

DATA INTEGRITY VALIDATION - PROACTIVE VS. REACTIVE

Douglas P. DeYoung, The Upjohn Co.

INTRODUCTION

This paper will compare two similar target animal safety studies and the data integrity methods used. The first study involved massive pre-submission data integrity validation requirements by the FDA's Center for Veterinary Medicine. The subsequent study included proactive methods employed for error identification and validation of data prior to analysis and submission, enabling significant reductions in the pre-submission checking requirements for this substantial (approx. eight million data values) bovine safety study. I will focus on the multiple facets of data integrity checking implemented in the second study including database design, data collection, electronic data entry, data transfer, and data analysis and on the system validation methods used to minimize post study program and data review requirements.

Purpose

I hope to provide you with some things to think about as to how you may be able to improve the validatability of systems. The methods of validation are extremely important, but to be effective the system must be validatable! I will compare some of the methods and results obtained by using a proactive approach to system and data validation. Overall system resources for programming, data validation, and report production with the proactive approach were more than 75% below those of the previous study. I would also solicit your feedback and suggestions for additional and alternative methods and approaches.

Project Overview

The two projects were target animal safety studies, involving different formulations of a

compound. Each study was to last three to four years. The statistical analysis and technical report production included analysis of the feed intake, production, physical and reproductive health of the animals and their offspring. Statistical analyses included frequency distributions of event occurrences, listing of bad or missing data values, arithmetic and least-squares means, statistical outliers and analysis of (co)variance. In addition, graphs of confidence intervals for a number of the variables were generated. The analysis includes some 285 different measurements for 150 animals in 5 different dose levels, and 7 phases. In all, the database consisted of over 8 million elements in 45 different source data sets with approximately 2 million data points actually to be analyzed. In the final report, it is anticipated that there will be nearly 11,000 tables and graphs.

THE REACTIVE APPROACH

Elements of the Reactive Approach

Conduct and analysis of the first project was done entirely by a contract research organization. The scope indeed, turned out to be magnitudes larger than originally perceived. The initial project was commenced with minimal time for planning and design of the analytic process and computer system. After the project was well under way and significant data was accumulated (but not entered into the system), design of the computer system was commenced, with various aspects of the system being developed by separate groups. Lacking a common design and coordination, multiple, less than compatible hardware and software elements resulted. Computer system validation and data entry integrity were not designed, but were implemented "as needed". The resulting computer system consisted of multiple

platforms, database designs, formats and variable names. In addition, virtually no documentation of the system was developed.

FDA Requested Data Validation

As completion of the first study report was approached, the FDA Center for Veterinary Medicine "suggested" that the company should go through all raw data and compare it to the computer database to validate the accuracy of the data entry process. ALL discrete data elements in the study data and all repeated measures for a randomly selected set (20%) of animals was reviewed.

Reactive Approach Results

Completing the statistical analysis and reports required approximately 12-14 months after the completion of the study segments. The data review required over 3,000 hours of reviewing the hard copy outputs of the data and verifying correctness to the raw data forms. All errors were corrected in the database and all affected tables and analysis reports were then recreated and updated. This resulted in a further delay of several months in submission. In addition, review of the project was slowed by the difficulty in understanding the analysis and data relationships and in navigation of the system and the data structures.

THE PROACTIVE APPROACH

The Proactive Approach Goals

Based on the experience with the former project, specific goals for the subsequent project were established. A system which provided the means to verify the accuracy of the data was considered a requirement. Additionally, we needed to minimize the manual writing/rewriting of data by employing electronic methods in both data entry and report production wherever possible. This reduced validation steps and simplified report production as well as decreased time and resource requirements. Inherent change in a

project of this size, duration and nature however, compelled us to develop a system with as much flexibility as possible. The analyses utilized also needed to be consistent across all aspects of the study for effective comparison, and reduced validation time. The capability to quickly analyze sub groups, create new analyses, and add, drop or change parameters was imperative to enable prompt reporting of interim as well final results and minimize time requirements for completing the final report. Since this was a GLP study, validation of the programming was required. Review by the FDA also implies the need for the system must be understandable by the non-programmer, non-statistician. As with any drug study, quality of the conduct and analysis must be considered an ultimate priority. Cost and time required for this project also mandated the analysis be complete and correct the first time.

