Subgroup Analysis in Regulatory Decision Making

Lilly Q. Yue and Heng Li

Division of Biostatistics, Center for Devices and Radiological Health,

U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Keyword: Subgroup analysis; Interpretation and reporting; Biomarker; Regulatory decision making

Abstract

This presentation discusses regulatory and statistical issues with subgroup analysis in regulatory decision making, for medical products ranging from therapeutic treatments to companion diagnostics. Such issues include interpretation and reporting of completed clinical trials and design of new studies where subgroup differences are expected.

1. Introduction

Patients in a clinical trial usually have different demographic, genomic, and disease characteristics, and the medical products under study may be safe and effective in certain types of patients, but ineffective or harmful in others. Therefore, subgroup analyses based on patients' baseline factors are routinely conducted as a formal component of the efficacy and safety assessment for regulatory submissions. Typically, the objective of the subgroup analysis is to investigate consistency/heterogeneity of the treatment effect across subgroups. As such, subgroup analysis plays a crucial role in the interpretation of the clinical trial findings and regulatory decision making. This presentation provides the regulatory basis of subgroup analysis, discusses issues with interpreting subgroup findings from completed trials and designing studies where subgroup differences are expected, briefly points out other issues to consider, and offers take-home messages.

2. Regulatory Basis of Subgroups Analysis

After analyses on the overall patient population have been conducted, additional analyses are routinely performed for subgroups based on demographic, disease severity and other relevant characteristics, to meet regulatory requirements. The Code of Federal Regulations [21 CFR 314.50(d)(5)(v)], *Content and Format of an Application – Clinical Studies Section (NDA)* states "The effectiveness data shall be presented by gender, age, and racial subgroups and shall identify any modifications of dose or dose interval needed for specific subgroups. Effectiveness data from other subgroups of the population of patients treated, when appropriate, such as patients with renal failure or patients with different levels of severity of the disease, also shall be presented." The International Conference on Harmonization *Statistical Principles for Clinical Trials* (ICH E9) discusses treatment-by-subgroup interaction and indicates "In some cases such interactions are anticipated or are of particular prior interest (e.g., geriatrics); hence a subgroup analysis or a statistical model including interactions is part of the planned confirmatory analysis. In most cases, however, subgroup or interaction analyses are exploratory and should be clearly identified as such; they should explore the uniformity of any treatment effects found overall."

U.S. FDA has issued a number of guidance documents referencing subgroup analyses (gender, race, age), including:

- Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products – Content and Format (CBER/CDER)
- Design Considerations for Pivotal Clinical Investigations for Medical Devices (CDRH/CBER)
- Study and Evaluation of Gender Differences in the Clinical Evaluation of Drugs (CDER)
- Evaluation of Sex Differences in Medical Device Clinical Studies - (CDRH/CBER)
- Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products. FDA draft guidance, issued on Dec. 2012 (CDER/CBER/CDRH) www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf
- Evaluation and Reporting of Age, Race, and Ethnicity Data in Medical Device Clinical Studies. Draft guidance, issued on June 20, 2016 (CDRH/CBER)
- Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product - Draft Guidance issued on July 15, 2016 (CDRH/CDER/CBER) http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM510824.pdf

In addition, an FDA White Paper, *Statistical Considerations on Subgroup Analysis in Clinical Trials,* written by FDA Subgroup Analysis Working Group, provides a good reference (Alosh et. al., 2015).

3. Interpreting Subgroup Findings from Completed Trials

There are different ways to define subgroups. For example, subgroups can be defined for reporting purposes based on subject demographics (e.g., gender, race, or age) where treatment differences are not expected but may be observed. Subgroups may be defined based on the geographic location of the investigative site (e.g., U.S. vs. outside U.S.) or the investigator conducting the study, in which case treatment differences are expected due to patient characteristics or medical practice, etc. Subgroups based on genomic markers may be of particular interest with regard to personalized medicine, usually with the expectation that one or more "marker-positive" subgroups will experience more benefit or less harm than others.

There are two scenarios to consider with a completed clinical trial: 1). A clinical trial met its objectives in the intended overall patient population, in terms of statistical and clinical significance on the key hypotheses regarding primary efficacy/safety endpoints; 2). A trial didn't meet its objectives in the intended overall patient population.

3.1 **Successful trial for the intended overall patient population**

When significant treatment effect is detected in the overall study population, the objective of subgroup analysis is typically to gain insight into the level of consistency/heterogeneity of the treatment effect across the subgroups. Relative consistency among the subgroups may provide evidence that the findings are robust over the intended patient population, while signs of heterogeneity may be used to inform clinical practice.

