There is no Perfect Dataset—Optimizing your Research By Linking VA Data Sources

Nader N Massarweh, MD MPH

VA HSR&D Center for Innovations in Quality, Effectiveness and Safety, Michael E DeBakey VA Medical Center

Michael E DeBakey Department of Surgery, Baylor College of Medicine
No Financial Disclosures

*There are lots of wrong ways to do research, but no one right way.

These represent my opinions.
Objectives

• Structure of Corporate Data Warehouse data

• Considerations when performing observational research
  • Internal validity
  • External validity
  • Causal Inference
  • Important types of bias

• How/why to link various data sources

• Discuss 2 examples of why to do this
Corporate Data Warehouse
What is CDW

- CDW = Corporate Data Warehouse
- National central repository of VA data
  - Clinical (EHR) and administrative data
- Relational database
  - Organized in groups of formally described and structured data
  - Multiple domains with subdomains
    - Unique data elements in each
Creation of the CDW

1. Acquire Data
2. Populate Warehouse
3. Create Marts
4. Access Information

VHA\textsuperscript{c} – VHA clinical systems
VHA\textsuperscript{a} – VHA administrative and financial systems

HDR – Health Data Repository
NPCD – National Patient Care Database
DSS – Decision Support System
ADR – Administrative Data Repository
DoD – Dept. of Defense
CMS - Centers for Medicare & Medicaid Services
## CDW Structure

<table>
<thead>
<tr>
<th>Domain A</th>
<th>Domain B</th>
<th>Domain C</th>
<th>Domain i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Table 1</td>
<td>Table 1</td>
<td>Table 1</td>
</tr>
<tr>
<td>Table 2</td>
<td>Table 2</td>
<td>Table 2</td>
<td>Table 2</td>
</tr>
<tr>
<td>Table 3</td>
<td>Table 3</td>
<td>Table 3</td>
<td>Table 3</td>
</tr>
<tr>
<td>Table 4</td>
<td>Table 4</td>
<td>Table 4</td>
<td>Table 4</td>
</tr>
<tr>
<td>Table 5</td>
<td>Table 5</td>
<td>Table 5</td>
<td>Table 5</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Table i</td>
<td>Table i</td>
<td>Table i</td>
<td>Table i</td>
</tr>
</tbody>
</table>

...
## CDW DOMAIN CHECKLIST

### CDW Production
- Allergy
- Appointment
- Beneficiary Travel
- Consult
- CPRS Orders
- Dental
- Emergency Dept. Int. Software (EDIS)
- Health Factors
- Health Benefits Request
- Immunization
- Inpatient
- Lab Microbiology
- Lab Chem
- Mental Health Assessment
- Non-VA Meds
- Outpatient
- Patient
- Patient Associated
- Patient Enrollment
- Patient Insurance
- Patient Record Flag
- PCMM (Primary Care Management Module)
- Pharmacy BCMA (Bar Code Medication Administration)
- Pharmacy Outpatient
- Pharmacy Patient
- Purchased Care (formerly fee)
- Radiology
- Recall Reminders
- Reengineered Primary Care Management Module (RPCMM)
- SPatient
- Staff
- SStaff
- Surgery PRE, INTRA, and POST
- VistA Waitlist
- VistA Compensation & Pension
- Vital Signs
- Women's Health

### CDW RAW*
- Bill Claims (CBO)
- CAPRI Audit Trail table
- CliniComp
- Echocardiogram
- Equipment Inventory
- FBCS (Fee Basis Claim System)
- IFCAP (Integrated Funds Control, Accounting, and Procurement)
- Intravenous meds (IV)
- Oncology
- Prosthetics
- Pulmonary Function Test (PFT)
- Talent Management System (TMS)
- Unit Dose (Pharmacy)
- VACAA (Veterans Choice Program Eligibility)

*CDW Raw data is data that has been pulled directly from the VistA sites and the data has not been verified or had business rules applied. It may not be current and there is limited documentation for users. Requests for CDW Raw data require additional time for creating data extracts and it is more difficult for studies to use.

