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

Christopher A. Cassa

S.B. Electrical Engineering and Computer Science, 2003
M.Eng. Electrical Engineering and Computer Science, 2004
Massachusetts Institute of Technology

SUBMITTED TO THE HARVARD-MIT DEPARTMENT OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOINFORMATICS AND INTEGRATIVE GENOMICS AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2008

Signature of Author: ____

Harvard-MIT Department of Health Sciences and Technology
August 18, 2008

Certified by: ____

Peter Szolovits
Professor of Health Sciences and Technology, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: ____

Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology

Abstract

Electronic transmission of protected health information has become pervasive in research, clinical, and public health investigations, posing substantial risk to patient privacy. From clinical genetic screenings to publication of data in research studies, these activities have the potential to disclose identity, medical conditions, and hereditary data. To enable an era of personalized medicine, many research studies are attempting to correlate individual clinical outcomes with genomic data, leading to thousands of new investigations. Critical to the success of many of these studies is research participation by individuals who are willing to share their genotypic and clinical data with investigators, necessitating methods and policies that preserve privacy with such disclosures.

We explore quantitative models that allow research participants, patients and investigators to fully understand these complex privacy risks when disclosing medical data. This modeling will improve the informed consent and risk assessment process, for both demographic and medical data, each with distinct domain-specific scenarios. We first discuss the disclosure risk for genomic data, investigating both the risk of re-identification for SNPs and mutations, as well as the disclosure impact on family members. Next, the de-identification and anonymization of geospatial datasets containing information about patient home addresses will be examined, using mathematical skewing algorithms as well as a linear programming approach. Finally, we consider the re-identification potential of geospatial data, commonly shared in both textual form and in printed maps in journals and public health practice. We also explore methods to quantify the anonymity afforded when using these anonymization techniques.

Table of Contents

Biographical Note

Christopher A. Cassa

POSITION TITLE

Fellow, Children's Hospital Informatics Program

Graduate Student, Harvard-MIT Division of Health Sciences and Technology

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Massachusetts Institute of Technology	S.B.	2003	Electrical Engineering and Computer Science
Massachusetts Institute of Technology	M.Eng.	2004	Electrical Engineering and Computer Science
Harvard-MIT Division of Health Sciences and Technology	Ph.D.	2008	Bioinformatics and Integrative Genomics

A. Positions and Honors

Appointments

2003- Pre-doctoral Fellow, Children's Hospital Informatics Prog., Boston, MA

Other Positions

2005- Member, International Society for Disease Surveillance

2005- Member, Committee on Public Health Practice, Research, International Society for Disease Surveillance

2005- Member, American Medical Informatics Association

2005- Member, Committee on Ethical Legal and Social Implications, and Public Health Informatics, American Medical Informatics Association

2006- National Library of Medicine Public Health Informatics Cohort

2006- MIT ACM/IEEE, Member

Honors

2005- Member, American Association for the Advancement of Science

2005- AAAS/Science Program for Excellence in Science

B. Selected peer-reviewed publications

Cassa CA, Schmidt BW, Kohane IS, Mandl KD. My sister's keeper?: genomic research and the identifiability of siblings. BMC Medical Genomics 2008, 1:32

Cassa CA, Wieland SC, Mandl KD. Re-identification of home addresses from spatial locations anonymized by Gaussian skew. International Journal of Health Geographics 2008, 7:45.

Cassa CA, Iancu K, Olson KL, Mandl KD. A software tool for creating simulated outbreaks to benchmark surveillance systems. BMC Med Inform Decis Mak. Jul 14 2005;5(1):22.

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;13(2):160-5.

Cassa CA, Olson KL, Mandl KM. System to generate semisynthetic data sets of outbreak clusters for evaluation of outbreak detection performance. MMWR Morb Mortal Wkly Rep. 2004; 53 Suppl:231.

Brownstein JS, **Cassa CA**, Mandl KD. No place to hide--reverse identification of patients from published maps. N Engl J Med. 2006 Oct 19;355(16):1741-2.

Mandl KD, Reis BY, **Cassa C**. Measuring outbreak detection performance using controlled feature set simulations. MMWR. 2004;53 (Supplement: Syndromic Surveillance: Reports from a National Conference, 2003):130-136.

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;5:56.

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

Brownstein JS, **Cassa, CA**, Kohane, IS, Mandl KD. Reverse Geocoding: Concerns about Patient Confidentiality in the Display of Geospatial Health Data. Presented by Dr. Brownstein at the 2005 American Medical Informatics Association Annual Symposium, Washington, DC, October 25, 2005.

Reis BY, Kirby C, Sprecher E, **Cassa CA**, Brownstein J, Simons W, Jordan L, Mandl KD Advanced Modular Design for Scalable Biosurveillance Systems. Advances in Disease Surveillance, Vol 1, 2006.

Zuberi B, Bertram AK, **Cassa CA**, Molina LT, Molina MJ. Heterogeneous nucleation of ice in $(NH_4)_2SO_4$-H_2O particles with mineral dust immersions. Geophysical Research Letters, VOL. 29, NO. 10, 1504, doi:10.1029/2001GL014289, 2002

C. Teaching Experience

MIT Courses

1.00 *Introduction to Computers and Engineering Problem Solving* (Spring 2004)

1.124 *Foundations of Software and Computation for Simulation* (Fall 2001, Spring 2002 as MIT Advanced Studies Program class)

1.264J / ESD.264J *Database, Internet, and Systems Integration Technologies* (Fall 2004, Fall 2007)

Harvard Medical School

Summer Scholars in Bioinformatics and Integrative Genomics (Summers 2005, 2006)

Scholars in Clinical Science Program Bioinformatics Module Coordinator and Lecturer (Summers 2004-2008)

Acknowledgments

I owe an enormous debt of gratitude to my thesis committee: Dr. Kenneth Mandl, Dr. Peter Szolovits, and Dr. Isaac Kohane. I would like to thank Ken, my research supervisor throughout my graduate career, for serving as a kind, patient, and caring mentor, always with my best interest and future in mind. Your attention to detail and dedication to your work has inspired me and helped improve mine. Thank you, Pete, for all of your help in improving both of my theses, and for always taking the time to chat about technology, research, and life. Thank you to Zak for helping to inspire interesting and relevant research and for your efforts to advance my career.

Shannon Wieland, John Brownstein, and Karen Olson: thank you for all of your energy, enthusiasm, and teamwork in our collaborative efforts. I would like to also thank those who have worked with me in publications, Ben Reis, Marc Overhage, Shaun Grannis, Brian Schmidt, and Karin Iancu, for all of the work that you have done to help me with my research. Thanks to every member of the Children's Hospital Informatics Program, including my office mates Fabienne Bourgeois and Lucy Hadden, and to Andrew Kiss, who has spent much time helping me with many projects.

To the students of HST and the Bioinformatics and Integrative Genomics program, thank you for making this an amazing place to do research and for being such wonderful colleagues in science and medicine.

I would like to thank Dr. John Tsitsiklis, Dr. Gilbert Strang, and Dr. David Altshuler for helpful discussions and advice in both mathematics and genetics.

To my friends: thank you for providing fun and respite from research, patiently enduring me, and appreciating me. Winter in Boston would not be worth it without you. To my students: thank you for keeping me enthusiastic about learning and teaching, and for providing opportunities to try and share the wisdom that many mentors have offered me.

And most important, thank you to my family, Fab, Carol, Jules, and Ari for your unending support, love and care. You have all helped me develop in so many ways, and enabled me to pursue every passion I have had in my life.

This work was partially supported by the following grants: R01LM007970-01 (NIH/NLM), R01LM007677 (NIH/NLM), R01LM009375-01A1 (NIH/NLM), R21LM009263-01 (Canadian IHR), 5U54LM008748-02.20 (NIH/NCBC).

List of Figures

List of Tables

Quiquid latine dictum sit altum viditur.

Whatever is said in Latin seems profound.

Chapter I: Introduction & Background

Introduction

Modern healthcare systems rely on transmission of protected health information for clinical, research, and public health purposes. This communication poses substantial risk to patient privacy, with the potential to disclose identity, medical conditions, and hereditary data. This cost in patient privacy must be carefully weighed and considered against the societal benefit for advancing the state of science and protecting public health. Additionally, allowing patients and practitioners to fully understand these risks when disclosing medical data will enable genuine informed consent in the era of personalized medicine. We also explore de-identification strategies – the removal of data that would help identify individuals from corresponding data set records – as well as re-identification techniques – the process of attempting to identify a specific individual or a set of individuals from de-identified data.

We explore disclosure risks of both demographic and medical data, each with distinct domain-specific approaches. We begin with genomic medical data, and investigate risk to patients and their relatives, both in the context of identifiability and disease status. We then change focus to spatial data, such as addresses, commonly included in demographic and clinical data sets, and investigate the ability to de-identify geographically encoded addresses in a manner that still maintains their usefulness in cluster detection. We also explore methods to quantify the anonymity afforded when using these anonymization techniques. We conclude with a discussion of reverse-identification techniques, including vulnerabilities that emerge when employing specific types of de-identification strategies.

While these specific approaches are tailored to different classes of clinical data, many share methodology and implications across those fields, particularly with respect to novel quantitative metrics for privacy and identifiability.

Genomic data have the potential to reveal a great deal about patients, ranging from phenotypic or disease propensity information, to paternity or lineage. Given the information content derived from familial records, we quantitatively model such data to help with the communication of privacy risks for relevant use cases. We hope this will encourage improved presentation of risk to patients in an informative, readable set of views and pedigree charts. There are also a number of legal and policy aspects to consider, including communication of otherwise confidential, but implicit data, and the sharing of derived familial medical data without direct consent.

Clinical data that are regularly recorded and stored in hospital data systems includes information from each part of the medical and payment process: patient identifiable demographics, insurance data (potentially with implicit or explicit employer information), laboratory results, physician and practitioner notes, and potential patient annotations. Each of these data types must be handled carefully, as the data contained in any piece of these may have the ability to assist in individually linking a record to a specific patient.

Ethical, Legal, and Social Implications (ELSI) of Personalized Medicine

The human genome project was initiated to explore and extract the shared genotypic sequence and basis for developing human characteristics and heritable health status [1]. Knowledge of the human genome sequence has led to the development of thousands of research studies and new fields of research, including functional genomics, epigenetics, and proteomics, among others. These range from those that attempt to discern the distribution of alleles throughout the world's populations in a variety of geographies [2, 3], to those that seek to identify the genomic location and function of genes that cause disease or disease propensity [4-6]. On top of these studies, there is a rich study of systems biology which integrates both genetic and protein networks whose complex interactions are difficult to model, but may be an effective way to study complex sets of genetic variants [7]. Most of this research ultimately seeks to help identify and cure disease in individual patients, a truly complex task called personalized medicine.

Personalized medicine is destined to improve treatment efficacy and outcomes for patients: if the most effective treatment possible for a specific patient can be selected and less effective or hazardous treatments can be left out [8], an individualized regimen has enormous potential [9-11]. Technology is evolving to enable personalized medicine to become a reality, including the evolution of research promoting inexpensive genotyping technology [12, 13] and companies

[14] offering inexpensive genotyping to the public. To make this information tractable and useful to patients, there are myriad companies offering informative personalized medical data about those genetic variants that have been observed [14-16] and even for full genome sequences [16]. There are also research projects aimed at making genomic data freely available on the web for exploration and research, including the Personal Genome Project [17, 18].

There are a number of contentious items in the personalized genetics and personalized medicine docket, including several ethical, social and legal ramifications that should be considered. Among those are the questions of whether personalized genomic medicine will fundamentally translate into a form of prophylactic medicine, where primary and secondary prevention will take the form of genetic screening, birth control measures, and pregnancy termination [19].

Prevention in human genetics does have an unpleasant history, including eugenics and sterilization. These items have been replaced with more sound preventive strategies, including routine newborn screening and community screening for at-risk carrier populations [20-22]. There is certainly a social and ethical risk of extension in this domain as available data linking a merely displeasing or disadvantageous characteristic to a genotype becomes available. Conversely, enhancement measures that use genomic data (vis-à-vis gene therapies) are equally hazardous when not treating a disease or disorder [23, 24]. Additionally, as the study of aging and gerontology take root in the age of personalized medicine, many ethical dilemmas will surely be raised; preventing Alzheimer's symptoms and providing more youthful, healthy life while aging is certainly desirably, but where will the bar be placed for termination of treatment [25]?

Perhaps the most important social implication to the public is the threat associated with sharing genetic data that might reveal personal or familial propensity to disease. The Genetic Information Non-Discrimination Act, recently passed (GINA, H.R. 493), will help protect individuals and their family members from financial consequences or forcible genetic testing by employers or health insurers [26]. This should allow the expansion of individual genetic testing and public screening efforts, but it does not solve all of the social issues associated with genetic testing. There are still many other places where discrimination may legally occur if a patient has a disease genotype, including the use of genetic testing in setting life, disability, and long-term care insurance premiums [27]. Familial genotypic sequences can be used to assist in forensic or criminal investigations for indirect identification of genotype, increasing the number of people who may be identified [28, 29]. Similarly, Freedom of Information Act (FOIA) [30] requests related to federally-funded genome wide association studies could potentially be used to identify research participants and their family members. Clinically, choosing the detail and type of disease propensity information that must be disclosed to patients and their potentially affected family members is also under debate [31, 32].

The current predictive power of genetic testing for approximately 1,500 monogenetic diseases is robust [33]. However, for much more common, polygenic disorders, the ability to predict disease propensity continues to be poorer from genetic testing than from family history [34]. This raises ethical and social concerns; should the public receive broad-based genetic screening until it has proven clinical value? Without appropriate diagnostic value, testing may provide false alarms and false hope, and also prove costly in unnecessary clinical follow-up. One specific example is the recent change made to clinical guidelines for prostate cancer screening, specifying that patients above age 75 receive more harm than benefit from such screenings [35-37].

Further, research in genetics – particularly for complex diseases – has generated a large number of irreproducible studies, creating a large set of incidental and dubious findings, coined the incidentalome [38-40]. There is no 'clearinghouse' that designates a SNP association with disease as clinically valid, and in this 'bleeding edge' research arena, there is a bias for journals to approve new positive associations between genetics and disease without substantial reproducibility [41-43]. Because of this, personalized medicine in the electronic medical record age may be filled with a lot of 'noise' for patients, as findings of dubious, not reproduced studies are rapidly disseminated to their electronic records with no proven clinical benefit.

Personalized Medicine and Personally Controlled Health Records

With all of its potential risks, the field of personalized medicine continues to grow, promising to have a significant impact on medical care. Legislation on the federal docket was recently considered (S.986 Genomics and Personalized Medicine Act of 2007) to broadly expand the funding for research targeted at research studies that will have future impact on individualized medical treatments. Because of this broad growth, studies have gathered large groups of participants interested in sharing their genomic data with researchers, including participants from the Framingham Heart Study and the Women's Health Initiative [4, 5, 7, 44-46].

In addition to these large-scale studies, enabling researchers to get a wide variety of linked clinical and genomic data sets from altruistic volunteers from the public who would share a subset of, or their entire genomes has enormous potential to advance science. One way to potentially reach these volunteers would be through electronic medical records, specifically patient controlled medical records (PCHRs) [47-51]. PCHRs differ from conventional electronic medical records (EMRs) [52] in several important ways: 1) PCHRs give patients complete control over what components of their medical records and data are shared with which clinicians 2) PCHRs have the ability to span many points of care, from disparate institutions, and 3) PCHRs may be patient owned (in some models), and if so, patients should be more comfortable with private medical and genomic data storage in those records. A recent article describes the mutual benefits such broad public participation could have for both patients and researchers in a controlled fashion [53]. The authors describe a Genetic Partnership Project (GPP) which would allow patients to share (likely through a PCHR interface) their genetic data with researchers, and then also allow patients to 'tune in' to updates on that research in an anonymous fashion.

Separately, there is interest in tapping the potentially enormous expanse of medical data that may be stored in corporate medical record storage systems, including

Google Health and Microsoft HealthVault. These systems plan to provide patients with the ability to gather and store their healthcare data from a variety of participating healthcare institutions [54]. Once these systems have a large set of consumers, they may potentially be in control of the largest available set of standardized electronic medical information. This information could similarly be used to create and consent research cohorts and there is much to be determined about how that process would work and whether it is ethical and would meet the high standards that are required for medical researchers [55].

Human Variation Data Sources and Information Content

Single Nucleotide Polymorphisms (SNPs) differ between members of a species (or between paired chromosomes in an individual). SNPs comprise up to 90% of all human variation [56], and individual SNP genotypes and geographical population frequencies of SNPs are becoming increasingly available in research repositories (Figure 1 and Figure 2). SNPs have the potential to help identify how genotypic diversity relates to phenotypic diversity, diseases, and outcomes.