However, the greater challenge in interpreting subgroup analysis findings arises when subgroup analyses were not included as part of a pre-specified and multiplicity-adjusted statistical plan. Some observed subgroup differences could be due to true heterogeneity and others are due to chance alone. A key challenge is how to distinguish true heterogeneity from random chance, especially when the number of subgroups is large and subgroup sample sizes are small. When a large number of subgroup analyses are conducted, and particularly when the treatment effect is not large, the probability is high that some subgroups will have treatment effect estimates in the opposite direction.

The consistency/heterogeneity of treatment effect across subgroups is usually assessed through statistical testing of treatment-by-subgroup interaction, quantitative or qualitative. With quantitative interactions, the magnitude of the treatment effect may vary across subpopulations, but the subgroup-specific treatment effects are in the same direction. Suspected quantitative interactions usually do not lead to restrictions on the population for which a product can be deemed efficacious; however, variations in observed treatment effect among subgroups need to be reported. With qualitative interactions, the treatment difference is nonzero in at least one subgroup but is zero or goes in the opposite direction in at least one other. In such a case, considerable concerns arise, as this implies the investigational treatment is no better than or even worse than the control for certain subgroups. It should be noted that in practice, clinical trials usually have low power to detect potentially important treatment-by-subgroup interactions, and failure to detect a significant interaction does not imply the absence of an important interaction. It is also known that observed subgroup difference, e.g., by sex, may be explained by other characteristics, e.g., body size.

If the treatment effect is consistent across the subgroups of interest, or the treatment effect varies in magnitude across the subgroups but the treatment effect is still favorable across the subgroups, the subgroup analyses can increase the confidence in the robustness of the study results in the overall patient population, and an appropriate regulatory decision would be to approve the product for the whole population. When there is significant treatment-by-subgroup interaction, regulatory approval decision may be restricted to a certain subgroup of the studied population.

For the primary efficacy/safety endpoints (and in some cases, important secondary endpoints as well), summary statistics should be reported for subgroups of interest. In particular, demographic subgroup analyses are expected as part of regulatory submissions. It is standard that one would report results by age, gender, and racial subgroups in product labeling. In 2012, legislation was passed that required FDA to report how subgroups were addressed in both labeling and in review across all its medical product areas. In 2014, FDA responded with an action plan to enhance the collection and availability of demographic subgroup data (Food and Drug Administration, 2014a).

3.2 **Failed trials for the intended overall patient population**

If a study didn't meet its objectives in the intended overall patient population, but some exploratory subgroup analyses identified one or more subgroups with apparent benefit, the product would likely not be approved for the subgroup(s), due to the concerns with the limited data and the post-hoc nature of the significant findings. The significant findings are considered as hypothesis generating that need to be confirmed by one or more new studies (Alosh et. al. 2015).

4. Designing Studies Where Subgroup Differences are Expected

In some studies, it is reasonable to expect that there is likely a meaningful qualitative treatment-by-subgroup interaction of treatment effect, based on previous studies or disease science. In such cases, important differences in the benefit-risk profile of a medical product are plausible or anticipated across subgroups of interest, and clinical trials could be designed to lead to a claim either for the overall patient population or a pre-specified target.

In such studies, planning is critically important. To ensure proper control of the Type I error rate, the multiplicity issues need to be appropriately handled, as there are two alternative paths of success (one for the overall study population and one for a targeted subgroup). Study power for both overall study population and the targeted subgroup should be considered, and a sufficient number of patients in the subgroup are needed for reliable subgroup results. In some situations, subgroup enrichment design (over-represented) may be reasonable to increase the power for a targeted subgroup, but the study findings for the overall patient population should be carefully interpreted. In addition, to better understand treatment effect for overall patient population, an evaluation of the complementary subgroup should be performed.

5. Subgroups Identified by Biomarker

Biomarker (Biological Marker) is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Biomarkers Definition Working Group, 2001).

In some disease processes, certain biomarkers may be used to identify patients or tumors that are more likely to respond to a therapeutic treatment or alternatively less likely to experience certain adverse reactions. For genetic markers that are not readily observable by subjects or investigators, a diagnostic medical test, e.g., in vitro diagnostic assay test kit or an imaging system, is used to measure the characteristics for classifying subjects into subgroups. As stated in the FDA Draft Codevelopment Guidance (2014c), "An in vitro companion diagnostic device ... is an in vitro diagnostic device (IVD) that provides information that is essential for the safe and effective use of a corresponding therapeutic product.", and " …. IVD companion diagnostics are, by definition, essential for the safe and effective use of a corresponding therapeutic product and may be used to: 1) identify patients who are most likely to benefit from the therapeutic product; 2) identify patients likely to be at increased risk for serious adverse reactions as a result of treatment with the therapeutic product; ….". It is usually expected that one or more "marker-positive" subgroups will experience more benefit or less harm from a particular therapy than others.

