On Designing a Web-Based Clinical Research Data Management System

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 800717 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat/dm

HEALTH EVALUATION SCIENCES RESEARCH CONFERENCE
UNIVERSITY OF VIRGINIA SCHOOL OF MEDICINE
26 SEPTEMBER 2001 UPDATED JANUARY 17, 2002

Thanks to Tony Rossini (U. Washington) and Aaron Mackey, Ken Scully, Rob Pates, Lori Elder, Bruce Carveth (UVa)

Outline

1. Type of needs
2. Advantages of Web as the interface
3. Components of a comprehensive web-based clinical trial data management system
4. External standards and requirements
5. Commercial vs. open source + local development
6. Components we should begin implementing
7. Proposed design

Needs

- Basic research
- Single institution observational clinical research
- Multi-institution obs. research
- UVa randomized trials and GCRC-type clinical observational research
- Multi-institution clinical trials
- Coordinating center for multi-center NIH clinical trials and epidemiologic/observational studies

Statisticians find

> unacceptably high error rates, usually from data values that would not have passed simple validation rules, or missing values in critically important fields. In such cases, statisticians quickly tire of rerunning statistical analyses that just reveal bad data rather than producing credible statistics. They tend to lose confidence in the quality of the data and spend more time examining data for errors before performing statistical analyses, that is, they are doing what data managers should have done earlier in the process.

—Ron Helms, *Drug Info J* 35:829;2001

Advantages of Web as the Interface

- Marks, Conlon, Ruberg *Stat in Med* 20:2683;2001
- U. Florida Division of Biostatistics has conducted the largest all Internet multi-center multi-country clinical trial
- INVEST–Phase IV hypertension study:800 sites, 9 countries, 22,000 patients so far
- No customized or proprietary software to install at site
- Local and worldwide interface to central database

Advantages, *cont.*

- Training conducted over Web with automatic database recording of completed training
- No voluminous paper stored a study site
- Study protocol always in one place and up-to-date
- Sites implement protocol changes enforced by system
- Each field on case report form linked to correct place in online protocol

Advantages, *cont.*

- Updated informed consent always printable from web
- Clinical data captured at health professional/subject interface
- Fields checked for
  - impossible values
  - missing values on mandatory items
  - values inconsistent with rest of form
  - values inconsistent with other forms stored for pt.

Advantages, *cont.*

- Wrong but legal values can still get by, but sites strive to get data capture correct first time while subject still present
- Data queries and monitoring greatly reduced
- Pharma companies estimate after-the-fact data queries average $15/query
- Make auditing based on less reliable data sources impossible (Florida punch ballots vs. Albemarle County touch screen)
- Allow entry of other data directly by labs

Advantages, *cont.*

- 24$\times$7 randomization
- e-IRB
- Better security than paper lying around, and no data or software exists at sites
- Site and study progress monitoring in real time
- Electronic fund transfers can be programmed and protocol-driven
- Time from last patient to database closure lessened

Disadvantages

- Some site personnel used for data entry
- Some patients uncomfortable when the health professional uses a computer in their presence
- Sites need medium- to high-speed Internet access
- Slowdown during peak Internet usage
- Audit trail ignores any paper components
- Need to safeguard against “man in the middle” SSL/SSH attacks

Components of a Comprehensive System for Clinical Trials

- Site training, administration, monitoring
- Recruitment of investigators and subjects
- Smart case report forms
- Capture or upload of external lab data
- Randomization
- Manage research pharmacy

IRB Components

- Electronic approval of site before allowed to enroll pts.
- Maintain informed consent documents electronically
- Monitor study conduct by getting online reports of individual and cumulative AEs
- E-remove site

Components, *cont.*

- Data security
- Study progress monitoring, reporting
- AEs result in automatic E-mail to sponsor safety group/PI
- General database queries
- Statistical analysis

Components, *cont.*

- Time/date/user ID stamp for data entry and updates
- Reviewable audit trail
- Multi-lingual
- Automatic electronic fund transfers to investigators
- Double data entry and visual data verification used only when paper CRFs are used and are not mandatory[a]

[a] Fong *Drug Info J* 35:843;2001

External Standards

- Compliance with FDA 21 CFR Part 11, Electronic Records: Electronic Signatures
- Good Clinical Data Management Practice; Society for Clinical Data Management
- MedDRA[a] will be the world standard for coding AEs[b]

  extreme specificity; $3000/year
- Database structure standards for databases to be submitted to FDA: CDISC[c]

---

[a] Medical Dictionary for Regulatory Activities; `www.meddramsso.com`

[b] Tremmel, Scarpone *Drug Info J* 35:845;2001; *Data Basics* Vol. 7 No. 3, Fall 2001 (newsletter of the Society for Clinical Data Management)

[c] Clinical Data Interchange Standards Consortium; `www.cdisc.org`

Commercial vs. Open Source + Local Development

- Commercial clinical trial data systems (e.g., Oracle Clinical) are
  - massively expensive
  - some components validated
  - liability protection
- Commercial e-clinical trial services (expensive)

