Joint DIA / PSI Statistics Journal Club: Multiregional Trials

Discussion of "Assessment of consistency of treatment effects in multiregional clinical trials" (Quan et al, DIJ 2010)

Paul Gallo

April 5, 2011

Hui Quan
Sanofi-Aventis

Mingyu Li
Celgene

Jeshua Chen
Merck

Paul Gallo
Novartis

Bruce Binkowitz
Merck

Ekopimo Ibia
Merck

Yoko Tanaka
Eli Lilly

Soo Peter Ouyang
Celgene

Xiaolong Luo
Celgene

Gang Li
Johnson & Johnson

Shailendra Menjoge
Boehringer Ingelheim

Steven Talerico
Merck

Kimiteshi Ikeda
Novartis

Key Words
Ethnic effect; Sample size; Power; Noninferiority; Interaction; Random effect model

Assessment of Consistency of Treatment Effects in Multiregional Clinical Trials

Multiregional clinical trials (MRCTs) present great opportunities but also challenges to the trial community. To address the challenges and fully realize the opportunities, a PhRMA MRCT Cross-Functional Key Issue Team (KIT) was formed in 2008. One of the work streams within the KIT particularly focuses on the assessment of consistency of treatment effects across regions. As the main objective of this work stream, this research explores a number of definitions for consistency assessments. We address the issues primarily for superiority trials with continuous endpoints, then extend briefly to noninferiority trials, random effect models, binary endpoints, and survival endpoints. Computations and simulations are used to study the properties of the proposed definitions, particularly the power for showing consistency. To illustrate applications of the methods, we use a trial example with a continuous endpoint. We discuss considerations for trial design as well as for data analysis. The consistency assessment relies heavily on the definition of regions and the number of regions. We recommend working with health authorities to define region in a manner that is sensible from a practical interpretation standpoint and also makes region consistency assessment a feasible undertaking.

INTRODUCTION

Heterogeneity resulting from differences in ethnicity, genetic factors, culture, and clinical practice across regions can present challenges for the conduct and interpretation of multiregional and bridging studies, or criteria for the assessment of similarity of treatment effect across regions. A number of different proposals for this assessment have been made (2–5), focusing on methodologies for a bridging study comparing treatment effects be-

Background

- Multiregional clinical trials (MRCTs) present *opportunities* and *challenges*.
- To attempt to realize the opportunities and address the challenges, a PhRMA Cross-Functional Key Issue Team on MRCTs was formed in 2008.
- The team formed several workstreams to focus on specific issues.
- This paper was produced by a workstream concentrating on *assessment of consistency* across regions.

Objectives and summary

- Summarize and characterize statistical tools available for addressing consistency
- Extend / generalize existing approaches
- Provide a basis for evaluating their properties
  - power, sample size; design implications
- Provide a framework to assist in making decisions as to which approaches might be used, and in what circumstances.

Multi-regional trial motivation

- Increasingly, clinical trials are run using patients from various regions worldwide.
  - More patients needed to demonstrate treatment advantages, as new treatments may have only incremental benefits vs existing therapies.
  - Local health authorities would like to see representation / evidence within their domains.
  - Varied settings may enhance confidence in observed effects.
  - Expanded markets interest trial sponsors.

Regulatory perspectives

- Regulatory viewpoints are an important driver in the current dialog on region consistency.
- ICH-E5 Q&A
  - "*. . . a multi-regional study should be designed with sufficient numbers of subjects so that there is adequate power to have a reasonable likelihood of showing an effect in each region of interest*."
  - "*It will be of interest also to examine consistency of effects across regions*."
- What does this mean? How do we quantify?

Challenges for design and interpretation

- Investigating region consistency is prone to the challenges arising in other subgroup contexts.
- How do we balance:
  - the poorer ability to control error rates
  - the possibility of increasing resources – perhaps *substantially* (*is this justified?*)
  - the appropriate reaction to visual signals that are within the realm of chance, but *important* if true.

