Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

NAACCR 2008-2009 Webinar Series
January 8, 2009

Q&A

Please submit all questions concerning webinar content through the Q&A panel

If the presentation is at full screen and you have a question, hit the escape key on your keyboard.
**Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination**

**Prizes!**

**Question of the Month!**
- The participant that submits the best question of the session will receive a fabulous Prize!
- Shannon and Jim will announce the winner at end of the session.

**Tip of the Month!**
- The participant that sends in the best tip related to the topic will win a spectacular prize!
- Shannon and Jim will announce the winner at the end of the session.
Today's Topics

• Confidentiality, Privacy and Disclosure of Cancer Data
  – Eric Holowaty, MD FRCPC MSc, Cancer Epidemiologist Cancer Care Ontario
• Central Cancer Registry Data Stewardship and Implications for Data Use
  – Jessica King, MPH, Biostatistician at the Centers for Disease Control & Prevention

Today's Topics

• Presenting Tabular Data Confidentiality Considerations
  – Glenn Copeland, MBA, Director Michigan Cancer Surveillance Program
• Microdata and Disclosure Risk
  – David Stinchcomb, MA, MS, Chief, Cancer Statistics Branch
• Empirical Assessment of Unique Records Using CINA Deluxe
  – Andy Lake, IMS

Confidentiality, Privacy and Disclosure of Cancer Data

General concepts and principles
### Concepts and principles
- Background
- Benefits and risks of data use and disclosure
- Definitions
- Legal and ethical framework
- Informed consent
- Framework for privacy protection

### Background
- Fiscal control and accountability
- PH surveillance
- Informatics
- Privacy concerns
Many Legitimate Beneficiaries of Access to Cancer Information

- Patients and Care Providers
- Public/Society
  - Research
  - PH surveillance
  - Accreditation
  - Fraud protection
- Commercial

Consequences of Overly-Restrictive Privacy and Security Measures

- avoidable clinical errors
- reduced future benefits from research
- reduced PH interventions
- less productivity; higher admin. costs
- erosion of public confidence in HC system

Hazards of Disclosure

- Improper use by authorized users
- By unauthorized users
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Definitions

- Privacy
- Confidentiality
- Data security
- Data stewardship

Definitions (cont’d)

- Identifiable/De-identified/Anonymized
- Disclosure risk
- Disclosure risk assessment
- Disclosure control
- Reasonableness standard

Disclosure risk scenarios

- Snoopy worker
- Sloppy worker
- Cancer cluster investigation
- Mapping report
- Case-Control interview study
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Ethical Framework for Cancer Surveillance

• Ethical Conduct

• Fundamental Ethical Framework
  • Respect for Autonomy
  • Non-maleficence
  • Beneficence
  • Equity or Justice

Legal Framework for Cancer Surveillance

• Constitutional law
• Disease-specific legislation
• Public health legislation
• FoIPoP legislation
• Health information legislation (e.g. HIPAA)
• Statistics legislation
• Common law

Informed Consent and Transparency

• Consent requirement is problematic for cancer surveillance and observational research
• Unclear scope of justification for excluding requirement for informed consent
• Patients and the public should be made aware of surveillance activities, partic. if consent is not required (Principle of Transparency)

Important role of REBs/IRBs in providing objective review of the merits and harms of disclosure
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Framework for Confidentiality Protection

- Determine users' requirements
- Understand the data
- Are there circumstances where disclosure is likely to occur?
- If so, would disclosure represent a breach of public trust, law or policy?
- If required select appropriate disclosure control methods
- Implement and disseminate


On Balancing Privacy and Research Access...

“The right to medical care should generally continue to include the responsibility to allow the information gained to be used for the benefit of others who develop a similar disease, or who are at risk of developing it.”

Sir Richard Doll

On the Importance of Research...

“Research shines a spotlight on ignorance. Most research projects cannot simply be turned off and on again, like a kitchen light. If the spotlights are turned off, many of them will stay off permanently.”

Prof. Michel Coleman
And now.....

Data Stewardship

Central Cancer Registry Data Stewardship and Implications for Data Users

NAACCR 2008-2009 Webinar Series

The Obligation

• To protect the privacy of cancer patients and their families and of providers...
• by ensuring the security of the data,
• while also ensuring that the data are available for appropriate analyses and purposes that further the battle against cancer.
Data Stewardship

- How your registry fulfills its obligation to protect privacy
- Starts with written policies and procedures, including:
  - Assignment of responsibilities
  - Rules and regulations regarding the handling and dissemination of confidential data

Important Parts of the Data Stewardship Policy

I. Data Steward
Person designated to:
- develop/maintain the policy to reflect applicable laws, agency regulations, etc., updating when necessary;
- assure that the registry complies with the Policy by monitoring procedures.

(Include a list of specific duties, e.g. tracking data requests/releases, maintaining Confidentiality Agreements, etc.)

