# Testimony to NCVHS
# Ad Hoc Workgroup for Secondary Uses of Health Data

**National Center for Health Statistics**

Hyattsville, MD

**July 17,2007**


**Clem McDonald, MD**

**Director, Lister Hill National Center**

**for Biomedical Communications**

**National Library of Medicine**

# The sources

# Many Sources for Electronic Data

- Lab data (almost always electronic
- Medication orders in patient and outpatient (ditto)
- Radiology reports (text)
- Pathology reports (text)
- Dictation (discharge summary)
- EKGs (tracings and data)
- Cardiac echoes
- Endoscopy
- More

# More Sources

- Administrative data– unfairly maligned
  - Coded diagnoses and treatments for ER and hospital encounters – identify the events
- Tumor registries – on order of 25 million records US wide over last 14 years
- Cardiology data bases (ACC, ATS , etc)
  - Millions of records documenting majority of catheterizations and by pass surgeries
- Federal ESRD base- Complete dialysis history
- Out patient medications- From pharmacy benefit managers.
- Pathology reports and paraffin blocks

# And Still More

- Medicaid –procedures and diagnoses and drug use
- Medicare- Diagnoses, procedures, and now, medication use
- Social Security death tapes-(define the big outcome)
- Lots of special federal collection instruments- OASIS – other nursing home data. disability, Medicare introductory exam, etc

# The Uses

# Secondary Uses

- **Public health**
  - Outbreak detection
  - Bioterrorism
  - Causal tracking

**Quality (and cost) control**

-- /performance improvement

-- Statistical feed back

-- Prospective performance assurance ( patient specific reminders)

**Personal health records**

- Personal health records

**Commercial uses**

- Feedback to physician
- *Marketing*
- *Market assessment*

# Research Uses

- - Epidemiology (in general)
- - Early discovery of drug toxicities( Viox )
- - Cost benefit and variation (Think Wennberg)
- - Value of new diagnostic and treatment technology
- - Recruitment of patients into studies
- - Longitudinal follow up
- - Large most simple clinical trials
  - Randomize and watch the Medicare encounters and Social Security death tapes

# The Issues

# We have a major shortage of evidence for Decision Making

- Clinicians are faced with zillions of decisions
- Research helps them with a smidgen of these
  - Preventive decisions – but even some of these (pneumonia vaccine) are soft
  - Some cardiovascular interventions
  - Some anticoagulation interventions
- Minimal help with special circumstances – age, co-morbidity
- Little help for decisions about diagnostic testing, surgery, use of devices
- Almost no help regarding cost benefits (Haynes)

# We worry about the small stuff

- Huge interest (and investment) in doing perfectly the 5% of care interventions which we know how to do.
- Not enough interest or investment in figuring out what to do about the 95% about which we know little or nothing.
- Many of these gaps could be filled in with the right population based data sets

# The prime directive

- Pull together the data that would let us take advantage of all of the data that now exists
- The same data – and the same effort to pull it together could be used for *Primary* as well as the secondary uses
- Could be used to asses the effect of the torrent of services we provide at so much cost to so many at so little (relative) benefit

# Have to overcome huge wall of entropy to pull the data together

- The economics of networks can provide the energy
- Regional reuse  for many purposes – clinical care – research, public health, quality etc, etc The RHIO model all providers can get to all- (Why put a machine in each office with all its costs and complexity)
- Many of those who do the standardizing work (over come entropy) get paid back (motivated) by the same effort done by others for them.

# Why is this hard

- Same reason that houses and desks become messy unless you invest work to organize it
- Entropy

# Major points of disorder

- Patient IDs across sources
  - Solutions- Patient linking algorithms
- Internal system data structures
  - Solution- Standard data structures in messagesLink observation/report IDs across sources
  - -- Encourage stacked data structures and master variable dictionary – for systems that don't

# The basic issues and problems do not vary with the use

# Exception 1

- When the use  requires data elements that are
  - Not collected at all at present
  - Or collected irregularly and not entered into anyone's computer
  - We don’t know how to collect some things
- Then someone has to absorb a new data collection cost
- Office testing systems are read like a thermometer- and written down on paper. Have to capture  dates, users, values and link results to a specific patient.

# Exception 2

-- When data is being sold

--Then the calculus about what is right becomes very complicated and unsure

# Flat versus stacked

- Think bingo cards vs playing cards
- The flat structure – defines its variables as column headers- and records limited set of results in  one “Bingo" card
- The stacked  structure—defines its variables as records in knowledge bases- and stores every result on a separate “cards” which can be shuffled and sorted different ways.

# Flat Data structure (Analytic Conceptualization)

| Pat ID | Name | surgery date | Hgb | DBP | # of BPU | Bypass Minute | Cholest |
|---|---|---|---|---|---|---|---|
| 1234-5 | Doe Jane | 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 |

# Think in terms of stacked not flat structures

- You can merge and sort the numbers in two decks of cards
- You can’t do that with bingo cards
- Encourage CMS to think stacked- so they can store all of their clinical data in one form (one file) . Horribly difficult to pool flat data

# Stacked Data Set Application Conceptualization

**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 | Cholesterol | 180 | | | St Francis | Dr Bloodbank |

# Linking patient identifiers across sources

- Different sources use “randomly” different patient identifiers
  - Not an issues in most countries- which have universal medical identifiers
- Solvable with linking strategies –
  - At least within restricted scopes of time & space
- Don’t make it worse-
- Action: Let researchers use one way hashes- (Vanderbilt)

# An Aside

- People shouldn't be so selfish
- Society pays everything for some people and some for everyone
  - ( tax free insurance)
- Their data should be made available for research
- Genetics alliance argues that people who do not allow their data *and* their DNA to be used for cure seeking research are selfish and short-sighted
- The risks of any big negatives are teeny compared to driving cars and snow skiing

# An Aside 2

- Issues and deciding what is right **much** tougher for some secondary uses
  - Marketing
  - Commercial use

# Deliver information in a standard data structure

- This problem is solved – for most of the clinical space
  - HL7
  - NCPDP
- Empirically - 98.5% to 99.5% of messages well formed and good.

