The *SCIENCE*™ In Analysis Databases:

The Value and Challenge of One-PROC-Away

Russell W. Helms, Ph.D.

Rho, Inc.

Chapel Hill, NC

RHelms@RhoWorld.com

www.RhoWorld.com

Presented to FDA/Industry Statistics Workshop, September 19, 2003, Bethesda, Maryland

The *SCIENCE* in Analysis Databases

- Outline:
  - Introduction: What is an analysis database?
    - In comparison, What is an Operational database?
  - One-PROC-Away
    - Why One-PROC-Away is so valuable
    - Why the alternative is dangerous
  - Scientific Issues: 90% Science, 10% Software
    - Examples of Scientific Issues solved and documented in an Analysis Database
  - 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
  - so that most requests can be handled by simple one-PROC "programs."
- The mantra: "Answers are one PROC away."
  - One PROC, method, function, …

What is an Analysis Database? Types of Analysis Databases

- A Study Analysis Database (SAD) contains data, metadata, documents, etc. from one study.
  - SAD construction is the focus of this presentation.
- A Compound Analysis Database (CAD) contains data, metadata, documents, etc., from all the clinical studies of the specific compound.
- A Product Analysis Database contains all of the SADs, and the latest version of the CAD.

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.

The Analysis Database Secret to Success: One-PROC-away

- Why 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.

The Analysis Database Secret to Success: 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
    - anwering 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.

What's the Alternative to One-PROC-away?

- In a *Relational Database* (e.g., 3rd normal form or hyper-normalized), each value or measurement is stored only once. Why?
  - Computer Science classes teach IT consultants to normalize.
  - This is a storage of convenience.
  - If data will be changing, storing a value in only one place facilitates updates. For instance, operational databases are highly normalized because they change.
  - If disk space is limited, storing a value only once preserves a limited resource.
  - There are other reasons…tradition…

What Are Views? Why use Views?

- However, to “use” the data, we must use a “view”.
  - A view is a method of rearranging data from a format for storage (normalized) to a format useful for analysis.
  - For instance, we might want to view a data display with one record per subject, displaying gender and ALT measurements at each visit.
    - From a normalized database, this might involve joining two datasets and transposing one of them.
    - The IT response: That’s “Just a few lines of code!”
      - Or (worse): A computer can do that!
      - SORT, MERGE, TRANSPOSE.
        - » If “All Goes Well”, maybe 12 lines of code…

