Some design and analysis issues for PRO's that might use IRT / CAT

Robert T. O'Neill Ph.D.

Director, Office of Biostatistics

CDER, FDA

DIA Workshop on Advances in Health Outcomes Measurement: Exploring the Current State and the Future Applications of Item Response Theory, Item Banks, and Computer-adaptive Testing
Bethesda – 25 June 2004

Two areas

- Item response theory (IRT)
  - Idea that an optimum test is constructed for each subject
  - Differential Item Functioning (DIF) - pre-screening and during RCT
    - Conditional probability of response for the same level of the latent variable differs for two groups
- Computer adaptive testing (CAT)

Some Questions about IRT / CAT to assess treatment effects in clinical trials

- Probability of correctly answering question (item) - how do you know ?
- Measuring change in endpoint response over time
  - What is controlled
- Controlling for differences among subjects
  - different items for different subjects
  - item content may vary within subject over study duration
- Multiple endpoints; structure of clinically relevant and meaningful treatment effects that may be differential across subjects (composites)

Trial Design Considerations

- Clinical relevant measures of treatment induced effects
  - effect size, correlation among effects, dimensionality of relevant
- Assessment of uncertainty in conclusions and interpretations - Controlling for chance findings
- Minimizing bias

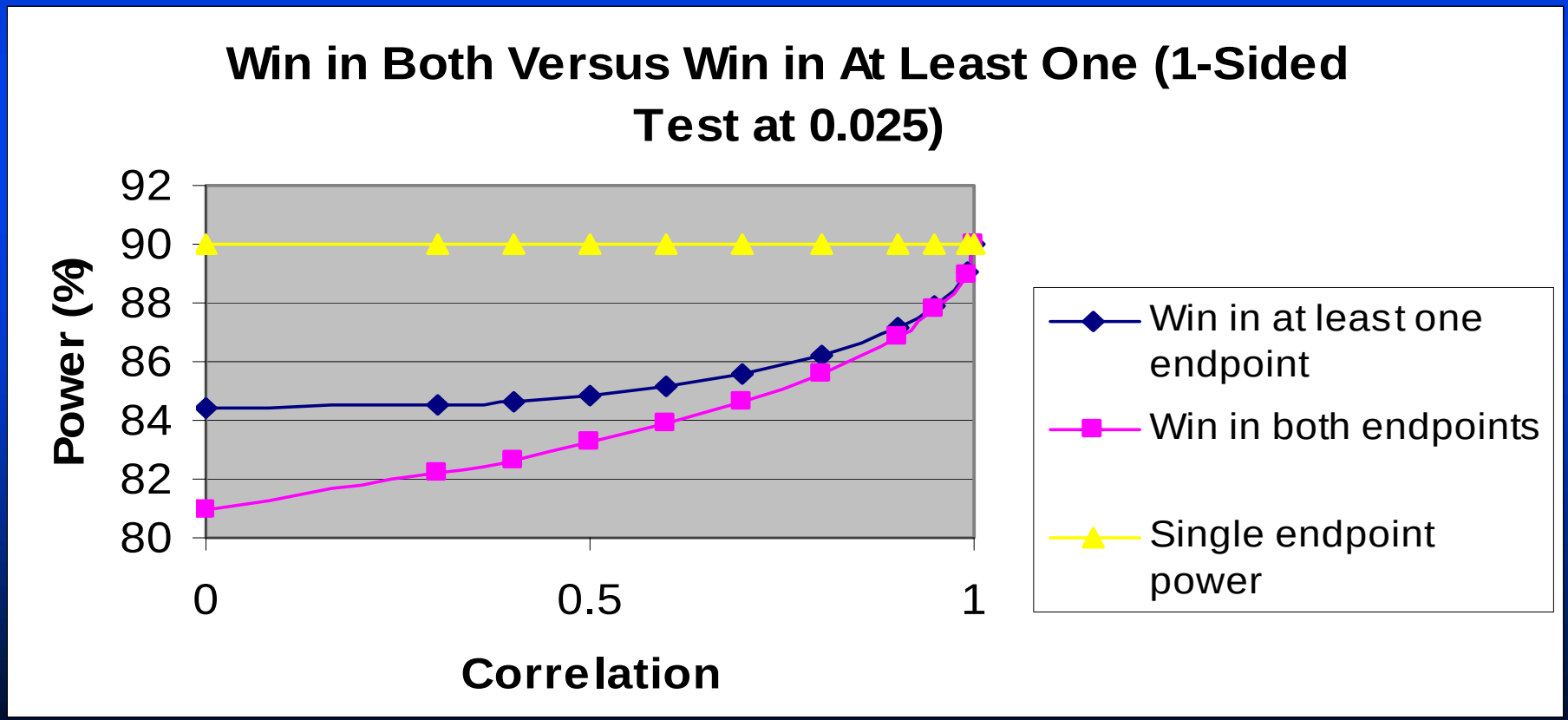
Data Analysis Considerations

- Multiple endpoints
- Missing Data
- Interpretation

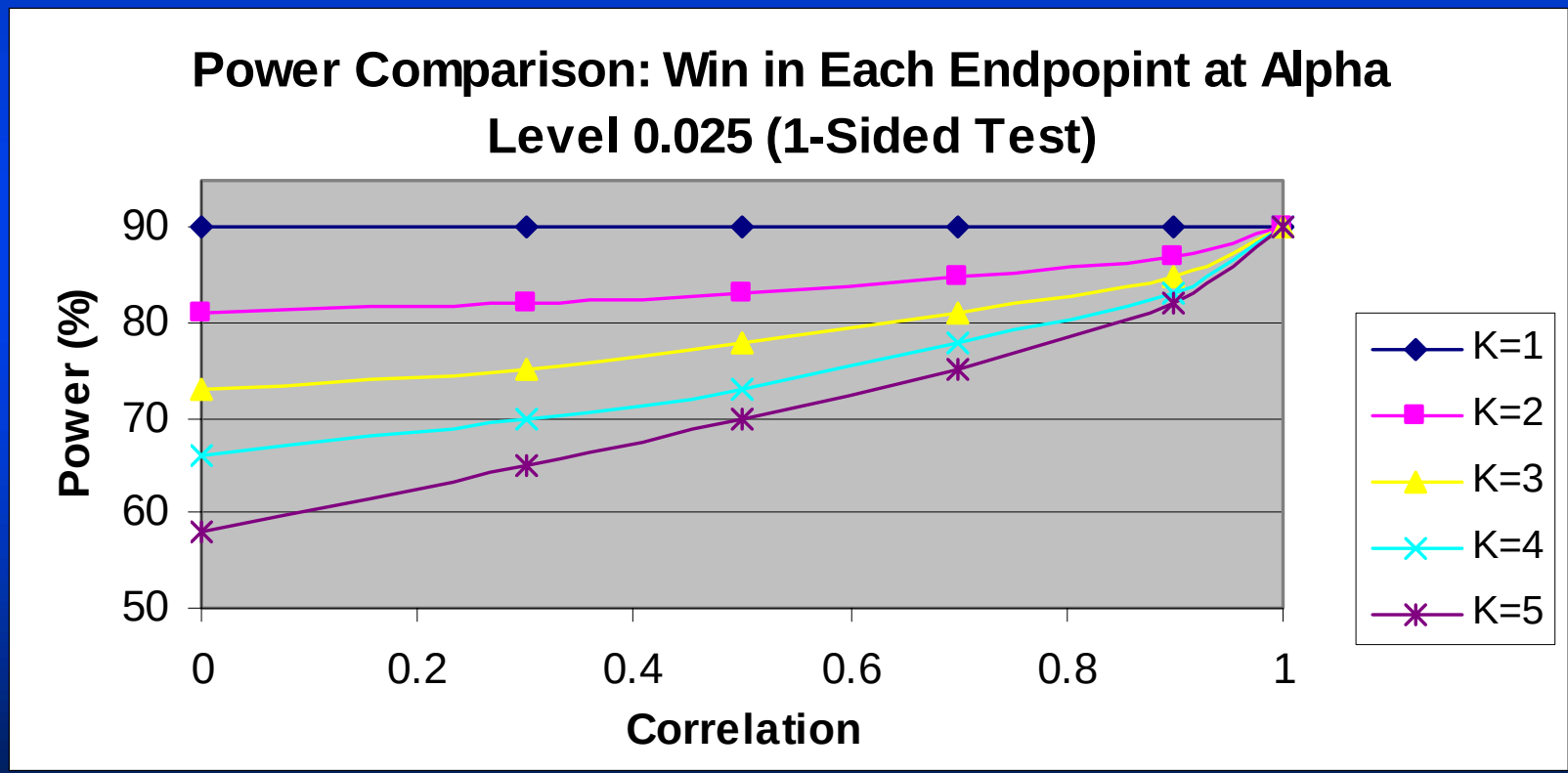
Multiple endpoints and controlling false positive conclusions

