Can you eat your cake and have it too?

Sharing healthcare data without compromising privacy or confidentiality

12th National HIPAA Summit
Concurrent Session IV
April 11, 2006

Krish Muralidhar
University of Kentucky

Rathindra Sarathy
Oklahoma State University

HIPAA Requirements

- "A major goal of the Privacy Rule is to assure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well being. The Rule strikes a balance that permits important uses of information, while protecting the privacy of people who seek care and healing."

  (Department of Health and Human Services Office of Civil Rights Privacy Brief, 2003, page 1)

NIH Requirements

- "… investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing."
- "… the rights and privacy of people who participate in NIH-sponsored research must be protected at all times."

Resolving Conflicting Goals

- Have your cake
  - Protect privacy and confidentiality of data
- …And eat it too
  - Make meaningful data available for purposes of analysis either within the organization or across organizations

Privacy

- Individuals claim that data about themselves should not be automatically available to other individuals and organizations, and that, even where data is possessed by another party, the individual must be able to exercise a substantial degree of control over that data and its use. This is sometimes referred to as 'data privacy' and 'information privacy'.

  – Introduction to Dataveillance and Information Privacy, and Definitions of Terms (http://www.anu.edu.au/people/Roger.Clarke/DV/Intro.html#Priv)

Privacy & Confidentiality Violations

- Identity Disclosure
  - Identity disclosure occurs when the identity of an individual can be inferred from the released data.
- Value Disclosure
  - Value disclosure occurs when the value of one or more variables can be inferred from the released data.

Meaningful Use …

- When data is distributed or shared for analysis purposes, a legitimate user is not interested in the values belonging to an individual record. The legitimate user is interested in analyzing the data at the aggregate level typically using standard statistical techniques.

… Legitimate users

- Our situation - make data available for legitimate users to perform legitimate analysis
- Unfortunately, a legitimate user can use the data to compromise privacy. Known as “snoopers” or “data intruders”

Not Hackers …

- NOT talking about preventing data from access by unauthorized users (such as hackers).
- Hacking is prevented using a different set of techniques

The Problem

- We want to make data available to authorized users for performing legitimate analyses
- We want to protect privacy by preventing snoopers from gaining information about the identity of an individual and/or the value of a particular attribute belonging to that individual

An Intra-Organizational Case

- Data has been gathered on individuals who have gone through tests for a specific disease in a particular department of a hospital.
- The hospital would like to analyze the costs associated with the office visits and lab tests from this department.
- Analyst is in the Accounting Department

Original Data

Patient #	Number of Office Visits	Number of Lab Visits	Original Data Set		
			Total Lab Charges	Total charges	Diagnosis
17098685	4	1	$1,979.54	$3,658.76	Negative
34469932	4	1	$1,823.28	$3,606.88	Negative
34787482	4	1	$2,982.10	$4,205.32	Positive
35425482	3	1	$653.54	$1,952.24	Positive
38941958	4	2	$1,090.47	$2,655.64	Negative
39205467	2	1	$1,620.66	$2,015.55	Negative
47318787	3	1	$1,080.53	$2,542.84	Negative
50335062	4	2	$1,349.87	$3,313.51	Negative
52329894	2	1	$256.08	$1,103.58	Negative
53022303	4	1	$1,096.95	$3,126.57	Negative
53140817	4	1	$1,580.04	$3,415.68	Negative
53168247	2	1	$1,819.77	$2,499.43	Negative
58019987	3	2	$1,783.64	$2,778.46	Negative
58335668	4	1	$1,344.25	$2,714.48	Negative
61779850	4	1	$1,784.11	$2,849.96	Positive
65728345	4	2	$1,058.46	$3,190.44	Negative
68996369	4	1	$1,140.26	$3,205.93	Negative
70101538	3	2	$1,978.21	$2,867.70	Negative
86485045	2	1	$1,591.13	$2,046.53	Negative
90757765	2	1	$1,567.82	$2,392.70	Negative

What is the big deal? Just remove the patient numbers

Number of Office Visits	Number of Lab Visits	Lab Charges	Total charges	Diagnosis
4	1	$1,979.54	$4,352.70	Negative
4	1	$1,823.28	$3,006.15	Negative
4	1	$2,982.10	$5,656.33	Positive
3	1	$653.54	$2,772.25	Positive
4	2	$1,090.47	$3,149.80	Negative
2	1	$1,620.66	$2,620.04	Negative
3	1	$1,080.53	$2,687.82	Negative
4	2	$1,349.87	$2,497.06	Negative
2	1	$256.08	$1,173.01	Negative
4	1	$1,096.95	$3,115.76	Negative
4	1	$1,580.04	$3,602.41	Negative
2	1	$1,819.77	$3,015.02	Negative
3	2	$1,783.64	$3,385.74	Negative
4	1	$1,344.25	$2,531.91	Negative
4	1	$1,784.11	$3,519.64	Positive
4	2	$1,058.46	$1,983.00	Negative
4	1	$1,140.26	$4,611.78	Negative
3	2	$1,978.21	$3,537.55	Negative
2	1	$1,591.13	$2,984.06	Negative
2	1	$1,567.82	$2,909.26	Negative

Rose in Accounting.. The legitimate user

- Rose is an analyst who gets the data. She is a legitimate user
- Rose needs to analyze the data to answer legitimate questions such as:
  - What is the average total charge for a patient who tests positive?
  - What is the relationship between lab charges and total cost?

