

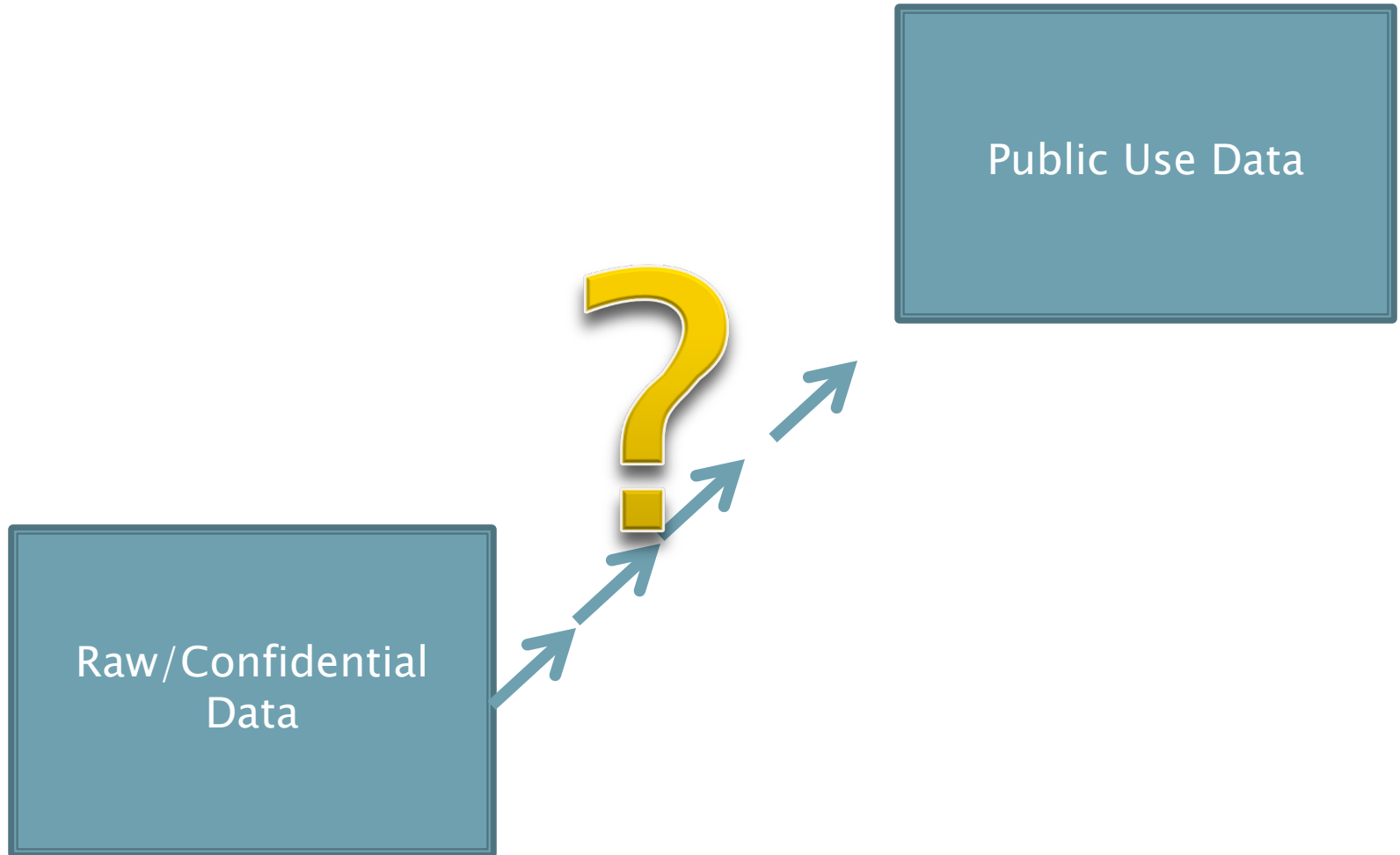
Assessing Identifiability in Public Use Data

Tiefu Shen, MD, PhD
Division of Epidemiologic Studies
Illinois Department of Public Health

Types of Public Use Data

- ▶ Tabular files
 - Aggregated data
- ▶ Microdata files
 - Record-level data

Note: Tabular files can be converted into microdata files, and vice versa.



