Tackling Practical Methodological Challenges of Using Electronic Data for CER & PCOR

Jay R. Desai  
*HealthPartners*

Michael Kahn  
*University of Colorado, Denver*

Russell E. Glasgow  
*National Cancer Institute*

Follow this and additional works at: [http://repository.academyhealth.org/webinars](http://repository.academyhealth.org/webinars)

Part of the [Health Information Technology Commons](http://repository.academyhealth.org/webinars) and the [Health Services Research Commons](http://repository.academyhealth.org/webinars)

**Recommended Citation**

[http://repository.academyhealth.org/webinars/7](http://repository.academyhealth.org/webinars/7)

This Video/Media is brought to you for free and open access by the Events at EDM Forum Community. It has been accepted for inclusion in Webinars by an authorized administrator of EDM Forum Community.
Tackling Practical Methodological Challenges of Using Electronic Data for CER & PCOR

Jay R. Desai, MPH, HealthPartners
Michael Kahn, MD, PhD, University of Colorado, Denver
Russell E. Glasgow, PhD, National Cancer Institute

December 18, 2012
Welcome

Erin Holve, Ph.D., M.P.H., M.P.P.

- Director of Research & Education, AcademyHealth
- Principal Investigator of the EDM Forum
AcademyHealth: Improving Health & Health Care

AcademyHealth is a leading national organization serving the fields of health services and policy research and the professionals who produce and use this important work.

Together with our members, we offer programs and services that support the development and use of rigorous, relevant and timely evidence to:

1. Increase the quality, accessibility and value of health care,
2. Reduce disparities, and
3. Improve health.

A trusted broker of information, AcademyHealth brings stakeholders together to address the current and future needs of an evolving health system, inform health policy, and translate evidence into action.
EDM Forum Updates

→ **New Issue Brief:**
  - *Building the Informatics Infrastructure for CER: A Review of the Grey Literature*

→ **Coming Soon:**
  - eGEMs, the new EDM Forum open-access, peer-reviewed e-publication
  - Upcoming Methods Papers and Issue Briefs:
    - *Effect Identification in CER* (December 2012)
    - *Getting Answers We Can Believe In: Considering Research Design and Methods When Using Electronic Medical Data for Clinical Research* (January 2013)
    - *Collecting and Using ePRO for CER and PCOR: Opportunities and Challenges* (January 2013)
  - Methods Webinars: Data Quality Reporting, Addressing Selection Bias, and more! (Spring 2013)
Connecting the Audio

- The audio will be delivered through your computer speakers. Speakers or headphones are required to hear the audio portion of the webinar.

- If you do not hear any audio now, check your computer’s speaker settings and volume.

- If you need an alternate method of accessing audio, please submit a question through the Q&A pod.
Technical Assistance

➔ Live technical assistance:
  – Call Adobe Connect at (800) 422-3623

➔ Refer to the ‘Technical Assistance’ box in the bottom left corner for tips to resolve common technical difficulties.
Questions may be submitted at any time during the presentation

To submit a question:
1. Click in the Q&A box on the left side of your screen
2. Type your question into the dialog box and click the Send button
Learning Objectives

- Present a “fit-for-use” conceptual model for data quality assessment and a process model for planning and conducting single-site and multisite data quality assessments.
- Describe selected conceptual and practical challenges related to development of multi-site diabetes and asthma registries, including the development of case definitions, validation of case identification methods, and variation in electronic health data sources.
Developing a comprehensive data quality assessment framework

Michael G. Kahn*,1,2, Marsha A. Raebel3,4, Jason M. Glanz3,5, Karen Riedlinger6, John F. Steiner3

1. Department of Pediatrics, University of Colorado Anschutz Medical Center, Aurora Colorado
2. Colorado Clinical and Translational Sciences Institute, University of Colorado, Aurora Colorado
3. Institute for Health Research, Kaiser Permanente Colorado, Denver Colorado
4. School of Pharmacy, University of Colorado, Aurora, Colorado
5. Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado
6. Northwest Kaiser Center for Health Research, Portland Oregon

