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

Kaitie Fernandez, Henry Bahnson, Jeremy Wildfire, Spencer Childress, Brett Jepson, Maya Barton, Kyle Breitschwerdt, Liz Goodman, John Lim, Meagan Spychala, Richard Addy, Edie Vandy, James Rochon

Rho, Chapel Hill, North Carolina

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, “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.

Methods

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 Mahalanobis distance measures across sites. Generally, regression models are applied to account for demographic characteristics and other important covariates, and the residuals from these models are used to identify outliers. The longitudinal model, for example, uses a mixed-effect model with fixed effects for overall trends and random effects to account for subject variability. These checks were implemented as self-contained programming constructs, utilizing a variety of programming languages, with easily-defined parameters to generate models that can be adapted for most studies and audiences.

Results

The suite of data checks is illustrated using a study from the NIAID Immune Tolerance Network. Fabricated data were also added to datasets to test robustness. We highlight the strengths of this approach and discuss some of its shortcomings.

Conclusion

This suite of statistical data checks is an effective tool for supplementing current processes and ensuring data accuracy. It can also focus resources on specific data fields and clinical sites for efficient risk-based monitoring strategies.

Funded by NIAID/NIH (HHSN272200800029C).