Identifiability

- ▶ Definition: The extent to which individuals in a data file can be uniquely identified
- ▶ Real world scenario: People who know or have information about the individuals can find them in the data file
- ▶ Real world consequence: Once the individuals are identified, their health conditions are revealed and privacy violated

Identifiability Risk of Data Files

.....A record is at risk of being identified if a respondent is unique in the database with respect to a set of identifying variables.....

*Statistical Policy Working Paper 22 (second version, 2005)
Report on Statistical Disclosure Limitation Methodology
Federal Committee on Statistical Methodology (FCSM)*

A Common Myth

“A file is safe if personal identifiers are removed”

- ▶ Why is it a myth?

- Presence of demographic, geographic and date/time variables
- These variables, when used in combination, can form unique values
- These unique values can be just as identifying as conventional personal identifiers

- ▶ Uniqueness of records = a file's identifiability

- Personal identifiers identify because of their uniqueness
- All data linkage programs rely on uniqueness of records

Which file can be identified more easily? And why?

DoB	Sex	Location	Race/ Ethnicity
06-22-1956	M	63511	NHW
05-12-1956	F	45321	NHW
12-23-1956	M	12345	HW
03-07-1951	M	45678	NHB
11-05-1923	F	67895	A
.....			
.....			

DoB	Sex	Location	Race/ Ethnicity
1956	M	Sangamon	NHW
1956	F	Sangamon	NHW
1943	M	Sangamon	HW
1951	M	Mason	NHB
1923	F	Mason	A
.....			
.....			

By quantifying a data file's risk of identifiability through record uniqueness, we can –

- Compare different data files
- Choose data files for public release (or FOIA release)
- Fine-tune a data file's identifiability risk

A SAS Algorithm for Assessing Record Uniqueness

- ▶ Created by the Division of Epidemiologic Studies, IDPH
- ▶ Adopted by some cancer registries as the best practice for releasing cancer public use data
- ▶ Currently used by many agencies, such as IDPH, IDHFS, et al for releasing non-confidential health-related data

State of Illinois
Rod R. Blagojevich, Governor
Department of Public Health
Eric E. Whitaker, M.D., M.P.H., Director

Guidelines for Generating and Maintaining Public Use Microdata Files

Fomoff, J.; Bengu, B.; Shen, T. and Lehnher, M. Simulated Attempt to Identify Individuals in State Cancer Registry Public Use Data File. *ISCR Quality Control Series 06.01*. Springfield, IL: Department of Public Health, May 2006.

SIMULATED ATTEMPT TO IDENTIFY INDIVIDUALS IN THE ILLINOIS STATE CANCER REGISTRY PUBLIC USE DATA FILE

May 2006

Linking Data for Health Services Research: A Framework and Instructional Guide



AHRQ
Agency for Healthcare Research and Quality
Advancing Evidence in Health Care • www.ahrq.gov

American Journal of Epidemiology Advance Access published December 20, 2006

DOI: 10.1093/aje/kwk049



American Journal of Epidemiology
Copyright © 2006 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Practice of Epidemiology

Method to Assess Identifiability in Electronic Data Files

Holly L. Howe¹, Andrew J. Lake², and Tiefu Shen³

¹ North American Association of Central Cancer Registries, Inc., Springfield, IL.

² Information Management Services, Inc., Silver Spring, MD.

³ Illinois State Cancer Registry, Springfield, IL.

Received for publication December 19, 2005; accepted for publication July 28, 2006.