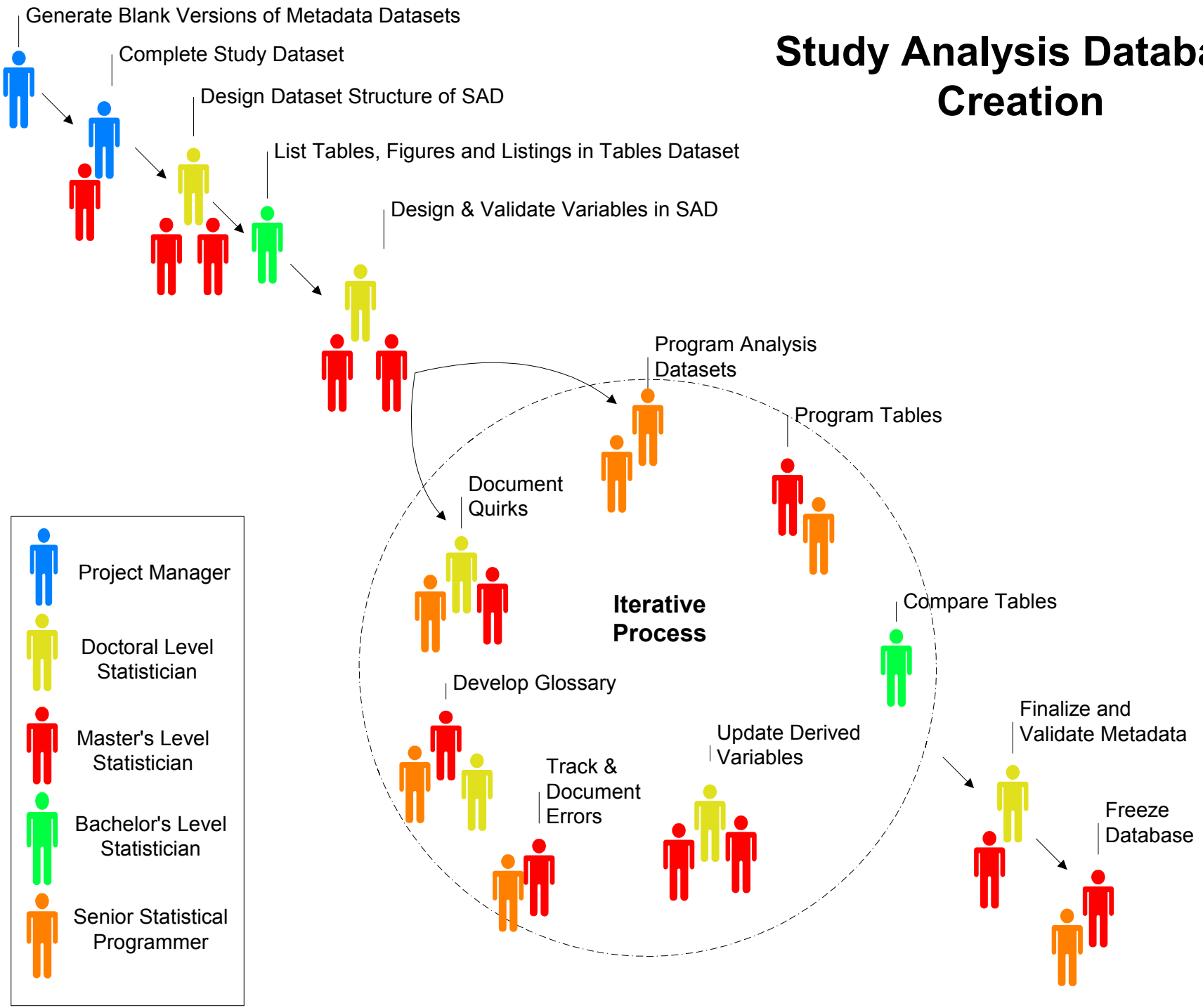
What Are Views? Why use Views?

- A normalized database is often best for commercial 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…
  - Very often, "All Goes Well", so reorganizing for analysis is often "just a few lines of code!"

Why NOT use Views for CT Data? 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.
  - This is NOT "just a few lines of code!"
    - e.g., 121 Pages of code…
- Creating an Analysis Database is a large *science* project.
  - Each step involves many research scientists.
  - Each step is documented thoroughly.
    - This documentation lives inside the database.

Study Analysis Database Creation



90% Science, 10% Software

- Analysis Database creation is primarily a ***science*** project.
- Series of decisions
  - Dozens in a SAD, hundreds in a CAD
  - Who? (Computer vs. Programmer vs. Scientist)
  - When? (Beforehand, while the study is fresh, or years later?)
  - Documented? (Not at all, on a piece of paper somewhere, or inside the database? Human-readable?)
- The following are real, fake examples.
  - Real: I have had to deal with each of them in the last year.
  - Fake: I am making up the specifics.

90% Science, 10% Software

- General categories of typical issues:
  - Missing Data
  - Extra Data
  - Protocol Violations (People not following the plan)
  - Planning Violations (Planning not anticipating the people)
- Example Displays: Lab Shifts, Change-from-Baseline, AE
  - Example Issues: Termination Visit Windowing, Unscheduled Visits, Multiple (linked?) Baselines and Endpoints, Competing Extremes, Population Definitions, Values Incompatible with Life, (Partially) Missing Date Comparisons

Lab Shift Table

Summary of Shift in Clinical Laboratory Evaluations
Population: Safety

			All Subjects at Visit 6 (N=XXX)		
			Low	Normal	High
ALT	Baseline	Low	XX	XX	XX
		Normal	XX	XX	XX
		High	XX	XX	XX
AST	Baseline	Low	XX	XX	XX
		Normal	XX	XX	XX
		High	XX	XX	XX
...					