Important Parts of the Data Stewardship Policy

II. Data Security/Confidentiality Protection
Should cover all aspects of maintaining security of the data, including:
- Building security – who goes in and out of the building, floor, or area where your data are in use and what precautions are taken to keep others from observing data.
- Program data security – what protects physical data (e.g. locked cabinets) and computerized data (e.g. passwords and timed screensavers)
- Staff security – training of staff, signed agreements (e.g. laptop security agreements), confidentiality agreements, signed annually to refresh security training
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

III. Decision Making Protocols
   - Who is responsible for:
     • making changes to policies and procedures
     • deciding how to handle data requests

   If a group, describe who its members are. Data Steward should preside.

IV. Description of Confidential Data
   • Directly identifies person or provider
   • Indirectly identifies person or provider
     - small numbers rules
     - mapping rules
     - population thresholds for suppression

V. Data Tracking System
   Method for keeping track of all confidential data released, to whom released, duration of access, etc.
   Track publications using data for verification purposes as well as to avoid duplication of papers
Example of Data Access and Use Tracking

<table>
<thead>
<tr>
<th>User</th>
<th>userid</th>
<th>email</th>
<th>Last IDA agreement</th>
<th>Access to SEER</th>
<th>Total # times accessed</th>
<th># times accessed last month</th>
<th># times accessed previous month</th>
<th>Date last accessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>jdoe1</td>
<td><a href="mailto:jdoe1@cdc.gov">jdoe1@cdc.gov</a></td>
<td>2007</td>
<td>yes</td>
<td>7</td>
<td>1</td>
<td></td>
<td>11/02/08</td>
</tr>
<tr>
<td>Jane</td>
<td>jdoe2</td>
<td><a href="mailto:jdoe2@cdo.gov">jdoe2@cdo.gov</a></td>
<td>2007</td>
<td>yes</td>
<td>18</td>
<td>3</td>
<td></td>
<td>12/04/08</td>
</tr>
</tbody>
</table>

Important Parts of the Data Stewardship Policy – Protecting it vs Using it

VI. Data Request Procedures
Forms to sign and submit
Maintain up to date info on who has access to what data and for how long
Decide who has right to give or deny access to data to users
Mandatory training for data users?

Important Parts of the Data Stewardship Policy – Protecting it vs Using it

VII. Data Release Procedures
Phased rollout to partners/individual customers such as universities, agencies, others/general public... level of data released to each may differ and data release/use agreements will also differ
Objectives

- Overview of Issues
- Types of Disclosures
- Strategies for Reducing Disclosure Risk
- References

Types of Disclosure

- Identity disclosure
  - Exposure of identifiers
    - Name, Address, SSN, other
- Attribute disclosure
  - Exposure of information about an individual
    - Cancer, type of cancer, severity
- Inferential disclosure
  - Exposure of information probably associated with an individual

Types of Disclosures

- Exact vs. Approximate Disclosure
  - Disclosing a fact or characteristic of an individual vs an individual's characteristic in a range, i.e. 45-54 years of age
- Probability-based vs. Certain Disclosure
  - Data indicate chance of having a characteristic where high certainty must be shielded
- Internal vs External Disclosure
  - Released data reveals confidential fact vs revealed via link to other data or knowledge
Considerations

- Numerator vs Denominator
  - Population Unique
- Disclosure Risk Tolerance Level
- Patient vs Facility Confidentiality
- Sensitivity/Risk
  - Demographic vs Clinical
- Public Use vs Restricted Access
  - Minimum necessary

**Attribute Disclosure - Certainty**

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Am. Indian</th>
<th>Asian/PI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>268</td>
<td>189</td>
<td>267</td>
<td>186</td>
</tr>
<tr>
<td>Under 20</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20-29</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>30-39</td>
<td>9</td>
<td>2</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>40-49</td>
<td>14</td>
<td>7</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>50-59</td>
<td>37</td>
<td>30</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>60-69</td>
<td>79</td>
<td>57</td>
<td>79</td>
<td>56</td>
</tr>
<tr>
<td>70 and over</td>
<td>124</td>
<td>92</td>
<td>124</td>
<td>90</td>
</tr>
</tbody>
</table>

**Attribute Disclosure - External**

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
<th>Am. Indian</th>
<th>Asian/PI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>268</td>
<td>189</td>
<td>267</td>
<td>186</td>
</tr>
<tr>
<td>Under 20</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>20-29</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>30-39</td>
<td>9</td>
<td>2</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>40-49</td>
<td>14</td>
<td>7</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>50-59</td>
<td>37</td>
<td>30</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>60-69</td>
<td>79</td>
<td>57</td>
<td>79</td>
<td>56</td>
</tr>
<tr>
<td>70 and over</td>
<td>124</td>
<td>92</td>
<td>124</td>
<td>90</td>
</tr>
</tbody>
</table>
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Attribute Disclosure - Inferential

