = **0.0076**, Cox model, p = 0.0065

Example - 2

- Trial results made public in May 1998 (approximately 20.4 months of median follow-up) and trial stopped early (all patients had completed treatment).

Example - 2

- The interim analysis results did not meet the pre-specified type I error rate (0.00005).
- Stopping rule only for taxol not for doxorubicin doses in the 3 x 2 design.
- Type I error not adjusted for 2 primary endpoints.
- FDA simulation: B-value – Prob (final analysis significant) = 0.6275 (DFS), 0.5441 (OS)

Example - 2

- Presented at ODAC in Sept 1999
- Drug was approved in 1999
- How to interpret p-value of the final analysis?

Definitions

Overall Survival:

Event = Death

Censor: at last date when patient was known to be alive if patient is lost to follow up or is still alive.

Time = Death/censor date – randomization date

Time to Progression:

Event = Disease Progression

Censor: at last date of evaluation for progression if patient is lost to follow up or has no documented progression (alive or *dead (informed censoring = biased estimates*)).

Time = Progression/censor date – randomization date

Definitions

Progression-Free Survival:

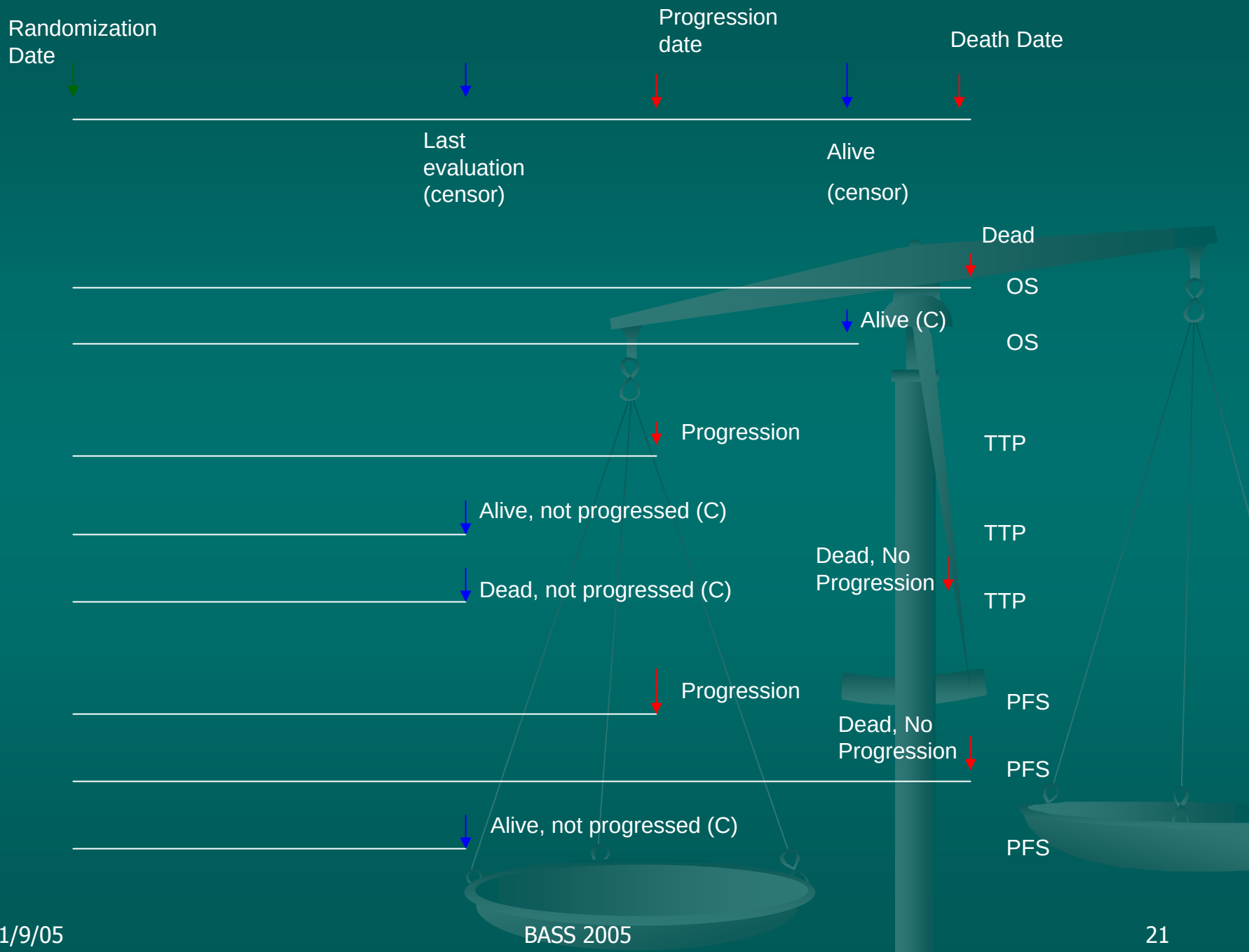
Event = Disease Progression or Death (competing risk and composite endpoint)

