1. Analyses and Displays Associated with Measures of Central Tendency – Focus on Vital Sign, Electrocardiogram, and Laboratory Analyte Measurements in Phase 2-4 Clinical Trials and Integrated Submission Documents

Version 1.0
Created 10 October 2013

A White Paper by the Computational Science Symposium Development of Standard Scripts for Analysis and Programming Working Group

2. Table of Contents

Section **Page**

3. Revision History

Version 1.0 was finalized 10 October 2013.

4. Purpose

The purpose of this white paper is to provide advice on displaying, summarizing, and/or analyzing measures of central tendency, with a focus on vital sign, electrocardiogram (ECG) quantitative findings, and laboratory analyte measurements in Phase 2-4 clinical trials and integrated submission documents. The intent is to begin the process of developing industry standards with respect to analysis and reporting for measurements that are common across clinical trials and across therapeutic areas. In particular, this white paper provides recommended tables and figures for measures of central tendency (e.g., means and medians) for a common set of safety measurements. Separate white papers address other types of data or analytical approaches (e.g., percentages of subjects meeting specified criteria).

This advice can be used when developing the analysis plan for individual clinical trials, integrated summary documents, or other documents in which measures of central tendency are of interest. Although the focus of this white paper pertains to specific safety measurements (vital signs, ECG quantitative findings, and laboratory analyte measurements), some of the content may apply to other measurements (e.g., different safety measurements and efficacy assessments). Similarly, although the focus of this white paper pertains to Phase 2-4, some of the content may apply to Phase 1 or other types of medical research (e.g., observational studies).

Development of standard Tables, Figures, and Listings (TFLs) and associated analyses will lead to improved standardization from collection through data storage. (You need to know how you want to analyze and report results before finalizing how to collect and store data.) The development of standard TFLs will also lead to improved product lifecycle management by ensuring reviewers receive the desired analyses for the consistent and efficient evaluation of patient safety and drug effectiveness. Although having standard TFLs is an ultimate goal, this white paper reflects recommendations only and should not be interpreted as "required" by any regulatory agency.

Detailed specifications for TFL or dataset development are considered out-of-scope for this white paper. However, the hope is that specifications and code (utilizing SDTM and ADaM data structures) will be developed consistent with the concepts outlined in this white paper, and placed in the publicly available Standard Scripts Repository.

5. Introduction

Industry standards have evolved over time for data collection (CDASH), observed data (SDTM), and analysis datasets (ADaM). There is now recognition that the next step would be to develop standard TFLs for common measurements across clinical trials and across therapeutic areas. Some could argue that perhaps the industry should have started with creating standard TFLs prior to creating standards for collection and data storage (consistent with end-in-mind philosophy), however, having industry standards for data collection and analysis datasets provides a good basis for creating standard TFLs.

The beginning of the effort leading to this white paper came from the FDA computational statistics group (CBER and CDER). The FDA identified key priorities and teamed up with the Pharmaceuticals Users Software Exhange (PhUSE) to tackle various challenges using collaboration, crowd sourcing, and innovation (Rosario, et. al. 2012). The FDA and PhUSE created several Computational Science Symposium (CSS) working groups to address a number of these challenges. The working group titled "Development of Standard Scripts for Analysis and Programming" has led the development of this white paper, along with the development of a platform for storing shared code. Most contributors and reviewers of this white paper are industry statisticians, with input from non-industry statisticians (e.g., FDA and academia) and industry and non-industry clinicians. Hopefully additional input (e.g., other regulatory agencies) will be received for future versions of this white paper.

There are several existing documents that contain suggested TFLs for common measurements. However, many of the documents are now relatively outdated, and generally lack sufficient detail to be used as support for the entire standardization effort. Nevertheless, these documents were used as a starting point in the development of this white paper. The documents include:

- ICH E3: Structure and Content of Clinical Study Reports
- Guideline for Industry: Structure and Content of Clinical Study Reports
- Guidance for Industry: Premarketing Risk Assessment
- Reviewer Guidance. Conducting a Clinical Safety Review of a New Product Application and. Preparing a Report on the Review
- ICH M4E: Common Technical Document for the Registration of Pharmaceuticals for Human Use - Efficacy
- ICH E14: The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential For Non-Antiarrhythmic Drugs
- Guidance for Industry: ICH E14 Clinical Evaluation of QT/QTc. Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs

The Reviewer Guidance is considered a key document. As discussed in the guidance, there is generally an expectation that analyses of central tendency are conducted for vital signs, ECGs quantitative findings, and laboratory analyte measurements. The guidance recognizes value to both analyses of central tendency and analyses of outliers or shifts from within reference limits to outside reference limits (below lower reference limit or above upper reference limit). We assume both will be conducted for safety signal detection. This white paper covers the central

tendency portion, with the expectation that an additional TFL or TFLs will also be created with a focus on outliers or shifts.

6. General Considerations

6.1. All Measurement Types

6.1.1. Importance of Visual Displays

Communicating information effectively and efficiently is crucial in detecting safety signals and enabling decision-making. Current practice, which focuses on tables and listings, has not always enabled us to communicate information effectively since tables and listings may be very long and repetitive. Graphics, on the other hand, can provide more effective presentation of complex data, increasing the likelihood of detecting key safety signals and improving the ability to make clinical decisions. They can also facilitate identification of unexpected values.

Standardized presentation of visual information is encouraged. The FDA/Industry/Academia Safety Graphics Working Group was initiated in 2008. The working group was formed to develop a wiki and to improve safety graphics best practice. It has recommendations on the effective use of graphics for three key safety areas: adverse events, ECGs and laboratory analytes. The working group focused on static graphs, and their recommendations were considered while developing this white paper. In addition, there has also been advancement in interactive visual capabilities. The interactive capabilities are beneficial, but are considered out-of-scope for this version of the white paper.

The recommendations in this white paper include the broad use of box plots. See Figure 6.1 for an explanation of what a box plot provides.

6.1.2. P-values and Confidence Intervals

There has been ongoing debate on the value or lack of value for the inclusion of p-values and/or confidence intervals in safety assessments (Crowe, et. al. 2009). This white paper does not attempt to resolve this debate. As noted in the Reviewer Guidance, p-values or confidence intervals can provide some evidence of the strength of the finding, but unless the trials are designed for hypothesis testing, these should be thought of as descriptive. Throughout this white paper, p-values and measures of spread are included in several places. Where these are included, they should not be considered as hypothesis testing. If a company or compound team decides that these are not helpful as a tool for reviewing the data, they can be excluded from the display.

Some teams may find p-values and/or confidence intervals useful to facilitate focus, but have concerns that lack of "statistical significance" provides unwarranted dismissal of a potential signal. Conversely, there are concerns that due to multiplicity issues, there could be over-interpretation of p-values adding potential concern for too many outcomes. Similarly, there are concerns that the lower- or upper-bound of confidence intervals will be over-interpreted. (A mean change can be as high as x causing undue alarm.) It is important for the users of these TFLs to be educated on these issues.

6.1.3. Conservativeness

The focus of this white paper pertains to clinical trials in which there is comparator data. As such, the concept of "being conservative" is different than when assessing a safety signal within an individual subject or a single arm. A conservative approach for defining outcomes, from a single arm perspective, is one that would lead to a higher number of subjects reaching a threshold or a larger change value compared to other approaches. However, a conservative approach for defining outcomes may actually make it more difficult to identify safety signals with respect to comparing treatment with a comparator (see Section 7.1.7.3.2 in the Reviewer Guidance). Thus, some of the outcomes recommended in this white paper may appear less conservative than alternatives, but the intent is to propose methodology that can identify meaningful safety signals for a treatment relative to a comparator group.

6.1.4. Measurements After Stopping Study Medication

Measurements collected after stopping medications under study (e.g., treatment under study and comparators) are common for various reasons. In some cases, "follow-up" phases are included to monitor patients for a period of time after study medication is stopped. Additionally, study designs where keeping subjects in a study (for the entire planned length of time) after deciding to stop medication early are becoming more popular. In these cases, subjects can be off study medication for an extended period of time.

Measurements post study medication can also arise not by design. For example, a subject can decide to stop study medication at any time, and then later attend the planned visit and the planned measurements are obtained. There is currently no standard approach on how to handle safety assessments post study medication. Some guidances contain advice on how long to collect safety measurements post study medication (e.g, 30 days post or, x half-lives). Any advice or decisions related to the collection of safety measurements post study medication should not be confused with how to include such data in displays and/or analyses. It is extremely important to document within the database for analysis the best estimate of the last date study treatment was taken as well as dates on which all numerical safety data were collected so that an accurate determination can be made of time of data collection relative to last dose of medication.

We recommend that the TFLs in this white paper generally exclude measurements taken during a "follow-up" phase. Separate TFLs can be created for the follow-up phase and/or the treatment and follow-up phases combined. We also recommend that the TFLs in this white paper exclude measurements taken after the visit which is considered the "study medication discontinuation" visit. In the study designs which keep subjects in a study for the entire planned length of time even after stopping medication, separate TFLs can be created for the "off-medication" time and/or the treatment and "off-medication" times combined. This enables the researcher to distinguish between drug-related safety signals versus safety signals that could be more related to discontinuing a drug (e.g., return of disease symptoms, introduction of a concomitant medication, and/or discontinuation- or withdrawal-effects of the drug) or due to subsequent therapy. We assume it is important to distinguish among these. Generally, at least some TFLs that include data from follow-up phases and/or "off-medication" time will be required, but not

usually as many as done for during treatment and not necessarily in the same format as provided in this white paper. For some compounds (e.g., compounds with a long half-life compared to the duration of the study, compounds used for a very short time like antibiotics), a more complete set of TFLs including such data may be required. The ease of interpretation from such TFLs will vary depending on the compound, disease, and/or design aspects, such as, the half-life of the compound, likelihood of taking alternative therapy, allowed concomitant medications during the observation period, etc. Analyses of numerical safety data subsequent to treatment discontinuation can serve two purposes related to safety assessment. First, if there has been an adverse change in central tendency of a given parameter during treatment and there is a return toward baseline on discontinuation (positive dechallenge on a treatment group [central tendency] basis) this is additional evidence of drug causation. Second, this pattern of resolution or improvement would provide important clinical information about the course of a potential adverse reaction, that it was transient. A persistant pattern of central tendency change following discontinuation might signal a more ominous pattern for the potential adverse reaction. Analyses appropriate to these considerations are outside the scope of this white paper.

