Re-identification Risk in Big Data in Health Care

Anand N. Vidyashankar
Department of Statistics, George Mason University

Joint work with Lei Li and Lucy J. Doyle

July 27, 2017

Contents

- Background on Health Care
- Regulatory Issues: HIPAA and HITECH
- Risk Measures
- Risk Estimation
- Simulation Study
- CMS Data Analysis
- Concluding Remarks

Background on Health Care

- Data confidentiality is an important area of scientific research within the field of statistics.
- In the health care industry, preserving the privacy of patient information is an important social and a legal issue.
- Current federal regulations in the United States, Canada, and Europe have had an enormous impact on businesses and agencies hosting patient level data.

Regulatory Issues: HIPAA and HITECH

- In the United States, the Health Insurance Portability and Accountability Act (HIPAA) was enacted into a public law, PL 104-191 on August 21, 1996.
- This law, which has both privacy and security components, mandates that agencies and businesses scientifically establish that the data are indeed 'confidential". Of course, it is impossible to separate the security and privacy aspects from one another. For instance, if a hacker breaches security and gains access to confidential data, then privacy has also been violated.

HIPAA and HITECH (contd.)

- The Health Information Technology for Economic and Clinical Health Act (HITECH Act) legislation was enacted in 2009 to stimulate the adoption of electronic health records (EHR) and supporting technology in the United States. HITECH became a law on Feb. 17, 2009, as part of the American Recovery and Reinvestment Act of 2009 (ARRA) economic stimulus bill.

- HITECH and HIPAA are separate and unrelated laws, but they do reinforce each other. For example, HITECH stipulates that technologies and technology standards created under HITECH do not compromise HIPAA privacy and security laws.

HIPAA and HITECH (contd.)

- It also requires that any physician and hospital that attests to meaningful use must also have performed a HIPAA security risk assessment as outlined in the "Omnibus rule," or 2013 digital update to the original 1996 law.

- HITECH established data breach notification rules; HIPAA's omnibus update echoes those rules and adds details such as holding health care providers' business associates accountable for the same liability of data breaches as the providers themselves.

Health Care Issues

- Health care organizations share information to improve patient care and produce products of services for increasing the efficiency of care.

