Paper 3118-2019

Statistical Data Checks to Identify Questionable and Suspicious Data

Kaitie Lawson, Rho, Inc.

ABSTRACT

Although Electronic Data Capture (EDC) has improved efficiency and timeliness in data entry and analysis in clinical trials, it has reduced the safeguards inherent in double data entry performed by dedicated professionals. EDC is vulnerable to inadequate training, transcription errors, fat-finger errors, negligence, and even fraud. Moreover, recent initiatives in risk-based monitoring are moving away from 100% on-site source data verification. Thus, supplemental data monitoring strategies are essential to ensure data accuracy for statistical analysis and reporting. We have developed a suite of statistical procedures to identify suspicious data values for individual subjects and across clinical sites. They include rounding errors and digit preference checks, univariate and bivariate outlier checks, longitudinal outlier checks within subjects, and variance checks. Generally, regression models are applied to account for demographic characteristics and other important covariates. The residuals from these models are used to identify outliers. The suite of data checks is illustrated using fabricated data. The strengths of this approach are highlighted and there is discussion of its shortcomings. This suite of statistical data checks is an effective tool for supplementing current processes and ensuring data accuracy. It can focus resources on specific data fields and clinical sites for efficient risk-based monitoring strategies.

INTRODUCTION

While electronic data capture (EDC) has improved efficiency and timeliness in data entry and analysis in clinical trials, it has also reduced the safeguards inherent in double data entry performed by dedicated professionals. EDC is vulnerable to inadequate training, transcription errors, negligence, and even fraud. Moreover, initiatives in “risk-based monitoring” are moving away from 100% on-site source data verification. Thus, supplemental data monitoring strategies are essential to ensure data accuracy for statistical analysis and reporting. We have developed a suite of statistical procedures to identify suspicious data values within individual subjects and across clinical sites. Rather than relying on vague, visual impression of ‘suspicious’ data, statistical methods such as Multinomial tests and (variance) tests are applied. The residuals from these models are used to identify the outliers for further review.

Using statistical techniques to identify suspicious data instead of relying solely on standard data management checks has numerous advantages. For example, the two checks that will be highlighted in this paper (Digit Preference and Variance Check) are not possible to perform within the traditional framework of EDC systems. By the end of this paper, the goal is that the reader will understand the advantages of implementing this additional level of data monitoring.

DIGIT PREFERENCE

The first statistical data check presented is the Digit Preference check. As the name suggests, this check is designed to determine if there are any data entry errors due to rounding or fraud in the last specified digit in a numeric variable. For example, sites or labs may have a tendency to round a lab value to either 0 or 5 as the last digit when recording data instead of giving the most accurate value. Fraudulent data enterers may tend to have a preference in the last digit. It is also plausible that a site might not follow the protocol in terms of the number of decimal places that need to be recorded (i.e., precision).

The macro we created to perform this check outputs stacked bar charts of the distribution of values for the desired digit of a numeric value. Time intervals can be specified as well as up to three different grouping variables may be chosen for display purposes.

STATISTICAL MODEL

The Chi-Square goodness-of-fit test is performed in 2 ways for the Digit Preference check. First, to test if the distribution of last digits within an individual bar is uniform (i.e., if the values of the last digit are distributed somewhat evenly between 0 and 9, as we would expect), the null hypothesis is that $p_0 = p_1 = .. = p_9 = 0.10$, where 0.10, where p_0 is the proportion of observations that have a 0 in the last digit place, p_1 is the proportion of observations that have a 1 in the last digit place, etc. The test statistic is:

$$\chi^2 = \sum_{i=0}^{9} \frac{(O_i - E_i)^2}{E_i}$$

This test has 9 degrees of freedom. O_i is the observed value and E_i is the expected value of the last digit value i.

Second, to test if the distribution of last digits within an individual bar differs from the distribution of last digits within the rest of the panel; the null hypothesis is that the distribution of the last digits within the bar of interest is the same as that of the rest of the data within the same panel. The test statistic is:

$$\chi^2 = \sum_{i=1}^{2} \sum_{k=0}^{9} \frac{(O_{ik} - E_{ik})^2}{E_{ik}}$$

This test also has 9 degrees of freedom. O_{ik} is the observed value and E_{ik} is the expected value of the last digit value k within the row I, where I =1 represents being part of the bar of interest, with I =2 being part of the rest of the data within the panel.

