Analysis Databases: An Academic Perspective from the "Real World"

David H. Christiansen, DrPH
Christiansen Consulting
dave_christiansen@hotmail.com

"Statistics: From Theory to Regulatory Acceptance"
2003 FDA/Industry Statistics Workshop

Sept 18-19, 2003
Bethesda, Maryland

"Statistics: From Theory to Regulatory Acceptance"

- Caveats and disclaimers
- The scientific process
- Science, statistics and experimental design
- Clinical trials and regulatory submissions
- Analysis datasets and databases
- Design principles
- Summary and excuses

Caveats

- I am not currently an academic
  - I am a recovering biostatistician and professor
- I am not currently employed by a pharma or biotech firm
  - I have worked on NIH and industry clinical trials and regulatory submissions
- Part of this talk relates to a review of submission analysis datasets being conducted for the FDA
  - The opinions expressed are my own and do not represent the FDA

The Scientific Process

- State a question as a testable proposition
- Define key terms precisely
- Hypothesize answers or solutions
- Gather empirical data under controlled conditions
- Analyze and synthesize data
- Test hypothesis using appropriate statistical procedures – accept or reject

Woolever and Scott,1988

From Science to Experiments

"Science is concerned with understanding variability in nature, statistics is concerned with making decisions about nature in the presence of variability, and an experimental design is concerned in reducing and controlling variability in ways which make statistical theory applicable to decisions about nature."

Winer

From Experiments to Clinical Trials

- An experimental design conducted on human subjects is a clinical trial, or
- A prospective study comparing the effect and value of intervention(s) against a control in human subjects – Friedman,Furberg, DeMets, or
- Testing in which preventive, diagnostic, or therapeutic agents are given to a human population under controlled conditions to determine the agents' safety and effectiveness – The Piping News Report Website

From Clinical Trials to Regulatory Acceptance

- **New Drug Application (NDA)**: A formal application to the FDA for approval to market a new drug product. When the investigational phase of a drug is completed, the manufacturer gathers together the results of all studies and submits them to the FDA in a New Drug Application. This application is reviewed in detail by a team of reviewers. The purpose of the NDA is to determine whether the drug meets the statutory standards for safety, effectiveness, labeling and manufacturing – The Piping News Report Website

Academic Freedom: Getting off the subject – Piping News

- The Piping News Report www.pipingnews.com
  - a point of reference for those individuals who's career revolves around the various industrial and commercial Design and Engineering fields.
  - FDA abbreviations - In an exuberant effort at being concise and efficient, various organizations (with the government at the top of the list) have created individualized pseudo languages. The confusing aspect is that there is no central Organization for the Establishment of Abbreviations and Acronyms (OEAA)
  - FDA Definitions and other definitions, e.g. Welding,
    - **peening - the mechanical working of metals using Impact blows.**

Regulatory Submissions

- The end result of a scientific process to answer a testable question
- Implemented as series of experiments using experimental designs to test statistical hypotheses
- Results are submitted to regulatory reviewers, who approval the treatment for a specified population based on the submitted data

Submission Data Sets Include:

- Case Report Tabulations (CRTs), formerly data listings for the
  - Safety Domain Data Sets referenced in guidelines:DEMO, CONMEDS, EXPOSURE, AE, DISPOSIT, LAB, ECG, VITAL, PE, MEDHIST
  - Domains related to efficacy measures, will vary by indication
- Analysis Data Sets
  - Not required in the Guidelines, but strongly suggested
  - Normally generated from the data values in the CRT (Domain) data sets

Uses for Submission Datasets (Domain and Analysis)

- Replicate or verify the sponsor's analyses, results, and conclusions
- Test the validity and robustness of the sponsor's analyses and assumptions (what if…)
- Audit the data for inconsistencies and errors

Primary Reviewer Tasks Involving Submission Datasets

- Statisticians
  - Replicate Analyses
  - Test assumptions
  - Perform alternative analyses
- Medical Reviewers
  - View data used for a specific table
  - View patient profiles
- Auditors
  - Compare source data values to CRFs or source documents
  - Verify derivations

Reviewers need sufficient data to answer these questions:

- Were the clinical trials sufficient in size and design to demonstrate the safety and efficacy of this compound?
  - Did the sponsor design the right trials?
- Were the designed trials conducted in accordance to the protocols?
  - Did the sponsor conduct the trials as designed?

