A Methodology for Quantitative Measurement of Quality and Comprehensiveness of a Research Data Repository (IDR Snapshot)

Vojtech Huser MD PhD
Acknowledgement

- Wisconsin Institute for discovery
  - Private branch: Morgridge Institute
- CTSA grant (NIH (NCRR) 1UL1RR025011)
  - U of Wisconsin-Madison + Marshfield Clinic
- Marshfield Clinic Research foundation
Agenda

- Introduction
- Methods
  - Design requirements
  - Measure components
- Results
  - Marshfield Clinic
- Discussion/Conclusion
Background

- Clinical Informatics
  - Subfield of Biomedical Informatics

Informatics = Technology + Information + People

Biomedical and Health Informatics

Bioinformatics (cellular and molecular)
Medical (Clinical) Informatics (person)
Public Health Informatics (population)

Research Informatics

Translational Informatics

Background

- Clinical Informatics
  - Subfield of Biomedical Informatics

Informatics = Technology + Information + People

Biomedical and Health Informatics

- Bioinformatics (cellular and molecular)
- Medical (Clinical) Informatics (person)
- Public Health Informatics (population)

Research Informatics

Translational Informatics


marshfieldclinic.org/birc
Introduction

- Advances
  - Federated research data warehouse
  - Future data growth
- Conduct of research
  - RCT vs. cohort and case-control studies
- IDR = Integrated Data Repository
- Choice of multiple institutions
- Quick evaluation of an IDR

Frueh FW. Back to the future: why randomized controlled trials cannot be the answer to pharmacogenomics and personalized medicine. Pharmacogenomics 2009;10(7):1077-81.
SHRINE: Query federation example

- Weber at al. (2009)
  - Application of Information Technology: The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories
    - *JAMIA 2009;16:624-63*

- Fusaro et al.
  - Electronic Medical Record Analysis Using Cloud Computing
    - *AMIA CRI summit, 2010*
Motivation

- **Motivation**
  - 10 institutions with IDR; budget allows only 3; How do you choose?
  - Tracking improvement within an organization

- **Assumptions (limitations)**
  - Goal: lifetime and complete EHR (+genetics, +behavior, +environment)
  - Improvement limits
  - General measures (rather then project specific)
  - Simple evaluation, pragmatic approach
  - Dichotomous approach to data sources (not profiling them)
  - Early stage (community interest solicitation)
Measure

- **Measures**
  - Qualitative description
  - Quantitative: Simple counts, Composite measure
- **Composite scores: APGAR, FICO, Google PageRank**
  - Score: A short-hand way to communicate about a complex concept
- **Data warehouse quality**
  - Size (1M vs. 10M), quality (billing data vs. structured EHR data)
- **Possible requirements**
  - Design issues
    - Completeness, Concision (simple), Measurability (objective, reliable), Independence
- **Our requirements**
  - Intuitive to interpret
  - Facilitates improvement
  - Fairness

Vojtech Huser, MD, PhD  
[linkedin.com/in/vojtechhuser](https://linkedin.com/in/vojtechhuser)
Data Structures (VDW)
Data Structures (i2b2)

Vojtech Huser, MD, PhD

linkedin.com/in/vojtechhuser
Data Structures (HealthFlow)

healthcareworkflow.wordpress.com
Data structures (data schemas)

- VDW
  - Several tables
  - Data domain specific structures
  - Optimized for queries, users

- i2b2, HealthFlow
  - One event table (+ attributes)
  - Generic data structures
  - Flexibility, extensibility
Event table (IDR Snapshot v1)

