

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

NAACCR 2008-2009 Webinar Series
January 8, 2009

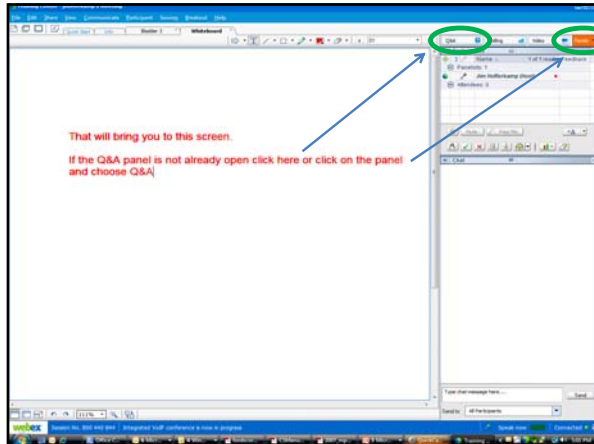


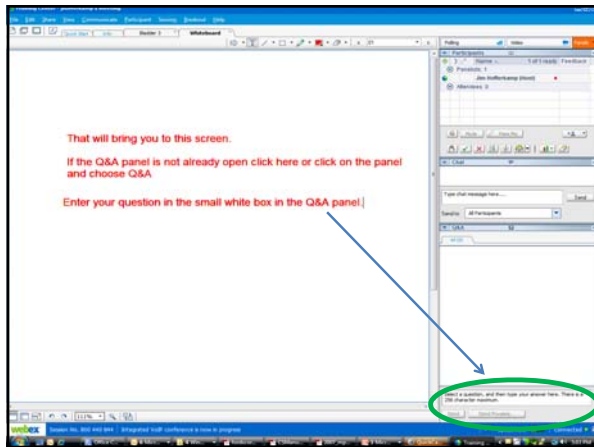
Q&A

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 key board





Prizes!

<p>Question of the Month!</p> <ul style="list-style-type: none"> 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. 	<p>Tip of the Month!</p> <ul style="list-style-type: none"> 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 Us
 - Jessica King, MPH, Biostatistician at the Centers for Disease Control & Prevention



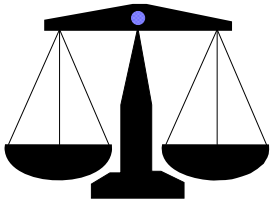
Today's Topics

- 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




NAACCR 2008-2009 Webinar Series




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



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




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




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

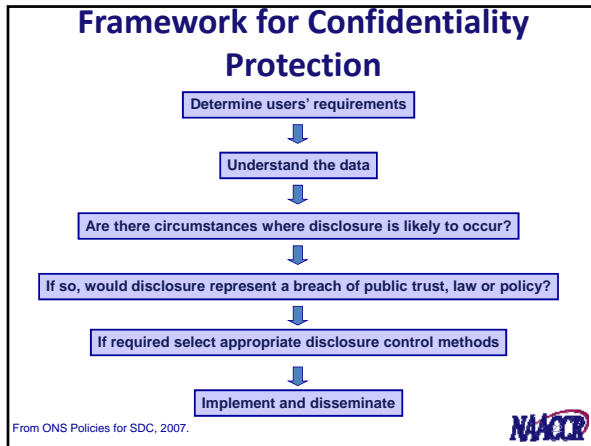


Informed Consent and Transparency

- 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






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



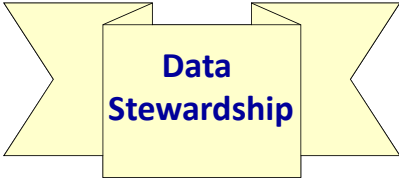
On the Importance of Research...

“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




Important Parts of the Data Stewardship Policy

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.



Important Parts of the Data Stewardship Policy

IV. Description of Confidential Data

- Directly identifies person or provider
- Indirectly identifies person or provider
 - small numbers rules
 - mapping rules
 - population thresholds for suppression




Important Parts of the Data Stewardship Policy

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

User	userid	email	Last SEER agreement	Access to 40401	Total # times accessed	# times accessed in last month	# times accessed in previous month	Date last accessed
John Doe	jdoe1	jdoe1@dcd.gov	2007	yes	7	4	1	11/02/08
Jane Doe	jdoe2	jdoe2@cdc.gov	2007	yes	18	8	3	12/04/08




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

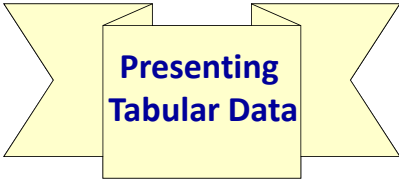
VII. Data Release Procedures
phased rollout to partners/individual customers such as universities, agencies, others/general public... level of data released to each may differ and data release/use agreements will also differ




Questions?