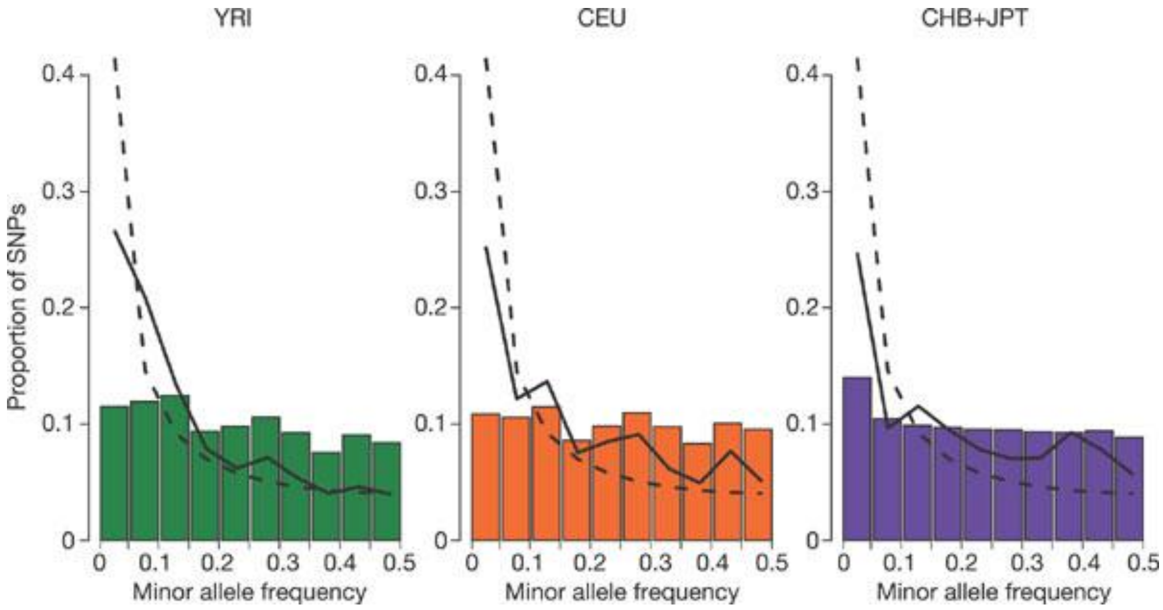


Figure 1: The proportion of SNPs by minor allele frequency, binned by 0.05 in frequency for the four HapMap populations (CHB + JPT were combined). The solid line represents the actual distribution from ENCODE SNPs and the dashed line describes the Minor Allele Frequency distribution expected for the standard neutral population genetics with random mating and fixed population size. [From A haplotype map of the human genome. *Nature* 437, 1299-1320]

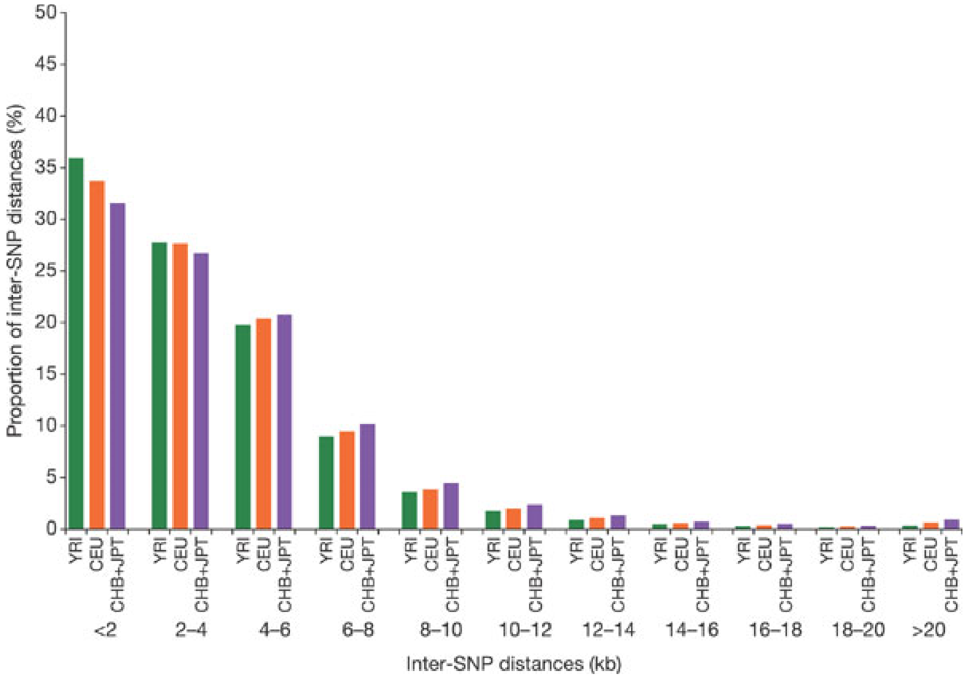


Figure 2: The proportion of inter-SNP distances in areas covered by the HapMap project, binned by inter-SNP distance (kb), for all SNPs with Minor Allele Frequency ≤ 0.05. [From A haplotype map of the human genome. *Nature* 437, 1299-1320]

SNP loci that are in linkage disequilibrium with one another can be grouped together to form haplotype blocks; groups of SNPs that have a population frequency of matching greater than would ordinarily be expected based on their distance from one another [57]. Linkage disequilibrium makes SNPs statistically dependent, and alters the information content when a set of SNPs are shared or published. SNP genotypes provide a variable amount of information which depends on the population frequencies at the loci in question and on the mutual linkage disequilibrium values between each SNP included in a data set.

The HapMap project has compiled sequencing and population frequency information that can be used to provide the most current and informative risk estimates for health data disclosure. The project, organized by the Harvard-MIT Broad Institute has compiled gene frequency values for a large selection of SNPs – loci in the genome that account for a great deal of genetic variability in populations [3]. The HapMap project also provides linkage disequilibrium data for several populations. Linkage disequilibrium is a covariance metric for each set of statistically dependent SNPs in the genome; the HapMap project has measured how likely it is that two SNP values would co-segregate together in a given population.

Biomedical data collection includes a wide variety of structured and unstructured values and measurements, including clinical phenotypes, DNA sequencing, demographics, family history, gene expression profiles, copy number variants, and proteomics data. It is very likely that there will be a wide variety of polymorphisms or variants that are associated with diseases.

With the breadth of genomic data types, as well as the data structures and identifiers that represent them, we have focused our efforts on creating models and metrics that utilize a limited set of the most informative genomic data for decision support. Because SNPs are both clinically informative and will be used for much future research, we have elected to focus our analysis efforts on population-specific SNP values at sequenced loci and familial relationships.

Research data that associate SNP alleles to health status and disease propensity is increasingly available, while comparable data for many other polymorphisms -- that are certainly relevant -- such as copy number variants, is not yet broadly available. Future projects will need to explore new genomic data sources that are available for populations and research cohorts and extend these techniques to them.

Measuring Risk of Identity Linkage using Genomic Data

When patients share their data with medical researchers, they expect that their identities and protected health information will be secured. There is a balance, however, between the need to protect patient identities and the imperative to publish supporting research data and to make available expensive genotyping assays from large publicly-funded studies with any researchers who might extract value from them.

Researchers who have attempted to de-identify protected health data have historically not been successful, as research subjects can often be re-identified uniquely or within a small group of individuals [58]. The use of a variety of publicly available de-identified data sources has aided in these re-identification efforts; many times these data sets can be joined together to link records and enrich the available information about each individual. Malin and Sweeney used a publicly available hospital discharge data set and combined it with voter records and census data to statistically link individuals within those data sources using zip codes, age, and gender. They were able to uniquely identify patients with rare genetic diseases including a third of all cycstic fibrosis patients, half of all patients with Huntington's disease, and even higher numbers of patients with more rare genetic disorders, that were admitted to hospitals in Illinois between 1990 and 1997.

These findings demonstrate that it is possible to directly link publicly available data sets down to clinical phenotypes and even individual-level DNA lesions. This certainly would bring alarm to some of the patients who had not even personally consented to the release of their healthcare data at Illinois hospitals. Given the complexity of genomic data, it may not be possible to provide an acceptable level of confidentiality or privacy in this form of medical research while publishing this data [59]. And to define what 'acceptable' means to patients adds additional complexity, as genomic data and the potential damage its disclosure might cause are not well understood [60-64].

Transfer or publication of genomic data poses unique privacy dangers. Irrevocable and unchanging as a fingerprint, any disclosure of patient genomic data poses a life-long risk for patients and their relatives; traditional data security mechanisms to cancel availability and access to previously disclosed genetic data are severely limited. Unlike fingerprints, however, which provide little direct information about patients when not linked with names, genomic test results contain information that encodes phenotypes, characteristics, and disease propensities. Hence, it will be increasingly possible to directly link sequence data with phenotypic data and inherently carry health care risk information [58].

Zhen Lin and Russ Altman [59] demonstrated that privacy decreases sharply with disclosure of a small number of SNP genotypes. In fact, with just 35-70 independent SNP genotypes, it is possible to uniquely identify any individual. Because DNA is so identifying, the authors contend that the ability to conduct meaningful medical research using genomic data will necessarily reduce the privacy afforded patients. They also characterize the sharp decline in privacy at a range of SNPs (which depends on the minor allele frequencies of those SNPs) at which an individual becomes uniquely identifiable, and demonstrate that this is well below the number of SNP genotypes that would likely be shared with researchers (Figure 3).

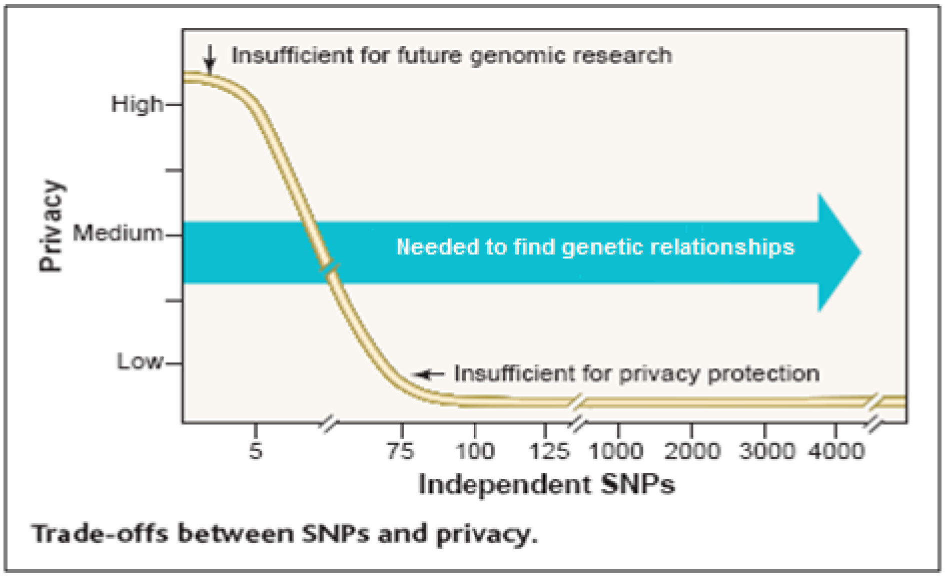


Figure 3: Lin and Altman, et al describe the trade-off between privacy level and number of SNP genotypes disclosed to researchers or clinicians. [From Genetics. Genomic research and human subject privacy. *Science* 2004, 305:183. (Revised Edition)]

The study explored the probability that two randomly-selected, unrelated individuals match on a group of M' SNPs that are statistically independent (not in linkage disequilibrium). The probability of two individuals matching at a single SNP is the sum of the probabilities of two homozygote major individuals matching, two heterozygote individuals matching, and two homozygote minor individuals matching in the population: $p(AA)^2 + p(Aa)^2 + p(aa)^2$. For a set of M' independent SNP matches (where we have a priori selected SNPs with population frequency of 10%), the probability of match, $\mu_j \leq 0.689$, $((0.9^2)^2 + (2*0.1*0.9)^2 + (0.1^2)^2)$, the probability of this set of matches happening by chance is:

$$\prod_{j=1}^{M'} \mu_j \leq 0.689^{M'}$$

It does not take a large value of M' to make this probability very small. Lin and Altman subsequently evaluated the probability that two people are the same given a set of matched SNPs in a fixed population size via Bayes' Theorem.

While these findings have launched valuable discussion, they have also led some researchers to believe there is no way to share a small amount of SNP data while precluding re-identification of patients. At present, based on this study, it appears that the sharing of small, but clinically relevant, sets of partially dependent SNPs is possible, with adequate threat assessment and updated population-specific SNP frequency data.

Attempted Interventions to Protect Genomic Privacy

Research groups have attempted to mitigate the threat to privacy that the publication of genomic data poses. The techniques include methods to blur or change the data, reducing granularity or resolution on the data, and aggregation techniques. All of these methods fail to improve the privacy afforded patients in any dramatic way [65, 66]. A summary of the attempts to date follows.

Using Binning to Maintain Confidentiality of Medical Data

Binning describes the process of aggregating elements in a data set into a more generic pool of field values with similar attributes. One study attempted to disregard exact genomic positions for a set of SNPs to increase the number of data sets that have the same sets of values [67]. The shortcomings of this approach were that the information that was subsequently available to researchers was substantially reduced; precise genomic location data are important for identifying the exact locus or lesion involved in a genetic disease process. Conversely, the privacy afforded by this technique is dubious when there is one predominant mutation that leads to sequencing in a genomic region; if a monogenetic disease locus is nearby, it is likely that any observed mutation within that region refers to that one specific, common, monogenetic lesion. This may also just slightly increase the size of the data set needed to uniquely identify a patient.

Disclose Frequencies and Aggregated Data Only

A variation on the above theme is to aggregate records, thereby binning at the patient level rather than by characteristics or fields within patient records. An example of this would be a population genetics description such as "Half of the patients in this study carried the homozygous major genotype AA while 40% carried Aa and 10% had aa." One shortcoming to this approach is that supporting clinical or phenotypic information, at the individual patient level, may help researchers gain insight on a genetic disease process. Additionally, clinical value for specific patients and ability to deconstruct research diminished

Anonymity by Pool Selection

The Human Genome Project (HGP) gathered a large number of samples from individuals who were brought through a thorough consent process. Then, the project anonymously selected a very small subset to create a consensus hybrid of several participant genomes to prevent the identity of participants from being known [68].

The HGP used participant pool selection as a privacy technique though it is of dubious privacy value. There are regions of the genome where SNP loci that have since been discovered (in larger pools of sequenced individuals) could help identify participants if samples were available for forensic analysis. Additionally, an unclear form of genome aggregation was used, which depends on the genomes that were used, in what proportion their derivative sample chromosomes were used and alignment technique statistics that were employed. This is not a clear form of privacy.

Use of Generalization Lattices

A more specific variant on the concept of binning is to use generalization lattices to de-identify data sets partially where it is either most prudent for privacy or where it will not substantively reduce clinical value. An example of this technique was described the use of genotypic base pair binning [69]. The most basic example would be to have two possible levels of generalization that cover the four DNA base pairs; A and G may be reduced to a representation of R; C and T reduced to Y (Figure 4); and at the next level of generalization, f1, R and Y may be more generally represented as N.

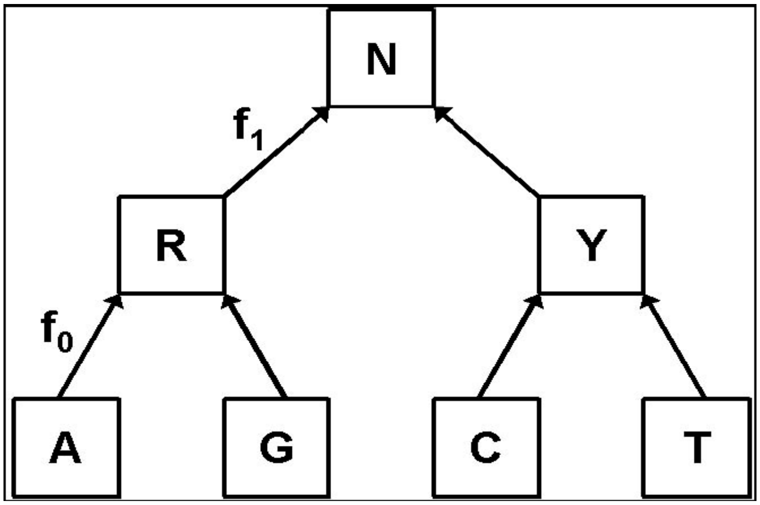
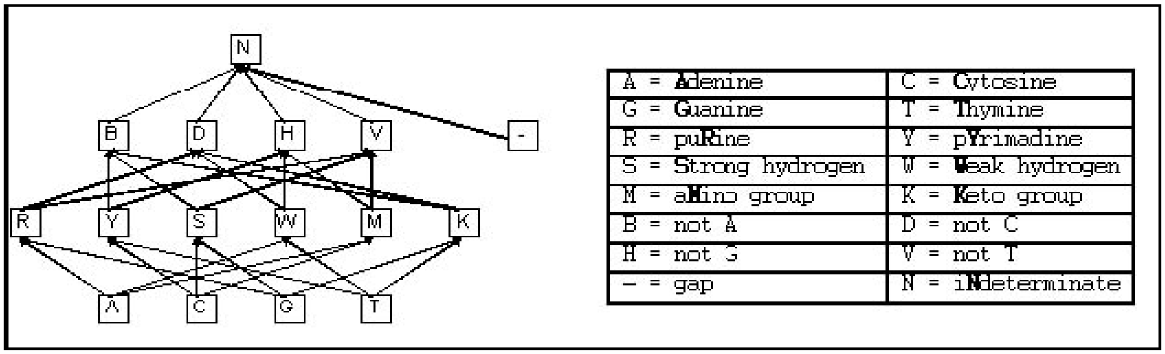


Figure 4: Protecting DNA Sequence Anonymity with Generalization Lattices. In this example, each purine (A, G) may be consolidated into R, and each pyrimidine (C, T) may be consolidated into a new base pair Y, both for generalization. All four base pairs can be generalized into N to reduce the information that is disclosed when publishing genomic data.

