Privacy and Identifiability in Clinical Research, Personalized Medicine, and Public Health Surveillance

Christopher Cassa

Children’s Hospital Informatics Program
Harvard-MIT Division of Health Sciences and Technology
Transmission of PHI Pervasive

- Pervasive in research, medicine, and public health investigations, posing risk to privacy
- Disclose identity, medical conditions, and hereditary data

Balance between privacy and research and public health
Publications Cited


• Cassa CA, Grannis SJ, Overhage M, Mandl KD. A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. *J Am Med Inform Assoc* 2006

• Wieland SC, Cassa CA, Berger B, Mandl KD. Revealing the spatial distribution of a disease while preserving privacy. *PNAS 2008* [In Review]

• Cassa CA, Iancu K, Olson KL, Mandl KD. A software tool for creating simulated outbreaks to benchmark surveillance systems. *BMC Med Inform Decis Mak.* 2005

• Cassa CA, Wieland SC, Mandl KD. Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *Int J Health Geogr.* 2008


• Brownstein JS, Cassa CA, Kohane IS, Mandl KD. An unsupervised classification method for inferring original case locations from low-resolution disease maps. *Int J Health Geogr.* 2006
Topics

1. Disclosure risks for genomic data and impact on family members

2. A) Anonymization of geospatial datasets containing patient home addresses
   B) Re-identification potential of geospatial data, commonly shared in both textual form and in printed maps
Data Rapidly Becoming Available

• Research studies publish sequencing and expression data for other investigators

• Public Studies:
  – NHLBI GWAS Framingham & Jackson Studies
  – EMR-based bio-repositories: eMERGE Network,
  – HapMap, 1000 genomes, Personal Genome Proj.

• Available to the public at large
Broad Fear of DNA Use in Society
What Protections are in Place?

• Genetic Information Non-Discrimination Act
  – Passed in 2008
• State Laws Protecting Similar Items
• Disclosure controls
Genomic Data Pose Unique Risks

• Discrimination Concerns
  – Insurance, workplace discrimination
  – Life, disability, and long term care insurance uncovered

• Genetic Knowledge and Personal Decision Making

• Implications for family members
Why Risk GATTACA?

• Correlate clinical outcomes with genomic data – likely requires large numbers of participants for common diseases
• Individual participation necessary – sharing genotypic and clinical data with investigators
• Methods to help individuals with risk assessment and to preserve privacy with such disclosures needed
Risk Disclosure Models

- Risk of Identity Linkage
- Risk of Aggregation
- Risk of Phenotypic Linkage
- Risk of Familial Linkage
Background

• Single Nucleotide Polymorphisms (SNPs) are genetic locations where at least 1% of the population has a different base pair

• SNPs distributed throughout the genome, responsible for much genetic diversity
Risk of Identity Linkage: Privacy Decreases Sharply with a Small Set of SNPs

- At a low number (35-70) of identified independent SNPs, the amount of privacy dramatically decreases.
- Match a hair or a soda can to a record.

Genomic Research and Human Subject Privacy
Zhen Lin, Art B Owen, Russ B Altman. Science. Vol.305, Iss. 5681; pg. 183
Risk of Aggregation: Combining Two Separate Genomic Datasets

- Extension of Risk of Identity Linkage
- With an overlapping set of SNPs and no supporting information, can one identify the whether two datasets came from one person?
Risk of Phenotypic Linkage: Identifying Phenotypes from Genotypes (and vice versa)

- Genomic data never “unlinked” to identity
  - Gender
  - Race/Ethnicity
  - Other physical characteristics
  - Propensity for diseases

Risk of Familial Linkage

- Siblings share 50% of contiguous chromosomal segments, and a larger fraction of genotypes.
- We share 25% of our DNA with our grandparents, aunts and uncles, and 12.5% with first cousins.