Electronic Data Methods (EDM) Forum Analytic Methods Webinar Series
Tackling Practical Methodological Challenges of Using Electronic Data for CER & PCOR
18 December 2012
Michael.Kahn@ucdenver.edu

Funding was provided by a contract from AcademyHealth. Additional support was provided by AHRQ 1R01HS019912-01 (Scalable PArtnering Network for CER: Across Lifespan, Conditions, and Settings), AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network), and NIH/NCRR Colorado CTSI Grant Number UL1 RR025780 (Colorado Clinical and Translational Sciences Institute).
Disclosures

Presentation based on EDM Forum commissioned paper:

**A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research**

*Michael G. Kahn, MD, PhD,*† Marsha A. Raebel, PharmD, ‡§ Jason M. Glanz, PhD, MS, ¶||
Karen Riedlinger, MPH, MT (ASCP), ‡ and John F. Steiner, MD, MPH‡

**Introduction:** Answers to clinical and public health research questions increasingly require aggregated data from multiple sites. Data from electronic health records and other clinical sources are useful for such studies, but require stringent quality assessment. Data quality assessment is particularly important in multisite studies to distinguish true variations in care from data quality problems.

**Key Words:** data quality, data quality assessment, single-site studies, multisite studies

*(Med Care 2012;50: S21–S29)*
What is the issue?

- Poor data quality can invalidate research findings
  - Cohort identification
  - Risk factors / exposures / confounders
  - Interventions
  - Outcomes

- Data quality in non-research settings even more problematic
  - Documentation practices
  - Workflow
  - Diligence to data quality

- Our focus: how to assess data quality systematically?
Measuring Data Quality

- Observed versus expected distributions
- Outliers
- Missing values
- Performance on data validity checks
  - Single attribute analysis
  - Double- / triple- / higher level attributes correlations
  - Physical / logical domain impossibilities
Why is a **systematic** data quality assessment framework useful?

- Data quality (DQ) assessment is standard practice
  - We know what features of DQ we have looked at
  - *We may not know what DQ features we have not looked at*

- A comprehensive evaluation of data quality is resource intensive
  - Focus on aspects that matter
  - If data use changes, are DQ assessments adequate?
Data Quality in **Electronic Medical Records**

- Data collection tools optimized for efficiency
  - Text templates
  - Copy/paste

- Minimal data validation checks
  - Min/Max limits
  - Pick lists
  - Required fields
Comparative Temporal Trends: Serum Glucose
Data Quality Assessment Stages

• Stage 1: initial assessments of source data sets prior to analysis
  – Simple global analyses, visualizations, descriptive statistics
  – Both single-site and multi-site

• Stage 2: Study-specific analytic subsets with complex models and detailed data validations focused on dependent and independent variables.
Data quality assessment lifecycles

1. Site level
   - Extraction from EMR
   - Data quality assessments

2. Multi-site level
   - Data quality assessments
   - Data merging

3. Final analytic data set
Defining Data Quality using “Fitness for Use”

“Data are of high quality if they are fit for their intended uses in operations, decision making, and planning. Data are fit for use if they are free of defects and possess desired features.”

– Joseph Juran

• Quality linked to intended use (context)
  – Not all tasks require highly accurate data
  – The same data may have different data quality with different intended uses
Beyond accuracy: What data quality means to data consumers
Wang, Richard Y; Strong, Diane M
Journal of Management Information Systems; Spring 1996; 12, 4; ABI/INFORM Global
pg. 5