There are also more complicated generalization lattices (Figure 5) that have been developed in order to reduce the amount of information that is disclosed when publishing a genotypic sequence.



A = Adenine	C = Cytosine
G = Guanine	T = Thymine
R = puRine	Y = pYrimadine
S = Strong hydrogen	W = Weak hydrogen
M = aMino group	K = Keto group
B = not A	D = not C
H = not G	V = not T
– = gap	N = iNdeterminate

Figure 5: More complex generalization lattices can be used to complexify and obscure the information content that is shared in a disclosed sequence. This example includes the four DNA base pairs as well as purine and pyrimidine generalizers used above, reverse identifiers (not A), amino group identifiers, keto group identifiers, among others.

All of these techniques simply make it slightly more difficult to re-identify the individual from a published genotype using this technique. It is possible to find a closed form solution for the anonymity provided by these techniques, which simply are another form of aggregation. This technique also shares the problem that published genotypes do (by definition) lose information content. This reduces the amount of information that is available for researchers to find a correlate or predictor of disease and would reduce the statistical significance of findings if not all genotypes are available at all loci.

Add Noise to a Genotypic Sequence

One technique that may be employed to reduce information in published data sets is to randomly skew a certain (perhaps unknown) fraction of genotypic values. The largest reduction of information content would come from blurring those loci that are either known rare variants (either rare SNP loci or mutations) so that a specific record is not so individually identifying. This certainly reduces the information content, but reduces the value of a research data set dramatically – the data being shared is intentionally being contaminated, potentially leading to false conclusions and missed findings.

This technique is perhaps the most dangerous because it can lead to false conclusions, and in fact is provides just as little protection to privacy. Altman, et al demonstrated that this technique still allows for identifiably with low numbers of independent SNPs, as described by a false negative and false positive rate of matching samples to individuals using a variety of skewing rates below (Figure 6).

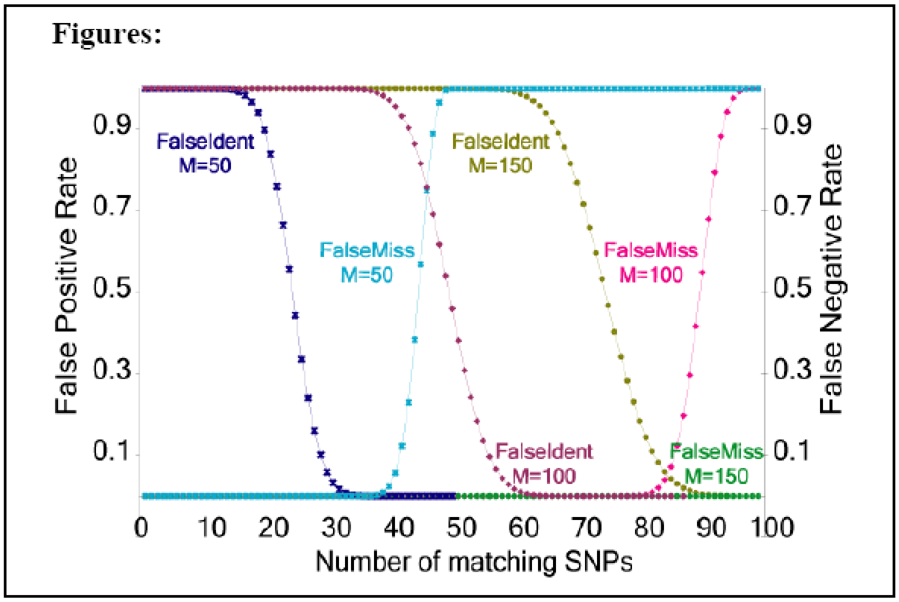


Figure 6: Introducing noise into SNP genotypes still results in identifiability. Ten percent random noise was added to a SNP data set, and at various numbers of SNP matches, the false negative and false positive rates of identification are graphed.

Synthesizing anonymized 'individuals' using statistical data associations

Recently, Lasko et. al. described a system to create anonymized records that contain data resembling authentic individual-level data sets. Using statistical associations within those data sets, he creates synthetic individual records with information and relevance for research purposes, while preserving patient privacy. Such systems will be a challenge with genomic datasets, however, because of the potential complexity of genetic interactions that will be explored in personalized medicine research. It will likely be the case that full contiguous genotypic data will be required, with associated potentially identifying clinical data to identify genomic network effects or subtle polymorphic variants acting in combination to create a larger effect.

Quantitative genomic disclosure risk models for patients and relatives

The need to calculate the information content in a genomic data set is acute; this will enable EMR and personalized medicine systems to model the degree of privacy afforded patients when they share a subset of their genomic data. Information theoretic tools are effective in characterizing the information content of sets of SNP data sets [70]. We have developed a set of four disclosure risk models that address important clinical sharing scenarios for patients and their relatives.

Risk of re-identification. One clinical and research privacy scenario is the disclosure of a set of genomic data that contains either SNPs or mutations. In this context, we describe a probability bound on how identifiable a set of SNP or mutation data is, under different sets of circumstances, such as whether it includes any phenotypic or population-specific data. This analysis should consider population-specific frequencies of the specific SNPs as well as the localized mutation rates and mutation types in the region of interest.

Risk of genome-gene inference. A related variation on the above theme is how readily two distinct, but overlapping sets (for example, where one set is a subset of the other, but the two are not disjoint sets) of genomic data can be combined with certainty to produce a more complete data set for one individual. With two sets of SNP data from a patient, one may identify whether the two data sets contain enough matched, overlapping base pairs to sufficiently determine whether the data sets came from the same individual, and with what probability. If it is possible to infer that the data sets are from the same individual, then to mitigate this threat, one might remove the most identifiable SNP values from the data set using a ranked list of the most informative loci. This issue is importance because if it is relatively easy to aggregate two distinct data sets from a patient with only minimal overlap between the two data sets, it allows a genomic test to be linked with other separately published or shared data, perhaps with clinical findings.

Risk of familial inference. Genetic data not only reveals information about those tested, but also about their family members, posing a considerable privacy risk for family members of those who would share their research data. On average, patients share half of their DNA with each parent and sibling, and a decreasing amount with other relatives. Given a patient's population demographic data, a set of a patient's SNPs, and a relationship with another person, we have quantified how likely it is that the 2nd person will have the same set of values at a set of SNP loci. Because we have additional knowledge of the specific relationship the first patient shares with the second, we are asking a question distinct from the original question of how likely it is that two patients should match at a set of loci.

We identify the familial information content within a set of proband SNPs: specifically how likely it is that a parent, sibling, and child will carry a specific set of SNP values based on proband genotypes, population-specific allele frequencies, and the familial relationship involved. We can also establish whether two individuals are related by evaluating a set of SNPs in both individuals, with certainty using closed form probabilities, if we consider independent SNPs.

Risk of genotypic-phenotypic linkage. Genotypic data predicts phenotypes in individuals, some of which are apparent through physical characteristics, clinically-observed values, and disease status. Using the genotypic and perhaps ethnicity identifiers in a patient controlled health record, one may identify some phenotypes from a specific set of patient SNP values [58]. Similarly, the reverse can be done, using phenotypic information we may derive likely patient genotypes. It is also possible to infer the population or ethnicity of a patient using a genomic sample with low numbers of SNP values if supporting population SNP frequency data are available.

Geographical Data Privacy in Public Health and Clinical Practice

The mapping of clinical and public health data is widespread in both academic research and public health practice [71]. While the study of the influence of geographical location on disease risk dates back to the mapping of yellow fever and cholera in the 1800's, research integrating maps and human health is an emerging field based on the wide availability geographic information system (GIS) software [72].

Ongoing disease surveillance and large research studies both rely on the ability to detect precise clustering patterns, but the privacy implications of sharing the necessary patient data carry risks for patients grouped in clusters with sensitive medical conditions or other protected health information. Both disease surveillance and research publication can utilize less-than-perfect spatial data that still illustrate the pattern of disease or incident clustering effectively, but afford patients increased privacy and anonymity.

Geographical disease surveillance systems designed to monitor public health threats have emerged that harvest data from a variety of sources, including emergency department and inpatient hospital visits, clinical diagnoses, lab results, over the counter drug purchases, and even orange juice and vitamin sales [73, 74]. These systems are designed to discover outbreaks of public health relevance that may be sparsely distributed geographically, before they would be noticed by an astute clinician or public health department. Web systems that mine a variety of news sources (open source media, Google News, CDC and WHO health alerts, among others) are also attempting to extract meaningful geographic information from reports and distill it into useful information that can help contextualize disease progression throughout regions [75-78].

GIS has broad applicability, and its use has been generally fueled by increased computing power, user-friendly software, and large geographic databases. The number of publications utilizing GIS for health research has grown at about 26% per year, four times the rate of increase for human health articles in general [72]. Patient address locations are mapped to identify patterns, correlates, and predictors of disease. These maps are often published electronically and in print [71]. A keyword search for the term “geographic” or “map” in the figure legends of five major medical journals from 1994-2005 identified 19 articles (including five from

NEJM) that include maps with patient addresses plotted as individual dots or symbols. In these papers, over 19,000 patient addresses are plotted on map figures.

The publication of disease maps with precise patient locations puts patient privacy in jeopardy. Guidelines for the display or publication of health data are needed to guarantee anonymity [79]. A common approach has been to map by administrative unit rather than home address. However, aggregation of data poses constraints on the visualization of disease patterns. Another method is spatial skewing or randomly relocating cases within a given distance of their true location. Skewing can allow a visualization that conveys the necessary information, while preserving privacy [80]. Both aggregation and skewing are systematic and reliable means of de-identification which are far safer, in terms of protecting identifiable health information, than simply reducing map resolution.

Anonymization of spatial data for disease surveillance

Patient re-identification from purportedly de-identified data can be accomplished with surprising ease. For example, Sweeney, et al. showed that 87% of individuals in a publicly available database were re-identified using five digit zip code, date of birth and gender alone [81]. There are well-described techniques for protecting the anonymity of individuals whose information resides in databases. Using these techniques, de-identification systems have been developed that remove personal data from database fields (for example, converting a date of birth to a year) [82] or from textual notes [83]. Uzuner, et. al. has also developed novel methodologies in de-identifying text and has worked on the NLP challenge problems addressing the same issue [84, 85]. Clifford, et. al. has also developed a system that promises to de-identify 94% of textual clinical notes [86]. The de-identification of databases has also been explored using several techniques [87, 88].

A metric for the ability to re-identify a patient in a data set is *k-anonymity*, where *k* refers to the number of people among whom a specific de-identified case cannot be reversely identified [82]. Spatial location information, whether stored as classic plain text address data or as geocoded longitude and latitude values, can potentially identify an individual or a markedly reduced set of candidate individuals. A common approach to de-identifying such data has been to use census tract or zip code rather than home address to protect anonymity. There are two important drawbacks to using location data that have been aggregated by political boundaries or administrative region. First, the loss of precise location may reduce sensitivity to detect clustering. Second, the ability to detect clustering may be diminished when some of the points cross administrative boundaries.

Previous investigators have attempted to mask geographic data by spatially skewing cases using, among others, affine and randomizing transformations [89, 90]. In this thesis, we describe a spatial anonymization algorithm based on skewing precise geocoded case locations using knowledge of local population characteristics. Skewing these patient addresses directly decreases the ability to re-identify, and thus increases the *k*-anonymity, of a case in a data set, as it will be much more difficult to determine what the actual patient's identity is once the address has been altered. Masking the identity of an individual in a densely populated urban area, for example, does not require as great a skew as one in a sparsely populated rural setting. Next, we measure the effect of anonymization intensity on outbreak detection, focusing on the sensitivity of spatial cluster detection. The goal is to provide individuals, institutions and public health authorities a comfort level with the sharing of skewed, and hence, anonymized data, rather than using raw, fully identifiable data. Further we aim to provide transparent information about the resulting diminution of spatial clustering detection.

Conclusion

Genomic medical testing and sharing mechanisms are quickly emerging and once these are codified, 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 practical or useful.

Chapter II: Genomic privacy: identifiability and familial risks

The use of integrated familial data is prevalent in genetic and genealogical studies, but has not yet reached its potential in clinical medicine, as its use poses substantial technical and policy challenges. Personal Health Record (PHR) systems currently lack the critical ability to incorporate such data, which is gathered at clinical encounters in a largely ad hoc fashion, without electronic standardization. Moreover, patients do not fully understand the benefits and potential risks involved in sharing such data with relatives, clinicians, or researchers [60-64]. The emerging use of PHRs presents an enormous opportunity for improvement, enabling patients to control the collection, extraction and disclosure of valuable genomic data.

Integration of familial genomic data in medical records has several tangible benefits for patients. First, familial data derived directly from family members' records is more likely to be accurate, complete, and up-to-date. Second, relatives may share genomic sequencing data with one another, which can be used to derive personalized disease propensity estimates [91]. The ability to derive genomic data poses risks to privacy when sharing clinical or genomic data with researchers: patients should understand the risks to their privacy as well as to family members' privacy when they share their data.

We describe and quantify the risks posed by these activities to address the challenges of curating and communicating the information content and disclosure risks of demographic, clinical and genomic familial data.

Ability to infer SNP genotypes from sibling genomic data

I am my sister's keeper

This section of the thesis was published in a manuscript entitled My sister's keeper?: genomic research and the identifiability of siblings, in BMC Medical Genomics with Brian Schmidt, Dr. Isaac Kohane, and Dr. Kenneth Mandl, from the Children's Hospital Informatics Program and Harvard-MIT Division of Health Sciences and Technology.

Abstract

Genomic sequencing of single nucleotide polymorphisms (SNPs) is increasingly prevalent, though the amount of familial information these sequences contain has not been quantified. We provide a framework for measuring the risk to siblings of a patient's SNP genotype disclosure, and demonstrate that sibling SNP genotypes can be inferred with substantial accuracy. Extending this inference technique, we determine that a very low number of matches at commonly varying SNPs is sufficient to confirm sib-ship, demonstrating that published sequence data can reliably be used to derive sibling identities. Using HapMap trio data, at SNPs where one child is homozygotic major, with a minor allele frequency ≤ 0.20, (N=452684, 65.1%) we achieve 91.9% inference accuracy for sibling genotypes. These findings demonstrate that substantial discrimination and privacy risks arise from use of inferred familial genomic data.

Background

Genomic data are increasingly integrated into clinical environments, stored in genealogical and medical records [92, 93] and shared with the broader research community [94, 95] without full appreciation of the extent to which these commodity level measurements may disclose the health risks or even identity of family members. While siblings, on average, share half of their contiguous chromosomal segments, well over half of a sibling's allelic values can be inferred using only population-specific allele frequency data and the genotypes of another sib. The informed consent process for research and clinical genomic data transmission must therefore include rigorous treatment of accurately quantified disclosure risks for all who will be impacted by such activity.

It is remarkably easy to positively identify a person with fewer than 40 independent, commonly varying SNPs, using a physical sample or a copy of those values [59]. As DNA sequences cannot be revoked or changed once they are released, any disclosure of such data poses a life-long privacy risk. Unlike conventional fingerprints, which provide little direct information about patients or relatives, SNP genotypes may encode phenotypic characteristics, which can link sequences to people [58]. Despite these privacy issues [65, 96], use of genetic sequencing is increasing in both forensics [97], and clinical medicine. The recent genetic fingerprinting provision in the renewal of the federal Violence Against Women Act [98], alone, may result in one million new sequenced individuals each year, markedly increasing the number of available links between identities and genotypes. This genetic fingerprinting has an impact on people beyond those directly sequenced--genetic testing partially reveals genotypes of siblings and other family members.

