Subgroup supportive evidence for consistency with the overall population efficacy

Mohamed Alosh, Ph.D.
Division of Biometrics III, OB, OTS, FDA*

BASS XX, November 6, 2013

* This presentation reflects the views of the author and should not be construed to represent FDA's views or policies

Collaboration

I. Mohammad Huque, Ph.D., Office of Biostatistics, FDA
II. Gary Koch, Ph.D., Department of Biostatistics, UNC, Chapel Hill

Outline

I. Introduction: Subgroup analysis and interpretation of study findings
II. Consistency of subgroup findings with those of the overall population:
    a. Subgroup reversal effect due to chance
    b. Treatment-by-subgroup interaction
III. Testing for interaction versus subgroup supportive role for the overall population findings
IV. Criterion on the “least benefited” subgroup findings for support of the overall population and implication on study design
V. Application
VI. Concluding remarks

I. Introduction: Subgroup analysis and interpretation of study findings

- ✓ Clinical trials enroll subjects who are expected to benefit from the treatment under investigation. Although the extent of benefits may vary among subjects, a certain degree of homogeneity in their response is assumed to justify enrolling all subjects who meet the enrollment criteria
- ✓ In reality, treatment effect may vary by subjects' baseline factors defining subgroups, including genomics, biomarkers, etc…

  ➔ It becomes important to evaluate consistency/heterogeneity of treatment effect across subgroups, as this would impact the interpretation of study findings and how treatment would be used
- ✓ Regulatory guidances call for assessing subgroup analysis findings for interpretation of study results

I. Introduction (cont'd):

- ✓ In practice, subgroup analyses are occasionally has been overdone and carried out based on data exploration and not supported by scientific rationale. This has resulted in spurious findings, which has led to negative views by critics of the subgroup analysis
- ✓ However, there are several examples in which subgroup analysis has played a positive role, and others where it has played a negative role, in the development of new therapies
- ✓ Need to have a balanced approach that takes into account the role intended for the subgroup analysis and the level of credibility required for supporting such a role
- ✓ Distinguish among the following objectives for subgroup analysis:

  a. Establishing an efficacy claim

  b. Supporting findings in the overall population

  c. Searching for differential responses through data exploration

  → "hypothesis generating"

I. Introduction (cont'd):

The credibility level of subgroup findings depends on the objective of the analysis and these can be considered based on a range of criteria

Criteria	**Role of subgroup analysis**		
	Establishing efficacy	Supporting study findings	Hypothesis generating
Pre-specification	✓	✓(preferable)	not required
Power	✓	✓(preferable)	not required
Type I error	strong control along with the population at α	$\alpha_s > \alpha$ (preferable)	not required
Other factors…			

II. Consistency of subgroup findings with those of the overall population

- ✓ Subgroup analysis aims to gain insight into the level of consistency or heterogeneity of the treatment effect across subgroups
- ✓ Relative consistency among the subgroups provides evidence that clinical trial findings are robust across the intended patient population, while signs of heterogeneity may be used to inform approval decisions or clinical practice
- ✓ For assessing consistency/heterogeneity in clinical trials that establishes efficacy in the total population, consider:
  a. Subgroups findings that are in the opposite direction to those of the trial
  b. Testing for treatment-by-subgroup interaction

II.a. Observed reversal subgroup findings

- ✓ It is expected that due to chance alone, partitioning the total population into subsets would result in some subgroup findings to be in the opposite direction to those of the overall population
- ✓ Learning of the extent of chance findings should temper the tendency to over-interpret subgroup findings
- ✓ The probability of chance findings can be calculated under some simplifying assumptions, including: similar treatment effect across subgroups, equal allocation for subgroups, normally distributed treatment response
- ✓ This probability can be calculated for a trend or for statistically significant findings in the subgroup and overall population. In addition, it can be calculated for one or more factors, with different levels, defining the subgroup (there are several publications in this regard)