<table>
<thead>
<tr>
<th>Patient_id</th>
<th>Date</th>
<th>Event Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000001</td>
<td>3065-01-25</td>
<td>Lab</td>
</tr>
<tr>
<td>0000001</td>
<td>3065-01-25</td>
<td>Report</td>
</tr>
<tr>
<td>0000001</td>
<td>3065-01-25</td>
<td>Dx</td>
</tr>
<tr>
<td>0000001</td>
<td>3065-01-25</td>
<td>Dx</td>
</tr>
<tr>
<td>0000001</td>
<td>3065-01-25</td>
<td>Dx</td>
</tr>
<tr>
<td>0000002</td>
<td>3003-01-05</td>
<td>Lab</td>
</tr>
<tr>
<td>0000002</td>
<td>3003-01-05</td>
<td>Dx</td>
</tr>
<tr>
<td>0000002</td>
<td>3003-01-05</td>
<td>Proc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0000002</td>
<td>3035-05-06</td>
<td>Lab</td>
</tr>
<tr>
<td>0000002</td>
<td>3035-05-06</td>
<td>Dx</td>
</tr>
<tr>
<td>0000002</td>
<td>3035-05-06</td>
<td>Report</td>
</tr>
<tr>
<td>0000002</td>
<td>3041-04-16</td>
<td>Lab</td>
</tr>
<tr>
<td>0000002</td>
<td>3041-04-16</td>
<td>Rx</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Event-DOB) + '3000-01-01'

code.google.com/p/IDRSnapshot
Results

- Events table (ev_measure01)
- Generation took 2.5 hours
- Size: 43 GB (includes additional info)

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
<th>Institution 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General dimension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>all events</td>
<td>0.5 B</td>
</tr>
<tr>
<td>G2</td>
<td>unique patients</td>
<td>1.7 M</td>
</tr>
<tr>
<td><strong>Data Dimension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>count of patients with at least one diagnosis and one laboratory result</td>
<td>0.755 M</td>
</tr>
<tr>
<td>D2</td>
<td>same as D1 plus at least one clinical report</td>
<td>0.720 M</td>
</tr>
<tr>
<td>D3</td>
<td>same as D1 plus at least one prescription</td>
<td>0.429 M</td>
</tr>
<tr>
<td>D4</td>
<td>D2 and D3 combined</td>
<td>0.429 M</td>
</tr>
<tr>
<td><strong>Lifetime dimension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>count of patients with at least one pediatric event</td>
<td>0.261 M</td>
</tr>
<tr>
<td>L2</td>
<td>count of patients with at least one pediatric event and one adult event</td>
<td>0.134 M</td>
</tr>
</tbody>
</table>

code.google.com/p/IDRSnapshot
Code

IDR snapshot can evaluate an integrated data repository (medicine)

Filename ▼ | Summary + Labels ▼
---|---
IDR_Snapshot_v1_abstract.pdf | IDR Snapshot v1 article (abstract) pdf art
IDR-v1.sql | IDR Snapshot version 1 SQL code sql Fe
IDR-v1-Snapshot-description.htm | IDR Snapshot version 1 description sql d

[Link to code.google.com/p/IDRSnapshot]
### Code

**idrsnapshot**

IDR snapshot can evaluate an integrated data repository (medicine)

<table>
<thead>
<tr>
<th>Filename</th>
<th>Summary + Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDR_Snapshot_v1_abstract.pdf</td>
<td>IDR Snapshot v1 article (abstract) pdf art</td>
</tr>
<tr>
<td>IDR-v1.sql</td>
<td>IDR Snapshot version 1 SQL code sql Fe</td>
</tr>
<tr>
<td>IDR-v1-Snapshot-description.htm</td>
<td>IDR Snapshot version 1 description sql d</td>
</tr>
</tbody>
</table>

[Visit code.google.com/p/IDRSnapshot](http://code.google.com/p/IDRSnapshot)
(select distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000101) and res_id in --has Lab
(select distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000102) and res_id in --has textual report
(select distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000500)

Dx, Lab and Rx (level D3)
select count(distinct res_id) from a142ev_measure01 e where res_id in --has Dx
select count(distinct res_id) from a142ev_measure01 e where e.evtype_cd = 100000101) and res_id in --has Lab
select count(distinct res_id) from a142ev_measure01 e where e.evtype_cd = 100000102) and res_id in --has Rx prescription in EHR
(select distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000800)