Open Source + Local Development

- No cost for software
- High quality
- Can share resulting system with partners or get good PR from giving to anyone
- Support from other users, not vendor
  - Internet help groups function very well
  - Likely that other users face same problem you encounter
- Local development does not have to be extremely expensive if modern tools are used
- Longer time to full functionality

System Validation

- CFR 21 Part 820.3: “confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled”[a]
- Not one-time process
- System changes must be vigorously tested
- Validation by vendors of commercial systems not sufficient; must validate particular instances of systems. But validating particular instances validates much of the rest of the system anyway.
- Test cases play a major role
- See Svindland, Regan *Drug Info J* 35:819;2001
- SOPs are essential to ensure consistency of data quality across studies, time, sites, data managers[b]

[a] `www.fda.gov/cdrh/comp/swareval.htm`

[b] Fong *Drug Info J* 35:843;2001

Components To Begin With

- Web data entry
- Error checking
- Tabular representation and other queries
- SQL database initialization and management
- Security, backup
- Audit trail
- Soon thereafter:
  - randomization module
  - parse metadata to compose data entry scripts for handheld devices
  - develop program to do batch import and error checking, particularly for data entered from handheld devices

Institution-wide Standards for Data Elements

- Need input from current data users
  - CDR, Cancer Center, MCC, etc.

- Use engines each excelling in one area
  - Linux operating system
  - APACHE web server
  - Netscape or IE for rendering, information exchange
  - Javascript for client-side interaction and field checks
  - PostgreSQL database engine
  - Perl for scripting, dynamic HTML generation, data checking and manipulation, database interface
  - Tangram for Perl database object mapping[a]
  - R for statistical analysis and graphics

[a] `http://www.soundobjectlogic.com/tangram/fs.html`

- Synergy with CDR, Bioinformatics—use same software tools
- Allow for very rich metadata specification, specified and stored using XML[a]
- Metadata runs the system
- Both system-wide and study-specific metadata for field-related objects
  - If 2 objects w/same name, use study-specific
  - Missing study-specific attributes for an object (e.g., range checks)—use system-wide attributes

[a] `http://www.acm.org/pubs/articles/journals/toit/2001-1-1/p110-yoshikawa/p110-yoshikawa.pdf` is a nice paper that presents methods for interfacing between XML data and SQL.

- Access privileges by user **function**
- Both hard and soft error checks
  - Hard: disallow saving record with current field values
  - Soft: warnings listed in a separate HTML frame

Weakness of Model

- Changes in data collection or structure (e.g., possible values of multiple choice fields but not changes in range checks) can be reflected easily by editing metadata
- But difficult to execute these changes in **existing** data in an automated way

Metadata for a Study

- Study short name
- Description
- URL base name for e-protocol
- URL (full or base) for other study documentation
  - used for documentation of fields or consistency checks
- Names of fields whose combinations uniquely identify subjects[a]
- List of names of CRFs comprising the study[b]
- User names and functions
- Name or parameters for randomization module, including name of field to hold treatment assignment in other tables

[a] This is usually a single field.

[b] Study metadata may name a CRF not defined for that study. In that case an entire CRF from the system area will be utilized.

Metadata for a Data Collection Module

- DCM name (e.g., PhysicalExam)[a]
- DCM description (used as heading for block of DCM data on form)
- Names of fields in DCM

[a] DCM is the term used by Oracle Clinical.

Metadata for a Subtable

- Subtables are like single choice fields except they are multivariate (e.g., date of onset of symptom and type of symptom when a patient can have a variable number of symptoms)
- Subtable name (e.g., CurrentMeds)
- Names of fields in subtable
- By convention the first name will be that of a field that will informally serve as an index for the subtable (e.g., date of symptom or date medication began)
- ID fields for subtables will automatically be defined as the unique ID of a CRF that names the sub-table

Metadata for a CRF

- Names of DCMs and field names not contained in DCMs, in order of presentation
- Names of subtables; subtables with variable numbers of rows shown will be inserted into the data collection form at the point defined by the order of DCM, field, and subtable names in the CRF metadata
- Skip rules
- Code for intra-table consistency check logic not anchored to a field
- Inter-table consistency checks for tables “above” CRF
- Names of derived fields not displayed to user

Metadata for a Field

- Name
- Label
- Units of measurement
- Type
  - integer, float, string, text, single choice, multiple choice, table look-up, etc.[a]
- Source
  - user entry (default)
  - computed (for user field is read only)
  - imported (read only)
- ChoiceList: name of choice list if choice field
- personal: used to mark that the field is a personal

[a] See the `details` document for table look-up. *string* refers to character strings having a maximum length of 255 bytes. *text* refers to a character string with no limits.

identifier[b]

- Len: max. field width if string[c]
- DisplayLen: max. display width (if not Len)
- MaxChoices: max. # choices allowed if multiple[d]
- Default value (fills field when form brought up)
- Hard range limits
- Soft range limits
- Intra-table consistency check logic for fields that are always used as a block (e.g., $\mathsf{sbp} \geq \mathsf{dbp}$)[e]
- Suffix of URL listed in study metadata, to be used

---

[b] Users provide a passphrase when they login; this is used to encrypt personal identifiers so that they are unreadable at the data management center.