Censor: at last date of evaluation for progression if patient is lost to follow up or is alive and has no documented progression

Time = Progression/death/censor date – randomization date

If many deaths prior to progression → biased estimates

- Neither TTP nor PFS are perfect and both are likely to give biased estimates.


Randomization Date
Progression date
Death Date
Last evaluation (censor)
Alive (censor)
Dead
OS
Alive (C)
OS
Progression
TTP
Alive, not progressed (C)
TTP
Dead, No Progression
Dead, not progressed (C)
TTP
Progression
PFS
Dead, No Progression
PFS
Alive, not progressed (C)
PFS

Challenges - 2

- In all our analysis methods we assume that censoring mechanism is independent of the outcome.
  - If this is not true then we have informed censoring
  - The usual methods will produce biased estimates if informed censoring
- Competing risk: If more than one process affects the final event → biased estimates for one cause if censored for another cause.
  - In such cases we may consider composite endpoint (event).

Challanges - 2

Determining Event Dates

★ Actual Tumor Progression

What if Imbalance in tumor/lesion assessment times between treatment arms?

Example – 3^3

- G3139 + DTIC vs. DTIC in 771 melanoma patients presented at ODAC May 2004 (Briefing package and FDA presentation by Yang)
- Failed primary endpoint of overall survival
- PFS a Secondary endpoint
- Primary analysis of PFS based on log-rank test with missing data imputed by LOCF

Results: Median PFS 74 vs. 49 days (DTIC), HR = 0.73, p-value=0.0003

IS THIS TRUE EFFECT?

Example – 3 (Contd.)

- Control group: DTIC (1000 mg/m2) administered by IV infusion over 60 minutes on Day1
- Treatment group: G3139 (7.0 mg/kg/day) administered by continuous IV infusion for 5 days (days 1 – 6) and DTIC (1000 mg/m2) administered by IV infusion over 60 minutes immediately upon completion of the G3139

Example – 3 (Contd.)

Progression Evaluation based on RESIST criteria:

Up to 10 Target lesions (Measurement: sum of longest diameters (LD))

All other lesions = non target lesions

Criteria for disease progression measured every 6 weeks:

≥ 20% Increase in sum of LD of Target lesions, or

Appearance of new lesions, or

Disease progression in non-target lesions

Example – 3 (Contd.)

- Progression of disease status was determined by target lesion measurements when at least 1 target lesion was measured at the visit. As a sensitivity analysis, incomplete lesion measurements were imputed by averaging the 2 measurements that were collected immediately before and after the missing data. If no data were available after the missing value, the missing data were imputed by carrying the last observation forward.
- For subjects whose response at last target lesion measurement was complete response, partial response, or stable disease, progression-free survival was censored at 60 days from last lesion measurement.
- *This method for censoring of missing data elements may introduce bias by adding variable time intervals to the endpoint.*

Example - 3 (Contd.)

- In the FDA analysis of the secondary endpoints, using a more conservative censoring procedure of censoring at last observation for missing data, the progression-free survival difference was 13 days, still highly statistically significantly different (because of the large sample size chosen to detect a survival difference.)
- A number of confounding factors create uncertainty in the interpretation of this measurement including variations in assessment timing and censoring of missing data.
- Simulations conducted by FDA reviewer suggested that in a large study, with very small changes in the interval between assessments, statistically significant differences may be observed which are in fact false positive.

Example – 3 (Contd.)