Lab Shift Table

- Easy Case:
  - Patient 100 ALT:
    - Baseline: Normal
    - Visit 2: (Labs Not Scheduled)
    - Visit 3: Normal
    - Visit 4: (Labs Not Scheduled)
    - Visit 5: High
    - Visit 6: High
  - Normal --> High
- Specs and programming aimed at such cases.
- “All goes well…”

Lab Shift Table

- "What-If": Early Termination Visit
- Patient 101 ALT:
  - Baseline: Normal
  - Visit 2: (Labs Not Scheduled)
  - Visit 3: Normal
  - Visit 4: (Labs Not Scheduled)
  - Visit 5: High, Dropout
  - Visit 6: Lost to follow-up
- How to count?
  - Missing, Normal → High, something else?
- Solutions:
  - Visit-Specific and Endpoint Lab Shifts (next page)
  - Cumulative Lab Shifts (two pages later)

Lab Shift Table

- "What-If": Termination Visit Windowing
  - Visit-Specific and Endpoint Lab Shifts
  - Patient 102 ALT:
    - Baseline: Normal
    - Visit 2: (Labs Not Scheduled)
    - Visit 3: Normal
    - Visit 4: (Labs Not Scheduled), High, dropout
    - Visit 5: Lost to follow-up
    - Visit 6: Lost to follow-up
- Does unplanned visit 4 measurement belong with Visit 3, 4, or 5?

Lab Shift Table

- "What-If": Cumulative Lab Shifts
  - Patient 598 ALT:
    - Baseline: Normal
    - Visit 2: (Labs Not Scheduled)
    - Visit 3: Low
    - Visit 4: (Labs Not Scheduled)
    - Visit 5: Normal
    - Visit 6: High
  - Does the "Low" or the "High" dominate?

Lab Shift Table

- "What-If": Cumulative Lab Shifts, Unscheduled Visits
  - Patient 432 ALT:
    - Baseline: Normal
    - Visit 2: (Labs Not Scheduled)
    - Visit 3: Normal
    - Visit 3.1 High
    - Visit 4: (Labs Not Scheduled)
    - Visit 5: Normal
    - Visit 6: Normal
  - Does the unscheduled visit between 3 and 4 count?

Lab Shift Table

- "What-If": Multiple Baselines
  - Patient 722 ALT:
    - Baseline1: High
    - Baseline2: Normal
    - Visit 2: (Labs Not Scheduled)
    - Visit 3: Normal
    - Visit 4: (Labs Not Scheduled)
    - Visit 5: Normal
    - Visit 6: Normal
  - Which Baseline should be used?
    - What about the other tests? Should the lab panels be linked? Would linking apply to Endpoint, too?

Lab Shift Table

- "What-If": Site differences in Labs in a Multi-Center study
  - Units
  - Normal Ranges
  - Clinically Significant ranges
  - To what extent can these be integrated?
- Who? When? Documented?

Change-from-Baseline Table

Summary of Sitting Systolic Blood Pressure (mmHg) by Study Visit
Population: Safety

		Visit 1 (N=XX)	Visit 2 (N=XX)	Visit 4 (N=XX)	Visit 6 (N=XX)
SiSBP	N	XX	XX	XX	XX
	Mean (SD)	XXX.X (XX.X)	XXX.X (XX.X)	XXX.X (XX.X)	XXX.X (XX.X)
	Median	XXX.X	XXX.X	XXX.X	XXX.X
	(Min, Max)	(XXX, XXX)	(XXX, XXX)	(XXX, XXX)	(XXX, XXX)
Change in SiSBP	N		XX	XX	XX
	Mean (SD)		XXX.X (XX.X)	XXX.X (XX.X)	XXX.X (XX.X)
	Median		XXX.X	XXX.X	XXX.X
	(Min, Max)		(XXX, XXX)	(XXX, XXX)	(XXX, XXX)

