


EMORY UNIVERSITY Department of Biomedical Informatics



EHR Data Quality for Healthcare Analytics: Not Secondary Anymore

Andrew Post, MD, PhD
Assistant Professor and Clinical Informatics Architect
Department of Biomedical Informatics
Emory Healthcare Information Services
Emory University

EMORY UNIVERSITY Department of Biomedical Informatics

Financial Disclosures

- None

EMORY UNIVERSITY Department of Biomedical Informatics

Secondary Use of Clinical Data

Collected for one purpose, used for another

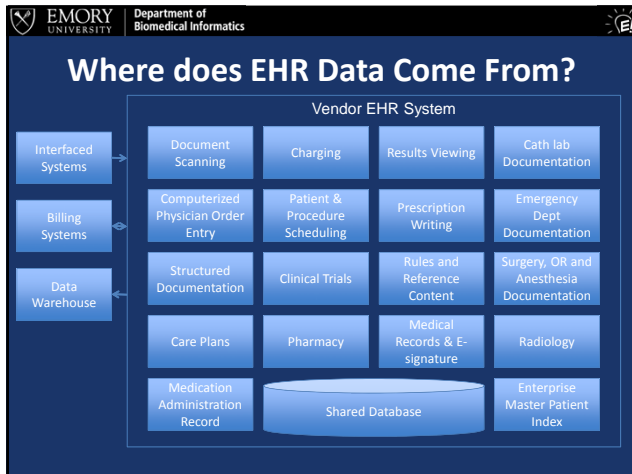
- Clinical and translational research
- Healthcare quality measurement and improvement
- Personal health records (PHRs)
- Health information exchange (HIE)
- Public health surveillance for emerging threats

EMORY UNIVERSITY Department of Biomedical Informatics

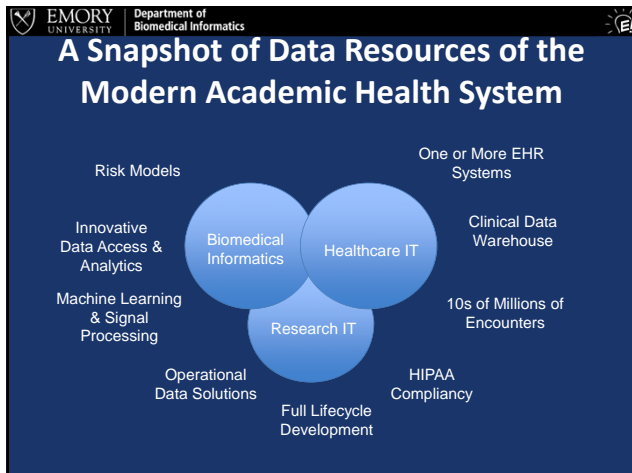
Challenges in Using Existing Data

- “Left censoring”: first instance of disease in record may not be when first manifested
- “Right censoring”: data source may not cover long enough time interval
- Data from other clinical (other health systems) and non-clinical (pharmacies) settings may not be captured
- Bias in testing and treatment
- Institutional or personal variation in practice or documentation styles
- Inconsistent use of coding or standards

Hersh, W. Secondary Use of Clinical Data from Electronic Health Records. Available from: <http://skynet.ohsu.edu/~hersh/secondary-use-pr.pdf> (presentation).



- ### Meaningful Use of an EHR
- Driven by 5 underlying goals for a healthcare system
 - Improving quality, safety and efficiency
 - Engaging patients in their care
 - Increasing coordination of care
 - Improving the health status of the population
 - Ensuring privacy and security
 - Three requirements – use of certified EHR tech
 - In a meaningful manner
 - Connected for health information exchange (HIE)
 - To submit information on clinical quality measures
- Hersh, W. Meaningful Use and All That: Update in Clinical Informatics. Available from: <http://skynet.ohsu.edu/~hersh/clinical-informatics-update.pdf> (presentation).



