Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

John F. Hurdle, Stephen C. Haroldsen, Andrew Hammer, Cindy Spigle, Alison M Fraser, Geraldine P Mineau, Samir J Courdy
Our objectives today

- Help you understand an important clinical research informatics problem: clinical research ascertainment
- Share with you approaches to this problem, with an emphasis on ours: leveraging public health sources (and pedigrees)
The outline...

- *Frame the problem*: finding study participants for studies large and small
- *Present* a brief background of prior work
- *Describe* our solution: exploiting public health, EDW, and pedigree data (UPDB and UPDBL)
- *Illustrate* how it works with an example from the paper
The problem

For a clinical/translational study to succeed, it needs an adequate sample (a cohort of participants)
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

The problem

For a clinical/translational study to succeed, it needs an adequate sample (a cohort of participants)

Recruit enough participants?

- Yes
- Stockholm

No

- Science weak or invalid
- Resources/time wasted
- IRB unhappy: no benefit means no risk is acceptable
Is recruitment really a problem?

University Utah IRB experience:

- Studies in 2007-2011 that were closed for any reason, including a successfully completed study.
Is recruitment really a problem?

University Utah IRB experience:

- Studies in 2007-2011 that were closed for any reason, including a successfully completed study.

- Note this is all types of human research studies (from small IIR studies up to NIH/Industrial RCTs)...important to consider study scale (more later)
Is recruitment really a problem?

University Utah IRB experience:

- Studies in 2007-2011 that were closed for any reason, including a successfully completed study.

- Note this is all types of human research studies (from small IIR studies up to NIH/Industrial RCTs)...important to consider study scale (more later)

- For the 726 studies for which we had adequate enrollment data, 182 (25.1%) met or exceeded their stated recruiting goals. The remaining 74.9% failed to meet their original recruitment expectations*

*This includes studies where sample size may have been originally overestimated (good) or so onerous to participants that the study was stopped (bad).
Is recruitment really a problem?

ClinicalTrials.gov might be illuminating

The Effectiveness of Lower Cyclosporine Doses for Psoriasis

This study has been terminated.
(Study terminated by the DSMB due to poor recruitment)

Sponsor:
University of Rochester

Collaborator:
National Institute of Arthritis and Musculoskeletal and Skin Diseases

Information provided by (Responsible Party):
Francisco Tausk, University of Rochester

Recruitment Information

Recruitment Status: Completed
Enrollment: 0
Completion Date: 
Primary Completion Date: 

ClinicalTrials.gov Identifier:
NCT00377325
First received: September 15, 2006
Last updated: June 20, 2012

NCT00000300
Why is that?

• Studies have specific inclusion/exclusion criteria
• Typical to oversample...dropouts
• Some populations overburdened with research requests (pool)
• Study procedures too onerous
Obvious way to find potential participants:

- Potential participant presents in clinic
- Practitioner word of mouth (referral)
- Flyers or other advertisements
- Rely on the Web (ClinicalTrials.gov, www.researchmatch.org...)
An obvious way for informaticists: use EHR/EDW data

- This is one goal of Meaningful Use
- At least as early as 1993, Tu et al.* were using EHR data to find HIV/AIDS trial participants
- Today
  - I2b2* architecture popular at academic med centers (has a cohort “cell”)
  - SHRINE* and eMERGE* also good examples

* See the paper for these references
Some problems with EHR/EDW ascertainment

• Big as they are, limited to one care system
  – Narrow or exotic eligibility criteria lead to small samples
  – May under-sample imp. geospatial locales
  – Tend to over-sample the overburdened

• Rarely have structured data for vital statistics, cancer cases, state-wide data
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

Utah’s solution: link multiple sources

- Public health data available in most states
Utah’ solution: link multiple sources

- Public health data available in most states
- Cancer registry data available in most states
Utah’ solution: link multiple sources

- Public health data available in most states
- Cancer registry data available in most states
- EDW data from our CTSA affiliates
Utah’ solution: link multiple sources

- Public health data available in most states
- Cancer registry data available in most states
- EDW data from our CTSA affiliates
- Unusual local resource: extensive pedigrees (most useful for familiality studies)
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

Utah Population Database (UPDB)

Facilitate High-Quality Health Related Research

35 Years of Research

~7 Million People

160 Approved Projects

http://www.huntsmancancer.org/updb
<table>
<thead>
<tr>
<th>Source</th>
<th>Linked to UPDB but Not Part of UPDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genealogies (Family History Library)</td>
<td>University of Utah Health Sciences Center EDW</td>
</tr>
<tr>
<td>Dept of Health Vital Records: Birth, Death, Marriages, Divorces</td>
<td>Intermountain Healthcare EDW</td>
</tr>
<tr>
<td>Dept of Health Inpatient Discharge, Ambulatory Surgery</td>
<td></td>
</tr>
<tr>
<td>Utah Cancer Registry, Cancer Registry of Idaho</td>
<td></td>
</tr>
<tr>
<td>Social Security Death Records</td>
<td></td>
</tr>
<tr>
<td>Utah Voter Registration</td>
<td></td>
</tr>
<tr>
<td>Driver License Division</td>
<td></td>
</tr>
</tbody>
</table>
Person-oriented records

- Multiple records for an individual are matched and a “person-oriented record” is created
Person-oriented records

- Multiple records for an individual are matched and a “person-oriented record” is created*
- Longitudinally-linked* data reflects many events for an individual.