### Other Data
- Lung Cancer Screening Demonstration Cohort
- OMOP Common Data Model (CDW Production/Raw Source)
- PSSG Geocoded Files
- SAS Fee (Formerly HERC FEE)
- VINCI NLP Output
Considerations when Conducting Observational Research
Observational Research

• Non-experimental research designs
  • Cohort
  • Case-control
  • Cross-sectional
  • Ecological

• Data sources
  • Retrospectively collected data
    • Chart review
  • Prospectively collected data analyzed retrospectively
    • VASQIP, VA-CCR
  • No group assignment or interference from the investigator
Pros and Cons

• Pros
  • Statistical power
  • Evaluation of real-world practice
    • Compared to data from RCT’s
  • Cost-effective
  • Efficiency

• Cons
  • Statistical power
    • *With great power comes great responsibility*
  • Many sources of bias
  • Limitations of each data source
Important Considerations

- **External validity**
  - Extent to which study findings are generalizable to other practice settings and other patient populations

- **Internal validity**
  - Extent to which an observed association is due to a true relationship between the independent and dependent variables

- Establish association or correlation, *not causation*
  - Bradford Hill’s principles for establishing causal inference
    - Strength of association
    - Consistency
    - Specificity
    - Temporality
    - Biologic gradient (i.e.: dose-response)
    - Experimental evidence
    - Plausibility
    - Coherence
    - Analogy
Sources of Bias

• Achilles’ heel of observational research
  • Factors that oppose Hill’s principles
    • Selection bias
    • Confounding by indication
    • Misclassification/information bias
    • Immortal time bias
    • Survivor treatment bias
    • Lead-time bias

• Two main methods of addressing bias
  • Statistical methodology
  • Study design
Addressing Bias Through Study Design
Linking Data Sources

• How to do it
  • Use a common variable across two different data sources
  • In most VA datasets, typically either SSN or scSSN
  • Within CDW, there are also common ID variables within Tables of a domain that allow for data linkage
Example of How

Data Source A
- 1A
- 1B
- 1C
- 2D
- 5J
- 2A
- 6K
- 9P

Data Source B
- 1B
- 2A
- 2C
- 1D
- 5A
- 5B
- 5J
- 6K
Linking Data Sources

• How to do it
  • Use a common variable across two different data sources
  • In VA datasets, typically either SSN or scSSN
  • Within CDW, there are also common ID variables within Tables of a domain that allow for data linkage

• Why to do it
  • Can potentially help investigators overcome the limitations of each individual observational data source
Available types of VA data sources

• Administrative/Operational data
• Surveys
• Electronic health record
• Finance/Cost
• Quality improvement/measurement
• CMS
• Cancer registry
Example of Why

• **Hypotheses**
  • Postoperative glucose level measurements in patients with elevated preoperative HbA1c are more vigilantly monitored with a lower threshold of hyperglycemia for insulin treatment
  • Elevated early postoperative glucose levels would have greater use than preoperative HbA1c for predicting complications and increased readmissions.
Putting this study together

VASQIP → Post-op SSI

Primary cohort

1.) Post-op Readmission
2.) Post-op insulin
3.) Post-op labs

Supplemental Data

CDW

• VASQIP good for:
  • Identification of surgical patients
  • Identification of post-operative complications

• No information on:
  • Post-op meds
  • Serial post-op labs
  • Readmissions

• CDW good for:
  • Medication utilization
  • Serial labs
  • (Re-)Admissions

• Not good for robustly identifying post-operative complications
The Finished Product