- Less than half the patients remained on study beyond 2 cycles (43 %) Much of this reflected disease progression from the restaging performed at that time.
- Since lesions were measured periodically, disease progression generally did not occur on the assessment date but rather prior to this date.

Example -3 (Contd.)

- If the intervals between two consecutive assessments (termed assessment intervals) are longer, the documented date of disease progression would tend to be delayed; hence, the observed progression-free survival time would tend to be prolonged.
- Similarly, if the first assessment date is delayed, the observed progression-free survival would also tend to be inappropriately prolonged even if the assessment intervals remain the same.

Example -3 (Contd.)

- In this study, although assessment schedules were intended to be the same (every 6 weeks) between the two treatment groups, because of the nature of treatment schedules (G3139 via 5-day and DTIC only 1-hour infusion), it was observed that the actual assessment schedule for patients in the G3139 + DTIC group appeared generally slightly behind that for patients in the DTIC group as summarized in Table.
- The first 3 assessments were chosen because most patients (~ 85%) had documented disease progression or death by the third assessment . The median times to each assessment appeared slightly longer (statistically significant) in the G3139 + DTIC group than in the DTIC group.

Example -3 (Contd.)

Assessment	G3139+DTIC	DTIC
First	N = 321 48 days (47, 49)	N = 311 43 days (42, 44)
Second	N = 135 94 days (92, 98)	N = 106 87 days (84, 89)
Third	N = 75 137 days (134, 146)	N = 67 129 days (125, 133)

Example – 3 (Contd.)

- Is the difference observed due to difference in the treatment start day of the first cycle between the two arms?

Example -3 (Contd.)

Adjusted Assessment	G3139+DTIC	DTIC
First	N = 321 41 days (41, 42)	N = 311 40 days (40, 41)
Second	N = 135 88 days (84, 91)	N = 106 83.5 days (82, 84)
Third	N = 75 131 days (127, 138)	N = 67 126 days (124, 130)

Example – 3 (Contd.)

- Simulation study by Yang under equal progression-free survival distributions with a forced delay in assessment by 2 days in the experimental arm compared to control arm in only first cycle and also in subsequent cycles.
- Chance of falsely inferring a difference in PFS was high
- Over estimation of median PFS

Example – 3 (Contd.)

- Another simulation study (Yang) with median PFS of 45 days in experimental arm vs. 41 days in the control arm and with a forced 2 days delay in assessment in the experimental arm in first cycle
- Results suggested median PFS of 86 days vs. 42 days and the power of rejecting the null to be close 1.

Conclusion: Trial results unlikely to be true

Challenges - 4: Subgroup Analyses

- When to conduct subgroup analyses
- Interpretation of subgroup analyses results
- Eligibility confirmed after entering the study
- Enriched population studies
- Imbalances between the treatment arms

ICH E-9[4] Guidelines

Section 5.7: Subgroups, Interactions and Covariates:

- ‘ In most cases, however, subgroup and interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall.’

Subgroup Analysis

- When the overall is positive further testing within subgroups OK
  - Closed testing procedure
- What if the subgroup is not positive – can you conclude no effect?

Subgroup Analysis

- Diagnosis results are not confirmed for eligibility before entering the study (example MDS patients) or diagnosis difficult
- Do you conduct the analysis in ITT population or subgroup with the intended indication? What are the consequences?
- Adaptive Design of intentional enrichment studies