Reviewers need sufficient data to answer these questions: (cont.)

- Did the trial data provide sufficient clinical evidence to conclude that this drug is safe and efficacious for the indication and the population proposed in the application?
  - Did the sponsor get the right results?
- Do the submitted data and documentation clearly describe the conduct and results of the trials?
  - Can the reviewer understand the data, results and conclusions?

Reviewers need sufficient data to answer these questions: (cont.)

- Is the clinical evidence of sufficient quality to ensure that the reported results are accurate and true?
  - Can the reviewer believe the data and analysis?

Clinical Trial Data Definitions

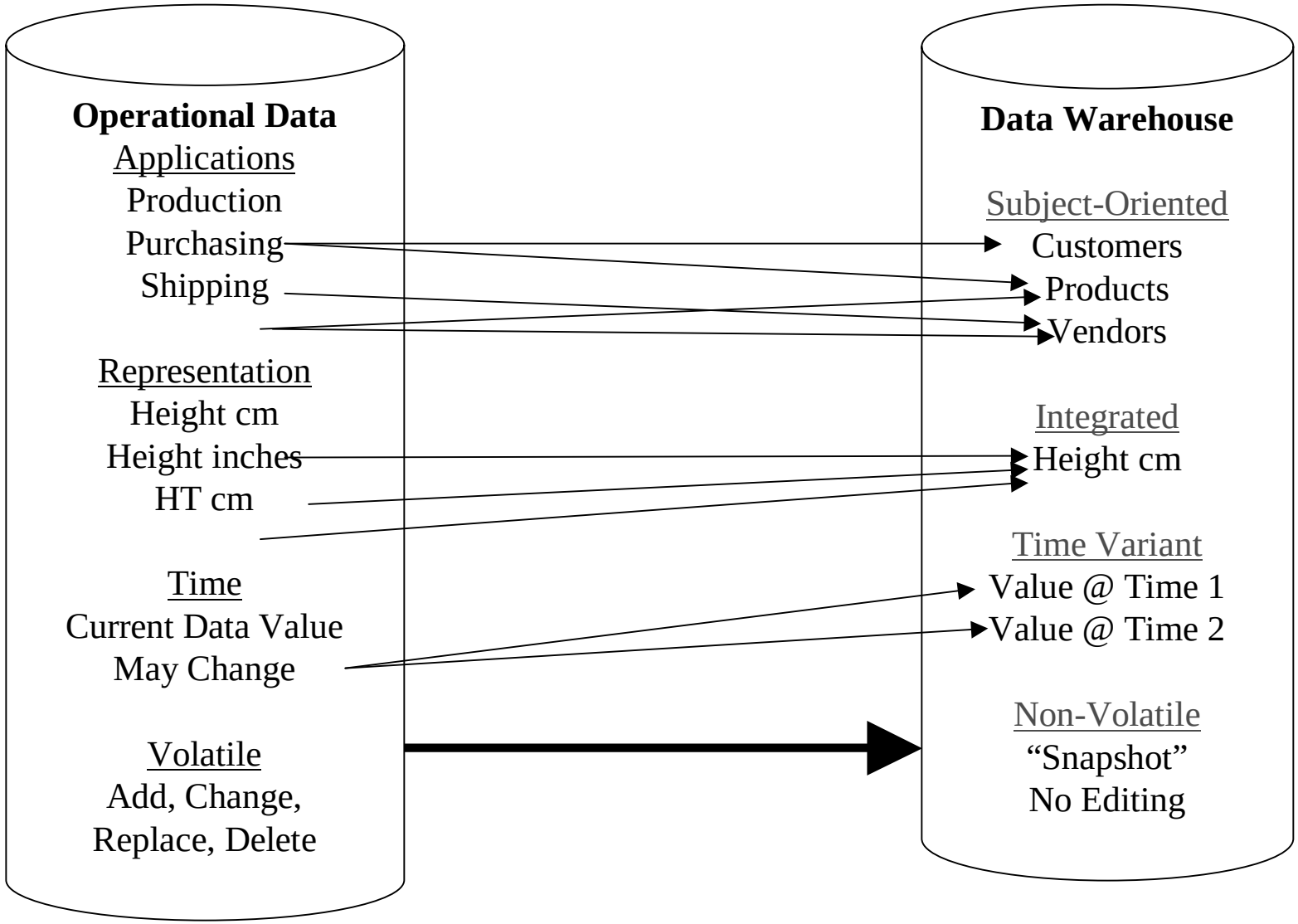
(*Adapted from: Meinert, CL, “Clinical Trials Dictionary”, Johns Hopkins Center for Clinical Trials, 1996.)