Change-from-Baseline Table

- All the same issues as the lab shift for the change-from-baseline table
  - e.g., Termination visit windowing, multiple baselines and linking, unscheduled visits, missing values
- What-if: Values Incompatible with Life
  - SiSBP values of 0, 20, 60, 80?
    - At visit 6?
    - At baseline?

Change-from-Baseline Table

- Population Definitions
  - A common definition: The “Safety Population” will consist of all subjects who took any study medication.
    - If in doubt, assume “Yes”, or assume “No”?
  - An implementation:
    - Safety = “Yes” if FirstDoseDate > ScreeningDate
    - Else Safety = “No”
  - How is FirstDoseDate defined?
    - Look at first visit or look at all visits?
    - Partially missing values for FirstDoseDate or ScreeningDate?
    - Implications for Compliance calculations? Time-on-treatment calculations? Time-to-event results?
    - Other, cascading implications?

Adverse Events Table

Summary of Adverse Events
Population: Safety

	Placebo (N=XX)		High Dose (N=XX)		All Subjects (N=XXX)	
	N	(%)	N	(%)	N	(%)
Number of Subjects With Adverse Events	XX	(XX)	XX	(XX)	XX	(XX)
Number of Subjects With Severe Adverse Events	XX	(XX)	XX	(XX)	XX	(XX)
Number of Subjects With Treatment-Related Adverse Events	XX	(XX)	XX	(XX)	XX	(XX)
Number of Subjects With Adverse Events at least possibly related to Study Drug	XX	(XX)	XX	(XX)	XX	(XX)
Total Number of Reports of Adverse Events	XX		XX		XX	
Total Number of Reports of Serious Adverse Events	XX		XX		XX	
Number of Subjects Discontinued Due to Adverse Events	XX	(XX)	XX	(XX)	XX	(XX)

Adverse Events Table

- Classification of medical conditions:
  - Pre-existing: before study treatment begins
  - Concurrent: starts before and continues into treatment period
  - Adverse Event (AE): starts after study treatment begins
- Missing Data ==> Missing Flag?
  - (Partially) Missing Start Date
  - (Partially) Missing End Date
  - (Partially) Missing Start-of-Treatment Date
  - (Partially) Missing End-of-Treatment Date
- Who? When? Documented?

CAD Issues

- Data Integration from Multiple Studies
  - Different purposes (e.g., phase I versus phase III)
  - Different units (e.g., labs)
  - Different outcomes (e.g., NIH vs. Glasgow stroke scales)
  - Different timing (e.g., weekly versus monthly)
  - Different tests
  - Different criteria
  - Different cultures (e.g., US vs EU vs. Japan)
    - e.g., AE reporting, instructions for patient diaries
  - …
- Who? When? Documented?

90% Science, 10% Software

- Decisions: Who? When? Documented?
  - Ignoring the issues or making the wrong decisions usually leads to the “wrong” answer.
  - Human judgment is valuable.
- After the scientific solution is created, a computer program is used to implement it.
- Could we anticipate all these things?
  - Maybe, if we were a lot smarter and didn’t mind 1,000 page protocols.
  - Every trial is different.

Conclusion

- With Clinical Trials data, use a One-PROC-Away Analysis Database instead of Views when…
  - The right (scientifically valid) answer is important
  - Understanding the answer is important
  - Getting the same answer every time is important
- And…
  - Disk space is cheap
  - The data are not expected to change

Conclusion

- There are many steps in the creation of analysis databases – SADs, CADs, Product 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

One PROC away!

Conclusion

- The Powerpoint 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!