Data Feeds

Typically encountered data feeds in health care that are shared are

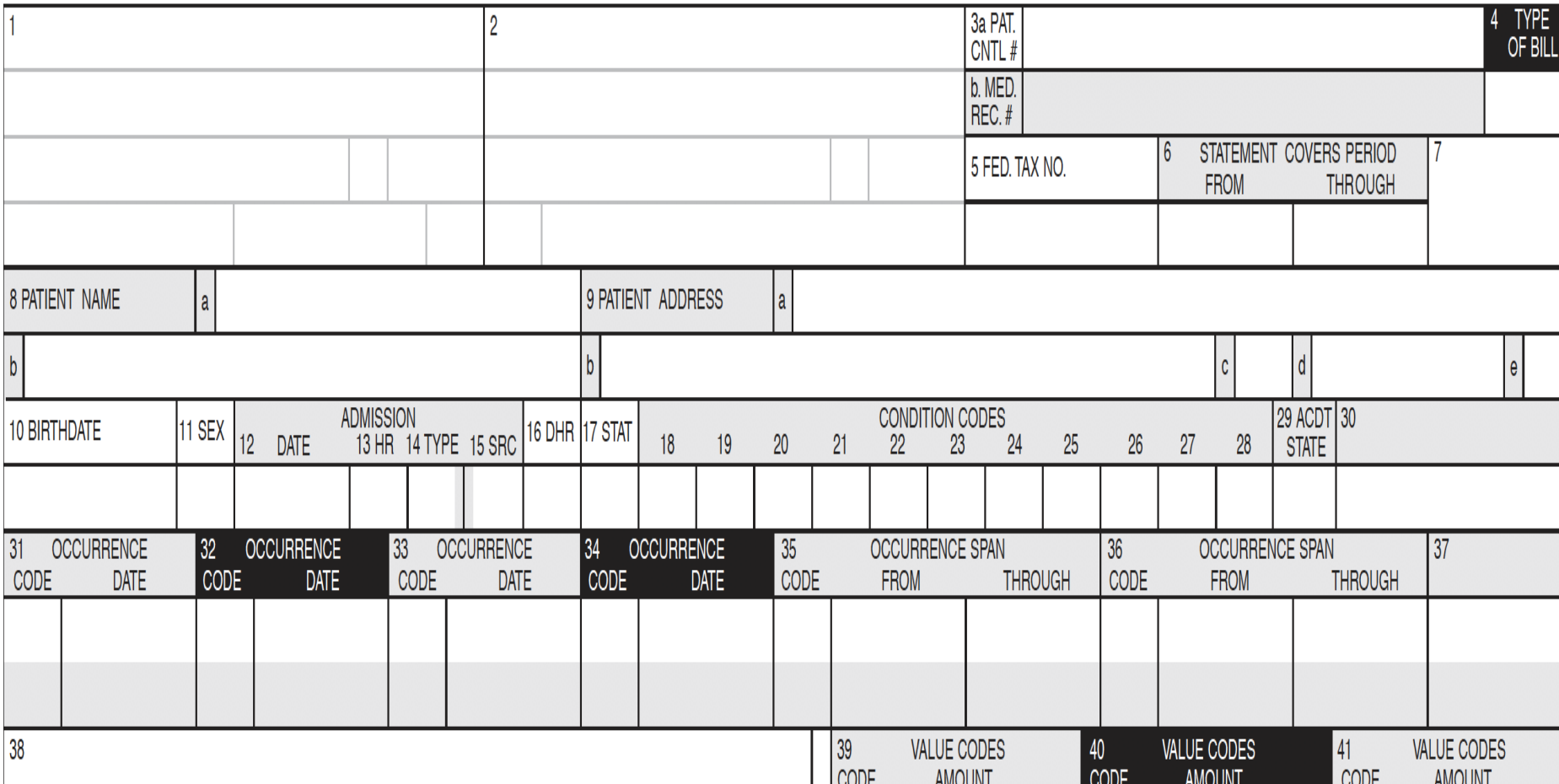
- UB04 Form
- HCFA 1500 Form
- NCPDP 10.6 Form

Data Feeds (contd.)

- In all these data feeds almost all direct identifiers are removed or encrypted.
- Our goal is to identify risk from the indirect identifiers.
- Of course, this depends on what kinds of information are shared.

Example of UB04 Form

Figure: Example of UB04 Form



Data Feeds (contd.)

- UB04 form collects detailed information about every patient, there are 81 categories in total.
- HCFA 1500 form has 33 categories.
- NCPDP form has 17 distinct categories.

Example of CMS Data

- Medicare Provider Utilization and Payment Data contains information about provider ID, name, gender, zip code and number of services, number of beneficiaries, payment amount, etc.