- **Source Data*** - Information collected and recorded about a subject (**raw data, operational data** or **primitive data**)
- **Derived or Computed Data*** - Transformation or reduction of one or more data items by a defined process or algorithm
- **Database*** - A collection of related data items, organized for ease and efficiency of use
- **Analysis Datasets** or **Analysis Files*** - A collection of source and derived data items, structured to facilitate data analysis
- **Analysis Database or Clinical Trial Data Warehouse** - Analysis files, metadata, documents and processes structured to facilitate the execution and reporting of clinical trials

“Commercial” Data Warehouse

- Classic Definition: “A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management’s decision- making process.” (Inmon, 95)
- The data are extracted from the operational database and transformed (re-structured and summarized) to populate the data warehouse.

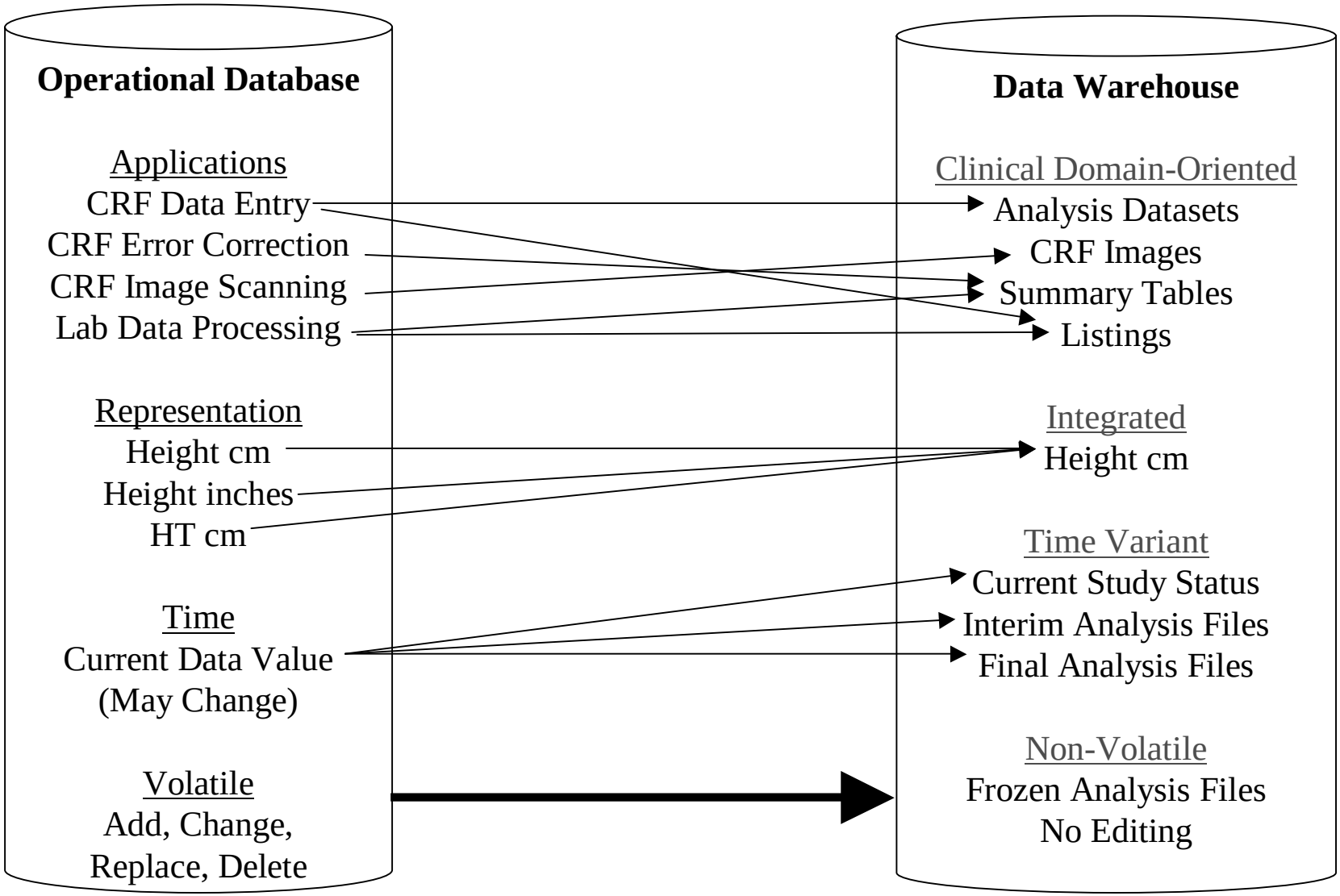
