



Uniting Rare Diseases

***Advancing Rare Disease Research:
The Intersection of Patient Registries, Biospecimen Repositories and Clinical Data***

Session I

Standards, Informatics and Technology:

Data Standardization for Rare Disease Research

Clement J. McDonald, M.D.
Director, Lister Hill Center for Biomedical Communications
U.S. National Library of Medicine
National Institutes of Health, HHS
Bethesda, MD

1/11/2010





Preface

- In this 20 minutes I will be over simplifying
- One of the goals is to standardize how we ask questions and define variables independent of computer systems or computer standards
- It won't help to have every variable in the the format the computer wants if every body invents their own set of questions for similar content



Background (2 of 3)

- One registry company offers registries and focuses (e.g. – bariatric surgery) EMR – so a thin line separates registries and specialized EMRs
- Registry has a narrow focus so can be more easily be pre-configured to the variables of interest to the registry .--- IF you know for sure what is needed



Background (3 of 3)

- Longitudinal studies (often) have problems:
 - Every time they change one question in a questionnaire they create a new table for the whole questionnaire,
 - Questions and measured variables change over time in many and subtle ways,
 - And they have drop outs and lost to follow up
- Registries face the same problems—



Data Collection

- Registries like medical records are empty when you first buy them
- They are like an egg carton without eggs - they have storage slots but no content
- Remember- the computer does not go out and gather data for you

Data collection costs

- It costs to collect data
 - Pharmacists consume 9 minutes (average) for patient med history needed for medication reconciliation
 - Physician order entry takes from 30 seconds to 2 minutes per order on average across different studies.
- The more granular the coding, the longer the data entry menus, and the longer the data entry time
- Answering discrete questions with menus costs more time than saying what you know (narrative)
 - Recall the mail survey W \$2 bill.

Why standardize?

- First - To minimize the work of data collection .
- With standards you can pull some of what you want from existing sources such as :
 - Your office EMR or your hospital's IT systems
 - Laboratory(s) results/records
 - Pharmacy – from benefits managers
 - Birth records and newborn screening records
 - Medicare –Medicaid (provides intriguing follow up opportunities)



Why standardize? – Stability over time

➤ Second

- To increase the likelihood that the data you collect today will be compatible with the data you collect 5 years from now
- Registries have to operate over a long time frame to learn anything . And things change over time.

Why standardize? – pooling

➤ Third

- To facilitate pooling of your data with related registries or longitudinal studies.

What is the crux of standardizing patient information

- It not just about technical standards-
- There is a pre-technical phase
- And the work in this phase is the same detailed work as needed to build a data collection form or survey instrument

The pre-technical phase

- IN this step you “decide what to collect
 - Realize that at the beginning you don’t always know what to collect (Ontario rules) – there is trial and error
 - And also that you have to be realistic- about how much you collect esp in a registry
- The work is a more formal than a pure paper collection instrument You have to have a place for everything and every thing in its place
 - You can’t scribble in the margins on a computer form
- If you want to add text comments- make a place for it



Questions and answers

- The data collection form boils down to a set of inter related questions and answers
 - To build the form you have to determine the data type of each question –
 - If it is numeric , you also have to specify the units of measure and the absolute range
 - If it is coded, then you have to specify the answer list , explicitly
 - And so on

Something like this from the ACC

HISTORY AND RISK FACTORS	
<p>_____ cm Height²⁷ <input type="radio"/> Missing</p> <p>_____ kg Weight²⁸ <input type="radio"/> Missing</p> <p>Family Hx CAD²⁹ <input type="radio"/> Missing</p> <p><input type="radio"/> No</p> <p><input type="radio"/> Yes</p> <p>CHF³⁰ <input type="radio"/> Missing</p> <p><input type="radio"/> No</p> <p><input type="radio"/> Yes</p>	<p>Diabetes³¹ <input type="radio"/> Missing</p> <p><input type="radio"/> No Diabetes</p> <p>(Treatments)</p> <p><input type="radio"/> No Treatment - OR -</p> <p>(select multiple below)</p> <p><input type="radio"/> Diabetes – Insulin</p> <p><input type="radio"/> Diabetes – Oral</p> <p><input type="radio"/> Diabetes – Diet</p> <p>Renal Failure³² <input type="radio"/> Missing</p> <p><input type="radio"/> No</p> <p><input type="radio"/> Yes – Dialysis</p> <p><input type="radio"/> Yes – No Dialysis</p>

A Standardizing by using what exists

- The first step is to find *validated* survey instruments that will cover your content needs, They are the best possible data collection forms.
- Next best is a set of questions that have been used on a large scale (in studies or administrative environments –and cover the content)
- The 1st saves you the validation work
- The 1st and 2nd guarantee you have some one to share with.