Lung Cancer Cases by Age, Race and Stage at Diagnosis
Lakeside County Residents, 2001-2005

<table>
<thead>
<tr>
<th>Total</th>
<th>White</th>
<th>Black</th>
<th>Am. Indian</th>
<th>Asian/PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>Late</td>
<td>Early</td>
<td>Late</td>
<td>Early</td>
</tr>
<tr>
<td>Total</td>
<td>268</td>
<td>189</td>
<td>196</td>
<td>3</td>
</tr>
<tr>
<td>under 20</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20-29</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30-39</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>40-49</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>50-59</td>
<td>37</td>
<td>30</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>60-69</td>
<td>79</td>
<td>57</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>70 and over</td>
<td>124</td>
<td>92</td>
<td>90</td>
<td>0</td>
</tr>
</tbody>
</table>

Disclosure Limitation Methods

• Population size
  – Census (100,000)
  – HIPAA (20,000)
• Cell Suppression
• Geographic Scale
• Collapse Rows and Columns

Threshold Rule

• Restrict tabular data
  – Limit cell frequencies to specified minimum
• For cells below minimum
  – Combine rows and/or columns
  – Suppress cell sizes below minimum and suppress complimentary cells
  – Rounding
  – Controlled tabular adjustment
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Disclosure Prevention Guidelines - NCHS

- Row or column total must not equal single cell
- Row, column or cell must not be less than 5
- Can not derive above from other tables

Censored Table - Threshold Rule 5+
Combine Rows and Columns

Lung Cancer Cases by Age and Stage at Diagnosis
Lakeside County Residents, 2001-2005

<table>
<thead>
<tr>
<th>Age at Diagnosis</th>
<th>Early</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>268</td>
<td>189</td>
</tr>
<tr>
<td>under 40</td>
<td>28</td>
<td>10</td>
</tr>
<tr>
<td>50-59</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>60-69</td>
<td>79</td>
<td>57</td>
</tr>
<tr>
<td>70 and over</td>
<td>124</td>
<td>93</td>
</tr>
</tbody>
</table>

Frequency of Incident Cancers in Five Ontario Counties 2002

<table>
<thead>
<tr>
<th>Cancer Site</th>
<th>County 1</th>
<th>County 2</th>
<th>County 3</th>
<th>County 4</th>
<th>County 5</th>
<th>All 5 Counties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Cavity and Pharynx</td>
<td>61</td>
<td>51</td>
<td>69</td>
<td>49</td>
<td>44</td>
<td>289</td>
</tr>
<tr>
<td>Esophagus</td>
<td>19</td>
<td>11</td>
<td>19</td>
<td>54</td>
<td>54</td>
<td>164</td>
</tr>
<tr>
<td>Stomach</td>
<td>32</td>
<td>7</td>
<td>41</td>
<td>117</td>
<td>117</td>
<td>397</td>
</tr>
<tr>
<td>Colon and Rectum</td>
<td>301</td>
<td>85</td>
<td>288</td>
<td>83</td>
<td>83</td>
<td>835</td>
</tr>
<tr>
<td>Pancreas</td>
<td>36</td>
<td>17</td>
<td>43</td>
<td>15</td>
<td>15</td>
<td>120</td>
</tr>
<tr>
<td>Lung and Bronchus</td>
<td>263</td>
<td>120</td>
<td>145</td>
<td>76</td>
<td>76</td>
<td>545</td>
</tr>
<tr>
<td>Melanoma of the Skin</td>
<td>50</td>
<td>24</td>
<td>26</td>
<td>26</td>
<td>26</td>
<td>108</td>
</tr>
<tr>
<td>Breast</td>
<td>464</td>
<td>109</td>
<td>350</td>
<td>75</td>
<td>75</td>
<td>1075</td>
</tr>
<tr>
<td>Lymphoid Tissues, NOS</td>
<td>74</td>
<td>10</td>
<td>64</td>
<td>15</td>
<td>15</td>
<td>164</td>
</tr>
<tr>
<td>Stomach</td>
<td>51</td>
<td>7</td>
<td>40</td>
<td>15</td>
<td>15</td>
<td>122</td>
</tr>
<tr>
<td>Testis</td>
<td>1916</td>
<td>1</td>
<td>44</td>
<td>4</td>
<td>4</td>
<td>1995</td>
</tr>
<tr>
<td>Urinary Bladder</td>
<td>85</td>
<td>12</td>
<td>13</td>
<td>68</td>
<td>19</td>
<td>197</td>
</tr>
<tr>
<td>Kidney and Renal Pelvis</td>
<td>61</td>
<td>11</td>
<td>18</td>
<td>40</td>
<td>10</td>
<td>140</td>
</tr>
<tr>
<td>Brain</td>
<td>48</td>
<td>11</td>
<td>26</td>
<td>10</td>
<td>10</td>
<td>105</td>
</tr>
<tr>
<td>Thyroid</td>
<td>126</td>
<td>35</td>
<td>35</td>
<td>11</td>
<td>11</td>
<td>282</td>
</tr>
<tr>
<td>Hodgkin Lymphoma</td>
<td>34</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
<td>86</td>
<td>34</td>
<td>34</td>
<td>90</td>
<td>25</td>
<td>255</td>
</tr>
<tr>
<td>Leukemia</td>
<td>54</td>
<td>10</td>
<td>70</td>
<td>14</td>
<td>14</td>
<td>188</td>
</tr>
<tr>
<td>All Others</td>
<td>26</td>
<td>6</td>
<td>24</td>
<td>9</td>
<td>9</td>
<td>68</td>
</tr>
<tr>
<td>All Sites</td>
<td>2584</td>
<td>471</td>
<td>1948</td>
<td>523</td>
<td>523</td>
<td>5870</td>
</tr>
</tbody>
</table>
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Frequency of Incident Cancers in Five Ontario Counties 2002