For the third example (a subject decides to stop study medication at any time and then later attends the planned visit to obtain the planned measurements), we recommend measures taken at the study medication discontinuation visit are included. Although some subjects may be off medication, the time is generally short in these situations. For this example, the inclusion of such measurements may more accurately reflect the safety profile of a compound versus their exclusion. In study designs with a long period of time between visits, an alternative approach may be warranted.

6.1.5. Measurements at a Discontinuation Visit

When creating displays or conducting analyses over time, how to handle data collected at discontinuation visits should be specified. Since a subject's discontinuation visit isn't always aligned with planned timing, it's not obvious whether to include these measurements in displays or analyses over time. Such measurements are "planned" per protocol, but not consistent with the planned timing. We generally recommend including measures taken at the discontinuation visit toward the next timepoint. The inclusion of such measurements may more accurately reflect trends over time for the compound than their exclusion. In study designs with a long period of time between visits, an alternative approach may be warranted.

6.1.6. Screening Measurements versus Special Topics

The focus of this white paper pertains to measurements as part of normal safety screening. For many compounds, some measurements are relevant to addressing a-priori special topics of interest. In these cases, it is possible that additional TFLs and/or different TFLs are warranted. TFLs designed for special topics are out-of-scope for this white paper. In addition, it is possible that additional TFLs are warranted when a safety signal is identified using the TFLs recommended in this white paper and/or the TFLs that focus on outliers or shifts (separate white paper). Additional TFLs that would be considered "post-hoc" for further investigation are considered out-of-scope.

6.1.7. Number of Therapy Groups

The example TFLs show one treatment arm versus comparator in this version of the white paper. Most TFLs can be easily adapted to include multiple treatment arms or a single arm.

6.1.8. Multi-phase Clinical Trials

The example TFLs show one treatment arm versus comparator within a controlled phase of a study. Discussion around additional phases (e.g., open-label extensions) is considered out-of-scope in this version of the white paper. Many of the TFLs recommended in this white paper can be adapted to display data from additional phases.

6.1.9. Integrated Analyses

For submission documents, TFLs are generally created from using data from multiple clinical trials. Determining which clinical trials to combine for a particular set of TFLs can be complex. Section 7.4.1 of the Reviewer Guidance contains a discussion of points to consider. Generally, when p-values are computed, adjusting for study is important. Creating visual displays or tables in which timepoints or treatment comparisons are confounded with study is discouraged. Understanding whether the overall representation accurately reflects the review across individual clinical trial results is important.

6.2. Laboratory Analyte Measurements

The following topics generally pertain to laboratory analyte measurements only. However, there could be situations in which a topic applies to another measurement type. In these cases, the discussion below may or may not apply.

6.2.1. Planned versus Unplanned Measurements

One topic that tends to be unique to safety (laboratory analyte measurements in particular) is the collection of unplanned measurements. Unplanned safety measurements can arise for various reasons. During a study, the clinical investigator sometimes orders a repeat test or "retest" of a laboratory test especially if he/she has received an unexpected value. The investigator may also request the patient return for a "follow-up visit" due to clinical concerns. In general, retests are repeat tests performed because an initial test result had an unexpected value. The repeat result may either confirm the initial test results, or (less commonly) suggest that a laboratory error occurred in the case of the initial result. Retests are often performed to verify that the action taken by the investigator (e.g., changing the dose of study drug as allowed by the protocol) has the desired effect (e.g., test results have returned to within reference limits). If such retests are conducted until desired measurement results have been reached, analyses from baseline to last observation, for example, would be biased toward "normality". Thus, we recommend including only planned measurements when creating displays or conducting analyses over time and when assessing change from baseline to endpoint. However, we recommend including planned and unplanned measurements when assessing maximum and minimum changes as these are intended to focus on the most extreme changes. (Including planned and unplanned measurements will also be recommended for analyses that focus on outliers or shifts, which will be a topic for a

separate white paper.) Of note, these recommendations can only be implemented if planned and unplanned measurements can be distinguished via data collection and identified in the analysis dataset. Using logic to compare laboratory collection dates with visit dates is possible, but doesn't always result in an accurate assessment of planned versus unplanned.

6.2.2. Transformations of Data

Another topic that tends to be debated for summaries or analyses of laboratory measurements is whether to transform the data, and if so, the appropriate transformation. Many laboratory measurements are not normally distributed. Generally, for routine safety screening, we recommend summarizing and analyzing laboratory measurements in the original scale. Reviewing data in the original scale is generally preferred by medical colleagues. We recommend descriptive statistics for laboratory measurements including mean, standard deviation, as well as minimum, q1, median, q3 and maximum. When providing p-values or confidence internals for a mean, the central limit theorem states that in many situations, the sample mean varies normally if the sample size is reasonably large.

Exceptions include laboratory measurements that are of special concern for a compound (considered out-of-scope for this white paper as discussed in Section 6.1.6). In these cases, special distributional considerations may be warranted. Other exceptions include cases when a transformation may facilitate the readability in a visual display.

Of note, if data are transformed, the transformation is really changing your response variable and null hypothesis. Thus, any inferences (e.g. CI's & hypothesis tests) made, are on the transformed data rather than the original scale. Also, when data are transformed, the transformed data should reside in the analysis dataset (e.g., ADaM).

6.2.3. Units

Units for laboratory measurements may be expressed in U.S. conventional units or Système International (SI) units. Of the reviewed regulatory guidances, a preference for SI units in the Summary of Clinical Safety is noted (ICH M4E). However, since the TFLs have multiple customers including internal physicians writing summaries of the data and providing oversight for safety in general, providing TFLs in both units is frequently warranted.

6.2.4. Central Versus Local Laboratories

In recent years, most large studies have utilized a central laboratory to ensure consistency in laboratory assessments across institutions. However, there are times when this is not feasible. For example, some studies may need to utilize local laboratories due to the nature of the study. There are also cases where the scheduled labs are done using a central laboratory, but ad-hoc local laboratory results are done as needed for patient care.

If the scheduled laboratory results are from a central laboratory, it seems appropriate to only consider these results in providing summary statistics for results over time. However, when looking for extremes for safety, one would not want to ignore the ad-hoc local results. This is consistent with the recommendation in section 6.2.1.

In some cases, multiple central laboratories (e.g., Europe and North America) are used in a single study, or perhaps vary across studies for integrated summaries. In these cases, consideration must be given as to whether the results should be scaled to allow the data to be combined. The approach will depend on the expected nature of the variability across laboratories and assays for different analytes. For example, hematology results can generally be combined from different laboratories without requiring any adjustments. On the other hand, assays for enzymes, such as the hepatic enzymes, can result in sufficiently different results (poor inter-assay reliability) that the results cannot be considered comparable. In some instances while the reference range for a given analyte can be slightly different across laboratories, the assays in the respective laboratories have good inter-assay reliability and results can be combined without concern. This latter case arrises when the same assay method is used in two laboratories but each laboratory uses a unique set of reference subjects to develop their own reference limits.

When some type of adjustment is required in order to combine information, a frequently used technique is to report the data as a percent above/below normal limits. Suppose the lower bound is L and upper bound is U. Then to evaluate a potential safety problem with increasing values, all values would be scaled as X/U. In some instances it may be necessary to consider an alternative which also takes into account differences in the range between the lower and upper bound of normal limits. As noted above, the choice will depend on the expected nature of the variability among laboratories. However, it is important to have a plan stated in advance as part of the Statistical Analysis Plan to assure that no question is raised about bias in selecting a method.

6.2.5. Above and Below Quantifiable Limits

Values above or below quantitative range include critical information and should not be discarded. If it is expected that a large portion of the values would be in the undetectable range (e.g., only abnormal results would be quantifiable), then it might be more appropriate to emphasize the percent abnormal and then provide information on the quantitative results only for those which are abnormal. For comparison of treatment groups a nonparametric rank test could be employed. For descriptive statistics quantiles would be appropriate, but presentation of the mean and standard deviation would not. If there are very few instances outside normal ranges, a simple rule such as assigning all those above the quantifiable range the largest observable value and all those below the quantifiable range the lowest observable value could be employed. Care should be taken before using zero for those below the quantifiable range if the lowest quantifiable value is far from zero since this could overweight these low values in any summary statistics or tests based on normal distribution assumptions. If the laboratory test is of known importance in evaluating the safety of the test agent, it will be important to pre-specify the approach planned to deal with these values and it may be necessary to do sensitivity analyses to confirm that conclusions do not change depending on how the out-of-quantitative-range values are handled.

6.3. ECG Quantitative Measurements

Special considerations for “thorough QT/QTc studies” are considered out-of-scope for this white paper.