Beyond Accuracy: What Data Quality Means to Data Consumers

RICHARD Y. WANG AND DIANE M. STRONG

RICHARD Y. WANG is Associate Professor of Information Technologies (IT) and Co-Director for Total Data Quality Management (TDQM) at the MIT Sloan School of Management, where he received a Ph.D. degree with an IT concentration. He is a major proponent of data quality research, with more than twenty papers written to
<table>
<thead>
<tr>
<th>Category</th>
<th>Domains</th>
<th>Technical definition</th>
<th>Clinical data redefinition</th>
<th>CER Adaptation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic: Data quality features that are</td>
<td>Accuracy</td>
<td>The extent to which data are correct, reliable, and free of error.</td>
<td>Data values represent the true state of a patient within the limitations of the measurement methods.</td>
<td>Mishandled specimen (e.g., potassium levels in hemolyzed serum); test performance against a gold standard</td>
<td></td>
</tr>
<tr>
<td>inherent to data alone</td>
<td>Objectivity</td>
<td>The extent to which data are unbiased and impartial.</td>
<td>The methods used to obtain data values are well described and represent best practices. Component values represent the total clinical measurement.</td>
<td>CLIA-approved laboratory tests. Use of standardized psychometric instruments to assess patient status</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Believability</td>
<td>The extent to which data are regarded as true, real, and credible</td>
<td>Independent measurements make clinical sense.</td>
<td>Sex agreement with sex-specific features (pregnancy, prostate cancer). Collections of related measures are physiologically consistent (AST, ALT, Bilirubin, PTT in liver failure) Clinical trials SOPs</td>
<td></td>
</tr>
<tr>
<td>Conceptual: Data quality features that are</td>
<td>Timeliness</td>
<td>The extent to which the age of the data is appropriate for the task at hand.</td>
<td>Serial measurements over time sufficient to detect clinical state</td>
<td>Blood pressure measurements for diagnosing hypertension linked in time to clinic visits.</td>
<td></td>
</tr>
<tr>
<td>relevant in the context of the task for which the data are to be used</td>
<td>Appropriate amount of data</td>
<td>The extent to which the quantity or volume of available data is appropriate.</td>
<td>Data are present or absent as expected</td>
<td>Degree and distribution of missingness.</td>
<td></td>
</tr>
</tbody>
</table>
How to **measure** data quality?

- Need to link conceptual framework with methods
- Maydanchik:

  ![Data Quality Assessment](image)
How to measure data quality?

- Maydanchik: Five classes of data quality rules
  - Attribute domain: validate individual values
  - Relational integrity: accurate relationships between tables, records and fields across multiple tables
  - Historical: time-vary data
  - State-dependent: changes follow expected transitions
  - Dependency: follow real-world behaviors

Data Quality Assessment METHODS

- Five classes of data quality rules → 30 assessment methods
  - Attribute domain rules (5 methods)
  - Relational integrity: (4 methods)
  - Historical: (9 methods)
  - State-dependent: (7 methods)
  - Dependency: (5 methods)

Time and change assessments dominate!!
EDM-F sponsored Data Quality Collaborative Workshop

- Held 6 December 2012
- Approximately 20 participants
- Focused on
  - A draft white paper on **data quality reporting**
  - Data quality metadata
  - Operationalizing data quality reporting
A consensus-based data quality reporting framework for observational healthcare data

Members of the EDM Forum Data Quality Collaborative

1 In alphabetical order by last name: Jeffrey Brown – Harvard Pilgrim Health Care Institute/Harvard Medical School; Michael Kahn – University of Colorado; Daniela Meeker – Rand Corporation; Meredith Nahm – Duke University; Patrick Ryan – OMOP/J&J; Lisa Schilling – University of Colorado; Nicole Weiskopf – Columbia University; Andrew Williams – Kaiser Permanente Hawaii.