With your genomic data how many SNP values can be identified for Parents, Siblings & Children.
No Genomic Privacy: Protective Strategies Inadequate

• Using Binning to Maintain Confidentiality
• Disclosing Aggregate Data (Frequencies)
• Use of Generalization Lattices
• Adding Noise to Genetic Data
• Creating Synthetic Individuals
• Anonymization by Pool Selection
Genomic Data in Medical Records

• SNPs
• Mutations
• Full Gene, Exome and Full Genome Sequencing
• Race/Ethnicity
• Family History Data, including Genetic Diagnoses
• Phenotypic clinical data including diseases and allergies
• Gene expression Profiles
• Proteomics Data
• More to come...
Genomic Inference: Identifying Sibling Genotypes

• Improving genotype inferences using Sib$_1$ genotype at one SNP (extensible to families)
• Confirming sibling relationship given matches at sets of SNP loci
• Measuring information provided by knowledge of Sib$_1$ genotype
• Relative risk for carrying a minor allele
• Experimental results
Improving Genotype Inferences Using Sib₁ genotype at One SNP

First sib is homozygous major at SNP A

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$p^4$</td>
<td>$2p^3q$</td>
<td>$p^2q^2$</td>
</tr>
<tr>
<td>Aa</td>
<td>$2p^3q$</td>
<td>$4p^2q^2$</td>
<td>$2pq^3$</td>
</tr>
<tr>
<td>aa</td>
<td>$p^2q^2$</td>
<td>$2pq^3$</td>
<td>$q^4$</td>
</tr>
</tbody>
</table>

For example, when Sib₁ is homozygous major, all possible parental genotypic candidates that involve one or both parent genotypes of ‘aa’ are excluded, as it is not possible to have a child with genotype ‘AA’ if either parent does not have at least one copy of the ‘A’ allele.
First sib is homozygous minor at SNP A

**Mother**

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>(p^4)</td>
<td>(2p^3q)</td>
<td>(p^2q^2)</td>
</tr>
<tr>
<td>Aa</td>
<td>(2p^3q)</td>
<td>(4p^2q^2)</td>
<td>(2pq^3)</td>
</tr>
<tr>
<td>aa</td>
<td>(p^2q^2)</td>
<td>(2pq^3)</td>
<td>(q^4)</td>
</tr>
</tbody>
</table>

Eliminate Crossed Boxes and Normalize
First sib is heterozygous at SNP A

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>$p^4$</td>
<td>$2p^3q$</td>
<td>$p^2q^2$</td>
</tr>
<tr>
<td>Aa</td>
<td>$2p^3q$</td>
<td>$4p^2q^2$</td>
<td>$2pq^3$</td>
</tr>
<tr>
<td>aa</td>
<td>$p^2q^2$</td>
<td>$2pq^3$</td>
<td>$q^4$</td>
</tr>
</tbody>
</table>

Eliminate Crossed Boxes and Normalize
Calculating $p(Sib_2 AA \mid Sib_1 AA)$ for one SNP

Nine possible parental genotypic combinations ($i$) at each SNP:

$$p(Sib_2 AA \mid Sib_1 AA) = \sum_{i=1}^{9} p(Sib_2 AA \mid \text{parental comb.} \ i) p(\text{parental comb.} \ i \mid Sib_1 AA)$$

$$= \sum_{i=1}^{9} \left( \frac{p(Sib_2 AA \cap \text{parental comb.} \ i)}{p(\text{parental comb.} \ i)} \right) p(\text{parental comb.} \ i \mid Sib_1 AA)$$

$Sib_1 AA$ and $Sib_2 AA$ refer to Sib 1 and Sib 2 genotypes ‘AA’, at the SNP in question, using HapMap SNP population frequencies, $p$ and $q$ for the SNP being evaluated.
Increase in Accuracy from Sib₁ Knowledge

The red section in the overlapping Venn diagram is the improvement from knowledge of the Sib₁ genotype in making the Sib₂ genotype inference.

\[ P(AA, Aa, aa) \quad P(AA | Sib1AA) \quad P(AA) \]
Example: Calculating \( p(Sib_2AA | Sib_1AA) \)

\[
= \sum_{i=1}^{4} \left( \frac{p(Sib_2AA \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} \right) p(\text{parental comb. } i | Sib_1AA)
\]

\[
= \left( \frac{p(Sib_2AA \cap AA_MAA_F)}{p(AA_MAA_F)} \right) p(AA_MAA_F | Sib_1AA) + \left( \frac{p(Sib_2AA \cap Aa_MAA_F)}{p(Aa_MAA_F)} \right) p(Aa_MAA_F | Sib_1AA)
\]

\[
= (1)(p^2) + \left( \frac{1}{2} \right)(pq) + \left( \frac{1}{2} \right)(pq) + \left( \frac{1}{4} \right)(q^2)
\]

\[
= p^2 + pq + \frac{q^2}{4}
\]

\[
= p^2 + \left[ pq + \frac{q^2}{4} \right]
\]
Example: Calculating $p(Sib_2X|Sib_1Y)$