6.3.1. QT Correction Factors

As noted in the ICH QT/QTc guidance (Section IA; Background), because of its inverse relationship to heart rate, the measured QT interval is routinely corrected by means of various formula to a less heart-rate-dependent value known as the QTc interval. Section IIIA of the same guidance provides a discussion of some of the various correction formulas and notes the controversy around appropriate corrections. Generally, we recommend that the TFLs include the corrected QT interval using Fridericia's method ($QTcF = QT/RR^{0.33}$). We believe the regulatory and medical environments are ready to accept the exclusion of Bazett's method from standard TFLs. We believe a second method would likely be warranted for a more complete evaluation. The second method could be one that is derived from a linear regression technique (Dmitrienko, et. al. 2005).

6.3.2. JT Interval

QTc is a biomarker with a long established history of being used to assess the duration of ventricular repolarization. However, QTc encompasses both ventricular depolarization and ventricular repolarization. The length of the QRS complex represents ventricular depolarization and the length of the JT interval, measured from the end of the QRS complex to the end of the T-wave, specifically represents ventricular repolarization. JT can be corrected for heart rate as with QT. Thus, when the QRS is prolonged (e.g., a complete bundle branch block), QTc should not be used to assess ventricular repolarization. The decision as to which basis for assessing potential changes in ventricular repolarization will be used should be based on the expected proportion of patients with widened QRS complexes for any reason in that study. It is worth noting that this proportion increases with the age of the patient population and the extent to which the population is expected to suffer cardiac disease.

7. Tables and Figures for Individual Studies

7.1. Multiple Measurements Over Time

For safety assessments that have multiple measurements over time (typically the case for vital signs, sometimes the case for ECGs and laboratory measurements), a box plot of the observed values and a box plot of change from baseline over time are recommended. See Figures 7.1 and 7.2. We chose to utilize the notch option, as notches are useful in offering a rough guide to significance of difference of medians (McGill, et. al. 1978). In Figure 7.1, observations that are outside of pre-defined reference limits are provided in red. Lines indicating the reference limits can be added to ease the review of the plots (not shown in the figure) when there is only one set of reference limits across the population included in the figure. In cases where limits vary across demographic characteristics and/or laboratories, lines for the lowest of the high limits and the highest of the low limits can be displayed, or lines for all limits can be displayed. However, these methods for handling lines for multiple limits are generally too confusing to the users of the plots.

We recommend the additional display of change from baseline to last observation (last non-missing observation in the treatment period) at the right-most side of the box plot of changes (Figure 7.2). A test for treatment differences can then be included in the summary table below the box plot, using ANCOVA containing terms for treatment and the continuous covariate of baseline measurement.

7.2. Single Pre- and Post-Treatment Measurements

For safety assessments that have single pre- and post-treatment measurements planned, a box plot displaying baseline and last observation (last non-missing observation in the treatment period) and a box plot of change from baseline to last observation are recommended. See Figure 7.3 (which displays the 2 box plots side-by-side).

7.3. Discussion

As described in Sections 7.1 and 7.2, the box plot is the recommended visual display. There are certainly multiple visual displays that can be used for central tendency. Another visual display for showing trends over time that was considered is one that displays means and standard error bars (See Figure 7.4). We recommend box plots instead since they have the advantage of easily showing additional summaries of interest beyond the mean for safety (median, min, max, quartiles). It also shows the impact of outliers on the central tendency. For safety assessments with single pre- and post-treatment measurements, a scatterplot was considered (see Figure 7.5 for an example). Scatterplots have the advantage of visually seeing individual subject data better while box plots show population data better. Scatterplots also have the advantage of visually seeing post-treatment outliers in the context of the subject's pre-treatment measurement. However, we believe it is easier to see group differences using a box plot than a scatterplot, which we believe is preferred in the context of central tendency assessment.

The disadvantage of box plots is that they have limited readability if there are multiple treatment arms and/or many time points such that having the data fit on one page becomes difficult. (Of course, this could be a limitation of other displays as well.) A particular box plot may also have limited readability when there's an extreme outlier which then squishes the box portion. Various techniques can be considered to handle this situation. Slashes can be used on the y-axis to provide separation, a transformation (e.g., log transformation) can be used for just the measurements in which this problem occurs, or the outlier is not shown (or "clipped") from the display but included in the summary statistics. The "slash-method" is generally preferred by medical colleagues, however may be more difficult to implement routinely.

Consideration was given on whether we could provide even less information on central tendency as the general recommendation, especially since marked outliers is typically of greatest interest (See Section 7.1.7.3.1 of the Reviewer Guidance). As previously noted, there is an expectation at least one analysis of central tendency is conducted. Consideration was given to recommend Figure 7.2 even when there are multiple measurements over time. It is a simpler display in which 2 box plots can be placed side-by-side in a single figure. However, having the ability to visually view potential time trends over time (when such data is collected) is generally desirable by medical colleagues. Having both the observed values (to visually see the potential impact of outliers on central tendancy) and change values are deemed useful and important for safety signal detection.

Consideration was given to provide even more information. First, we considered change from baseline to last visit for completers only. Since this analysis excludes data from subjects who could not tolerate the drug and discontinue early, we feel its value is limited for signal detection. Second, we considered change from baseline to last visit using repeated measures methodology. This method is more computational intensive and special consideration must be given to the appropriate covariance structure. Thus, we believe this is best reserved for topics of special concern. Third, we considered providing p-values in the summary table for Figure 7.2 for each timepoint. We believe the inclusion of p-values to this level might take focus away from simply reviewing the trend across time, and the notches in the box plots already provides a sense of potential differences.

For Figure 7.2, when a p-value is included to assess treatment differences in baseline to endpoint change, there is debate whether to use non-parametric versus parametric methods (i.e., Mann-Whitney-Wilcoxon test vs. t-test). The Mann-Whitney-Wilcoxon (MWW) test is frequently used as it is a test that is resistant to outliers or extreme values. A significant MWW test can be thought of as showing a difference in medians. Yet, for safety signal screening purposes, which most of the routine safety analyses are about, outliers/extreme values are of interest. Very often, the drug may have a strong effect on a small portion of the population. These extreme values are of interest for further exploration. A p-value from a test of mean differences is more sensitive than from a MWW test to detect differences when extreme values exist. The TFLs focusing on outliers or shifts may not be helpful in identifying such situations if the changes are still under the specified threshold. These situations may not be obvious in box plots over time if the changes are scattered across time. Thus, providing p-values using parametric testing can be a

useful tool while reviewing large quantities of information for safety signal screening purposes. The primary concern with using parametric testing versus non-parametric testing in the context of safety signal detection is if an outlier exists for a comparator patient but not in any subject part of the treatment group. Even if a subset of the subjects in the treatment goup change by a moderate amount, parametric testing may not indicate a potential signal. However, such cases would be apparent in the box plots, so a safety signal should not be missed. Also as noted above, the notches in the box plots are useful in offering a rough guide to significance of difference of medians.

8. Tables and Figures for Integrated Summaries

8.1. Timing of Measurements Varies Across Studies

For safety assessments in which the timing of measurements varies across studies, box plots of baseline (last non-missing observation across pre-treatment measurements), last observation (last non-missing observation in the treatment period), minimum baseline (minimum non-missing value across pre-treatment measurements), minimum during treatment (minimum non-missing value during the treatment period), maximum baseline (maximum non-missing value across pre-treatment measurements), and maximum during treatment (maximum non-missing value during the treatment period) are recommended. When the number of studies is small (e.g. $\leq$5), box plots for each study and studies combined are recommended. See Figures 7.6A-C.

Additionally, box plots of change from baseline to last observation, change from the minimum baseline value to the minimum value during the treatment period, and change from the maximum baseline value to the maximum value during the treatment period are recommended. See Figures 7.7A-C. The box plots of change can then include the following p-values for treatment differences:

- Change from baseline to last observation; includes all subjects who have both a baseline and post-baseline observation; ANCOVA containing terms for treatment, study, and the continuous covariate of baseline measurement
- Change from the minimum value during the baseline period to the minimum value during the treatment period; includes all subjects who have both a baseline and post-baseline observation; ANCOVA containing terms for treatment, study, and the continuous covariate of baseline measurement
- Change from the maximum value during the baseline period to the maximum value during the treatment period; includes all subjects who have both a baseline and post-baseline observation; ANCOVA containing terms for treatment, study, and the continuous covariate of baseline measurement

When the number of studies is large, it becomes less reasonable to display by-study results, and then the summaries can be combined on a single display. See Figure 7.8.

8.2. Timing of Measurements is Consistent Across Studies

For safety assessments in which the timing of measurements is consistent across studies, Figures 7.6 and 7.7 are still recommended. Figures 7.1 and 7.2 can be considered, however, additional statistics should be added to the table utilizing methods controlled for study to alert users for potential paradoxes in the data.

8.3. Discussion

For safety assessments in which the timing of measurements varies across studies, box plots over time are not recommended when some time points have some studies while other time points have other studies. When time is confounded with study, it is easy for the data to be misinterpreted. Thus, summaries of baseline to last observation, maximum observation (when

increases are of interest), and minimum observation (when decreases are of interest) are recommended. These are convenient summaries for signal detection purposes.

Assessments of maximum and minimum changes is desirable since potentially meaningful changes can occur at different times for different patients. Changes to minimum and maximum values provide an average of the range of changes, and the differences in these averages between treatment and control is considered useful for safety signal detection purposes. The quantitative displays of changes to minimum and maximum values supplement the displays for outliers and shifts, which also focus on extreme changes. In particular, they are considered sensitivity analyses associated with the displays for outliers and shifts. Displays for outliers and shifts may miss safety signals if the thresholds aren't in the location where changes may be occurring.