II.a. Observed reversal subgroup findings

- ✓ The overall conclusion is that chance findings increase with the increase in the number of subgroups and with unequal allocation of subgroups; such findings are much higher for trend compared to significant results
- ✓ Aside from the chance factor, observed heterogeneity in treatment effect across subgroups can be also truly related to subject characteristics
- ✓ The challenge in subgroup analysis is to separate chance findings from those due to true heterogeneity as the consequences of the two are different. True heterogeneity impacts interpretation of study findings and, consequently, the prescription population
- ✓ Consider statistical testing to account for chance factors. Testing for treatment-by-subgroup interaction is commonly used in subgroup analysis

II.b. Testing for treatment-by-subgroup interaction

- ✓ Testing for treatment-by-subgroup interaction considers differences in treatment effect between two or more subgroups
- ✓ Distinguish between quantitative vs. qualitative interactions
- ✓ It is well reported in the literature that clinical trials are generally underpowered for detecting interactions, as they are usually powered for establishing efficacy claims in the total population and not for detecting interaction
- ✓ The extent of power for interaction depends on the magnitude of treatment effect which the test is aiming to detect, allocation of subjects in subgroups and number of categories for the factor defining the subgroups

III. Testing for interaction and supportive role for subgroup

- ✓ Testing for interaction on its own might not address the question of whether a subgroup finding supports those of the overall study findings since:
  - A non-significant interaction does not imply the absence of the interaction as the study is not powered for this test. Additionally, this case usually is not of concern in practice
  - While a significant interaction is indicative of a difference between subgroups, a significant difference can be driven by the relative sizes of the subgroups, precision of estimates, etc…
  - Thus testing for interaction might not assess the extent of benefits for a subgroup, or whether subgroup findings support those of the overall population. Further, it does not consider other factors to weigh the benefits and risks of the treatment for the various subgroups for determining the prescription population

III. Testing for interaction and supportive role for subgroup

- ✓ When "substantial" heterogeneity in treatment effect is observed across subgroups, a significant treatment effect in the overall population can be driven by a favorable subgroup and without benefits for other subgroups
- ✓ In principle, treatment should be limited to the subgroups whose benefits outweigh their risks (adverse events). Thus, for such a trial the concern is identification of the subgroup(s) who truly benefit from the treatment
- ✓ Testing for treatment-by-subgroup interaction, as it is related to comparing effect between subgroups, may not be sufficient to address this concern
- ✓ Obviously, one should focus on the "least benefited" subgroup(s), whether the trial was designed with a targeted subgroup in mind or not

IV. Criterion on the "least benefited" subgroup findings to support those of the overall population

- ✓ What level of evidence is expected for a subgroup finding to be supportive of those of the overall population ?
- ✓ Koch and Schwartz (2013)* noted that for this supportive role, consideration of the Type II error rate is more relevant than the Type I error rate. By taking into account the applicable sample size, they suggested that, for a study powered at 0.90, the 1-sided *p*-value for the subgroup should be compared with: $\alpha_s \approx 0.05$ for a subgroup with substantial majority (≥0.70), $\alpha_s \approx 0.25$ for a subgroup with clear minority (≤0.30), and $\alpha_s \approx 0.15$ when the subgroup includes about half of the patients

* Koch and Schwartz (2013): An overview of statistical planning to address subgroups in confirmatory clinical trials (to appear)

IV. Criterion on the "least benefited" subgroup findings to support those of the overall population

- ✓ In the same vein, Alosh and Huque (2013)* noted that efficacy results for the "least benefited" subgroup meet, at a minimum, a threshold parameter, α_c, called the consistency criterion, which can be determined based on clinical considerations taking into account adverse events, toxicity and others
- ✓ Obviously, a more flexible approach can be built by taking advantage of the two proposals; that is, by considering α_s, which takes into account the relative size of the subgroup, and then modify it, if needed, to take into account toxicity and adverse events and other clinical inputs

* Alosh and Huque (2013). Multiplicity considerations for subgroup analysis subject to consistency constraints. Biometrical Journal, 55 (3) 444-462