<table>
<thead>
<tr>
<th>Cancer Site</th>
<th>County 1</th>
<th>County 2</th>
<th>County 3</th>
<th>County 4</th>
<th>County 5</th>
<th>All 5 Counties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Cavity and Pharynx</td>
<td>61</td>
<td>32</td>
<td>30</td>
<td>28</td>
<td>21</td>
<td>156</td>
</tr>
<tr>
<td>Esophagus</td>
<td>19</td>
<td>9</td>
<td>11</td>
<td>19</td>
<td>58</td>
<td>117</td>
</tr>
<tr>
<td>Stomach</td>
<td>30</td>
<td>7</td>
<td>41</td>
<td>9</td>
<td>58</td>
<td>157</td>
</tr>
<tr>
<td>Colon and Rectum</td>
<td>301</td>
<td>85</td>
<td>266</td>
<td>53</td>
<td>807</td>
<td>207</td>
</tr>
<tr>
<td>Pancreas</td>
<td>36</td>
<td>11</td>
<td>43</td>
<td>15</td>
<td>120</td>
<td>186</td>
</tr>
<tr>
<td>Lung and Bronchus</td>
<td>840</td>
<td>156</td>
<td>242</td>
<td>84</td>
<td>157</td>
<td>840</td>
</tr>
<tr>
<td>Melanoma of the Skin</td>
<td>58</td>
<td>16</td>
<td>15</td>
<td>19</td>
<td>58</td>
<td>157</td>
</tr>
<tr>
<td>Breast</td>
<td>1026</td>
<td>156</td>
<td>256</td>
<td>102</td>
<td>526</td>
<td>2026</td>
</tr>
<tr>
<td>Liver</td>
<td>160</td>
<td>25</td>
<td>44</td>
<td>12</td>
<td>160</td>
<td>244</td>
</tr>
<tr>
<td>Lymphatic tissue, NOS</td>
<td>76</td>
<td>11</td>
<td>19</td>
<td>58</td>
<td>120</td>
<td>157</td>
</tr>
<tr>
<td>Uterus</td>
<td>280</td>
<td>117</td>
<td>242</td>
<td>58</td>
<td>120</td>
<td>580</td>
</tr>
<tr>
<td>Uterus, NOS</td>
<td>90</td>
<td>16</td>
<td>15</td>
<td>120</td>
<td>157</td>
<td>253</td>
</tr>
<tr>
<td>Kidney and Renal Pelvis</td>
<td>52</td>
<td>11</td>
<td>41</td>
<td>15</td>
<td>120</td>
<td>186</td>
</tr>
<tr>
<td>Brain</td>
<td>48</td>
<td>11</td>
<td>29</td>
<td>10</td>
<td>102</td>
<td>152</td>
</tr>
<tr>
<td>Thyroid</td>
<td>126</td>
<td>9</td>
<td>36</td>
<td>11</td>
<td>120</td>
<td>202</td>
</tr>
<tr>
<td>Hodgkin Lymphoma</td>
<td>34</td>
<td>7</td>
<td>7</td>
<td>120</td>
<td>157</td>
<td>157</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
<td>88</td>
<td>34</td>
<td>90</td>
<td>120</td>
<td>157</td>
<td>253</td>
</tr>
<tr>
<td>Leukemia</td>
<td>54</td>
<td>21</td>
<td>70</td>
<td>14</td>
<td>186</td>
<td>253</td>
</tr>
<tr>
<td>All Others</td>
<td>26</td>
<td>8</td>
<td>24</td>
<td>9</td>
<td>84</td>
<td>157</td>
</tr>
<tr>
<td>All Sites</td>
<td>2584</td>
<td>671</td>
<td>1046</td>
<td>157</td>
<td>5870</td>
<td>8070</td>
</tr>
</tbody>
</table>