Health Information Exchanges

EMORY AND GRADY JOIN GEORGIA'S FIRST STATEWIDE ELECTRONIC HEALTH INFORMATION EXCHANGE

News Release

Emory and Grady Join Georgia's First Statewide Electronic Health Information Exchange

Secure network improves patient-centered care and health outcomes

Atlanta, Georgia (April 3, 2014) – Emory Healthcare and Grady Health System join Georgia's Department of Public Health and the state's Medicaid program in connecting to the Georgia Health Information Network (GAHIN), the statewide health information exchange network that electronically connects Georgia hospitals, physicians and clinicians to safely and securely exchange patient health information. Connecting to GAHIN allows Emory and Grady to augment continuity of care by enhancing patient care coordination among its authorized providers and affiliates and improves access to a patient's information for diagnosis and treatment.

"Traditionally, patient health information has been difficult to share across care settings," said

<http://www.gahin.org/media/press-release/emory-and-grady-join-georgia-s-first-statewide-electronic-health-information>

EMORY UNIVERSITY Department of Biomedical Informatics

Big Data

- Storage capacity has outstripped advances in the speed of reading data from disk and the amount of RAM in typical computers for processing it
- Data management challenges around volume, velocity and variety – Doug Laney, META Group, 2001
- “Big Data” requires specialized tools because it doesn’t fit into RAM
- Big Data tasks usually involve data reuse for a new purpose

EMORY UNIVERSITY Department of Biomedical Informatics

Data Variety (Complexity) is the Challenge

“...no greater barrier to effective data exist than the variety of inconsistent, non-aligned data structures, semantic...”

Doug Laney, “Managing the Data Volume Velocity Variety,” Group, 2001

EMORY UNIVERSITY Department of Biomedical Informatics

Data Science/eScience

<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>

EMORY UNIVERSITY Department of Biomedical Informatics

Data Science Characteristics

80% of the work
– Aaron Kimball

90% of their work
– Scientists at Univ. of Washington
(the other 10% is science)

- Get and clean lots of data, often from more than one source
- Define a data model describing the “typical” regions in the data
- Look for “outliers” – data that doesn’t fit the model
- Look for individual outliers – credit card transaction involving an usually large amount of money
- Look for outliers in sequence data
 - sequence of credit card transactions at businesses near highways in neighboring states
 - spike in tweets about cold symptoms
- Typical behavior may change with time, so update the model using new data frequently

Hospital Readmissions within 30 Days
 5 years of data from local data warehouse – 230,000 hospital encounters
 5 years of data from the UHC Clinical Database – ~20 million hospital encounters

<p>Relevant Medical Record Data</p> <ul style="list-style-type: none"> • Admissions/discharges • Diagnosis codes (18,012) • Procedure codes (4,665) • APR DRGs (318) • MS DRGs (750) • Medication orders (32,936) • Laboratory test results (3,117) • Vital signs • Geographic information 	<p>Variables of Interest</p> <ul style="list-style-type: none"> • Chronic co-morbidities • Specialty providing care • Variables derived from clinical results <ul style="list-style-type: none"> – Obesity – Diabetes/uncontrolled diabetes – End-stage renal disease (ESRD) – Pressure ulcer – Sickle cell disease/sickle cell crisis • Temporal variables <ul style="list-style-type: none"> – Multiple previous MI – Multiple 30-day readmissions – Chemotherapy prior to surgery – Recent previous encounter
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Automated Chart Abstraction

Post AR & Harrison JH. J Am Med Inform Assoc 2007;14(5):674-83

Abstraction of Data into Intervals

Abstraction of Intervals into Information

Elev. BP in Hypertensive on Diuretic (temporal pattern)

Knowledge Base

Standard Cohort Definitions (<http://www.phekb.org>)

PheKB - a knowledgebase for discovering phenotypes from electronic medical records

Home | Phenotypes | Implementations | Tools | Groups | MERGE Network | Contact Us

Phenotypes > Severe Early Childhood Obesity > Implementations by Phenotype

Implementations by Phenotype

View Implementations

Severe Early Childhood Obesity - Vanderbilt

Phenotype: Severe Early Childhood Obesity

Algorithm Type: Class, Control

Implementation Details: We performed manual review of 75 potential cases and 40 potential controls based on the ICD9 Exclusion Criteria prior to review of BMI and height criteria.

a. Class Criteria: Manual review of 75 charts identified using the ICD9-based case exclusion criteria revealed that 74/75 were potential cases. PPV = 98.7

b. Control Criteria: Manual review of 40 charts identified using the ICD9-based case exclusion criteria revealed that 38/40 were potential controls. PPV = 95.0

The height and BMI percentiles data for potential cases and controls were then evaluated for all potential cases and controls. There are 4 severe early childhood obesity cases and 4 controls meeting the height and BMI criteria with QIMC data available from the Vanderbilt site.