At each locus in a child's genome, each parent transmits only one of his or her two chromosomes. If we have the genotype of one child, and would like to use that information to help infer the genotype of a sibling, we consider both the known parental genotypes (for the alleles they have transmitted to their first sibling,) and also consider those chromosomes they have but have not transmitted. We assume that the unknown parental alleles are drawn from a reference population, such as one of the HapMap populations. Now, considering the genotype of the inferred sibling (2nd child), with probability 0.25, the sibling will receive the same 2 chromosomes transmitted to the first child, in which case they will have the same genotype. With probability 0.25, the inferred sibling will receive both previously untransmitted chromosomes, in which case the sibling will have the same genotype distribution as the reference population. If only one of the same chromosomes is transmitted, then one chromosome will be the same and the other will be drawn from the population.

Methods

To quantify the risk of SNP disclosure to relatives, we demonstrate a model for inferring sibling genotypes using proband SNP data and population-specific allele frequency databases, such as the HapMap [99, 100]. We also evaluate the probability that two people, in a selected pool of individuals, are siblings given a match at an independent subset of SNPs, and show that this number can be made remarkably low with appropriate SNP selection.

Enhanced ability to infer sibling genotypes

First, consider the case where one sibling's genotype is known to be *'AA'*, and the goal is to determine the probability that a second sibling's genotype will also be *'AA'* at that locus. Because there is additional knowledge—the familial relationship between the two sibs—the prior probability of the second sib carrying a specific genotype at a selected SNP will be altered under the new constraint. A conditional probability expression that sums over the nine possible parental genotypic combinations (for example, maternal genotype *'Aa'* with paternal genotype *'AA'*) at a single SNP, each denoted as *i* can be used:

$$p(Sib_2AA|Sib_1AA) = \sum_{i=1}^{9} p(Sib_2AA|parental\ comb.i)p(parental\ comb.i|\ Sib_1AA)$$

$$= \sum_{i=1}^{9}\left(\frac{p(Sib_2AA \cap parental\ comb.i)}{p(parental\ comb.i)}\right)p(parental\ comb.i|\,Sib_1AA)$$

where Sib_1AA and Sib_2AA refer to Sib_1 and Sib_2 genotypes *'AA'* at a selected SNP, respectively.

With unknown parental genotypes, we would calculate $p(Sib_2AA)$ considering all nine possible parental genotype combinations, but knowledge that Sib_1 has genotype *'AA'* allows exclusion of any parental combinations where either parent has genotype *'aa'*, as that would require the transmission of at least one copy of the *'a'* allele to Sib$_1$, if non-paternity and new mutations are excluded.

For example, when the child 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. In this case, there are four possible parental genotypic combinations:

$$= \sum_{i=1}^{4} \left(\frac{p(Sib_2 AA \cap parental\ comb.i)}{p(parental\ comb.i)} \right) p(parental\ comb.i \,|\, Sib_1 AA)$$

$$= \left(\frac{p(Sib_2 AA \cap AA_M AA_F)}{p(AA_M AA_F)} \right) p(AA_M AA_F | Sib_1 AA)$$

$$+ \left(\frac{p(Sib_2 AA \cap AA_M Aa_F)}{p(AA_M Aa_F)} \right) p(AA_M Aa_F | Sib_1 AA)$$

$$+ \left(\frac{p(Sib_2 AA \cap Aa_M AA_F)}{p(Aa_M AA_F)} \right) p(Aa_M AA_F | Sib_1 AA)$$

$$+ \left(\frac{p(Sib_2 AA \cap Aa_M Aa_F)}{p(Aa_M Aa_F)} \right) p(Aa_M Aa_F | Sib_1 AA)$$

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

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

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

which allows calculation directly from the SNP population frequencies. Before knowledge of the Sib_1 genotype was used, $p(Sib_2AA)$ would have been the Hardy-Weinberg frequency for major homozygotes, p^2. However, with the Sib_1 genotype, $p(Sib_2AA|Sib_1\ AA)$, the additional constraint increases the probability to $p^2+pq+(q^2/4)$, increasing inference accuracy by $pq+(q^2/4)$.

The remaining entries in the probability vector, $p(Sib_2Aa|Sib_1\ AA)$, and $p(Sib_2aa|Sib_1\ AA)$, can then be calculated just as we have done for $p(Sib_2AA|Sib_1\ AA)$ above. Again, these probabilities have been generated without any actual knowledge of the parent genotypes. If the Sib_1 genotype were instead *'Aa'* or *'aa'*, the above technique can similarly be used (with a different combination of possible parental genotypes) to calculate the two other probability vectors, [$p(Sib_2AA|Sib_1\ Aa)$, $p(Sib_2Aa|Sib_1Aa)$, $p(Sib_2aa|Sib_1\ Aa)$] and [$p(Sib_2AA|Sib_1\ aa)$, $p(Sib_2Aa|Sib_1\ aa)$, $p(Sib_2aa|Sib_1aa)$].

HapMap SNP population frequencies, p and q, for each selected SNP, can be used to calculate the probabilities of each parental combination, i. Once these values have been calculated, the genotype of the first sibling eliminates possible parental genotypic candidates (Figure 7A-C), and the remaining probabilities are normalized.

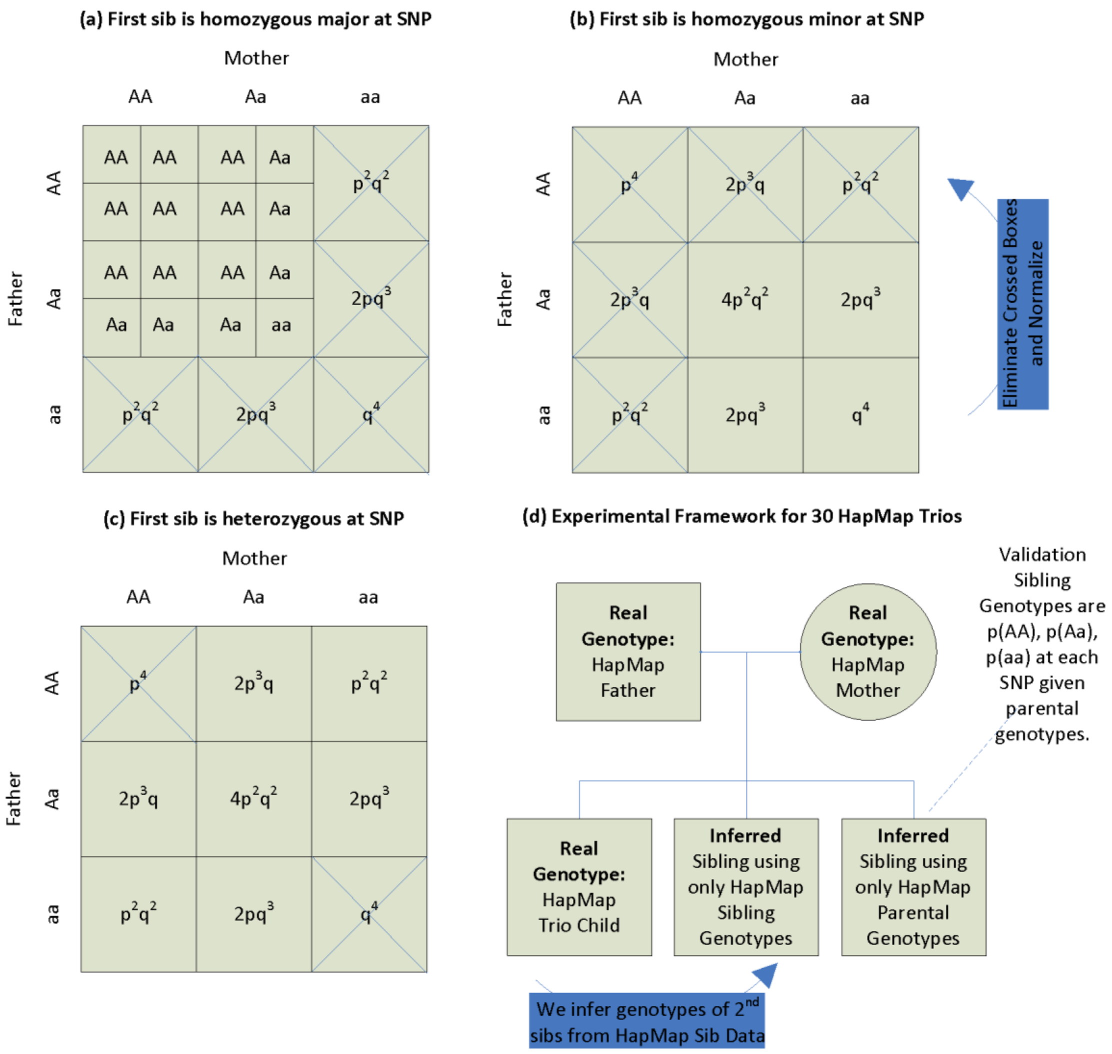

(a) First sib is homozygous major at SNP
Mother
AA
Aa
aa
Father
AA
Aa
aa
AA AA AA Aa
AA AA AA Aa
AA AA AA Aa
Aa Aa Aa aa
p^2q^2
$2pq^3$
p^2q^2
$2pq^3$
q^4
(b) First sib is homozygous minor at SNP
Mother
AA
Aa
aa
Father
AA
Aa
aa
p^4
$2p^3q$
p^2q^2
$2p^3q$
$4p^2q^2$
$2pq^3$
p^2q^2
$2pq^3$
q^4
Eliminate Crossed Boxes and Normalize
(c) First sib is heterozygous at SNP
Mother
AA
Aa
aa
Father
AA
Aa
aa
p^4
$2p^3q$
p^2q^2
$2p^3q$
$4p^2q^2$
$2pq^3$
p^2q^2
$2pq^3$
q^4
(d) Experimental Framework for 30 HapMap Trios
Real Genotype: HapMap Father
Real Genotype: HapMap Mother
Validation Sibling Genotypes are p(AA), p(Aa), p(aa) at each SNP given parental genotypes.
Real Genotype: HapMap Trio Child
Inferred Sibling using only HapMap Sibling Genotypes
Inferred Sibling using only HapMap Parental Genotypes
We infer genotypes of 2nd sibs from HapMap Sib Data

Figure 7: (a-c) Refining mechanism for homozygous major SNPs: when the first sibling is homozygous major (a), homozygous minor (b), or heterozygous (c) at a given SNP, this constrains the possible parental genotypes; in the first case, five of nine parental genotypic combinations can be eliminated (crossed boxes). Using HapMap CEPH SNP population frequencies, *p* and *q*, the probability frequencies are populated for the remaining squares, and normalized. The probability that subsequent sibs will be homozygous major, heterzygous, or homozygous minor can then be calculated using the probabilities that parents would contribute specific allelic values. (d) For each of 30 HapMap CEPH trios, the Sib$_1$ genotype and the SNP population frequencies are used (without the parent genotypes) to infer *p('AA')*, *p('Aa')*, and *p('aa')* for subsequent siblings. Those probabilities are then validated against those that would be expected given only the parental genotypes at each SNP.

Measuring the information content of Sibling genotype data

When calculating the probability of a specific Sib$_2$ genotype given a known Sib$_1$ genotype, it is possible to directly measure the benefit of the proband genotype information in improving Sib$_2$ inferences. This involves measuring the difference between the prior Hardy-Weinberg probability for the genotype, given only population frequencies, and the posterior probability, as calculated by the conditional expression above. To measure the information content provided by the first sibling's genotype, we propose the use of a likelihood ratio test statistic, comparing models where two individuals are known to be siblings versus two individuals that are known to be unrelated. There are a total of nine possible likelihood ratios, $\Lambda_{Ind_1, Ind_2\ genotypes}$, for each of the possible individual genotypic combinations, such as $Ind_1 AA$:

$$\Lambda_{Ind_1, Ind_2\ genotypes} = \frac{p(Ind_2\, genotype | Ind_1 genotype \cap siblings)}{p(Ind_2\, genotype | Ind_1 genotype \cap unrelated)}$$

$$= \frac{p(Sib_2\, genotype | Sib_1 genotype)}{\left(\frac{p(Ind_2\, genotype) \cap p(Ind_1 genotype \cap unrelated))}{p(Ind_1 genotype \cap unrelated)}\right)}$$

$$= \frac{\sum_{i=1}^{9} p(Sib_2\, genotype | parental\ comb. i) p(parental\ comb. i |\ Sib_1 genotype)}{\left(\frac{p(Ind_2 genotype)\ \cap p(Ind_1 genotype \cap unrelated))}{p(Ind_1 genotype \cap unrelated)}\right)}$$

$$= \frac{\sum_{i=1}^{9} \left(\frac{p(Sib_2 genotype \cap parental\ comb. i)}{p(parental\ comb. i)}\right) p(parental\ comb. i |\ Sib_1\ \ genotype)}{\left(\frac{p(Ind_2\, genotype) \cdot p(Ind_1 genotype) \cdot \left(1 - \frac{1}{N}\right)}{p(Ind_1 genotype) \cdot \left(1 - \frac{1}{N}\right)}\right)}$$

$$\cong \frac{\sum_{i=1}^{9} \left(\frac{p(Sib_2 genotype \cap parental\ comb. i)}{p(parental\ comb. i)}\right) p(parental\ comb. i |\ Sib_1\ \ genotype)}{p(Ind_2 genotype)}$$

The denominator becomes *$p(Ind_2$ genotype)*, which is either p^2, $2pq$, or q^2. This is intuitive; when considering two unrelated individuals, the probability that the 2nd has a specific genotype can only be identified using the population frequencies for that genotype. The numerator is the posterior probability expression derived in Table 1, also in terms of *p* and *q*.

Table 1: Sib_2 inference error reduction when Sib_1 genotype is known. The error reduction depends only on the allele frequencies, and at all frequencies, the error is reduced, improving the quality of genotypic inference.

Sib_2	Sib_1	Prior Prob.	Posterior Prob.	Error Reduction
AA	AA	p^2	$p^2 + pq + \frac{1}{4}q^2$	$\lvert p^2 – [p^2 + pq + \frac{1}{4}q^2]\rvert$
Aa	AA	$2pq$	$pq + \frac{1}{2}q^2$	$\lvert 2pq – [pq + \frac{1}{2}q^2]\rvert$
aa	AA	q^2	$\frac{1}{4}q^2$	$\lvert q^2 – [\frac{1}{4}q^2]\rvert$
AA	Aa	p^2	$\frac{1}{2}p^2 + \frac{1}{4}pq$	$\lvert p^2 – [\frac{1}{2}p^2 + \frac{1}{4}pq]\rvert$
Aa	Aa	$2pq$	$\frac{1}{2}p^2 + (2/3)^{-1}pq + \frac{1}{2}q^2$	$\lvert 2pq – [\frac{1}{2}p^2 + (2/3)^{-1}pq + \frac{1}{2}q^2]\rvert$
aa	Aa	q^2	$\frac{1}{4}pq + \frac{1}{2}q^2$	$\lvert q^2 – [\frac{1}{4}pq + \frac{1}{2}q^2]\rvert$
AA	aa	p^2	$\frac{1}{4}p^2$	$\lvert p^2 – [\frac{1}{4}p^2]\rvert$
Aa	aa	$2pq$	$\frac{1}{2}p^2+pq$	$\lvert 2pq – [\frac{1}{2}p^2+pq]\rvert$
aa	aa	q^2	$\frac{1}{4}p^2 + pq + q^2$	$\lvert q^2 – [\frac{1}{4}p^2 + pq + q^2]\rvert$

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 Sib_1 genotype (Figure 8, Figure 9, & Figure 10).

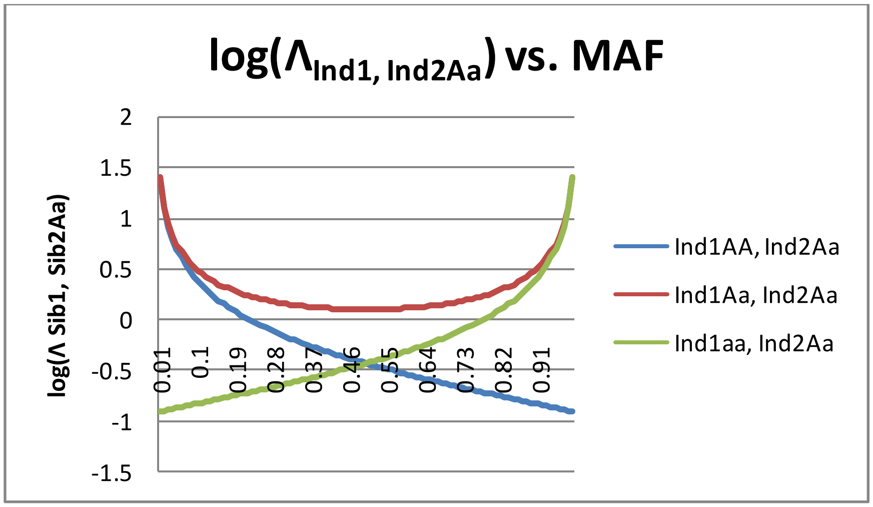


Figure 8: Log likelihood ratio test statistic for sibling inferences: for each Sib_2 genotype, the log likelihood ratio for each possible Sib_1 inference is shown versus Minor Allele Frequency (MAF). These charts describe how informative the Sib_2 genotype of 'Aa' is, when inferring each Sib_1 genotype.