Frequency of Incident Cancers in Five Ontario Counties

<table>
<thead>
<tr>
<th>Adequately Suppressed Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Site</td>
</tr>
<tr>
<td>Oral Cavity and Pharynx</td>
</tr>
<tr>
<td>Esophagus</td>
</tr>
<tr>
<td>Stomach</td>
</tr>
<tr>
<td>Colon and Rectum</td>
</tr>
<tr>
<td>Pancreas</td>
</tr>
<tr>
<td>Lung and Bronchus</td>
</tr>
<tr>
<td>Melanoma of the Skin</td>
</tr>
<tr>
<td>Breast</td>
</tr>
<tr>
<td>Liver</td>
</tr>
<tr>
<td>Lymphatic tissue, NOS</td>
</tr>
<tr>
<td>Uterus</td>
</tr>
<tr>
<td>Uterus, NOS</td>
</tr>
<tr>
<td>Kidney and Renal Pelvis</td>
</tr>
<tr>
<td>Brain</td>
</tr>
<tr>
<td>Thyroid</td>
</tr>
<tr>
<td>Hodgkin Lymphoma</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
</tr>
<tr>
<td>Leukemia</td>
</tr>
<tr>
<td>All Others</td>
</tr>
<tr>
<td>All Sites</td>
</tr>
</tbody>
</table>

Frequency of Incident Cancers in Five Ontario Counties

<table>
<thead>
<tr>
<th>Primary and Complimentary Suppression Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer Site</td>
</tr>
<tr>
<td>Oral Cavity and Pharynx</td>
</tr>
<tr>
<td>Esophagus</td>
</tr>
<tr>
<td>Stomach</td>
</tr>
<tr>
<td>Colon and Rectum</td>
</tr>
<tr>
<td>Pancreas</td>
</tr>
<tr>
<td>Lung and Bronchus</td>
</tr>
<tr>
<td>Melanoma of the Skin</td>
</tr>
<tr>
<td>Breast</td>
</tr>
<tr>
<td>Liver</td>
</tr>
<tr>
<td>Lymphatic tissue, NOS</td>
</tr>
<tr>
<td>Uterus</td>
</tr>
<tr>
<td>Uterus, NOS</td>
</tr>
<tr>
<td>Kidney and Renal Pelvis</td>
</tr>
<tr>
<td>Brain</td>
</tr>
<tr>
<td>Thyroid</td>
</tr>
<tr>
<td>Hodgkin Lymphoma</td>
</tr>
<tr>
<td>Non-Hodgkin Lymphoma</td>
</tr>
<tr>
<td>Leukemia</td>
</tr>
<tr>
<td>All Others</td>
</tr>
<tr>
<td>All Sites</td>
</tr>
</tbody>
</table>
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Suppression Software

- NCHS – Data Protection Utility (DPUT)
  The US National Center for Health Statistics has sponsored the development of disclosure limitation software for two-way tables by OptTek Systems, Inc.
  jgonzalez@cdc.gov

- CASC -- t - Argus
  The Centers for Excellence – Statistical Disclosure Control project has developed software tools that work to protect tabular data.
  http://neon.vb.cbs.nl/CASC/

DPUT Software Functions

- cell suppression
- controlled rounding
- unbiased controlled rounding
- controlled rounding - subtotal constraints
- synthetic substitution
  (controlled tabular adjustment)

Cell Suppression

- Remove the value of disclosure cells as well as a sufficient number of neighboring cells so the disclosure cells can’t be deduced
- Totals left unchanged
- Requires specifying a disclosure rule, i.e. a non-zero cell is a disclosure cell if it falls below a threshold (base) e.g. n=5
- Provide sufficient disclosure protection while minimizing the amount of information lost due
Controlled Rounding

- Rounding table frequencies using the threshold value as the base in such a way that the resulting frequencies add to the total

- Uses linear programming to restrict results to row and column totals
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Synthetic Substitution (Controlled Tabular Adjustment)

- Developed by Dandekar and Cox (2002) as an alternative to complementary cell suppression.
- Uses a threshold rule(s) to determine how cells can be modified.
- All sensitive cells, \( a_{ij} \leq \text{"Base"} \) (B) are set = 0 or \( (a_{ij} + B) \).
- All other cells can be adjusted such that:
  \( (a_{ij} - B) < \text{new value} < (a_{ij} + B) \).
- A "noise" factor is used to randomize the results of synthetic substitution.
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

**Frequency vs Magnitude Data**

- Frequency data
  - The number of units in a cell
- Magnitude data
  - Aggregate quantity of interest
  - Measures something other than membership
- Number of lung cancer cases
  - Lung cancer incidence – frequency
  - Patient load, market share - magnitude
Magnitude Suppression Rules