Case	True Positive	False Positive
Predicted Class (Observation)	74.00 Correctly classified	1.00 Incorrectly classified
Positive Predictive Value	98.67%	

Control	True Positive	False Positive
Predicted Class (Observation)	38.00 Correctly classified	1.00 Incorrectly classified
Positive Predictive Value	97.5%	

Standard Quality Measures

NATIONAL QUALITY FORUM

0421 Preventive Care and Screening: Body Mass Index (BMI) Screening and Follow-Up

Measure Description: Percentage of patients aged 18 years and older with a documented BMI during the current encounter or during the previous six months AND when the BMI is outside of normal parameters, a follow-up plan is documented during the encounter or during the previous six months of the encounter.

Normal Parameters: Age 40 years and older BMI <= 27 and >= 30
Age 18 - 39 years BMI <= 18.5 and >= 25

Measure Statement: Patients with a documented BMI during the encounter or during the previous six months, AND when the BMI is outside of normal parameters, follow-up is documented during the encounter or during the previous six months of the encounter with the BMI outside of normal parameters.

Denominator Statement: All patients aged 18 years and older.

Exclusions: A patient is identified as a Denominator Exclusion (DE) and excluded from the Total Denominator Population (TDP) in the Performance Measurement (PM) calculation for any one or more of the following reasons:

- Patient is pregnant
- Patient refused BMI measurement (includes height and/or weight)
- An other reason documented in the medical record by the provider why BMI calculation or follow-up plan was not implemented
- Patient is in a long-term care facility or other setting where the use of the measure, and its follow-up treatment would jeopardize the patient's health status.

NA Adjustment: No

Classification: National Quality Strategy Domains: Health and Well-Being

Measure Status Contact Information: An additional measure user contact information page is located here.

Found in Portfolio(s): 2012 Joint Commission Family of Measures, ACC - ACOG Measures, AHA/ACC Measures, Diabetes Care and Measures, Health and Well-Being, Healthy Work-Care-Community and Quality of Life, Measure Management Dashboard

<http://www.qualityforum.org/Home.aspx>

Shared Data Models

<http://omop.org/CDM>

Observational Medical Decision Partnership

Common Data Model

The purpose of the Common Data Model (CDM) is to standardize the format and content of the observational data, so standardized applications, tools and methods can be applied to them. This page explains the Common Data Model. It also provides a collection of Data ETLs for a number of popular databases.

1. Common Data Model

This is the latest CDM in Version 4.0. In addition to person, condition, drug, procedure and visit information, it now includes encounter information. This and related health economics, cost data and hospital treatment indicator tables, including financial source tables, comprehensive electronic and hardware tables.

- CDM Specifications v4.0
- CDM Table DDL (DDL format)

This is the draft specification for Version 3 of the OMOP Common Data Model. These documents are draft and are for public comment from the OMOP community through 30 June 2014.

- CDM Draft Specifications v3
- CDM changes from CDMv2 to CDMv3

2. Extract, Transform and Load (ETL)

This is the most time consuming part of creating a database in OMOP format. You need to write a script or program to extract your data to meet the specifications. To ease this process and get things organized we provide a template for loading:

- ETL Mapping Template - Word version
- ETL Mapping Template - PDF version

In addition to mapping and transforming of the data to the CDM, the content also has to conform to the Standard Terminology. This is performed through a process called Terminology Mapping described in the Standard Terminology. The following is a list of existing ETL implementations for a number of popular databases. All of them are Open Source. However, note that some of them are conforming to previous versions of the CDM. Please consider sharing your updated ETL with the OMOP community as well.