Operational vs Warehouse



Clinical Trial Data Warehouse

- A specialized data warehouse (DW), sometimes called a data mart
- Similar to a commercial DW, but some components may have different emphasis
- CAUTION: Commercial DW software may not be optimal for clinical trial DW

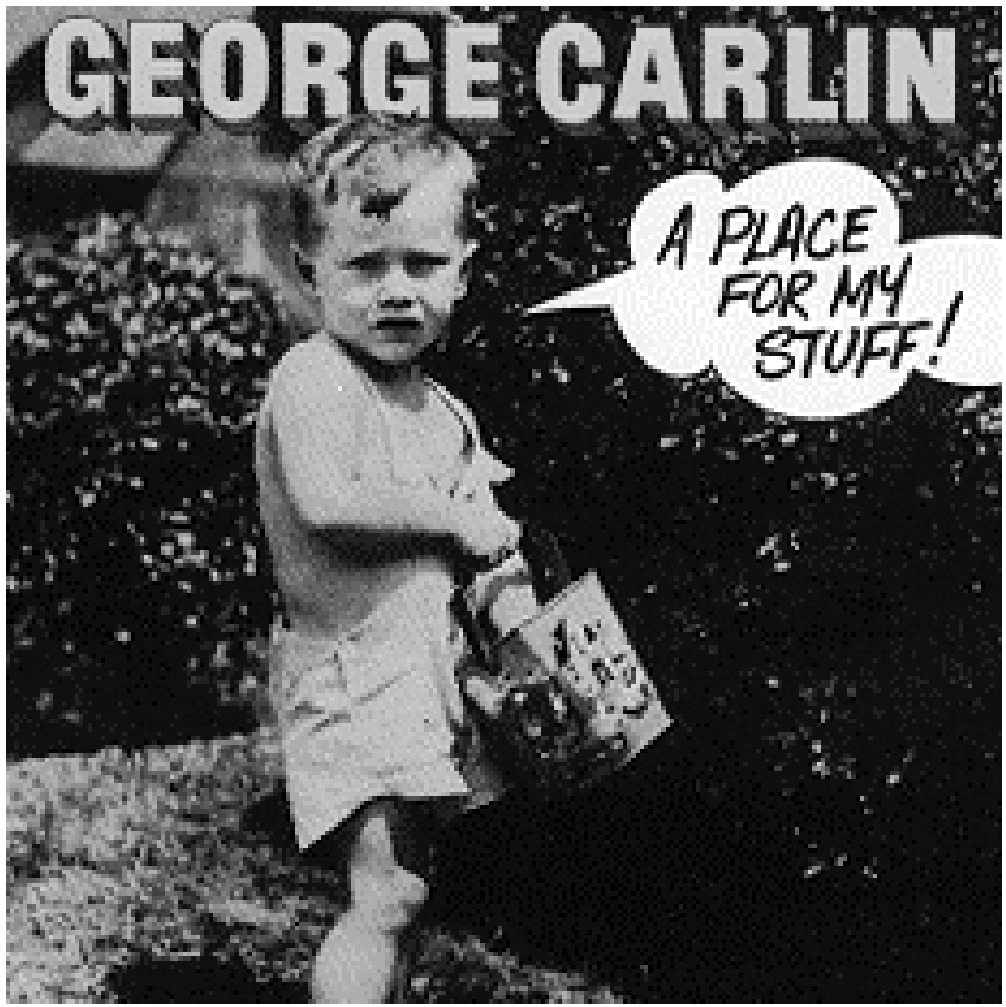
Clinical Trial Data Warehouse Components



“Commercial” Data Warehouse

- Classic Definition: “A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management’s decision-making process.” (Inmon, 95)

Alternative Definition- 1980



What Kind of “Stuff”: Contents a Clinical Trial DW

- Clinical Domain-oriented datasets
  - Source data items (CRF or raw data)
  - Derived data items (computed variables)
  - Summarized data items (endpoints, means…)
- Data Displays
  - Summary Tables
  - Listings (domain and patient profiles)
  - Results of Statistical Analyses
  - CRF and other images (X-rays, gels…)

Contents of a Clinical Trial DW (cont.)