DATA SOURCE

We will use the same dataset and outcome for all the following examples. Data was randomly generated to have the following properties:

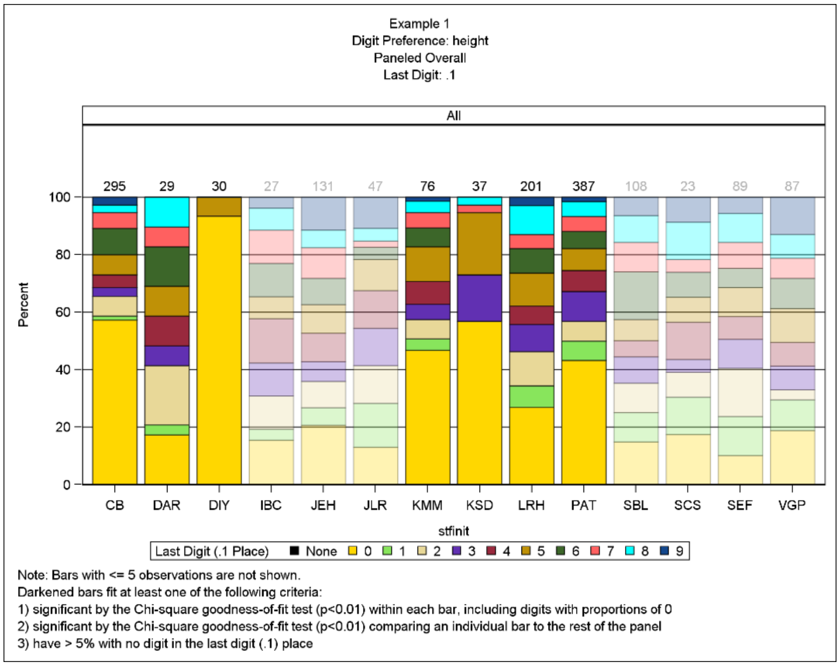
- Data for Height in meters was generated on each participant one time (e.g., one observation per participant).
- Data was generated by 14 total staff members ‘entering data’.
- Data was generated for 4 sites.

- Data was ‘collected’ from 2015 to 2017.

For all the following examples, we would like to examine the 10th place (i.e., 0.1) of the height variable summarized in various ways.

EXAMPLE #1 – OVERALL BY SITE STAFF

Since we are looking for potentially fraudulent data as entered by an individual staff member, we first start with examining the data in an overall fashion by the initials of the staff member who entered the data into EDC. Figure 1 displays the distribution of the 10th place of the height variable overall staff member initials.



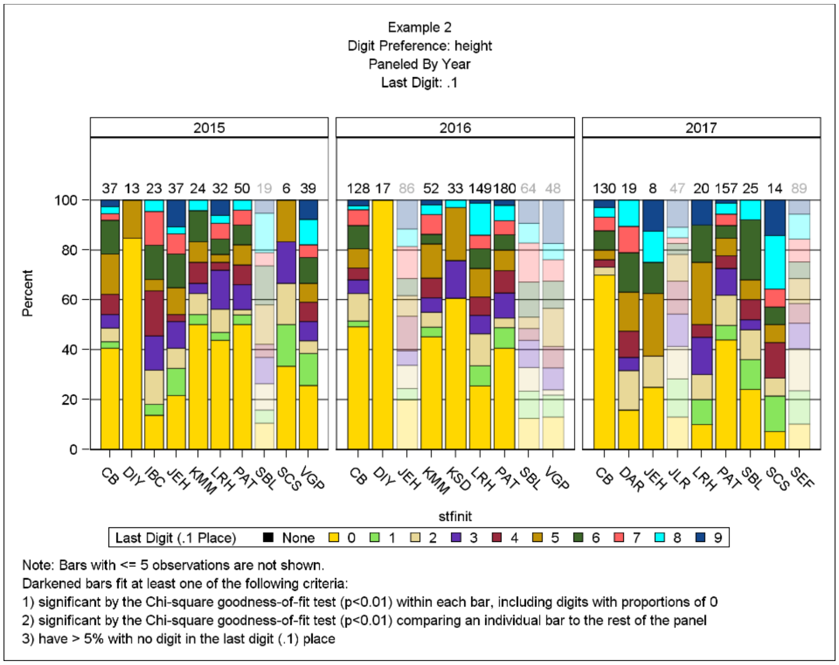
Note: Bars with <= 5 observations are not shown.
Darkened bars fit at least one of the following criteria:
1) significant by the Chi-square goodness-of-fit test (p<0.01) within each bar, including digits with proportions of 0
2) significant by the Chi-square goodness-of-fit test (p<0.01) comparing an individual bar to the rest of the panel
3) have > 5% with no digit in the last digit (.1) place