WHERE TO LOOK FOR EXISTING SURVEY INSTRUMENTS

Data Standardization – Rare Disease Research

1/11/2010

Clem McDonald, Lister Hill Center, NLM



Some places to look for existing questions and answers

- The literature- (obviously) look for validated surveys
- NHANES – wide spectrum of survey instruments and questions honed over decades
- American College of cardiology American thoracic society for cardiac measures and disorders



More sources

- PROMIS –for variety of functional assessments
- PhenX- broad range of measures for GWAS studies
- Federal Assessment forms (MDS, OASIS, CARE – some parts are reusable -further they *may* become accessible for research purposes



NCBI's DbGap

- 100's of longitudinal studies-
- lists all of the questions and the possible answers
- Includes some very large studies (Framingham)
- Also provides data with appropriate approval
- Good source of question for special disorders

<http://www.ncbi.nlm.nih.gov/gap>



NCBI'S dbGaP -- Framingham

ses

cardiovascular disease (CVD) is the leading cause of death and disability in the United States. In 1990, Framingham Heart Study -- under the direction of the late Dr. William B. Kannel at the National Heart, Lung, and Blood Institute -- launched a novel and ambitious project in which the Framingham Heart Study was designed to learn about the general causes of CVD. The rates for CVD had been declining in the 20th century and had become

the common factors or risk factors following its development over the past 50 years. The Framingham Heart Study participants who had not yet experienced a heart attack or stroke.

Level data

[\(DUC\)](#)

[Access](#)

[Individual level data](#)

[Support Template](#)

Search Within This Study

Search for:

Go

Associated Substudies

 [Framingham SHARe](#)
 [Non-invasive Tests](#)

 [CT](#)
 [Ankle-arm Blood Pressure](#)
 [Bone Related](#)
 [Ultrasound](#)
 [ECG](#)
 [Hearing](#)
 [Pulmonary Function Test](#)
 [MRI](#)
 [Vascular](#)
 [Eye](#)
 [Sleep Study](#)

Welcome Trust UK Biobank follow 500K people for 30+years

- Will follow ½ million people for 30+ years
- Focus on the 8 commonest diseases
- They have done the work of combing through literature for variables to follow and survey instruments to collect them
- See Report of UK population Biomedical Collection Protocol Workshop held at the Royal College of Physicians April 17, 2001 (with Burroughs Wellcome support).



E.G collection instruments for DM

- Socio economic variable
- Alcohol use, smoking, exercise
- UK diabetes questionnaire
- Rose angina questionnaire
- Birth weight
- Infectious Hx
- BMI
- Vital signs, step test
- Lots of lab tests



Other Welcome- Choice of instruments

- for chronic Pulm disease
 - FEV1
 - MRC breathlessness questionnaire
 - Asthma questionnaire (more than one option)
- Mental health disorders
 - Mini Mental Status
 - “crystallized intelligence” e. g. NART
 - General health questionnaire (GHQ)
 - Depression – BDI or CES-D
 - Lots more



Other subjects that are well covered due to work in technical standards

- Laboratory tests
- Anthropomorphic measurements
- Medications



What we really need

- Catalogue of all of the variables that are used widely in one catalogue to make it easy to pick the ones you want.
- Don't have yet– but were working on it



There are some new things under the sun

- If you look hard – at least some of your data space will be covered by existing questionnaires
- But some will not
- So you will have to build your own
- We'll say a few words at the end about that



TECHNICAL STANDARDS



Message standards

- Technical standards exist for the data structures needed to carry completed data collection instruments (questionnaires)
- Within the US clinical settings HL7 version 2.x is king. Laboratory results, radiology reports, clinician dictation and many other kinds of results are delivered via HL7 in almost every hospital and large clinic. From the measured message flows in central Indiana- we estimated > 3 billion HL7 result message delivered in the US per year.