Syntax is not the problem

  - The 0.5% to 1.5% bad are egregious violations of the standard.

# Good News About Tapping into these Sources

- Almost every clinical system marketed to hospitals or large group practices can pump out data and do so in a standard message format –Using HL7 –
- Enables a  Vulcan “Mind Meld” among clinical systems  (and other systems)

# Linking variable across sources

- Different sources invent “random” codes for the same observation (entropy).
- Need a universal code system attached to messages so that the same observations from different places can be linked. (Can do nothing without that)
- With such a universal code system numeric observations are ready for use.
- Then  a hemoglobin is a hemoglobin and a pulse is a pulse where ever it comes from--- to finish the “mind meld’

# Grappling with Observation (variable identifiers)

- LOINC  provides Good coverage of many spaces- producers need a push
  - eg 800-900 LOINC codes cover 99.3% of the lab message volume  (Vreeman & Fennel)
  - Action:  should require LOINC for at least these
  - Action: produce mapping guide for this too

# Grappling with Observation (variable identifiers) more

- Action: instrument and package vendors should deliver (or publish) the LOINC codes for their tests (2-10 measures per package insert

- The full problem is the same as
  - Standardizing all data collection forms

# Example HL7 Message with LOINC – CBC

**Patient level**

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

**Order/report**

**BR|||H9759-0^REG_LAB|20725^ ROUTINE CBC**

**Discrete results**

OBX|2|NM|63^RBCs^L~**789- 8^RBC^LN**||4.9|M/mm3| 4.0-5.4

OBX|3|NM|60^ Hb^L~**718-7^HGB^LN**||12.4|g/dL|12.0- 5.0||||F|

OBX|4|NM|61^Crit^L~**20570-8^HCT^LN**||50|%|35-49|H|||F|

OBX|5|NM|875^MCV^L~**30428-7^MCV^LN**||81|fL|80-94||||F|b

# Messages and Medical Records

- The  medical record is like a modular housing.
- The message defines the structure of the module.
- Have to have a place in the medical record for each item in the messages
- Won't need anything  else assuming everything must be sent or received in the message

# The Message Storage Structure: Action Items

- Stop looking for “gaps” that provide excuses for a new approach to what is already standardized- We are almost done !!!
  - Squish  CDSC and HL7 together where they overlap (much does )

# Focus on the senders of data not the receivers

- The receivers can’t make the messages more standard and easier to digest
- Only the senders can do that
- Incent the senders to deliver good messages
- Incent them to use standard codes for at least the common established concepts

# Q1 –regarding facilitating quality data gathering

- Make it easier to gather  needed data
- Encourage Data exchange
- Too much work required to capture hand held lab machine data in the office.
- Make it easier
  - Build in LOINC codes to output
  - Scan patient ID in
  - HL7 outbound – or
  - Print  2-D bar code for scanning results date,

# 1. Current privacy regs

- Administrators over interpret HIPAA
  - They say “HIPAA says no” when it doesn’t
- In general leave HIPAA alone-
  - It was hard fought. Provides good protection. Widely understood- now. Don’t start over

# Q2 Security questions- A

- Don’t need anything tighter for care or most secondary uses
  - However, I worry about selling data

# Q2. Security questions- B

- One tweak to consider
  - Allow some form of one way hash
- Would like to link patients AND use only de-identified data (Can be done under appropriate conditions under current rules
- But would actually be less risk to privacy-if each side could hash strong 1 way
- Especially impartial for up-dating
- Could be less need to move identified registry data around

# 3. Uses where protection might be low

- Where data is sold – things get complicated
- Assume it can only be de-identified
- Many opportunities for abuse with complex data sets
- Patient is not the only one with interest
- Sources become greedy
- Changes everyone's thinking
- Could risk patient's cooperation with high minded things.

# Special cases

- For research, we should make it easier
- Lose great opportunities to help each other – and our children by withholding our data (at least for de-identified data)
- Should not require consent for de-identified or limited set data
- All of health care is “paid” in part by society (Tax deductions to health insurance)
- Some say we are just being greedy buggers by forbidding its use

# Q4. Other issues

- RHIOs are more or less a sine quo non for most 2ndary uses.

# Q6-Collect for other uses

- Yes- from  Regenstrief experience
- Use for
  - Public health – state required case detection
  - Research-  De-identified research that does not compare sources, Identified with IRB approval
  - Magnificent performance project (Marc Overhage)  involving payers, hospitals and office practices-

# Use of data

- For re-search
  - De-identified- (text scrubbed) forbidden fields removed
  - Do not make public . Only qualified researchers  who pass IRB testing can access ( There are issues beyond patient privacy that matter)
  - Limited data sets
  - IRB – and contributor- approved identified research

# More

- Link for clinical use
- De-identify for research