Table 3. Complications by Preoperative Hemoglobin A1c

<table>
<thead>
<tr>
<th>No. (%)</th>
<th>Preoperative HbA1c, %</th>
<th>Peak 48-h Postoperative Glucose Level, mg/dl.</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal (&lt;5.7)</td>
<td>Prediabetic (5.7-6.4)</td>
<td>Diabetic (≥6.5)</td>
</tr>
<tr>
<td>Any postoperative complication</td>
<td>1346 (27.3)</td>
<td>2248 (26.1)</td>
<td>2683 (28.1)</td>
</tr>
<tr>
<td>Infectious complications</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wound infection</td>
<td>270 (5.5)</td>
<td>445 (5.2)</td>
<td>594 (6.2)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>76 (1.5)</td>
<td>125 (1.5)</td>
<td>104 (1.1)</td>
</tr>
<tr>
<td>UTI</td>
<td>89 (1.8)</td>
<td>173 (2.0)</td>
<td>220 (2.3)</td>
</tr>
<tr>
<td>Sepsis</td>
<td>95 (1.9)</td>
<td>171 (2.0)</td>
<td>199 (2.1)</td>
</tr>
<tr>
<td>Postdischarge outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Readmission within 14 d</td>
<td>494 (10.0)</td>
<td>867 (10.1)</td>
<td>1038 (10.9)</td>
</tr>
<tr>
<td>Readmission within 30 d</td>
<td>716 (14.5)</td>
<td>1222 (14.2)</td>
<td>1453 (15.2)</td>
</tr>
</tbody>
</table>

Abbreviations: HbA1c, hemoglobin A1c; UTI, urinary tract infection.

Figure 2. Postoperative Glucose Level Surveillance and Hyperglycemic Treatment

A. Mean no. of postoperative insulin checks by closest preoperative HbA1c

B. Proportion of patients receiving postoperative insulin by peak postoperative glucose

A. Mean number of postoperative insulin checks by closest preoperative hemoglobin A1c (HbA1c). Error bar represents mean 95% CI. B. Proportion of patients receiving postoperative insulin by peak postoperative glucose level. The gray shaded area represents glucose levels less than 5.7%
Infectious Postoperative Complications Decrease Long-term Survival in Patients Undergoing Curative Surgery for Colorectal Cancer

A Study of 12,075 Patients

Avo Artinyan, MD, MS,*† Sonia T. Orcutt, MD,† Daniel A. Anaya, MD,*†‡ Peter Richardson, PhD,‡§ G. John Chen, MD, PhD, MPH,‡§ and David H. Berger, MD, MHCM*†‡

**Hypotheses**

- Postoperative complications result in decreased long-term survival after radical resection for CRC after excluding early mortality
- Infectious complications, particularly severe postoperative infections, are more strongly associated with long-term survival than noninfectious complications
Putting this study together

**VA-CCR**

- **Primary cohort**
- **VA-CCR good for:**
  - Identification of cancer patients
  - Disease stage
- **No information on:**
  - Complications
  - Preoperative comorbid conditions

**CRC diagnosis**

**Supplemental Data**

1. Post-op complications
2. Preoperative comorbid conditions

**VASQIP**

- **VASQIP good for:**
  - Complications
  - Robust identification of preoperative comorbidities
- **Not good for cancer staging information**
The Finished Product

Stage 1

- No Complication (MS 50.0 months)
- Any Complication (MS 44.3 months)

$p=0.004$

Stage 2

- No Complication (MS 44.5 months)
- Any Complication (MS 38.2 months)

$p<0.001$

Stage 3

- No Complication (MS 41.9 months)
- Any Complication (MS 34.2 months)

$p<0.001$

Stage 4

- No Complication (MS 41.9 months)
- Any Complication (MS 34.2 months)

$p<0.001$
Closing points

• VA data can support a robust, informative, and impactful observational and comparative effectiveness research enterprise

• Linkage of VA data sources is a useful and available research technique to provide more robust observational research data
  • Complementary data from a variety of VA data sources can help to overcome limitations and sources of bias
Questions?

massarwe@bcm.edu