Using Large Databases in Observational Studies
ISPOR Student Chapter Webinar

Diana Brixner, RPh, PhD
Professor and Chair Pharmacotherapy
Executive Director, Pharmacotherapy Outcomes Research Center
L.S. Skaggs Pharmacy Institute
Director of Outcomes Personalized Health Care Program
University of Utah, Salt Lake City, Utah

September 24, 2013

Agenda

- The value of real world data (RWD)
- Databases for observational studies
- Case Studies
  - Metabolic Syndrome
  - Oncology

Efficacy / Effectiveness

- Efficacy
  - RCT
  - High internal validity
  - Limited generalizability

- Effectiveness
  - Observational studies
  - High external validity
  - Lack of Controls

Information is needed beyond RCTs...

Efficacy and safety in a small population with a restricted study protocol

Decision makers need real world information to make health care decisions for large populations within defined budgets

Changing Landscape of Information

- RCTs for registration
- For reimbursement
  - and (continuing) market access
  - Supplemental RCTs
  - Observational Studies
    - Patient Registries
    - Patient Reported Outcomes
    - Administrative Claims Databases
    - Electronic Health Records
  - Modeling

Data sources for Outcomes
Research Studies: Advantages and Disadvantages of each

RCTs: Gold Standard

- Eliminate bias and confounders through randomization
- Evidence level “A,” for assessing efficacy of therapeutic agents

However . . .
- Can take years to complete
- Expensive
- Protocol driven
- Usually vs. placebo or traditional therapies

Supplements to RCTs

- PROs
- Resource use and cost
  However . . .
  - Not powered for secondary endpoints
  - Short time frame
  - Protocol driven
**Pragmatic Trials: Advantages**

- Eliminate randomization issues and management of blinded data
  - cohort design
  - case control design
- Larger populations; greater power
- Less expensive
- Often against relevant comparators
- Can be used where RCT maybe impossible or unethical

**Pragmatic Trials: Disadvantages**

- Concerns about bias
  - Patient selection
- May increase positive treatment effects
- Large size
- High cost
- Lack of standardization

**Patient Registries**

- Collect data on current treatment patterns and medical resource utilization across a disease state
- Provides baseline data for cost of illness analysis; economic modeling, observational or retrospective study design
- Does not compare between therapies.
- No treatment optimization information

**Patient Reported Outcomes**

- QoL (How do you feel?)
- Productivity / Absenteeism
- Can be disease specific
- Patient preference / Patient satisfaction
- Activities of Daily Living
- Global Assessment Scale
- Visual Analog Scale
- Often complimentary to other types of data collection

---


Walter J. Hierholzer, Jr, M.D., AJM 1991; 91:3B-21S

**Retrospective Databases:**

- Primarily insurance claims for reimbursement
  - Federal government (DOA, VA)
  - State government (Medicaid)
  - Private healthcare (insurers (MCO, PBMs))
- Inexpensive and can be done quickly
- Large populations over long periods of time
- Captures real world consumption patterns


**Retrospective Databases:**

- Internal validity
  - Diagnostic information
  - Compliance variance
  - Confounders
  - Severity of illness
- External Validity
  - Generalizability (i.e. Medicaid database)
  - Plan design
  - Regional practice patterns
  - Cost differences


**EMR Data: The Advantages**

- Deep clinical information directly from the patient
- Used for clinical documentation
- Unique data elements: diagnoses, labs, vitals, BMI, adverse events
- All patients regardless of payment status
- Near real-time access
- Rapid analysis of millions of patients


**EMR Data: The Disadvantages**

- Limited universe of users
- Lack of continuity in care
- Unbalanced use of free text vs. discrete data elements
- Disparate users unconnected physically and intellectually
- Prescription orders not fills
Database Examples

- Administrative Claims Databases
  - TruVen (formerly MedStat)
  - Optum Insight (formerly Ingenix)
  - IMS Lifelink Pharmetrics Plus (formerly Pharmetrics)
- GE Centricity MQIC: Electronic Medical Record Database; now owned by Quintiles
- Inovalon (Quality driven database)
- Comprehensive Health Insights (a Humana company)