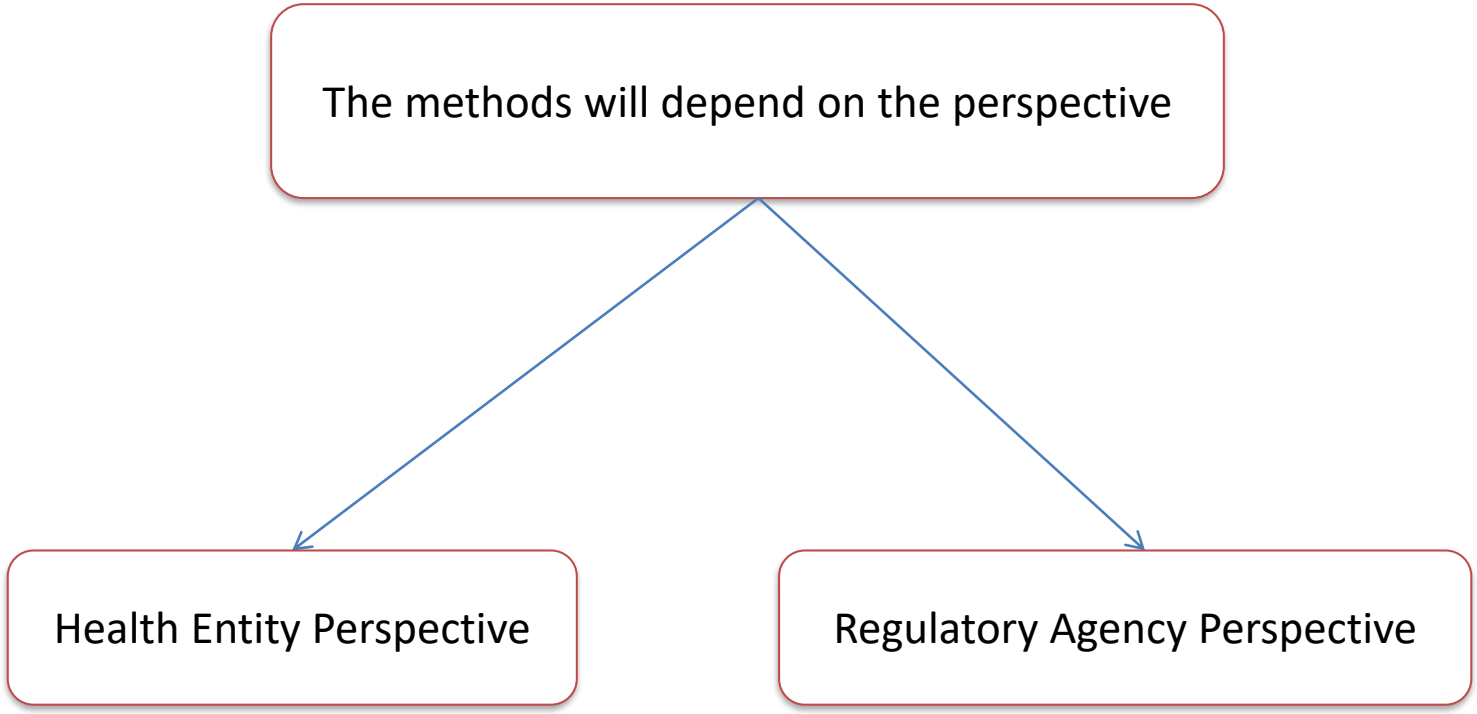
Goal

- In today's presentation, we will describe how to estimate risk in these varied data sets.

Steps for Risk Estimation

- 1. Identify the unprotected variables that can be combined or linked with other information to identify patients (Mining step).
- 2. Use an appropriate risk measure.
- 3. Evaluate the statistical properties of the risk measures.
- 4. Estimate the risk and provide confidence intervals.
- 5. Facilitate decision making.

Steps for Risk Estimation (contd.)



Identifying Keys

- Keys represent categories of records. They are based on either demographic or health status or disease information.
- For, example, age group can be split into 0-20, 21-30, ..., > 70, which has 5 keys.

Identifying Keys (contd.)

- Understanding of health care data is critical to identifying keys.
- Some understanding of the disease epidemiology provides more information.

Risk Measure for Aggregated Data

- Let J be the number of keys, which is known.
- Let F_{ij} be the number of people in *ith* record *jth* key in the population, $i = 1, \ldots, N$, $j = 1, \ldots, J$. So the data structure has this format.

$$\boldsymbol{F} = \begin{pmatrix} F_{11} & F_{12} & \ldots & F_{1J} \\ F_{21} & F_{22} & \ldots & F_{2J} \\ \ldots & \ldots & \ldots & \ldots \\ F_{N1} & F_{N2} & \ldots & F_{NJ} \end{pmatrix}$$

Risk Measures for Aggregated Data (contd.)

- Each record can be treated as a multiway contingency table, and we assume conditionally independence across different records.
- Each record may represent patient, zip code, disease, time, etc.

Risk Measures for Aggregated Data (contd.)

- Similarly, let f_{ij} be the number of people in *ith* record *jth* key in the sample, $i = 1, \ldots, n$, $j = 1, \ldots, J$. So the data structure has this format.
- $$\boldsymbol{f} = \begin{pmatrix} f_{11} & f_{12} & \ldots & f_{1J} \\ f_{21} & f_{22} & \ldots & f_{2J} \\ \ldots & \ldots & \ldots & \ldots \\ f_{n1} & f_{n2} & \ldots & f_{nJ} \end{pmatrix}$$
- f_{ij} depends on the sampling mechanism.