Example - 4^5

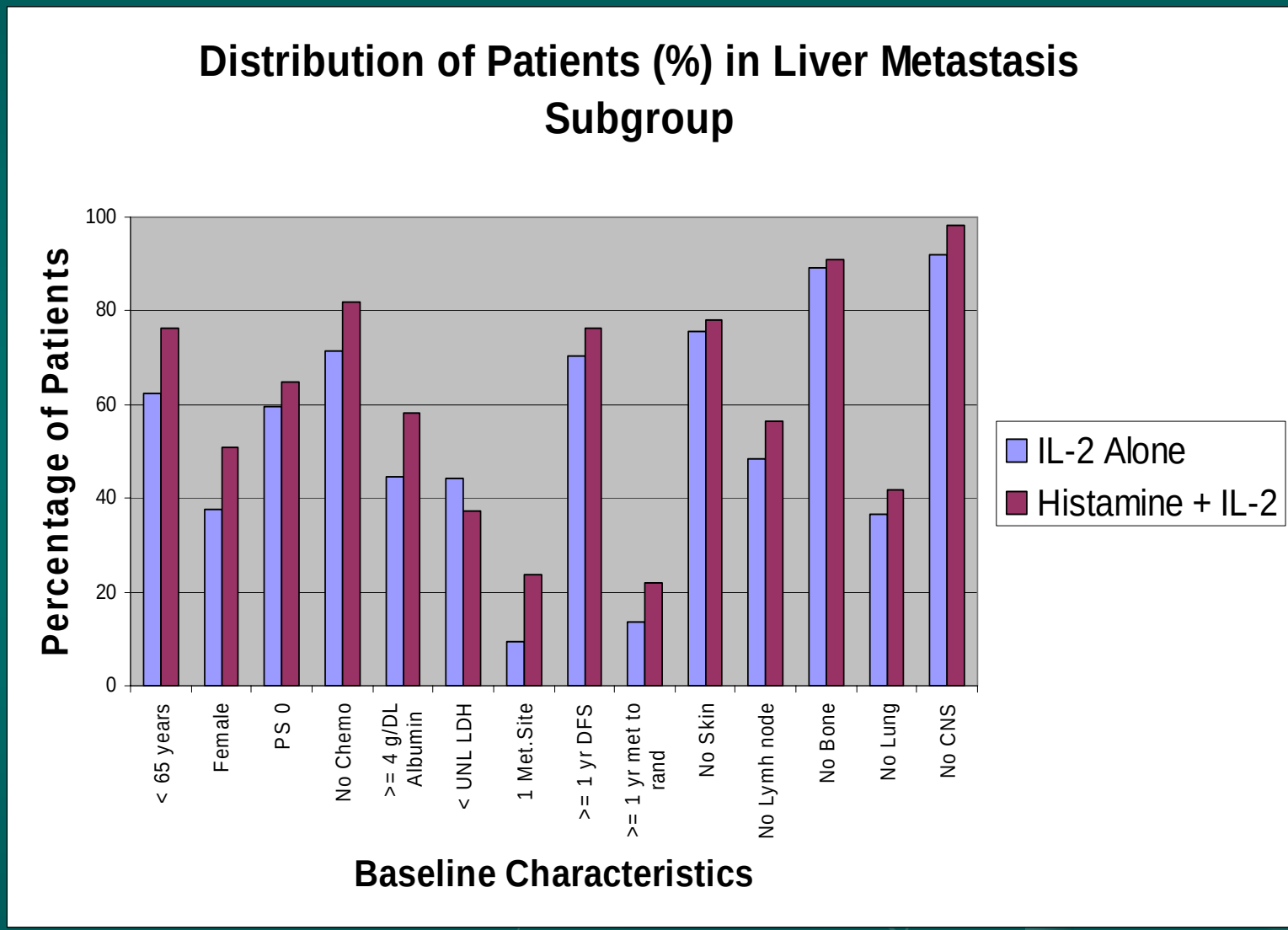
- Randomized open-label study of histamine dihydrocholoride + IL-2 vs. IL-2 alone in 300 advanced metastatic melanoma patients
- Study failed to demonstrate survival benefit in the ITT population.
- Sponsor claimed survival efficacy in the subgroup of patients with liver metastasis
- Randomization not Stratified for liver metastasis or no liver metastasis
- Imbalances favoring Histamine + IL-2 arm

KM ESTIMATES OF MEDIAN DURATION OF SURVIVAL (IN MONTHS) - LM and No LM SUBGROUPS (3/8/00 and 9/8/00)