- Using the same technique, we can calculate all possible $p(Sib_2X|Sib_1Y)$
- Prior probability is Hardy-Weinberg equilibrium value
- Posterior includes knowledge of $Sib_1$ genotype

<table>
<thead>
<tr>
<th>$Sib_2$</th>
<th>$Sib_1$</th>
<th>Prior Prob.</th>
<th>Posterior Prob.</th>
<th>Error Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>AA</td>
<td>$p^2$</td>
<td>$p^2 + pq + \frac{1}{4}q^2$</td>
<td>$</td>
</tr>
<tr>
<td>Aa</td>
<td>AA</td>
<td>$2pq$</td>
<td>$pq + \frac{1}{2}q^2$</td>
<td>$</td>
</tr>
<tr>
<td>aa</td>
<td>AA</td>
<td>$q^2$</td>
<td>$\frac{1}{4}q^2$</td>
<td>$</td>
</tr>
<tr>
<td>AA</td>
<td>Aa</td>
<td>$p^2$</td>
<td>$\frac{1}{2}p^2 + \frac{1}{4}pq$</td>
<td>$</td>
</tr>
<tr>
<td>Aa</td>
<td>Aa</td>
<td>$2pq$</td>
<td>$\frac{1}{2}p^2 + (\frac{2}{3})^{-1}pq + \frac{1}{2}q^2$</td>
<td>$</td>
</tr>
<tr>
<td>aa</td>
<td>Aa</td>
<td>$q^2$</td>
<td>$\frac{1}{4}pq + \frac{1}{2}q^2$</td>
<td>$</td>
</tr>
<tr>
<td>AA</td>
<td>aa</td>
<td>$p^2$</td>
<td>$\frac{1}{4}p^2$</td>
<td>$</td>
</tr>
<tr>
<td>Aa</td>
<td>aa</td>
<td>$2pq$</td>
<td>$\frac{1}{2}p^2 + pq$</td>
<td>$</td>
</tr>
<tr>
<td>aa</td>
<td>aa</td>
<td>$q^2$</td>
<td>$\frac{1}{4}p^2 + pq + q^2$</td>
<td>$</td>
</tr>
</tbody>
</table>

Cassa, Kohane, Mandl, BMC Medical Genomics 2008
Avg. Error Reduction by Sib₁ Genotype

Inference Error Reduction vs. Minor Allele Frequency

- Avg. Sib1AA
- Avg. Sib1Aa
- Avg. Sib1aa

Minor Allele Frequency vs. Average Percentage Error Reduction
Measuring Information Provided by Knowledge of Sib₁ Genotype

Measure information provided by knowledge of Sib₁ genotype using the ratio between the inference with knowledge and without:

\[ \Lambda_{\text{Ind}_1,\text{Ind}_2 \text{ genotypes}} = \frac{p(\text{Ind}_2 \text{ genotype} | \text{Ind}_1 \text{ genotype } \cap \text{siblings})}{p(\text{Ind}_2 \text{ genotype} | \text{Ind}_1 \text{ genotype } \cap \text{unrelated})} \]

The log of this odds ratio can then be used as a statistic for measuring relatedness, depending only on the SNP allele frequency and the Sib₁ genotype.
Probabilistic Maneuvering

\[ \Lambda_{\text{Ind}_1, \text{Ind}_2 \text{ genotypes}} = \frac{p(\text{Ind}_2 \text{ genotype} | \text{Ind}_1 \text{ genotype } \cap \text{siblings})}{p(\text{Ind}_2 \text{ genotype} | \text{Ind}_1 \text{ genotype } \cap \text{unrelated})} \]

\[ = \frac{p(\text{Sib}_2 \text{ genotype} | \text{Sib}_1 \text{ genotype})}{\left( \frac{p(\text{Ind}_2 \text{ genotype}) \cap p(\text{Ind}_1 \text{ genotype } \cap \text{unrelated})}{p(\text{Ind}_1 \text{ genotype } \cap \text{unrelated})} \right)} \]

\[ = \sum_{i=1}^{9} \left( \frac{p(\text{Sib}_2 \text{ genotype } \cap \text{parental comb.} \cdot i)}{p(\text{parental comb.} \cdot i)} \right) \cdot \frac{p(\text{parental comb.} \cdot i | \text{Sib}_1 \text{ genotype})}{p(\text{Ind}_2 \text{ genotype})} \]