* See the paper for these references
Person-oriented records

- Multiple records for an individual are matched and a “person-oriented record” is created*
- Longitudinally-linked* data reflects many events for an individual.
- Composite record of the most accurate and complete demographic information
Person-oriented records

- Multiple records for an individual are matched and a “person-oriented record” is created*
- Longitudinally-linked* data reflects many events for an individual.
- Composite record of the most accurate and complete demographic information
- About 6.8 million unique individuals created
Person-oriented records

- Multiple records for an individual are matched and a “person-oriented record” is created*
- Longitudinally-linked* data reflects many events for an individual.
- Composite record of the most accurate and complete demographic information
- About 6.8 million unique individuals created
- New records are linked to composite person-records*

* See the paper for these references
How does that help with ascertainment?

• Create a secure subset of UPDB that is directly available to students and researchers

• Provide a query tool that is pre-research and does not require IRB approval

• Return de-identified, summary data to the researchers: is there a cohort?
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

Diagnosis coverage

- Birth:
  - UPDB only until 1978
  - UPD/L-UPDB from 1978 to 2009

- Death:
  - UPDB only until 1966
  - UPD/L-UPDB from 1966 to Present

- UCR:
  - UPD/L-UPDB from 1996

- Inpatient:
  - UPD/L-UPDB from 1996
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

The UPDB Limited Query Tool provides summary data on a subset of data contained in a limited version of the Utah Population Database.

UPDB Limited was created at the Huntsman Cancer Institute by the Pedigree Population Resource and Research Informatics.

**Announcements**

- UPDB Limited data updated.
- UPDB Limited data synchronized to UPDB as of 17 October 2012

**Help Videos Available**

- Single Data Source Query
- Multiple Data Source Query
- Complex Query
- Cohort Based on Relatives
- Selection vs. Aggregation

**UPDBL Training Classes**

- Seminars on the Utah Population Database (UPDB)
A demo from the paper...
(click here if ReadyTalk not compatible to video...)

Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

![Image of UT Population Database with cohort builder interface](image)

<table>
<thead>
<tr>
<th>Diagnosis Criteria Applies to:</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utah Cancer Registry (ICDO)</td>
<td></td>
</tr>
<tr>
<td>Utah Inpatient Claims (ICD9)</td>
<td></td>
</tr>
<tr>
<td>Utah Death Certificates (ICD8, ICD9 &amp; ICD10)</td>
<td></td>
</tr>
</tbody>
</table>

**Diagnosis**
- Complication of device, implant or graft
- Coronary atherosclerosis and other heart disease
- ICD9
- OTHER FORMS OF CHRONIC ISCHEMIC HEART DISEASE
- Other and ill-defined heart disease

**Patients of ages**
- to
- at event

**Whose diagnosis**
- Primary
- Secondary
- Either
- Includes
- Does not include
- Any
- All

**Add Criteria**
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

1) Find persons matching your query criteria (below):
   Patients that meet all of the following criteria:
   - Is a living female
   - Had an inpatient diagnosis that includes any of the following ICD9 coded diagnoses: 414, 414.0, 414.00, 414.01, 414.03, 414.06, 414.2, 414.8, 414.9

2) Who have a:
   - First or Second Degree Relative or Cousin
   - First or Second Degree Relative
   - First Degree Relative
     - Sibling
     - Child
     - Parent

3) You may add additional demographic, birth, diagnosis and procedure criteria to this relative later
   * Due to the nature of queries involving relationships, certain kinds of queries may take 5-10 minutes to complete
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database
Security...

- UPDBL returns deidentified, summary data >5
- User agreement
- Final query constraints worked out with consultant from the Inter-University Consortium for Political and Social Research, UMichigan

**Health Districts**

The 29 Utah counties are collapsed into 12 Utah Health Districts.

Six are single counties:
- Davis
- Salt Lake
- Summit
- Tooele
- Utah
- Wasatch

Six are groups of counties:
- Bear River (Box Elder, Cache, Rich counties)
- Central Utah (Juab, Millard, Piute, Sevier, Wayne, Sanpete counties)
- Southeastern Utah (Carbon, Emery, Grand, San Juan counties)
- Southwest Utah (Garfield, Iron, Kane, Washington, Beaver counties)
- TriCounty (Daggett, Duchesne, Uintah counties)
- Weber-Morgan (Weber, Morgan)
UPDBL adds value to our CTSA federated tools
Identifying clinical/translational research cohorts: ascertainment via querying an integrated multi-source database

Closing points

• Other investigators have used SEER registries linked to other sources
• So local SEERs are a place to start
Now its time for questions...thanks for watching