HL7- is the ISO shipping container for resu



Stacks of 20' & 40' Shipping Containers

An HL7 v2 observation message with a cargo of CBC results

Patient level

```
PID|||0999999^6^M10||TEST^PATIENT^||1992022  
5|F||B|4050 SW WAYWARD BLVD |
```

Order/report level t

```
-OBR|||H9759-0^REG_LAB|24358-4 ^Hemogram^LOINC
```

- Discrete Results

```
OBX|2|NM||789-8^RBC^LOINC||4.9|M/mm3| 4.0-5.4..
```

```
OBX|3|NM|718-7^HGB^LOINC||12.4|g/dL|12.0- 5.0|..
```

```
OBX|4|NM||20570-8^HCT^LOINC||50|%|35-49|H||F|
```

```
OBX|5|NM||30428-7^MCV^LOINC||81|fL|80-94|N||F|
```

HL7 sends results in a “table”

- One row per result
- A column for questions (The yellow one)
- A column for the answers (the orange one)
- I used an example with numerically valued answers because it fits on a single slide.
- But it can carry question/answer pairs for questions with multiple choice answers (coded), free text answers, even images as answers
- It can carry nested questionnaires



HL7 messages represent a stacked data structure

- Each value gets its own row.- with other attributes about where it was collected, when, units , normals
- This is the way EHRs , laboratory systems and pharmacy systems, are built . The same is true for any system that faces lots of possible questions
- It is also what you see in the CDSC result messages and V3 and CDA versions of HL7



Its not the same as the typical research DB

- Which will use a flat structure- with one record per visit
- And the questions defined by the column



Flat – structure

Pat ID	Name	surgery date	Hb	DBP	# of BPU	Bypass Minutes	Cholesterol
1234-5	Doe , Jan	12May95	13	95	3	80	180
9999-3	Jones , T	1Aug95	12.5	88	2	90	230
8888-3	Doe Sam	4June95	16	78	0	80	205

Stacked structure

Operational Data Base: One Record Per Observation

Pt ID	Relevant Date	Observation ID	Value	Units	Normal Rang	Place	Observer
Doe J	12-May-95	Hemoglobin	13	mg/dl	12.5-15	St Francis	Dr Smith
Doe J	12-May-95	Hemoglobin	11.5	mg/dl	12.5-15	St Francis	Dr Smith
Doe J	12-May-95	Dias BP	95	mm/Hg	80-140	St Francis	Dr Smith
Doe J	12-May-95	Dias BP	110	mm/Hg	80-140	St Francis	Dr Smith
Doe J	13-May-95	Bypass minutes	80	min		St Francis	Dr Sleepwell
Doe J	12-May-95	Diagnosis	CHF- 365 (ICD9)			St Francis	Dr Bloodbank

The point

- Don't jump to the flat structure for registries without some consideration of the alternative
- One data base using the stacked structure will accommodate many kinds of registries
- It will also accommodate change -like inserting cards in a deck
- It fit better with data from clinical systems
- However, the two are not incompatible with each other- as long as you are aware.



WHERE DO CODES (VOCABULARY) FIT IN



For clinical observations

- If you will accept a very over simplified version
- For results and diagnostic order message the key code systems are LOINC (the question – the orange) and SNOMED CT (the answer). SNOMED CT also serves important roles in specialized fields within many HL7 messages
- For drug orders and dispensing records Rx.Norm and Rx.terms (that identify drugs and ingredients are the key codes (but that is another topic.)



THREE NLM SUPPORTED VOCABULARIES



LOINC

- LOINC Codes recommended by US federal government and other countries (Canada, Germany, China ,etc) for laboratory results and other content
- Provides codes for lab tests, clinical measurements, diagnostic reports and survey instruments (as packages- with all the parts connected)
 - E.g. PHQ-9 , PHQ-2, OASIS, MDS, CARE, etc
 - Working on PhenX variables and PRMISE
- RELMA DISC + pig



SNOMED CT-

- Also recommended widely and internationally
- provides a unified approach for most clinical answers (organisms, anatomic parts, specimens, diagnoses and symptoms) . It does also provide codes for some observations
- More than 300K codes and hierarchical relations
- Has an elegant formalism



Rx.Norm

- To complete the NLM triple thread
 - US recommendation for drug ordering , medication profile, etc. Rx.Norm provides codes for drugs at the clinical drug and ingredient level. FDA provides related codes
 - Clinical level includes the strength and dosage form
 - E.g. Ampicillin 500mg oral capsules
 - Includes brand names and generic
- RX.Terms- a subset tailored to ease ordering (CMS)
 - AMIA 2008 Fall meeting Kin Wah Fung – paper
 - <http://wwwcf.nlm.nih.gov/umllicense/rxtermApp/rxTerm.cfm>



Web sites for code systems

➤ ICD -

http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

➤ LOINC - <http://loinc.org/>

➤ SNOMED CT

http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

➤ RX.NORM -

<http://www.nlm.nih.gov/research/umls/rxnorm/>

- Rx.Terms -

<https://www.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm>

Data Standardization – Rare Disease Research

Clem McDonald, Lister Hill Center, NLM

APPROACH WHEN TO BUILDING YOUR OWN QUESTION



Realize that you won't find everything in existing coding system

- You will be able to find codes for most standard clinical measurements and laboratory and other common tests- and their answers
- You will not find existing codes for everything – esp new kinds of survey instruments