EMR Provider Members & Clinical Database

Contributing Members
Opt-In Participation

- 635 Member Institutions
- 33,000+ Providers
- 49 States & DC

QEMR Database

- 35M Patients
- 228M+ Office Visits
- 1B+ Documents

35+ Million Unique Patients and Growing
Unique Clinical Data Elements

Problems/Diagnoses
- Current, personal history, family history (ICD9 codes)
- Assessments: Change in status

Medications & Prescriptions, including OTCs
- Medications documented (dose, frequency, refill)
- Prescriptions written (frequency, refills)

Observations
- Vital signs & measures (Blood pressure, weight, BMI)

Chief Complaint/Reason for Visit
- Reason for office visit, Symptoms

Disease Counts

Examples
- Hypertension: 2,383,000
- Hyperlipidemia: 2,315,000
- Depression: 1,224,000
- Cardiovascular Disease: 1,048,000
- GERD: 1,029,000
- Diabetes: 981,000
- Asthma: 776,000
- Osteoarthritis: 625,000
- COPD: 369,000
- BPH: 335,000
- ADD/ADHD/HKD: 196,000
- Bipolar Disorder: 108,000
- Rheumatoid Arthritis: 88,000
- Alzheimers: 37,000
- Parkinsons: 36,000

Note: Actual disease counts will vary per query “filters”

Comprehensive Health Insights
A Humana Subsidiary: Integrated Data Sources

Many of these data originate in the Medicare space (approximately 60%) making this the nation’s largest private sector Medicare claims database.

Data are fully-integrated by the use of a unique patient identifier assigned to every member which remains constant regardless of benefit design, gap in coverage, or change in employer.
Commercial Data Distribution

Gender Distribution

- 15% Male
- 13% 0-14
- 16% 15-24
- 5% 25-34
- 18% 35-44
- 16% 45-54
- 5% 55-64
- 17% 65-74
- 16% 75+

Integrated Data

Database contains:
- Longitudinal, cross-matched, granular data
- Patient records can be linked across eligibility, medical, pharmacy and lab data

Data Fields include:
- Eligibility
- Encrypted Provider ID
- Pharmacy
- Medical
- Diagnostic
- Test
- Lab

Humans utilizes:
- NDC
- ICD-9
- HCPCS
- CPT-4

Claims can be segmented by:
- Inpatient
- Outpatient
- Emergency Room
- Physician Office

This unique resource is readily, and exclusively, available through CHI

Integrated Data: Beyond claims

New Data Types

- Database contains:
  - More medical and Rx benefit information
  - Wellness

- Data Fields include:
  - Census-based measures
  - Household information
  - Annual Max Amt

- Recently added variables, still being evaluated for completeness and level of information
  - All data share unique encrypted identifier

The MORE2 Registry®

Medical Outcomes Research for Effectiveness and Economics Registry (MORE2 Registry®)

- Inovalon’s MORE2 Registry® is a data warehouse that empowers informed insight into national and regional healthcare trends such as hospital utilization, medication adherence, chronic disease prevalence, and treatment effectiveness.

- This warehouse of healthcare data contains information derived from more than 6 billion medical events generated by over 85 million unique and de-identified individuals nationwide. Encompassing more than 2.6 billion member-months of data.

- The registry goes beyond claims data to include information about demographics, enrollment, diagnoses, procedures, pharmacy and laboratory results and represents a significant mix of commercial, Medicare, and Medicaid managed care plan members. The MORE2 Registry® is augmented regularly and is growing at a rate in excess of 40% annually.
MORE® Registry® Overview

Through the extensive service delivery relationships across its client base, Inovalon has assembled datasets that are among the most extensive and broadly representative health care datasets in the marketplace—containing more than 6 billion medical events and related data from more than 540,000 physicians, 220,000 clinical facilities, and 85 million unique members—expanding approximately 2.7% compounding monthly (18.6% annually).