And now.....




**Presenting
Tabular Data**




Presenting Tabular Data
Confidentiality Considerations

NAACCR
2008-2009 Webinar Series




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 a 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 individuals characteristic in a range, ie 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

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

	Total		White		Black		Am. Indian		Asian/PI	
	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
Total	268	189	267	186	1	1	0	1	0	1
under 20	2	1	2	1	0	0	0	0	0	0
20-29	3	0	3	0	0	0	0	0	0	0
30-39	9	2	8	2	1	0	0	0	0	0
40-49	14	7	14	7	0	0	0	0	0	0
50-59	37	30	37	30	0	0	0	0	0	0
60-69	79	57	79	56	0	1	0	0	0	0
70 and over	124	92	124	90	0	0	1	0	1	0



Attribute Disclosure - External

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

	Total		White		Black		Am. Indian		Asian/PI	
	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
Total	268	189	267	186	1	1	0	1	0	1
under 20	2	1	2	1	0	0	0	0	0	0
20-29	3	0	3	0	0	0	0	0	0	0
30-39	9	2	8	2	1	0	0	0	0	0
40-49	14	7	14	7	0	0	0	0	0	0
50-59	37	30	37	30	0	0	0	0	0	0
60-69	79	57	79	56	0	1	0	0	0	0
70 and over	124	92	124	90	0	0	0	1	0	1



Attribute Disclosure - Inferential

Lung Cancer Cases by Age, Race and Stage at Diagnosis

Lakeside County Residents, 2001-2005

	Total		White		Black		Am. Indian		Asian/PI	
	Early	Late	Early	Late	Early	Late	Early	Late	Early	Late
Total	268	189	267	186	1	1	0	1	0	1
under 20	2	1	2	1	0	0	0	0	0	0
20-29	3	0	3	0	0	0	0	0	0	0
30-39	9	2	8	2	1	0	0	0	0	0
40-49	14	7	14	7	0	0	0	0	0	0
50-59	37	30	37	30	0	0	0	0	0	0
60-69	79	57	79	56	0	1	0	0	0	0
70 and over	124	92	124	90	0	0	0	1	0	1



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



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

Age at Diagnosis	Total	
	Early	Late
Total	268	189
under 40	28	10
50-59	37	30
60-69	79	57
70 and over	124	92



Frequency of Incident Cancers in Five Ontario Counties 2002

Cancer Site	County 1	County 2	County 3	County 4	County 5	All 5 Counties
Oral Cavity and Pharynx	61	12	25	49	9	156
Esophagus	19	6	11	19	3	58
Stomach	52	3	7	41	14	117
Colon and Rectum	301	50	85	288	83	807
Pancreas	36	9	17	43	15	120
Lung and Bronchus	283	45	120	316	76	840
Melanoma of the Skin	80	17	24	42	25	188
Breast	464	48	106	309	78	1005
Cervix Uteri	23	1	6	14	5	49
Corpus and Uterus, NOS	76	8	17	82	19	202
Ovary	51	7	10	40	12	120
Prostate	385	68	117	308	78	956
Testis	19	1	6	14	3	44
Urinary Bladder	85	12	13	68	19	197
Kidney and Renal Pelvis	61	11	18	58	12	160
Brain	48	7	11	29	10	105
Thyroid	138	9	9	35	11	202
Hodgkin Lymphoma	34	3	8	7	2	54
Non-Hodgkin Lymphoma	88	14	34	92	25	253
Leukemia	54	10	21	70	14	169
All Others	26	3	6	24	9	68
All Sites	2384	344	671	1948	523	5870



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



	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	797	309	838	366	955	2965
Row 2	348	742	86	453	881	2410
Row 3	252	271	324	795	174	1816
Row 4	284	858	743	793	423	3101
Row 5	12	875	700	955	772	2914
Row 6	953	871	366	747	681	3618
Row 7	137	188	927	721	680	2143
Row 8	143	793	782	527	916	2541
Row 9	560	647	633	527	987	3354
Row 10	3477	9542	5002	5748	6817	20586