- Criteria for characterizing the treatment induced effects - multiple endpoints
  - Primary , co-primary
  - Secondary - follow hierarchy
  - Composite (how chosen, how components change with treatment)
- Controlling the chances of false positive conclusions is a function of which multiplicity strategy is pre-specified in the protocol

Statistical Implications: "Any vs All" Power Comparison Case of K=2 endpoints:



Loss in Power when win in all endpoints K=# of endpoints



What do the Europeans think about this ?

The European Agency for the Evaluation of Medicinal Products
Evaluation of Medicines for Human Use

London, 19 September 2002
CPMP/EWP/908/99

COMMITTEE FOR PROPRIETARY MEDICINAL PRODUCTS (CPMP)

POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS

DISCUSSION IN THE EFFICACY WORKING PARTY	January 2000
TRANSMISSION TO CPMP	July 2001
RELEASE FOR CONSULTATION	July 2001
DEADLINE FOR COMMENTS	October 2001
DISCUSSION IN THE EFFICACY WORKING PARTY	June 2002
TRANSMISSION TO CPMP	September 2002
ADOPTION BY CPMP	September 2002

POINTS TO CONSIDER ON MULTIPLICITY ISSUES IN CLINICAL TRIALS

1. INTRODUCTION

Multiplicity of inferences is present in virtually all clinical trials. The usual concern with multiplicity is that, if it is not properly handled, unsubstantiated claims for the effectiveness of a drug may be made as a consequence of an inflated rate of false positive conclusions. For example, if statistical tests are performed on five subgroups, independently of each other and each at a significance level of 2.5% (one-sided directional hypotheses), the chance of finding at least one false positive statistically significant test increases to 12%.

This example shows that multiplicity can have a substantial influence on the rate of false positive conclusions which may affect approval and labelling of an investigational drug whenever there is an opportunity to choose the most favourable result from two or more analyses. If, however, there is no such choice, then there can be no influence. Examples of both situations will be discussed later. Control of the study-wise rate of false positive conclusions at an acceptable level α is an important principle and is often of great value in the assessment of the results of confirmatory clinical trials.

A number of methods are available for controlling the rate of false positive conclusions, the method of choice depending on the circumstances. Throughout this document the term 'control of type I error' rate will be used as an abbreviation for the control of the family-wise type I error in the strong sense, i.e., there is control on the probability to reject at least one true null hypothesis, regardless which subset of null hypotheses happens to be true. The issue of setting an appropriate type I error level on a submission level when this includes the need for more than one confirmatory trial is discussed in a separate Points-to-Consider document (CPMP/2330/99 Points to Consider on Application with 1.) Meta-analyses and 2.) One Pivotal study).

This document does not attempt to address all aspects of multiplicity but mainly considers issues that have been found to be of importance in recent European applications. These are:

- Adjustment of multiplicity – when is it necessary and when is it not?
- How to interpret significance with respect to multiple secondary variables and when can a claim be based on one of these?
- When can reliable conclusions be drawn from a subgroup analysis?
- When is it appropriate for CPMP to restrict licence to a subgroup?
- How should one interpret the analysis of "responders" in conjunction with the raw variables?
- How should composite endpoints be handled statistically with respect to regulatory claims?

There are further areas concerning multiplicity in clinical trials which, according the above list of issues, are not the focus of this document. For example, there is a rapid advance in methodological richness and complexity regarding interim analyses (with the possibility to stop early either for futility or with the claim of effectiveness) or stepwise designed studies (with the possibility for adaptive changes for the future steps). However, due to the importance of the problem and the amount of information specific to this issue it appears appropriate that a separate document may cover these aspects.

