Pitfalls and Realities of Working with Big Data

Reflections of a PBRN "Big Data" Newbie

Disclosure Statement

- I have no relevant financial relationships with the manufacturer of any commercial products and/or provider of commercial services discussed in this CME activity
- I do not intend to discuss an unapproved or investigative use of a commercial product or device in my presentation

Presentation Objectives

- Review:
  - What EHRs are
  - What data they provide
- Advantages for PBRN clinicians
- Discuss some challenges to the use of EHR data for research
- Suggest possible approaches to addressing these challenges

What is the Electronic Health Record (EHR)?

- In its simplest form, the EHR is analogous to the individual patient chart system
- EHR is first and foremost a tool for documenting the care of patients
  - Clinical use
  - Charge capture and billing
- EHRs have been built for data input, not for data extraction

What data are in the EHR?

- Narrative (text)
- Structured data
  - Anything with a numerical value
    - Heights, weights, BP
    - Lab results
  - Anything with a code
    - Medications
    - Diagnostic codes
    - Laboratory tests
  - Anything that's *properly* templated

Narrative data from a SOAP note

- Subjective: Johnny brought in by his mother today for a 3 year check. Things going well except for a recent eczema flare-up. In Little Shavers day care. His mother is pregnant and his "given up smoking." The family has moved into a 100 yr old farm house with peeling paint and... Johnny speaks in full sentences and is toilet trained.
- Objective: Alert cooperative 3 yr old boy ….
- Assessment: Healthy three year old with mild atopic dermatitis. Needs no immunizations. Should screen for Pb.
- Plan: Lead level ordered. RTC in one yr.