\[ = \sum_{i=1}^{9} \left( \frac{p(\text{Sib}_2 \text{ genotype } \cap \text{parental comb.} \cdot i)}{p(\text{parental comb.} \cdot i)} \right) \cdot \frac{p(\text{parental comb.} \cdot i | \text{Sib}_1 \text{ genotype})}{p(\text{Ind}_2 \text{ genotype})} \]

\[ = \underbrace{\sum_{i=1}^{9} \left( \frac{p(\text{Sib}_2 \text{ genotype } \cap \text{parental comb.} \cdot i)}{p(\text{parental comb.} \cdot i)} \right) \cdot \frac{p(\text{parental comb.} \cdot i | \text{Sib}_1 \text{ genotype})}{p(\text{Ind}_2 \text{ genotype})}}_{\text{expression}} \]
Measuring Information Provided by Knowledge of Sib$_1$aa Genotype

\[ \log(\Lambda_{\text{Ind}_1, \text{Ind}_{2aa}}) \] vs. Allele Frequency

Minor Allele Frequency

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
Measuring Information Provided by Knowledge of Sib$_1$AA Genotype

$log(\Lambda_{\text{Ind1, Ind2AA}})$ vs. Allele Frequency

Minor Allele Frequency

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
Measuring Information Provided by Knowledge of Sib$_1$Aa Genotype

\[
\log(\Lambda_{\text{Ind1, Ind2Aa}}) \text{ vs. Allele Frequency}
\]

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
Confirming Sibling Relationship
Given Matches at SNP Loci

• Probability that two people in a pool of size $N$ are siblings calculated using a version of Bayes’ Theorem, if they have matching alleles at $M$ independent SNP loci.

\[
p(sibs|\text{match at } M \text{ loci}) = \frac{p(\text{match at } M \text{ loci}|sibs) p(sibs)}{p(\text{match at } M \text{ loci}|sibs)p(sibs) + p(\text{match at } M \text{ loci}|\neg sibs)p(\neg sibs)}
\]

\[
= \frac{[p(\text{both } AA|sibs) + p(\text{both } Aa|sibs) + p(\text{both } aa|sibs)]^M \left(\frac{1}{N}\right)}{[p(\text{both } AA|sibs) + p(\text{both } Aa|sibs) + p(\text{both } aa|sibs)]^M \left(\frac{1}{N}\right) + p(\text{match}|\neg sibs)^M \left(1 - \frac{1}{N}\right)}
\]
Confirming Sibling Relationship Given Matches at SNP Loci

\[ a \] \( p(\text{sib} | \text{match at } M \text{ independent SNPs}) \) vs.
Minor Allele Frequency (N=100,000)

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
Confirming Sibling Relationship Given Matches at SNP Loci

[b] \( p(\text{sib} | \text{match at } M \text{ independent SNPs}) \) vs. Minor Allele Frequency (\( N=10,000,000 \))

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
Confirming Sibling Relationship Given Matches at SNP Loci

\[ \text{[c] } p(\text{sib} \mid \text{match at } M \text{ independent SNPs}) \text{ vs. Minor Allele Frequency (N=6,000,000,000)} \]

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
How Many Genotypic Inferences Should We Expect to Get Correct?

• Guesses can be treated as a random variable with $p$ as the average % of success, as long as SNPs selected are independent.

• If $n$ guesses are considered (i.e. $n$ SNPs are genotyped and used for sib inference), what is the probability that $k$ of those will be correct,

$$p(k, n, p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$F(k; n, p) = P(X \leq k) = \sum_{j=0}^{k} \binom{n}{j} p^j (1 - p)^{n-j}$$

Example: $n = 100$ SNP inferences, $p = 0.8$ of correct inferences

What is the probability of at least $k = 75$ correct guesses
Inference Experimental Results

• 700,000 SNPs on 3 chromosomes (2,4,7), 30 HapMap CEPH trio datasets were used.

• For each SNP, the child’s genotype was used to infer the genotype of another sib at that locus using a refining strategy and SNP population frequencies.

• Results were validated using the expected probabilities $p(AA)$, $p(Aa)$, $p(aa)$ of children from the parents in the HapMap trios.
30 HapMap Trios

Validation Sibling Genotypes are p(AA), p(Aa), p(aa) at each SNP given parental genotypes.