ETL

- ETL Mapping for V4.0
- ETL Code for V4.0

Tools

- ETL Mapping for V4.0
- ETL Code for V4.0
- HSQL Mapping for V4.0

Computing Derived Measures with Shared Tools

<http://eurekaclinical.org>

Specify patient features of interest

Spreadsheets

CSV

Clinical databases

Phenotype Editor

Action	Name	Description	Type	Created Date	Last Modified
+	High Blood Pressure	High blood pressure. Includes signs for all patients and patients with CDD or diabetic.	Visit Threshold	02/06/2013	07/31/2013
+	2 consecutive high blood pressure values within 180 days	2 consecutive high blood pressure values within 180 days.	Frequency	02/06/2013	02/15/2013
+	Digestive procedure with prior chemo	Digestive procedure with prior chemo	Sequence	02/14/2013	02/14/2013
+	Reasonism	A reasonism.	Frequency	02/15/2013	02/15/2013

Compute them in millions of patient records

High BP (category) > Hypertension (category) > Second Hypertension (temporal slice) > Hypertension (category)

On Diabetic (category) > Elevated BP (category) > High Systolic BP (category) > High Diastolic BP (category)

Time

Post AR, Kurc T et al. AMIA Annu Symp Proc 2013;1160-69

EMORY UNIVERSITY Department of Biomedical Informatics

Cohort Discovery and Data Extraction

<http://www.i2b2.org>

i2b2 Query & Analysis Tool

Find Patients | Analysis Tools | Message Log | Help | Logout

Query Tool

Query Name:

Group 1: 45-64 years old
Group 2: Cancer
Group 3: Encounter with subsequent

one or more of these AND one or more of these AND one or more of these

Run Query New Query 3 Groups New Group

Query Status

Navigation Terms: Gender, Language, Marital Status, Race, Religion, Discharge Disposition, Hospital Readmissions, ICD9 Diagnostic Codes, ICD9 Procedure Codes, Laboratory Tests, Medication, User-defined Derived Variables, Cancer, Chemotherapy encounter, Chronic kidney disease (CKD), Chronic obstructive pulmonary disease, Diabetes, Encounter in last 180 days, Encounter in last 90 days, Encounter with subsequent 30-day readmission, End-stage renal disease (ESRD), First 2 MI

EMORY UNIVERSITY Department of Biomedical Informatics

HIEs for Population Studies

Application of Information Technology

The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories

GRIFFIN M. WEBER, MD, PhD, SHAWN N. MURPHY, MD, PhD, ANDREW J. McMURRY, MS, DOUGLAS MACFADDEN, MS, DANIEL J. NIGREN, MD, MS, SUSANNE CHURCHILL, PhD, ISAAC S. KOHANE, MD, PhD

Abstract The authors developed a prototype Shared Health Research Information Network (SHRINE) to identify the technical, regulatory, and political challenges of creating a federated query tool for clinical data repositories. Separate Institutional Review Boards (IRBs) at Harvard's three largest affiliated health centers approved use of their data, and the Harvard Medical School IRB approved building a Query Aggregator Interface that can simultaneously send queries to each hospital and display aggregate counts of the number of matching patients. Our experience creating three local repositories using the open source Informatics for Integrating Biology and the Bedside (i2b2) platform can be used as a road map for other institutions. The authors are actively working with the IRBs and regulatory groups to develop procedures that will ultimately allow investigators to obtain identified patient data and biomaterials through SHRINE. This will guide us in creating a future technical architecture that is scalable to a national level, compliant with ethical guidelines, and protective of the interests of the participating hospitals.

■ J Am Med Inform Assoc. 2009;16:624-630. DOI 10.1197/jamia.M3191.

EMORY UNIVERSITY Department of Biomedical Informatics

Secondary Use – Where Are We?

1. Comprehensive electronic health record
2. Comprehensive clinical data warehouse ←
3. Shared data representations for research and quality
4. Shared definitions of derived measures of interest in research and quality
5. Tools for cohort discovery and export into analytics and clinical research tools
6. Health information networks for research and quality

EMORY UNIVERSITY Department of Biomedical Informatics

QUESTIONS AND DISCUSSION