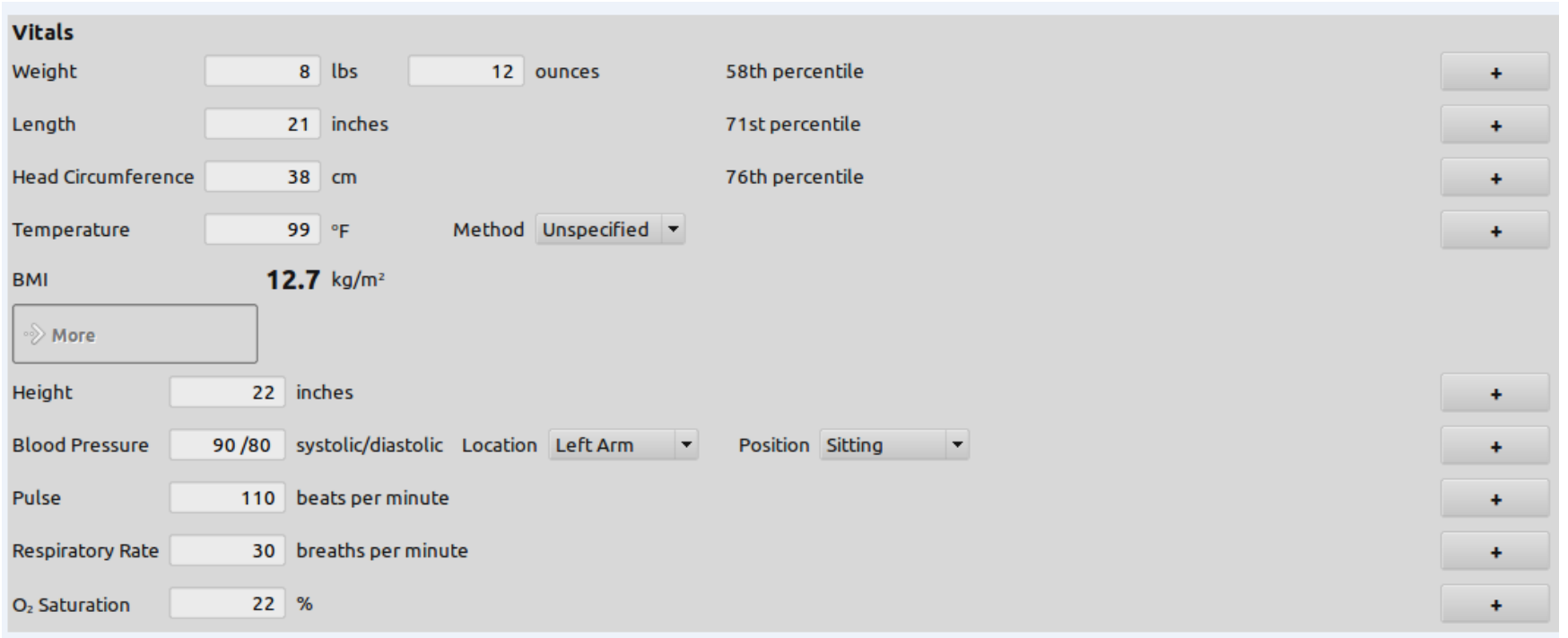
Data from narratives

- Very difficult to extract from the EHR
  - Inconsistently entered (i.e., often missing)
  - Not standardized (e.g., “year old,” “yr. old,” “y.o.”)
  - Challenging for machines to read/interpret (natural language processing – a work in progress)

Structured data

- Anything with a numerical value
  - Heights, weights, BP
  - Lab results
- Anything with a code
  - Medications
  - Diagnosis codes
  - Laboratory tests
- Anything that's *properly* templated

Structured data: vital signs template



Structured data:
a social history template

Social History:

Childcare:{childcare:20299}.

- Parent
- Relative
- Daycare ***
- In-home care ***
- ***

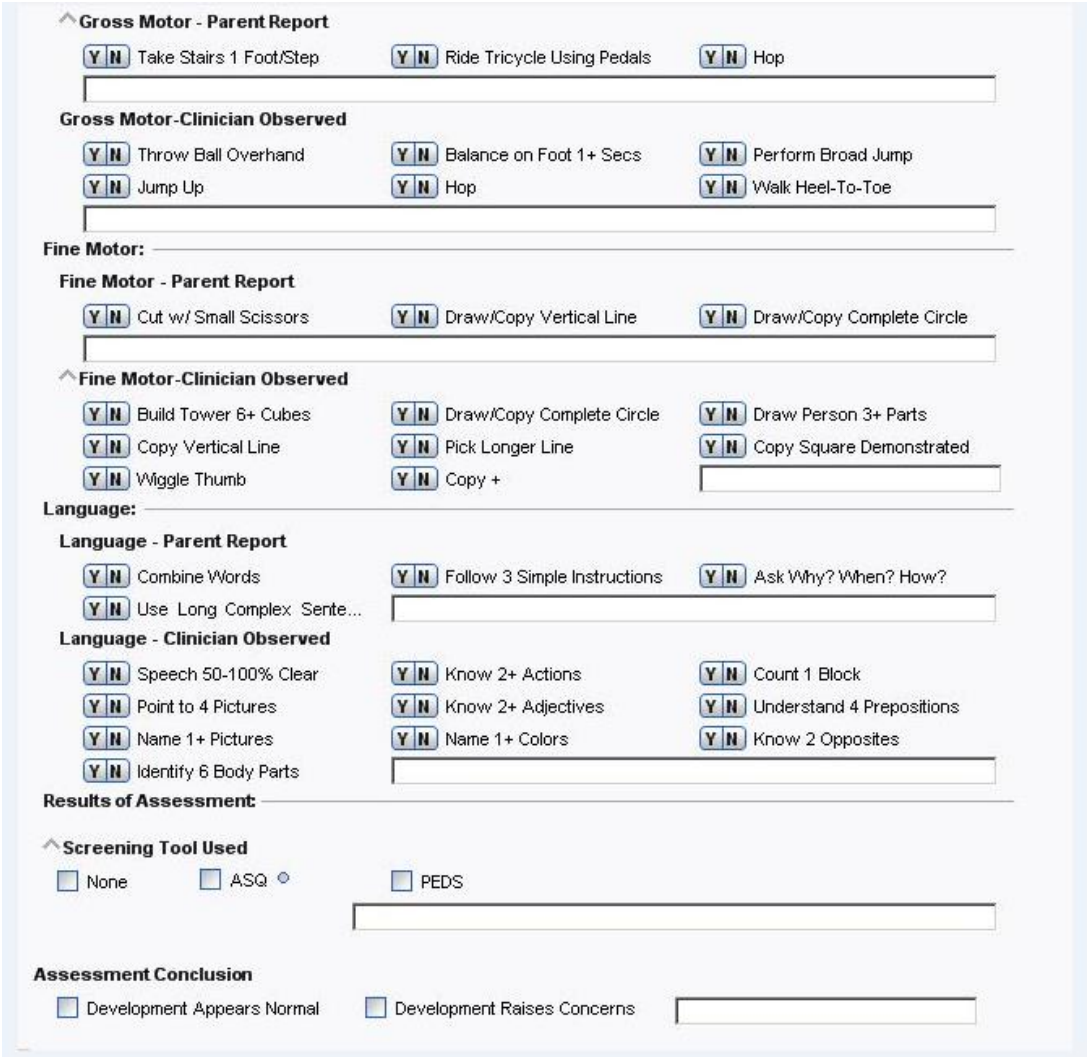
Secondhand smoke exposure? {second hand smoke:20467::"No"}.