Risk Measures for Aggregated Data (contd.)

- Health Entity Perspective:

$$R_1 = R_p(r) = \sum_{j=1}^{J} P\left(1 \leq F_j \leq r\right),$$

  where r is a small number.

- R_1 measures the probability that the population has very few patients/subjects within each key.

- $r = 1$, it is referred to as population unique in the census literature.

Risk Measures for Aggregated Data (contd.)

- Regulatory Agency Perspective:

$$R_2 = \sum_{j=1}^{J} E\left(\frac{1}{F_j}|f_j\right).$$

- R_2 measures the posterior risk of re-identification given the sample size in the released information for the j^{th} key.

Risk Measures for Sharing of Records

- $$R_3 = Pr\left(PU\middle|SU\right) = \frac{\sum_{j=1}^{J} P(F_j = 1, f_j = 1)}{\sum_{j=1}^{J} P(f_j = 1)}$$

  where PU refers population unique, and SU refers sample unique.

- $$R_4 = P(CM\middle|UM) = \frac{\sum_{j=1}^{J} P(f_j = 1)}{\sum_{j=1}^{J} F_j P(f_j = 1)}$$

  where CM refers correct match, and UM refers unique match.

Brief Literature Review

In the context of census data,

- Bethlehem, J. G.(JASA, 1990) considered Poisson Gamma(PG) model to fit counts and identify the risk.

- Elsayed and Skinner (JOS, 2006) assumes PG log linear model to calculate the risk $E(\frac{1}{F_j}|f_j = 1)$, where $f_j|\lambda_j \sim Poi(\lambda_j)$.

Brief Literature Review (contd.)

- C.J Skinner and M.J.Elliot (JRSS(B), 2002) considered record based mechanism risk measure, also they assume poisson model for population and binomial resampling mechanism for sample.

Proposed Models for Risks

- We allow correlation amongst different keys.
- We allow different means amongst different keys.

Proposed Models for Risks (contd.)

- Modified Poisson Gamma(PG) model:

$$F_{ij}|\lambda_i, \theta_{ij} \overset{i.i.d}{\sim} Poi(\lambda_i \theta_{ij})$$
$$\theta_{ij} \overset{i.i.d}{\sim} Ber(p_j)$$
$$\lambda_i \overset{i.i.d}{\sim} Gamma(\alpha, \beta)$$
$$i = 1, \ldots, N;\ j = 1, \ldots, J$$

Proposed Models for Risks (contd.)

- Modified Poisson Lognormal(PL) model:

$$F_{ij}|\lambda_i, \theta_{ij} \overset{i.i.d}{\sim} Poi(\lambda_i \theta_{ij})$$
$$\theta_{ij} \overset{i.i.d}{\sim} Ber(p_j)$$
$$\lambda_i \overset{i.i.d}{\sim} Lognormal(\mu, \sigma^2)$$
$$i = 1, \ldots, N;\ j = 1, \ldots, J$$

Proposed Models for Risks (contd.)

- Modified Negative Binomial Lognormal(NBL) model:

$$F_{ij}|\lambda_i, \theta_{ij} \overset{i.i.d}{\sim} NB(r, \theta_{ij} q_i)$$

$$\theta_{ij} \overset{i.i.d}{\sim} Ber(p_j)$$

$$log \frac{rq_i}{1 - q_i} \overset{i.i.d}{\sim} N(\mu, \sigma^2)$$

$$i = 1, \ldots, N; \; j = 1, \ldots, J$$

Estimation Methods

- Maximum Likelihood Estimator (MLE): $\hat{\theta} = \underset{\underset{\sim}{\theta} \in \Theta}{argmax} L(\underset{\sim}{\theta}, \boldsymbol{X})$,

  where $L(\underset{\sim}{\theta}, \boldsymbol{X})$ is the likelihood function of $\underset{\sim}{\theta}$ for given $\boldsymbol{X}$.

