One-PROC-Away: The Essence of an Analysis Database

Russell W. Helms, Ph.D.
Rho, Inc.
Chapel Hill, NC
RHelms@RhoWorld.com
www.RhoWorld.com

Presented to ASA/JSM: San Francisco, August 2003
One-PROC-Away Analysis Databases

• Outline:
• Introduction to Analysis Databases
  – What is an Analysis Database?
  – Comparison of Clinical Trial Databases (What an Analysis Database isn’t)
• Why Clinical Trial Data Warehouses have failed
• Why one-PROC-away Analysis Databases succeed
• How much is too much? Where one-PROC-away isn’t worthwhile
• Conclusion
What is an Analysis Database?

• To a user who is a programmer or statistician, an analysis database is:
  – a collection of integrated [SAS] datasets,
    • This includes derived variables (e.g., change from baseline to endpoint) and statistics (e.g., mean baseline score, averaged over multiple baseline visits).
  – with excellent documentation and metadata
    • metadata is data/information about data (the other Essence).
    • Not the topic of this presentation.
  – so that most requests can be handled by simple one-PROC “programs.”

• The mantra: “Answers are one PROC away.”
  – One PROC, method, function, …
What One-PROC-Away Means

The primary objectives of an Analysis Database are:

• To facilitate rapid, scientifically valid analysis without additional data manipulation.
  – Goal: data files that are “one PROC away” from scientifically valid answers.
    • “One PROC away” files can be analyzed directly, without data manipulations.
    • “One PROC away” and scientific validity imply that
      – much thought has gone into the design of the database and many preparatory processing and validation steps have been performed.
  • Sometimes easy, sometimes impossible. Usually difficult: need judgment to decide when worthwhile.
Differences among Types of Clinical Trials Databases

• What are the differences among:
  – An analysis database,
  – An operational (CRF, “domain”) database, and
  – A data warehouse?

• The primary difference is in ease of use for analysis.
  – An analogy can help clarify this difference.
Differences among Types of Clinical Trials Databases

An Operational Database is something like crops or livestock in a field…
Differences among Types of Clinical Trials Databases

The CRF or “operational” database system is efficient for “growing,” field-processing, and harvesting data…
Differences among Types of Clinical Trials Databases

But a lot of additional preparation is required before the data are ready to “serve.”
Differences among Types of Clinical Trials Databases

A “data warehouse” is a lot like a freezer....
Differences among Types of Clinical Trials Databases

Like frozen food, the data in a data warehouse are well preserved, but significant additional preparation is still needed.
Differences among Types of Clinical Trials Databases

An analysis database is comparable to a fine meal – all the preparations have been completed; everything is ready for use.
Comparison of CT Databases:
Analysis Database

• **Goal:** To facilitate rapid, scientifically valid analysis without additional data manipulation.

• **Structure:** based on anticipated queries and statistical analyses.
  – Each dataset is structured to facilitate a different set of analyses (per-patient, per-visit, AE’s, …)
  – Each record contains data collected, derived, computed and integrated from many different CRF pages,
  – Many values exist in multiple formats and locations to facilitate analysis.

• **Preparation:** designed by a combination of clinical scientists, biostatisticians, and epidemiologists, based on scientific objectives of the analysis
  – Thinking, specifying (metadata), solving scientific issues, programming
Comparison of CT Databases:

Operational Database

- **Goal:** facilitate data capture, error detection and correction, etc.
- **Structure:** incorporates the organization of a study’s CRF.
  - The data is usually “normalized”—so that each value only occurs once. The database changes as the data are cleaned, and changes only need to be made in one place.
- **Preparation:** set up by data processors, using the study protocol and CRF, to create tools for data entry (or other data capture), error detection and correction, etc.
  - The content is created by data processors who enter data, comments, corrections, etc., during data management processes.
Comparison of CT Databases:
Data Warehouse

• Goal: Serve as a repository for data files from a variety of sources, and sometimes to serve as a source of data files for future (integrated) analyses.
• Structure: primarily inherited from individual source datasets.
  – Primarily copies of original data files from individual studies.
    • A variable often has different names, formats, and other characteristics in different data files.
    • Data structures typically vary from study to study.
• Preparation: created by IT/programmers who “populate” the data warehouse by:
  – Creating new directories for files from new studies,
  – Copying the files into the directories, and
  – Creating integrated data files upon request.
Why Clinical Trial Data Warehouses Fail

• The goals do not match!

• The goals of a Data Warehouse are (Kimball, ’96):
  1. The data warehouse provides access to corporate or organizational data.
  2. The data in a data warehouse is consistent.
  3. The data in a data warehouse can be separated and combined by means of every possible measure in the business.
  4. The data warehouse is not just data, but also a set of tools to query, analyze, and present information.
  5. The data warehouse is the place where we publish used data.
  6. The quality of the data in the data warehouse is a driver of business reengineering.

• These apply to commercial data, NOT clinical data.
Why Clinical Trial Data Warehouses Fail: They are not one-PROC-away

- The 80/20 rule:
  - In clinical data, typically 80% of the “work” of an analysis is restructuring the data for analysis, 20% is performing the analysis.