The authors developed the Record Uniqueness (RU) software program to assess electronic data files for risk of

The Algorithm Calculates

- ▶ For each variable and variable combination
 - Percent of unique records
 - Percent of records with 5 or less occurrences
- ▶ Each variable's contribution to the change in percent unique records

Two Measures from the Algorithm

- ▶ Unique records
 - Records with frequency equal to 1 as defined by key variables (i.e., records in small bins of $n=1$)
- ▶ Records occurring five times or fewer
 - Records with frequency equal to or less than 5 as defined by key variables (i.e., records in small bins of $n \leq 5$)
 - The number also indicates how many records would be suppressed if data are released as a tabular file

A Data File and Its Frequencies

Obs	AGE	SEX	RACE	ZIP
1	23	M	W	34950
2	45	M	B	23687
3	45	F	B	23687
4	32	M	A	34950
5	56	M	B	36475
6	37	F	W	23485
7	67	M	W	23687
8	56	F	W	23687
9	56	F	B	34950
10	56	F	A	36475
11	34	M	W	36475
12	56	F	W	36475
13	45	M	W	23687
14	56	F	B	23687
15	58	M	W	23485

Variable	Unique Records	Records Occurring Five Times or Fewer
AGE	6 (6/15=40%)	9 (9/15=60%)
SEX	0 (0/15=0%)	0 (0/15=0%)
RACE	0 (0/15=0%)	7 (7/15=47%)
ZIP	0 (0/15=0%)	9 (9/15=60%)
AGE*SEX	8 (8/15=53%)	15 (15/15=100%)
.....		
AGE*SEX *RACE*ZIP	15 (15/15=100%)	15 (15/15=100%)

Number and percent of cases within small bins

Obs	which	case_1	pct_1	case_5	pct_5	v_list
1	1	6	40.000	9	60.000	age
2	3	0	0.000	7	46.667	race
3	4	0	0.000	9	60.000	zip
4	5	8	53.333	15	100.000	age SEX
5	6	8	53.333	15	100.000	age race
6	7	7	46.667	15	100.000	age zip
7	8	2	13.333	15	100.000	SEX race
8	9	3	20.000	15	100.000	SEX zip
9	10	5	33.333	15	100.000	race zip
10	11	11	73.333	15	100.000	age SEX race
11	12	9	60.000	15	100.000	age SEX zip
12	13	13	86.667	15	100.000	age race zip
13	14	11	73.333	15	100.000	SEX race zip
14	15	15	100.000	15	100.000	age SEX race zip

Variables and their influence on confidentiality

Obs	Variable	Weight	P_Change
1	age	1.08750	196.685
2	zip	0.53570	70.865
3	race	0.51195	66.854
4	SEX	0.21910	24.496

After grouping age to intervals and replacing ZIP with County

Number and percent of cases within small bins

Obs	which	case_1	pct_1	case_5	pct_5	v_list
1	1	1	6.6667	5	33.333	age
2	3	0	0.0000	7	46.667	race
3	5	2	13.3333	9	60.000	age SEX
4	6	3	20.0000	15	100.000	age race
5	7	1	6.6667	9	60.000	age county
6	8	2	13.3333	15	100.000	SEX race
7	9	0	0.0000	15	100.000	SEX county
8	10	2	13.3333	15	100.000	race county
9	11	4	26.6667	15	100.000	age SEX race
10	12	5	33.3333	15	100.000	age SEX county
11	13	5	33.3333	15	100.000	age race county
12	14	4	26.6667	15	100.000	SEX race county
13	15	7	46.6667	15	100.000	age SEX race county