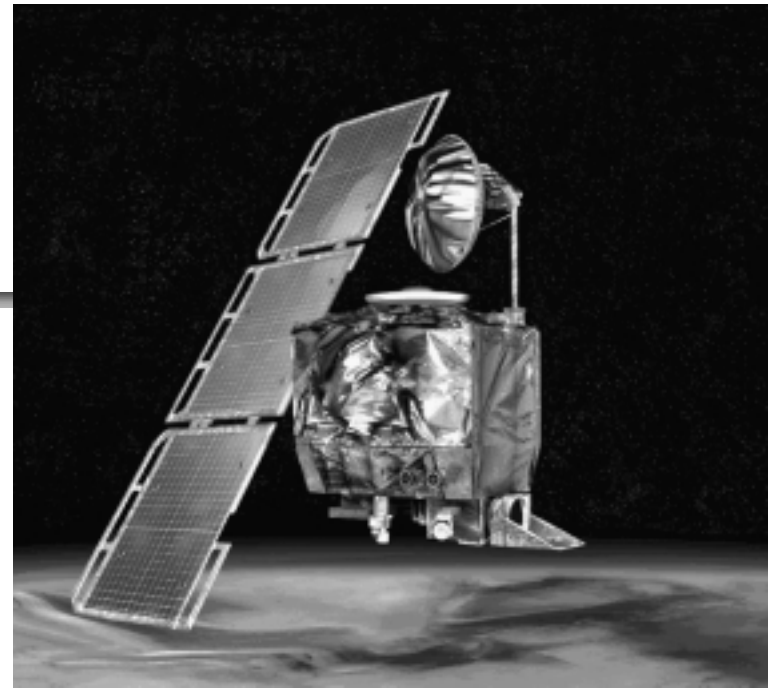
- Metadata
  - Description of datasets
  - Transformation programs
    - Reshape
    - Summarize
    - Compute new values
  - Analysis programs
  - Documentation
  - Hyperlinks among components

Another Academic Aside: How Important is Metadata?

Event	Time	D	S	F
Begin	9/23/99 02:01:00	121,900,000	12,300	143.878
End	9/23/99 02:17:23		9,840	

Event	Time	D	S	F
Start	19990923 05:01:00	196,200,000	5.5	640
Finish	19990923 05:17:23		4.4	

In this case $125,000,000: Mars Climate Orbiter



Mars Orbit Insertion Burn	**M/D/Y HH:MM:SS PDT (Earth Receive Time, 10 min. 49 sec. Delay)**	**Distance (miles)**	**Speed (miles/hr)**	**Force (Pounds)**
Begin	9/23/99 02:01:00	121,900,000	12,300	143.878
End	9/23/99 02:17:23		9,840	
Mars Orbit Insertion Burn	**YYYYMMDD EDT (Earth Receive Time, 10 min. 49 sec. Delay)**	**Distance (km)**	**Speed (km/sec)**	**Force (Newtons)**
Start	19990923 05:01:00	196,200,000	5.5	640
Finish	19990923 05:17:23		4.4	

Contents of a Clinical Trial DW (cont.)

- Study Status Data and Reports
  - Clinical center status
    - Projected vs actual active sites
  - Subject Status
    - Screening, enrollment, randomization, drop-outs
  - Data flow
    - Inventory of records, forms, samples, etc.
    - Forms received, processed, corrected
  - Data quality
    - Error rates, edit rates, center performance, etc.
    - Data quality report
  - Blinding of study is an issue