- But Data Warehouses primarily inherit their structure from operational database!
  - Operational databases are highly normalized because they change.
  - Computer Science classes teach IT consultants to normalize.
  - Commercial data warehouse software expects normalized data.
  - But commercial data are “simpler” than clinical trials data:
    - many rows, few columns
    - simpler relationships
    - Possible to force users to “use correctly” (human research is very different), so there are few “science projects” in figuring out how to analyze them…
  - Normalized data is not one-PROC-away!
Why Clinical Trial Data Warehouses Fail: They are not one-PROC-away

- Data Warehouses are sold as a software solution.
- But Analysis Database creation is primarily a science project: 90% science, 10% software
  - General categories of typical scientific issues:
    - Missing Data
    - Extra Data
    - Protocol Violations (People not following the plan)
    - Planning Violations (Planning not anticipating the people)
    - Data integration conversions
- Series (hundreds or thousands) of decisions
  - Who? When? Documented?
Why Analysis Databases Succeed: They are one-PROC-away

• The 80/20 rule:
  – In clinical data, 80% of the “work” of an analysis is restructuring the data for analysis, 20% is performing the analysis.
  – In a one-PROC-away Analysis Database, the 80% is done and validated ahead of time.
    • This encourages extremely rapid analysis.
    • This allows statisticians to focus on statistics.
    • This can facilitate and accelerate the drug development process, especially study report creation and submission (e.g., NDA, CTD) creation and support.
Why Analysis Databases Succeed: They are one-PROC-away

• Doing the work ahead of time means…
  – Spend less time getting the answers when they are very expensive.
  – A turn-around time of minutes for fully validated answers when…
    • answering FDA questions post-submission,
    • answering competitor’s challenges, and
    • answering marketing questions
  – Facilitates sensitivity analysis.
  – Facilitates exploratory analysis, e.g., leading up to a Phase-III.
  – Even years later, scientifically valid analysis is quick and easy because…
    • All the difficult, annoying data manipulations have been done, and
    • All the difficult, annoying data problems and scientific issues have been resolved and/or documented.
    • Most programs consist of “one PROC” to create output (ODS), then formatting output. They are easy to understand and easy to reproduce.
How much is too much? AE example

Table XX

Number and Percentage of Subjects with Adverse Events by System Organ Class

<table>
<thead>
<tr>
<th>System Organ Class</th>
<th>Total</th>
<th>n</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Subjects</td>
<td>xxx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastrointestinal disorders</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>DRY MOUTH</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>CONSTIPATION</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>NAUSEA</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>Infections and infestations</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>INFLUENZA</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>PNEUMONIA NOS</td>
<td>xx</td>
<td>(xx.x)</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AE example

• AE: System-Organ-Class Table
• Why it is hard: the percentages in this analysis require one dataset structure for numerators, a different structure for denominators
  – Made worse by subsetting on demographics
  – Made even worse by considering severity or other common ordinal outcomes
• We invented a one-PROC-away structure for this problem, but we abandoned it.
  – Too disk-space and processor-time intensive.
  – Non-intuitive data structure confused users.
AE example

- True one-PROC-away for AE’s is too expensive.
- But it doesn’t matter because…
  - We can solve the scientific issues ahead of time (e.g., extra variables to solve partially missing date comparisons).
  - Then the required manipulation is standard. So we could write our own “PROC” using standard macros (or SAS could write PROC AE…)
- Be careful with the data torture!
  - Solve the scientific issues, but create a data structure that makes sense and is intuitive.
- Use some judgment. Experience helps.
How much is too much?  
Structures and Variables

• One-PROC-away from what?  
  – “Vast majority of scientifically valid questions”.

• Guessing all the questions that people might ask is very hard and expensive.

• We may gain much advantage from the fact that there are really only a dozen or so analysis structures. Use them.
  – But what variables should we include?  
  – Anything needed for planned analysis. But unplanned?
  – We call these Repository variables.
Repository variables

- What Repository variables should we include?
- Definitely include “Selection” variables—they are easy to create.
  - Demographics, population variables, treatment assignments, etc. should be on every dataset, allowing for easy subset analysis.
- Probably include:
  - “Sensitivity” variables—different decisions for some scientific problems so that you can evaluate the impact of them.
    - E.g., One SBP variable with all data, another with SBP=0 values set to missing
  - Centered variables for modeling.
Repository variables

• Probably include (continued)
  – Common flags for analysis. For example:
    • AE’s that are ongoing at end of study.
    • AEs that were stopped the last day of medication intake.
    • We gain a lot by the ability to combine many simple flags.

• Other, results-based questions are very hard.
  – E.g., median response over a series of diary entries when the primary analysis is based on means.
  – If these questions get asked, you’ll wish you had them, but you’ll hate the process of making & validating them.
  – Up-front costs and complexity versus rapid turnaround.

• Experience helps in this middle ground.
Conclusion

• There are many steps in the creation of analysis databases that require acute **scientific** judgment.
  – Some decisions are “no-brainers” that do not require judgment.
  – As in other areas of science, many really important analysis database decisions require extensive training, experience, deep insight, and other characteristics of some humans, aided by computers.

• This intensive scientific process leads to a database that is (all together now)

  One PROC away!
5. Conclusion

• The presentation is available on the internet at:
  www.RhoWorld.com

• Email questions, comments, etc. to:
  RHelms@RhoWorld.com

• UNIX types: you don’t have to use capital letters…
• Thank you!