Observed effect reversals

- There are numerous literature quantifications of *effect reversals*, i.e., observing region effects in the wrong direction, given *true efficacy and homogeneity*:
  - Hung *et al* (*Pharm Stat.* 2010)
  - Marschner (*Clin Trials* 2010)
  - Li *et al* (*DIJ* 2007)
  - Senn (*Drug Development* text, *Multicenter Trials)*
- Basically, the chance is distressingly high . . .

MHLW proposals

Much constructive dialog followed from proposals made in 2007 Japanese guidance.

- Method 1
  - Observed effect in Japanese patients achieves a specified fraction of the effect in the full study population (½ is "*appropriate*")
- Method 2
  - "*Similar tendency*": all region effects exceed a particular value, e.g., 0.
- Desire 80% chance of achieving these.
  - see Quan et al (*Pharm Stat* 2010), Kawai et al (*DIJ* 2007).

Characterizing consistency methods

Structurally, approaches tend to be of 2 types.

- Methods that tend to conclude consistency until there is sufficient evidence to the contrary
  - e.g., interaction tests
- Methods requiring a certain strength of signal of similarity in order to conclude consistency
  - MHLW proposals
- Which type is appropriate for a given situation? Where should the *burden of evidence* lie?

“Definition 1”

- In order to conclude consistency, each of *s* regions should achieve a proportion, π, of the observed overall effect.
  - π = 0: MHLW Method 2.
  - It may be sensible to select π as a function of the number of regions, e.g., $\pi = 1/s$.
  - Note: this lower bound induces an upper bound as well:
    - e.g., effects in 2 equally-sized regions must not differ by a factor > 3.

“Definition 2”

- In order to conclude consistency, each region should achieve a common pre-specified constant value ($b \geq 0$).
  - $b = 0$: MHLW Method 2
  - b = design effect size / s may be reasonable.

“Definition 3”

- In order to conclude consistency, demonstrate through hypothesis testing that each region achieves a proportion, π, of the overall effect.
  - Larger-than-conventional significance levels may be needed to make this approach viable
  - π = 0: essentially, significance within each region
  - α’ = 0.5: reduces to Definition 1.

“Definition 4”

- In order to conclude consistency, a test for treatment-by-region interaction must not yield a significant result.
  - Similarly as in other subgroup contexts, lack of power may sensibly lead to a choice of significance level that is larger than conventional, e.g. 0.10.

“Definition 5”

- In order to conclude consistency, tests for individual regions having effects lower than the overall effect must all not yield significant results.
  - Because multiple tests are being performed, using a significance level such as of 0.1 / *s*, where *s* is the number of regions, may be sensible.

Comparing / evaluating the approaches

- The paper demonstrates how to evaluate the probability of claiming consistency for each of these approaches
  - a number of illustrations are presented
  - region sample sizes are addressed.
- These are evaluated unconditionally, as well as *conditionally*, given that the overall treatment effect is significant
  - although for Definitions 4 and 5, the conditional and unconditional probabilities are the same.

Sample results

s=3 ($\alpha = 0.025$, $\delta = 0.25$, $\sigma = 1$)

(f_1, f_2, f_3)	(u_1, u_2, u_3)	Uncond.	Cond.	Uncond.	Cond.	Uncond.	Cond.
$1-\beta$ =0.8, N=252		Definition 1 π =1/3		Definition 2 $b = \delta / s$		Definition3 π =0, α' =0.3	
(1/3,1/3,1/3)	(1,1,1)	67	76	64	76	64	77
(0.2,0.2,0.6)	(1,1,1)	62	69	59	70	56	66
(0.1,0.45,0.45)	(1.9,0.9,0.9)	72	81	67	80	68	80
$1-\beta$ =0.9, N=337		Definition 1 π =1/3		Definition 2 $b = \delta / s$		Definition3 π =0, α' =0.3	
(1/3,1/3,1/3)	(1,1,1)	76	81	72	78	76	82
(0.2,0.2,0.6)	(1,1,1)	69	73	66	72	66	72
(0.1,0.45,0.45)	(1.9,0.9,0.9)	80	85	75	82	79	85

More . . .

