The Ontario Cancer Data Linkage Project (‘cd-link’)
A new data release mechanism for cancer health services research in Ontario

Craig Earle, MD MSc FRCPC
Director, Health Services Research
Program for Cancer Care Ontario & the Ontario Institute for Cancer Research

Objective

• Describe a new data release mechanism for cancer HSR in Ontario
### Cancer data in Ontario

<table>
<thead>
<tr>
<th>Institute for Clinical Evaluative Sciences (ICES)</th>
<th>Cancer Care Ontario</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Ontario Cancer Registry</td>
<td>• Ontario Cancer Registry</td>
</tr>
<tr>
<td>• Vital Statistics</td>
<td>• Vital Statistics</td>
</tr>
<tr>
<td>• Cytobase, OBSP</td>
<td>• Cytobase, OBSP</td>
</tr>
<tr>
<td>• OHIP claims</td>
<td>• ColonCancerCheck</td>
</tr>
<tr>
<td>• Pharmacy/ODB data</td>
<td>• OCRIS (incl. staging)</td>
</tr>
<tr>
<td>• CIHI DAD, NACRS</td>
<td>• New Drug Funding Program</td>
</tr>
<tr>
<td>• Home Care database</td>
<td>• Radiation data</td>
</tr>
<tr>
<td>• Census/LHIN geographic data</td>
<td>• OPIS searchable records</td>
</tr>
<tr>
<td>• HOBIC</td>
<td>• Wait Time Information System</td>
</tr>
<tr>
<td>• Other registries</td>
<td>• ISAAC (patient-reported outcomes)</td>
</tr>
<tr>
<td>-- Diabetes, stroke, MI…</td>
<td></td>
</tr>
<tr>
<td>• Provider databases</td>
<td></td>
</tr>
<tr>
<td>-- Physicians, allied providers, hospitals, and other institutions</td>
<td></td>
</tr>
<tr>
<td>• Surveys</td>
<td></td>
</tr>
<tr>
<td>-- Canadian Community Health Survey, National Population Health Survey, Ontario Health survey…</td>
<td></td>
</tr>
</tbody>
</table>

Minutes from a meeting about fostering collaborative health services research in Ontario, 2005
cd-link goals

1. To make standing linkages of existing data sources available as an infrastructure resource for cancer health services researchers
2. To put de-identified linked data directly into the hands of researchers

Linked Data Sets:
SEER-Medicare data
(Surveillance, Epidemiology, & End Results)
- Tumor registry (diagnosis)
- Medicare claims (treatment)
- Death index (outcomes)
- Census data (ecological SES)
- Hospital files, AMA files (provider information)
- Area Resource File
- Capacity to link other data:
  - Sociological measures, specific cohorts, geocoding, accreditation
⇒ De-identified
Principles

Balance personal protection vs public good
1. Re-identification probability
2. Mitigating controls in place
3. Motive & capacity to re-identify
4. Extent of potential privacy invasion
   (Khaled El Emam)

Available data sets

• CIHI – Discharge abstract database (DAD)
• CIHI – National Ambulatory Care Reporting System (NACRS)
• Home Care Database
• Ontario Drug Benefit Claims (ODB)
• Ontario Health Insurance Plan Claims Database (OHIP)
• CytoBase (Cervical Screening)
• Ontario Breast Screening Program (OBSP)
• Ontario Cancer Registry Information System (OCRIS)
• Registered Persons Data Base (RPDB)
cd-link Procedures

cd-link procedures: Submit a proposal

- Rationale & objectives
- Data required and justification
- Planned analyses
- Expected products
- Describe data custodian resources
- Timeline
- List research staff
Review

1. Privacy
2. Feasibility
3. (Novelty)

Not:
- To approve the methods
  - Rely on peer review, data complexity, transparency
  - Prioritization

⇒ approved (4 weeks)

Data Use Agreement (DUA)

- Purpose limitation
- Confidentiality/re-identification/linkage/re-contact
- Security: password protection, encryption, public access, removable media
- Research ethics approval
- Limitation on onward transfer/sharing with 3rd parties
- Cell size suppression
- Pre-publication review
- Acknowledgement (not co-authorship or endorsement)
- Ownership of data
- Returning/destroying data
- Breach notification enforcement
- Responsibility to educate anyone touching the data
- Signed confidentiality agreement with anyone touching the data
- Threat of surprise audits