Multiple secondaries

Composites

PRO's may be related to the clinical endpoint in many ways

- Surrogate: measures how a patient feels, functions, or survives and should be calibrated by outcomes determining defined clinical benefit that reflects the changes predicted by the earlier measured surrogate
- Surrogacy : how much of the treatment effect is captured by the surrogate
  - independence, correlation with other endpoints
  - another descriptor of effect or another dimension of effect

Subgroups: Heterogeneity of the disease process and outcomes

- Subgroup differences , stratification, choice and use of covariates to increase statistical power
- When are observed treatment differences real - difficult but
  - In RA patient groups in this heterogenous disease have different levels of clinical response
  - But PRO's may be more sensitive to change than traditional measures

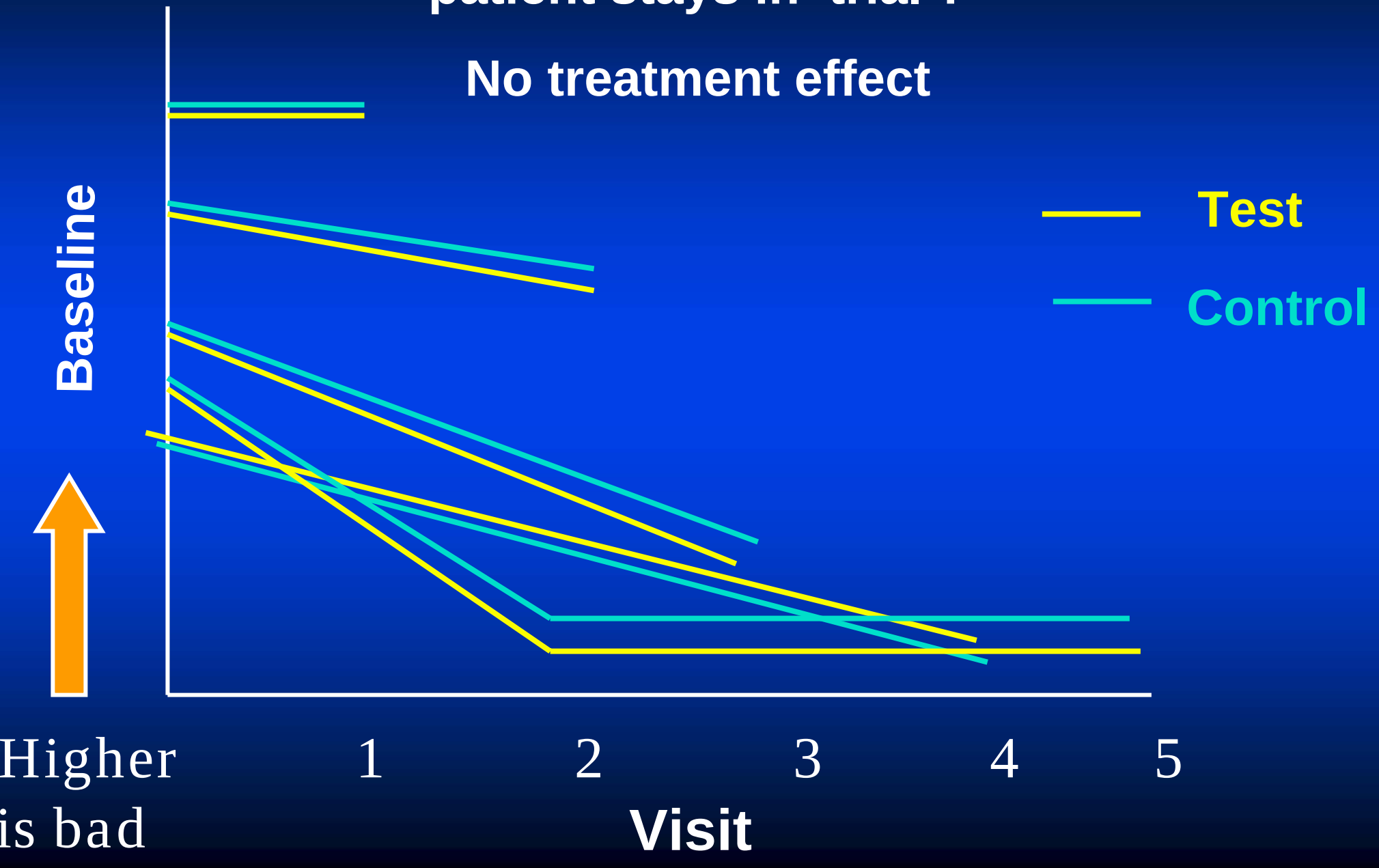
Missing data due to withdrawal from a trial prior to planned completion

- PRO's are very likely predictors of satisfaction with assigned treatment and with staying in a trials

What is unique about missing data in clinical trials ?