Figure 1 Height Overall by Staff Initials

The study name (‘Example 1’), variable of interest (‘height’), interval (‘Overall’) and digit of interest (‘.1’) are all displayed in the titles of the generated figure. The grouping variable staff initials (‘stfinit’) are displayed on the x-axis. The number of observations contributing to each bar are annotated on the top of each bar. The 10 possible choices for the last recorded digit (i.e., 0 through 9) each have a corresponding color in the generated graphic. From this overall look, we can see that staff member ‘DIY’ only ever used the 0 and 5 in the 10th place when entering data. This might be a cause for concern, but thankfully this person only contributed 30 observations in total. Of the 14 total staff members, half of them had potentially questionable choices for the 10th place. It seems like most of the reason these data enterers are being flagged is due to the overuse of 0. Finally, no one had a missing entry for the 10th place.

EXAMPLE #2 – BY SITE STAFF OVER TIME

After investigating the overall pattern of the data, the next step should be to investigate if there is a time trend to these data entry anomalies. Looking at this data over time might tell us if there was a higher likelihood that these errors happened at the beginning or maybe the end of the trial. Figure 2 displays the distribution of the 10th place of the height variable by staff member initials and year of data entry.


Example 2
Digit Preference: height
Paneled By Year
Last Digit: .1

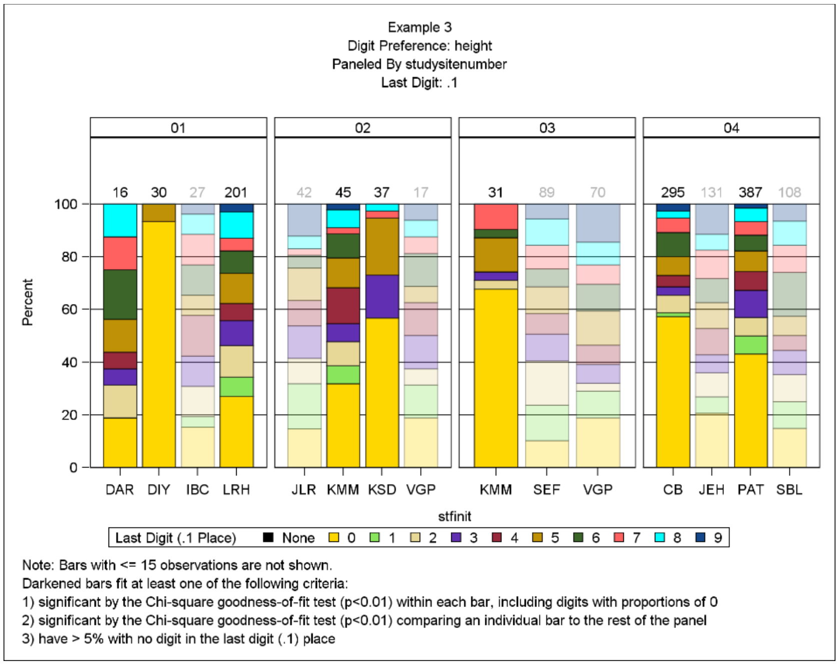
Note: Bars with <= 5 observations are not shown.
Darkened bars fit at least one of the following criteria:
1) significant by the Chi-square goodness-of-fit test (p<0.01) within each bar, including digits with proportions of 0
2) significant by the Chi-square goodness-of-fit test (p<0.01) comparing an individual bar to the rest of the panel
3) have > 5% with no digit in the last digit (.1) place

Figure 2. Height by Staff Initials and Year of Data Entry

There are a few interesting things to note when reviewing this data over time. One, not every data enterer contributed to data entry in the trial for all three years. For example, data enterer 'DIY' that was highlighted in the first example, only entered data in 2015 and 2016. This highlights the nature of turnover this trial experienced in terms of site staff entering data. We can also observe that, with the exception of one staff member ('CB'), the number of 0s in the 10th place diminished over time. This observation speaks to the improvement of recording in terms of precision. Maybe the sites ended up with a newer/more precise scale, or maybe there were supplemental site trainings that focused on precision of data entry.

