Paper 131-26

Data Modeling in Clinical Data Analysis Projects

Gajanan Bhat, PAREXEL International Corporation
Shy Kumar, DataFarm, Inc.

ABSTRACT

Development of data model in clinical trial management environment is an ongoing process but of utmost importance for a good clinical information management. Companies are trying to come up with efficient global standards in data modeling to work with. Global standards are also important and imminent in view of the new electronic submission guidelines of Food and Drug Administration (FDA) for New Drug Application. Most important factors to consider when designing the analysis and reporting data structure are standard data structures, metadata, documentation, and data normalization. Often, the normalized data structure is in conflict with the raw data structure obtained from several clinical data sources. The main objective of this paper is to provide a data model for the analysis-ready data using transformed data structure explaining several levels of data normalization from the Data Warehousing perspective. The paper discusses specific areas such as developing data model, features of data model, defining source data structures, defining derived variables, and also discusses the strengths and weakness of this modeling approach.

INTRODUCTION

When data (referred to later as raw data) from the Case Report Forms (CRFs) are entered into a Clinical Database System (CDBS), it tends to be organized in a more vertical format. This means that there is more than one record per patient in a given data table or data set. Though this structure is desirable for the data collection, entry and cleanup process, it is not always analysis-friendly. That is, it is not easy to perform statistical analysis needed to produce reports from the data. If used raw data, it involves duplicating efforts of transforming the data in each report and also adopts a different approach in each report. In many cases, there is usually an added complication that the physical platform and machine on which the CDBS resides is not the platform on which the statistical programming is done. Also, the collaborative research often brings the situation of obtaining raw data from different sources that are very often structured differently even if the information collected is very similar.

Many companies in recent years have been striving to produce global and project standards. These standards are the basis for the data sets from which analyses and reports are produced. The aim of these standards is to provide naming and definition for variables that occur in any study that the company uses. Food and Drug administration has brought the new electronic submission standards recently and requires the new drug sponsors to comply with the electronic submission standards for new drug application (NDA) as soon as possible. This calls for improvement in data modeling, metadata, and documentation. Thus, the paper discusses few standards of data structure and provides a model that can be adopted to work with for a clinical project.

One of the main sections of the NDA is the case report Tabulation where in sponsors are required to provide data definitions for the database that they submit as part of the NDA submission. This paper discusses those standards and provides a model that can be adopted to work with for a clinical project.

Clinical Data Environment

Clinical Trial Management describes the general process that the clinical data pass through from data collection until data is analyzed for the statistical reports. Project standards augment the global standards already defined to include extra data sets and variables that are specific to a particular clinical project. The standard data sets and variables for the project are then used to define the analysis data sets for the actual study. Once the raw data structure is known and the analysis data structure has been specified, the data must be run through a transformation process to convert it from one format to the other.

Normalization

For reasons of both data integrity and performance, it is the goal of a production database, typically, to represent each fact or data item in only one place. The data redundancy not only causes potential errors in data maintenance; it also requires added storage. Normalization is the technical name for the process that reveals and then eliminates data redundancy. The normalization in RDBMS is a key factor of a good database design. The normalized database in the relational database system gives advantages such as elimination of redundancy, ease of use, update, and modification. Each table in a normalized database has a primary key, which is a field or fields that uniquely identifies each record in the table. The tables may also have one or more fields that serve as foreign keys, which are fields with the same name as a primary key field in another table. Normalization is done by matching the tables with first, second, and then, third normal forms.

Clinical Data Complexity

Analysis data structures in clinical projects do not completely conform to normalization in real life for many reasons. Main reasons are attributed to the way programmers and statisticians use the data at that stage to create the reports and the nature and the source of data. Normalization, as discussed in the previous section, makes the task of updating and modification in the original data tables easier as it eliminates duplication of information. However, the main place of update and modification of data is the original CDBS and not the analysis database. The other important reason is that statistical programming needs not the completely normalized structure in the analysis data structure. This will pose an extra overhead of combining data sets to create meaningful and workable data sets for the production of Tables and Listings. Thus, it leads to normalization and de-normalization at the same time. Also, it is better to keep some basic information such as visit number, visit date, etc. in every analysis data set.

FDA Guidelines for Electronic Submission

In order to expedite the drug approval process not compromising the approval standards and also with the advantage of the current information technology, FDA requires sponsors to provide the regulatory NDA submission in electronic format. This provides many advantages such as global standards in submission formats, increased efficiency gain in sponsors' R&D efforts, faster and more accurate reviews and approval. However, this requires many deviations from the current submission practices, and calls for new standards in formats, data models, and metadata. One of the main sections of the NDA submission is Case Report Tabulation (CRT). This section includes Patient Profiles and SAS databases in transport files that include data definition files.

The data definition file is mainly a metadata file describing data structure, data model, definitions of contents, sources, etc. This requires a detailed data model to be developed in clinical data management and analysis projects.

CLINICAL DATA MODEL

The Role of Metadata

Before beginning data modeling from the data warehousing perspective, an extremely thorough understanding of the user's requirements is necessary. This includes a thorough knowledge and assessment of metadata. Aside from the obvious concerns of data cleanliness, integrity, and integration, the understanding of data value cardinality and distribution pattern are of utmost importance in determining various keys and indexing of final data models.