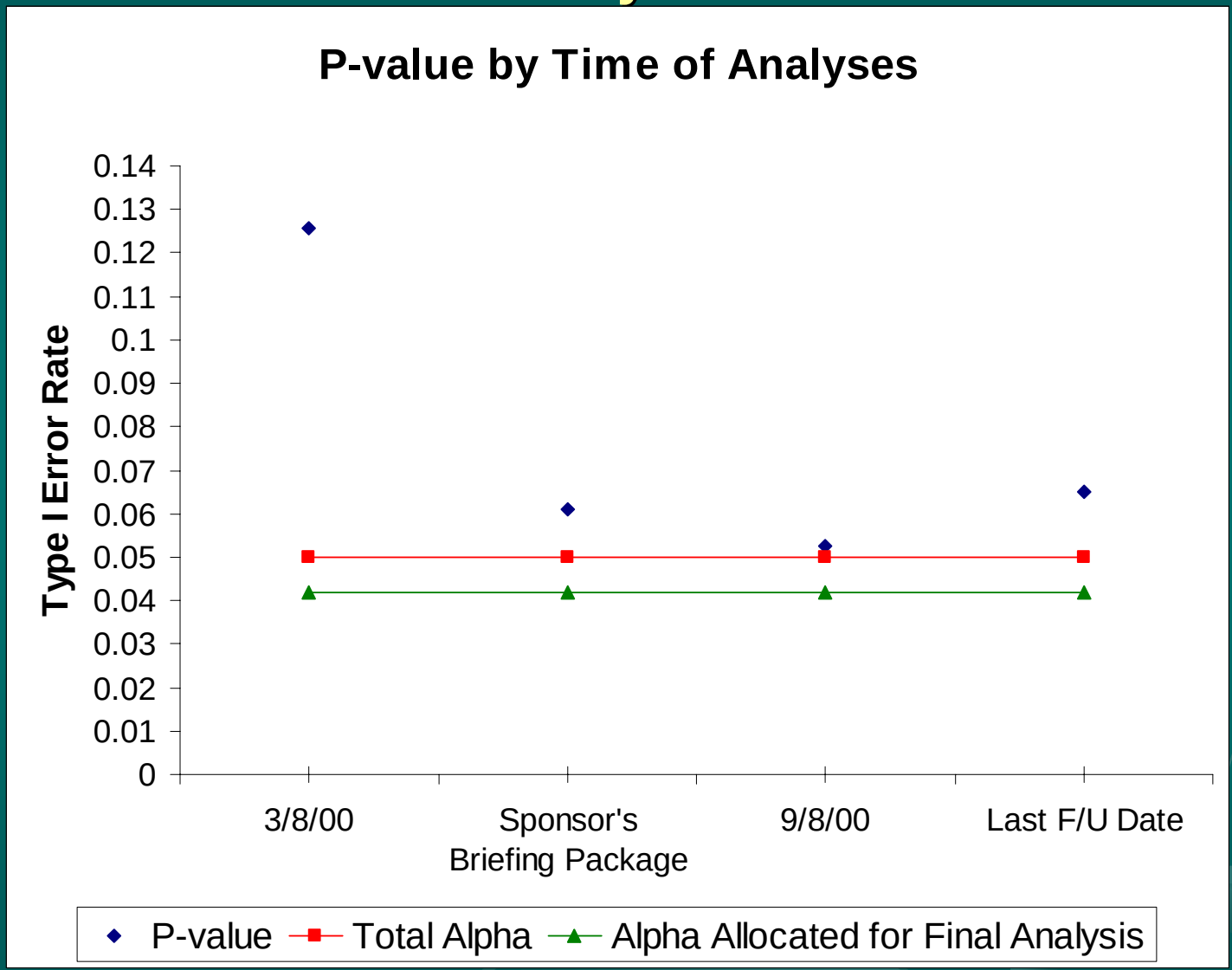
Population	IL-2	IL-2 + Histamine	Hazard Ratio[1] (95% C.I.)	P-value[2] (Log-rank test)
LM				
N	74	55		
# Dead[3]	***69***	***42***		
Median[3] (95% C.I.)	5.0 (3.9, 6.7)	9.2 (6.4, 12.7)	0.568 (0.383, 0.835)	0.0040
# Dead[4]	***72***	***46***		
Median[4] (95% C.I.)	5.0 (3.9, 6.7)	9.2 (6.4, 12.7)	0.572 (0.392, 0.835)	0.0033
No LM				
N	79	97		
# Dead[3]	***57***	***75***		
Median[3] (95% C.I.)	10.3 (8.6, 12.3)	8.7 (6.6, 10.4)	1.142 (0.811, 1.600)	0.4493
# Dead[4]	***67***	***80***		
Median[4] (95% C.I.)	10.3 (8.6, 12.3)	8.7 (6.6, 10.4)	1.047 (0.756, 1.452)	0.7808