- Minimum Hellinger Distance Estimator (MHDE):

$$\hat{\theta} = \underset{\underset{\sim}{\theta} \in \Theta}{argmax} \int (f(x; \underset{\sim}{\theta}))^{1/2} (h_n(x))^{1/2} dx,$$

  where $h_n(x)$ is an empirical estimate of $f(x; \underset{\sim}{\theta})$.

  - MHDE facilitates model mis-specification analysis.

Simulation Study

For Modified PG model, we assume there are $J = 20$ keys, true $\alpha = 4$, true $\beta = 2$, and true p_j are generated i.i.d from $U(0, 1)$, $j = 1, \ldots, J$. iteration size 5000. $R_1(r)$ denotes the risk from health entity perspective, $R_2(\nu_j)$ denotes risk from regulatory perspective, where ν_j is the binomial parameter.

Table: Modified PG model: MLE vs MHD

		$R_1(1)$	$R_1(3)$	$R_2(\nu_j = 0.1)$	$R_2(\nu_j = 0.3)$	$R_2(\nu_j = 0.5)$
MLE	Estimator	3.326	7.707	8.881	10.76	12.94
	95% CI	(3.041, 3.419)	(7.399, 8.039)	(8.485, 9.280)	(10.40, 11.11)	(12.63, 13.25)
MHD	Estimator	3.226	7.707	9.077	10.81	12.94
	95% CI	(2.911, 3.430)	(7.483, 8.151)	(8.654, 9.467)	(10.40, 11.17)	(12.49, 13.22)

CMS Data Analysis

- Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use Files (PUF) 2015.
- Data size: 2.5 *Gbs*.
- Randomly sample 1 million records.

CMS Data Analysis (contd.)

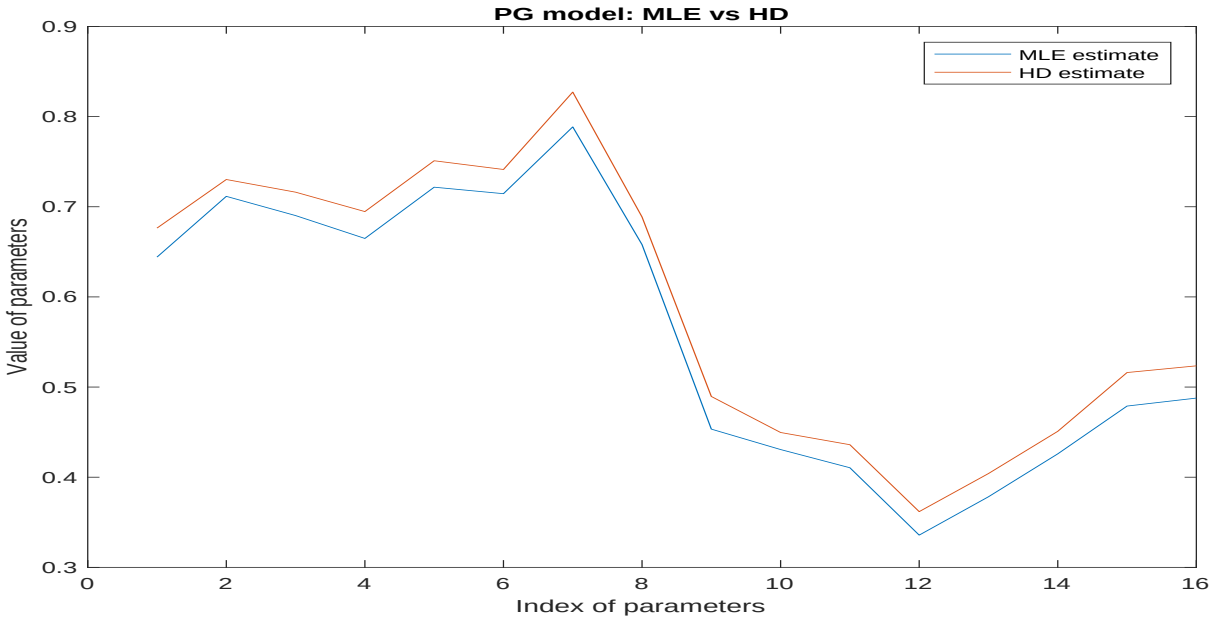
- Our focus is cardiology, which has 48148 records.
- We take key variables: number of services(split into 4 categories) and average medicare payment amount (split into 2 categories from median). Hence there are $J = 8$ keys.