Data Model Overview

One of the most important decisions affecting the modeling efficiency is the schema with which you choose to model your data. This also depends on the level of granularity (level of details) that is required for the project. The user requirements may vary depending on the type of user. In the clinical data analysis projects, from the data warehousing perspective, the summarized or granular structure is preferable, where as less normalized data structure is preferable from the statistical analysis perspective to reduce the overhead involved in creating reports.

Two subject-based data models are of widely use. They are Star Schema and Snowflake Schema. In the Star schema, relational tables are used to model the subject of interest. The schema utilizes implicit modeling of relationships. Because it holds large volumes of numeric factual data about the subject, the Fact table (mostly Keyvar data) is typically the largest table in a subject-based schema. The Fact table is surrounded by many "Dimension" tables (such as Demog, Vital Signs, AE, LAB, MEDS, etc).

Another popular subject-based schema is the Snowflake schema. The difference is that it is more normalized that Star schema. This reduces redundancy, but at a query construction and performance cost. The Snowflake schema is recommended when data storage requirements are extremely large. The attributes defined for each data element in the clinical database are described in the Table 1.

Table 1. Clinical Database Attributes

Column Name	Description
Data set name	Name of the data set
Description/Label	Description of the data set regarding its purpose, contents, and key information/Label given to the data set
Data set program	Name of the program
SAS Engine	Version of SAS Engine used to create the data set
Observations	Number of observations/records in the data set
Variables	Number of variables/columns in the data set
Index	Names of the indexes defined in the data set
Compression	Yes/No
Protection	Yes/No
Primary key	Unique identifier variable(s) used to distinguish the records
Sort by	Name of the variable(s) that the data set is sorted by
Variables	List of the all the variables
Type	Data type: Numeric/Character/Date/user defined
Length	Length of the variables (8 for numeric)
Format	The formats associated with the variables
Informat	The informat associated with the variables to read in
Label	Labels associated with the variables

Keys

There are several considerations when selecting primary keys. They should uniquely identify individual patients and records and should be as short as possible. The primary keys must be identical across each of the patient level data files i.e. the variables have the same length and type. Most used variables as primary keys in the clinical databases are Protocol and Patient ID.

Indexing

One of the primary determinants of query performance against granular (lowest level data) is how effectively indexes are used to aid in data location and retrieval. Indexing is not a simple topic and performance can be degraded if is not careful in creating and using them. The paper does not discuss the general indexing information such SAS system indexing using Btree indexing system.

Naming Convention

Standard naming convention is important for many reasons. The names of the variables/data sets should point out the purpose and nature of variables. Also, the standard makes it easier to identify and compare across different data sets.

Defining the Data Structure

Consider the following Vital Signs Data set example in Table 2.

Table 2. Denormalized vs. Normalized Data Structure

Denormalized Data Structure

Visit	Systolic	Diastolic	Weight
1	128	84	161
2	125	82	158
3	121	81	159

Normalized Data Structure

Visit	Parameter	Value
1	Systolic	128
1	Diastolic	84
1	Weight	161
2	Systolic	125
2	Diastolic	82
2	Weight	158
3	Systolic	121
3	Diastolic	81
3	Weight	159

Invariably the question arises, "Why not just use denormalized or rather more robust data sets with all the data in one observation instead of Fact and Dimension tables (data sets)?" This situation most often creates far too much redundant data for project Database in the cases where in considerably large storage is required. The models using Star Schema and such represents a viable option to allow reasonably good query performance and elimination of some redundant data.

The analysis data structure must be optimized to facilitate reporting and analysis process. The most important factor to consider when designing the analysis and reporting data structure is data normalization. The efficiency and ease of extracting the data from the database and use them for reporting are very important. Depending on the nature of data, denormalized data structure is better suited for some data sets such as Vital Signs where as normalized data structure is better suited for some data sets such as labs. For example, Single Vital Signs listings and tables include all parameters where as there will be separate listings/tables for each lab or efficacy parameters.

CONCLUSION

This paper describes in brief, the transformation of clinical data from CDBS to analysis-friendly structure that would facilitate reporting and analysis process. Also, the paper gives a brief account of the FDA guidelines in terms of data model for the new electronic submission. The paper also compares the normalized as well as denormalized data structure for its strengths and weakness. Depending on the way the clinical outputs are produced, analysis data structure

should be a combination of both structures. The main goal of clinical data structure for statistical analysis is the efficiency and ease of use. This influence the data structure rather than traditional relational database structure and normalization process.

CONTACT INFORMATION

Gajanan Bhat
PAREXEL International Corporation
195 West St.
Waltham, MA 02145
Phone: 781-434-4617
E-mail: Gajanan.Bhat@PAREXEL.com

Shy Kumar
DataFarm, Inc.
275 Boston Post Road, Suite 2
Marlboro, MA 01752
Phone: 508-624-6454
E-mail: Shy@datafarminc.com
URL: http://www.datafarminc.com