Rose (the snooper) wants to find out …

- If a particular patient, Joe Schmo tested positive
- Can Rose succeed in violating Joe's privacy, when there are no patient numbers?

Recent Computer Associates Ad

When the original data is released (even without identifiers)

- Rose knows that Joe Schmo was charged $1784.11 for Lab tests and incurred $3519.64 total charges from this department

  … there is only one patient with $1784.11 for Lab tests and $3519.64 total costs in the data set

  … and the patient tested Positive
- Rose knows Joe Schmo tested positive for the disease
- If Rose has similar information about other patients, their privacy will be violated
- There may be many other people like Rose

Protect Data using Encryption?

- Encryption is not a solution
  - Encryption only prevents unauthorized users from viewing the data
  - It cannot be used to prevent disclosure to authorized users
    - Encrypted data cannot be analyzed
    - Decrypted data provides no protection

Rose – A greater threat?

- Rose knows a lot more about the data and the organization and could infer information that an outsider could not
- Rose may have been doing this for years and nobody ever found out … and if Rose is careful, nobody ever will
- Rose could be anyone
  - Rose could be working in Benefits or payroll or marketing.
  - She could be a VP, manager, data entry operator, secretary, nurse, lab technician

Wall Street Journal (2/13/2006)

Technology

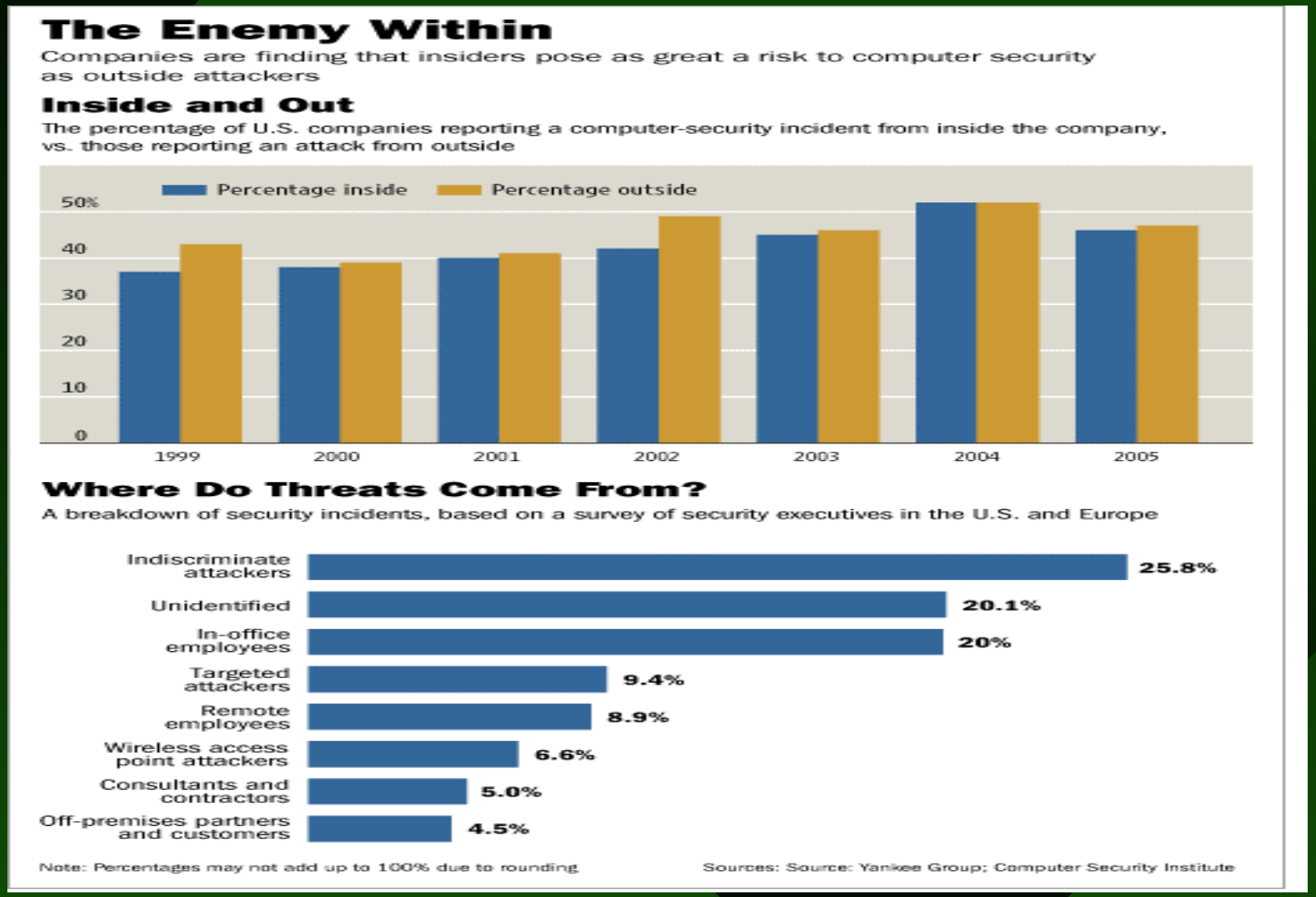
THE JOURNAL REPORT

THE WALL STREET JOURNAL.