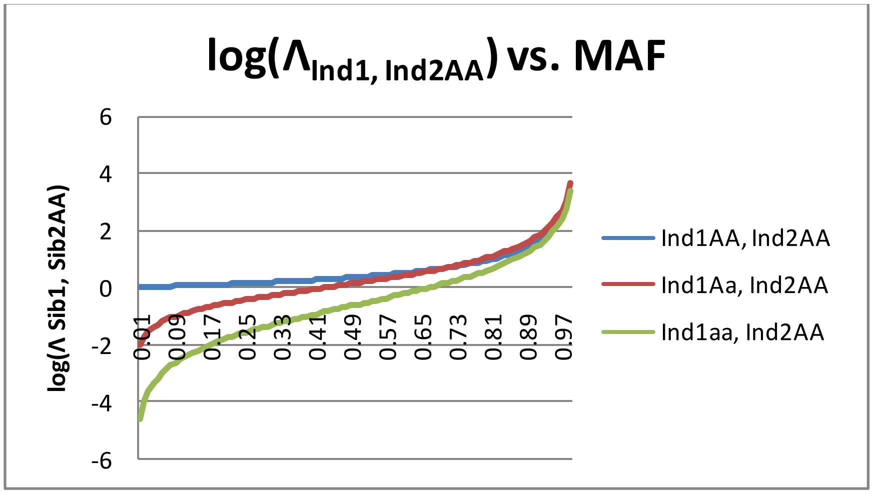


Figure 9: Log likelihood ratio test statistic for sibling inferences: for each Sib_2 genotype, the log likelihood ratio for each possible Sib_1 inference is shown versus Minor Allele Frequency (MAF). These charts describe how informative the Sib_2 genotype of 'AA' is, when inferring each Sib_1 genotype.

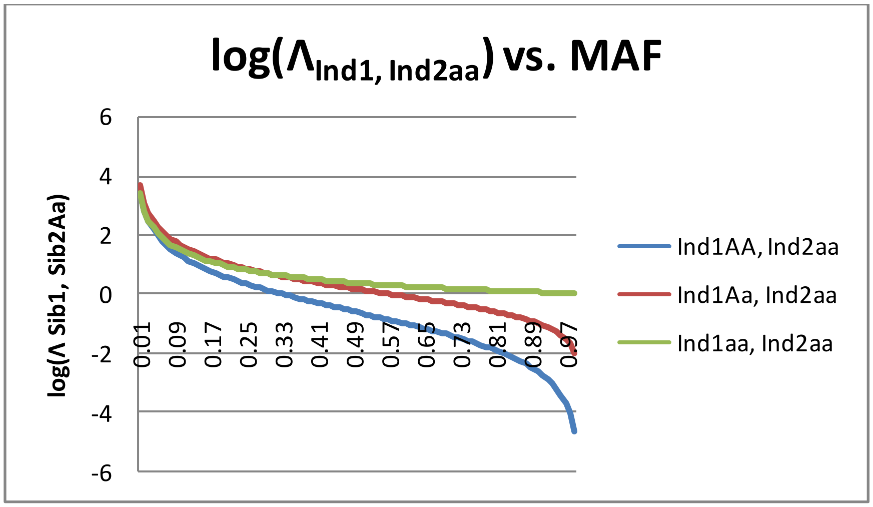


Figure 10: Log likelihood ratio test statistic for sibling inferences: for each Sib_2 genotype, the log likelihood ratio for each possible Sib_1 inference is shown versus Minor Allele Frequency (MAF). These charts describe how informative the Sib_2 genotype of 'aa' is, when inferring each Sib_1 genotype.

The allele frequency, *p*, that maximizes this statistic can then be found numerically for each $\Lambda_{\text{Ind}_1,\text{Ind}_2 \text{ genotypes}}$ expression, to identify which allele frequencies and conditions are most informative for genotypic inferences. These results are below in Table 2.

Table 2: Finding the Minor Allele Frequency (MAF) that maximizes the log likelihood ratio test statistic for each Sib_2 genotypic inference type. The *maximizing MAF* is the allele population frequency at which the most information will be derived about the Sib_2 genotype from Sib_1 under that Sib genotypic combination. Note: There are two equally maximizing MAF values for Log($\Lambda_{\text{Sib1Aa,Sib2Aa}}$), 0.01 and 0.99, both resulting in a value of 1.407.

Sib_2	Sib_1	Maximizing MAF	Log($\Lambda_{\text{Ind1,Ind2 genotypes}}$)
AA	AA	0.01	3.407
Aa	AA	0.01	3.699
aa	AA	0.01	3.389
AA	Aa	0.99	1.396
Aa	Aa	0.01, 0.99	1.407
aa	Aa	0.01	1.396
AA	aa	0.99	3.389
Aa	aa	0.99	3.699
aa	aa	0.99	3.407

Confirming sib-ship with two non-matching sets of SNP genotypes

The above inference technique can be extended to confirm sib-ship in two non-matching samples of SNP sequence data. Given a set of matches at *M* independent loci from a pool of *N* individuals, an expanded form of Bayes Theorem can be used to calculate *p(sibs|match at M loci)* directly, where *!sibs* refers to two individuals not being siblings:

$$p(sibs|match\ at\ M\ loci)$$

$$= \frac{p(match\ at\ M\ loci|sibs)\ p(sibs)}{p(match\ at\ M\ loci|sibs)p(sibs) + p(match\ at\ M\ loci|!\,sibs)p(!\,sibs)}$$

$$= \frac{[p(both\ AA|sibs) + p(both\ Aa|sibs) + p(both\ aa|sibs)]^M \left(\frac{1}{N}\right)}{[p(both\ AA|sibs) + p(both\ Aa|sibs) + p(both\ aa|sibs)]^M \left(\frac{1}{N}\right) + p(match|!\,sibs)^M \left(1 - \frac{1}{N}\right)}$$

p(match|!sibs) can be calculated for each SNP using the population frequency; it is the probability that two unrelated individuals in the population would share the same genotype, *'AA'*, *'Aa'*, or *'aa'*. The expression *p(match|!sibs)* is effectively the same as *p(match)* as long as the sample pool, *N*, is large enough, as the probability of sib-ship is very low in a large pool. For three different pool sizes, (*N=100,000;10,000,000;6,000,000,000)*, we have created a sib-ship probability surface that varies with the number of matched SNPs and minor allele frequency (MAF) of those SNPs (Figure 11a-c) and published supporting values for these probabilities in Table 3. These estimates use the selection of M independent SNPs all with the same a priori known MAF. For SNPs that commonly vary in the population, a small number of genotypic matches are required to confirm sib-ship.

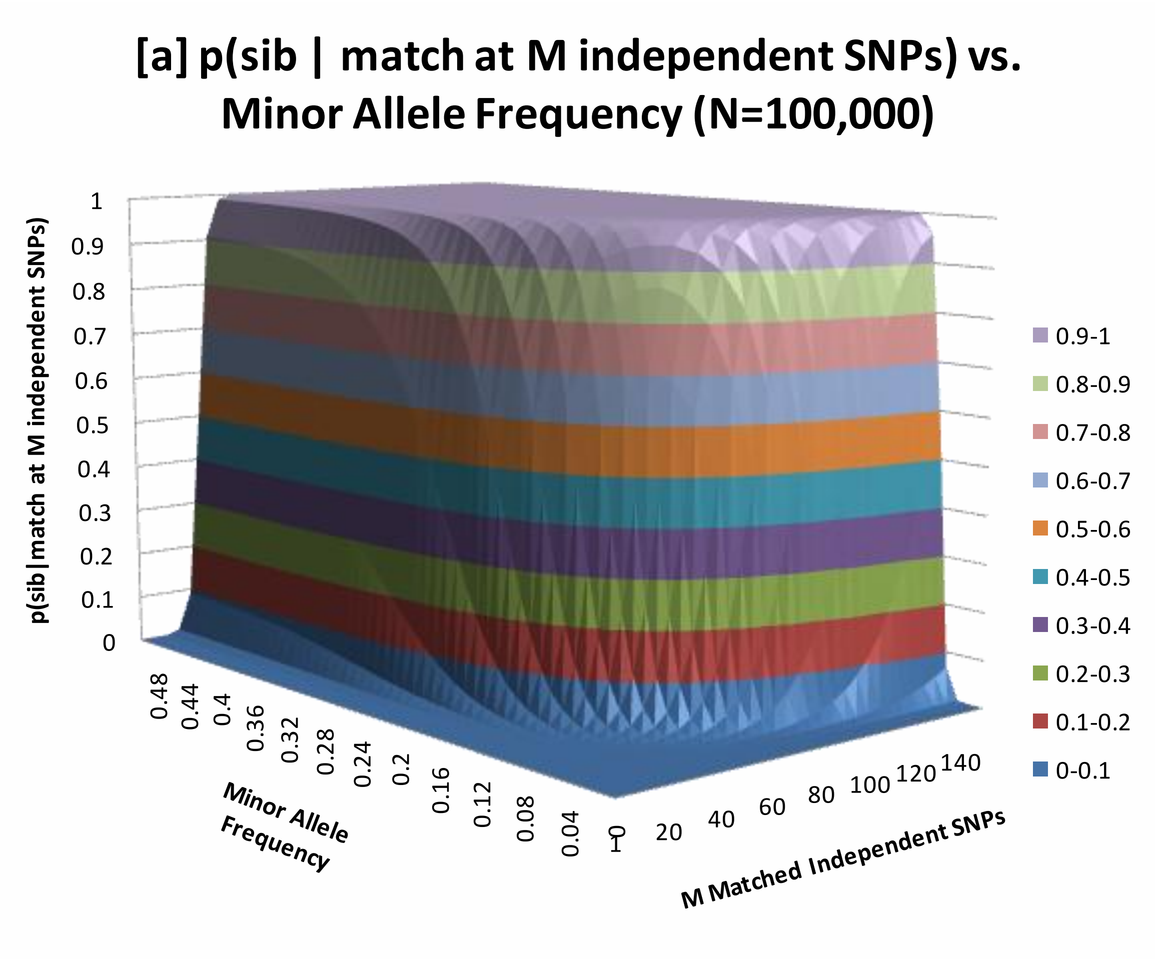


Figure 11-a: Sib-ship identifiability surfaces: these surfaces describe the probability of sib-ship as a function of *M*, the number of matched independent SNPs (between two individuals) and Minor Allele Frequency (MAF). We show this across three sample size pools--N=(a)100,000; (b)10,000,000; (c)6,000,000,000 people. At high MAFs even very large increases in the potential sample pool size will not prevent sib-ship confirmation with relatively few matched SNPs. For example, if loci with MAF=0.25 are selected, the number of matched SNPs to confirm sib-ship with p=0.999 is 50 with a candidate pool of 100,000 and increases to only 80, in a group of 6 billion.

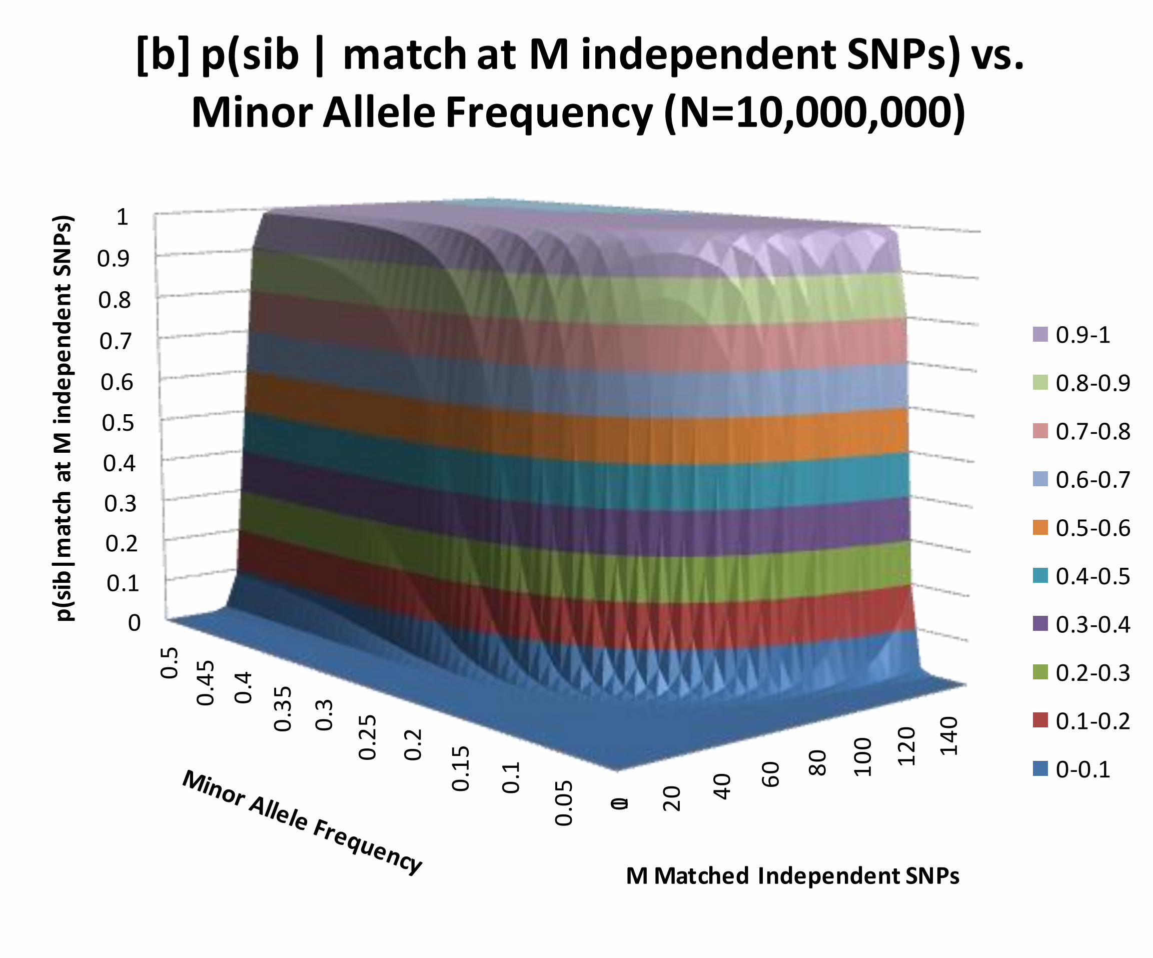


Figure 11-b: Sib-ship identifiability surfaces, continued. Population size (N) = 10,000,000.

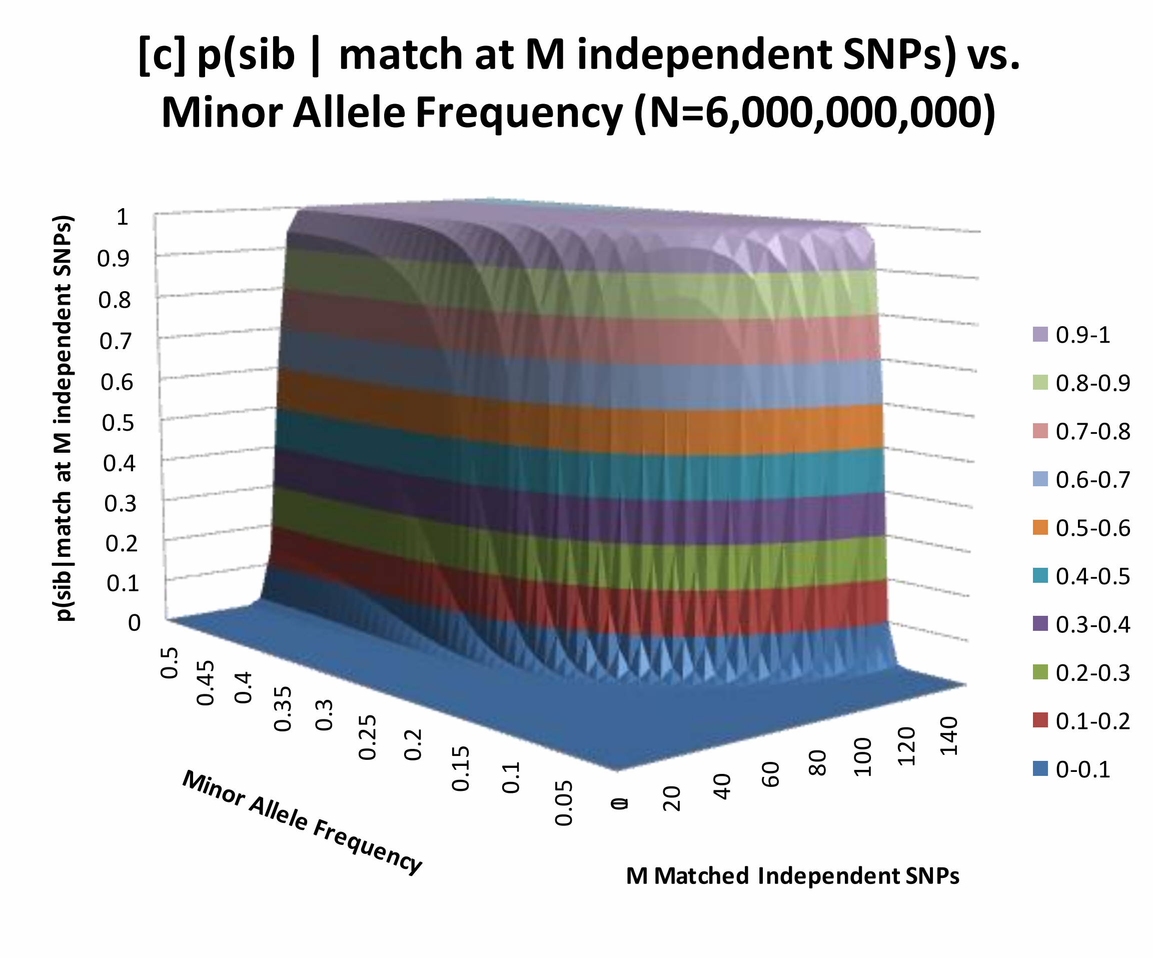


Figure 11-c: Sib-ship identifiability surfaces, continued. Population size (N) = 6,000,000,000.