2. Problem Statement

Electronic health records (EHRs) support the capture of detailed clinical, operational, and administrative data as part of routine clinical and administrative processes. Electronic data from administrative sources, billing and claims databases, drug fulfillment, specialized registries and patient-reported data expand the scope and richness of available patient-level data. As EHR use becomes the norm, the availability of electronic data from a variety of practice and patient settings provides an opportunity to transform the spectrum of clinical research, allowing critical insights to be gained from clinical data.
<table>
<thead>
<tr>
<th>Item#</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Original data source</td>
</tr>
<tr>
<td>Data origin</td>
<td>1</td>
</tr>
<tr>
<td>Data capture method</td>
<td>2</td>
</tr>
<tr>
<td>Original collection purpose</td>
<td>3</td>
</tr>
<tr>
<td>2.</td>
<td>Data custodian information</td>
</tr>
<tr>
<td>Data provenance</td>
<td>4</td>
</tr>
<tr>
<td>Data custodian’s database model</td>
<td>5</td>
</tr>
<tr>
<td>Data custodian’s data dictionary</td>
<td>6</td>
</tr>
<tr>
<td>DATA PROCESSING</td>
<td></td>
</tr>
<tr>
<td>Data extraction, including use of natural language processing, specifications</td>
<td>7</td>
</tr>
<tr>
<td>Mappings from original values to standardized values</td>
<td>8</td>
</tr>
<tr>
<td>Data management organization’s data transformation routines, including constructed variables</td>
<td>9</td>
</tr>
<tr>
<td>Data validation routines</td>
<td>10</td>
</tr>
<tr>
<td>Audit trail</td>
<td>11</td>
</tr>
<tr>
<td>DATA ELEMENTS CHARACTERIZATION Documentation</td>
<td></td>
</tr>
<tr>
<td>Data model verified</td>
<td>12</td>
</tr>
<tr>
<td>Single element data domain descriptive statistics</td>
<td>13</td>
</tr>
<tr>
<td>Temporal data</td>
<td>14</td>
</tr>
<tr>
<td>Multiple variables cross validations</td>
<td>15</td>
</tr>
<tr>
<td>STUDY SPECIFIC DATA QUALITY Documentation (as applied by investigators or analytic team)</td>
<td></td>
</tr>
<tr>
<td>Data cleansing/customization</td>
<td>17</td>
</tr>
<tr>
<td>Data quality checks of key variables used for cohort identification</td>
<td>18</td>
</tr>
<tr>
<td>Data quality checks of key variables used for outcome categorization</td>
<td>19</td>
</tr>
</tbody>
</table>
Developing a comprehensive data quality assessment framework

Questions?

Electronic Data Methods (EDM) Forum Analytic Methods Webinar Series
Tackling Practical Methodological Challenges of Using Electronic Data for CER & PCOR
18 December 2012
Michael.Kahn@ucdenver.edu

Funding was provided by a contract from AcademyHealth. Additional support was provided by AHRQ 1R01HS019912-01 (Scalable PArtnering Network for CER: Across Lifespan, Conditions, and Settings), AHRQ 1R01HS019908 (Scalable Architecture for Federated Translational Inquiries Network), and NIH/NCRR Colorado CTSI Grant Number UL1 RR025780 (Colorado Clinical and Translational Sciences Institute).
Extra credit slides: A formative proposal

- President’s Council of Advisors on Science and Technology (PCAST)
  - Recommended mandatory “metadata” tags attached to all HIT data elements
    - Metadata are descriptions of the data
    - PCAST proposed tags: data provenance, privacy permissions/restrictions
The Office of the National Coordinator should:

- Move more boldly to ensure that the Nation has electronic health systems that are able to exchange health data in a universal manner based on metadata-tagged data elements. In particular, ONC should signal now that systems will need to have this capability by 2013 in order to be deemed as making “meaningful use” of electronic health information under the HITECH Act.