[1] Hazard Ratio = Histamine + IL-2 / IL-2 alone; [2] Unadjusted P-value;
[3] Cut-off date 3/8/2000 per NDA submission; [4] Cut-off date 9/8/2000 per updated submission

Example - 4: Imbalances



Histamine Dihydrochoride: Multiple Survival Analyses



Example - 5[6]

- A randomized, open-label study of standard WBRT/oxygen, with or without RSR13, in patients with brain metastases (Control arm N = 267; RSR13 arm N= 271)
- Study failed to demonstrate survival benefit in the ITT population. Sponsor claimed efficacy in the subgroup of patients with breast cancer primary
- Imbalances in baseline characteristics between the treatment arms in the subgroup

Example - 5: Imbalances within Breast Primary Subgroup– Important Factors

Characteristic	WBRT	RSR13 + WBRT
Bidirectional Area of Baseline Brain Lesions (mm^2)		
Mean (S.D.) Median (Range)	882 (695) 699 (17 – 3588)	762 (706) 579 (16 – 2936)
Number of Brain Lesions 3 or more	74.1%	56.7%
Extracranial Mets. 3 or more	40%	31.7%

None of these were individually statistically significant;

P-value for Brain lesions (single vs. multiple) = 0.07)

Example - 6[7]

- Randomized study of adjuvant therapy with eloxatin in combination with infusional 5-FU/LV vs. infusional 5-FU/LV alone in 2246 patients with stage II or III colon cancer.
- Primary endpoint DFS in ITT was statistically significant (599 events, HR = 0.76, p=0.0008)
- In 1347 Stage III patients DFS significant (452 events, HR = 0.75, p=0.002)
- In 899 Stage II patients DFS not significant (147 events, HR = 0.80, p=0.179)
- Indicated only in Stage III patients.

Challenges with Symptom Improvement - LCS Score

	Not at all	A little bit	Some-what	Quite a lot	Very much
1. I have been short of breath	0	1	2	3	4
2. I am losing weight	0	1	2	3	4
3. My thinking is clear	0	1	2	3	4
4. I have been coughing	0	1	2	3	4
5. I have a good appetite	0	1	2	3	4
6. I feel tightness in my chest	0	1	2	3	4
7. Breathing is easy for me	0	1	2	3	4

Symptom Improvement

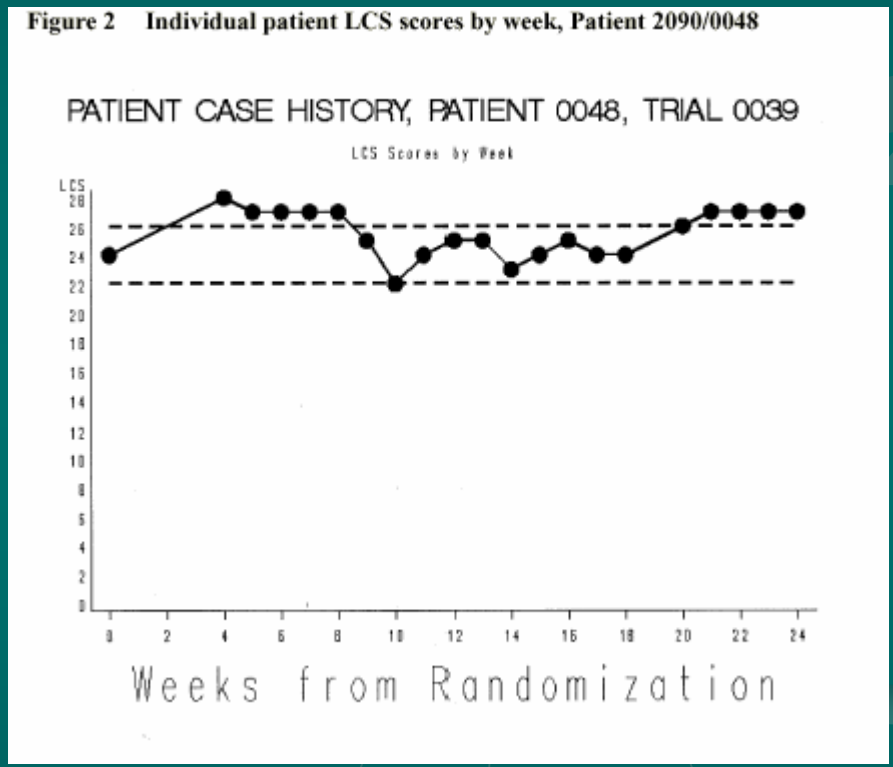
- LCS sub-scale total score = 28 $\Rightarrow$ No symptom
- Sponsor definition of symptomatic patient: Baseline total LCS score $\leq$ 24
- Symptom improvement defined by sponsor as: Increase in the total LCS score by $\geq$ 2: An increase from a baseline score of 24 to 26 is an improvement - so also an increase from a baseline score of 4 to 6