EXAMPLE #3 – BY STAFF AND SITE

Up to this point, we have been reviewing the data with the assumption that all the data enters were at the same site. We know there are multiple sites for this trial; therefore, it is of interest to see if there are any noticeable trends within or across sites. Maybe one site wasn't trained as well as the other sites. There might be one site that is so poor at entering data that we might want to think about not using them in the future. Figure 3 displays the distribution of the 10th place of the height variable by staff member initials and site number.



Note: Bars with <= 15 observations are not shown.
Darkened bars fit at least one of the following criteria:
1) significant by the Chi-square goodness-of-fit test (p<0.01) within each bar, including digits with proportions of 0
2) significant by the Chi-square goodness-of-fit test (p<0.01) comparing an individual bar to the rest of the panel
3) have > 5% with no digit in the last digit (.1) place

Figure 3. Height by Staff Initials and Site Number

Again, there are a few interesting aspects that are highlighted when looking at the data by site and data enterer. First, there was one person worked at two different sites. Staff member ‘KMM’ worked at Site 02 and Site 03. Second, Site 04 had far more participants compared to the other sites combined. Finally, on the whole, it looks like Site 01 potentially had the worst data entry since staff member ‘DIY’ (who was highlighted in Figure 1) only used the 0 and 5 place and other staff members ‘DAR’ and ‘LRH’ are also highlighted for having differences in the 0.1 place for data entry.

VARIANCE CHECK

The second statistical data check presented will be the Variance Check. This check is designed to detect significant differences in variances between groups of observations (e.g., between sites). Differences in variances might be an indication that the protocol is not being followed correctly or that the data may be fabricated. Fraudulent data enterers may be able to enter data such that mean differences are apparent and make sense within the confines of a trial; however, it is much less likely that they will be able to fabricate data with reliable variances.

Taking covariates of interest (as fixed effects) into consideration, and accounting for the intra-class correlation among observations for the same group (random effects), pairwise differences are tested on the resulting residuals. The macro we created to perform these checks output a two-panel graphic of grouped box-and-whiskers plots and a heat map quantifying variance differences between groups.

STATISTICAL MODEL

Data within naturally occurring groups (e.g., family, staff member, or clinical site) tend to be more correlated with other members of the same group compared to members of a different group. Any analysis therefore needs to account for this intra-class correlation for valid between-group comparisons.

A mixed effects model takes this sort of intra-class correlation into account and includes two types of effects, fixed and random. Fixed effects describe the impact of known measured covariates (e.g., age, sex, weight, etc.). A random group effect (e.g., site, data enterer, etc.) is added to introduce the intra-class correlation among subjects in the same group. The basic model is:

$$Y_{ai} = \mu + a_a + \beta_1 x_{ai1} + \cdots + \beta_p x_{aip} + \varepsilon_{ai}$$

Here, Y_{ai} is the outcome for the *i-th* individual at the *a-th* site, μ is the overall intercept, the $\{x_{aik}\}$ are covariates accounting for variability in the outcome, and a_a is the random effect term for the *a-th* site. We assume that the $\{ a_a \}$ are i.i.d., normal random variables and independent of the $\{\epsilon_{ai}\}$.

The analysis is conducted in two steps. First, a mixed effects model is fit using PROC MIXED with the RANDOM statement included to produce the intra-class correlation among the observations in each defined group adjusted for the fixed covariates of interest. Scaled residuals are output from this model.

Second, the scaled independent residuals are then analyzed using PROC GLM. The HOVTEST option compares all pairwise comparisons of the defined groups using the Brown-Forsythe test.

Finally, to protect against type-I errors, PROC MULTTEST is applied to adjust the p-values for the number of pairwise comparisons.