- $(n,k)$ rule
  - $n =$ minimum cell size
  - $k =$ no respondent $>$ XX percent
- $p$ percent rule
- $pq$ rule

Federal Agency Practices

<table>
<thead>
<tr>
<th>Agency</th>
<th>Procedures to Protect Tabular Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census</td>
<td>Data Swapping</td>
</tr>
<tr>
<td></td>
<td>Query Rules</td>
</tr>
<tr>
<td></td>
<td>Threshold Rule 4+</td>
</tr>
<tr>
<td></td>
<td>$(n,k) (1/6)$</td>
</tr>
<tr>
<td>National Center for Health Statistics</td>
<td>Threshold Rule 4+</td>
</tr>
<tr>
<td>Dept of Education</td>
<td>Data Swapping</td>
</tr>
<tr>
<td></td>
<td>Data Coarsening</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Agency for Healthcare Research and Quality</td>
<td>Threshold Rule 4+</td>
</tr>
<tr>
<td>Social Security Administration</td>
<td>Threshold Rule 5+</td>
</tr>
<tr>
<td></td>
<td>Marginals, 3+ cells</td>
</tr>
<tr>
<td>Internal Revenue Service</td>
<td>Threshold Rule 3+</td>
</tr>
<tr>
<td>National Science Foundation</td>
<td>Varies by risk</td>
</tr>
</tbody>
</table>

CDAC Checklist

Statistical Policy Office - OMB

Purpose
To guide reviewing disclosure-limited data products

Reflects current standards of Census and NCHS

Section 4 – Tabular data
Section 5 – Magnitude data
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

CDAC Checklist

http://www.fcsm.gov/committees/cdac/

- Dimensions of table
- Geographic level of detail
- Sample or census
- External sources
- Values suppressed?
- Any cells contain domain
- Secondary suppression process
- Audit of suppressed table
- Noise factor
- Additional methods used
- Coordination of disclosure

References

- Annotated Bibliography on Confidentiality Protection in Data Release
  www.naaccr.org/confidentiality/index.asp

  http://www.fcsm.gov/working-papers/spwp22.html

- Checklist on Disclosure Potential of Proposed Data Releases, July 1999
  http://www.fcsm.gov/committees/cdac/

References

- Gonzalez JF, Cox L; Software for Tabular Data Protection; Statist. Med. 2005;24:659-669
- Computational Aspects of Statistical Confidentiality Project
  http://neon.vb.cbs.nl/CASC
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

And now.....

Microdata and Disclosure Risk

Microdata and Disclosure Risk

NAACCR 2008-2009 Webinar Series

Microdata Release Basics

- Microdata:
  - File of individual case records
  - We assume that explicit (direct) identifiers have been removed – name, SSN, address, etc.

- Types of exposure risk:
  - Identity exposure – revealing the identity of a previously unknown cancer patient
  - Attribute exposure – revealing additional information about a known cancer patient
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Microdata Intrusion Scenarios

- An individual or organization
  - Seeking to identify cancer patients for product sales (insurance, treatment options)
  - Seeking to know more about an acquaintance with cancer
- Access to commonly available resources
  - Internet locators: e.g., AnyWho, WhitePages, PeopleSearch
  - Casual observation: people in a small town

Indirect Identifiers

- Variables on a microdata file that could be used for indirect identification
- Examples: age, race, sex, birthplace, marital status, etc.
  - Things a casual observer could know and/or could be linked with common internet resources
- Often referred to as “keys”
- CDAC checklist – section 3 on microdata

Microdata Risk Assessment

- Population uniques:
  - Number or percent of records with a combination of key values that is unique in the population
  - Assessment of identity exposure risk
- Sample uniques:
  - Number or percent of records with a combination of key values that is unique on the file
  - Assessment of attribute exposure risk
- Unique (N=1) versus small number (N ≤ 5)
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Risk Assessment Tools

• Population uniques
  – No direct method without access to complete population microdata
  – The NCI/SEER program is working on an estimation method based on the Census Bureau’s Public-Use Microdata Sample (PUMS)

• Sample uniques
  – NAACCR record uniqueness program
  – CASC -- µ-Argus (http://neon.vb.cbs.nl/CASC/)

Microdata Risk Reduction

• Common risk reduction techniques for microdata:
  – Classification
  – Top and bottom coding
  – Field suppression
  – Swapping
  – Shuffling

Classification

• Combining values into groups or categories
• Also known as “global recoding”
• Can be used with either numeric or categorical variables
• Examples:
  – Group age into 5-year age groups
  – Combine detailed race groups into “white”, “black”, and “other”
  – Rounding of income to nearest $10,000
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Top and Bottom Coding
• Eliminate high or low extreme values
• Useful for variables with outliers or long tails
• Examples:
  – Combine youngest and oldest ages: under 20, 85+
  – Group high incomes: over $1,000,000 per year

Suppression
• Global suppression
  – Eliminate columns and/or rows for the entire file
  – Useful for data with little analytic value
  – Example: provide only the rows and columns needed for the specific research project
• Local suppression
  – Individual values for specific records
  – Example: suppress age for small racial subgroups

Swapping and Shuffling
• Swapping: exchange data values between two records
• Shuffling: perturbation of a numeric field with preservation of rank order correlation
• Not often used with cancer surveillance data
  – Significant impact on data utility for most cancer microdata files
Mapping Microdata

- A fully-qualified residence address is a direct identifier (household)
- A dot map based on residence addresses from microdata — is this an identity exposure risk?