Example - 7[8]

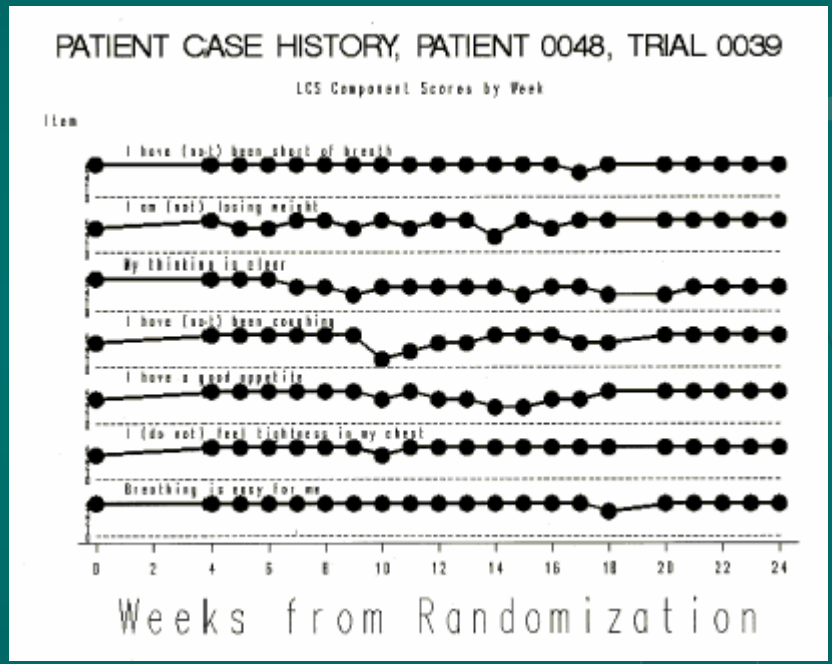
- 250 mg ZD1839 Treatment Arm:
  - 44/102 (43.1 %) with symptom improvement per sponsor on the LCS scale
  - 32 (31.4 %) with symptom improvement in LCS, FACT and TOI
- 500 mg ZD1839 Treatment Arm:
  - 41/114 (36 %) with symptom improvement per sponsor on the LCS scale
  - 20 (17.5 %) with symptom improvement in LCS, FACT and TOI