The Dangers Within

The biggest threats to information security often don't come from hackers. They come from a company's own employees. Here's how you can stop them.

Wall Street Journal (2/13/2006)



We want …

- Rose to have access to the data to perform legitimate analyses, but prevent Rose from gaining any unauthorized information about individual records in the data set

But, de-identification is not enough

- Just removing all identifiers (such as Name, Address, SS#, etc.) from the data prior to release is insufficient in most cases
  - Using a combination of characteristics, we could possibly identify an individual in a data set

Re-identifying

Using Categorical data

- One common re-identification procedure is to identify someone using their demographic characteristics
  - Race, Gender, Profession, Education
  - Age

Using Numerical data

- Numerical information often uniquely identifies an individual. The greater the number of numerical variables, the greater the probability that an individual has a unique set of values.
- Numerical data may pose a greater threat to privacy than categorical data

Possible solutions

- Release no data
- Release subsets of the data
  - Limited to demographic variables
  - Limited to only a few numeric variables
  - Limited to a few records
- All of these inherently reduce the analytical value of the data. In many cases, the data that is released is practically useless for analysis purposes.

Data Masking

- Techniques intended to facilitate the sharing and dissemination of useful data without compromising privacy or confidentiality
- Data masking protects the data from snoopers

Data Masking Using General Additive Data Perturbation (GADP)

- GADP - A methodology for modifying the values of numerical confidential attributes such that
  - Disclosure risk is minimized (privacy maximized)
  - The mean and covariance of the original and perturbed data are identical (usefulness)
- Powerful because the vast majority of statistical analysis rely primarily on the mean and covariance.

Masked Data Set

Number of Office Visits	Number of Lab Visits	Masked Lab Charges	Masked Total charges	Masked Diagnosis
4	1	$1,581.78	$2,808.65	Negative
4	1	$1,469.28	$4,544.06	Positive
4	1	$1,628.59	$3,978.08	Negative
3	1	$1,677.97	$3,555.93	Negative
4	2	$914.89	$2,885.16	Negative
2	1	$1,859.75	$3,598.42	Negative
3	1	$1,541.97	$3,165.00	Positive
4	2	$1,203.69	$1,787.76	Negative
2	1	$1,563.13	$3,269.09	Negative
4	1	$226.55	$1,709.03	Negative
4	1	$2,065.44	$3,472.92	Negative
2	1	$330.54	$1,311.83	Negative
3	2	$1,003.51	$2,652.52	Negative
4	1	$1,287.79	$4,043.78	Negative
4	1	$2,074.09	$4,501.63	Negative
4	2	$1,871.36	$3,894.80	Negative
4	1	$2,408.50	$4,239.16	Positive
3	2	$2,267.20	$3,332.90	Negative
2	1	$1,210.50	$1,941.71	Negative
2	1	$1,394.19	$2,418.85	Negative

- For the purposes of this illustration, we have left the order of the patients the same as in the original data set. In practice, we would randomly sort the data prior to release.

Rose wants to find out if …

- If a particular patient, Joe Schmo (Patient# 61779850), tested positive. Rose know that Joe Schmo was charged $1784.11 for Lab tests and $3519.64 total (office visits and lab tests) from this department
- But there are no such amounts in the data set!
- Rose cannot identify this individual from the masked data set

What if Rose attempts to identify Joe Schmo by…

- Finding the lab charges closest to $1784.11 in the masked data set?
  - The closest value in the masked data set is $1859.75 which DOES NOT BELONG TO JOE SCHMO
  - She would incorrectly identity Patient # 39205467
- Finding the total charges closest to $3519.64 in the masked data set?
  - The closest value in the masked data set is $3472.92 which DOES NOT BELONG TO JOE SCHMO
  - She would incorrectly identity Patient # 53140817

Aha … How about …

- Computing the office visit charges (difference between total charges and lab charges) closest to $1735.53 ($3519.64 – $1784.11) in the masked data set?
  - The closest value in the masked data set is $1623.03 ($3165.00 – $1541.97) which DOES NOT BELONG TO JOE SCHMO
  - She would incorrectly identity Patient# 47318787

Rose is very frustrated …

- She has tried three different approaches and has got three different results. If Joe Schmo is one of these three which one is he? More importantly, is Joe Schmo even one of these three? What if he is someone else?
- We have achieved our disclosure risk objective.

