Secondary Use of Data Resources for Research and Education

Chip Shaw, EdD, MPH
Executive Director Clinical Research Data Warehouse
What is Secondary Use

Data gathered and recorded previously for purposes other than the current project.

- Usually historical and already assembled
- Require no access to respondents or subjects
- Cost and time savings
Secondary Data Sources

Public data
- Data sets available to anyone.
- May require permission from the owner to access the data

Registry data
- A collection of information about individuals, usually focused around a specific diagnosis or condition.

Administrative data
- Data generated at each patient encounter for billing and claims purposes.

EMR data
- Data sets extracted from electronic medical record systems.
Public Data

Publicly available research data.

Measuring the Prevalence of Diagnosed Chronic Obstructive Pulmonary Disease in the United States Using Data From the 2012-2014 National Health Interview Survey

Trends in Obesity and Severe Obesity Prevalence in US Youth and Adults by Sex and Age, 2007-2008 to 2015-2016

Craig M. Hales, MD1; Cheryl D. Fryar, MSPH1; Margaret D. Carroll, MSPH1 et al
Registry Data

Clinical data used to improve the quality and safety of the care, compare the effectiveness of different treatments, and to monitor the safety of implanted devices.
Administrative Data

Claims data used to study health care delivery, benefits, harms, and costs.
Secondary Use of Clinical Data

- Billing and Cost Analysis
- Morbidity and mortality reporting
- Quality
  - HEDIS Reporting
  - Continuous quality improvement
- Patient safety reporting
  - Adverse event reporting
- Clinical Trials
  - Cohort identification
  - Post-marketing information on drugs and devices
- Clinical Research
- Health population statistics
- Public Health
  - Bio-surveillance
  - Disease reporting
  - Disease registries
- Education
  - Develop data sets for simulation
  - Student research projects
  - Data sets for biostatistics, epidemiology, etc.
A de-identified patient database from participating healthcare institutions with time-stamped and sequenced information on pharmacy, laboratory, admission and billing data from all patient care locations. As of 2018, the Health Facts database contained

- Over 65 million patients
- Patient information from 750 healthcare facilities across the United States
- Over 500 million encounters
- 4.7 billion laboratory results
- Detailed pharmacy, laboratory, billing and registration data as far back as 2000
- 684 million orders for nearly 4,500 drugs by name and brand.
Cerner Health Facts Data

Annals of Epidemiology
Volume 26, Issue 2, February 2016, Pages 151-154.e4

Brief communication
Temporal trends of esophageal disorders by age in the Cerner Health Facts database

Jessica L. Petrick PhD, MPH, Tuyet Nguyen MD, Michael B. Cook PhD
https://doi.org/10.1016/j.annepidem.2015.11.004

The American Journal of Cardiology
Volume 119, Issue 11, 1 June 2017, Pages 1809-1814

Effect of Transient and Sustained Acute Kidney Injury on Readmissions in Acute Decompensated Heart Failure

Benjamin J. Freda DO, Alexander B. Knee MS, Gregory L. Braden MD, Paul F. Visintainer PhD, Charulhas V Thakar MD
https://doi.org/10.1016/j.amjcard.2017.02.044
Cerner Health Facts Data Includes:

- Billed Charges
- Medications
- Labs and Microbiology
- Diagnosis and Procedure
- Facility Type
- Admission and Discharge
- Patient Demographics
- Clinical Assessment
Facility Type Data

Facilities Type data includes:
- US Census Region and Division (all represented)
- Bed size category
- Teaching status
- Urban/rural community setting
- Part of hospital system
- Acute care status
- Cardiac cath lab status
  - Full cath lab
  - Diagnostic cath lab only
- Statistically derived cost-to-charge ratio available
Admissions and Discharge Data

Admissions and Discharge data includes:
- Type of encounter
- Physician specialty
- Date/time of admission
- Admission type (e.g., elective)
- Admission source
- Date of discharge
- Discharge disposition
Patient Demographics Data