After grouping age to intervals and replacing ZIP with County

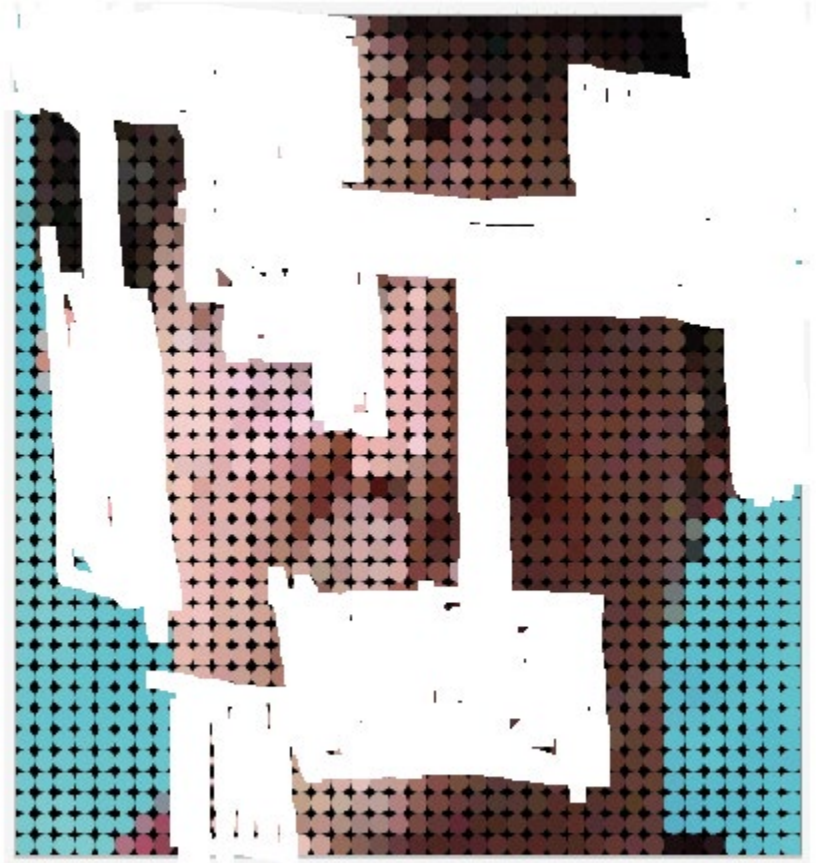
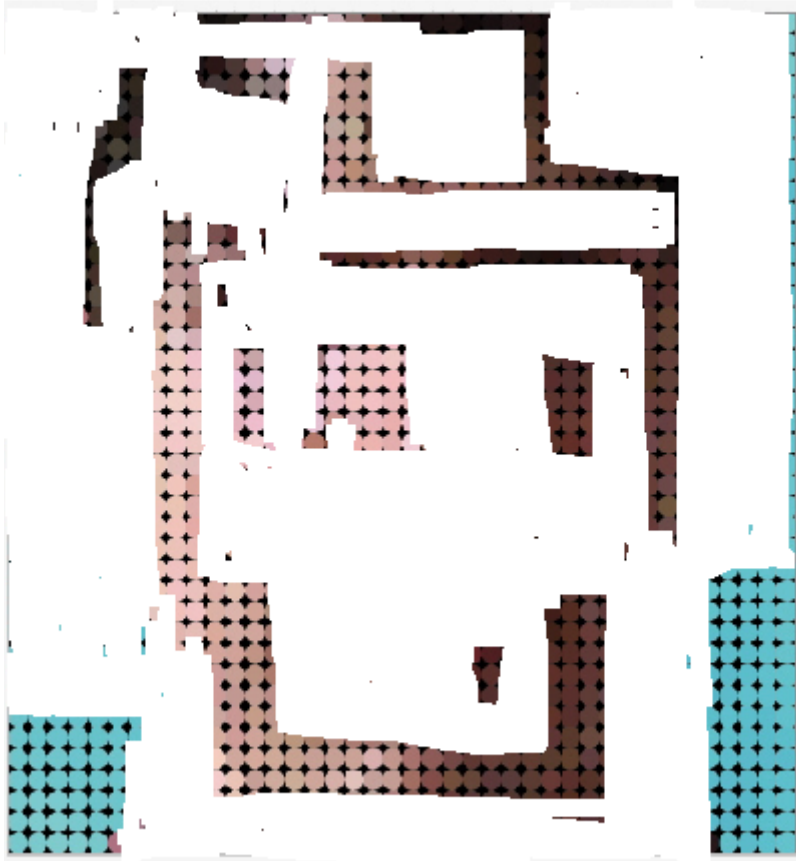
Variables and their influence on confidentiality

Obs	Variable	Weight	P_Change
1	race	0.93442	154.573
2	age	0.79436	121.301
3	SEX	0.70380	102.142
4	county	0.54713	72.829