IV. Supportive role of the "least benefited" subgroup and trial design

- ✓ When substantial heterogeneity (quantitative interaction) across subgroups is expected, either based on scientific rationale and/or early clinical trials, an appropriately designed trial can have the objectives of establishing efficacy claims in the overall population and/or in the pre-specified targeted subgroup
- ✓ The design for such trials should consider the following issues for interpretation of study findings:
  - a. Multiplicity considerations
  - b. Subgroup power and subgroup enrichment
  - c. Level of support expected from the "least benefitted" subgroup to the overall study findings, as this would impact the size of the subgroup

V. Application

- Consider a Phase 3 trial that examined the long-term survival advantage for pegylated liposomal doxorubicin (PLD) compared with topotecan (Top) in women with recurrent and refractory epithelial ovarian cancer (Gordon et al. 2004). Patients were stratified prospectively according to the response to initial platinum-based chemotherapy. Table 2 summarizes the efficacy results from Gordon et al.(2004, *J. Clinical Epidem*. 63:1298-1304)
-

	PLD	Topotecan	Hazard Ratio 95% C.I.	p-value
Overall population (n)	(239)	(235)		
Median survival time (weeks)	62.7	59.7	1.216 (1.000, 1.478)	0.05
Platinum sensitivity				
Platinum sensitive (n_1=109, n_2=110)	107.9	70.1	1.432 (1.066-1.923)	0.017
Platinum refractory (n_1=130, n_2= 125)	NR	NR	1.069 (0.823-1.387)	0.618

n_1 and n_2 are the number of patients in the PLD and topotecan subgroups, respectively.

NR: not reported

V. Application

- ✓ With significant results for PLD compared to Top in the overall population and with different results for the two subgroups, the issue is whether results for the platinum refractory (the least benefited) subgroup support those of the overall population, as this would impact the prescription population. For this we consider:

(i) Testing for interaction:

Using the reported *p*-values one can get the corresponding Z-statistics for the platinum sensitive subgroup (Z_s) and platinum refractory subgroup (Z_c). Then by using the equation:

$$Z_{\text{int}} = \sqrt{1-K}Z_s - \sqrt{K}Z_c = (0.538)^{1/2}\ 2.3867 - (0.462)^{1/2}\ 0.4987 = 1.4117$$

with (1-*K*) being the proportion of subjects in the least benefited subgroup, which leads to: $p_{\text{int}} = 0.079$; which is not significant at the 1-sided level of 0.025 or 0.05 (working with the log HR resulted in $p_{\text{int}} = 0.073$)

This suggest the test treatment could be prescribed for both subgroups

V. Application (cont'd)

(ii) Support criterion proposed by Koch and Schwartz (2013) denoted by α_s. By applying Koch and Schwartz equation (8):

$$Z_{\alpha_S} = (n_S / n)^{0.5}(Z_\alpha + Z_\beta) - Z_\beta = 1.0626,$$

Then for study powered at 0.80, with n_s/n =0.462, we get α_s=0.144, 1-sided. Thus with *p*-value for this subgroup =0.309 > α_s, we conclude that the treatment effect in this subgroup is not in "harmony" with that of the overall population, in contrast to the interaction test

(iii) The consistency criterion,α_c, is determined based on clinical considerations taking into account the relative safety and toxicity of the two treatments, which they are different according to Gordon et al. (2004)

VI. Concluding remarks

i. Assessing whether findings of the “least benefited” subgroup are supportive of those of the overall population is important for interpreting the overall study findings and determining how the treatment will be used

ii. Evaluation of subgroup chance findings and testing for treatment-by-subgroup interaction can be useful to address certain aspects of heterogeneity across subgroups, but these might not be sufficient for determining whether a subgroup finding is supportive of the overall study findings

iii. Consider proposals for setting a “threshold level” for findings of the “least benefited” subgroup to support those of the overall population

iv. The above threshold level can take into account the relative size of the “least benefited” subgroup as well safety considerations based on clinical inputs

V. Concluding remarks

v. The relative size of the “least benefited” subgroup should be considered at the design stage to aid in interpreting study findings. This is in addition to ensuring that the trial designed with establishing efficacy claim in the targeted subgroup have a sufficient number of subjects in its targeted subgroup

vi. In practice, reasonable judgment should be exercised on selecting the number of subgroups, their sizes, and expected support levels; as otherwise, the conduct of the clinical trial can be impractical.