Patient Demographics data includes:

- Patient identifier
- Age at admission
- Race
- Gender
- Payer
Clinical Assessment Data

Clinical Assessment data includes:
- Height
- Weight
- Blood pressure
- Heart rate
- Pulse
- Temperature
- LOC / Glasgow Coma Scale
- BMI

- Respiratory rate
- Smoking status
- Alcohol use
- Pregnancy status
- Allergies
- Apgar
- Symptoms
- Pain assessment
Diagnosis and Procedure Data

Diagnosis and Procedure data includes:
- 1º and 2º diagnoses and procedure codes (ICD-9, ICD-10)
- Supplemental diagnoses and procedures (outside inpatient encounters)
- Discharge diagnoses, procedures
Laboratory Data

General Labs data includes:
- Procedure name
- Specimen source and type
- Date/time of order, collection, and completion
- Result and unit of measure
- Type of result
- Normal ranges, if applicable
- Ordering physician specialty
- Treatment setting
Medication Data

Medication data includes:
- Drug name
- Drug class
- Ordering physician specialty
- Treatment setting (on order)
- Dose, route of administration
- Order frequency
- Total quantity dispensed
- Start and discontinuation dates
Billing Data

Billed Charges data includes:
- Total billed charges
- Cost to charge ratios
## Cerner Health Facts Metrics

### Metrics Name

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Patients</td>
<td>69M</td>
</tr>
<tr>
<td>Inpatient Encounters</td>
<td>22M</td>
</tr>
<tr>
<td>Emergency Encounters</td>
<td>66M</td>
</tr>
<tr>
<td>Outpatient Encounters</td>
<td>431M</td>
</tr>
<tr>
<td>Lab results</td>
<td>4.7B</td>
</tr>
<tr>
<td>Microbiology</td>
<td>192M</td>
</tr>
<tr>
<td>Medication orders</td>
<td>684M</td>
</tr>
<tr>
<td>Total diagnoses</td>
<td>971M</td>
</tr>
<tr>
<td>Total procedures</td>
<td>119M</td>
</tr>
<tr>
<td>Clinical events</td>
<td>5.3B</td>
</tr>
</tbody>
</table>

### Metrics Name

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Systems</td>
<td>85</td>
</tr>
<tr>
<td>Total Facilities</td>
<td>750</td>
</tr>
<tr>
<td>Total Inpatient Facilities</td>
<td>388</td>
</tr>
<tr>
<td>Geography - Midwest</td>
<td>177</td>
</tr>
<tr>
<td>Geography - Northeast</td>
<td>235</td>
</tr>
<tr>
<td>Geography - South</td>
<td>263</td>
</tr>
<tr>
<td>Geography - West</td>
<td>174</td>
</tr>
</tbody>
</table>

### Metrics Name

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bed Size - ≤ 5</td>
<td>80</td>
</tr>
<tr>
<td>Bed Size - 6 to 99</td>
<td>136</td>
</tr>
<tr>
<td>Bed Size - 100 to 199</td>
<td>70</td>
</tr>
<tr>
<td>Bed Size - 200 to 299</td>
<td>48</td>
</tr>
<tr>
<td>Bed Size - 300 to 499</td>
<td>31</td>
</tr>
<tr>
<td>Bed Size - 500+</td>
<td>23</td>
</tr>
<tr>
<td>Type - Non-Teaching</td>
<td>279</td>
</tr>
<tr>
<td>Type - Teaching</td>
<td>108</td>
</tr>
</tbody>
</table>

### Number of Encounters for Key Conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>Encounters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>17.3M</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>7.1M</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>5.1M</td>
</tr>
<tr>
<td>Diabetes</td>
<td>10.1M</td>
</tr>
<tr>
<td>Asthma</td>
<td>3.3M</td>
</tr>
<tr>
<td>COPD</td>
<td>2.7M</td>
</tr>
</tbody>
</table>
## Cerner Health Facts De-Identification