Table 3: Probability of sib-ship for three pool sizes. In a sample pool of size *N*, provided below, the probability that two individuals are siblings given a match at a subset of SNPs is charted as a function of *M*, the number of independent SNPs that they match at, and the minor allele frequency, *q*, which is known a priori (from population frequency estimates) and is the same for all *M* SNPs. Non-matches are not considered here, and requires separate principle and analysis.

_N_=100,000

Q	M=1	M=10	M=20	M=30	M=40	M=50	M=60	M=70	M=80	M=90
0	0.00001	1E-5	1E-5	1E-5	1E-5	1E-5	1E-5	1E-5	1E-5	1E-5
0.05	1.1E-5	2.67E-5	7.11E-5	0.000189	0.000505	0.001345	0.003578	0.009482	0.024886	0.063706
0.1	1.21E-5	6.64E-5	0.000441	0.002923	0.019099	0.114527	0.462126	0.850907	0.974301	0.996045
0.15	1.31E-5	0.000148	0.002194	0.031572	0.325877	0.87757	0.990679	0.999366	0.999957	0.999997
0.2	1.4E-5	0.000287	0.008152	0.190701	0.871059	0.994863	0.99982	0.999994	1	1
0.25	1.47E-5	0.000472	0.021816	0.512966	0.980292	0.999574	0.999991	1	1	1
0.3	1.52E-5	0.000666	0.042483	0.747176	0.994946	0.999924	0.999999	1	1	1
0.35	1.55E-5	0.000823	0.063574	0.848341	0.997835	0.999974	1	1	1	1
0.4	1.57E-5	0.000924	0.078846	0.88788	0.998637	0.999985	1	1	1	1
0.45	1.58E-5	0.000975	0.086919	0.902796	0.998898	0.999989	1	1	1	1
0.5	1.58E-5	0.000989	0.089295	0.906621	0.998961	0.999989	1	1	1	1

N=10,000,000

Q	M=1	M=10	M=20	M=30	M=40	M=50	M=60	M=70	M=80	M=90
0	1E-7	1E-7	1E-7	1E-7	1E-7	1E-7	1E-7	1E-7	1E-7	1E-7
0.05	1.1E-7	2.67E-7	7.11E-7	1.89E-6	5.05E-6	1.35E-5	3.59E-5	9.57E-5	0.000255	0.00068
0.1	1.21E-7	6.64E-7	4.41E-6	2.93E-5	0.000195	0.001292	0.008518	0.053991	0.274896	0.715775
0.15	1.31E-7	1.48E-6	2.2E-5	0.000326	0.004811	0.066884	0.515231	0.940333	0.995739	0.999711
0.2	1.4E-7	2.87E-6	8.22E-5	0.002351	0.063279	0.659483	0.982308	0.999372	0.999978	0.999999
0.25	1.47E-7	4.72E-6	0.000223	0.010423	0.332172	0.959166	0.999099	0.999981	1	1
0.3	1.52E-7	6.66E-6	0.000443	0.028705	0.663129	0.992431	0.999886	0.999998	1	1
0.35	1.55E-7	8.24E-6	0.000678	0.052974	0.821712	0.997374	0.999968	1	1	1
0.4	1.57E-7	9.25E-6	0.000855	0.073378	0.879899	0.998527	0.999984	1	1	1
0.45	1.58E-7	9.76E-6	0.000951	0.084983	0.900612	0.99887	0.999988	1	1	1
0.5	1.58E-7	9.9E-6	0.00098	0.088497	0.905783	0.998951	0.999989	1	1	1

N=6,000,000,000

Q	M=1	M=10	M=20	M=30	M=40	M=50	M=60	M=70	M=80	M=90
0	1.6E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10	1.67E-10
0.05	1.8E-10	4.44E-10	1.18E-9	3.16E-9	8.42E-9	2.24E-8	5.98E-8	1.6E-7	4.25E-7	1.13E-6
0.1	2.0E-10	1.11E-9	7.35E-9	4.89E-8	3.25E-7	2.16E-6	1.43E-5	9.51E-5	0.000631	0.00418
0.15	2.1E-10	2.47E-9	3.66E-8	5.43E-7	8.06E-6	0.000119	0.001768	0.025594	0.280299	0.852397
0.2	2.3E-10	4.78E-9	1.37E-7	3.93E-6	0.000113	0.003217	0.084701	0.726254	0.987023	0.999542
0.25	2.4E-10	7.87E-9	3.72E-7	1.76E-5	0.000828	0.037674	0.648979	0.988676	0.999758	0.999995
0.3	2.5E-10	1.11E-8	7.39E-7	4.93E-5	0.00327	0.179341	0.935717	0.99897	0.999985	1
0.35	2.5E-10	1.37E-8	1.13E-6	9.32E-5	0.007623	0.387598	0.981185	0.999767	0.999997	1
0.4	2.6E-10	1.54E-8	1.43E-6	0.000132	0.012063	0.530447	0.990523	0.999897	0.999999	1
0.45	2.6E-10	1.63E-8	1.59E-6	0.000155	0.014878	0.595717	0.993092	0.999929	0.999999	1
0.5	2.6E-10	1.65E-8	1.63E-6	0.000162	0.01577	0.613392	0.993675	0.999936	0.999999	1

Modeling a series of SNP inferences using a binomial distribution

A binomial distribution can be used to represent a series of sibling genotypic inferences, such as the probability of correct inferences at 50 SNP loci, if each inference meets specific criteria. Independent inferences can be treated as a random variable with probability *p* of success, as long as independent SNPs are selected, with the same minor allele frequency and Sib$_1$ genotype.

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

where *p(k,n,p)* refers to the probability that *k* correct inferences were made out of *n* attempted inferences when the probability of success for each inference attempt is *p*.

The cumulative binomial measures the probability of reaching up to *k* successes in *n* trials with probability *p* of success at each attempt:

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

If *n* guesses are considered (i.e. *n* SNPs are genotyped and used for sib inference), *1-F(k,n,p)* is the probability that at least *k* of those will be correct.

We already know the expectation of the number of SNP genotypes that will be correctly inferred from the above section (simply the probability of correct inferences at a MAF multiplied by the number of inferences). This cumulative binomial measure helps clarify the probability of guessing a at least a specific number of SNPs correctly.

For example, if we take a set of $n = 100$ SNP inferences where our ability to correctly infer sibling SNP genotypes is $p = 0.8$, and we would like to know what the probability of at least $k = 75$ correct guesses is $(1 - F(k,n,p))$, we can calculate that it is 0.912.

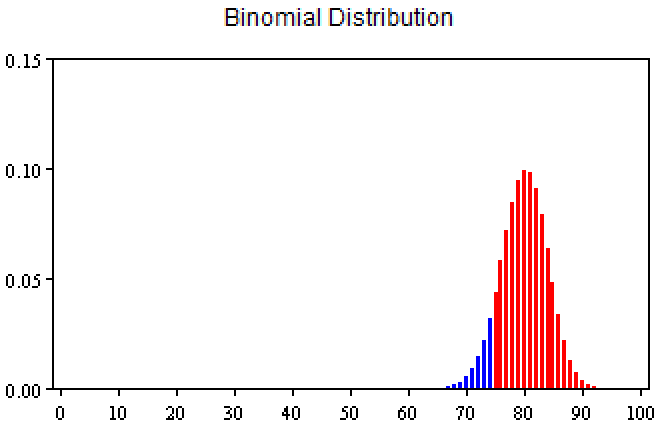


Figure 12: The binomial distribution for number of correct SNP genotype inferences. In this example, we attempt inference of 100 SNP genotypes, each with probability 0.8 of success. We would like to know what the probability is of correctly inferring at least 75 (red) is 0.912. This can be calculated using the 1-F(k,n,p) cumulative binomial distribution formula in the above section.

HapMap CEPH and global population SNP genotypes and allele frequency data

The demographic data used in this project are population-specific SNP allele frequencies from the CEPH HapMap population, Utah residents with ancestry from northern and western Europe, and the global SNP allele frequencies (from all populations that participated in the HapMap) [99]. The HapMap project has compiled allele frequency values for a large selection of SNPs – loci in the genome that account for a great deal of genetic variability in populations. Within the CEPH population, there are 30 familial trios, each containing one mother, father, and child. Additionally, the individual genotypes of the 90 CEPH trio participants are directly used in this study. One limitation of this population specific allele frequency database is the small size of each HapMap population – the CEPH population contains 90 participants, and as such, each trio child contributes 1/90th of the allele frequency data that are used in the study.

Validating the sibling genotype probability vector using parental genotypic data

To validate the results of the refining strategy on inferring the second sibling genotype, the authentic parental genotypes are used to create the probability vector *p('AA')*, *p('Aa')*, *p('aa')* at the SNP being evaluated, for the children the pair would be expected to have. For each of the trio pairs at each of the SNPs being tested, the probability vector was calculated.

Error reduction calculation

The error reduction measurement identifies the extent to which inference error is reduced. For example, when trying to infer the probability that Sib_2 has genotype *'AA'* at a specific SNP, we calculate the absolute value of the difference between our best inference and the Hardy Weinberg probability for Sib_2 to have genotype *'AA'*, using population-specific allele frequency data and the Sib_1 genotype,

$|p(Sib_2AA|Sib_1\ genotype)\text{-}p(Sib_2AA)|$. This value is specifically the percentage improvement to the probability value from the new data, when inferring the specific event that Sib_2 will have genotype *'AA'* and Sib_1 will have the specific genotype in question.

Any change to $p(Sib_2AA)$ must also correspond with the opposite change in the sum of $p(Sib_2Aa)$ and $p(Sib_2aa)$. To accurately represent the overall error reduction by Sib_1 genotype, with any of three possible Sib_2 genotypes, the average of the three values is measured. For example, where the Sib_1 genotype is *'AA'*, the overall average improvement (and error reduction) is the average of $|p(Sib_2AA) - p(Sib_2AA|Sib_1AA)|$, $|p(Sib_2Aa) - p(Sib_2Aa|Sib_1AA)|$, and $|p(Sib_2aa) - p(Sib_2aa|Sib_1AA)|$. The percentage improvement is graphed in Figure 14 for each possible Sib_1 genotype.

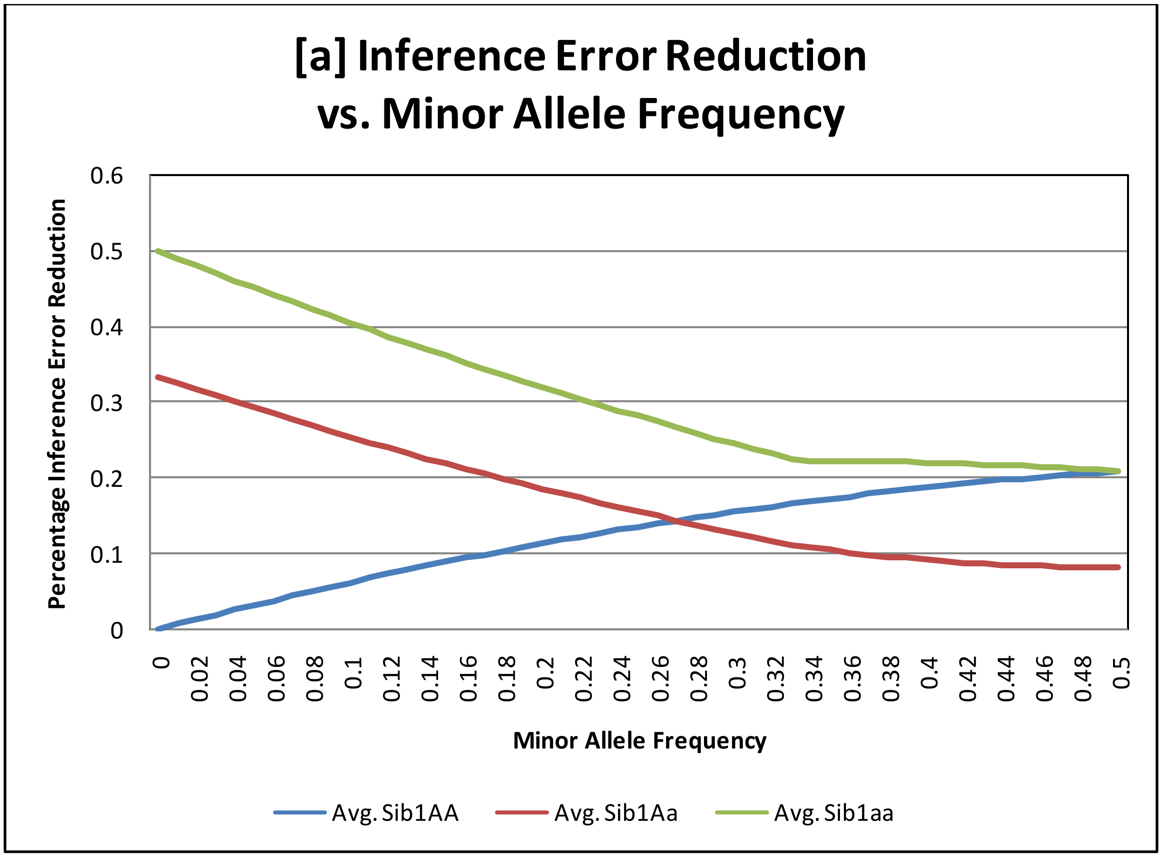


Figure 13: The error reduction, in the form of percentage improvement in inference accuracy for subsequent siblings, when one sibling's (Sib_1) genotype is available (for each possible Sib_1 genotype).

Scoring metric for calculating correct fraction of inferences

To ascertain whether the inferences are helpful for producing correct answers, a scoring metric was used to calculate the fraction of correct SNP inferences, in our empirical inference validation study. For each SNP inference, the scoring metric provides a full point when the plural entry in the inference vector, (the maximum of *p('AA')*, *p('Aa')*, and *p('aa')*, and thus the predicted sib genotype), matches the plural entry in the parental validation vector (the empirical most likely genotype). Given the parental genotype values, it is possible, and not infrequent, that a validation probability vector has two matching plural values, for example, if *p('AA')* = *p('Aa')* = 0.5. When this is the case, one half point was awarded if the plural value in the inference vector matched one of the two validation choices, to signify that one of the two equally likely candidates was chosen.

There are some conditions that arise from use of a simple scoring metric, where it becomes difficult to score well. For example, a heterozygous Sib_1 will likely result in a 0.5 score for inferences. A score of 1 point would be possible if one parent had a genotype of *'AA'* and the other had genotype *'aa'*, making the probability that the parents would have a child with genotype *'Aa'* equal 1. Most remaining parental combinations would not result in the probability of child genotype *'Aa'* equal to 1, and would likely result in only a half point. These values can be adjusted using machine learning techniques or more robust decision making algorithms, but those are out of the scope of this work.

Results

Validation of SNP genotype inference using HapMap trio data

We then empirically infer sibling genotypic sequences from HapMap trio child genotypes using the above technique. At 700,000 SNP loci on chromosomes 2, 4, and 7, in each of 30 HapMap CEPH trios, the trio sibling, Sib_1, known genotypes are combined with the CEPH and global HapMap SNP allele frequencies to produce genotypic inferences of a hypothetical sib, Sib_2, at these loci. The inference method produces three genotypic probabilities for Sib_2 (or subsequent siblings): $p(Sib_2AA|Sib_1\ genotype)$, $p(Sib_2Aa|Sib_1\ genotype)$, and $p(Sib_2aa|Sib_1\ genotype)$ for each SNP, which we call the SNP probability vector.

