Data Integration and Data Quality: Pharmaceutical Industry Case.

Sergiy Sirichenko, Vadim Tantsyura, Olive Yuan, Ph.D.
(Regeneron Pharmaceuticals, Inc., Tarrytown, NY)

Max Kanevsky (Pinnacle21, Plymouth Meeting, PA)

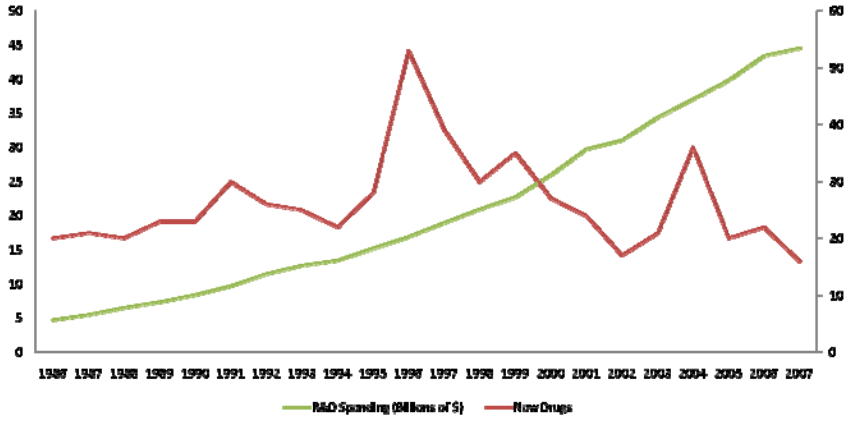
Agenda

- Current process
- Current definition of DQ in pharma
- Data integration issue and examples
- Recommendations

Introduction

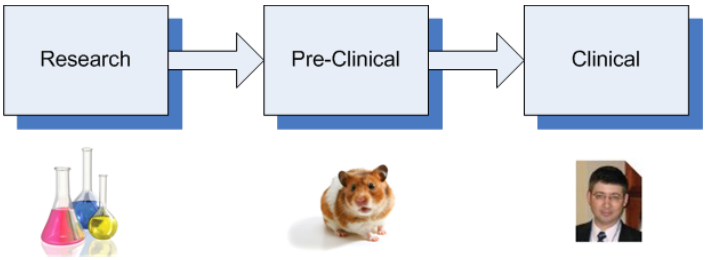
- "The review and approval of new pharmaceuticals by federal regulatory agencies is contingent upon a trust that the clinical trials data presented are of sufficient integrity to ensure confidence in the results and conclusions presented by the sponsor company." (Society for CDM, Charter of the Committee for Standards for GCDMP, 1998.)

U.S. Drug Industry Spending on Research and Development vs. New Drug Approvals by FDA (1986 – 2007)

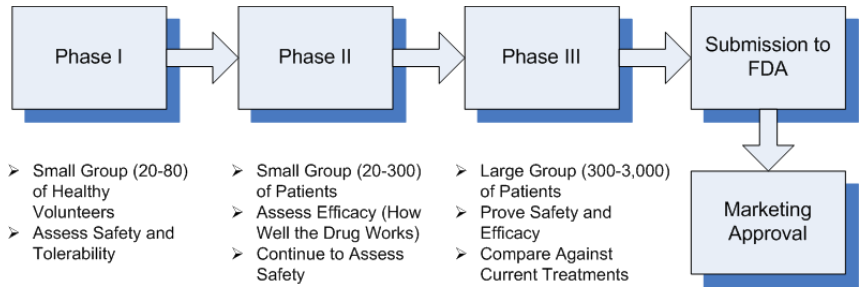


Source: Pharmaceutical Research and Manufacturers of America, *Pharmaceutical Industry Profile 2006,* http://www.phrma.org/files/2008 Profile.pdf and FDA Center for Drug Evaluation and Research, *CDER Drug and Biologic Approval Reports,* http://www.fda.gov/cder/rdmt/