s=3 ($\alpha = 0.025$, $\delta = 0.25$, $\sigma = 1$)

(f_1, f_2, f_3)	(u_1, u_2, u_3)	Uncond./Cond.	Uncond./Cond.
$1-\beta = 0.8$, N=252		Def. 4, ε=0.1	Def. 5, α'=0.1/3
(1/3, 1/3, 1/3)	(0.25, 0.55, 2.2)	31	48
(1/3, 1/3, 1/3)	(0.3, 0.3, 2.4)	20	39
(1/4, 1/4, 1/2)	(0.25, 2.65, 0.55)	23	47
(1/4, 1/4, 1/2)	(0.5, 0.7, 1.4)	75	76
$1-\beta = 0.9$, N=337		Def. 4, ε=0.1	Def. 5, α'=0.1/3
(1/3, 1/3, 1/3)	(0.25, 0.55, 2.2)	20	37
(1/3, 1/3, 1/3)	(0.3, 0.3, 2.4)	10	27
(1/4, 1/4, 1/2)	(0.25, 2.65, 0.55)	12	35
(1/4, 1/4, 1/2)	(0.5, 0.7, 1.4)	70	72

Other topics

- Brief discussion of other issues:
  - random effect models
  - non-inferiority studies
  - alternate data structures: binary, time-to-event
  - other approaches: tests for qualitative interaction, range-based tests.

Example

- MRCT is designed to evaluate effect of an investigational drug on HbA1c.
  - 558 patients in 2:1 (active : placebo) allocation
  - α=0.025, >99% power for δ=0.005, σ=0.013
  - 4 regions

- We'd like to determine the minimum regional sample size so that there's an 80% chance of showing consistency, using methods such as described in this paper.

Example

If there is true underlying regional consistency:

	Uncond.	Cond.	Uncond.	Cond.	Uncond.	Cond.
	Def. 1, π=1/4		Def. 2, $b = \delta / s$		Def. 3, α'=0.4, π=0	
$f_1 < f_2 = f_3 = f_4$	0.14	0.13	0.23	0.18	0.09	0.08
$f_1 = f_2 < f_3 = f_4$	0.18	0.17	0.24	0.21	0.13	0.13
$f_1 = f_2 = f_3 < f_4$	0.20	0.20	0.24	0.22	0.16	0.16

Discussion

- Approaches differ in various ways:
  - where they put the "burden of proof" – on a hypothesis of consistency or inconsistency?
    - depends on the context: what are the *implications* of failing a test?; what are the *prior expectations* about potential for regional inconsistency?
  - Hypothesis test based (e.g., Defs. 3-5) versus numerically based (e.g., Defs. 1-2)
    - but *all procedures* have associated error rates (for the numerically based procedures they are just not the basis, and perhaps less apparent).

Discussion

- Statistical properties, and numerical limitations, should always be kept in mind in interpreting results of consistency investigations.
  - False signals may arise by chance, as in other subgroup contexts.
  - Inconsistency may fail to be demonstrated even if true, due to lack of power.

Discussion

- A signal / indication of inconsistency is not the end of this process.
  - Rather, it's an initial step in an investigation to understand why it occurred, and to properly interpret its implications
    - e.g., is the signal of regional difference perhaps due to other identifiable covariates, confounded with region, that are the real effect modifiers?
    - What are the implications for regulatory decisions or medical practice?

Discussion

- Definition of region is obviously *key*
  - but not the focus of this paper.
  - Another PhRMA group workstream is addressing this issue.
  - While the definition should be driven by science and not numbers, the implications for the statistical properties must be kept in mind
    - e.g., the behavior of some statistical approaches may deteriorate more rapidly with increasing number of regions / decreasing region size.

Discussion

- *Pre-specification* is also key.
  - "An investigation of regional consistency in the final trial data will be most meaningful if the consistency definition and methodological approach to be used are clearly prespecified in the protocol."
  - Hung *et al* 2010: *"The additional benefit from pre-specification is to force the trial planners to think through the factors that may have to be considered in defining the regions."*