Patient LCS Profile - An Example

Figure 2 Individual patient LCS scores by week, Patient 2090/0048

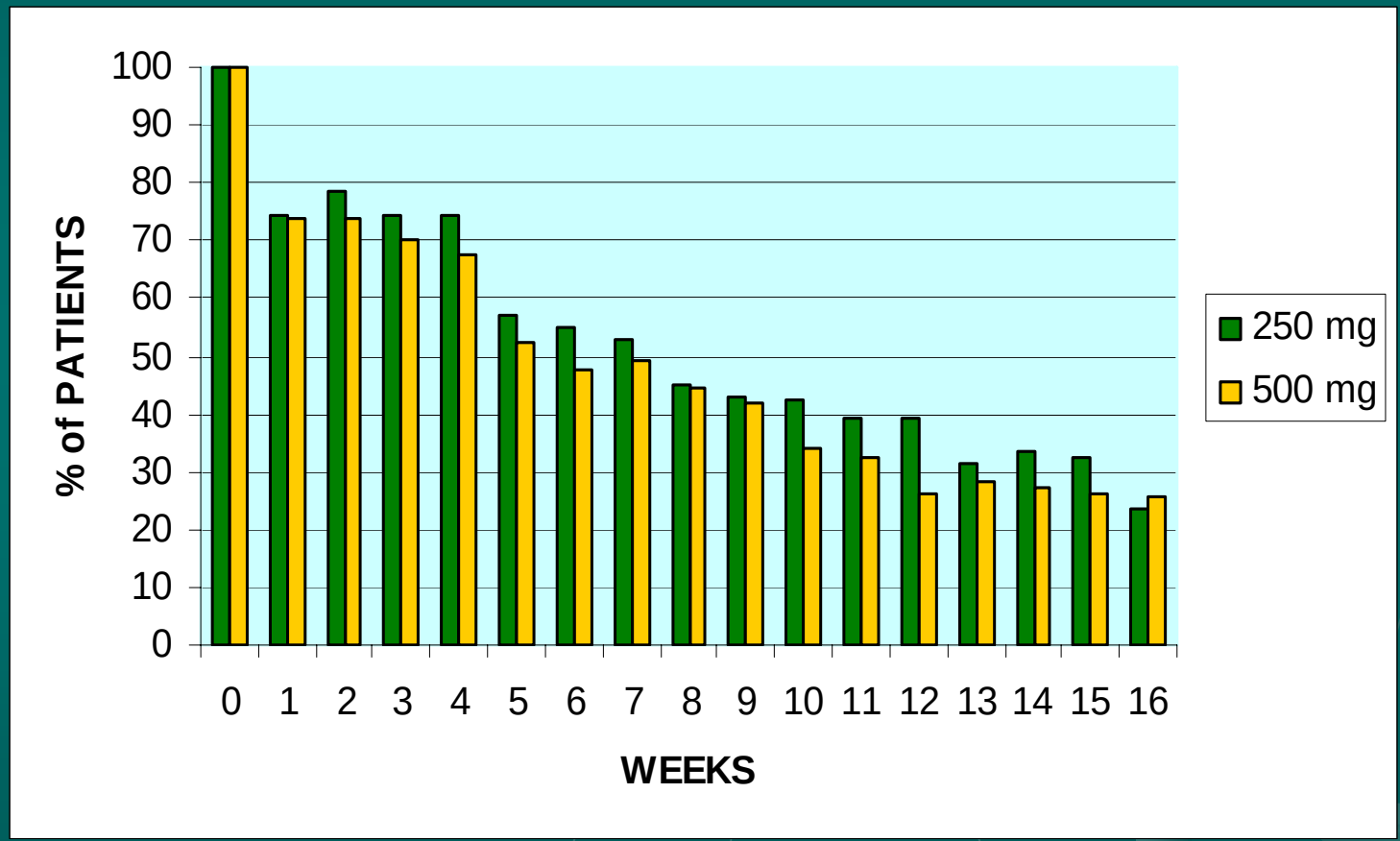


Patient LCS Profile - An Example



1. Short of Breath
2. Losing Weight
3. Thinking is Clear
4. Been Coughing
5. Good Appetite
6. Tightness in Chest
7. Breathing Easy

% of Patients Evaluated Over Time



Challenges with Surrogates

- Surrogate independent of treatment
- Surrogate independent of baseline risk (demographics, behavior, pharmaco-genomics, etc.)
- Accuracy, Precision and Timing of measurement of surrogate

Summary

- We have many statistical challenges. To list a few:
  - Interpretation of p-value
  - Estimation of effect
  - Endpoint definition & evaluation
  - Interpretation of subgroup analyses
  - Interpretation of symptom benefit

References

1. NEJM 352: 2487-2498, 2005
2. Oncology Drug Advisory Committee Meeting Sept 17, 1999
3. Oncology Drug Advisory Committee Meeting, May 3, 2004
4. ICH Harmonised Tripartite Guideline, E9: Statistical Principles for Clinical Trials
5. Oncology Drug Advisory Committee Meeting, Dec 13, 2000
6. Oncology Drug Advisory Committee Meeting, May 3, 2004
7. Oxaliplatin product drug label
8. Oncology Drug Advisory Committee Meeting, Sept 24, 2002