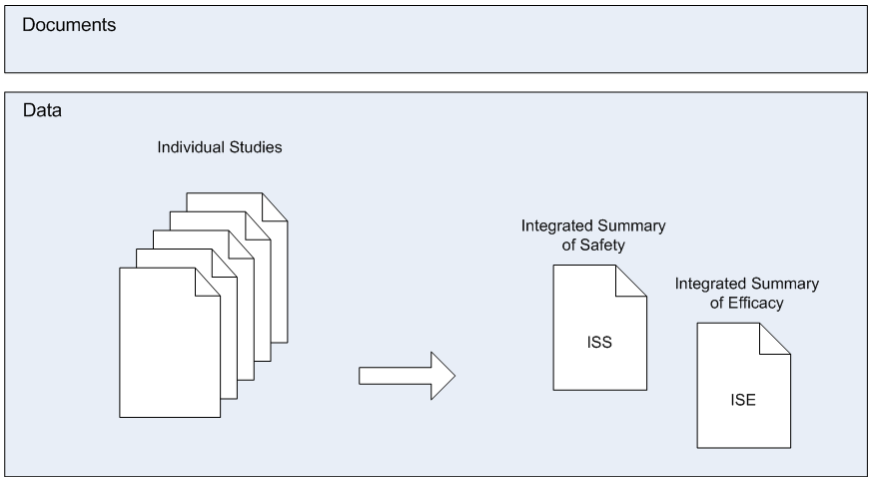
Drug Development Process



Clinical Trials



Submission to FDA



Definition of Quality

- SCDM adapted the IOM definition:
  - (sufficient) "quality data is data that support conclusions and interpretations equivalent to those derived from error-free data" (Institute of Medicine, Roundtable Report, 1999)
- **Risk-based approach**: Quote from Janet Woodcock (Science Board: FDA's New Bioresearch Monitoring Initiative, 04 November, 2005):
  - "High-quality Clinical Trial Data:
    - Support integrity of clinical research enterprise
    - Support confidence of public/patients in human studies
    - Provide evidentiary base for product approvals and medical practices"

FDA Guidance for pharmaceutical industry:

Computerized Systems Used in Clinical Trials (April 1999).

Data quality: attributes

- attributable
- original
- accurate
- contemporaneous
- legible

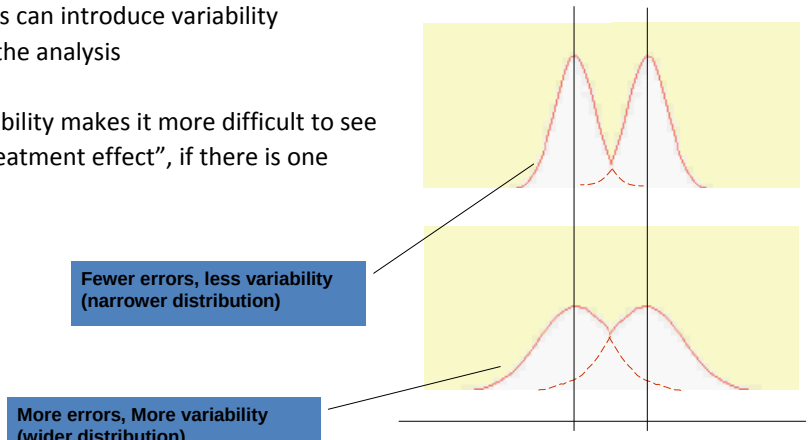
Common definitions of DATA error in CDM

- **Very often "conveniently" defined as "mismatch between database and the source".** (This definition is misleading.)
- **GCDMP definition** (v4, p. 77)
  - "A data error occurs when a data point inaccurately represents **a true value**… This definition is intentionally broad and includes errors with root causes of misunderstanding, mistakes, mismanagement, negligence and fraud."
- What is **true value**?
  - The foundation of traditional data manager is the belief that "source document" represents true value (in 99+% of cases source represents "true value")
  - Let's keep in mind this is not always the case (example: a value inconsistent with life)
- The following are NOT examples of DATA errors (according to the definition above):
  - Lack of compliance of procedures to regulations
  - Lack of compliance of practices to written documentation

Why are Errors Undesirable?

Errors can introduce variability into the analysis

Variability makes it more difficult to see a "treatment effect", if there is one



Meredith Nahm, MS, CCDM
Director, Clinical Data Integration, Duke Clinical Research Institute
Author, GCDMP Measuring Data Quality and Assuring Data Quality Sections

Assuring Clinical Trial Data Validity: The Current Process

- **The complexity of the design and the amount of data collected have important influences on data quality**
  - Design of protocol
  - CRFs
  - Data collection systems
- **Training is critical to ensuring that the protocol is followed correctly and the CRFs are properly completed**
  - Clinical investigator
  - Study personnel
- **Clinical site monitoring (can consume 15 to 30 percent of overall trial costs)**
- **Industry data QA procedures**
  - Assembly of all the data from trial
  - Entry of the information into databases
  - Evaluation of the data for quality
  - Audits of clinical sites

