Using EMR Data: Methods, Epidemiology and Biostat

David Thiemann, MD
Associate Professor of Medicine, Epidemiology and Health Sciences Informatics
Medical Director, JHM Center for Clinical Data Analysis

27 February 2018

Common Research Uses for EMR Data

- Identify cohort for manual chart review
  - Eligibility for randomized trial
  - Case-control or cohort studies
- Large-database epidemiology, long-term outpatient followup
  - Large simple trials, survival
  - Caveat: JHM is NOT a closed system, lousy survival/outcome data, lots of flawed impossible research proposals
- EMR system develop/engineering
  - Forms, decision support modules, .DLL, BPA
  - Profile/improve provider experience/behavior
- Big Data, Machine Learning, Artificial Intelligence, Natural Language Processing

Sipping from the EMR Fire Hose: The Data Minefield

- Junk science is easy
  - Load an Epic report into stat package-- voila!!
  - Ignore all underlying assumptions/confounding/bias, systematic/secular variability, unexplained heterogeneity, coding errors
- Honest solid EMR-based research is hard
  - Manual chart review is easier to design/control
- Force yourself to be cautious, hypervigilant
  - Everyone makes mistakes; good analysts constantly explore the data, re-test, extend, correct.

Don’t Start with Data!! Learn Clinical Workflow—How Data Gets Into EMR

- Visit units/clinics. Befriend senior RNs. Talk to end-users at every level, patient to CMA to attending.
- Learn how each individual data element gets into the system, by site, user, temporal epoch.
- Get screenshots!! Lots!!!
- Don’t re-invent the wheel.
  - Partner with QA/QI analysts, advanced RN projects, Armstrong Institute.
  - Seek out legacy reports/code.

Workflow Data Pitfalls (I)

- Varying usage of same form/data element
  - Site- or unit-specific variability (eg BMC vs JHH)
  - Varying time/scenario for completion (ED vs floor, pre vs post-op)
  - Overloaded operators: Variable usage (eg pre/post-event)
  - Completed variably by different users (triage, RN, CMA, MD)
- Undocumented events
  - CAUTI, CLABSI—lines get slammed in, fall out
- Undocumented knowledge
  - Scanned documents—pharmacogenomics in rheum
  - Ignored sepsis alerts—because sepsis already Dx + Rx

Workflow Data Pitfalls (II)

- Unrecognized dependencies/intercurrent events—timing is critical!!
  - Hgb pre vs. post-transfusion
  - Pain assessment scores interact with narcotic dosing
  - Identifying unbiased events/measures is hard
- Confounding by indication
  - Anemia, transfusion and PCI

EMR Data Quality/Complexity Vastly Complicate Good Research

- Analysis needs static data; EMR is a moving target
  - Data elements constantly re-defined—even labs
    - New assay, analyzer—different code
  - ICD9 coding NOT fairly comparable to ICD10
  - Private vs faculty providers
  - Decision support (CDS/BPAs) constantly evolving
    - Revised triggers, targets, wording, actions

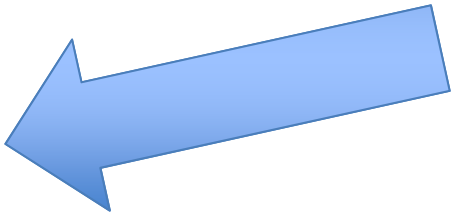
Coding (ICD/CPT) Gotchas

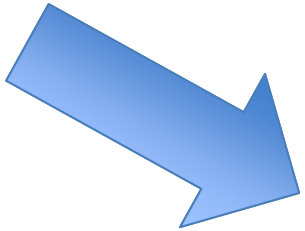
- ICD (Dx/procedure) and CPT (professional fee) codes are key basics, BUT:
- Don’t blindly query for ICD codes
  - Code existence doesn’t mean it’s used (or used accurately)—especially outpatient
  - Always compare to actual DB code frequency
- CPT limitations
  - Won’t capture unbilled house-staff procedures
  - Won’t capture private-physician charges
- Temporal evolution—codes get discontinued/superseded annually

EMR/Epic Data Nuances

- Always expect/test for missing/bad data
  - SQL “group by” query on every variable, by month/year
- Understand/analyze underlying source data structures, not just extract
  - Lab components may have different units, names by site and/or temporal epoch
  - Compound meds have recursive data structure
  - Flowsheets need intermediate data structures/joins
- Don’t blindly rely on orders
  - Orders not always completed, often modified by ancillary (esp radiology, pharmacy), timing uncertain
  - Focus on hard events—med admin, lab results, flowsheets

From DB to Analytic File





Data Cleaning/Management: 90% of Work (Analysis Is Easy!!)

- Use robust database tools that enforce data typing and primary keys, NOT Excel
- Start with separate files—labs, meds, vitals, imaging, whatever
- Use endogenous Epic classes rather than individual identifiers if possible
  - Lab component groups, pharm class and subclass
  - Using specific labels inevitably misses key data
- Carry primary keys throughout data transform steps
- Try multiple methods. Do they get the same cohort/result?
- Annotate code, document the whole ETL (extract-transform-load) process

Methods/Epi/Biostat Considerations

- Bias/confounding ubiquitous in EMR data. Address them.
  - Reviewers are getting smart about data science, shred sloppy work
- Methods section needs to describe in detail data cleaning, inclusion/exclusion criteria, missing data rates
  - Critical flaws typically buried in data cleaning, not fancy stat methods
- Fully describe (with table) excluded patients/observations, with reasons
- Test/prove robustness with multiple methods:
  - Vary propensity criteria, use alternate controls
  - Sensitivity analysis--“what-if” methods
- Don’t focus on AUC (area under curve) or sensitivity/specificity—low bars, no practical utility
  - Use reclassification indexes, predictive value (PPV/NPV), number needed to treat (NNT).

Creating an Flat File For Multivariate Analysis

- Begin by creating a single huge file of every conceivably useful variable; subset later
- Use explicit, defensible selection criteria for 1-to-many possibilities
- Create indicator/dummy variables for context/method/qualifier
  - Who entered data, where, interval (eg pre/post-op, ED/floor), intercurrent events (eg transfusion), datetime intervals
    - Missing data often has meaning; use dummy variables
- Explore/test surrogates for key unmeasured traits—frailty, survivability, nutritional/cognitive status, 6-minute walk
- Don’t use stepwise regression

Cautions

- Don’t use MRN as primary key-unstable
  - Use Epic enterprise id (E-ID) or study ID instead
- Don’t extract large DB directly to flat file
  - No redo capability, no inclusion/exclusion control
- Separate raw and analytic (no PHI) files
  - Calculated intervals, no dates/identifiers
- Beware of lead-time, lag-time and surveillance bias

Notes on Clinical Decision Support (CDS)

CDS cartoon

Be Realistic

- Maintainability
  - Everyone builds custom CDS. No one maintains.
  - If Epic has a reasonable native tool, enterprises will (and should) use that—even if yours is better.
- Bolt-on modules need 2-way data interfaces
  - Develop with Epic Clarity (or DW), not HL7/web services
  - Never use data polling—always event-driven model
  - Plan from the outset to feed real-time data back to Epic (often via hidden flowsheet rows). Systems that hoard data in standalone silos are stillborn.

User-Centric CDS Design

- Pop-ups are by definition bad decision support
  - Better to have users do things right the first time
  - Embed in-line CDS in workflow—relevant results, defaults, decision trees
- Generic “be careful” or “consider” guidance is useless (and annoying)
  - CDS needs to provide specific, transparent actions, links, buttons, ordersets
- Don’t expect users to enter data