MORE® Registry® contains:
- Patient Demographics
- Inpatient, Procedure, Discharge and Emergency Room Data
- Health Risk Assessment Data
- Practitioner Profile Data
- Claim Diagnostic Data
- Eligibility and Enrollment Data
- Encounter Data
- Pharmacy Data
- Pathology Data
- Procedural Data
- Durable Medical Equipment Data
- Payment Data
- Benefits Data
- Value-added Reference Data

MORE® Registry® Snapshots

Representative of the broad U.S. insured population, Inovalon’s datasets enable deep insight into a vast array of disease, co-morbidity, diagnostic, therapeutic, outcomes, and related financial inquiries.

Geographic Distribution

Gender Distribution

Age Distribution

MORE® Registry® Data*

Unique de-identified persons with enrollment: 86,002,257
Population with medical coverage: 78,585,538
Population with a medical claim: 68,821,580
Population with a medical and Rx claim: 51,634,914
Population with a medical and Rx claim: 49,779,057
Population with a medical and pharmacy claims: 20,091,431

Access to HEDIS and Star Quality Measures

Inovalon is the market and technological leader for Quality Measurement & Reporting, with responsibility for more than 800 NCQA data submissions annually. Using NCQA Certified Software, Inovalon offers full-service HEDIS reporting, including calculation of measures using administrative data and supplemental record review for hybrid reporting.

2013 Star Rating contains 37 Part C measures and 18 Part D measures

Inovalon provides clients access to inbuilt industry standard HEDIS and Star quality measures that can be run on the entirety or any custom subset of the MORE® Registry®
UU Patient Centered Retrospective Research Registries

- Collaboration with University of Utah:
  - Information Technology and Bioinformatics
  - Utah Population Database
  - HCI/UHOSP investigators and clinicians
  - Enterprise Data Warehouse (EDW)

- Contain longitudinal data on patient cohorts from 1995 to current including clinical, survival and charge data.

- Used to develop models to predict cost-effective outcomes of new pharmaceuticals and diagnostic tests

Data Sources

- EDW
  - Driver’s license UT
  - MPI#
  - Marriage certificate
  - Death certificate
  - Birth certificate
  - University of Utah population database

- UPDB
- University of Utah Health Sciences Patient Record Database
- Integrated the comprehensive electronic medical record (EMR) including all cancer cases across the University of Utah Health Sciences (UUHS), the Huntsman Cancer Institute and the Hospital (HCI/HCH).
- Clinical data included in electronic notes, lab orders and results, medication orders, and in certain cases death certificates data.
- If the patient died in UUHS-HCI/HCH, EDW has death certificate data, including date of death and cause(s) of death (as an ICD-9 code).
- The health care text data in the EMR of a subject originates from the physician created notes section in a subject’s EMR.

Academic Medical Center Patient Record Database

- EDW record
- Inpatients
- Labs
- MPI#
- Outpatients
- Procedures
- Others
- Rx
- UPDB record
- Death certificate
- Birth certificate
- Marriage certificate
- MPI #
- Driver’s license UT

Utah Population Database (UPDB)

- Where the death occurred external to UUHSC services
  - match subjects between EDW and UPDB by the master patient index (MPI).
  - The UPDB can be as much as 1 year behind in linking to death
  - The UPDB has been linked to the Utah Cancer Registry records from 1966 ~ 2011
  - Death certificates dating back to 1904.
- Grows annually through updates from the Utah Department of Health for births, deaths, marriages, and divorces, as well as records from the Utah Driver’s License Division, in addition to new information from the UCR.
- The UPDB has more than 14.9 million records and has 6.5 million unique individuals based upon person oriented record linking analysis.
- Under jurisdiction for the UU-HCI Resource for Genetic and Epidemiological Research (RGE).
Cancer Staging Methodology

• Included patients based on the criteria describing the different stages of cancer.
• Defined stage at diagnosis based on the electronic physician’s notes in the EHR.
  q based upon World Health Organization (WHO) criteria.
  q For the course of disease, the stages to determine progression or remission and response are identified in the charts and validated by a clinical expert

Charlson Comorbidity Index (CCI)

• The Charlson Comorbidity Index (CCI) encompasses 17 medical conditions (ICD-9) weighted 1–6.
• From the weighted conditions, a sum score can be tallied to yield the total comorbidity score.
• To account for increasing age, one point is added to the CCI score for each decade of life over the age of 50 (1 point for 51-60, 2 points for 61-70, 3 for 71-80 and 4 for >80.).
• Thus, possible CCI scores range from 0 to 34.