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Any other unique identifying number, characteristic, or code</td>
<td>System-assigned numbers are provided to records. These numbers are mapped to Health Facts and are not identifiable.</td>
</tr>
<tr>
<td>Biometric identifiers, including finger and voice prints</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Certificate or license numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Date (except year) directly related to an individual (for example, date of birth (DOB), discharge date, date of death) and all ages over 89 and all elements of dates (including year) indicative of age, except that such ages and elements may be aggregated into a single category of age 90 or older.</td>
<td>Dates are shifted by a consistent value across a single patient record.</td>
</tr>
<tr>
<td>Device identifiers and serial numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Email addresses</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Fax numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Full face photographic images and any comparable images</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Geographic subdivisions smaller than the census division</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Health plan beneficiary numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>IP address numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Medical record numbers (MRNs)</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Names</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Social Security numbers (SSNs)</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Telephone numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Vehicle identifiers and serial numbers, including license plate numbers</td>
<td>Not extracted for Health Facts.</td>
</tr>
<tr>
<td>Web URLs</td>
<td>Not extracted for Health Facts.</td>
</tr>
</tbody>
</table>
Cerner Health Facts Data

LIMITATIONS

- Bias
  - “Sick people have more data”
- Selection Bias
- Confounding

- Data Quality
  - Misclassification
  - Missing data

ADDRESSING LIMITATIONS

- Validation against a Gold Standard
- Inclusion & exclusion criteria
- Case and Control Selection
Inclusion Exclusion Criteria-
Data request for a hypertension control study

- Which of the ICD-9-CM, ICD-10-CM codes for hypertension should be included?
- Should outpatient, inpatient, and emergency department encounters be included?
- Should automated blood pressure monitoring data be included?
- Should orders, medication reconciliation, and fulfillment data be used?
- How far back in time should data be evaluated?
- Should rolling year versus calendar year be used?
- Should deceased patients be included?
- Should perioperative data be included?
- Should hypertension in the gestational period be included?
Data Set Production - Type 2 Diabetes Case and Control selection

Northwestern University Type 2 diabetes mellitus (T2DM) algorithms for extracting both T2DM cases and T2DM controls from the electronic medical record (EMR).

dbGaP Study Accession: phs000888.v1.p1
Using Diabetes Inclusion Exclusion Phenotype

Conceptual Data Model of Cerner Health Facts

S. Piri, et al., A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a mode..., Decision Support Systems (2017), http://dx.doi.org/10.1016/j.dss.2017.05.012
## Data Set Processing

### Aggregation

<table>
<thead>
<tr>
<th>Hosp</th>
<th>Encounter</th>
<th>Labs / Meds</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>Lab Name 1 (Variant 1)</td>
<td>value</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td>Lab Name 1 (Variant 2)</td>
<td>value</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td></td>
<td>Lab Name 2</td>
<td>value</td>
</tr>
</tbody>
</table>

### Table Transposition

<table>
<thead>
<tr>
<th>Encounter</th>
<th>Labs / Meds</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Lab Name 1</td>
<td>Some Value</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>Lab Name 2</td>
<td>Some Value</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>Lab Name 3</td>
<td>Some Value</td>
</tr>
</tbody>
</table>
Identifying Appropriate Data

- Define the question you want to study
- Specify the population you want to study
- Specify what variables you want to include in your analysis
  - Exclusion and inclusion criteria
- What kind of data is most appropriate for your research
- Contact the Clinical Research Data Warehouse at CRDW@ttuhsc.edu and we can help to obtain the proper data and datasets
Clinical Research Data Warehouse

The Clinical Research Data Warehouse provides TTUHSC investigators with a single source for obtaining access to vast amounts of clinical data available in various systems at TTUHSC and in various databases throughout the United States. The data is available for educational and research tasks including preparatory research and data mining. The goal in creating the clinical research data warehouse is to accelerate clinical research that may potentially result in life-changing medical solutions for West Texas and beyond.

Cerner Health Facts  De-identified Patient Data  Public Use Data Sets