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	797	309	838	366	955	2965
Row 2	348	742	86	453	881	2410
Row 3	252	271	324	795	174	1816
Row 4	284	858	743	793	423	3101
Row 5	12	875	700	955	772	2914
Row 6	953	871	366	747	681	3618
Row 7	137	188	927	721	680	2143
Row 8	143	793	782	527	916	2541
Row 9	560	647	633	527	987	3354
Row 10	3477	9542	5002	5748	6817	20586

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

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	309	838	366	555	2053	3053
Row 2	742	95	453	591	2399	2990
Row 3	348	158	797	768	2074	2646
Row 4	252	271	324	795	174	1856
Row 5	234	958	743	733	425	3133
Row 6	12	875	700	595	772	2914
Row 7	593	871	366	747	681	3058
Row 8	137	188	527	721	662	2335
Row 9	143	793	782	513	916	2547
Row 10	560	647	633	527	907	3374
Col Sums	3477	5542	5302	5748	6817	29398

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	210	843	305	555	2076	2999
Row 2	360	180	795	765	2079	2699
Row 3	290	270	325	785	175	1865
Row 4	295	860	740	795	425	3115
Row 5	10	875	700	595	775	2915
Row 6	365	870	365	745	680	3025
Row 7	125	110	525	725	660	2345
Row 8	145	795	780	515	915	2550
Row 9	560	645	635	530	905	3375
Col Sums	3475	5545	5300	5750	6815	29395

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, $a_{ij} \leq$ "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.

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	1	309	630	364	555	2559
Row 2	797	742	86	453	681	2959
Row 3	348	158	3	797	758	2014
Row 4	252	271	324	795	174	1856
Row 5	284	856	743	793	423	3101
Row 6	12	875	793	595	772	2914
Row 7	363	871	366	747	681	3018
Row 8	127	138	527	721	680	2143
Row 9	143	703	782	4	916	2548
Row 10	560	647	633	527	967	3734
Col Sums	3477	8542	5002	5749	6817	20506

	Col 1	Col 2	Col 3	Col 4	Col 5	Row Sums
Row 1	1	309	630	364	555	2559
Row 2	797	742	86	453	681	2959
Row 3	348	158	3	797	758	2014
Row 4	252	271	324	795	174	1856
Row 5	284	856	743	793	423	3101
Row 6	12	875	793	595	772	2914
Row 7	363	871	366	747	681	3018
Row 8	127	138	527	721	680	2143
Row 9	143	703	782	4	916	2548
Row 10	560	647	633	527	967	3734
Col Sums	3477	8542	5002	5749	6817	20506

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

Agency	Procedures to Protect Tabular Data
Census	Data Swapping Query Rules Threshold Rule p-percent
National Center for Health Statistics	Threshold Rule 4+ (n,k) (1,6)
Dept of Education	Data Swapping Data Coarsening Accuracy
Agency for Healthcare Research and Quality	Threshold Rule 4+
Social Security Administration	Threshold Rule 5+ Marginals, 3+ cells
Internal Revenue Service	Threshold Rule 3+
National Science Foundation	Varies by risk



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



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.naacr.org/confidentiality/index.asp
- Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper #22, FCSM, CDAC, December 2005 (Revised)
<http://www.fcsm.gov/working-papers/spwp22.html>
- Checklist on Disclosure Potential of Proposed Data Releases, July 1999
<http://www.fcsm.gov/committees/cdac/>

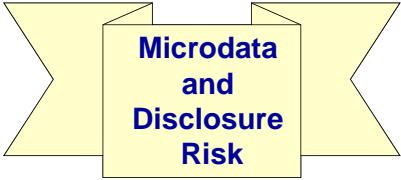


References


- 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>



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



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)



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



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



Reverse Geocoding Example

- Enter latitude and longitude and click "Submit Point"
- Address is returned



– ASTHO Headquarters



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



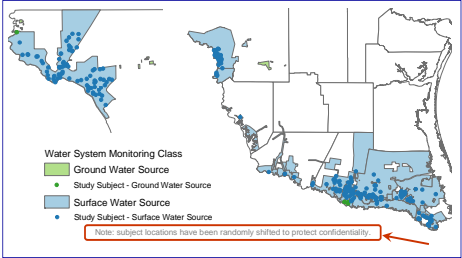
Risk Reduction for Dot Maps

- 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




Geo-Masking Example

- Randomly shifted dot map:


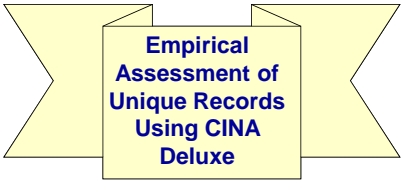


Other Risk Reduction Methods

- 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**

Andrew Lake
Information Management Services Inc.



Acknowledgements

- Dr. Tiefu Shen, Illinois Cancer Registry



Acknowledgements

- Dr. Tiefu Shen, Illinois Cancer Registry
- Dr. Holly Howe



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




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



Methodology

Step 1

- Generate a frequency distribution for a variable combination.




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

Race	Frequency
Chinese	150
Japanese	50
Asian NOS	4
Other Asian	1

Methodology Uniqueness - 1 Variable

Race	Frequency
Chinese	150
Japanese	50
Asian NOS	4
Other Asian	→ 1

Methodology
Uniqueness - 1 Variable

Race	Frequency
Chinese	150
Japanese	50
Asian NOS	→ 4
Other Asian	→ 1

Methodology

Variable	Unique Records	Unique Records in Groups of 5 or Less
Race	1 (.65%)	5 (3.2%)


Methodology

Variable	Unique Records	Unique Records in Groups of 5 or Less
Race	1 (.65%)	5 (3.2%)

↑

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




Methodology
2 Variables

Two Variables: Age, Race

Frequency Distributions

- ✓ Age
- ✓ Race
- ✓ Age x Race




Methodology
2 Variables

Variable	Unique Records	Unique Records in Groups of 5 or Less
Race	1 (.65%)	5 (3.2%)
Age	2 (1.3%)	7 (4.5%)
Age x Race	11 (7.9%)	23 (14.3%)

Methodology
2 Variables

Variable	Unique Records	Unique Records in Groups of 5 or Less
Race	1 (.65%)	5 (3.2%)
Age	2 (1.3%)	7 (4.5%)
Age x Race	11 (7.9%)	23 (14.3%)