Case Studies

Linking Treatment and Disease in Obesity

Methods and Results from:
Objective & Design

- Evaluate the association of cardiometabolic risk factors with BMI as recorded in an EMR database
- Observational cross-sectional study of patients treated in primary care physician practice settings identified from the GE EMR between January 1996 and December 2005

Identification of Study Population From GE EMR

GE EMR Population ≥20 years old on last activity date
1996 to 2005
N=3,216,323

Patients with 2 years continuous EMR activity
N=1,529,639
47.56% of Population

 Patients with BMI ≥ 18 kg/m²
N=499,593

 Patients with BMI ≥ 18 kg/m² and zero cardiometabolic risk factors
N=209,363

 Patients with BMI ≥ 18 kg/m² and one or more cardiometabolic risk factors
N=289,960

Cardiometabolic Risk Factors Identified by Treatment, Diagnosis, or Both (N=209,633)

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Individual Risk Factors N=157,586</th>
<th>Multiple Risk Factors N=52,047</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESRD Triglycerides</td>
<td>3,667 1,339 2,580 252 12,162 2,650</td>
<td>2,950 30,343 8,31</td>
</tr>
<tr>
<td>Low HDL</td>
<td>1,201 0 1,201 0 4,866 0 4,866 0</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>141,852 2 44,922 128,925 5 50,742 2 14,993 47,444 11,69 5</td>
<td></td>
</tr>
<tr>
<td>Type 2 Diabetes</td>
<td>10,866 5,885 7,232 2,251 42,876 0 2,215 33,728 6,013</td>
<td></td>
</tr>
</tbody>
</table>

Patients with Individual and Multiple Cardiometabolic Risk Factors

Study Population: Patients with BMI ≥ 18 kg/m² and one or more cardiometabolic risk factors
N=209,363

Elevated Triglycerides
N=3,667

Low HDL
N=1,201

Hypertension
N=141,852

Type 2 Diabetes
N=10,866

Elevated Triglycerides
N=12,162

Low HDL
N=4,866

Hypertension
N=50,742

Type 2 Diabetes
N=42,876
Distribution of Cardiometabolic Risk Factors: Normal Weight (BMI 18-27) vs. Overweight (BMI >27)

What We Needed To Build The Model

Methods

GE EMR – Prevalence of Risk Factors by BMI

- Stratified by BMI group
  - 25-26.9, 27-29.9, 30-34.9, or ≥35
- Stratified by sex
- Risk factors (RF) identified by lab values, diagnosis codes, and drug use
- RFs included diabetes (DM), hypertension (HTN), and hyperlipidemia
- Stratified by # of RF and specific combinations
  - Zero RF
  - One RF – DM Only, HTN Only, DYS Only
  - Two RF – DM+HTN, DM+DYS, HTN+DYS
  - All three RF – DM+HTN+HLD
GE EMR – Risk Factor Identification Timeline

At least 365 days of activity prior to BMI date
BMI ≥ 25 in patients aged 20-64
Lab values identified ± 90 days
Prescription orders identified within 365 days
ICD-9 codes identified within 2 years

NHANES – Prevalence of CV Events by Risk Factors

- Identified from questionnaire responses in 2007-2008
- Included patients aged 20-64 with BMI ≥ 25
- Five possible CV event outcomes were included for each risk factor combination: no event, myocardial Infarction (MI), angina, heart failure (HF), or stroke
- Prevalence was determined for each of the CV RF combinations
- Due to data limitations:
  - BMI and sex groups were collapsed
  - CV events for patients with DM and DM+HLD were for those aged ≥ 20 years with BMI ≥ 25
  - Patients with multiple events may have been “double-counted”

MarketScan – CV Event and Risk Factor Costs

- Identified from Commercial Claims Database, does not include Medicaid or Medicare claims
- Inclusion criteria:
  - Age 20-64
  - ICD-9 for overweight, obesity (unspecified), or morbid obesity
  - ≥ 365 days of activity after first event or risk factor claim
  - ≥ 1 inpatient claim was required for CV event claim
- Costs extracted when CV event or risk factor was primary or secondary ICD-9 code
- Prescription drug treatments were not included