The ability to correctly infer a sibling genotype from a trio child genotype can be validated by comparing whether the best estimated genotype, using only the sibling genotype and population frequencies, matches the best estimated genotype using the parental genotypic data (Fig. 1D). While there are CEPH families where multiple children are genotyped, there are not many, and to get the statistical power necessary for our analysis, we needed to use the trios and impute sibling values. We do this by comparing the plural, largest, value in the SNP probability vector, with the plurality value in the SNP probability vector that would be expected given the parental genotypes and Mendelian Inheritance. The fraction of correct inferences for SNPs where the Sib_1 is homozygous major or heterozygous versus minor allele frequency are graphed in Figure 14A and Figure 14B, respectively. There were insufficient SNPs where the trio child was homozygous minor, so they have been excluded from this analysis.

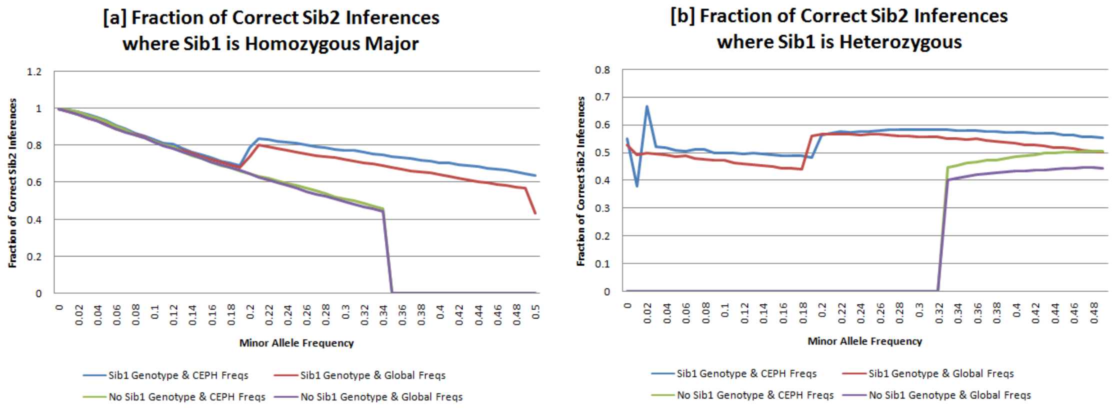


Figure 14: Fraction of correct Sib2 inferences: the fraction of Sib2 SNPs that can be correctly identified when Sib1 is (a)homozygous major or (b)heterozygous. Each line represents use of distinct data--inclusion or exclusion of Sib1 genotypes, and use of population-specific or global allele frequency data. Without Sib1 genotypes, homozygous major inferences would always be incorrect at Minor Allele Frequency (MAF) ≥ 0.33 and heterozygous inferences would always incorrect at MAF ≤ 0.33. At many allele frequencies, use of Sib1 genotypes dramatically improves Sib2 inferences.

For inferences at SNPs where the trio child, Sib_1, was homozygous major, with MAF < 0.05 (N=300512,43.2%), we are able to correctly infer the genotype of other siblings, e.g. Sib_2, with 98.5% accuracy when using population-specific allele frequency data. At SNPs with MAF < 0.20 (N=452684,65.1%) we achieve 91.9% average accuracy. For SNPs where the first sibling is heterozygous, with MAF > 0.20 (N=125796,18.1%), it is possible to infer the correct genotype of the second sibling with 57.7% average accuracy. Without Sib_1 genotypes, all inferences for homozygous major SNPs with MAF ≥ 0.33 and heterozygous SNPs with MAF ≤ 0.33 would be incorrect when validated against plural parental values. At these allele frequencies, as well as others, use of Sib_1 genotypes markedly improves Sib_2 inferences.

Deriving propensity to disease from sibling SNP data

Additionally, sibling SNP data can be used to quantify an individual's disease propensity through genotypic inference, without that individual's actual sequence data. For example, the likelihood ratio test statistic above may also be used to describe relative risk, using a multiplicative model.

$$\Gamma_{Sib_2 genotype \ \ |Sib_1 genotype} = \frac{\text{probability with sibling knowledge}}{\text{probability without sibling knowledge}}$$

$$= \frac{p(Sib_2 genotype|Sib_1 genotype)}{p(Sib_2 genotype)}$$

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

For example, the relative risk of Sib_2Aa, carrying one copy of the disease allele 'a', is provided by information from the Sib_1aa genotype:

$$\Gamma_{Aa\,|Sib_1 aa} = \frac{p(Sib_2 Aa|Sib_1 aa)}{p(Sib_2 Aa)}$$

$$= \frac{½\mathrm{p}^2 + \mathrm{pq}}{2\mathrm{pq}}$$

$$= \frac{½\mathrm{p} + (1 - \mathrm{p})}{2(1 - \mathrm{p})}$$

$$= \frac{1 - ½\mathrm{p}}{2 - 2\mathrm{p}}$$

In this example, at MAF=0.01, the relative risk of genotype 'Aa' is 25.25, given information that Sib_1 carries genotype 'aa' at that locus. However, at MAF=0.5, the relative risk of genotype 'Aa' is 0.75, given information that Sib_1 carries genotype 'aa', explaining that the risk of having the genotype 'Aa' is reduced at this MAF. This may seem counterintuitive, as the risk of carrying a disease allele is actually higher at this MAF, but Sib_2 carrying genotype 'Aa' is lower than in the control population, while the relative risk of carrying the disease allele with genotype 'aa' is higher.

$$\Gamma_{aa\,|Sib_1 aa} = \frac{p(Sib_2 aa|Sib_1 aa)}{p(Sib_2 aa)}$$

$$= \frac{\tfrac{1}{4}\mathrm{p}^2 + \mathrm{pq} + \mathrm{q}^2}{\mathrm{q}^2}$$

$$= \frac{\tfrac{1}{4}\mathrm{p}^2 + \mathrm{p}(1-\mathrm{p}) + (1-\mathrm{p})^2}{(1-\mathrm{p})^2}$$

At MAF 0.5, $\Gamma_{aa\,|Sib_1 aa}$ is 2.25, demonstrating that it is more likely that a disease allele will be carried by Sib_2 in genotype 'aa' than in the control population given the Sib_1 genotype.

The explicit probability of developing a disease is also altered. If an individual with genotype *'Aa'* at a specific locus has a probability p_d of developing a disease by age *a*, and that individual has a probability p_s of having that genotype given his sibling's genotype at that locus, his probability of developing that disease by age *a* is $p_s.p_d$. This can easily be extended to multiple independent loci, important for diseases in which a set of common or rare variants dictates disease likelihood[6, 101]. As SNPs are both clinically informative and there is a wealth of supporting allele frequency data, they have been the focus of our analysis, however there are other genomic data types which should be considered in a rigorous privacy and propensity analysis, including copy number variant and mutation data.

Discussion

These findings demonstrate that substantial discrimination and privacy concerns arise from use of inferred familial genomic data. While the Genetic Information Nondiscrimination Act of 2008 (GINA, H.R. 493), recently passed into law, would mitigate the threat of direct discriminatory action by employers or insurers [26], there will continue to be other uses of genomic data that pose privacy risks, including the use of genetic testing in setting life, disability, and long-term care insurance premiums [27, 102, 103]. Familial genotypic sequences can be used to assist in forensic or criminal investigations for indirect identification of genotype, increasing the number of people who may be identified [28, 29]. Similarly, Freedom of Information Act (FOIA) [30] requests related to federally-funded genome wide association studies could potentially be used to identify research participants and their family members. Clinically, choosing the detail and type of disease propensity information that must be disclosed to patients and their potentially affected family members is also under debate [31, 32, 104].

Quantifying the information content of disclosed genomic data will add clarity to the informed consent process when a patient shares genotypic data for research use. For research investigations, it is conceivable that a subject would want to limit the impact of her genomic disclosure on her family members, or be asked to have a discussion with specific family members before proceeding. Providing subjects with different levels of genomic anonymity based on their sequence data, along with an estimate of the probability of re-identification and familial impact for each of those anonymity levels, will allow patients to trade off altruistically motivated sharing [105] with privacy consideration, especially when they volunteer to share all the variants in their genome [17].

While the inference accuracy rates are very high, particularly for inferences where Sib_1 has a homozygous major genotype, we would like to caution that some of these findings are not always highly informative. For example, if the MAF is 0.01, where 99% of the alleles in the population are the major allele, the prior probability for a homozygous major allele is 0.99*0.99 = 0.98. If Sib_1 has a homozygous major allele, the posterior probability of observing a homozygous major allele in another sibling is (¼ + ¼*0.99*0.99 + ½*0.99) ≈ 0.99. In this case, the difference between prior and posterior probabilities is only 0.01, and knowledge of the Sib_1 genotype provides very little information, as most accuracy comes from the allele frequency in the population.

However, homozygous minor alleles are much more informative. With a MAF of 0.2, if Sib_1 has a homozygous minor genotype, the probability of Sib_2 having the same genotype, given only the reference population is 0.04. Given that Sib_1 has a homozygous minor genotype, Sib_2 will have a homozygous minor allele with probability of (¼ + ¼*0.2*0.2 + ½*0.2) = 0.36, which is quite different from the prior probability of 0.04.

One limitation of this study is that the population-based estimates for MAF rely on the HapMap study population sizes, which, at present, are small, though these types of sources will continue to expand. For example, the CEPH population contains 90 participants, so each trio child contributes 1/90th of the allele frequency data used in the study. This approach also depends on the independence of the loci considered, and would need to be adapted for SNPs that are in linkage disequilibrium. Extending this study to include linked SNP loci is possible, using the haplotype block information for HapMap populations that is available. To ensure that SNPs are independent, linkage data from the HapMap population can be used to confirm independence, and SNPs that are far from one another may be selected. Additionally, this approach does not consider the possibility of genotypic errors, which may be common on some platforms. An adjustment using a binomial probability distribution could be used to account for possible errors.

Conclusions

Technologies for sequencing large numbers of SNPs are rapidly dropping in cost, which will help realize the promise of personalized medicine, but pose substantial personal and familial privacy risks. While electronic storage and transmission of genetic tests is not yet a common component of medical record data, these tests will soon be stored in electronic medical records and personally controlled health records [50]. This mandates the need for improved informed consent models and access control mechanisms for genomic data. The increasingly common practice of electronically publishing research-related SNP data requires a delicate balance between the enormous potential benefits of shared genomic data through NCBI and other resources, and the privacy rights of both sequenced individuals and their family members.

Ability to infer SNP genotypes from parental or child data

Similarly, improvements to genomic inferences are possible for paternal relationships: knowledge of a parent's genotype can improve the accuracy of estimates of a child's genotype. For example, consider the case where a child's mother is known to have genotype 'AA' at a variant locus. In this case, we can alter the probability that the child will have genotype 'AA', 'Aa' or 'aa' at that locus, given knowledge of the maternal genotype. For example, the probability that a child has genotype 'AA' given that the mother has genotype 'AA' at a specific locus can be directly calculated:

$$p(Child\ AA|Mother\ AA) = \frac{p(Child\ AA \cap Mother\ AA)}{p(Mother\ AA)}$$

$$= \sum_{i=1}^{3} \left(\frac{p((Child\ AA \cap Mother\ AA)|paternal\ genotype\ i)p(paternal\ genotype\ i)}{p(Mother\ AA)} \right)$$

$$= \left(\frac{p((AA_C \cap AA_M)|AA_F)p(AA_F)}{p(AA_M)} \right) + \left(\frac{p((AA_C \cap AA_M)|Aa_F)p(Aa_F)}{p(AA_M)} \right)$$

$$+ \left(\frac{p((AA_C \cap AA_M)|aa_F)p(aa_F)}{p(AA_M)} \right)$$

$$= \frac{(1)(p^4) + \left(\frac{1}{2}\right)(2p^3q) + (0)(p^2q^2)}{(p^2)}$$

$$= (p^2) + (pq)$$

$$= p^2[+pq]\ .$$

Using only the population allele frequencies, it is possible to determine the improvement of a SNP inference given maternal or paternal genotype at that locus. Before knowledge of the maternal genotype was included, *p(Child AA)* would have been the Hardy-Weinberg frequency for major homozygotes, p^2. However, with knowledge of the maternal genotype, *p(Child AA|Mother AA)*, the additional constraint increases the probability to *p2+pq*, increasing inference accuracy by *pq*.

Now consider the opposite case, where we attempt to infer the genotype of the mother given a known child genotype. Consider the analogue of the above example, where we would like to determine the probability that the mother has genotype 'AA' at a locus given that her child has genotype 'AA':

$$p(Mother\ AA|Child\ AA) = \frac{p(Mother\ AA \cap Child\ AA)}{p(Child\ AA)}$$

$$= \frac{p(Mother\ AA \cap Child\ AA)}{p(Mother\ AA)}$$

Because *p(Mother AA) = p(Child AA)* with no other knowledge, we can substitute it in the denominator, as follows:

$$= \frac{p(Mother\ AA \cap Child\ AA)}{p(Child\ AA)}$$

$$= p(Child\ AA|Mother\ AA)$$

Thus, if non-paternity and new mutations are excluded, *p(Child X|Mother Y)* = *p(Mother X|Child Y)*, where *X* and *Y* are genotypes, and where *X* may be the same genotype as *Y*.

For all of the possible combinations of known parent genotypes and possible inferred child genotypes, the prior and posterior probabilities are enumerated in Table 4 below.

Table 4: Error reduction on genomic inference when the genotype of one known parent is known.

Child	Known Parent	Prior Prob.	Posterior Prob.	Error Reduction
AA	AA	p^2	$p^2 + pq$	$\lvert p^2 – [p^2 + pq]\rvert$
Aa	AA	2pq	$pq + q^2$	$\lvert 2pq – [pq + q^2]\rvert$
Aa	AA	q^2	0	$\lvert q^2\rvert$
AA	Aa	p^2	$\frac{1}{2}p^2+\frac{1}{2}pq$	$\lvert p^2 – [\frac{1}{2}p^2+\frac{1}{2}pq]\rvert$
Aa	Aa	2pq	$\frac{1}{2}p^2+pq+\frac{1}{2}q^2$	$\lvert 2pq – [\frac{1}{2}p^2+pq+\frac{1}{2}q^2]\rvert$
Aa	Aa	q^2	$\frac{1}{2}pq+\frac{1}{2}q^2$	$\lvert q^2-[\frac{1}{2}pq+\frac{1}{2}q^2]\rvert$
AA	aa	p^2	0	$\lvert p^2\rvert$
Aa	aa	2pq	$p^2 + pq$	$\lvert 2pq – [p^2 + pq]\rvert$
Aa	aa	q^2	$pq + q^2$	$\lvert q^2-[pq + q^2]\rvert$

Likelihood ratio test statistic for paternity and information content

The likelihood ratio test statistic explored above for siblings can be employed for inferences that use other familial data for inferences. This technique describes both how informative the genotypic inference technique is in each case, and at each MAF, and can also be used as a statistic for likelihood of paternity. The likelihood ratio again compares two models –one where the known parent genotype is considered and one where it is not. There are a total of nine possible likelihood ratios, $\Lambda_{Ind_1, Ind_2\ genotypes}$, for each of the possible individual genotypic combinations. Consider an example test statistic where we explore the benefit of Maternal genotype knowledge, *MotherAA*, on the inference of a *ChildAA* genotype:

$$\Lambda_{Ind_1, Ind_2\ genotypes} = \frac{p(Ind_2\, genotype|Ind_1 genotype \cap paternal\ relationship)}{p(Ind_2\, genotype|Ind_1 genotype \cap unrelated)}$$

$$= \frac{p(Child\ AA|Mother\ AA)}{\left(\frac{p(Ind_2\, genotype) \cap p(Ind_1 genotype\ \cap unrelated))}{p(Ind_1 genotype \cap unrelated)}\right)}$$

$$= \frac{\left(\frac{p(Mother\ AA \cap Child\ AA)}{p(Child\ AA)}\right)}{\left(\frac{p(Ind_2\, genotype) \cdot p(Ind_1 genotype) \cdot \left(1 - \frac{1}{N}\right)}{p(Ind_1 genotype) \cdot \left(1 - \frac{1}{N}\right)}\right)}$$

$$= \frac{\sum_{i=1}^{3} \left(\frac{p((Child\ AA \cap Mother\ AA)|paternal\ genotype\ i)p(paternal\ genotype\ i)}{p(Mother\ AA)}\right)}{\left(\frac{p(Ind_2 genotype) \cdot p(Ind_1 genotype) \cdot \left(1 - \frac{1}{N}\right)}{p(Ind_1 genotype) \cdot \left(1 - \frac{1}{N}\right)}\right)}$$

$$\cong \frac{\sum_{i=1}^{3} \left(\frac{p((Child\ AA \cap Mother\ AA)|paternal\ genotype\ i)p(paternal\ genotype\ i)}{p(Mother\ AA)}\right)}{p(Ind_2\, genotype)}$$

As above, we have a denominator that becomes *p(Ind$_2$ genotype)*, which is either p^2, $2pq$, or q^2. This intuitively makes sense, when considering two unrelated individuals, the probability that the 2nd has a specific genotype can only be identified using the population frequencies for that genotype. The numerator is the posterior probability expression derived in Table 4, also in terms of *p* and *q*. The log of this odds ratio can then be used as a statistic for measuring relatedness when the nature of the relationship is a priori known to be paternal.