Can People Be Identified from Dot Maps?

- Maps can be scanned and geo-referenced
- Can read latitude and longitude from map
- Reverse geocoding can convert to an address
Reverse Geocoding Example

- Enter latitude and longitude and click “Submit Point”
- Address is returned

Dot Map Case Study

- Recent NEJM study identifying addresses from a dot-map:
  - Dot map of 550 patients in Boston
  - Able to identify 432 addresses (79%)

Source: Brownstein et al, NEJM, 2006

Risk Reduction for Dot Maps

- Aggregation
  - Loss of geographic information
  - Artificial boundaries
  - Assumed homogeneity
- Derived spatial data
  - Example: NAACCR request for distance from patient’s residence to hospital
- Geo-masking – moving the point locations
  - Random distance within circular buffer
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Geo-Masking Example

- Randomly shifted dot map:

Other Risk Reduction Methods

- Data use agreements
  - Limit use to specific purpose (usually research)
  - Explicitly prohibit attempts to identify individuals
  - Require use of appropriate safeguards
  - Insure agents follow same restrictions

- Education
  - All data users
  - Repeat messaging

- Enforcement, audits

And now.....

Empirical Assessment of Unique Records Using CINA Deluxe
Empirical Assessment of Unique Records Using CINA Deluxe

Andrew Lake
Information Management Services Inc.

Acknowledgements

• Dr. Tiefu Shen, Illinois Cancer Registry

Acknowledgements

• Dr. Tiefu Shen, Illinois Cancer Registry
• Dr. Holly Howe
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Acknowledgements

• Dr. Tiefu Shen, Illinois Cancer Registry
• Holly Howe, NAACCR
• IMS Staff, Dave Roney

Objectives

• Why Record Uniqueness?
• Methodology
• Available Tools From NAACCR
• Application to CINA Deluxe
• Guidelines for Applying Record Uniqueness

Why Record Uniqueness?

• Balance Between Access to Data and Patient Privacy
Why Record Uniqueness?

- Balance Between Access to Data and Patient Privacy
- Confidentiality
- Re-Identify Existing Patients

NAACCR CINA Deluxe Advisory Group

- Guidelines For Data
  - Researcher Files - No more than 20% of all records should be unique in groups of 5 or less for any given combination of variables.
  - Public Use Files – No more than 5% of all records should be unique in groups of 5 or less for any given combination of variables.
NAACCR CINA Deluxe Advisory Group

Default Variable Set:
- Age
- Sex
- Race
- Year of Diagnosis
- Primary Cancer Site
- Geographic Area

Methodology

Step 1
- Generate a frequency distribution for a variable combination.
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

**Methodology**

**Step 1**
- Generate a frequency distribution for a variable combination.

**Step 2**
- Count the number of records with a frequency of one (unique records).
- Count the number of records with a frequency of 5 or less (unique records in groups of 5 or less).

**Methodology**

**Uniqueness - 1 Variable**

<table>
<thead>
<tr>
<th>Race</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>150</td>
</tr>
<tr>
<td>Japanese</td>
<td>50</td>
</tr>
<tr>
<td>Asian NOS</td>
<td>4</td>
</tr>
<tr>
<td>Other Asian</td>
<td>1</td>
</tr>
</tbody>
</table>

**Methodology**

**Uniqueness - 1 Variable**

<table>
<thead>
<tr>
<th>Race</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>150</td>
</tr>
<tr>
<td>Japanese</td>
<td>50</td>
</tr>
<tr>
<td>Asian NOS</td>
<td>4</td>
</tr>
<tr>
<td>Other Asian</td>
<td>1</td>
</tr>
</tbody>
</table>
## Methodology

### Uniqueness - 1 Variable

<table>
<thead>
<tr>
<th>Race</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>150</td>
</tr>
<tr>
<td>Japanese</td>
<td>50</td>
</tr>
<tr>
<td>Asian NOS</td>
<td>4</td>
</tr>
<tr>
<td>Other Asian</td>
<td>1</td>
</tr>
</tbody>
</table>