Steps- when you have to roll your own

- Figure it out on paper
- Favor measures over descriptors
- Be careful about how you construct – measures
 - As multiple questions or one
- Pay close attention to existing measures that are close but not quite right
- Work with a NLM to facilitate a common structuring - and get us to a standard –computer based catalogue of variable



Favor measures/scales over descriptors

Hyperlipidemia as a lesson

- In the 1970's a national data collection group measured cholesterol and recorded the trait - hyperlipidemia (yes/no)
- They kept the descriptor and tossed the measure
- What was the harm? The definition had crisp sharp edge: "Cholesterol > 300"
- A few years later the normal became < 250 – and it changes didn't stop
- **ALWAYS KEEP THE MEASUREMENTS-**



Lots and lots of variation in question style

- How to make get different large studies and administrative groups to use the same exact question to for the same purpose
- Different answer lists for same dimension
- Lots of different styles-
- Not time to cover– will give one example



More Examples: Question/Answer Variation

Example from WHI

Education at screening?

- Didn't go to school
- Grade school (1-4 years)
- Grade school (5-8 years)
- Some high school (9-11 years)
- High school diploma /GED
- Vocational or training school
- Some college or Associate Degree
- College graduate or Baccalaureate Degree
- Some post-graduate or professional
- Master's Degree
- Doctoral Degree
- (PhD, M.D., J.D., etc.)

Example from Eye Study

1. What is the highest level of school you completed?
2. Grade 11 or less
3. High school graduate
4. Some college or associate degree
5. Bachelor's degree
6. Postgraduate work



Example Comparisons with WHI's Study

Example from Eye Study

What is your current Marital Status

1. Never married
2. Divorced/separated
3. Widowed
4. **Married**

Example from WHI

F20 Current marital status

What is your current marital status?

(Mark the one that best describes you)

- Never Married
- Divorced or Separated
- **Presently Married**
- Widowed
- **Marriage-like Relationship**



One question - different answer lists

- Could tolerate different answer lists for different contexts – if they came from one universe-e.g SNOMED CT

THE END





Center
Drive

Library
of
Medicine

Thank you!

Alternative Styles 2

➤ Dichotomous questions about many states

- COPD — yes, no (+,- many forms of negative)
- CHF — yes, no
- Stroke — yes, no

• Pick all that apply
– (answer all that apply)

- COPD [x]
- CHF []
- Stroke [x]
- None []



Pros and Cons 2

- The first provides more information (reputedly)
- The second is easier and faster for the users
- Looks like one question to the user – I prefer it
- Should settle this question with empirical comparisons of user time cost



Alternatives: Specialized questions versus companion questions

➤ Specialized

- Creatinine during last hospital stay = 3.1 mg/dl
- Creatinine post cath = 3.1 mg/dl

• Generalized with companion variable

- Creatinine = 3.1 mg/dl
- Associated event = Post cath result



Pros and Cons 3

- Case 1- Better for data collector and analyzer
 - Prescribes what is to be entered
 - Easier for analysis of the given study
- Case 2 - better for standardization
 - Isolates differences; keeps commonalities across data sets
 - Facilitates data pooling
 - Provides direct linking to existing clinical care variables
- Solution – Can have both
 - Name the question as needed, then transform as needed for the study the two question for communication and pooling

Style more

- But there can be more to measuring BP
 - Cuff size
 - Method
 - Alternative locations
 - Relation to exercise



Style still more

- So define as a panel – with many optional elements
- Diastolic BP
- Systolic BP
 - Cuff Size
 - BP method (auscultatory manual, auscultatory auto, oscillometry, etc)
 - BP vendor and model name (esp when delivered automatically)
 - BP Serial number (when delivered automatically)
 - Always want time stamped
 - Who took (maybe)
 - Where measured (maybe) – e.g. home/office/hospital

Two step vs. one step question

- From STS – Two step
 - Angina? yes /no If yes:
 - Angina type? – stable/unstable
- Versus - one step
 - Angina? None/stable/unstable
- From STS – one step
 - Radial artery used? No, radial/left/right/both
- When one-answer can be part of next question it saves a separate user response, and removes a source of differences between questions

What's it for

- What questions are being asked and what analysis is planned
- Makes a difference in what you collect and how much
- Realize the deep versus wide conundrum
 - If you collect hoards of variables – you need even larger hoards of patients for analyses
 - There are trade offs
 - Fewer variables on more patients is usually a better bet.