Dx, Lab, Report and Rx (level D4)
select count(distinct res_id) from a142ev_measure01 e where res_id in --has Dx
(select distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000101) and res_id in --has Lab
select count(distinct res_id) from a142ev_measure01 e where e.evtype_cd = 100000102) and res_id in --has textual report
select count(distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000500) and res_id in --has Rx prescription in EHR
select count(distinct res_id from a142ev_measure01 e where e.evtype_cd = 100000800)

Vojtech Huser, MD, PhD
code.google.com/p/IDRSnapshot
# Cross institution comparison

<table>
<thead>
<tr>
<th></th>
<th>Institution 1</th>
<th>Institution 2</th>
<th>Institution 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General dimension</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>0.5 B</td>
<td>1.9 B</td>
<td>1.3 B</td>
</tr>
<tr>
<td>G2</td>
<td>1.7 M</td>
<td>3.2 M</td>
<td>0.8 M</td>
</tr>
<tr>
<td><strong>Data Dimension</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>0.7 M</td>
<td>2.1 M</td>
<td>0.4 M</td>
</tr>
<tr>
<td>D2 (report)</td>
<td>0.7 M</td>
<td>1.6 M</td>
<td>0</td>
</tr>
<tr>
<td>D3 (prescription)</td>
<td>0.4 M</td>
<td>0</td>
<td>0.7 M</td>
</tr>
<tr>
<td>D4</td>
<td>0.4 M</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Lifetime Dimension</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>0.3 M</td>
<td>1.1 M</td>
<td>0.3 M</td>
</tr>
<tr>
<td>L2</td>
<td>0.2 M</td>
<td>0.6 M</td>
<td>0.05 M</td>
</tr>
<tr>
<td>IDR Snapshot v2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

[Code: code.google.com/p/IDRSnapshot](code.google.com/p/IDRSnapshot)
Monitoring

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Institution 1</th>
<th>Institution 2</th>
<th>Institution 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>General dimension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>0.5 B</td>
<td>1.9 B</td>
<td>1.3 B</td>
</tr>
<tr>
<td>G2</td>
<td>1.7 M</td>
<td>3.2 M</td>
<td>0.8 M</td>
</tr>
<tr>
<td>Data Dimension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>0.7 M</td>
<td>2.1 M</td>
<td>0.4 M</td>
</tr>
<tr>
<td>D2 (report)</td>
<td>0.7 M</td>
<td>1.6 M</td>
<td>0</td>
</tr>
<tr>
<td>D3 (prescription)</td>
<td>0.4 M</td>
<td>0</td>
<td>0.7 M</td>
</tr>
<tr>
<td>D4</td>
<td>0.4 M</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lifetime Dimension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>0.3 M</td>
<td>1.1 M</td>
<td>0.3 M</td>
</tr>
<tr>
<td>L2</td>
<td>0.2 M</td>
<td>0.6 M</td>
<td>0.05 M</td>
</tr>
<tr>
<td>IDR Snapshot v2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

[link: code.google.com/p/IDRSnapshot]
What’s in it for me?

- Use it
  - Download our SQL code
  - Use it to evaluate your IDR

- Collaborate with us
  - Paper in writing (3 institutions involved so far)
  - IDR Snapshot version 2

IRB note: aggregate numbers

code.google.com/p/IDRSnapshot
Future work

- Learn from IDR of other institutions
- IDR Snapshot version 2
  - Event table structure
  - HL7 vMR (lite, SQL format)
- Delphi method to extend measures
  - Qualitative measures (transfers, insurance, PHR, family history)
  - Institutional context
  - Quantitative measures (general level, more complex “average patient” pattern)
- Structured data
  - Beyond claims data (e.g. nursing homes, decision support audit trails)
- Specialty snapshot
  - Prioritizing CDS interventions
  - CKD example (dialysis data)

code.google.com/p/IDRSnapshot
Thank you

- huser.vojtech@marshfieldclinic.org
- http://code.google.com/p/idrsnapshot

- Slides: http://www.linkedin.com/in/vojtechhuser
- http://marshfieldclinic.org/birc
- Questions?