We infer genotypes of 2nd sibs from actual HapMap Sib Data
Scoring Genotypic Inferences

- Results come in the form of:
  - $p(AA)$, $p(Aa)$, $p(aa)$ for the inferred sibs
- Validation data comes in the form of:
  - $p(AA)$, $p(Aa)$, $p(aa)$ given actual parent genotype.
- If we ‘called’ the correct expected genotype, we get a full point.
- If we ‘called’ one of the two matched 0.5/0.5 genotypes we get a half point.
Results of Genomic Inferences

For SNPs where Sib$_1$ was **homozygotic major:**
- Minor allele frequency < 0.05 (N=300512, 43.2%), we can infer Sib$_2$ with 98.5% accuracy
- Minor allele frequency < 0.20 (N=452684, 65.1%), we can infer Sib$_2$ with 91.9% accuracy

For SNPs where Sib$_1$ was **heterozygotic:**
- Minor Allele Frequency > 0.20 (N=125796, 18.1%), it is possible to infer the correct genotype of the second sibling with 57.7% average accuracy.
Percentage of Correct Inferences

[a] Fraction of Correct Sib2 Inferences where Sib1 is Homozygous Major

[b] Fraction of Correct Sib2 Inferences where Sib1 is Heterozygous

Cassa, Kohane, Mandl. BMC Medical Genomics 2008
Relative Risk for Sibling Carrying a Specific Genotype

Sibling SNP data can be used to quantify an individual’s disease propensity through genotypic inference, without that individual’s actual sequence data.

\[ \Gamma_{Sib_2 \text{ genotype} \mid Sib_1 \text{ genotype}} = \frac{\text{probability with sibling knowledge}}{\text{probability without sibling knowledge}} \]

\[ = \frac{p(Sib_2 \text{ genotype} \mid Sib_1 \text{ genotype})}{p(Sib_2 \text{ genotype})} \]

\[ = \sum_{i=1}^{9} \left( \frac{p(Sib_2 \text{ genotype} \cap \text{parental comb. } i)}{p(\text{parental comb. } i)} \right) \frac{p(\text{parental comb. } i \mid Sib_1 \text{ genotype})}{p(Sib_2 \text{ genotype})} \]
Relative Risk for Sibling Carrying a Specific Genotype

For example, the relative risk of \( Sib_2 Aa \), carrying one copy of the disease allele ‘a’, is provided by information from the \( Sib_1 aa \) genotype:

\[
\Gamma_{Aa|Sib_1 aa} = \frac{p(Sib_2 Aa|Sib_1 aa)}{p(Sib_2 Aa)}
\]

\[
= \frac{\frac{1}{2}p^2 + pq}{2pq}
\]

\[
= \frac{\frac{1}{2}p + (1 - p)}{2(1 - p)}
\]

\[
= \frac{1 - \frac{1}{2}p}{2 - 2p}
\]
Topics

1. Disclosure risks for genomic data and impact on family members

2. A) Anonymization of geospatial datasets containing patient home addresses
   B) Re-identification potential of geospatial data, commonly shared in both textual form and in printed maps
Revealing Addresses from Published Maps

Brownstein, Cassa, Mandl
NEJM Oct.
2006
Re-identified 79% of points from low resolution map

Brownstein, Cassa, Mandl NEJM Oct. 2006
Background

• The use of protected health information for spatial analysis is common and critical for
  – Exchange of health data in health record networks
  – Disease detection and surveillance systems
  – Identifying etiology, patterns, correlates, and predictors of disease
Key Concept: $k$-Anonymity

• Degree of anonymization is defined in terms of $k$-anonymity – where each patient is not identifiable among $k$ other patients.

Current Anonymization Methods

- **Simple aggregation**: Eliminate entire data fields (such as zip code, birth date, street address)
- **Truncation**: Remove portions of those fields (i.e. remove the last two digits of the zip code)
- **Geographically skew**: random changes to geocoded address data
- **Transformation**: Other affine transformations (translations, reflections, dilations preserving colinearity)
- **Geographical aggregation**: K-nearest neighbor ‘mixing’
Population Density-Based Anonymization Algorithm

• The goal is to preserve location information without endangering the patient’s privacy.
• Shifting by 1 mile in a rural area would yield a very different anonymization level than shifting by 1 mile in downtown Manhattan.
• Census data can be used to adjust skew of longitudes/latitudes based on population density.
• Gaussian weightings and randomizations are used to maintain maximum information while decreasing identifiability.
A Gaussian Approach to Anonymization
Distribution of Distances from Original Points
[Average Distances Moved: 0.0587, 0.1168, 0.1762, 0.2354 km]

Cassa, Grannis, Overhage, Mandl, JAMIA 2006
Authentic
Anonymized
Application Area: Cluster Detection and Disease Surveillance