- Yes, advised to smoke outside
- No
- ***

Lead risk: {lead:20086}.

- not done
- 0/5
- 1/5
- 2/5
- 3/5
- 4/5
- 5/5

Structured data: parental report developmental history template



Gross Motor - Parent Report

Y N Take Stairs 1 Foot/Step | Y N Ride Tricycle Using Pedals | Y N Hop

Gross Motor-Clinician Observed

Y N Throw Ball Overhand | Y N Balance on Foot 1+ Secs | Y N Perform Broad Jump

Y N Jump Up | Y N Hop | Y N Walk Heel-To-Toe

Fine Motor:

Fine Motor - Parent Report

Y N Cut w/ Small Scissors | Y N Draw/Copy Vertical Line | Y N Draw/Copy Complete Circle

Fine Motor-Clinician Observed

Y N Build Tower 6+ Cubes | Y N Draw/Copy Complete Circle | Y N Draw Person 3+ Parts

Y N Copy Vertical Line | Y N Pick Longer Line | Y N Copy Square Demonstrated

Y N Wiggle Thumb | Y N Copy +

Language:

Language - Parent Report

Y N Combine Words | Y N Follow 3 Simple Instructions | Y N Ask Why? When? How?

Y N Use Long Complex Sente...

Language - Clinician Observed

Y N Speech 50-100% Clear | Y N Know 2+ Actions | Y N Count 1 Block

Y N Point to 4 Pictures | Y N Know 2+ Adjectives | Y N Understand 4 Prepositions

Y N Name 1+ Pictures | Y N Name 1+ Colors | Y N Know 2 Opposites

Y N Identify 6 Body Parts

Results of Assessment:

Screening Tool Used

☐ None ☐ ASQ ☐ PEDS

Assessment Conclusion

☐ Development Appears Normal ☐ Development Raises Concerns

Plus and minus of structured data

- Plus: data are much more readily extracted
- Minus: typically, data take longer to enter
- Very highly structured EHR data entry can lead to work-around (e.g., switch to text) or clinician rebellion (e.g., omit)

What do EHR data offer to researchers?

- Data ≠ information
- Some “research quality” information can be found in the EHR (e.g. diagnostic code for “pneumococcal pneumonia” – ICD 481) means that clinician judged patient to have pneumococcal pneumonia
- Much information must be derived from some EHR data (e.g.,…)
  - Pediatric BMI can be derived from data on birth date, date of visit, gender, height, and weight
  - Pediatric hypertension derived from data on birth date, date of visit, blood pressure, gender, and height *on three occasions*

How accurate are EHR data?

- First we must ask, how accurate are data in *paper charts*?
  - Only moderately accurate, when compared to reports of standardized patients
    - Therapies prescribed: 68% accurate
    - Lab tests ordered: 64% accurate
    - History obtained: 29% accurate
    - Physical exam performed: 31% accurate

Are EHR data more or less accurate than paper chart data?

- For some things (*medications* and *laboratory tests* ordered), the EHR might be *more* accurate
- For other things (e.g., history and physical exams completed from templates with default settings), they might be *less* accurate

Example of infant physical exam template that could default to normal findings

Physical Exam:

Growth parameters are noted and {are, are not:2101676901::"are"} appropriate for age.