System design objectives

One objective was to develop a computer system which provided consistency throughout all analyses. In addition, the system should enable efficient programming and development of the system. Finally, the system should provide for efficient production of the final report. A second objective was reduction of development time so as to enable complete development in a time frame that would provide analysis and feedback during the study trial period, not just upon it's conclusion.

Due to the magnitude of the project, it was apparent from the beginning that we could not identify and define exactly what would be needed prior to the commencement of the study. It would be necessary to develop a system which provided for change and minimize the impact of additions and changes to the study parameters.

Based on the significant verification requirements of the prior study, it was mandatory that we provide mechanisms to validate the integrity of the database and

computer system. We expected not only to avoid the mammoth manual checking required in the former study, but to facilitate the approval process. It would be important that we could validate the developed system utilizing known results, to assure its integrity.

As with any business, delays in getting the product to market are costly. Time required to produce or update the technical report had to be significantly reduced.

Elements of Proactive Approach

The second study was very similar in scope and was to follow virtually the same analysis and report format as the initial study. This time, the company chose to do the analysis and report generation in house. Considerably more time (approx. 4 months) was provided prior to the commencement of the study for coordinating the design of the study with those conducting the study, the study monitor, statisticians, and computer personnel.

Programming would be done so as to easily modify and reuse modules. This was accomplished by developing program modules for each analysis and evaluation element which would then be used in every program to accomplish that function. Program modularity provided multiple benefits including significant reduction in development time, consistency of analysis, reduced validation requirements, and consistency and ease of modification.

Data integrity was to be validated using multiple methods at each point of data handling, as will be discussed later. Uniformity throughout the project through consistent technologies, single programming language, and identified methods and approaches in the computer programs and macros. In addition, the final technical reports would be produced using programming rather than manual entry, eliminating potential errors and providing for uniform output formatting.

Proactive Approach Methods

The initial system design period prior to study commencement was critical to development of an effective system. SOP's were developed for computer hardware and software, data transfer methods and transferred data integrity validation procedures. Documentation of system design, file structures, variable definitions, and a data dictionary provided a basis for communication. This was critical to provide consistency between those conducting the study, data entry personnel, medical and statistical personnel at Upjohn doing the monitoring and analysis of the study and computer systems personnel. This also facilitated modification to the study design. System and analysis procedures were also thoroughly documented.

SAS® macros provided the mechanism to achieve the program modularity desired, but, did not provide flexibility which would allow them to be utilized in different ways, without writing multiple routines. By using macro variables, we were able to develop one macro for each statistical analysis and provided the flexibility to identify varying elements (e.g. different models in proc GLM). Further, by utilizing SAS macros which could be included in any program, changes were able to be implemented in a single location and carried through into all calling programs.

Data sifting and separation and output methods likewise were developed as macro programs for consistency and simplification. Again, we made significant use of macro variables for values which change (e.g. table numbers, number of variables measured). Additionally, it was determined that all analysis programs should go directly to the raw data sets and not use permanent calculated data sets. This eliminated the possibility that interim data sets did not match the current raw data sets, especially for during interim study analysis.

Data verification would be critical to the success of the study and was to be

incorporated at all stages necessary to assure validity of the data utilized in the reports. Likewise, the computer system needed to be validated to assure the integrity of the report. I would like to discuss these items in further detail in the remainder of the paper.

DATA INTEGRITY AND VALIDATION PROCEDURES

Data Sources

Data was collected both manually and electronically and by a variety of personnel and mechanisms including field technicians, veterinarians, feed dispensers, scales, various laboratories and weather stations.

Data Flow Events

Events affecting and accomplishing data creation, each creating the potential for error as well as the opportunity for verification, included all of the following: the scheduling of daily tasks for each animal in the study, data collection from multiple sources as noted earlier, tracking of samples sent to independent labs and those to be analyzed by the study organization, electronic downloading of data, manual data entry, electronic range checking within the data entry programs, visual verification of entered data, data transfer to Upjohn, transfer of data and electronic field verification checking, graphical analysis of outliers and statistical analysis.

Data Integrity Personnel

Along with the potential for error, multiple personnel provided diversity of data integrity. Field technicians, data entry personnel and checkers, the contract organization quality assurance director, Upjohn data technicians, programmers and statisticians all had responsibility for data and system integrity.