MarketScan – Cost Identification Timeline

At least 365 days of follow up activity and at least one inpatient claim for CV events
Claim for overweight, obesity (unspecified), or morbid obesity and age 20-64
Claim for CV event or risk factor
Decision Tree Model

- Developed using Microsoft® Office Excel® 2007 and TreeAge Pro© 2009
- Probabilities were determined by proportions of each group
  - For number of risk factors and CV events, one-way sensitivity analyses were performed using 95% CI around proportions
- Payoffs were defined as cost of CV events
  - Three scenarios devised to address uncertainty
    - 1st Scenario (Base Case) – ignored cost of risk factors
    - 2nd Scenario – assumed CV event costs already included cost of risk factors and “No Event” cost became the additive cost of risk factors in the pathway
    - 3rd Scenario – assumed cost of risk factors was not included in in CV event cost, so cost of risk factors was added to CV and No Event costs

---

Model Schematic

GE EMR – Distribution of Sex By BMI Group

NHANES – Prevalence of CV Events by Number of Risk Factors

*The “No CV Event” category comprises 97.8% of 0 RF, 86.9% of 1 RF, 73.3% of 2 RF, and 45.2% of 3 RF groups.
MarketScan – Cost of CV Events and Risk Factors

<table>
<thead>
<tr>
<th>CV event or Risk Factor</th>
<th>N</th>
<th>One-year Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td>1,204</td>
<td>$42,563</td>
</tr>
<tr>
<td>Angina</td>
<td>539</td>
<td>$5,002</td>
</tr>
<tr>
<td>HF</td>
<td>1,289</td>
<td>$40,517</td>
</tr>
<tr>
<td>Stroke</td>
<td>635</td>
<td>$17,054</td>
</tr>
<tr>
<td>HTN</td>
<td>54,23</td>
<td>$1,511</td>
</tr>
<tr>
<td>DM</td>
<td>24,64</td>
<td>$3,896</td>
</tr>
<tr>
<td>HLD</td>
<td>46,11</td>
<td>$1,136</td>
</tr>
</tbody>
</table>

Conclusions

- Average cost per patient increased as BMI increased
- Increased BMI → Increased # of risk factors
- Increased # risk factors → Increased CV events
- Increased CV events → Increased cost

Comparison of Cost Scenarios

<table>
<thead>
<tr>
<th>BMI Group</th>
<th>Ave. Cost/Patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>25-26.9</td>
<td>$1,122</td>
</tr>
<tr>
<td>27-29.9</td>
<td>$1,323</td>
</tr>
<tr>
<td>30-34.9</td>
<td>$1,749</td>
</tr>
<tr>
<td>≥35</td>
<td>$2,383</td>
</tr>
</tbody>
</table>

1. Assumed no cost incurred from RF (cost of no event was $0)
2. Assumed patients with no CV event incurred costs of RF and cost of CV event already included cost of RF (cost of no CV event was additive cost of RF in pathway)
3. Assumed patients with no CV event incurred costs of RF and RF cost was not included in cost of CV events (additive cost of RF in pathway were added to no CV event and CV event cost)
Limitations

- Looked at prevalence, did not take into account the timing of CV events
- Did not address patients with multiple events
- Did not include patients ≥65
- Did not assess changes in weight and how this effects change in risk and events
- Only non-fatal CV events addressed
- Did not include cost of prescription treatments
- Did not include other risk factors (e.g., smoking status, exercise) as inputs

Overall Conclusions/Final Thoughts

- Obesity is a major health care problem in the United States
- Modeling clinical and economic outcomes requires access to various different type of data sources
- Over time we anticipate these sources becoming more integrated
- Collaborative research between organizations can accelerate this progress

Summary

- Real World Data is becoming integrated into the decision and policy making process worldwide
- There are increasing options for real world data sources
- Using the appropriate data source to address value based research questions is critical in generating relevant evidence for decision making