Disclosure Risk is minimized

- GADP minimizes disclosure risk (and maximizes privacy) by making the masked values "conditionally independent" of the original values
- That is, a *knowledge of the masked values provides no knowledge about the original values*
- For illustration purposes, we have left number of office and lab visits unmodified, but they could also be masked

But what of Rose the analyst?

- The masked data maintains a high level of usefulness
  - Responses to many typical questions using the masked data will be identical as that using the original data
  - Very small differences may be observed because of rounding

What is the average total charge for a patient who tested positive?

Diagnosis	Average of Total Cost
For patients Testing Negative	3009.59
For Patients Testing Positive	3982.74
Overall Average	3155.56

Original Data

Diagnosis	Average of Masked Total Cost
For patients Testing Negative	3009.59
For Patients Testing Positive	3982.74
Overall Average	3155.56

Masked Data

What is breakdown of average lab costs by number of visits?

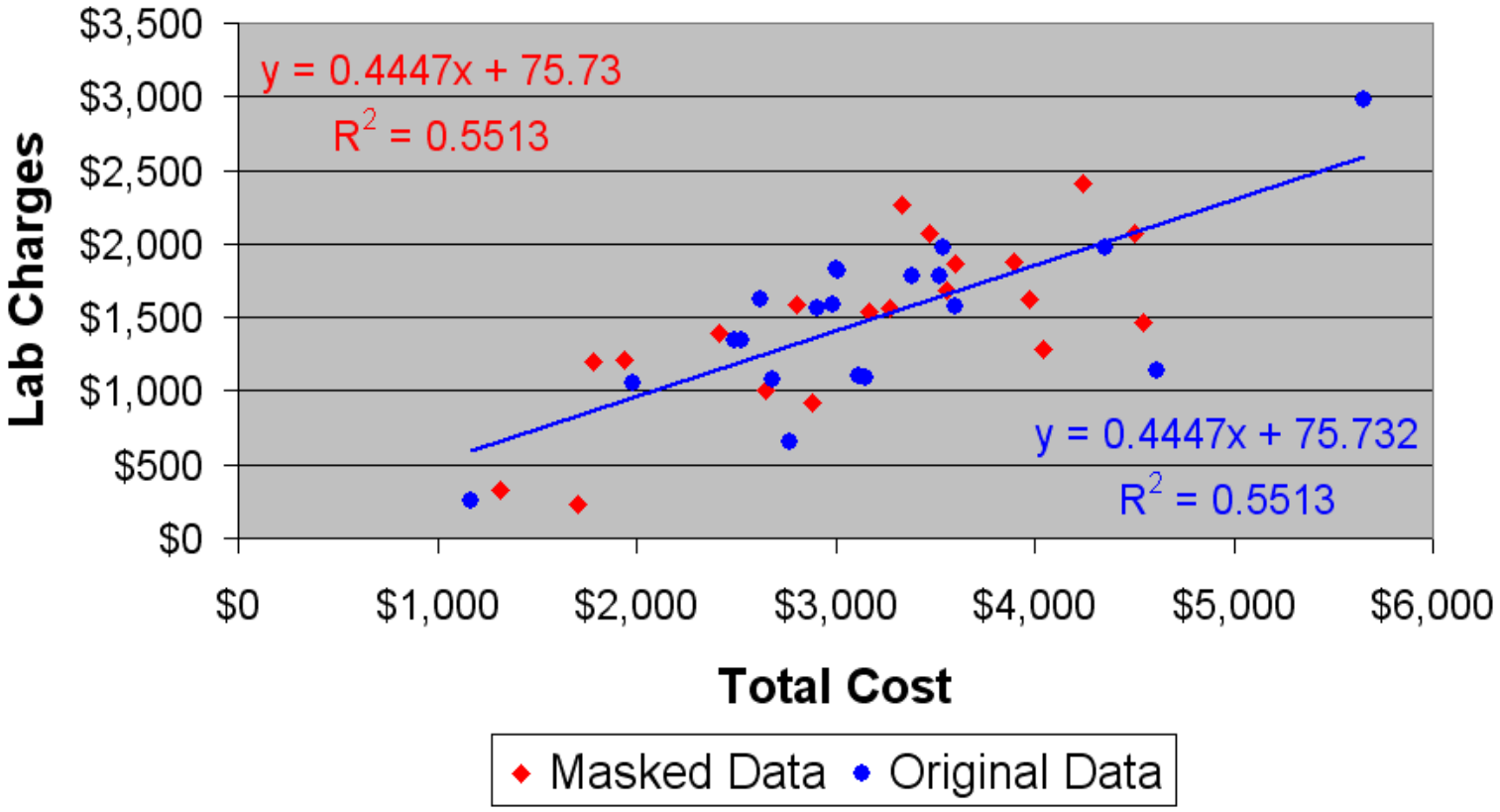
Number of Lab Visits	Average of Lab Charges
1	1488.00
2	1452.13
Overall Average	1479.04

Original Data

Number of Lab Visits	Average of Masked Lab Charges
1	1488.00
2	1452.13
Overall Average	1479.04