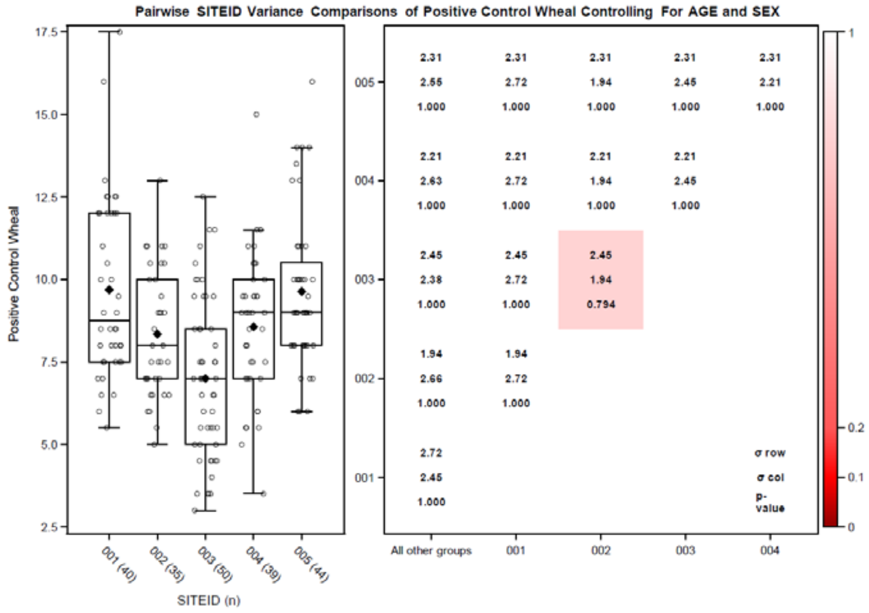
DATA SOURCE

We will use the same dataset and similar outcomes for all the following examples. Data was randomly generated to have the following properties:

- Data for skin prick test wheal sizes in millimeters was generated on each participant one time (e.g., one observation per participant).
- Data was generated for 5 different sites ‘entering data’.
- Skin prick test data is known to be associated with age and gender. Data for age and gender have also been included in the simulation.

EXAMPLE #1 – NO DIFFERENCES IN VARIANCE

We first want to see if there are any potential differences across sites related to the Positive Control allergen. In asthma and allergy trials, Positive Control allergens are included in skin prick test panels to (1) ensure that a subject has not taken any antihistamines prior to the skin testing and (2) assess the skin reaction to other allergens in relation to a standard histamine. Figure 4 displays the distribution and heat map for Positive Control skin prick test wheals by site ID adjusted for age and gender.



	All other groups	001	002	003	004
005	2.31 2.55 1.000	2.31 2.72 1.000	2.31 1.94 1.000	2.31 2.45 1.000	2.31 2.21 1.000
004	2.21 2.63 1.000	2.21 2.72 1.000	2.21 1.94 1.000	2.21 2.45 1.000	
003	2.45 2.38 1.000	2.45 2.72 1.000	2.45 1.94 0.794		
002	1.94 2.66 1.000	1.94 2.72 1.000			
001	2.72 2.45 1.000				

Asterisks indicate significant group variance comparisons after controlling for Sex, Calculated age in months and multiple comparisons.

Figure 4. Positive Control Wheal Sized by Site ID Controlling for Age and Gender

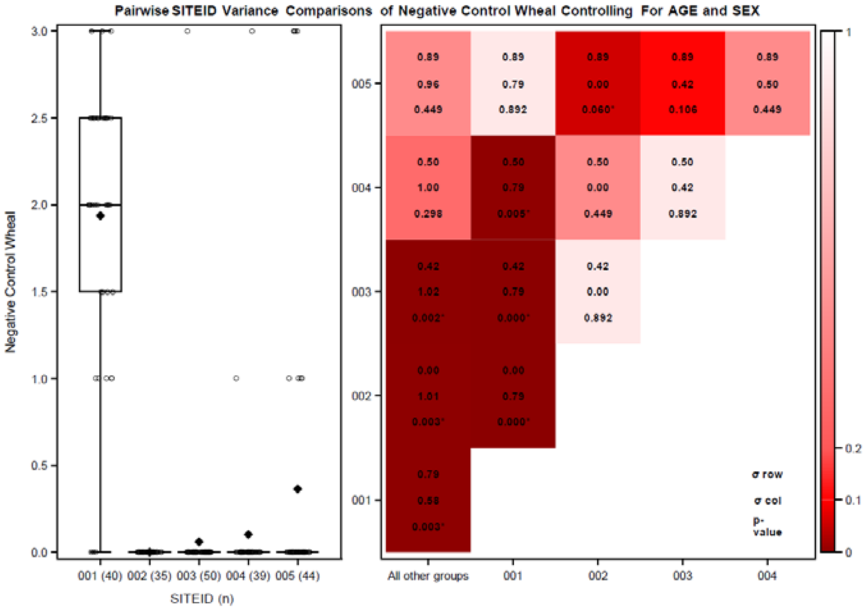
The grouping variable of interest ('SITEID'), outcome of interest ('Positive Control Wheal'), and covariates of interest ('AGE' and 'SEX') are all displayed in the title of the generated figure.