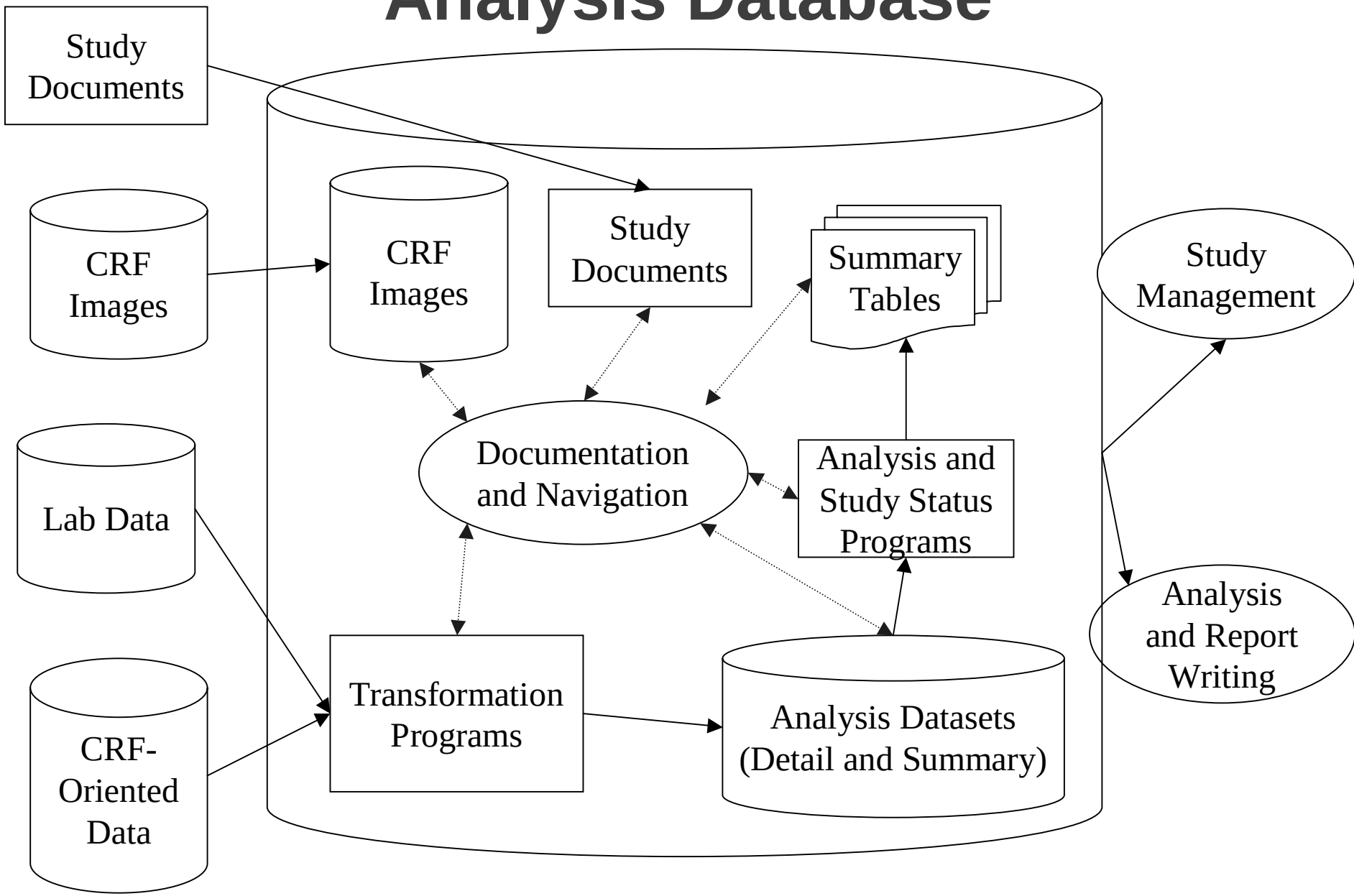
Data Warehouse Definition Problem!

- Study data and study status data is changing during the conduct of the study (it is volatile, which violates the classic definition of a DW)
- Solution is easy
  - Call it an Analysis Database instead

More Stuff: Why Not Store Study Documents in the Analysis Database?

- Protocol and Manuals of Operation
- Annotated CRFs
- Investigator CVs
- Statistical Analysis Plan
- Reports
  - Interim
  - Final
- References, Bibliographies

Analysis Database



Analysis Databases: Potential Scope of Use

- NIH research
  - Basic research
  - Clinical trials
  - The Biomedical Information Science and Technology Initiative (BISTI)
  - Genomics
- Academia
  - Basic Research
  - Clinical trials

Analysis Databases: Potential Scope of Use (cont)

- Drug Development Industry
  - Pharma and biotech developers
  - CROs
- FDA/Regulatory
  - Drug approval
  - Safety analysis initiative
  - ICH Electronic Common Technical Document (eCTD)

Analysis Database Uses and Activities

- Study management
- QC
- Data transformation
- Statistical computation
- Browsing, exploring
- Analysis
- QA/Auditing
- Documentation
- Regulatory review

Analysis Database Users and Customers

- Data Managers/CRAs/Study Team
  - Study management, QC, QA
  - Adaptive and interim analyses
- Programmers
  - Transformation
  - Statistical computation
- Statisticians (Researcher, Sponsors and Reviewers)
  - Browsing, exploring
  - Statistical computation
  - Analysis

Analysis Database Users and Customers (cont.)