Masked Data

Relationship between Total Cost and Lab Charges



Implications for Practice

- Useful data can be shared without compromising privacy or confidentiality
- The accounting department should be informed about the masking
- Disclaimer should be added when the data is shared with external entities

  "The results of from analyzing this data would be identical to the results using the original data for those statistical procedures for which the mean vector and covariance matrix are sufficient statistics"

Secure Data sharing between Organizations

- XYZ Corporation requests data from their healthcare provider Goodhealth regarding their employees to see if they provide additional benefits to employees.
- The specific data is the amount of un-reimbursed medical expenses and prescription drug expenses. They have also requested Gender and information on whether these individuals have supplementary insurance
- The primary purpose for which this data is being requested is legitimate, even laudable
- Goodhealth must still worry about the data being used for illegitimate purposes. They have to protect the data

The Data (subset)

Insurance ID #	Gender	Supplemental Insurance?	Unreimbursed Prescription Drug Expenses	Unreimbursed Medical Expenses
1328305	Male	No	$500.46	$1,155.83
1409893	Male	Yes	$493.92	$1,377.02
1813206	Male	Yes	$634.32	$1,299.13
2444802	Female	No	$508.84	$1,354.33
2467321	Male	No	$525.22	$1,107.10
2517537	Male	Yes	$526.76	$1,235.12
2684770	Female	No	$594.36	$1,181.63
2870951	Female	No	$443.69	$1,300.61
2978232	Female	No	$597.20	$1,465.07
3333418	Male	No	$546.01	$1,497.05
3676413	Male	No	$480.45	$1,316.39
3859648	Female	Yes	$378.61	$1,219.56
4065955	Male	Yes	$518.33	$1,521.56
4281840	Female	Yes	$471.69	$1,028.49
4328143	Male	No	$503.79	$1,330.52
5388916	Male	Yes	$485.52	$872.84
5404673	Female	Yes	$368.55	$1,252.72
5537339	Male	No	$676.18	$1,652.26

XYZ Corporation's Perspective

- The availability of the data could have significant policy implications. Analyzing this data could provide important insights that will allow them to provide better coverage.
- Impacts
  - The organization (cost savings)
  - The employees (better coverage)
  - Society

GoodHealth's Perspective

- Allowing XYZ corp. to analyze the data may result in
  - Improved relations with XYZ Corp & its employees
  - Potential cost savings
  - Social benefits

But …

- If the data is misused and disclosure occurs, it has the potential to cause harm to
  - The employees
  - Both organizations
  - Society

Who actually makes the decision?

- Currently, the data sharing decisions are most often driven by technical issues (fear of disclosure)
- Decisions are based on suggestions of “technicians” who often err on the side of caution. This results in either not sharing valuable data or sharing data with reduced value
- Sacrifices benefits due to perceived privacy risks
- Our masking methods allow policy makers to make decisions that focus on benefits, while retaining a high level of privacy
- Decision should be made by evaluating the costs and benefits from sharing the data, while being assured of privacy

Goodhealth de-identifies the data

Data without Identifiers (subset)

Gender	Supplemental Insurance?	Unreimbursed Prescription Drug Expenses	Unreimbursed Medical Expenses
Male	No	$500.46	$1,155.83
Male	Yes	$493.92	$1,377.02
Male	Yes	$634.32	$1,299.13
Female	No	$508.84	$1,354.33
Male	No	$525.22	$1,107.10
Male	Yes	$526.76	$1,235.12
Female	No	$594.36	$1,181.63
Female	No	$443.69	$1,300.61
Female	No	$597.20	$1,465.07
Male	No	$546.01	$1,497.05
Male	No	$480.45	$1,316.39
Female	Yes	$378.61	$1,219.56
Male	Yes	$518.33	$1,521.56
Female	Yes	$471.69	$1,028.49
Male	No	$503.79	$1,330.52
Male	Yes	$485.52	$872.84
Female	Yes	$368.55	$1,252.72
Male	No	$676.18	$1,652.26

But Goodhealth knows…

- Even if the data is released without identifiers, it may still be possible to identify some or all individuals in the released data based on
  - Information from the supplemental insurer or medical reimbursement account
  - Survey of employees within the organization

So Goodhealth Masks the Data…

  - For this illustration Gender and Supplemental Insurance information remain unmodified

..and shares the Masked Data (same subset as the original data)

Gender	Supplemental Insurance?	Unreimbursed Prescription Drug Expenses	Unreimbursed Medical Expenses
Male	No	$482.67	$1,424.78
Male	Yes	$359.06	$1,071.86
Male	Yes	$624.47	$1,301.06
Female	No	$434.07	$1,335.30
Male	No	$665.40	$1,929.30
Male	Yes	$371.39	$1,102.37
Female	No	$491.96	$1,294.86
Female	No	$497.09	$1,271.22
Female	No	$555.71	$1,341.88
Male	No	$476.31	$878.05
Male	No	$397.87	$1,099.87
Female	Yes	$412.86	$950.93
Male	Yes	$561.08	$1,121.90
Female	Yes	$587.63	$1,093.22
Male	No	$509.79	$1,093.16
Male	Yes	$475.66	$1,119.50
Female	Yes	$539.06	$1,351.51
Male	No	$611.77	$1,299.23

How useful is the data to XYZ?