Once it is decided to include an assessment of minimum and maximum changes, there is debate on how to define baseline when more than one pre-treatment measure is collected per protocol. We recommend taking the minimum value across the baseline period for change to minimum, and taking the maximum value across the baseline period for change to maximum. This approach is not currently common across the industry, but is recommended as a means to continuously improve analytical approaches for detecting more meaningful changes. This is preferred over the last measurement during the baseline period since this minimizes the effect of normal variation and generally reflects more clinically meaningful changes of interest. For example, assume there are two pre-treatment assessments where the first assessment is in the high range and the second assessment is the normal range. Also assume the subject has a value in the high range during treatment. If the last pre-treatment value is used as baseline, the subject contributes a high change toward the central tendency measure. This is inconsistent with what medical colleagues generally feel is appropriate. Another method to minimize the effect of variation is to take the average of pre-treatment measurements. We recommend the minimum and maximum value during the pre-treatment period over the average since it minimizes the effect of normal variation to a greater degree.

9. Example Language for Statistical Analysis Plans

9.1. Individual Study

Values at each visit (starting at randomization) and change from baseline to each visit for laboratory tests, vital signs, physical characteristics, and ECG parameters will be displayed in box plots (notched box for each treatment with outliers displayed) for subjects who have both a baseline and a result for the specified visit. Individual measurements outside of reference limits will also be displayed using distinct symbols overlaying the box plot. Baseline will be the last non-missing observation in the baseline period. Original-scale data will be used for the display. Unplanned measurements will be excluded. Displays using both SI and U.S. conventional units will be provided (when different). The following summary statistics will be included in a table below the box plot: N, mean, standard deviation, minimum, Q1, median, Q3, and maximum. P-values and confidence limits will not be included in the summary statistics at the bottom of the box plot. Box plots will be used to evaluate trends over time and to assess a potential impact of outliers on central tendency summaries.

Change from baseline to last observation will also be summarized within the box plot of changes (rightmost column) for subjects who have both baseline and at least one post-baseline result. Baseline will be the last non-missing observation in the baseline period. The last non-missing observation in the treatment period will be used as the last observation. Original-scale data will be used. Unplanned measurements will be excluded. A p-value will be included in the summary statistics at the bottom of the box plot for this assessment, using an ANCOVA model containing terms for treatment and the continuous covariate of baseline measurement.

Laboratory tests include all planned analytes as defined in the protocol, excluding those collected in a reflex manner (only collected under certain circumstances). Vital signs include systolic blood pressure, diastolic blood pressure, pulse, and temperature. Physical characteristics include weight and BMI. ECG parameters include heart rate, PR, QRS, QT, corrected QT using Fredericia's correction factor ($QTcF=QT/RR^{0.333}$), and corrected QT using a large clinical trial population based correction factor ($QTcLCTPB=QT/RR^{0.413}$; Dmitrienko AA, Sides GD, Winters KJ, Kovacs RJ, Rebhun DM, Bloom JC, Groh W, Eisenberg PR. Electrocardiogram reference ranges derived from a standardized clinical trial population. DRUG INF J 39:395-405; 2005) When the QRS is prolonged (for example, a complete bundle branch block), QT and QTc should not be used to assess ventricular repolarization. Thus, for a particular ECG, the following will be set to missing (for analysis purposes) when QRS is ≥120: QT, QTcF and QTcLCTPB.

9.2. Integrated Summary

For laboratory tests, vital signs, physical characteristics, and ECG parameters, the following box plots (notched box for each treatment with outliers displayed) will be created, including each study and studies combined in a single display:

- Baseline (last non-missing observation across pre-treatment measurements) and last observation (last non-missing observation in the treatment period)

- Minimum baseline (minimum non-missing value across pre-treatment measurements) and minimum during treatment (minimum non-missing value during the treatment period)
- Maximum baseline (maximum non-missing value across pre-treatment measurements) and maximum during treatment (maximum non-missing value during the treatment period).
- Change from baseline to last observation
- Change from the minimum baseline value to the minimum value during the treatment period
- Change from the maximum baseline value to the maximum value during the treatment period

Original-scale data will be analyzed. Unplanned measurements will be excluded for baseline and last observations. Planned and unplanned measurements will be included for minimum and maximum observations. Analyses will be provided in both SI and U.S. conventional units (when different). Individual measurements outside of reference limits will also be displayed using distinct symbols overlaying the box plot. The following summary statistics will be included in a table below the box plot: N, mean, standard deviation, minimum, Q1, median, Q3, and maximum. The box plots of change will include the following p-values for treatment differences:

- Change from baseline to last observation; includes all subjects who have both a baseline and post-baseline observation; ANCOVA containing terms for treatment, study, and the continuous covariate of baseline measurement
- Change from the minimum value during the baseline period to the minimum value during the treatment period; includes all subjects who have both a baseline and post-baseline observation; ANCOVA containing terms for treatment, study, and the continuous covariate of baseline measurement
- Change from the maximum value during the baseline period to the maximum value during the treatment period; includes all subjects who have both a baseline and post-baseline observation; ANCOVA containing terms for treatment, study, and the continuous covariate of baseline measurement

Summaries and analyses of minimum and maximum values are for sensitivity assessment purposes and will only be discussed if deemed relevant in the overall discussion of the safety profile of the compound.

Laboratory tests include all planned analytes as defined in the protocol, excluding those collected in a reflex manner (only collected under certain circumstances). Vital signs include systolic blood pressure, diastolic blood pressure, pulse, and temperature. Physical characteristics include weight and BMI. ECG parameters include heart rate, PR, QRS, QT, corrected QT using Fredericia's correction factor ($QTcF=QT/RR^{0.333}$), and corrected QT using a large clinical trial population based correction factor ($QTcLCTPB=QT/RR^{0.413}$; Dmitrienko AA, Sides GD, Winters KJ, Kovacs RJ, Rebhun DM, Bloom JC, Groh W, Eisenberg PR. Electrocardiogram reference ranges derived from a standardized clinical trial population. DRUG INF J 39:395-405; 2005) When the QRS is prolonged (for example, a complete bundle branch block), QT and QTc

should not be used to assess ventricular repolarization. Thus, for a particular ECG, the following will be set to missing (for analysis purposes) when QRS is ≥120: QT, QTcF and QTcLCTPB.

10. References

Amit O, Heiberger RM, and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. Pharmaceut. Statist. 2008; 7: 20–35. doi: 10.1002/pst.254.

Crowe BJ, Xia A, Berlin JA, Watson DJ, Shi H, Lin SL, et. al. Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: a report of the safety planning, evaluation, and reporting team. Clinical Trials 2009; 6: 430-440.

Biological Variation: From Principles to Practice. Callum G. Fraser. Washington, DC: AACC Press, 2001, 151 pp.

Dmitrienko AA, Sides GD, Winters KJ, Kovacs RJ, Rebhun DM, Bloom JC, Groh W, Eisenberg PR. Electrocardiogram reference ranges derived from a standardized clinical trial population. Drug Inf J 2005; 39:395-405.

McGill R, Tukey JW, and Larsen WA. Variations of Box Plots. The American Statistician 1978; 32(1): 12-16. doi:10.2307/2683468.JSTOR 2683468.

Rosario LA, Kropp TJ, Wilson SE, Cooper CK. Join FDA/PhUSE Working Groups to help harness the power of computational science. Drug Information Journal 2012; 46: 523-524.

11. Acknowledgements

The key contributors include: Mary E. Nilsson, Wei V. Wang, Charles M. Beasley, Jr., and Qi Jiang.

Additional contributors and members of the white paper project within the CSS Development of Standard Scripts for Analysis and Programming Working Group include: Sasha Ahrweiler, Kirk Bateman, Simin Baygani, Nancy Brucken, Jean-Marc Ferran, Dany Guerendo, Harprit Dosanjh, Lina Jorgensen, Fabien Linay, Raphael Noirfalise, Musa Nsereko, Frank Senk, Jack Shostak.

Acknowledgement to others who provided text for various sections, review comments, and/or participated in discussions related to methodology: Joshua Betcher, Sally Cassells, Lai Shan Chan, Brenda Crowe, Mary Anne Dellva, Damon Disch, Michele Ennis, Kaylani Kothapali, Craig Mallinkrodt, Beth Pangallo, and Mike Smith.