ICES Confidentiality Agreement
Data Request Form

• Define Cohort

• Datasets
  – Variables

HIPAA 18 restricted variables

1. Name
2. MRN
3. HIC
4. Geographic units < 20,000
5. Dates (except year)

• 6. phone #
• 7. fax #
• 8. e-mail address
• 9. SSN
• 10. license #
• 11. account #
• 12. VIN
• 13. device serial #
• 14. URL
• 15. IP address
• 16. Biometrics
• 17. photos
• 18. any other unique identifying code
De-identification

<table>
<thead>
<tr>
<th>Name</th>
<th>OHIP</th>
<th>DOB</th>
<th>Sex</th>
<th>Dx</th>
<th>DoDx</th>
<th>Adm dt</th>
<th>MD</th>
<th>Census med income</th>
<th>DoD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lynn Foma</td>
<td>123456</td>
<td>1/7/46</td>
<td>F</td>
<td>NHL</td>
<td>3/9/07</td>
<td>7/4/07</td>
<td>35429</td>
<td>61,435</td>
<td>9/9/07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>95135</td>
<td>1946</td>
<td>F</td>
<td>NHL</td>
<td>2007</td>
<td>117</td>
<td>5384</td>
<td>61,000</td>
<td>184</td>
</tr>
</tbody>
</table>

No longer PHI. Not human subjects research.

Privacy Analytics Risk Assessment Tool (PARAT)

Measures:

- **Prosecutor Risk** (Nosy Neighbor Risk): the probability of a single record being re-identified if the intruder has background information about a single individual

- **Marketer Risk**: The expected number of records that would be re-identified if the registry is matched with another database (exact matching)

- Uses a globally optimal k-anonymity algorithm to ensure that the probability is below a pre-defined threshold (the default is 0.2)
Example Risk Assessment

The percentage of records with a high probability of re-identification

Re-identification risk for the file compared to a threshold

The quasi-identifiers and their number of equivalence classes

Levels of data sensitivity

<table>
<thead>
<tr>
<th>Identifiable record-level data</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identified record-level data</td>
</tr>
<tr>
<td>Aggregate data</td>
</tr>
<tr>
<td>Previously published data</td>
</tr>
</tbody>
</table>
## Levels of data sensitivity

<table>
<thead>
<tr>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiable record-level data</td>
</tr>
<tr>
<td>De-identified record-level data</td>
</tr>
<tr>
<td>“Risk-Reduced De-identified Data”</td>
</tr>
<tr>
<td>Aggregate data</td>
</tr>
<tr>
<td>Previously published data</td>
</tr>
</tbody>
</table>

## Get primary, de-identified data

Within 6 weeks of receipt of
- DUA, confidentiality agreements
- data request form, and
- eventually, $ (cost-recovery)
After analysis

• Submit all manuscripts for pre-submission review
  – Privacy (>5/cell)
  – MOH & CCO review
• Destroy data when DUA term is up
  – Can submit proposals for other projects for the same data before DUA expires
  – Can get extensions on DUA as well

First release

• Occurred March 25, 2010
• A second request is in process

• Initially, CCO data only available to investigators at academic institutions in Ontario
  – Expected to expand
Conclusion

• Privacy and research are both public goods

• With the proper safeguards in place, both can be optimized

“Positive sum (win-win) paradigm”
Dr. Ann Cavoukian
(Ontario Information and Privacy Commissioner)

Future directions

• Provide analytic support
  – Web page
    [www.ices.on.ca](http://www.ices.on.ca) => 'About us' header => 'cd-link' on left sidebar
  – Data users workshops

• Expand to include other data sources
  – CCO/ICES data sharing agreement
  – Other provinces, countries
  – A model for other diseases

• Improve data quality (e.g., registry quality)
• Remote access
Acknowledgements

- David Henry
- Jan Hux
- Pam Slaughter
- Refik Saskin
- Hong Lu
- Karey Iron
- Derek Browne
- Nelson Chong
- Kathy Sykora
- Don DeBoer

- Terry Sullivan
- Kamini Milnes
- Pamela Spencer
- Alwin Kong

...for the cd-link planning committee

craig.earle@ices.on.ca