A companion in vitro diagnostic device and its corresponding therapeutic product are reviewed and approved according to applicable regulatory requirements (Food and Drug Administration 2014b). The therapeutic product is evaluated in one or more subgroup(s) defined by the companion diagnostic. The companion diagnostic is reviewed for adequate performance characteristics, including its ability to correctly classify subjects into subgroups. Subgroup misclassification results from incorrect test results used to determine the biomarker status, false positives or false negatives.

Misclassification error can have serious implications on the composition of patients enrolled into a clinical study. If there is evidence that a therapy may be beneficial in the test positive subgroup and harmful in the test negative subgroup, subjects with false-positive results may be harmed by the therapy, and subjects with false-negative results may be deprived of beneficial therapy. False-positive results could lead to underestimation of treatment effect size, and false-negative results could result in underestimation of the proportion of subjects who are more likely to respond. Therefore, prospective validation, analytical and clinical validation, of biomarker assay measured or determined by an IVD companion diagnostic is critical (evaluated by CDRH) (Pennello, 2013, Food and Drug Administration, 2014c).

FDA Draft Codevelopment Guidance (Food and Drug Administration, 2014c) discusses two types of biomarker-based Clinical Trial Design: Trial A is designed to evaluate treatment and marker effects, and their interaction, by stratifying randomization based on marker status, as determined by an IVD. Trial B is designed to evaluate treatment effects in a targeted population by selecting only those who are test-positive. Other biomarker-based clinical trial designs include marker strategy design, adaptive enrichment design, etc. In such biomarker-based clinical trials, challenges with multiplicity issues arise.

Multiplicity issues associated with the control of Type I error rate may arise due to the use of high-dimensional genomic data, multiple candidate biomarkers, or possible multiple assays for the same biomarker (Food and Drug Administration, 2014c). Multiplicity issues could be associated with multiple tests of treatment effect of a therapy in overall patient population, marker-positive subgroup, and marker -negative subgroup. There are multiple approaches proposed to handle the multiplicity issues and one example is Marker Sequential Test (MaST) design (Freidlin et. al. 2014). It sequentially tests the treatment effect in the subgroups and the overall population, while controlling the relevant type I error rate. First, the marker-positive subgroup is tested at a reduced significant level α_1, $\alpha_1 < \alpha$. If it is significant, the marker-negative subgroup is tested at the level α; If it is not significant, the overall population is tested at the level $\alpha_2 = \alpha - \alpha_1$.

6. Concluding Remarks

For a completed clinical trial, subgroup analyses should be appropriately performed, interpreted and reported. The treatment-by-subgroup interaction test or modeling should be conducted and descriptive analysis (tables and plots) needs to be performed. With anticipated subgroup difference, a clinical trial could be appropriately designed to make a claim for either overall patient population or a pre-specified target subgroup. In such a case, design strategy, study power and multiplicity control should be carefully considered. Furthermore, benefit/risk balance should be assessed for subgroups. In addition, other issues to consider include whether different non-inferiority margins for different subgroups are needed for an active control non-inferiority trial, how to investigate safety across subgroups, especially when the serious adverse event is of low frequency, and how to incorporate the safety information in the subgroup benefit/risk assessment.

Acknowledgement

The authors would like to thank Drs. Gene Pennello and Tinghui Yu for their help with the presentation.

References

1. Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S., Yue, L. (2015) Statistical Considerations on Subgroup Analysis in Clinical Trials. *Statistics in Biopharmaceutical Research,* 7(4), 286-303.

2. Biomarkers Definition Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology Therapeutics*, 69(3), 89–95.

3. Food and Drug Administration (2014a), "FDA Action Plan to Enhance the Collection and Availability of Demographic Subgroup Data," available at: http://www.fda.gov/downloads/RegulatoryInformation/Legislation/FederalFoodDrugandCosmeticActFDCAct/SignificantAmendmentstotheFDCAct/FDASIA/UCM410474.pdf.

4. Food and Drug Administration (2014b), "Guidance for Industry and Food and Drug Administration Staff- In vitro Companion Diagnostic Devices," available at:http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM262327.pdf.

5. Food and Drug Administration (2014c), Principles for Codevelopment of an In Vitro Companion Diagnostic Device with a Therapeutic Product - Draft Guidance issued on July 15, 2016 available at http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM510824.pdf

6. Freidlin, B., Korn,E.L., Gray, R. (2014) Marker Sequential Test (MaST) Design. *Clinical Trials*, 11, 19-27.

7. Pennello, G. (2013) Analytical and Clinical Evaluation of Biomarkers Assays: When Are Biomarkers Ready for Prime Time? *Clinical Trials*, 10, 666–676.