Aggregate Summary Statistics

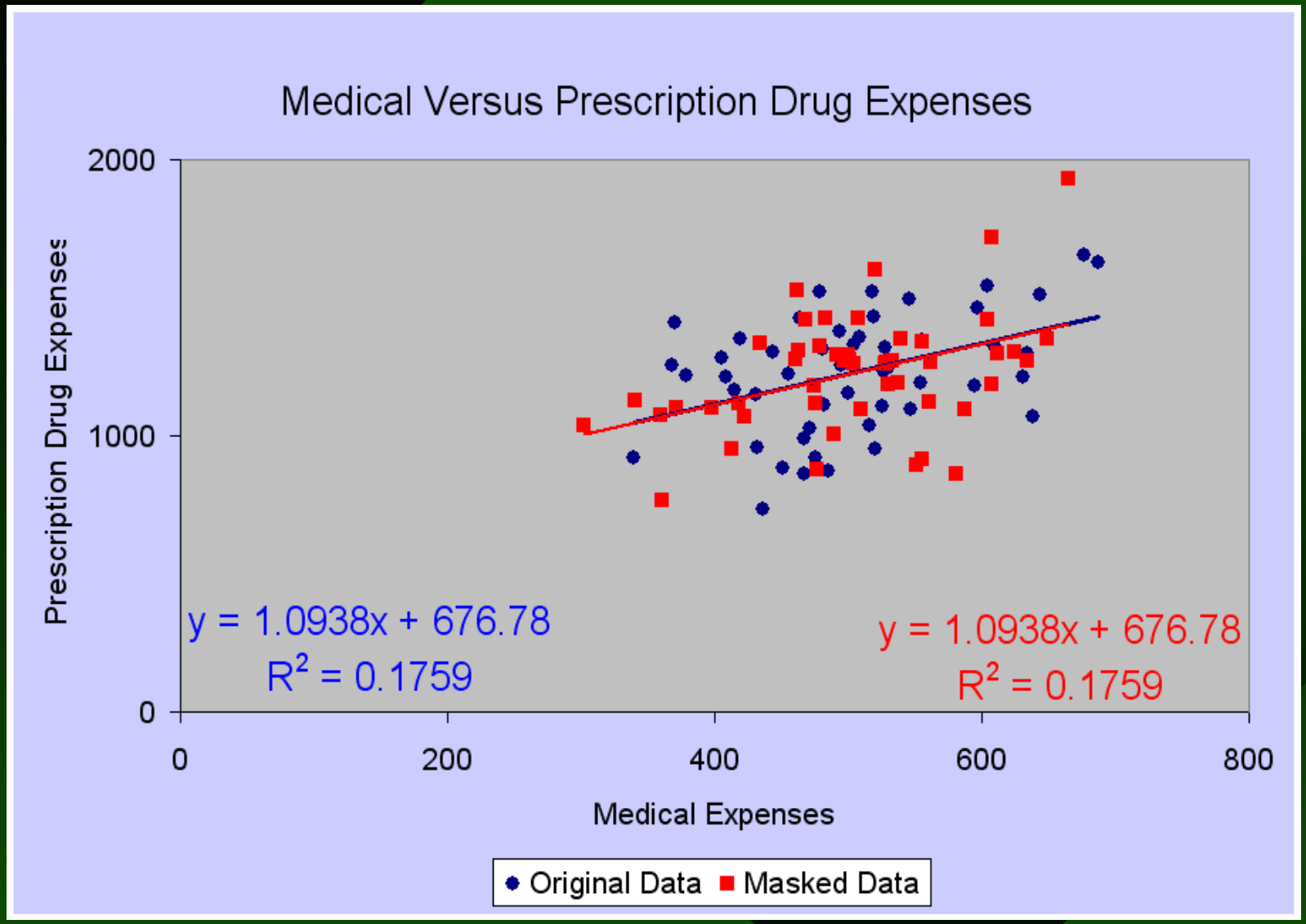
		Original Data	Masked Data
Unreimbursed Prescription Drug Expenses	Mean	$504.59	$504.59
	Standard Deviation	$81.96	$81.96
Unreimbursed Medical Expenses	Mean	$1,228.71	$1,228.71
	Standard Deviation	$213.73	$213.73

Is the average medical expenses of those who have supplemental insurance statistically different from those who do not?

Results of ANOVA Procedure (Original Data)						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	189558	1	189557.92	4.441013	0.040335	4.042647
Within Groups	2048808	48	42683.49			
Total	2238366	49				

Results of ANOVA Procedure (Masked Data)						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	189558	1	189557.92	4.441013	0.040335	4.042647
Within Groups	2048808	48	42683.49			
Total	2238366	49				

Using both data sets, we reach the same conclusion … "Yes there is a statistically significant difference in the average medical expenses between those who have supplemental insurance and those who do not".