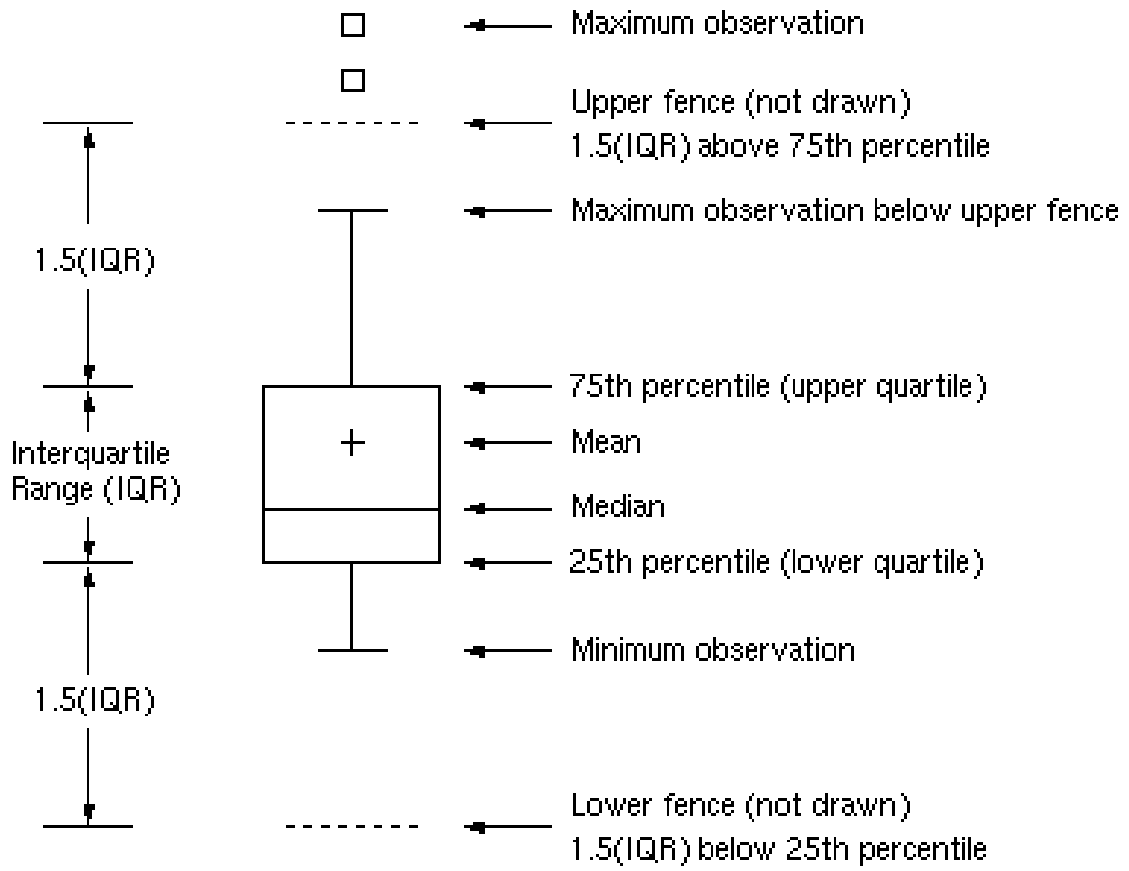
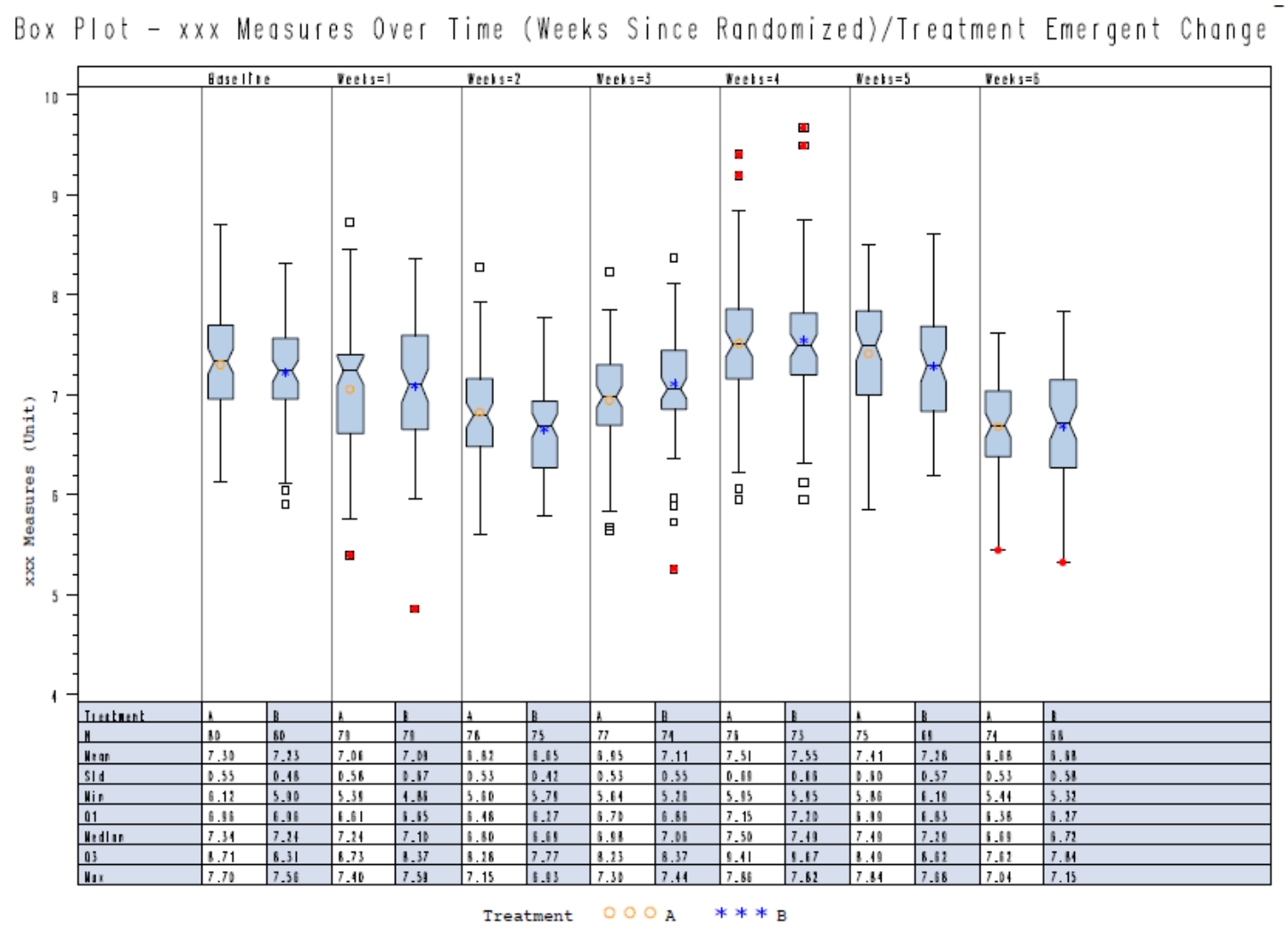
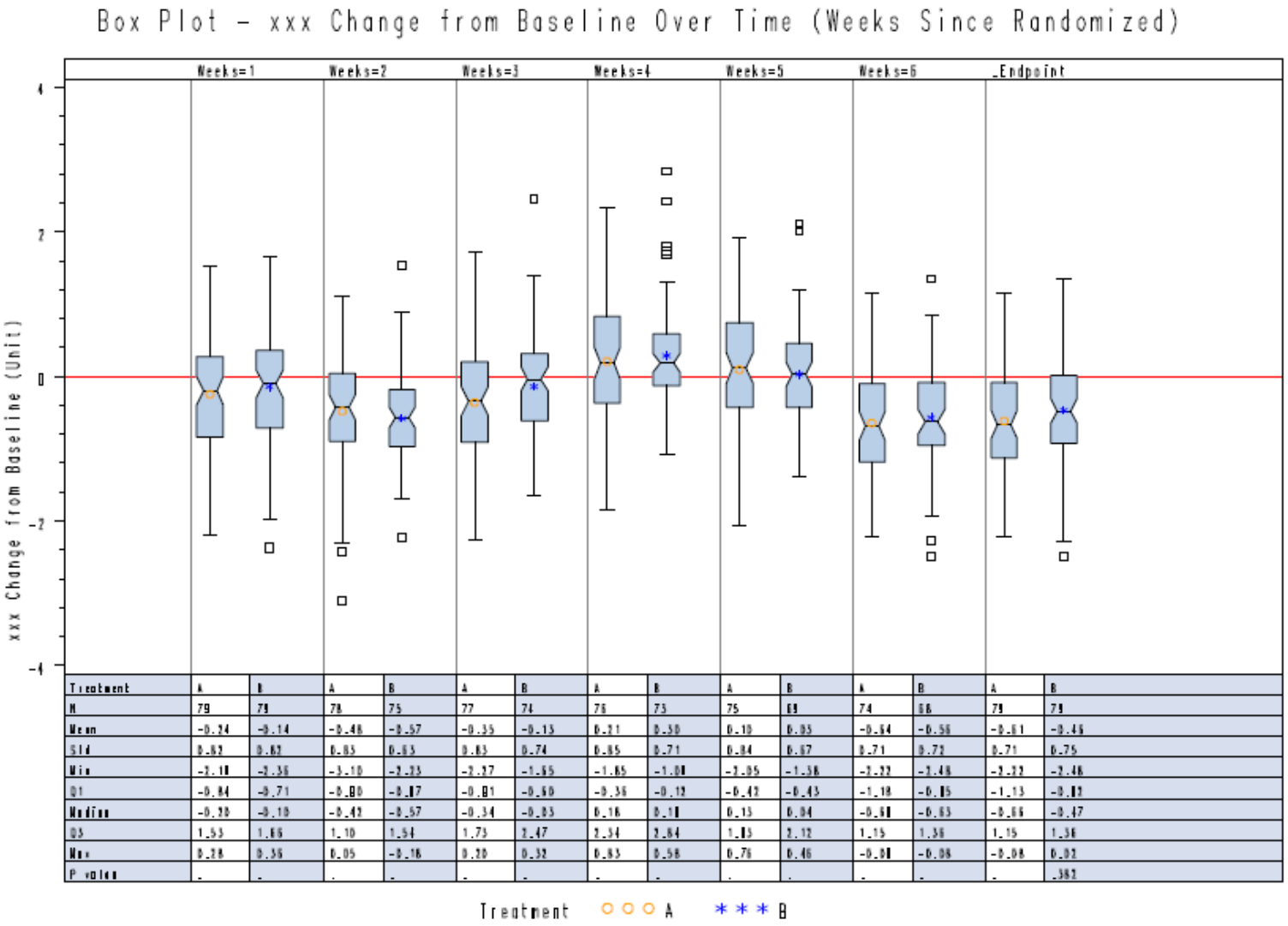


Figure 6.1. **Explanation of Box Plot (copied from SAS/STAT(R) 9.2 User's Guide, Second Edition – Styles of Box Plots)**