Methodology
N Variables

Number of Variables	Number of Frequencies
1	1
2	3

Methodology
N Variables

Number of Variables	Number of Frequencies
1	1
2	3
3	7

Methodology
N Variables

Number of Variables	Number of Frequencies
1	1
2	3
3	7
5	31

Methodology N Variables

Number of Variables	Number of Frequencies
1	1
2	3
3	7
5	31
N	$2^N - 1$

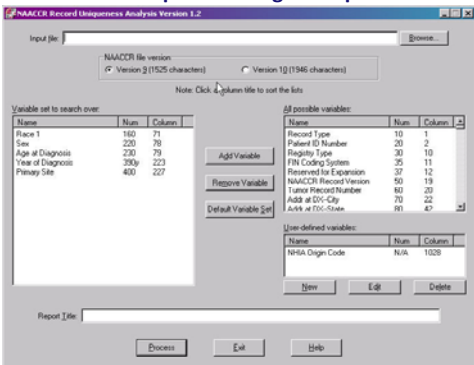
Available Tools

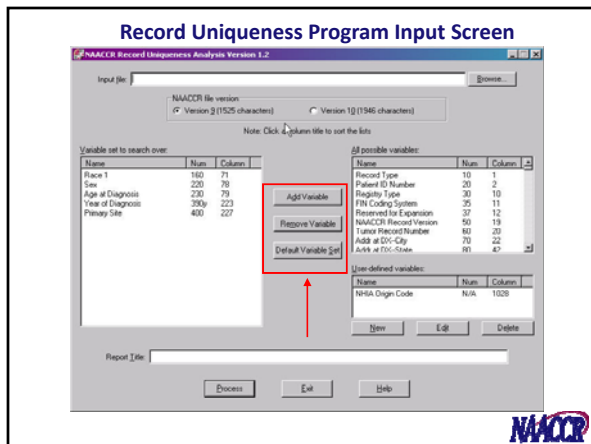
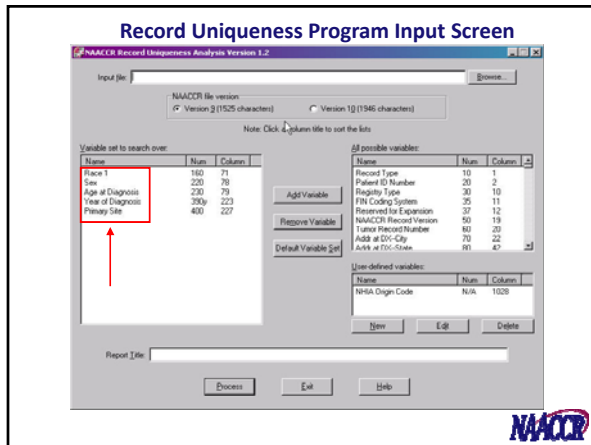
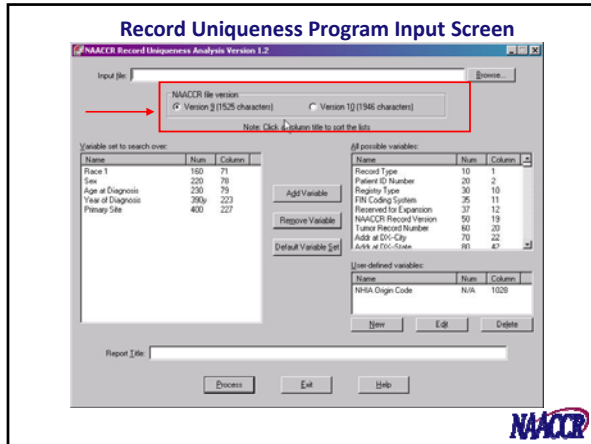
- NAACCR Record Uniqueness Program
- NAACCR Record Uniqueness SAS Macro

www.naacccr.org
(under registry standards/registry operations)



Record Uniqueness Program Input Screen





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

```

NAACCR Record Uniqueness Macro (v1.0) 04/2005      09:08 Sunday, June 5, 2009
Basic Results - Counts and Percentages of Unique Variables
Record Uniqueness Example #2
Site Recoded to SEER Site Categories

```

Total Cases	Unique Cases	Percent Unique Cases	Unique Cases in Groups of 5 or Less	Percent of Cases Unique in Groups of 5 or Less	Variable Set
160,605	0	0.00	23	0.01	sex, year_dx, age
160,605	0	0.00	5	0.00	state, site_rec, sex
160,605	0	0.00	89	0.05	state, sex, year_dx, age
-	-	-	-	-	-
-	-	-	-	-	-
160,605	4,209	2.62	19,927	12.42	state, site_rec, sex, year_dx, age
160,605	5,090	3.17	19,279	12.01	state, site_rec, race, year_dx, age
160,605	6,004	3.74	21,766	13.56	site_rec, sex, race, year_dx, age
160,605	8,290	5.16	28,622	17.77	state, site_rec, sex, race, year_dx, age



SAS Macro Output

```

NAACCR Record Uniqueness Macro (v1.0) 04/2005
Advanced Results - Variable and Weight Coefficients
Record Uniqueness Example #2
Site Recoded to SEER Site Categories

```


Variable	Weight
site_rec	5.6123
age	3.6788
race	2.1340
year_dx	1.9597
sex	0.7676
state	0.7022



**Application
Initial Run**


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

Total Cases	Unique Cases in Groups of 5 or Less	Percent of Cases in Groups of 5 or Less
160,505	45,581	28.4




**Application
Initial Run**

Variable	Weight
Site	6.3
Age	2.7
Year Dx	2.0
Race	1.2
State	0.7
Sex	0.6



**Application
Initial Run**


Variable	Weight
Site	6.3
Age	2.7
Year Dx	2.0
Race	1.2
State	0.7
Sex	0.6



Application
Run After Aggregation

Variable List
Site Recode, Race, Age, Sex, Year DX, State


Total Cases	Unique Cases in Groups of 5 or Less	Percent of Cases in Groups of 5 or Less
160,505	28,522	17.8



Application
Run After Aggregation


Variable List
Site Recode, Race, Age, Sex, Year DX, State

Total Cases	Unique Cases in Groups of 5 or Less	Percent of Cases in Groups of 5 or Less
160,505	28,522	17.8




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