What Causes Record Uniqueness to Increase

- ▶ More variables
 - For example, age and sex compared to age, sex, and ZIP code
- ▶ More specificity (value) in variables
 - For example, age in 10-year intervals compared to age in exact years

Mosaic Effect



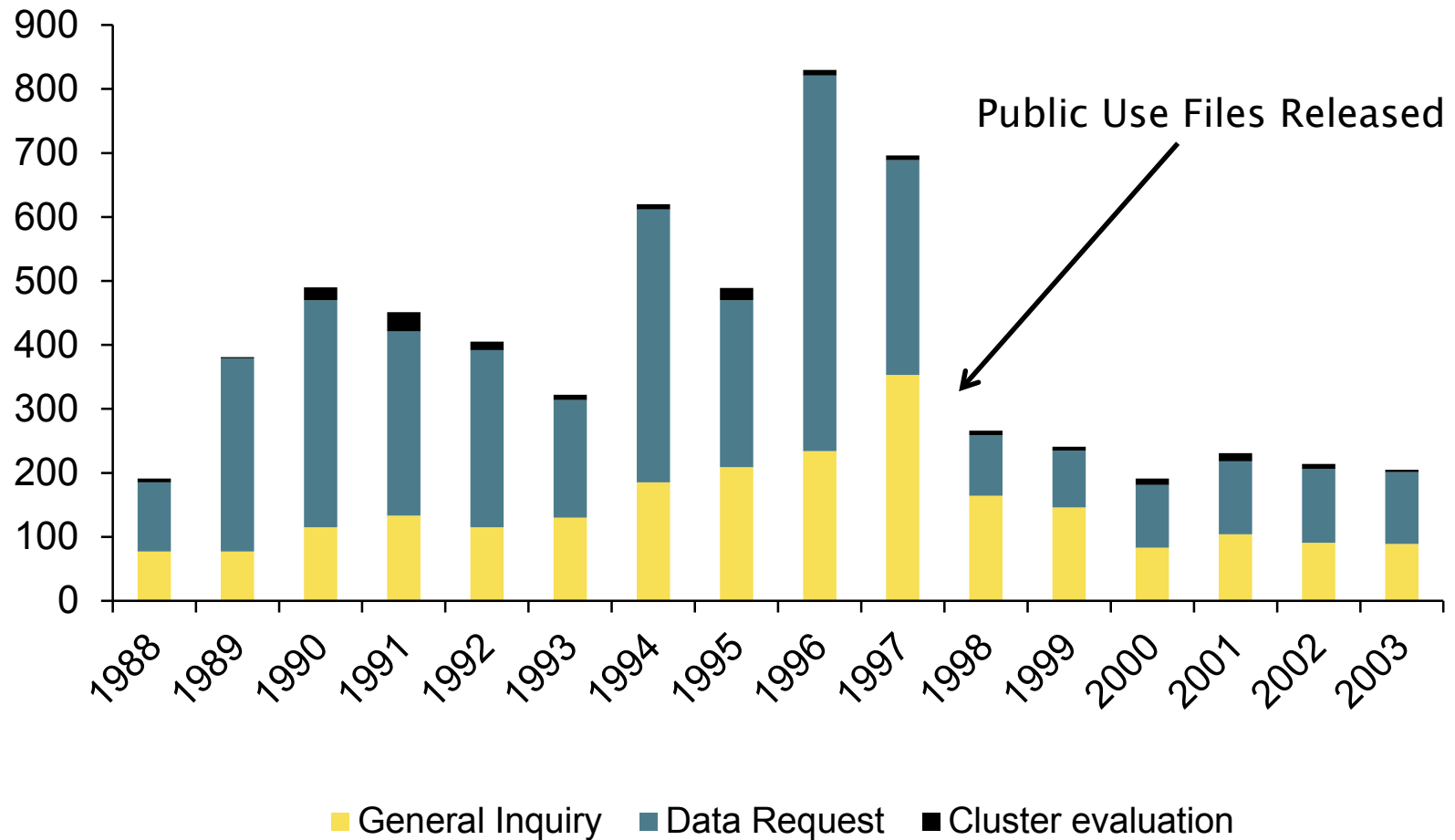
When to Use the Algorithm

- ▶ Release data as public use data files
- ▶ Compare files for identification risks
- ▶ Adjust a file to balance risk and benefit
- ▶ Follow the minimum risk principle in releasing data