- Act to establish initial minimal standards for the metadata associated with tagged data elements, and develop a roadmap for more complete standards over time.
Extra credit slides: A formative proposal

• CER community defines metadata tags that describe data quality for data elements and data sets
  – Simple distributions (mean, median, min, max, missingness, histograms)
    • ala OMOP OSCAR
  – More comprehensive set of measures derived from this framework
• Data quality metadata tags attached to all data sets for interpretation/evaluation by end-user
Dimension 1: Attribute domain constraints

<table>
<thead>
<tr>
<th>Data quality rule</th>
<th>Definition</th>
<th>Assessment methods</th>
<th>Data quality assessment methods examples</th>
</tr>
</thead>
</table>
| Attribute domain      | Rules that validate individual attribute values based on restrictions on allowed values. | Attribute profiling                                                                | • Basic aggregate statistics (counts, means, medians, minimum and maximum values  
• Examination of highest and lowest values  
• Value distributions (histograms) |
| constraints           |                                                                           | Optionality                                                                       | • Number of NULL (missing) values.  
• Use of default values to denote missingness (9999, 1/1/1900) |
| Format                |                                                                           |                                                                                   | • DDMYYYY versus MMDDYYYY versus YYYYMMDD  
• SSN: NNN-NN-NNNN      |
| Valid values          |                                                                           |                                                                                   | • Gender = ‘M’ or ‘F’ or ‘U’ only  
• Route of administration = ‘PO’, ‘IM’, ‘IV’, ‘other’ only  
• Blood glucose cannot be a negative value  
• Age cannot be > 120 years |
| Precision             |                                                                           |                                                                                   | • Units of measure  
• Rounding rules  
• Date/time precision |
### Dimension 2: Relational integrity rules

<table>
<thead>
<tr>
<th>Data quality rule</th>
<th>Definition</th>
<th>Assessment methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relational integrity rules</td>
<td>Rules that ensure accurate relationships between entities (tables), instances (records), and attributes (fields) across multiple tables</td>
<td>Identity:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Different unique IDs (keys) refer to distinct things (person, place, concept or event)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The same unique ID (e.g. SSN) refers to same entity (e.g. person)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reference:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• A reference in one table to data in another table points to a row that exists in the second table.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cardinality:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Relationship cardinality profiling – the count of the actual number of occurrences for each relationship in the database.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• References to data in a table refer to no more than the allowed number of occurrences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inheritance:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Entities are grouped into types and subtypes correctly (e.g., all patients, parents, spouses, and employees are also persons)</td>
</tr>
<tr>
<td>Data quality rule</td>
<td>Definition</td>
<td>Assessment methods</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------</td>
<td>------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Historical data rules</td>
<td>Rules involving time-varying data</td>
<td><strong>Currency</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Retention</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Granularity</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Continuity</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Timeline patterns</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Value patterns</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Event dependencies</td>
<td><strong>Event dependencies</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Event conditions</td>
<td><strong>Event conditions</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Event attributes</td>
<td><strong>Event attributes</strong></td>
</tr>
</tbody>
</table>
**Dimension 4: State-dependent rules**

<table>
<thead>
<tr>
<th>Data quality rule</th>
<th>Definition</th>
<th>Assessment methods</th>
<th>Data quality assessment methods examples</th>
</tr>
</thead>
</table>
| State-dependent objects rules | Rules that ensure that changes in the lifecycle of an object follow expected transitions | State-transition profiling | • Examination of valid transition states over time.  
• A terminal state cannot be followed by another state (e.g., an expired patient cannot be readmitted).  
• A valid action or event is associated with a corresponding state transition (e.g., cardiac arrest precedes patient state transition to expired). |
|                    |            | State domain       | • An object’s state can only be a valid value (e.g. either admitted or discharged from hospital). |
|                    |            | Action domain      | • The set of actions that can be applied to an object can only be a valid value (e.g. a test can be performed or cancelled, but not both). |
|                    |            | Terminator domain  | • States in which an object can start or stop its lifecycle can only be a valid value. |
|                    |            |                    |                                          |
|                    |            | State-actions      | • Each action is consistent with the change in state that it engenders (e.g., myocardial infarction cannot lead to subsequent “no cardiac disease” state). |
|                    |            | Continuity         | • Sequence or timing of start of each state record must follow the end of the previous state record. |
|                    |            | Duration           | • The minimum or maximum length of time an object can stay in a specific state (e.g., admission for myocardial infarction cannot be less than 1 day).  
• Zero-length rule: End date of a state must be later than the start date. |
### Dimension 5: Attribute dependency rules