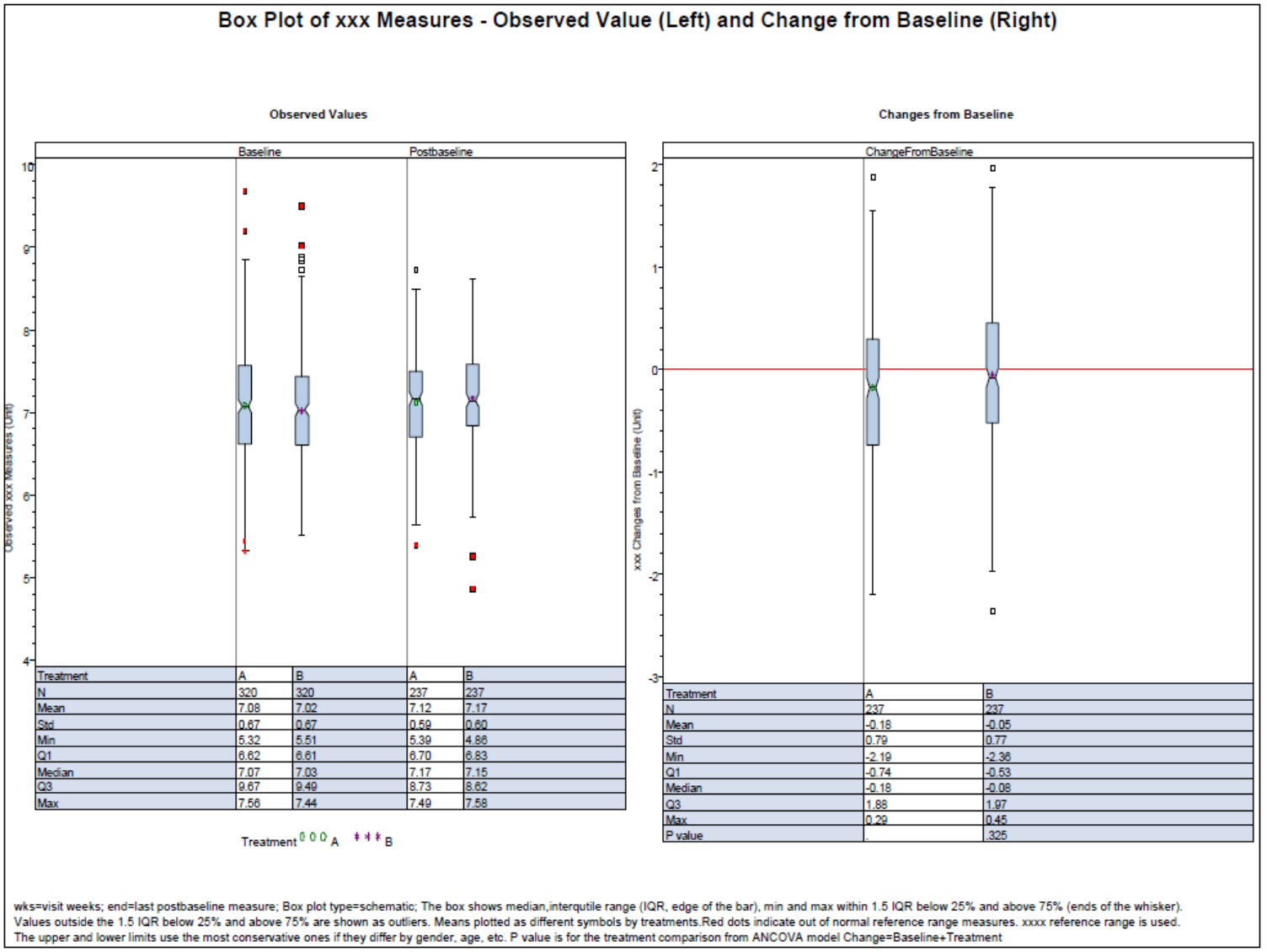
Treatment	A	B	A	B	A	B	A	B	A	B	A	B	A	B
N	80	80	78	78	76	75	77	74	76	73	75	69	74	66
Mean	7.30	7.23	7.06	7.00	6.82	6.85	6.95	7.11	7.51	7.55	7.41	7.26	6.86	6.88
Std	0.55	0.46	0.56	0.67	0.53	0.42	0.53	0.55	0.80	0.80	0.80	0.57	0.53	0.58
Min	6.12	5.90	5.36	4.86	5.80	5.79	5.64	5.28	5.85	5.85	5.80	6.10	5.44	5.32
Q1	6.86	6.80	6.61	6.65	6.46	6.27	6.70	6.86	7.15	7.10	6.90	6.85	6.56	6.37
Median	7.34	7.24	7.24	7.10	6.80	6.86	6.98	7.06	7.50	7.49	7.40	7.20	6.89	6.72
Q3	8.71	8.31	8.73	8.37	8.26	7.77	8.23	8.37	8.41	8.67	8.40	8.02	7.02	7.84
Max	7.70	7.56	7.40	7.58	7.15	8.03	7.30	7.44	7.86	7.82	7.84	7.86	7.04	7.15

Figure 7.1. Box Plot – Observed Values of xxx Over Time



Treatment	A	B	A	B	A	B	A	B	A	B	A	B	A	B
N	79	79	78	75	77	74	76	73	75	69	74	66	79	79
Mean	-0.24	-0.14	-0.46	-0.57	-0.35	-0.15	0.21	0.30	0.10	0.05	-0.64	-0.56	-0.61	-0.46
Std	0.82	0.82	0.85	0.63	0.83	0.74	0.85	0.71	0.84	0.67	0.71	0.72	0.71	0.75
Min	-1.18	-2.36	-3.10	-2.23	-2.27	-1.65	-1.85	-1.08	-2.05	-1.36	-2.22	-2.48	-2.22	-2.48
Q1	-0.84	-0.71	-0.80	-0.87	-0.81	-0.60	-0.36	-0.13	-0.42	-0.43	-1.18	-0.85	-1.13	-0.82
Median	-0.30	-0.10	-0.42	-0.57	-0.34	-0.05	0.16	0.18	0.13	0.04	-0.61	-0.63	-0.66	-0.47
Q3	1.53	1.66	1.10	1.54	1.73	2.47	2.34	2.84	1.03	2.12	1.15	1.36	1.15	1.36
Max	0.28	0.36	0.05	-0.16	0.20	0.32	0.83	0.58	0.76	0.46	-0.01	-0.08	-0.08	0.02
P value	.	.	.	.	.	.	.	.	.	.	.	.	.	.382

Figure 7.2. Box Plot – Change in xxx Over Time



Treatment	A	B	A	B
N	320	320	237	237
Mean	7.08	7.02	7.12	7.17
Std	0.67	0.67	0.59	0.60
Min	5.32	5.51	5.39	4.86
Q1	6.62	6.61	6.70	6.83
Median	7.07	7.03	7.17	7.15
Q3	9.67	9.49	8.73	8.62
Max	7.56	7.44	7.49	7.58

Treatment	A	B
N	237	237
Mean	-0.18	-0.05
Std	0.79	0.77
Min	-2.19	-2.36
Q1	-0.74	-0.53
Median	-0.18	-0.08
Q3	1.88	1.97
Max	0.29	0.45
P value		.325