Benefits of Releasing Public Use Data

- ▶ Transparency, accountability and efficiency
- ▶ Community empowerment
- ▶ Easy data access for researchers
- ▶ Relief to Institutional Review Board (IRB)
- ▶ Relief to programs

Number of Data Requests Dropped after Releasing Public Use Data Files – Div of Epi



Rules for Releasing Data

	Public Use Data (PUD)	Customized & Limited Data (CLD)	Confidential Data
Purpose/ Application	Release to anyone, including the general public	Release under discretion to partners or researchers	Release to bonafide researchers or users
Record uniqueness by key variables*	$\leq 5\%$	$> 5\% - \leq 20\%$	$> 20\%$
Data use agreement/Assurance required	No	Yes	Yes
IRB review required	No	No	Yes

What Are the Key Variables

- ▶ Demographic variables
 - age, sex, race, ethnicity, education, etc.
- ▶ Geographic variables
 - ZIP code, city, county, etc.
- ▶ Chronologic variables
 - date of birth, date of death, etc.
- ▶ Program-specific variables
 - e.g., date of DX for cancer (for ISCR)

Relation to HIPAA Privacy Framework

- ▶ Complete de-identification/safe harbor approach
 - List of 18 personal identifiers that must be removed
 - Very conservative
- ▶ Limited data set approach
 - List of 16 personal identifiers that must be removed
 - A data use agreement must be established
 - Very liberal without the data use agreement
- ▶ Statistical de-identification
 - Done by a qualified statistician/expert
 - Use accepted methodology

Survey of Other Approaches

- ▶ HIPAA personal identifiers (PIs)
 - Safe harbor (18 PIs) – too restrictive
 - Limited Data (16 PIs) – too liberal
- ▶ Cell suppression
 - Equivalent to limiting bin sizes (and probability of identifying a record in the cell)
 - Tend to be inflexible and restrictive
- ▶ Scoring systems (e.g., rule of ones)
 - Arbitrary
 - Hard to generalize
- ▶ Based on population size
 - Inflexible and often too restrictive
- ▶ Because I said so!
 - Arbitrary and risky

HIPAA Safe Harbor's 18 Personal Identifiers

Names	All geographic subdivisions smaller than a state
All elements of dates (except year)	Telephone number
Fax number	Electronic mail addresses
Social security numbers	Medical record numbers
Health plan beneficial numbers	Account numbers
Certificate and license numbers	VINs and license plate numbers
Medical device IDs and series #s	Internet universal resource locators (URLs)
Internet protocol (IP) addresses	Biometric IDs incl. fingerprints and voice prints
Full-face photographic images etc.	Any other unique identifying numbers, characteristic, or code

HIPAA Limited Data Set's 16 Personal Identifiers

Names	Postal address info., other than town or city, state, and ZIP Code
Telephone number	Fax number
Electronic mail addresses	Social security numbers
Medical record numbers	Health plan beneficial numbers
Account numbers	Certificate and license numbers
VINs and license plate numbers	Medical device IDs and series #s
Internet universal resource locators (URLs)	Internet protocol (IP) addresses
Biometric IDs incl. fingerprints and voice prints	Full-face photographic images etc

Some References

- ▶ Am J Epidemiology. 2007;165:597–601. Method to assess identifiability in electronic data files
- ▶ IDPH. 2007. Guidelines for generating and maintaining public use microdata files
- ▶ IDPH ISCR Quality Control Series. 2006 06:01. Simulated attempt to identify individuals in the Illinois State Cancer Registry Public Data File
- ▶ Report on Statistical Disclosure Limitation Methodology. Federal Committee on Statistical Methodology STATISTICAL POLICY WORKING PAPER 22 (Second version, 2005)
- ▶ Science. 2015, 30:347(6221):536–9. Unique in the shopping mall: on the reidentifiability of credit card metadata