The grouping variable site ID is displayed on the x-axis. The number of observations at each site are annotated in the parentheses after each site ID. The plot on the left is a simple box-and-whiskers plot that displays all the data and provides a nice visual representation of the distribution of the data at each site. We can see that there are differences in mean values with specifically Site 003 having the smallest average wheal sizes. The heat map on the right displays all the pairwise variance comparisons of interest. The number displayed at the top of each cell corresponds to the variance for the group displayed in that row. For example, the variance for Site 005 is 2.31 which is displayed as the top number in each cell across the row for Site 005 (top of the figure). The number in the middle of each cell corresponds to the variance for the group displayed in that column. For example, the variance for Site 001 is 2.72 which is displayed as the middle number in all the cells in the Site '001' column. The number at the bottom of every cell is the p-value that results from the comparison between the variance of the group in the row to the variance of the group in the column. The first column 'All other groups' generates the variance by combining the other groups when the group in the current row is excluded. For example, in the bottom left hand cell, the variance for Site 001 is 2.72 and displayed as the number in the top of that cell. The variance of all the other Sites (i.e., Site 002, 003, 004, and 005) combined is 2.45 and displayed as the middle number in that cell. The red color in the heat map gets darker the smaller the p-value is indicating that the comparisons of the variances are statistically significantly different.

From these plots we can see that overall the variances of the Positive Control Wheals do not differ by site even though the means of the wheals do seem to be different.

EXAMPLE #2 – DIFFERENCES IN VARIANCE

Now first want to see if there are any potential differences across sites related to the Negative Control allergen. In asthma and allergy trials, Negative Controls are included in skin prick test panels to ensure that a subject does not have a reaction. If a subject does have a large skin reaction to a Negative Control Wheal, then we may need to exclude them from the trial as their skin may be excessively sensitive and could potentially change treatment interpretations. We expect that the Negative Control Wheals will almost all be zero or maybe 1 or 2mm. Figure 5 displays the distribution and heat map for Negative Control skin prick test wheals by site ID adjusted for age and gender.



Asterisks indicate significant group variance comparisons after controlling for Sex, Calculated age in months and multiple comparisons.

Figure 5. Negative Control Wheal Sizes by Site ID Controlling for Age and Gender

For this example, we can see that there is a very noticeable difference in how the Negative Control wheals have been recorded at Site 001 compared to all the other sites. For some reason, the Negative Control Wheal outcomes at Site 001 are larger than all the other sites. This could be a training problem (e.g., the site staff pushed the skin prick test too hard), a true difference in study population at Site 001 (e.g., the participants at that site are inherently more allergic), or there could be a contaminant in the extract used as the Negative Control at Site 001. There are a few more interesting aspects of this data. One, as indicated by the filled in black diamond on the box-and-whiskers plot, the mean at Site 005 is larger than the mean at Sites 002, 003, and 004. Two, in addition to having a much larger variance than the other sites, Site 001 is also the only site that recorded Negative Control Wheals to 0.5mm. Clearly, something is not quite right with Site 005 and some investigation as to why this Site is vastly different from then others needs to be addressed.

IMPLEMENTATION STRATEGY

Hopefully, the potential utility of these checks as illustrated in the above examples has been well highlighted. The following steps and actions are a basic guideline for implementing these data checks:

1. Identify the Data Fields for Checking

   This first step might seem trivial, but is of the upmost importance. Knowing which data has been collected and how it should be analyzed is a team effort that benefits from having input from multiple disciplines (e.g., statistics, programming, data management, etc.). Directly observed data on each subject like height and weight are good candidates for the digit preference check. For the variance check, any continuous data that can be quantified into groups could be a potential candidate. The goal is to be as inclusive as possible. Because of resource limitations, many trial outcomes are not scrutinized during on-site monitoring visits. This process, therefore, may well be the only opportunity to investigate the quality of these outcomes.