Figure 7.3. Box Plot – Observed Values and Change in xxx Over Time

Figure 7.4. **Mean of xxx Over Time (Weeks Since Randomized)**

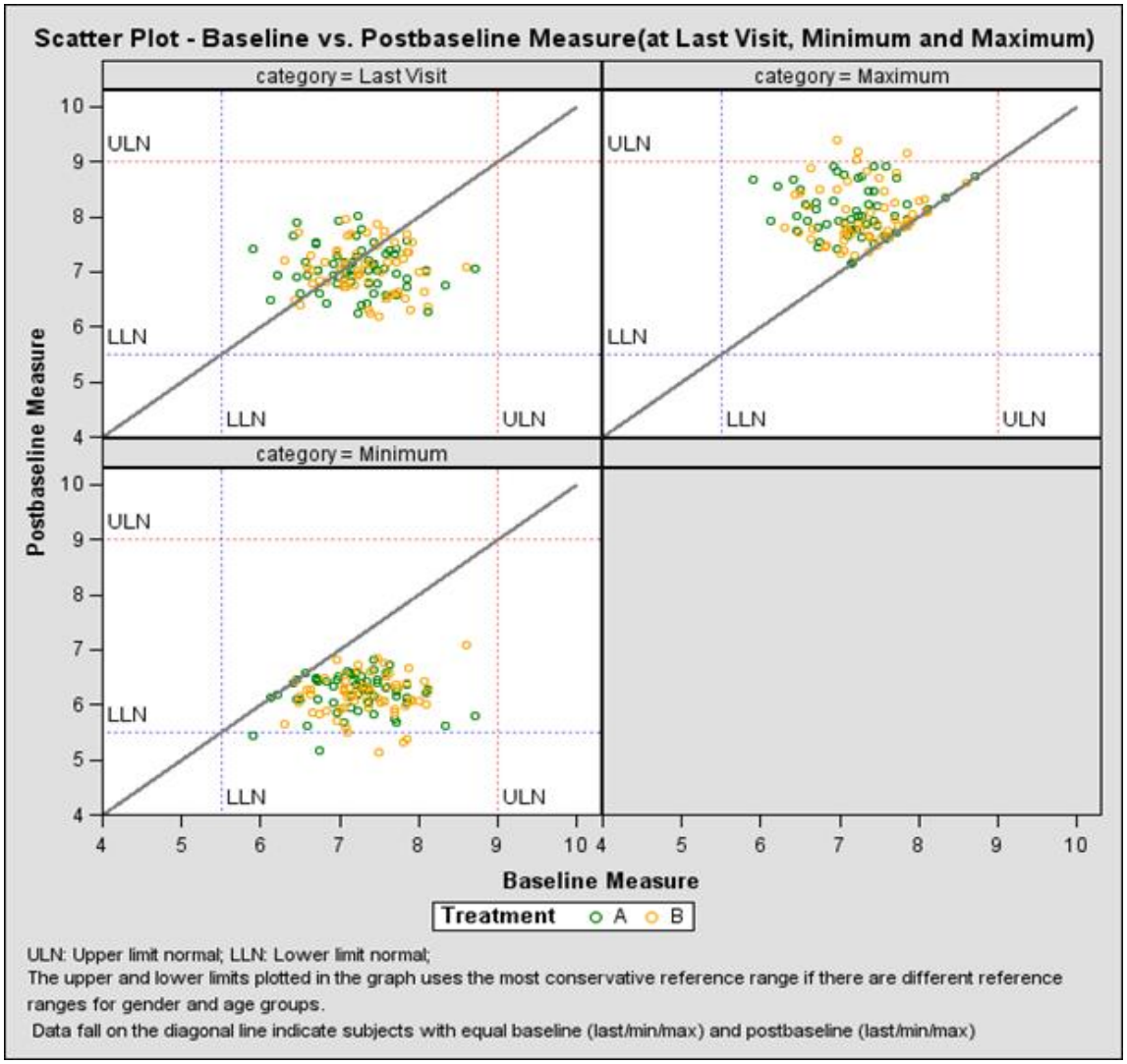
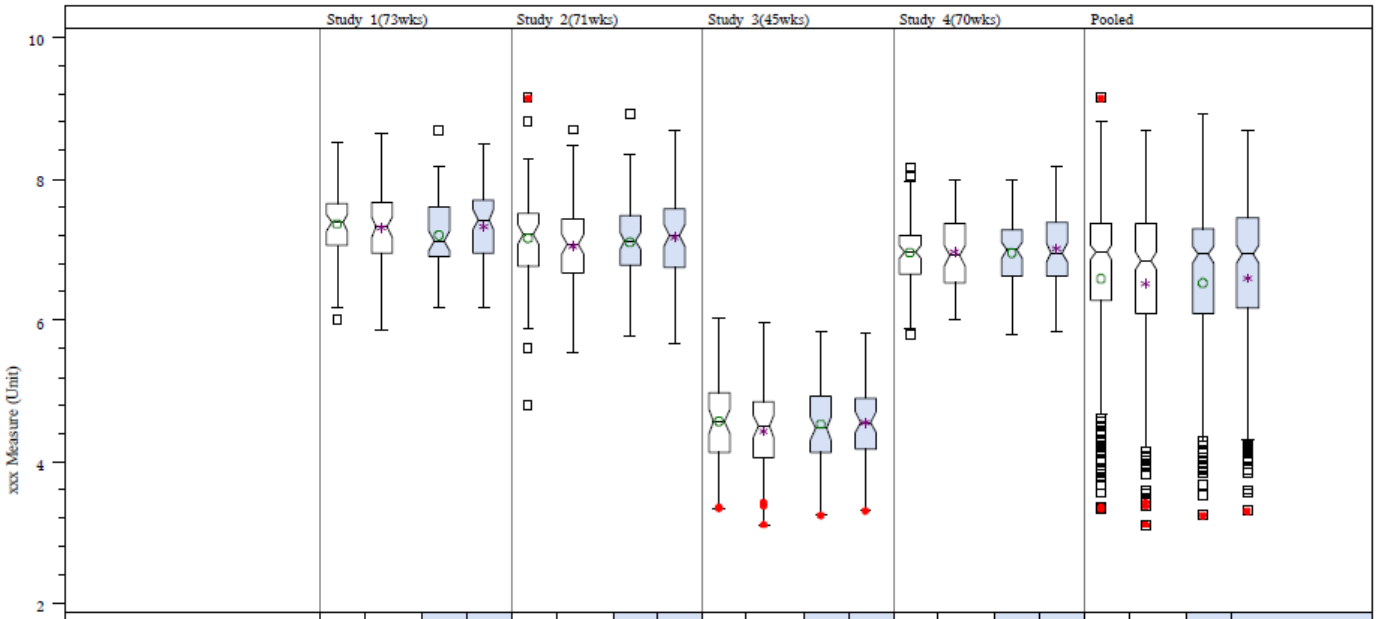


Figure 7.5. Scatterplot of Baseline Versus Endpoint Measurements for xxx



Period	BSE	BSE	PST	PST	BSE	BSE	PST	PST	BSE	BSE	PST	PST	BSE	BSE	PST	PST	BSE	BSE	PST	PST
Treatment	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
N	60	60	60	60	73	73	73	73	59	59	59	59	83	83	83	83	275	275	275	275
Mean	7.37	7.31	7.21	7.34	7.17	7.05	7.11	7.18	4.58	4.44	4.53	4.55	6.96	6.98	6.96	7.02	6.59	6.52	6.53	6.60
Std	0.49	0.60	0.57	0.52	0.69	0.64	0.62	0.66	0.60	0.56	0.55	0.53	0.46	0.51	0.46	0.51	1.20	1.24	1.18	1.21
Min	6.00	5.86	6.19	6.17	4.80	5.55	5.77	5.67	3.34	3.11	3.25	3.31	5.81	6.01	5.80	5.84	3.34	3.11	3.25	3.31
Q1	7.07	6.95	6.91	6.95	6.76	6.67	6.79	6.75	4.14	4.06	4.13	4.18	6.66	6.53	6.62	6.63	6.28	6.09	6.10	6.18
Median	7.38	7.32	7.13	7.42	7.22	7.08	7.12	7.20	4.58	4.50	4.47	4.56	6.96	6.91	6.99	6.96	6.96	6.84	6.93	6.94
Q3	8.53	8.63	8.69	8.49	9.15	8.70	8.92	8.70	6.04	5.97	5.84	5.81	8.16	7.99	7.98	8.19	9.15	8.70	8.92	8.70
Max	7.66	7.67	7.60	7.69	7.51	7.44	7.49	7.58	4.98	4.86	4.93	4.89	7.21	7.38	7.27	7.38	7.37	7.38	7.30	7.45

Treatment ○ ○ ○ A ＊ ＊ ＊ B

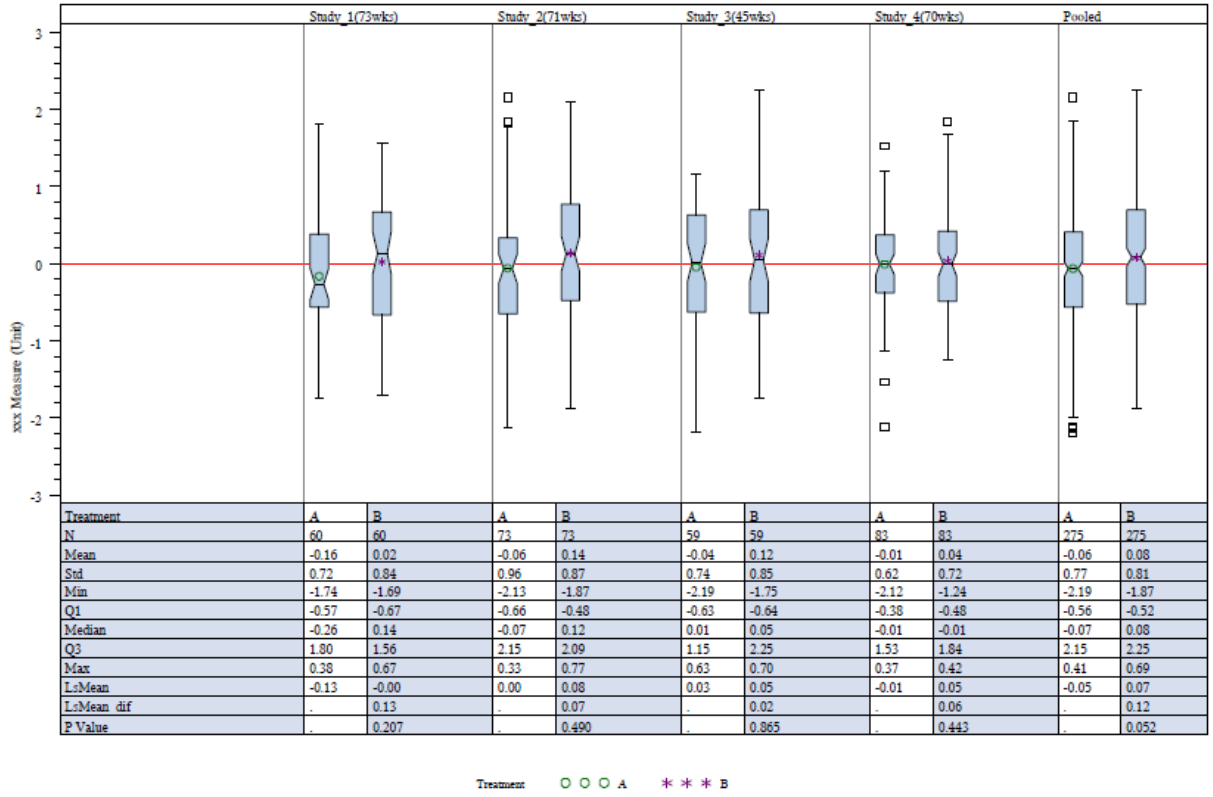
Box plot type=schematic, the box shows median,interqutile range (IQR, edge of the bar), Min and Max within 1.5 IQR below 25% and above 75% (ends of the whisker).
Values outside the 1.5 IQR below 25% and above 75% are shown as outliers. Means ploted as different symbols by treatments. Red dots indicate out of normal reference range measures.
The upper and lower limits use the most conservative ones if they differ by gender, age, etc. xxxx reference range is used. Baseline and postbaseline boxes are framed in different colors.
BSE=baseline PST=post-baseline