Assuring Clinical Trial Data Validity: The Current Process (cont'd)

- **FDA data analysis (includes clinical and statistical review)**
  - Checking and verification of data from important analyses submitted by the sponsor
  - Performance of exploratory analyses to answer questions that emerge from the review
- **FDA data QA evaluation**
  - Auditing of CRFs to verify the accuracy of tabulated data
  - Evaluation of follow-ups on reported AEs
  - Verification of primary outcome measure at the CRF level
  - An overall assessment of data quality is developed. If serious questions regarding overall data integrity are not resolved, FDA will not approve the application
- **FDA clinical study audit program**
  - A thorough on-site review of these sites is conducted by trained FDA inspectors. Record keeping, adherence to the protocol, informed-consent procedures, and other aspects of the study are assessed. If objectionable conditions are found, a report (FDA Form 483) is provided tot the PI at the conclusion of the audit.
- **FDA enforcement activities**
  - If an investigator found to have serious or repeated problems in performing clinical studies, FDA will take steps to debar the individual from performing trials for regulatory purposes. In cases of fraud, criminal prosecution may be pursued.

Main Thesis:

- Individual study data can be perfect (e.g. 100% clean), but when you integrate them the quality suffers

Data Integration Issue

- FDA collects all data from all studies, but their actual usage is limited to individual studies
- To perform meta analysis data integration is needed. This is a very time and resource consuming process that makes the collected data much less useful.
  - Example: FDA blood pressure drugs analysis
  - Analogy: Cryptography

Why clinical data are different?

- No required data standards in Pharma
- Usually each study is designed separately and physician rather than statistician is its owner
- Different data collection systems

Data

- Structure
- Terminology
- Content

Standards: CDISC, HL7, ISO, etc.

Structure

- Different structure is
  - a potential source of errors during integration and
  - obstacle for integration itself
- In addition, lack of standards leads to errors during data collection

Examples:
- Gender: M=1, F=2 vs. F=1, M=2
- Date formats: 07-08-12, 08-12-07, 12-07-08

Structure: (cont'd)

- Adverse Events:
  - Headache volunteered
  - Headache elicited from checklist (recall bias)
- Smoking
  - Dichotomy
    - Smoker
    - Non-smoker
  - Quantification or Qualification:
    - Smokes less than 1 pack per year
    - Quit smoking within past year
    - Smokes less than 1 pack per week, 2-4 pack, more…
    - Smokes cigarettes, cigars, pipes;

Terminology

- Definition- the same term is understand in the same way by different people in different places at different times

Examples:

- Myocardial infarction
- Age in China
- Age in raw data and SAE reporting
- Same name for different lab analytes: Lymphocyte count and percentages
- Same name for bilirubin in hematology and in urinalysis

Contents: Subject Race example

Study A	Study B	Integrated
White	*White*	*White*
Black	*Black*	*Black*
Other	*Asian*	*Other*
	Other	

Any coding leads to loosing of information

Contents: Subject Ethnicity example

Old way	Current way	
Race:	Race: Ethinicity:	
White	*White*	*Hispanic*
Black	*Black*	*Non-hispanic*
Hispanic	……	

Any coding leads to loosing of information

Content: AE Causality example

Study A	Study B	
Related	*Definitely*	
Not Related	*Definitely Not*	
	Possibly	*?*
	Probably	*?*
	Unlikely	*?*

Depends on your choice the analysis results can be different

Content: Study Day example

Study A:	*Day 1*		*Day 28*
Study B:	*Day 1*	*Day 21*	*Day 35*

Integration: *D21+D28* or *D35+D28* ?

Depends on your choice the analysis results can be different

Do not forget during analysis they are still apples and oranges!

Data can be perfect, but source populations are different

Phase 1: Healthy volunteers
Phase 2: Ideal patients and different diseases
Phase 3: Real patients with targeted disease

10 apples and 90 oranges → They are fruits → More than 90% of fruits are orange in color

Sometimes less data means better quality

Conclusion

- Issue:
  - Data integration has bearing on data quality that is under-researched in pharmaceutical industry now.
- Recommendations:
  - Use existing standards. If none, develop your own ones
  - Design individual studies keeping integrating data base as final goal in mind