Concerns Addressed

Problems which were addressed by this system included: omitted data collection; missed data

entry; incorrect data entry; lost lab data; analysis inconsistencies and post study data validation impact.

Computer system integrity

Characteristics which facilitated the validation of the computer system included the above data validation methods which did provide significant system validation. In addition, the use of a single software technology (SAS) reduced potential errors introduced in the interfacing of unlike technologies. As noted earlier, the consistent use of modular programs minimized the validation requirements which would result if each program were unique. The decision to do data filtering within the programs increased the validation requirements therein, but significantly reduced validation requirements for separate programs and the need to constantly validate the consistency of multiple data sets.

Major hurdles

Major hurdles which have inhibited the implementation of this system were primarily in the area of communication and follow through. Decisions and alterations to the study conduct were frequently not passed to all necessary parties in a timely fashion. Significant within this category was frequent failure to pass changes in the database to appropriate data entry and field personnel, and failure to adequately prepare site personnel with regard to data interpretation and entry requirements. Data entry requirements were greater than anticipated and data entry lagged by up to several months. This prevented effective use of programs intended to monitor the ongoing study and provide interim analysis, as well as delay validation of the system with "live" data. Lastly, as anticipated, the study continued to change and the system required modifications and enhancements, sometimes impacting the entire system. In general however, the relative ease in implementing modifications

confirmed the value of the approaches utilized and also further validated the system.

Proactive approach results

FDA response has not been received so we can only evaluate the effect based on what we hope would occur. The data integrity checking provisions have provided data sets with very low error rates and provided a mechanism to further identify errors. The report production process required approximately 1 week from receipt of final data, compared to 6-8 months on the first study. Programming and data verifications have been accomplished (and documented) and were completed ahead of schedule and with programming resources 40% below projections and more than 60% below the first project. Changes can generally be accomplished to the system in minutes, not days. Total system resources for programming, data validation, and report production with the proactive approach were reduced by more than 75% from the previous study. It is expected that further savings will result in the FDA review process.

Keys to success

Planning and design

Having some time up front to develop a good plan for the system design, and continuous refinement of the design of the system are critical to providing an effective, flexible system. The design would not have been successful without the consistent involvement of representatives of all aspects of the project including study director, study monitor, data entry personnel, quality assurance, statisticians and computer systems personnel.

Standards/consistency

Standards are imperative with a project of any size, but especially critical when dealing with one of this magnitude and with the wide dispersment of personnel involved. Naming conventions were established with a target of consistent,

meaningful variable names which clarified function.

Documentation

System documentation was developed, constantly updated and followed by the entire project team. Program in-line documentation is critical to understanding the programs, especially as the use of SAS macros and macro variables tends to reduce visibility of those functions within the programs. Also, documentation of the function of macro routines and the definition and default values for macro variables is necessary to provide maximum usefulness.

Modularity/flexibility

The modularity of programming provided multiple benefits including significantly reduced programming costs, flexibility, and reduced validation time and cost. Our goal was to use available developed macros wherever possible, both for analysis and output. Where unique situations were involved, utilizing existing macro routines as a starting point provided consistency and increased the speed of development.

Testing/validation

The data integrity and system validation that was accomplished throughout the development and conduct of the system provided the ability to successfully produce the final technical report in dramatically reduced (-95%) time and with significantly increased confidence in the quality.

Teamwork and communication

The importance of communication within the project team to provide awareness of applications and potential refinements has been the foundation on which the benefits have been obtained. Maintaining consistency of labeling and programming approach likewise are critical to successful implementation. Weekly project team meetings facilitated communication and project review.

SUMMARY

Through the design and development of a computer system which included re-usable programming modules, and multi-faceted data and system validation techniques, we were able to provide a validatable system in significantly reduced time, with fewer resources, which provided fast, accurate report generation upon study completion.

ACKNOWLEDGEMENTS

I wish to thank H. Jeffrey Rostami and Mohammed J. Al-Adhami for their significant roles in formulating and implementing these ideas and their technical review of this paper.

TRADEMARK NOTICE

SAS is a registered trademark of SAS Institute Inc. in the USA and other countries.

AUTHOR

Douglas P. DeYoung
The Upjohn Co.
301 Henrietta St. 9162-32-2
Kalamazoo, MI 49001-0199
(616) 384-9661 FAX (616) 385-5286