- Monotonically Missing data is potentially an outcome by itself
  - Why ? - It can be a surrogate for patient preference, acceptability with therapy, and can potentially be unpredictive of where the subject would be in the future (where no observations are taken or available)
- With monotone missing data, the ‘dropout mechanism’ is very likely informative
- It’s possible to plan to collect information during study prior to a patient withdrawal from treatment, prior to study completion but post treatment withdrawal (conditioning)

Are slope and baseline predictive of how long a patient stays in trial ?



Evaluation of dependency of outcome and time on study

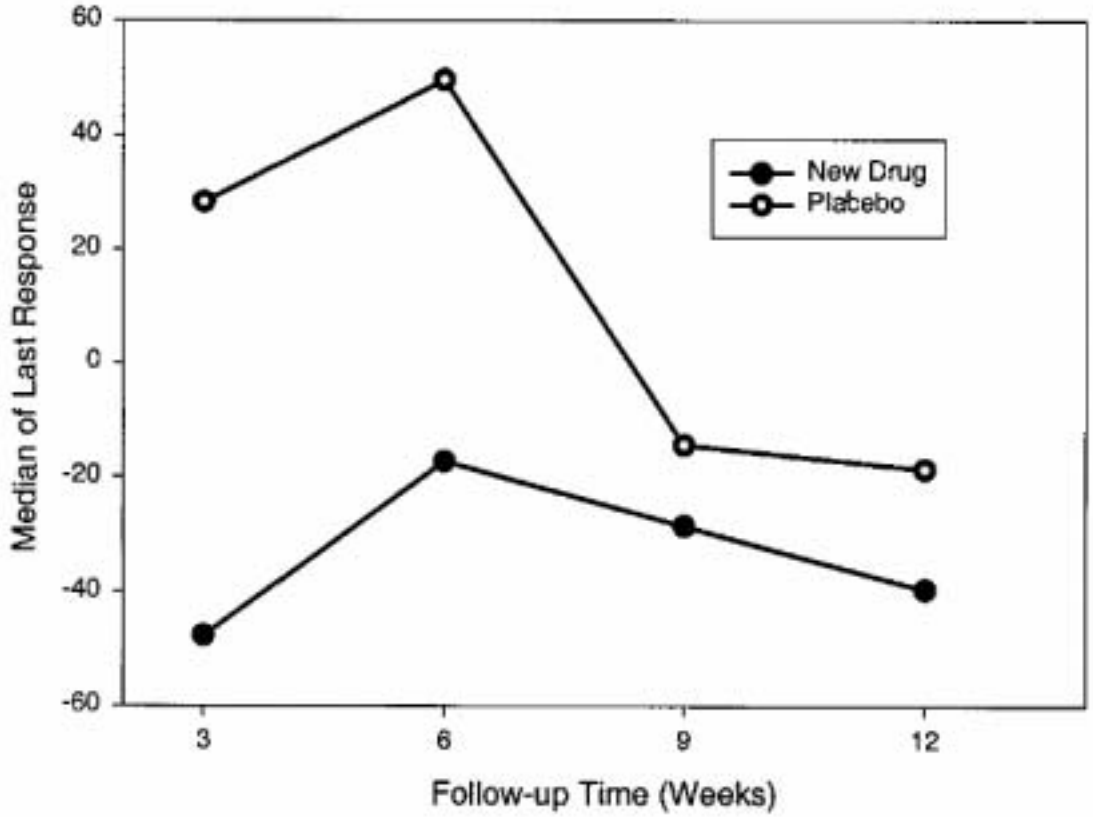


Figure 2. Relationship between patients' follow-up time and last response (per cent change of β-agonist use).

LEARNING, PRIVATE INFORMATION AND THE ECONOMIC EVALUATION OF RANDOMIZED EXPERIMENTS

Tat Y. Chan and Barton H. Hamilton
John M. Olin School of Business
Washington University in St. Louis
Campus Box 1133
One Brookings Drive
St. Louis, MO 63130
USA