2. Timeliness of Data Checks

   Identifying questionable data early in the data collection phase is important. Questionable data may reflect misunderstanding in the data requirements, and resolving these misunderstandings before they become ingrained is essential. Moreover, if errors are not captured within a certain amount of time, the window of opportunity for correction will close (e.g., the subject goes off treatment and begins doing something that cannot be reversed). You can imagine what might happen if the skin prick test data shown for the digit preference check was not reviewed until all the data had been entered. At this point, it would be too late to intervene and try to understand how the differences at Site 001 had happened and could be fixed. Although the data might not be able to be corrected, at least the team would be aware to include Site as a covariate and that the assumption of equal variances will not be applicable for any analyses related to the Positive Control Wheal outcome.

3. Frequency of Data Checks

   Checks can be performed at specific intervals in calendar time; in anticipation of specific study milestones (e.g., an upcoming site visit, a DSMB report, or database lock), on an as-needed basis, or any combination of the above. Additionally, different checks may be performed at different frequencies depending on how quickly new data accumulate. The important points are:

   - Enough new data must have accumulated since the last data check to make running the check worthwhile
   - Performing the checks must not overburden data management and site management teams with too much activity (i.e., occur too often)
   - Performing the checks must not overburden the clinical sites (or other stakeholders) with too many data checks at a time (i.e., sending many requests too infrequently)

4. Running Programs and Identifying Questionable Data

   Each program is run on the identified data fields as determined in Step 1 above, starting at the point determined in Step 2 above, and according to the frequency determined in Step 3 above.

   Each of the statistical data checks presented in this paper generates a graphic showing all the data used in the checking algorithm. The specific questionable data are identified or otherwise highlighted in the graphic.

   The graphic should be uploaded to a central location (e.g., a study website) where it is accessible to all stakeholders. If a website is not available, then emailing results to all pertinent stakeholders could be acceptable; however, special care for safeguarding potentially identifying information and distribution of sensitive material should be of the utmost importance.

5. Query Distribution and Resolution

   Prior to sending queries to a site, everyone reviewing the data should determine if there is already a reasonable explanation for the questionable data. For example, one site may be using a different assay than the other sites or the site has already provided a comment or deviation that explains the questionable data.

   The interesting aspect of the two data checks presented in this paper is that it would be difficult for a traditional data manager to issue queries into an EDC system based on the results. These two checks are geared more towards a global sense of compliance with the protocol and potential fraudulent data entry opposed to other checks that would flag individual outliers. It is recommended that the graphics generated from these checks be reviewed with the study team as a whole to identify how/if the data that is highlighted can be addressed.

CONCLUSION

Although an EDC system already has a variety of range checks, consistency checks, and cross-field checks, this process is still vulnerable to transcription errors from source documents and even fraudulent data entry. This paper describes two supplemental procedures to identify data collection and data entry errors that may elude standard range and consistency checks. The Digit Preference and Variance Check help identify, visualize, track, and resolve questionable data. Statistical models are applied to identify questionable data by evaluating them relative to the data provided by the other subjects or data from the same subject. The data displays generated from these checks can be provided to study sites and other stakeholders for verification and resolution of questionable or suspicious data and can also help with the interpretation of the quality of the data. These checks not only help with cleaning the data, but also provides a basis for risk-based monitoring by identifying specific data fields, staff members, or sites that may require more thorough on-site monitoring.

ACKNOWLEDGMENTS

This work was performed as part of the Statistical and Clinical Coordinating Center for the National Institute of Allergy and Infectious Diseases (NIAID), Division of Allergy, Immunology, and Transplantation (DAIT) of the National Institutes of Health (NIH) grant UM2AI117870. Furthermore, this work was performed with the help of many of my colleges including (but not limited to): Jim Rochon, Brett Jepson, Spencer Childress and Liz Goodman.

RECOMMENDED READING

Kirkwood AA, *et al*. 2013. "Application of methods for central statistical monitoring in clinical trials." *Clin Trails,* 10: 873-806.

Venet D., *et al*. 2012. "A statistical approach to central monitoring of data quality in clinical trials." *Clin Trials,* 9:705-13.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kaitie Lawson
Rho, Inc.
kaitie_lawson@rhoworld.com