Very Useful …

- With our perturbation procedure, the results of any statistical analyses for which the mean vector and covariance matrix are sufficient statistics using the masked data would be IDENTICAL to the results using the original data
- More importantly, decisions based on the results of analyzing the masked data will be the same as that using the original data

..and also Secure!

- The procedure minimizes risk of disclosure since the original and perturbed values are (conditionally) independent of one another
- In simple terms … It is practically impossible for snoopers to identity individuals and predict the original values, using the masked values, even with advanced record linkage procedures

Implications for Policy Makes

- There is a viable alternative to not sharing the data
- You may ..
  - Decide to share the data
  - Decide to share some segments of the data
  - Decide not to share the data
- *But the decision can be based on the relative costs and benefits and not based exclusively on the fear of disclosure*

What about other types of analyses?

- The perturbation procedure is not effective when non-traditional types of analyses are to be performed on the data
- We need other procedures in cases where the analyses to be performed is non-traditional (such as data mining).

Other Data Masking Approaches

- Data Shuffling
  - A procedure where the original values of the confidential variables are “switched” among the observations.
  - Maintains the original values of the variables, linear and monotonic non-linear relationships, and minimizes disclosure risk
  - Results of analyses are very similar (but not identical to) the analyses performed on the original data

An Example Data Set

Identifier Key	Variable 1	Variable 2
1	63.09	4240.09
2	25.64	2978.64
3	38.43	3449.14
4	81.00	4997.88
5	25.49	2704.33
6	80.49	4164.08
7	97.74	4177.89
8	87.68	4263.78
9	60.12	3963.70
10	61.15	4581.37

Shuffling

- Assign the original values to different records in the data set
  - Maintain relationships
  - Minimize disclosure

Shuffled Variable 1	Shuffled Variable 2
38.43	2704.33
61.15	4177.89
81.00	3963.70
87.68	4997.88
97.74	4263.78
25.49	2978.64
63.09	4164.08
80.49	4581.37
60.12	4240.09
25.64	3449.14

Can we identify the shuffled values?

Identifier Key	Variable 1	Variable 2
1	63.09	4240.09
2	25.64	2978.64
3	38.43	3449.14
4	81.00	4997.88
5	25.49	2704.33
6	80.49	4164.08
7	97.74	4177.89
8	87.68	4263.78
9	60.12	3963.70
10	61.15	4581.37

Identifier Key	Shuffled Variable 1	Shuffled Variable 2
?	38.43	2704.33
?	61.15	4177.89
?	81.00	3963.70
?	87.68	4997.88
?	97.74	4263.78
?	25.49	2978.64
?	63.09	4164.08
?	80.49	4581.37
?	60.12	4240.09
?	25.64	3449.14

No …

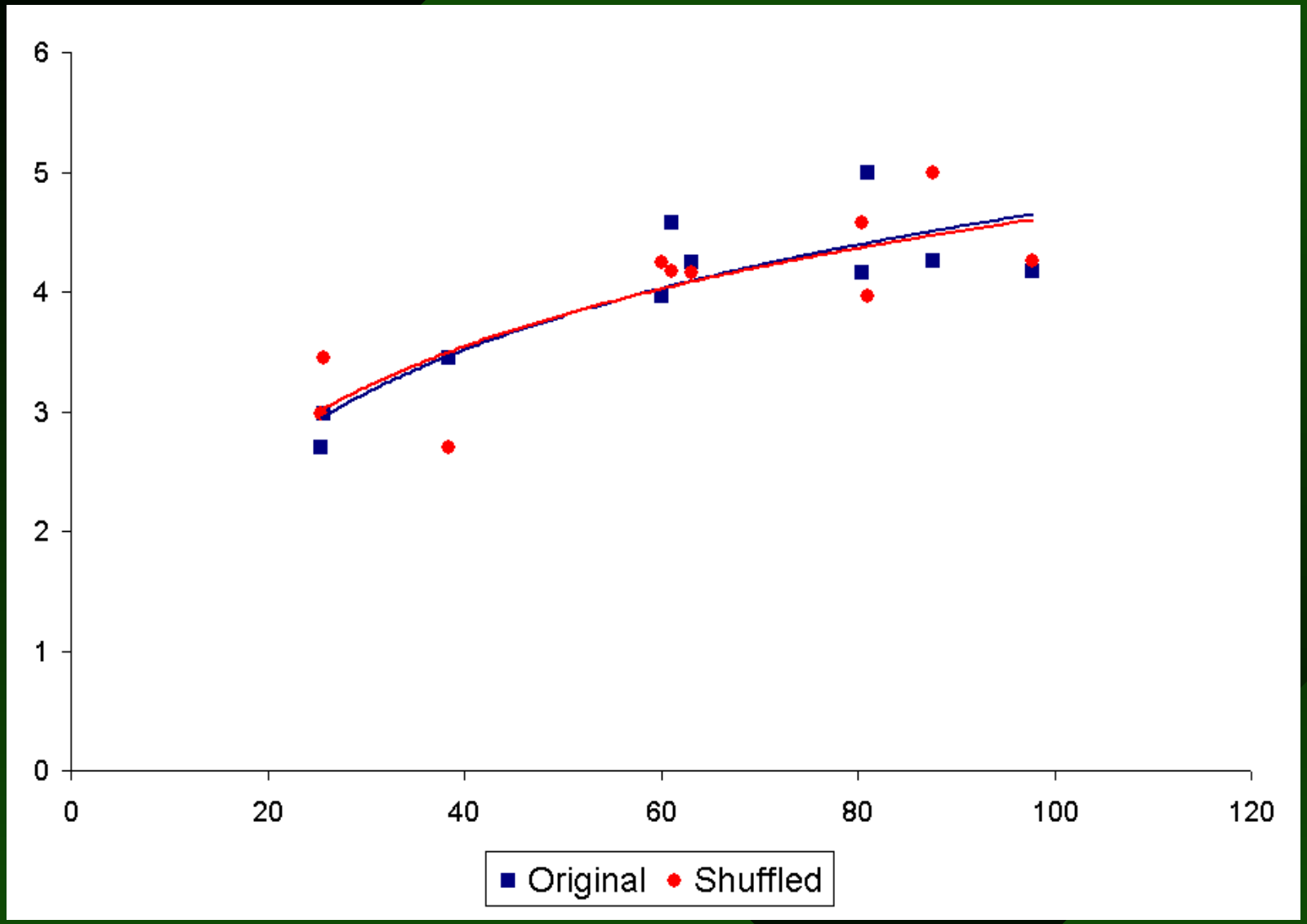
Identifier Key	Variable 1	Variable 2
1	63.09	4240.09
2	25.64	2978.64
3	38.43	3449.14
4	81.00	4997.88
5	25.49	2704.33
6	80.49	4164.08
7	97.74	4177.89
8	87.68	4263.78
9	60.12	3963.70
10	61.15	4581.37