chan@olin.wustl.edu; hamiltonb@olin.wustl.edu

July 2003

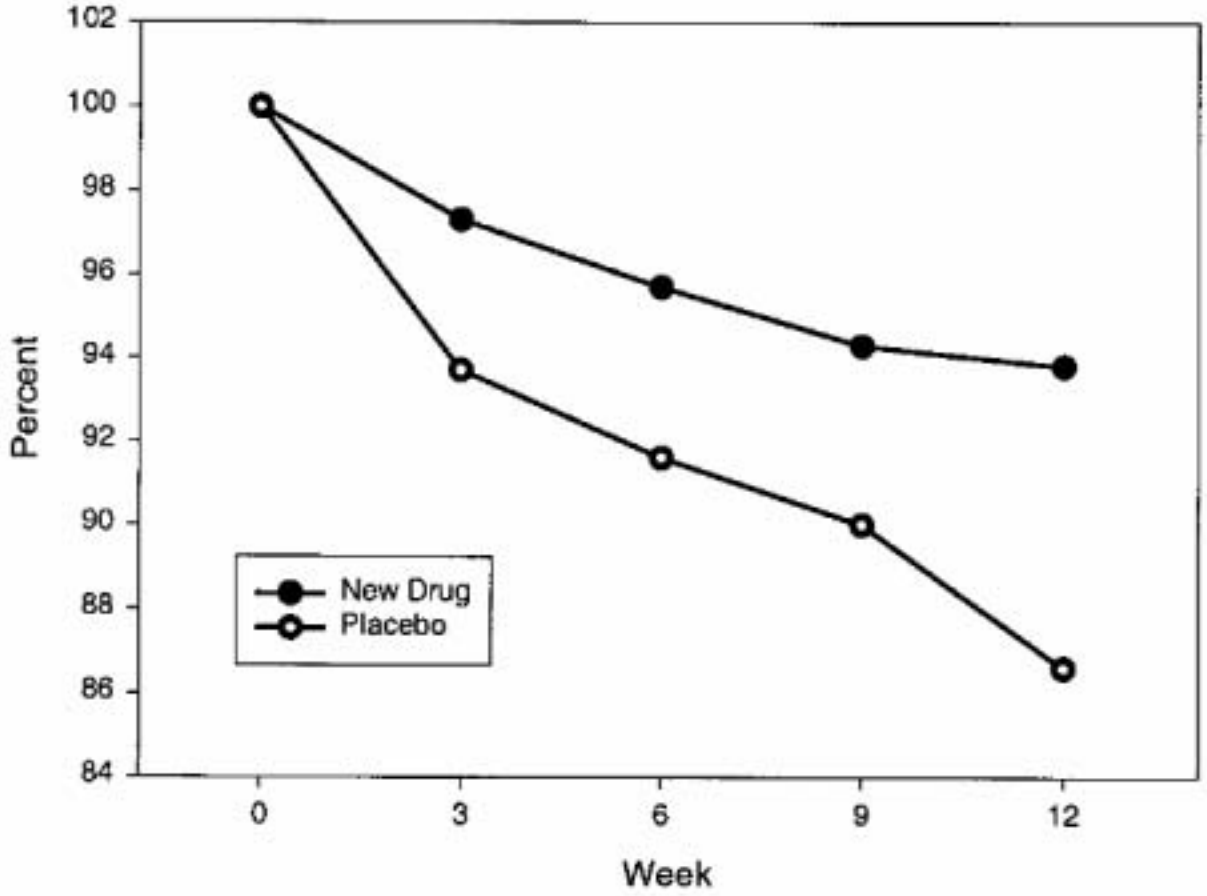


Figure 1. Per cent of patients who stayed in the study (anti-asthma clinical trial example).

Placebo group had a higher drop- out rate (or, equivalently, shorter follow- up time) than the new drug group.

Many drop outs due to worsening asthmatic symptoms, hence were considered to be dependent drop- outs .

STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:1197–1214

FIGURE 3
CD4 COUNT PROFILES, BY TREATMENT ARM

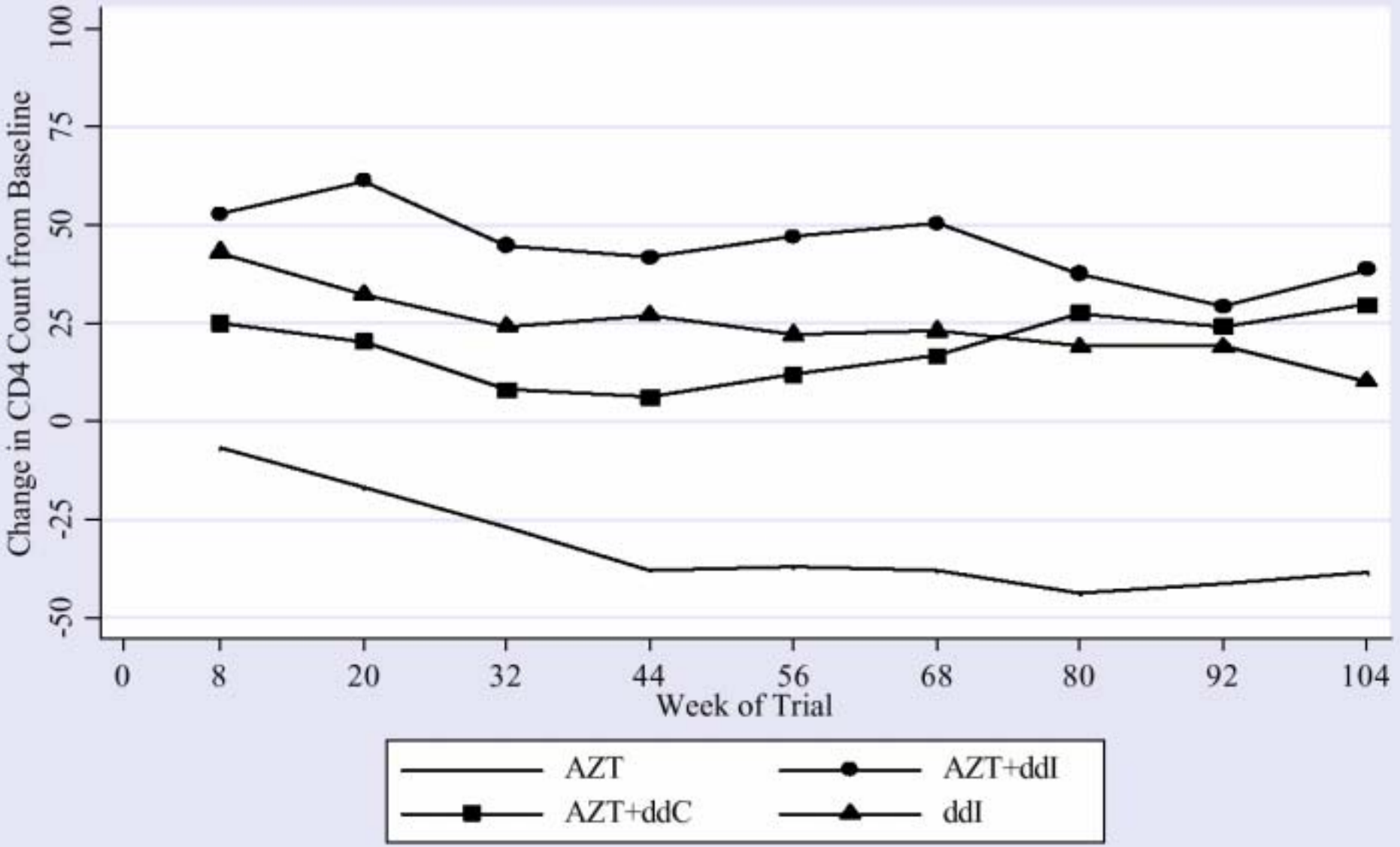
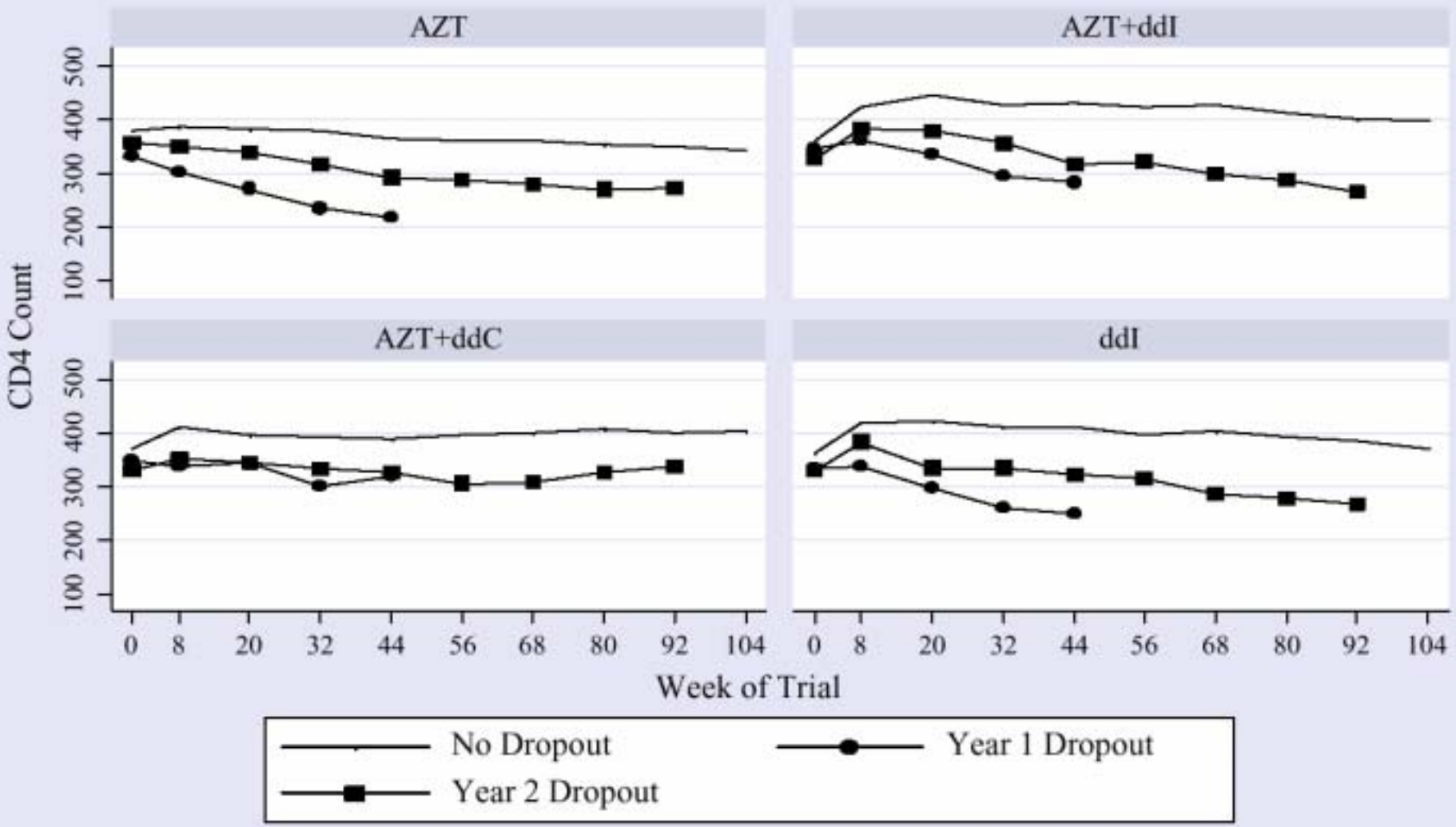