<table>
<thead>
<tr>
<th>Data quality rule</th>
<th>Definition</th>
<th>Assessment methods</th>
<th>Data quality assessment methods examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute dependency rules</td>
<td>Rules for describing real world objects</td>
<td>Redundant attributes</td>
<td>• Same attributes in different data sources should have identical values</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Historical measures of a non-time varying attribute should have identical values</td>
</tr>
<tr>
<td>Derived attributes</td>
<td></td>
<td>• An aggregated value must equal the total of the atomic level values.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• An aggregated value must follow appropriate rules when component data elements are missing</td>
<td></td>
</tr>
<tr>
<td>Partially dependent attributes</td>
<td></td>
<td>• The allowed values of one attribute are limited by the assigned value of another attribute (e.g., sex = female eliminates prostate cancer as a valid diagnosis).</td>
<td></td>
</tr>
<tr>
<td>Conditional optionality</td>
<td></td>
<td>• The allowed values of one attribute determines if another attribute must be (cannot be) present (e.g., a discharge disposition = “to home” implies expiration date = null).</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Mutually exclusivity – the presence of any value in one attribute precludes (requires) a value in another attribute</td>
<td></td>
</tr>
<tr>
<td>Correlated attributes</td>
<td></td>
<td>• Values in one attribute changes the likelihood of values in another attribute (e.g., sex = male and age = 65 and smoker = yes increases the likelihood of discharge diagnosis = myocardial infarction)</td>
<td></td>
</tr>
</tbody>
</table>
MULTI-SITE DIABETES DATALINK USING ELECTRONIC HEALTH RECORDS: IDENTIFICATION, VALIDATION, AND REPRESENTATIVENESS

Jay Desai
HealthPartners Institute for Education and Research

EDM Forum Webinar
December 18, 2012
DIABETES AND ASTHMA CASE IDENTIFICATION, VALIDATION, AND REPRESENTATIVENESS WHEN USING ELECTRONIC HEALTH DATA TO CONSTRUCT REGISTRIES FOR COMPARATIVE EFFECTIVENESS AND EPIDEMIOLOGIC RESEARCH

Jay Desai MPH (HealthPartners Institute for Education and Research)
Pingsheng Wu PhD (Vanderbilt University)
Greg Nichols PhD (Kaiser Permanente Northwest)
Tracy Lieu MD, MPH (Harvard Pilgrim)
Patrick O’Connor MD, MPH (HealthPartners Institute for Education & Research)
SUPREME-DM DataLink

This project is supported by grant number R01HS019859 from the US Agency for Healthcare Research and Quality.
DEVELOPING A HEALTH DATABASE OR REGISTRY

• How good is the case definition?
  • Gold Standard?
  • Sensitivity, Specificity, Predictive Positive Value

• Building Confidence and Understanding
  • Understand contribution of EHD sources
  • Variation in EHD data sources

• Population Representativeness

• Case Retention
DIABETES CASE IDENTIFICATION

• At least 6 months of continuous enrollment
• Meet following criteria within a 24 month time frame

• Two elevated lab criteria on separate days, any combination
  • Fasting plasma glucose ≥ 126 mg/dL
  • Random plasma glucose ≥ 200 mg/dL
  • HbA1c ≥ 6.5%
  • OGTT ≥ 200 mg/dL (only 1 needed)
• ICD-9 250.xx, 357.2, 366.41, 362.01-07 (EMR or claim)
  • Two outpatient visit diagnosis on separate days
  • One inpatient visit diagnosis
• At least 1 diabetes drug pharmacy dispense/claim
  • Insulin
  • Oral agents (metformin & TZD must have other criteria met)
THE GOLD STANDARD CONUNDRUM