General: {general:20514::"Alert","Good tone/color","Appears stated age"}
Growth: {growth:20515::"Normal interval growth"}
Head/Neck: {head/neck:20940::"Normocephalic","Fontanel soft/flat","Neck supple"}
Eyes: {eyes:20955::"Red reflex present","No discharge"}
Ears: {ears:20518::"Canals clear","TMs clear","Light reflex present"}
Nose: {nose:20753::"Nares patent","No discharge"}
Mouth: {mouth:20942::"Palate intact","No thrush"}
Nodes: {nodes:20521::"No adenopathy or tenderness"}
Chest: {chest:20522::"BS Clear/ R=L","No retractions"}
CVS: {CVS:20956::"RRR","No murmur","Brachial=femoral pulses"}
Abdomen: {abdomen:20948::"Soft","Non-tender","No HSM/mass","No hernia"}
GU: {gu:20526::"Normal genitalia"}
MSK: {MSK:20954::"Full ROM","Normal hips"}
Skin: {skin:20950::"No rash","No diaper rash"}
Neuro: {neuro:20529::"Good tone","Motor skills intact"}

Accuracy of EHR data?

- **The fact that data are electronic does not confer upon them a higher epistemological status**
- EHR data depend on what the clinician chooses to document in the act of caring for the patient, and for billing – information accuracy can't be assumed
- Large scale retrospective data extractions are equivalent to "chart review on steroids"

Can we get population data *directly* from an EHR

- Not possible – real time needs of EHRs require different database architecture than is needed for research
- Data must first be extracted, transformed, and loaded (ETL) into a data warehouse, which allows for analytical processing
- Reminder: EHRs have been built for data input, not data extraction

The challenge of multiple EHR products and vendors

- ~100 products currently use in primary care pediatrics
- Products are not interoperable
  - Written in different programming languages
  - Different database architectures
- For research, data must be found, extracted, and standardized

What if we all used the same EHR?

- Data extraction would be easier, but still somewhat problematic
- Even if we all used the same EHR product, the systems are designed to be highly flexible and responsive to local needs, so…
- When you've seen one implementation of (Epic, Allscripts, eClinicalWorks, etc.), you've seen one

Other challenges that EHR data present vis-à-vis research

- Case identification
  - Who has asthma?
    - ICD codes (sensitive, but not specific)
    - Problem lists (specific, but not sensitive)
    - Pulmonary function tests (inconsistently applied)
    - Rx for bronchodilators (sensitive, but not specific)
    - Rx for inhaled steroids (specific, but not sensitive)
    - Combination of the above

Other challenges that EHR data present vis-à-vis research

- Defining an episode of acute illness
  - For example, to study treatment breakthrough in OM one must:
    - Find *new* OM encounters where patient was on antibiotics
    - Review prior and subsequent encounters
    - Review Rx orders, prescription start and end dates, other diagnoses
- Computer programming and data analysis required to do this are highly challenging

Irony re EHRs & research

- These challenges are actually not at all unique to EHR data; they exist for any studies based on medical record data
- Irony
  - It may be *as or more* challenging to employ informatics techniques to define cases of disease, episodes of illness, and denominator populations than to use human coders to perform the same tasks
  - In any event, the computer-based process ultimately needs to validated by a human decision-maker who reviews a sample of records by hand

Context: the biggest challenge?

- Clinical information is inextricably entangled with the “context of its production”
- What the data were collected for affects how the data can be used
- Simple example: patient weight & height
  - Units and granularity: not major problems
  - General pediatric office versus endocrinologist office
    - How height measured (stadiometer versus yardstick)
    - How weight measured (clothed or unclothed)
- Intended use: are we tracking large groups of patients over time or establishing standards?
- If the latter, then height and weight data collected in the context of delivering care will not necessarily do

Context: the biggest challenge?

- Context is especially problematic when extracting information involving substantial coding discretion on the part of the clinician
  - Workload involved in finding a code
    - "Anemia" versus "transient erythroblastopenia of childhood"
  - Billing context (code affects payment)
  - Codes used locally or idiosyncratically (e.g., purulent rhinitis)

Context: an extreme position*

- “Data shall only be used for the purpose for which they were collected”
- “If no purpose was defined prior to collection of the data, then the data should not be used”

*van der Lei J. Use and abuse of computer-stored medical records. *Meth Inform Med.* 1991;30:79-80.

The law of medical information: Berg and Goorman

- “The further information has to be able to circulate (i.e. the more diverse contexts it has to be usable in),the more work is required to disentangle the information from the context of its production. The question that then becomes pertinent is; who has to do this work, and who reaps the benefits?”