FIGURE 5
CD4 COUNT PROFILES, BY TREATMENT AND ATTRITION GROUP



Compare performance characteristics against known instruments used in several drug / disease areas

- Arthritis
  - ACR20; DAS (disease activity score)
- Schizophrenia
  - Workshop on Clinical Trial Designs for Neurocognitive Drugs for Schizophrenia ; April 23, 2004 sponsored by NIMH, FDA, MATRICS

Rheumatoid Arthritis

ARTHRITIS & RHEUMATISM
Vol. 48, No. 3, March 2003, pp 625–630
DOI 10.1002/art.10824

An Index of the Three Core Data Set Patient Questionnaire Measures Distinguishes Efficacy of Active Treatment From That of Placebo as Effectively as the American College of Rheumatology 20% Response Criteria (ACR20) or the Disease Activity Score (DAS) in a Rheumatoid Arthritis Clinical Trial

T. Pincus,[1] V. Strand,[2] G. Koch,[3] I. Amara,[4] B. Crawford,[5] F. Wolfe,[6] S. Cohen,[7] and D. Felson[8]

Schizophrenia (Some messages)

- Eliminate from entry, subjects who are at ceiling - that is they cannot improve from baseline - and even if they do, it may not be a clinically meaningful change.
- Distinguish between changes in cognition that are not secondary to changes in symptom - ie a drug has both anti-psychotic and cognition impact
- Need to maintain stability of the condition to tease out cognitive changes
- FDA's concern: a small effect on a cognitive measure that we don't know how to interpret what its clinical impact is; Low levels of cognitive changes impacting ability to show functional changes.
- Domains; Need two co-primary ?; symptoms, functional

Concluding remarks

- We will need considerably more experience and understanding of IRT / CAT methods and applications
- Choose a few areas and demonstrate how it would work - only then can informative evaluations occur
- It is very early to commit to developing a data bank to use in drug trials - others may be further along in their thinking