CMS Data Analysis (contd.)

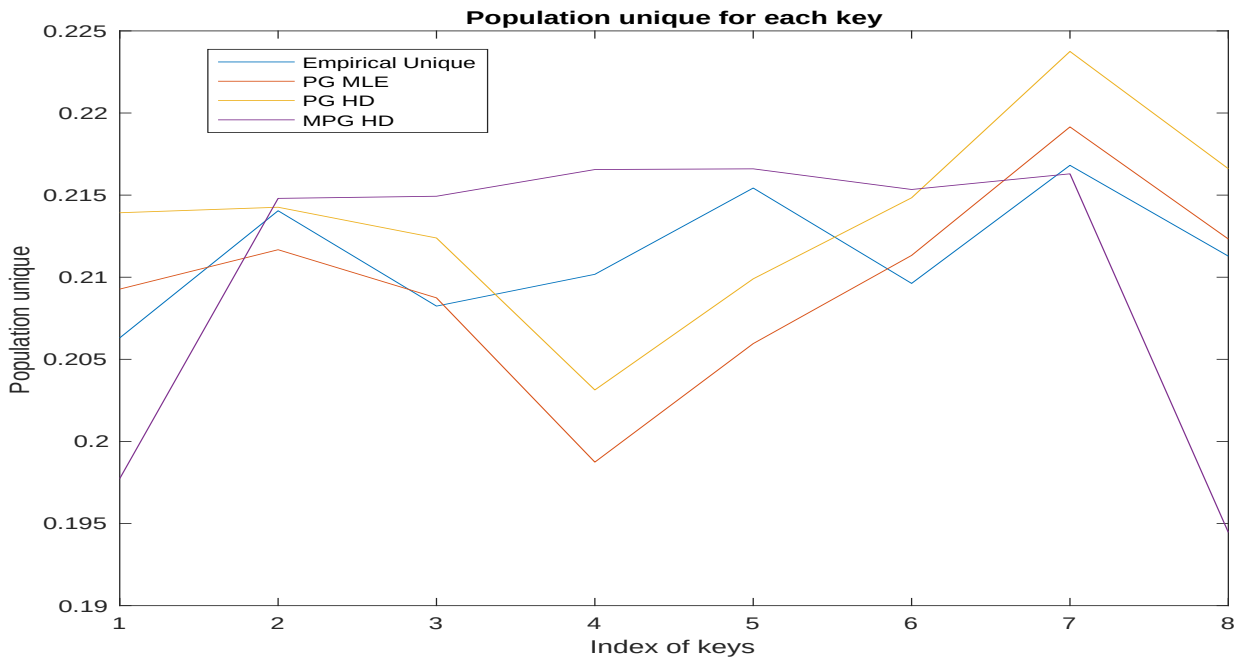
- We treat these 48148 as population records. Besides, $\boldsymbol{F}$ is a 3616×8 matrix.
- Assume the common PG model:

$$F_{ij}|\lambda_{ij} \overset{i.i.d}{\sim} Poi(\lambda_{ij})$$
$$\lambda_{ij} \overset{i.i.d}{\sim} Gamma(\alpha_j, \beta_j)$$
$$i = 1, \ldots, n;\ j = 1, \ldots, J$$

MLE and MHD Estimates



Population Unique for Each Key



CMS Data Risk Estimates

Table: Risk Estimate of CMS Data

		$R_1(1)$	$R_2(\nu_j = 0.5)$
MLE	Estimator	1.677	5.518
	95% CI	(1.658, 1.697)	(5.504, 5.531)
MHD	Estimator	1.719	5.543
	95% CI	(1.702, 1.738)	(5.529, 5.557)

Concluding Remarks

- We developed a principled approach for risk estimation and decision making within health care organization.
- We introduced risk metrics that take into account the kind of data that are being shared.
- Our methods are theoretically justified.
- Our methods allow policy development within an health care organization.

Thank you!