Identifier Key	Shuffled Variable 1	Shuffled Variable 2
9	38.43	2704.33
6	61.15	4177.89
2	81.00	3963.70
7	87.68	4997.88
10	97.74	4263.78
3	25.49	2978.64
8	63.09	4164.08
1	80.49	4581.37
4	60.12	4240.09
5	25.64	3449.14

Is it useful …

- It might seem that we have simply rearranged the data randomly
- Not true … the data have been rearranged in such a fashion that they maintain relationships

Identifier Key	Rank of Variable 1	Rank of Variable 2	Rank of Shuffled Variable 2	Rank of Shuffled Variable 2
1	6	7	7	9
2	2	2	8	4
3	3	3	1	2
4	8	10	4	7
5	1	1	2	3
6	7	5	5	6
7	10	6	9	10
8	9	8	6	5
9	4	4	3	1
10	5	9	10	8
Rank order Correlation	**0.745455**		**0.745455**	



Shuffling is useful and secure

- Even for small data sets, we can develop an effective shuffled data set
  - Individual values are unmodified
  - Maintains non-linear relationships (in addition to linear relationships)
  - Minimizes disclosure risk
  - Effectiveness improves with the size of the data set

GADP or Shuffling?

- Both GADP and Shuffling are secure because both are based on the conditional distribution approach.
- Guarantees that the extent of disclosure is minimized to the information provided by the non-confidential variables...the masked values provide no additional information
- We have confirmed their theoretical disclosure characteristics using sophisticated tools such as record-linkage and canonical correlation
- GADP can be used for a vast majority of statistical analysis
- Shuffling can be used for more complex analysis

Have your cake and eat it too!

- Conventional wisdom dictates that it is not possible to share or disseminate data without compromising usefulness and/or privacy and/or confidentiality
- This is not necessarily true. For specific types of analyses, data can be shared or disseminated without compromising privacy or confidentiality

Sample of Our Work in this Area

- Muralidhar, K. and R. Sarathy, "Data Shuffling: A New Procedure for Masking Numerical Data," *Management Science* (Forthcoming).
- Sarathy, R. and K. Muralidhar, "Secure and Useful Data Sharing," Decision Support Systems (Forthcoming).
- Muralidhar, K. and R. Sarathy, "A Theoretical Basis for Perturbation Methods," Statistics and Computing, 13(4), 329-335, 2003.
- Sarathy, R. and K. Muralidhar, "The Security of Numerical Confidential Data in Databases," Information Systems Research, 13(4), 389-403, 2002.
- Muralidhar, K., R. Parsa, and R. Sarathy, "An Improved Security Requirement for Data Perturbation with Implications for E-commerce," Decision Sciences, 32(4), 683-698, 2001.
- Muralidhar, K. and R. Sarathy, "Security of Random Data Perturbation Methods," ACM Transactions on Database Systems, 24(4), 487-493, 1999.
- Muralidhar, K., R. Parsa, and R. Sarathy, "A General Additive Data Perturbation Method for Database Security," Management Science, 45(10), 1399-1415, 1999.

Please Visit:

http://gatton.uky.edu/faculty/muralidhar/maskingpapers/

for more details of these and other procedures as well as actual data sets that illustrate the application of these procedures