Avg. Cluster Sensitivity/Specificity vs. Avg. Distance to Original Point

Cassa, Grannis, Overhage, Mandl, JAMIA 2006
Novel Estimate of K-Anonymity

- 68.26% patients in $\sigma$ (1SD) miles from center
- Multiply the local population density by the area, $[\pi \sigma^2_1]$
- Then multiply by the probability that the patient would have been moved into that region, 0.6826.
- Repeat each $[\pi(\sigma^2_n - \sigma^2_{n-1})]$
Average $k$-Anonymity Achieved vs. Average Distance Moved [km]

Cassa, Grannis, Overhage, Mandl, JAMIA 2006
Anonymization Using Linear Programming

The decision variables are the transition probabilities $P_{ij}$ of assigning a patient in location $i \in A$ to a new location $j \in B$.

Constraint equations specify conditions that must be satisfied by the decision variables $P_{ij}$.

$$0 \leq P_{ij} \text{ for all } i \in A \text{ and } j \in B$$

In addition, every case must be moved somewhere, so

$$\sum_j P_{ij} = 1 \text{ for all } i \in A$$
Constraints

The risk of linking any randomized location with any original patient should be small. We specify the probability that any location from the randomized data set originated from any specific individual in the underlying population is at most $\xi$:

$$P_{ij} \cdot \frac{n_i}{N} \leq \frac{n_i \cdot \xi}{s} \cdot \sum_{k \in A} \frac{n_k}{N} \cdot P_{kj} \quad \text{for all } i \in A \text{ and } j \in B$$

Objective Function is the expected distance that a patient is moved, to be minimized:

$$\frac{\sum_{i \in A} \sum_{j \in B} d_{ij} \cdot n_i \cdot P_{ij}}{N}$$
Anonymization Using Linear Programming

• Linear programming technique to anonymize address data has several advantages:
  – Finds the mathematically optimal solution
  – Moves points a smaller distance on average
  – No unreasonable locations for points

• Downsides:
  – Most points are not moved very far, so while it is mathematically sound, it may be easy to find cases
Welcome to the Patient Anonymizer ver. beta 3. To begin:
1. Choose an input file and output filename. (You may enter the paths manually or use the file browser.)
2. Select input and output file types. Also choose the appropriate delimiter if using a CSV file.
3. Choose XML tags or CSV positions of your file for the required fields. Enter birthdate format used in your file.
4. Select anonymization level and click convert to begin anonymization.
Topics

1. Disclosure risks for genomic data and impact on family members

2. A) Anonymization of geospatial datasets containing patient home addresses
   B) Re-identification potential of geospatial data, commonly shared in both textual form and in printed maps
Anonymization Vulnerabilities

• Explored two classes of anonymization vulnerabilities:
  – Those published in disease maps in journals and in public health practice
  – Those that are more identifiable with multiple versions of the same cases anonymized
Identifying Original Addresses Using Multiple Copies of Anonymized Data
Equivalent of a Less Stringent Anonymization Strategy

The average of \( n \) anonymized data points with original location \((x_0, y_0)\)
is 
\[
\frac{1}{n} \sum_{i=1}^{n} L_i
\]
; a two-dimensional Gaussian random variable with mean 
\((x_0, y_0)\) and covariance matrix
\[
\begin{bmatrix}
\frac{\sigma^2}{n} & 0 \\
0 & \frac{\sigma^2}{n}
\end{bmatrix}.
\]

Inferred data is the same as Gaussian anonymized data with standard deviation of \(\sigma/\sqrt{n}\), a less stringent Gaussian skew anonymization level.
Identifying Original Addresses Using Multiple Copies of Anonymized Data

One Dataset of 10,000 Addresses and Unique Identifiers was Created

The same dataset is anonymized 50 times

Anonymized Datasets 1 & 2 are averaged

Anonymized Datasets 1, 2, & 3 are averaged

Anonymized Datasets [1,n] are averaged

Cassa, Wieland, Mandl. IJHG 2008
Average Distance to Original Point Decreases with Number of Averaged Anonymized Versions

Cassa, Wieland, Mandl. IJHG 2008
Conclusion

• PHI sharing mechanisms are quickly emerging and once in place, they can be used in concert with clinical medical records to achieve a wide variety of innovative health promotion and surveillance goals.

• There are associated ethical and social risks that must be monitored effectively, and privacy decision-making and security for these documents must be improved for adoption to be safe and useful.