Figure 7.6A. Box Plot – Last Baseline and Last Postbaseline Measures

Figure 7.6B. Box Plot – Minimum Baseline and Minimum Postbaseline Measures

Figure 7.6C. Box Plot – Maximum Baseline and Maximum Postbaseline Measures

Laborary Analysis - Box Plot of Change from Last/Min/Max Baseline for Last/Min/Max Postbaseline Measure
label=last BSE to last PST



Treatment	A	B	A	B	A	B	A	B	A	B
N	60	60	73	73	59	59	83	83	275	275
Mean	-0.16	0.02	-0.06	0.14	-0.04	0.12	-0.01	0.04	-0.06	0.08
Std	0.72	0.84	0.96	0.87	0.74	0.85	0.62	0.72	0.77	0.81
Min	-1.74	-1.69	-2.13	-1.87	-2.19	-1.75	-2.12	-1.24	-2.19	-1.87
Q1	-0.57	-0.67	-0.66	-0.48	-0.63	-0.64	-0.38	-0.48	-0.56	-0.52
Median	-0.26	0.14	-0.07	0.12	0.01	0.05	-0.01	-0.01	-0.07	0.08
Q3	1.80	1.56	2.15	2.09	1.15	2.25	1.53	1.84	2.15	2.25
Max	0.38	0.67	0.33	0.77	0.63	0.70	0.37	0.42	0.41	0.69
LsMean	-0.13	-0.00	0.00	0.08	0.03	0.05	-0.01	0.05	-0.05	0.07
LsMean_dif	.	0.13		0.07	.	0.02		0.06	.	0.12
P Value	.	0.207	.	0.490	.	0.865	.	0.443	.	0.052

Treatment ○ ○ ○ A ✳ ✳ ✳ B

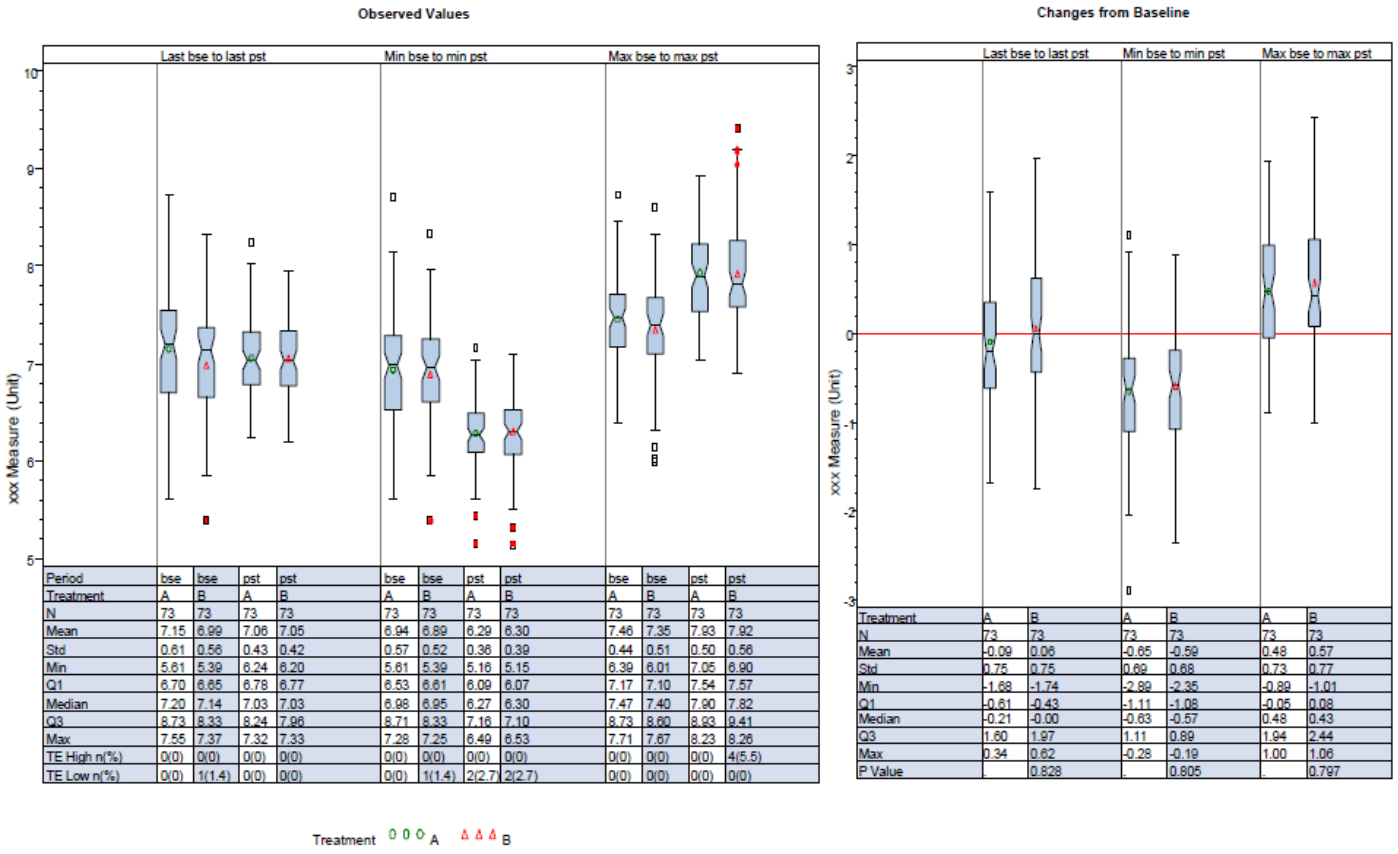
BSE=baseline; PST=postbaseline; Box plot type=schematic; The box shows median,interqutile range (IQR, edge of the bar), Min and Max within 1.5 IQR below 25% and above 75% ends of the whisker).Values outside the 1.5 IQR below 25% and above 75% are shown as outliers. Means ploted as different symbols by treatments. The upper and lower limits use the most conservative ones if they differ by gender, age, etc. P value is for the treatment comparison from the ANCOVA model Change=Baseline+Treatment+Study

Figure 7.7A. Box Plot – Change From Last Baseline to Last Postbaseline Measures

Figure 7.7B. Box Plot – Change From Minimum Baseline to Minimum Postbaseline Measures

Figure 7.7C. Box Plot – Change From Maximum Baseline to Maximum Postbaseline Measures

Laborary Analysis - Box Plot of Last/Min/Max Baseline vs. Last/Min/Max Postbaseline Measure



Period	bse	bse	pst	pst	bse	bse	pst	pst	bse	bse	pst	pst
Treatment	A	B	A	B	A	B	A	B	A	B	A	B
N	73	73	73	73	73	73	73	73	73	73	73	73
Mean	7.15	6.99	7.06	7.05	6.94	6.89	6.29	6.30	7.46	7.35	7.93	7.92
Std	0.61	0.56	0.43	0.42	0.57	0.52	0.36	0.39	0.44	0.51	0.50	0.56
Min	5.61	5.39	6.24	6.20	5.61	5.39	5.16	5.15	6.39	6.01	7.05	6.90
Q1	6.70	6.65	6.78	6.77	6.53	6.61	6.09	6.07	7.17	7.10	7.54	7.57
Median	7.20	7.14	7.03	7.03	6.98	6.95	6.27	6.30	7.47	7.40	7.90	7.82
Q3	8.73	8.33	8.24	7.96	8.71	8.33	7.16	7.10	8.73	8.60	8.93	9.41
Max	7.55	7.37	7.32	7.33	7.28	7.25	6.49	6.53	7.71	7.67	8.23	8.26
TE High n(%)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	4(5.5)
TE Low n(%)	0(0)	1(1.4)	0(0)	0(0)	0(0)	1(1.4)	2(2.7)	2(2.7)	0(0)	0(0)	0(0)	0(0)

Treatment	A	B	A	B	A	B
N	73	73	73	73	73	73
Mean	-0.09	0.06	-0.65	-0.59	0.48	0.57
Std	0.75	0.75	0.69	0.68	0.73	0.77
Min	-1.68	-1.74	-2.89	-2.35	-0.89	-1.01
Q1	-0.61	-0.43	-1.11	-1.08	-0.05	0.08
Median	-0.21	-0.00	-0.63	-0.57	0.48	0.43
Q3	1.60	1.97	1.11	0.89	1.94	2.44
Max	0.34	0.62	-0.28	-0.19	1.00	1.06
P Value	.	0.828	.	0.805	.	0.797

bse=baseline; pst=postbaseline; Box plot type=schematic; The box shows median,interqutile range (IQR, edge of the bar), min and max within 1.5 IQR below 25% and above 75% ends of the whisker).Values outside the 1.5 IQR below 25% and above 75% are shown as outliers. Means plotted as different symbols by treatments. Out of normal reference range data is plotted as red dots overlay the boxplot. Reference lines plotted the lowest of the upper limited normal and the highest of the lower limited normal. P value is for the treatment comparison from the ANCOVA model Change=Baseline+Treatment

Figure 7.8. Box Plot – Last/Minimum/Maximum Baseline and Last/Minimum/Maximum Postbaseline Measures and Change from Last/Minimum/Maximum Baseline to Last/Minimum/Maximum Postbaseline Measures