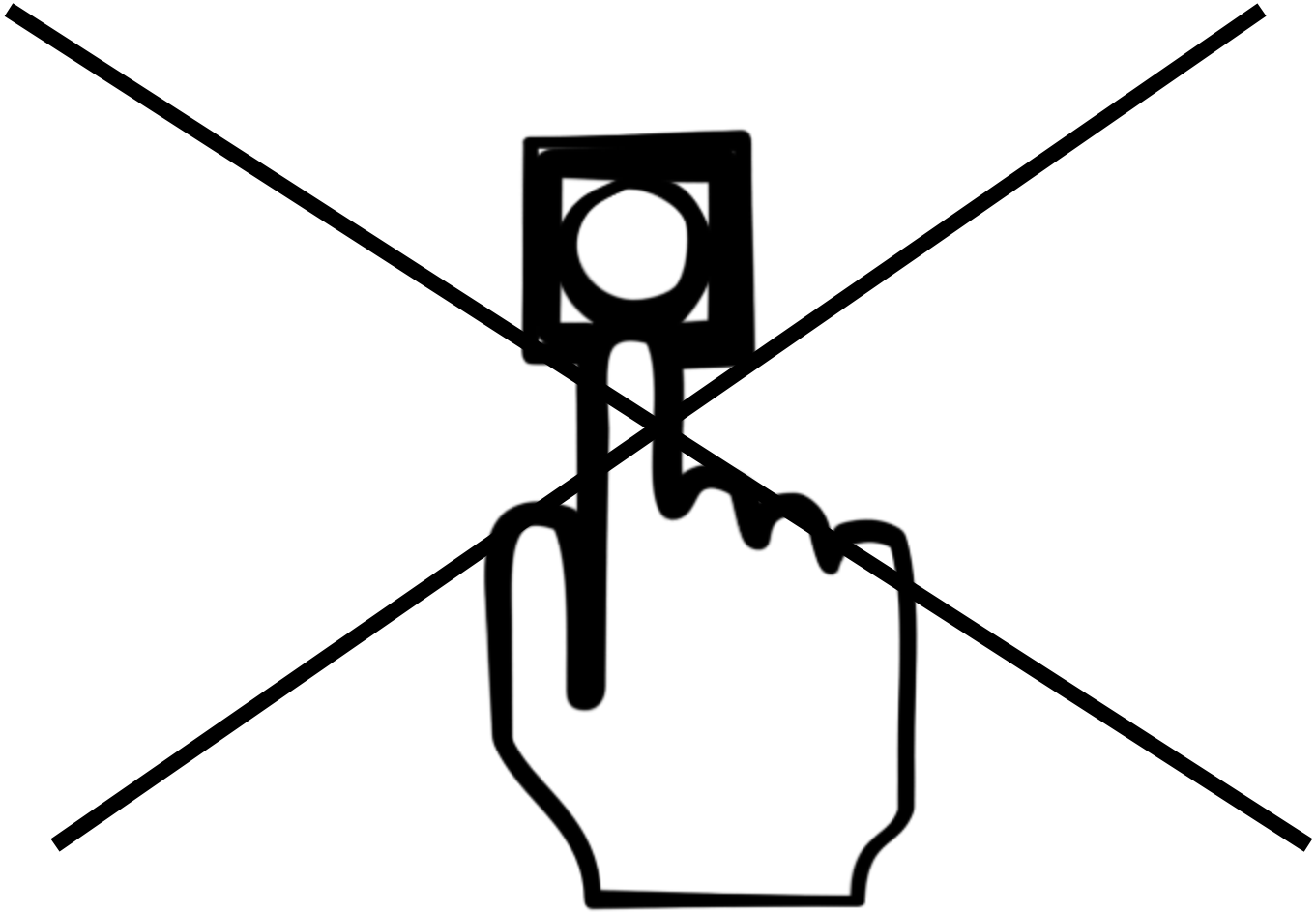
Int J Med Informatics 1999; 56:51–60

“Who has to do this work, and who reaps the benefits?”

- Some of the work can be done up front, by the clinician, by structuring data entry
- But if the clinician has to do some of the work, then the clinician must reap some of the benefits

Someday, all you'll have to do is push a button, and…

Someday, all you'll have to do is push a button, and…



Potential solutions to the problem of using EHR data for research?

I will leave this to my fellow presenters!!

Summary

- EHRs provide narrative and structured data that must be extracted, standardized and loaded in order to be analyzed
- PBRN clinicians can be spared the pain of data entry, but..
- Many challenges remain in using EHR data for research