## Methodology

### Unique Records

<table>
<thead>
<tr>
<th>Race</th>
<th>Unique Records</th>
<th>Unique Records in Groups of 5 or Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>1 (.65%)</td>
<td>5 (3.2%)</td>
</tr>
</tbody>
</table>
Methodology
2 Variables
Two Variables: Age, Race

Methodology
2 Variables
Two Variables: Age, Race
Frequency Distributions
✓ Age

Methodology
2 Variables
Two Variables: Age, Race
Frequency Distributions
✓ Age
✓ Race
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Methodology
2 Variables

Two Variables: Age, Race

Frequency Distributions

- Age
- Race
- Age x Race

Methodology
2 Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unique Records</th>
<th>Unique Records in Groups of 5 or Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>1 (.65%)</td>
<td>5 (3.2%)</td>
</tr>
<tr>
<td>Age</td>
<td>2 (1.3%)</td>
<td>7 (4.5%)</td>
</tr>
<tr>
<td>Age x Race</td>
<td>11 (7.9%)</td>
<td>23 (14.3%)</td>
</tr>
</tbody>
</table>
### Methodology

#### N Variables

<table>
<thead>
<tr>
<th>Number of Variables</th>
<th>Number of Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
</tr>
</tbody>
</table>
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

### Methodology

#### N Variables

<table>
<thead>
<tr>
<th>Number of Variables</th>
<th>Number of Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>N</td>
<td>$2^{n-1}$</td>
</tr>
</tbody>
</table>

### Available Tools

- NAACCR Record Uniqueness Program
- NAACCR Record Uniqueness SAS Macro
  
  www.naaccr.org
  
  (under registry standards/registry operations)
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Record Uniqueness Program Input Screen

Record Uniqueness Program Input Screen

Record Uniqueness Program Input Screen
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

Record Uniqueness Program Input Screen

Record Uniqueness Program - Results

Record Uniqueness Program - Results
SAS Macro

- Some SAS experience is recommended
- Recommended for large files or up to 9 variables
- Easily create frequency distributions
- Use of formats allows many iterations with different aggregations
- Runs Very Quickly!

SAS Macro Output

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>age_rec</td>
<td>0.6322</td>
</tr>
<tr>
<td>age</td>
<td>0.5788</td>
</tr>
<tr>
<td>race</td>
<td>2.1743</td>
</tr>
<tr>
<td>year_dx</td>
<td>1.9907</td>
</tr>
<tr>
<td>sex</td>
<td>0.7673</td>
</tr>
<tr>
<td>state</td>
<td>0.7222</td>
</tr>
</tbody>
</table>
### Application Initial Run

#### Variable List

Site, Race, Age, Sex, Year DX, State

<table>
<thead>
<tr>
<th>Total Cases</th>
<th>Unique Cases in Groups of 5 or Less</th>
<th>Percent of Cases in Groups of 5 or Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>160,505</td>
<td>45,581</td>
<td>28.4</td>
</tr>
</tbody>
</table>

### Application Initial Run

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>6.3</td>
</tr>
<tr>
<td>Age</td>
<td>2.7</td>
</tr>
<tr>
<td>Year Dx</td>
<td>2.0</td>
</tr>
<tr>
<td>Race</td>
<td>1.2</td>
</tr>
<tr>
<td>State</td>
<td>0.7</td>
</tr>
<tr>
<td>Sex</td>
<td>0.6</td>
</tr>
</tbody>
</table>

### Application Initial Run

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>6.3</td>
</tr>
<tr>
<td>Age</td>
<td>2.7</td>
</tr>
<tr>
<td>Year Dx</td>
<td>2.0</td>
</tr>
<tr>
<td>Race</td>
<td>1.2</td>
</tr>
<tr>
<td>State</td>
<td>0.7</td>
</tr>
<tr>
<td>Sex</td>
<td>0.6</td>
</tr>
</tbody>
</table>
Managing and Minimizing the Disclosure Risk of Cancer Data for Research and Dissemination

**Application Run After Aggregation**

**Variable List**
- **Site Recode**: Race, Age, Sex, Year DX, State

<table>
<thead>
<tr>
<th>Total Cases</th>
<th>Unique Cases in Groups of 5 or Less</th>
<th>Percent of Cases in Groups of 5 or Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>160,505</td>
<td>28,522</td>
<td>17.8</td>
</tr>
</tbody>
</table>

**Guidelines For Using Record Uniqueness**

- Use Variables Related To Confidentiality (Default Variables)
- NAACCR Recommends Use for All Files
- Aggregate Before Eliminate
Questions?

Thank you for participating in today’s webinar!

• The next webinar is scheduled for 2/5/2009, and the topic is ‘Collecting Cancer Data: Pharynx’.
• Forward questions from today’s webinar to Shannon or Jim.
• Contact us at
  – Shannon Vann – svann@naaccr.org; 217-698-0800 X9
  – Jim Hofferkamp – jhofferkamp@naaccr.org; 217-698-0800 X5