- Clinicians (Researchers, Sponsors and Reviewers)
  - Browsing, exploring
  - Analysis
- Auditors (Sponsors and Reviewers)
  - Review documentation
  - Verify transformations and derivations
  - Validate source data

Designing an Analysis Database

- Why isn’t this a “No-Brainer”?
- Analysis database characteristics
- Underlying principles
- Outstanding design issues

Why isn't this a "No-Brainer"?

- The theory is simple, but the application is hard
- Analyses database serves many masters
- Different drug classes and indications require different analysis files
- No set of analysis files can anticipate all possible exploratory analyses
- This is not an Information Technology issue, it requires medical, statistical and regulatory expertise
  - Not hardware or software, but "wetware"

Characteristics of an Analysis Database

- Redundancy is OK
- Clear documentation is essential
- Datasets one PROC away from results
- Good navigation
- Should improve communication and cut review time
- Should answer more questions than it raises
- Subset of Analysis Database forms forms the basis for the regulatory submission

General Principles for Regulatory Submission of Datasets

- Clear communication of the content, source and quality of the data is the highest priority.
- Utilize standard data models and variables where they exist and are applicable.
- Remember that clinical trials are unique research experiments and it is highly unlikely that any standard will cover all data for all studies.

Clear communication of the content, source and quality of the data is the highest priority.

- When faced with competing alternatives, give higher weight to the one that provides the clearer communication of the content, source and quality of the data to a reviewer.
- Data quality is an outstanding issue

Utilize standard data models where they exist and are applicable

- The Clinical Data Interchange Standards Consortium (CDISC) has standards for
  - Submission Domain Standard Datasets (SDS)
  - Analysis Data Models (ADaM)
  - Operational Data Models (ODM)
  - Laboratory Data
- HL-7 and CDISC are collaborating on data and documents standards
- www.cdisc.org

Utilize standard data models where they exist and are applicable (cont)

- ICH and FDA have guidance on a Electronic Common Technical Document (eCTD)
  - ICH E3, E6 and E9 provide some models
  - XML allows navigation and “smart” datasets
- Do not “force” standards into situations that compromise the scientific integrity and clarity of the data.

Issues: Imputation

- CDISC is currently developing approaches for documenting imputation methods
  - Metadata can describe what method is used
  - How to identify specific data items is under discussion
- Partial dates
- ~~Last observation carry forward~~
- Statistical imputation methods

Issues: Identification of Analysis Populations

- Options
  - Status flags
  - Separate variables
  - Separate records
  - Separate datasets
- Choice depends on:
  - Statistical analysis
  - Study design
  - Dataset structure
  - Priority of ease-of-use vs ease-to-create

Issues: Submission of SAS Programs

- Purpose?
  - Replicate analysis
  - Exploratory analysis
  - Auditing
- Which SAS programs?
  - Dataset creation programs
  - Analysis programs
- How will programs be used?
  - As documentation
  - As “code fragments”
  - Execute in FDA environment

Issues:
Submission of SAS Programs – cont.

- Sponsors/CRO work flows vary
- Proprietary programs
- Dataset size restrictions in Guidelines
- Standardized report programs are complicated
- Macros are difficult to transport and understand
- Need to start dialogue with FDA statisticians

Summary

- The Analysis Database concept can be applied to a broad range of scientific activities
- The scientific and economic need for a uniform approach is increasing
  - Data is becoming more expensive
  - Data is becoming more expansive, e.g., genetic data in clinical trials

Summary (cont)

- Use existing standards, where applicable
  - When standards don't exist – THINK
  - For regulatory submission – think like a reviewer.
- A consistent, clear model for analysis database will increase the clear communication of the science
  - Better communication means less "peening" (mechanical working of the data with a hammer)