[c] This may not be needed. At any rate, it is only needed if < 255.

[d] Omitted$\rightarrow$no limit

[e] All consistency check objects include a flag denoting whether the check is a hard or soft check.

for defining the field[f]

- Field prompt (default = Label)
- Optional:
  - field used in support of FDA application
  - field is a clinical endpoint[g]
  - condition of subject for measurement
  - acquisition method (lab download)
  - date of validation
  - how validated

[f] An absolute URL should also be allowed, when a definition exists outside the study protocol. A typical relative URL will be `#fieldname` where this tag is used inside the online protocol HTML document.

[g] In general this is not helpful, as many endpoint variables are also collected at baseline.

Metadata for ChoiceList

- ChoiceList: name of list
- Choices: comma-sep. array of choices
- Order: array of display orders (0=suppress)
- Code: array of codes to display to the left of Choices
  - These codes are not stored with the data but are used only for display and for optional concatenation in front of choice labels when constructing analysis files

Overview of Proposed Design

- Relational DB model with medium-client Web interface
- Metadata parser to generate
  - SQL database definition commands
  - JavaScript code for client-side field and within-table consistency checks
  - Dynamic HTML generator for data entry forms
  - Perl code to fetch data, repeat field and within-table cons. checks, do inter-table consistency checks, interface with SQL database
  - First draft of case report form if paper form needed
- Parser handles inheritence from system metadata when attributes unspecified in study

Project Organization and Funding

- Funding exists in DHES for developing a particular database for a two-site observational clinical study, an ideal test framework
- Initial development will probably involve writing Perl and SQL code to implement the database
- I.e., writing the code that will later be generated by XML parsers
- Future studies using a clinical data management core in DHES would need to fund the following personnel
  - 0.10-0.15 of a systems programmer if project is routine
  - 0.2 of a data manager (for small projects;

0.4-1.0 for larger projects)[a]

- funds for data entry at clinic or lab

- Database management, including writing metadata for data elements, will be done centrally at DHES; researchers will fund appropriate portions of salaries of DHES personnel for data management and system administration, in addition to parts of salaries of their own personnel, for data entry

[a]The data manager will serve these roles: At study start-up the data manager will implement the CRF and quality checks in the database system, and will compose the electronic protocol in HTML. Following this implementation, the data manager will do query generation/resolution by communicating with the sites, prepare administrative reports to monitor study progress and data quality/completeness, make improvements in quality checking specifications, and prepare analysis files for use by the statisticians.

Example of Advantages of Model

- $\hat{\Omega}$ Distributed computing project at AT&T Lucent[a] provides interface between S and XML
- Convert metadata in XML to S object
- Use this object for smart import of data into R

[a] `www.omegahat.org`

Example S Commands after Import

```
label(sbp)
'Systolic Blood Pressure'

units(sbp)
'mmHg'

mlevels(symptoms)
'headache'  'dyspepsia'  'leg cramps'
'diarrhea'
# levels of multiple choice var.

gi ← symptoms %in%
      c('dyspepsia','diarrhea')
# %in% operator: union of choices
head.leg ←
  symptoms==c('headache','leg cramps')
# == operator: intersection of choices
# checkedN(symptoms) for number selected

showSource(sbp)
```

`showSource` would pop up a window containing the page in the protocol where `sbp` is defined

On Designing a Web-Based Clinical Research Data Management System

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences

There is an increasing need in my department and in the School of Medicine in general for expanding our capabilities for collecting and managing data originating from clinical research projects conducted both inside and outside UVa. These projects involve observational patient-oriented research and randomized clinical trials, and to a lesser extent, basic biomedical research.

With the rise of the Internet has come the ability to enter research data remotely without installing or maintaining any software on the research personnel's computers, using standard Web browsers. As described in Marks *et al*. (*Stat in Med* 20:2683;2001) there are many advantages to conducting clinical research through the Web, chief among these being catching data errors during initial data capture. Simultaneously, the rapid availability of free high-quality high-efficiency open source database engines, Web servers, operating systems, and scripting languages such as Perl, Python, PHP, and Zope, and of the statistical computing and graphics language R, has given us an amazing number of tools without being subject to the whims or licensing fees of large profit-oriented companies such as Oracle and Microsoft.

This talk will describe the needed elements of a comprehensive Web-based clinical data management system such as those provided by Contract Research Organizations, and will overview a plan for implementing the

research data component of such a system. Issues such as new worldwide standards for coding diagnoses and adverse events and standards for database structure for studies to be reviewed by FDA will be mentioned. A method for developing and implementing study- or institution-wide standards for data elements will be described.

See `http://software.biostat.washington.edu/statsoft/snake/clintrial` for more thoughts on designing clinical database systems.