- Biological gold standard for diabetes identification:
  - Elevated blood glucose levels

- With good care management glucose levels may be below the threshold for diabetes diagnosis

- Remission due to substantial weight loss, bariatric surgery

- Comparative validity
  - Medical record documentation
  - Self-report
  - Claims-based diagnosis codes
VALIDITY CHARACTERISTICS

No good way to truly determine sensitivity, specificity, and predictive positive value
TAILORING DIABETES CASE DEFINITION TO SPECIFIC RESEARCH QUESTIONS

High Sensitivity
- **Maximize inclusion of potential cases**
- Observational studies
- Surveillance
- Population-based quality metrics
- CER
  - Attenuate results but may have broader generalizability

High Predictive Positive Value
- **Maximize identification of ‘true’ cases**
- Studies involving subject interventions
- Registries that guide clinical interactions
- Accountability tied to providers or systems
- Intervention studies
- CER
  - Stringent case identification
  - Potential selection bias
BUILDING CONFIDENCE IN CASE IDENTIFICATION

• Vary time frames using same case identification criteria
  • Shorter time frames: more confident but capture fewer cases
  • Longer time frames: less confident but may capture more cases
  • What is ideal, especially with no gold standard?

• Periodic recapture
DYNAMIC AND STATIC COHORTS

• Dynamic Cohort
  • Cumulative case identification over multiple years
  • Enter as new case or care system member
  • Leave due to death or disenrollment

• Static Cohort
  • Identification over defined time period and followed
  • Enter…none.

Figure 1. Dynamic & Static Diabetes Prevalence at one DataLink Health System (2008-9)
BUILDING CONFIDENCE IN CASE IDENTIFICATION

- Prioritizing case identification criteria
  - Assign probabilities of ‘true case’

- Prioritize data sources
  - More independent data sources identification… more confidence.
2008-9 DIABETES CASE IDENTIFICATION AT A DATALINK HEALTH SYSTEM
DIFFERENTIAL USE AND CHARACTERISTICS OF ELECTRONIC HEALTH DATA SOURCES

- There is wide variation across health systems regarding the primary source for case identification.

- There may be selection bias associated with specific data sources.

- This could affect case-mix and therefore results.
INITIAL CASE IDENTIFICATION:
2 ELEVATED BLOOD GLUCOSE LEVELS (LAB)
CER studies?
- Assess relative effectiveness of various treatments and systems of care in defined patient populations.

Uninsured
- No: if population defined based on insurance claims
- Probably: if population defined based on EMR

Units of analysis
- Patient, Provider, Clinic, Health System

Large multi-site registries more likely to provide representative ‘units of analysis’
- HIE potential to include smaller, less integrated systems
PERCENT RETENTION OF 2002 INCIDENT DIABETES COHORT AT ONE DATALINK HEALTH SYSTEM

The graph shows the percent retention of a 2002 incident diabetes cohort at One Datalink Health System over the years 2002 to 2010. The retention rate decreases over time, with a retention rate of 52.8% in 2002, 37.5% in 2004, and 9.7% in 2010. The graph also indicates an increase in death and disenrollment rates over the same period.
COMPARING SELECTED CHARACTERISTICS OF 2006 INCIDENT COHORT BY RETENTION STATUS THROUGH 2010:

<table>
<thead>
<tr>
<th>Baseline characteristics</th>
<th>Retained Cohort</th>
<th>Lost to disenrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>51%</td>
<td>50%</td>
</tr>
<tr>
<td>18-44 years</td>
<td>15%</td>
<td>27%</td>
</tr>
<tr>
<td>45-64 years</td>
<td>52%</td>
<td>55%</td>
</tr>
<tr>
<td>65+ years</td>
<td>32%</td>
<td>16%</td>
</tr>
<tr>
<td>Current smoker</td>
<td>15%</td>
<td>19%</td>
</tr>
<tr>
<td>BMI ≥ 30 kg/m²</td>
<td>60%</td>
<td>62%</td>
</tr>
<tr>
<td>HbA1c &lt; 8%</td>
<td>86%</td>
<td>78%</td>
</tr>
<tr>
<td>LDL-c &lt; 100 mg/dl</td>
<td>51%</td>
<td>43%</td>
</tr>
<tr>
<td>SBP &lt; 140 mmHg</td>
<td>80%</td>
<td>79%</td>
</tr>
<tr>
<td>DBP &lt; 90 mmHg</td>
<td>94%</td>
<td>90%</td>
</tr>
</tbody>
</table>
SUMMARY

- No realistic EHD gold standard for many conditions.

- When designing a registry,
  - Think multi-purpose
  - Maximize case capture so that a variety of case definitions can be derived depending on specific study needs.

- Consider developing several case definitions with different levels of confidence [sensitivity & PPV].
SUMMARY

- For CER studies we are interested in defined patient populations, providers, clinics, health systems...

- The greater the diversity of health systems participating in a disease registry the better...the more representative.

- EMR-derived registries may include uninsured and be most representative.
SUMMARY

- Cohorts developed using insurance claims have substantial attrition due to disenrollment over time.

- Important to include demographic and clinical characteristics of retained population compared to those loss-to-follow-up.

- CER studies requiring long follow-up to outcomes may be challenging if based on secondary use of EHD data.
  - Improve as health systems get regionally connected so patients can be tracked across systems (HIE’s)?
Reflections and Non-Technical Issues to Kick Off Discussion

Russell E. Glasgow, Ph.D.
Deputy Director, Implementation Science
Division Cancer Control and Population Sciences
National Cancer Institute

December 18th, 2012
EDM Forum Analytic Methods Webinar
Quality Control in the EHR for CER

- Use of frameworks, strategies and recommendations in these presentations is important and will clearly improve data quality

- The ‘devil is in the details’

- Focusing on ‘for what purpose’ is key - one site, all; sensitivity, specificity, etc.
  - all have advantages, limitations, and often unintended consequences
Lessons Learned

• **Kahn et al:** it is impossible and cost prohibitive to conduct comprehensive QI and cleaning on all data.

• **Desai et al:** the potential to create quality disease registries is a potential game changer.

• What is NOT in the EHR is also equally important...
  ...and these key domains are
  - Patient reported measures
  - Social and environmental factors (expose-ome)
• Supplement existing health utilization and biometrics with more complete and standardized data (e.g., patient demographics, literacy, numeracy, preferences, goals, health behaviors)

• Better and standardized GIS and environmental data in EHR datasets

• Generalizability: Potential for ‘rapid learning health care organizations’ providing real time quality data on hundreds of thousands of patients
Summary

• Recommendations are sound, helpful, important..... and may not go far enough

• Both disease registries and quality EHR data are necessary components of informed, progressive, rationale and possibly equitable health care systems

• EHR data are likely much better for some questions than others- they are not a panacea
Questions for Discussion

• What are the pros and cons of EHR data for:
  - recruitment for trials?
  - natural experiments?
  - post-market surveillance?
  - data for simulation models?
  - pragmatic studies?

• What types of additional data not routinely in the EHR are most important, and how can they be collected?
Submitting Questions

To submit a question:
1. Click in the Q&A box on the left side of your screen
2. Type your question into the dialog box and click the Send button
Continue the Discussion!
www.edm-forum.org

- Medical Care supplement
- Issue Briefs
- CER Project Profiles
- Descriptions of eHealth data initiatives for research & QI
- Event archives
- Wiki glossary
- Webinars

Sign up at edmforum@academyhealth.org
Thank You

*Please take a moment to fill out the brief evaluation which will appear in your browser.*