A Suite of Statistical Data Checks to Identify Questionable and Suspicious Data in a Central Management System

Kaitie Fernandez, Henry Bahnson, Spencer Childress, Brett Jepson, Katy Jaffee, Kyle Breitschwerdt, Liz Goodman, John Lim, Meagan Spychala, James Rochon

Rho, Chapel Hill, NC

Overview

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, 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.

Methods

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 impressions of "suspicious" data, the following statistical methods are applied:

1. Regression Models (multivariate and longitudinal)
2. Multinomial Tests
3. Cook's and Mahalanobis Distances

These models allow us to account for demographic characteristics and other important covariates that may account for their distribution. The residuals from these models are used to identify the outliers for further review. The longitudinal model, for example, uses a mixed-effect model with fixed effects for overall trends and random effects to account for subject variability.

Advantages Over Standard Checks

Using statistical techniques to identify suspicious data instead of relying solely on standard data management checks have numerous advantages.

Digit Preference
- Not possible in standard checks

Bivariate
- Fewer queries generated compared to range checks

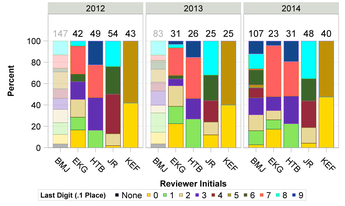
Longitudinal Mixed Effects Model
- Highlights improbable data that standard data checks would miss
- Does not query logical data that fall outside of reference ranges

Mahalanobis Distance
- Illustrates how sites are performing compared to one another
- Compresses high-dimensional data into a few key values to query

Digit Preference

Purpose
- Determine if there are data entry errors or fraud by analyzing the last digit in a numeric variable

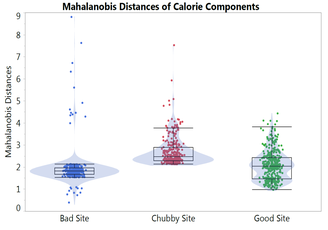
Statistical Method
- Multinomial test: $\Pr(\mathbf{x})_0 = N! \prod_{i=0}^{9} {\pi_i^{x_i}}/{x_i!} \quad \mathbf{x} = (x_0, x_1, \ldots, x_9), \sum_{i=0}^{9} x_i = N$
- H_0: $\pi = (\pi_0, \pi_1, \ldots, \pi_9)$ and $\pi_0 = \pi_1 = \cdots = \pi_9 = 0.10$
- Darkened bars represent reviewers with a distribution significantly different from the null hypothesis ($p < 0.001$)

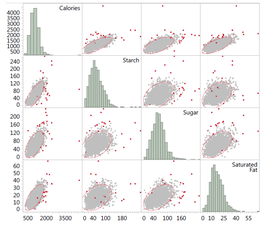


Mahalanobis Distance

Purpose
- Identify potential outliers and inliers in highly correlated multivariate data

Statistical Method
- Distance: $D_M = \sqrt{(Y_i - \bar{Y})^T S^{-1} (Y_i - \bar{Y})}$ Y_i = data for the ith row $\bar{Y}$ = row of means
- S = estimated covariance matrix
- MANCOVA methods are used to compare distances across sites.

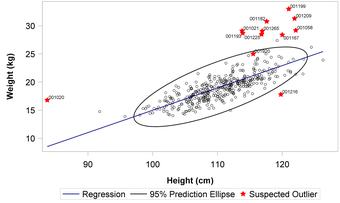




Bivariate

Purpose
- Identify unusual values by taking advantage of the correlated nature of data

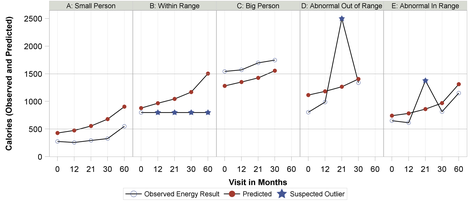
Statistical Method
- General Liner Model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$
- Cook's Distance: measures effect of deleting an observation; composite of studentized residual and leverage



Longitudinal Mixed Effects Model

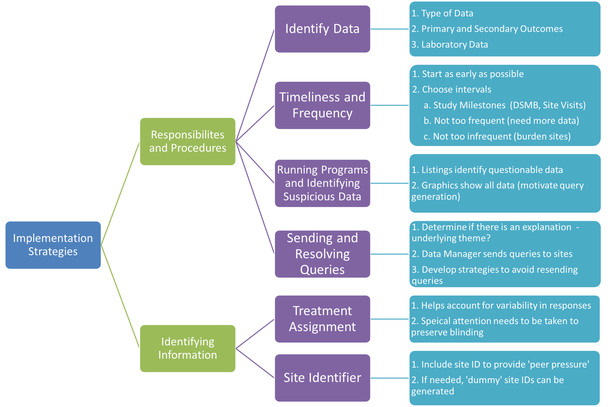
Purpose
- Identify unusual values over time by taking advantage of the correlated nature of data

Statistical Method
- Mixed Effects Model: $Y_{ij} = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_p x_{ij} + Z_{ij} d_{ij} + \varepsilon_{ij}$
- Fixed Effects: Age, Gender, Height, Weight
- Random Effects: Intercept (± slope)
- p = # of covariates of interest
- i = # of subjects
- j = number of time points



Implementation

Ensuring data quality takes cross-functional collaboration. Many different functional areas (Biostatistics, Data Management, Statistical Programming, and Clinical Operations) are involved with the implementation and maintenance of the data checks process. The graphic below outlines essential processes and decisions that need to take place for efficient implementation of the statistical data checks.



Future Endeavors

We plan to build our suite of statistical checks using different statistical methodologies (e.g. logistic regression, decision trees, clusters analyses) and create user and training guides so these methods can be practiced more widely across studies. We plan to implement these checks early on and during the study process for our federally funded projects.

Rho®
Giving flight to research

NIAID

This project has been funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, under contracts HHSN272200800029C and HHSN272200900057C .