An Introduction to the EDW

Daniel Schneider
Manager of Research Analytics

Biostats Seminar Series
09/13/2016
Agenda

• What is the EDW?
• What types of research can it help with?
• Important issues in clinical data warehousing
• Q/A
What is the EDW?
The EDW Provides a Campus Data Platform for Your Approved Projects

Participating Entities

Contribute
Report
Analyze

EDW

Load
Cleanse
Enhance

Campus Data

Provide Campus Data Platform

Analyze Data and Create Reports
What is the EDW?

The EDW is a SQL database that contains data from many of the transactional systems used by the campus.
What is the EDW?

Clinical data from NMG (EPIC)
Billing data from NMG (IDX/Epic)
Clinical data from NMH (Cerner)
Billing data from NMH (PRIMES)
More clinical systems (ancillary)
Research systems from FSM
What is the EDW?

The end goal of the EDW is to provide a more comprehensive dataset and options for interacting with and understanding it.
What is the EDW?

• Types of data in the EDW
  – Discrete data elements
    • Demographics, lab values, dates, diagnosis codes, form/flowsheet data, etc.
  – Text data
    • All clinical notes, pathology reports, image impressions, etc.
  – Messaging data
    • Phone calls, mychart messages, etc.
  – Audit level data
    • Chart accesses, logins, modifications

• Types of data NOT in the EDW
  – Images
  – Scanned documents, attachments
What is the EDW?
Research Uses of the EDW
Research Uses of the EDW

- Study Feasibility
- Clinical Trial Subject Recruitment
- Retrospective Chart Review
- Prospective, Longitudinal Studies
Study Feasibility

• Aggregate counts of patients meeting defined criteria
• No approval needed for prep-to-research data containing no PHI
• i2b2 is currently available and is being actively developed

Example:

“How many patients we've had over the years with ovarian cancer and a recurrence in the brain...”
Clinical Trial Subject Recruitment

• Real-time scheduling data from NMG
• Primary care physician contact information
• Daily updates to the subject population

Example:

“Patients with the diagnoses of peripheral arterial disease and/or coronary artery disease...”
Retrospective Chart Review

- Data in the EDW reaches back about to 1998...
- ... but becomes robust after 2000
- Discrete fields can be used to efficiently define and capture data elements for a population
- Complex NLP can be leveraged for free-text, though not as reliably as discrete fields

Example:

“The study is evaluating patients at NMG Ophthalmology with the following ICD9 codes between Jan 1, 2004 and Jan 1, 2011...”
Prospective, Longitudinal Studies

- Disease, other study registries can be built to keep track of a population, events and outcomes
- Data can be integrated with custom applications

Example:

“For consented patients provide the following data points for the first available Cardiac Cath Report: 1. Mean RA  2. Mean PA  3. PA Saturation %  4. Mean PCW  5. Mean LV  6. Fick Output...”
Getting Started
Getting Started…

- Submit Data Request using the EDW Portal
- Work with an EDW Analyst to refine requirements
- Service Agreement
  - Data Steward approvals
- Reports retrieved from the EDW Portal

It’s an iterative process!
Requesting an EDW Report

- IRB and NMHC approval is needed
- Discrete data elements are best
- Be able to send screen shots of the data elements you’re looking for
- Free text is difficult to work with
  - NLP can be used but is time consuming (expensive) and rarely as accurate as a discrete field
- Recharge rate of $75/hour
What is a Data Steward?

• The data in the EDW is still owned by the contributing organizations

• Data Stewards are representatives from these organizations and approve data and access requests

More information on the EDW Wiki
https://confluence.nubic.northwestern.edu/display/ED/Data+Stewards
NM EDW Self Service: i2b2

https://edwapps.nmff.org/NMEDWI2B2/Query
Nomenclatures and Standardized Vocabularies

Common national standards:
- RXNorm (medications)
- ICD-9, ICD-10 (diagnoses)
- SNOMED (Systematized Nomenclature of Medicine) (a bit of everything)
- LOINC (Logical Observation Identifiers Names and Codes) (labs)

Pros:
- Everyone starts to speak the same language
- Multi-site projects are enabled

Cons:
- Granularity is often lost because of the Lowest Common Denominator problem
- Not everything can fit into an existing vocabulary (smoking status, for instance)
Project One
Project One

What is Project One?

- Project One is the implementation of a single electronic medical record across Northwestern Medicine using Epic.
  - Project One is more than a technology tool -- it is a major commitment to the future of Northwestern Medicine, enabling our premier, integrated academic health system to deliver world-class, compassionate care.
## Guiding Principles

- **Patients first, safety always**
  (The needs of our patients and families will break ties)
- **Prioritize the use of Epic**
  (The Epic system will be the prioritized application when rationalizing between vendors)
- **Reduce the application portfolio**
  (Project One will purposefully reduce the application environment)
- **Make it efficient and keep it simple**
  (Solutions will prioritize efficiency and simplicity in design)
- **Improve the caregiver experience**
  (Caregiver value of the IT system will increase through design and implementation)
- **Drive toward an enterprise standard solution**
  (Where Epic does not provide a solution, an enterprise solution will be prioritized)
- **Solve for the health system**
  (Health system then region ... excellence across all)
- **Changes to existing applications will only occur if the changes are consistent with the Project One goals**
  (Changes should:
    - Have workflows consistent with the Project One goals
    - Be for regulatory, quality, patient safety or financial requirement
    - Be approved by Project One leadership)
Important Issues in Clinical Data Warehousing
Important Issues in Clinical Data Warehousing

GIGO: Garbage In, Garbage Out

- “We don’t make the news, we just report it.”
- If a system has a 43yo man with ovarian cancer, so will the EDW

When to clean data? Inbound vs outbound

- If data are cleaned as they arrive in the EDW, faith is sacrificed. One needs to be able to see the 43yo man with ovarian cancer in the EMR
- Few researchers want men with ovarian cancer in their reports; so we exclude the outliers when data are sent to the requester

Data provenance

- One should be able to trace any datum in the EDW all the way back to the source system
Future Directions
Future Directions

• Integrated Data Structures (IDS)
  – Incorporate new source data (Cadence, KishHealth, Marianjoy, etc.)

• Predictive Analytics/Biostatistics
  – Integrate R into our analytics infrastructure
  – Utilize advanced data visualizations

• Self-Service Query and Annotation
  – REDCap
  – Abstractor

• Predictive Analytics
  – RAPID

• Advanced Natural Language Processing
  – Will Thompson
Quick Resource Reference
NMEDW@northwestern.org
http://edw.northwestern.edu
https://grants.nubic.northwestern.edu/
Questions?