Deriving paternal and child propensity to disease from SNP data

Additionally, paternal SNP data can be used to quantify an individual's disease propensity through genotypic inference, without that individual's actual sequence data. As above, the likelihood ratio test statistic above may also be used to describe relative risk, using a multiplicative model.

$$\Gamma_{child\ \ genotype\ \ \ |paternal\ \ \ genotype} = \frac{\text{probability with paternal knowledge}}{\text{probability without paternal knowledge (control)}}$$

$$= \frac{p(child\ genotype|paternal\ genotype)}{p(child\ genotype)}$$

$$= \frac{\sum_{i=1}^{3}\left(\frac{p(paternal\ genotype \cap parental\ comb.\,i)}{p(parental\ comb.\,i)}\right)p(parental\ comb.\,i|\ child\ genotype)}{p(child\ genotype)}$$

In this example, there are only three elements in the summation rather than nine, because there are only three possible parental genotype combinations (*i*) when one parental genotype is fixed.

Risk of re-identification analysis of mutation data

Introduction

Sequencing of an individual's DNA may reveal single nucleotide variants that have not been documented or previously identified as SNPs. These variants include nonsense and missense mutations, insertions or deletions, and other lesions. Presence of such mutation data in a shared or published sequence substantially increases the ability to identify the individual whose data are shared.

In the case of a de novo germline mutation, we can evaluate privacy implications for carrying that mutation. We first explore general identifiability issues for mutant loci, and how likely a match would be among 1000 people. If a mutation is not de novo, we will need to adjust our estimate with population genetics using population size and estimates of prevalence in the population.

De novo germline mutations

For de novo germline mutations that are not distributed widely in the population, we explore the use of region-specific mutation frequency information in the genome to estimate how common such a mutation might be. We will be gender-neural in our analysis, and consider autosomes only.

Generally, our approach will be to treat a mutation as a rare allele with frequency q. We will estimate the allele frequency of that mutant in a specific population using locus-specific and mutation type information, and estimate *p('Aa')*, the probability of heterozygotes in the population. We then calculate the probability of a match of that mutated base pair in a second person, m_i, and then calculate the probability that those two people are the same given a match observed with probability m_i .

Let the population frequency of the mutant variant at locus i be q_i:

$$q_i = (r_{region\ ,type}\ ) \cdot P_{sub\ -type}$$

where $r_{region,type}$ is the region-specific, type specific mutation rate per base pair, per generation and $P_{sub\text{-}type}$ is the probability of the specific sub-type of the mutation class, $r_{region,type}$ (normalized by type). An example of $r_{region,type}$ is the 'Transition mutation rate in a CpG locus' and an example of the $P_{sub\text{-}type}$ is 'A→G mutation rate for all transition mutations at a CpG locus'.

Mutation type and region-specific data sources

There are a number of data sources for location-specific and type-specific mutation rate data. For our study, we selected mutation rates that included location-specific data – specifically whether the locus in question is in a CpG region – and mutation type-specific data – whether this mutation was a transition or transversion mutation, or a length mutation. These are calculated using a population genetics approach with a divergence time of 5 mya, an ancestral effective population size of 10^4, a generation length of 20 yrs, and rates of molecular evolution, detailed in Table 5. There are also a growing number of population-specific mutation databases that are available for this type of analysis, segregated by nationality, racial, and ethnic group [106, 107].

Table 5: Mutation rate estimates. These mutation rate estimates are suitable for $r_{region,type}$, as they are rates for regions (considering a CpG or non-CpG locus) and mutation type. Rates calculated on the basis of a divergence time of 5 mya, ancestral population size of 10^4, generation length of 20 yrs, and rates of molecular evolution. [http://www.genetics.org/cgi/content/full/156/1/297/T4]

Mutation type	Mutation rate
Transition at CpG	1.6 x 10-7
Transversion at CpG	4.4 x 10-8
Transition at non-CpG	1.2 x 10-8
Transversion at non-CpG	5.5 x 10-9
All nucleotide substitutions	2.3 x 10-8
Length mutations	2.3 x 10-9
All mutations	2.5 x 10-8

Specific mutation type information, $P_{sub\text{-}type}$, was collected and interpreted using data from the Cardiff Human Gene Mutation Database (Table 6, Table 7, & Table 8). We can calculate $p_{sub\text{-}type}$, which must be normalized among all main type mutations ($r_{region,\ type}$). This information is collected and curated from reports of rare mutation findings that are *not* SNPs – usually occurring in just one or two families – so counts of mutation loci are statistically acceptable in this case, even though some mutations may occur with some multiplicity, for estimates of mutation type specific rates.

Table 6: HGMD Statistics for Missense Mutations ($P_{sub-type}$). This table details the counts of each identified missense mutation from the Cardiff Human Gene Mutation Database.

Wild type	G	T	A	C	Total
Guanine	--	2228	7140	2290	11658
Thymine	1481	--	1045	3609	6135
Adenine	2947	734	--	1048	4839
Cytosine	1619	4785	1376	--	7780

Table 7: HGMD Statistics for Nonsense Mutations ($P_{sub-type}$). This table details the counts of each identified nonsense mutation from the Cardiff Human Gene Mutation Database.

Wild type	G	T	A	C	Total
Guanine	--	1009	1028	0	2037
Thymine	224	--	325	0	549
Adenine	0	273	--	0	339
Cytosine	499	3178	727	--	4817

Table 8: HGMD Statistics for All Transition Missense Mutations ($P_{sub-type}$). This table details the counts of each identified transition missense mutation from the Cardiff Human Gene Mutation Database.

Wild type	G	T	A	C	Total
Guanine	--	***	7140	***	7140
Thymine	***	--	***	3609	3609
Adenine	2947	***	--	***	2947
Cytosine	***	4785	***	--	4785

The Cardiff Human Gene Mutation Database endorses a technical correction that can compensate for this bias in counts, published in the American Journal of Human Genetics [108]. Krawczak et al created tables describing the results of the meta-analysis of base-pair substitutions in the Human Gene Mutation Database. These tables contain the relative clinical observation likelihood (RCOL) value for each mutation type that is adjacent to each possible mono and dinucleotide sequence, flanked by 0-4 other nucleotides [109]. There is also a table that enumerates the RCOL values for amino acid substitutions by a chemical difference metric, and also contains an RCOL value for nonsense mutations. These data may serve as additional sources for mutation identifiability research, and can be used to create estimates of q_i as described above.

Probability of finding a match in rare mutation alleles

Technically, the probability that two people carry the same allele must include the possibility that either of them is homozygotic minor or heterozygotic at the locus. For de novo mutations, carrying a homozygotic minor genotype is extremely unlikely, so we exclude this for very small frequencies. We then can estimate that the probability that two unrelated people with mutations matching at any locus, m_i , is the frequency of heterozygotes, from $2p_i$ (1-p_i):

$$\mu_i = p(match\ homozygote\,minor) + p(match\ heterzygote)$$

$$= (q_i{}^2)^2 + \big(2(p_i)(q_i)\big)^2$$

$$\cong \big(2(p_i)(1-p_i)\big)^2 + \varepsilon$$

$$\cong \big(2p_i{-}2p_i{}^2\big)^2 + \varepsilon$$

$$\cong (4p_i{}^2) + \varepsilon$$

Probability that two people are the same given a match at M mutant base pairs

We then evaluate the probability of a match at M mutant loci using Bayes' Theorem. For this approach, all M mutant loci must be statistically independent, which in this case means is that they cannot be in linkage disequilibrium, or otherwise correlated in their likelihood to occur.

Suppose that an adversary assumes a conservative model that research subjects are uniformly sampled from a population of *N* people. The probability that a person is subject *i*, given that they share a set of *M* mutant loci is:

$$p(same|match) = \frac{p(match|same)p(same)}{p(match|same)p(same) + p(match|!same)p(!same)}$$

$$= \frac{1\left(\frac{1}{N}\right)}{1\left(\frac{1}{N}\right) + \left(\prod_{i=1}^{M} \mu_i\right)\left(1 - \frac{1}{N}\right)}$$

The probability that two samples came from the same individual *p(same|match)* is exceedingly high given a match at just a few rare mutant variants.

Likelihood of identifying an individual out of 10000 genotyped at that locus

One way to contextualize how individually identifying a mutation is would be to describe how likely it is that two people within a specific population pool would match. For example, "how identifiable is a Caucasian male with a missense A→G mutation in a CpG locus among 10000 others?"

With genome region-specific information, (the mutation is at a CpG locus,) and knowledge of the mutation type (a transition mutation, because it is a purine/purine mutation,) we can evaluate this example by first calculating the 'minor allele frequency' estimate for this mutant variant:

$q_i = (p_{\text{sub-type}})(r_{\text{region,type}})$

$q_i = (0.386)(1.6 \text{ x } 10^{-7})$

$q_i = \boldsymbol{6.2 \times 10^{-8}}$

Next, we can use that minor variant allele frequency to calculate the probability that two people would match

$\mu_i = (2(6.2 \times 10^{-8})(1\text{-}6.2 \times 10^{-8}))^2 + \varepsilon$

$\mu_i = (1.54 \times 10^{-14}) + \varepsilon$

$$p(same|match) = \frac{p(match|same)p(same)}{p(match|same)p(same) + p(match|!\,same)p(!same)}$$

$$= \frac{1\left(\frac{1}{N}\right)}{1\left(\frac{1}{N}\right) + (\prod_{i=1}^{M} \mu_i)\left(1 - \frac{1}{N}\right)}$$

$$= \frac{1\left(\frac{1}{10^4}\right)}{1\left(\frac{1}{10^4}\right) + (1.54 \cdot 10^{-14})\left(1 - \frac{1}{10^4}\right)}$$

$$= 0.99999 \cong 1$$

If the probability of one person matching is $m_i = 1.54 \times 10^{-14}$, the likelihood of a identifying a match in 10000 people is an example of the birthday problem:

$$p(n) = 1 - !p(n), approximated\, by\; 1 - e^{\left(\frac{-(n(n-1))}{2(1.54x10^{14})}\right)}$$

$$p(10000) = 0.000000325;$$

$$14{,}599{,}883\; people\, required\, for\; 50\%\; chance\; of\; a\; match$$

This answer should be close to the answer in an African population, as long as neutral assumptions hold because the localized mutation rate should not be different among populations.

Chapter III: Anonymization of data for transmission and disease surveillance

A Context-Sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection

The work in this section was published as a research manuscript in the Journal of the American Medical Informatics Association with Dr. Shaun Grannis and Dr. Marc Overhage of the Regenstrief Institute in Indianapolis, IN, and Dr. Kenneth Mandl from the Children's Hospital Informatics Program and the Harvard-MIT Division of Health Sciences and Technology.

Abstract

Cases were emergency department (ED) visits for respiratory illness. Baseline ED visit data were injected with artificially-created clusters ranging in magnitude, shape, and location. The geocoded locations were then transformed using a de-identification algorithm that accounts for the local underlying population density. 12,600 separate weeks of case data with artificially created clusters were combined with control data and the impact on detection of spatial clustering identified by a spatial scan statistic was measured. The anonymization algorithm produced an expected skew of cases which resulted in high values of data set k-anonymity. De-identification that moves points an average distance of 0.25km lowers the spatial cluster detection sensitivity by less than 4%, and lowers the detection specificity less than 1%. A population-density based Gaussian spatial blurring markedly decreases the ability to identify individuals in a data set while only slightly

decreasing the performance of a commonly used outbreak detection tool. These findings suggest new approaches to anonymizing data spatial epidemiology and surveillance.

Introduction

The use of spatial clustering algorithms in epidemiology and disease surveillance presents privacy challenges for researchers and public health agencies because of the identifiability risk associated with transmission of patient address information. The emerging science of spatial outbreak detection [110-113] is based on the recognition of unexpected clustering among cases. There is an inherent tension between the requirement for precise patient locations to accurately detect an outbreak and the need to protect patient privacy. Case locations that are identified using home address or a portion of that address, such as the zip code or census tract, increase the risk of breaching patient confidentiality. While identifiable data can be shared for public health activities, the barriers to and inherent risks of such exchange could be minimized if privacy preservation were optimized with respect for the intended use of the information.

We describe a novel method for anonymizing individuals in public health data sets, by transposing their spatial locations through a process informed by the underlying population density. Further, we measure the impact of the skew on detection of spatial clustering as measured by a spatial scanning statistic.

Background

Patient re-identification from purportedly de-identified data can be accomplished with surprising ease. For example, 87% of individuals in a publicly available database were re-identified using zip code, date of birth and gender alone [81]. There are well-described techniques for protecting the anonymity of individuals whose information resides in databases. Using these techniques, de-identification systems have been developed that remove personal data from database fields (for example, converting a date of birth to a year) [82] or from textual notes [83].

A metric for the ability to re-identify a patient in a data set is *k-anonymity*, where *k* refers to the number of people among whom a specific de-identified case cannot be reversely identified [82]. Location information, whether stored as classic plain text address data or as geocoded longitude and latitude values, can potentially identify an individual. A common approach to de-identifying such data has been to use census tract or zip code rather than home address to protect anonymity. There are two main drawbacks to using location data that have been transformed to a count of points within an administrative region. First, the loss of precise location may reduce sensitivity to detect clustering. Second, the ability to detect clustering may be diminished when some of the points cross administrative boundaries.

Previous investigators have attempted to mask geographic data by spatially skewing cases using, among others, affine and randomizing transformations [89, 90]. We describe a spatial anonymization algorithm based on skewing precise geocoded case locations using knowledge of local population characteristics. Skewing these patient addresses directly decreases the ability to re-identify, and thus increases the *k*-anonymity, of a case in a data set, as it will be much more difficult to determine what the actual patient's identity is once it has been altered. Masking the identity of an individual in a densely populated urban area, for example, does not require as great a skew as one in a sparsely populated rural setting. Next, we measure the effect of anonymization intensity on outbreak detection, focusing on the sensitivity of spatial cluster detection. The goal is to provide individuals, institutions and public health authorities a comfort level with the sharing of skewed, and hence, anonymized data, rather than using raw, fully identifiable data. Further we aim to provide transparent information about the resulting diminution of spatial clustering detection.

Methods

A. Overview

Cases were emergency department (ED) visits for respiratory illness from an urban, academic, pediatric, tertiary care hospital over a five week period from 12/30/2001 to 02/02/2002. Institutional Review Board approval at Children's Hospital Boston was granted. Home addresses of patients were cleaned to correct data entry errors using software (ZP4, Semaphore Corp., Aptos, CA) and then converted to geographic coordinates using geocoding software (ArcGIS 8.1, Environmental Systems Research Institute, Inc., Redlands, CA). ED visit data were injected with artificially-created clusters that varied in magnitude, shape, and location [114]. The geocoded locations of all points (real addresses and artificial cluster-points) were then transposed using a de-identification algorithm that skews the location based on the underlying population density. The impact on detection of spatial clustering as identified by a spatial scan statistic [115] was measured.

B. Population-Density Based Gaussian Spatial Skew

We blurred the spatial location of patient home addresses by a distance informed by the underlying population density near the home of each patient. The patient's home address, represented by latitude/longitude coordinates, was skewed using a random offset based on a Gaussian distribution whose standard deviations are inversely correlated to the local area's population density. The use of local demographic data enables our anonymization system to transpose patients in densely populated areas by a smaller distance than patients who live in more rural areas. Hence addresses can be skewed minimally while maintaining a specified *k*-anonymity.

C. Census Block Groups

Producing de-identified data sets based on local population densities requires state-wide, location-specific population density data, which are readily available from the US Census Bureau. Our de-identification system identifies each patient's census block group for which the total population per square kilometer by age is available [116]. Due to variability in the available Census 2000 block group data set, data were pre-processed to constrain maximum and minimum population density values and correct missing or improperly formatted values.

D. Gaussian Randomization

Optimally, individual points will be skewed by a minimal distance to obscure identity, while preserving spatial information. Transforming a data set using a Gaussian probability distribution function results in most cases being moved only a small distance, because the Gaussian probability distribution function is strongly weighted about its mean (center) value. We have developed a bivariate Gaussian anonymization scheme that uses two randomly selected values, σ_x and σ_y, the standard deviations of normal distributions, which are used to generate the distances for patient displacement in each dimension, randomly selected from the Gaussian distributions described above, [117]. When cases are moved d_x, d_y, they may be moved outside the boundaries of their original census block groups. Selecting distinct standard deviations in each direction helps protect against steep population gradients that are purely in either dimension, however this is generally not necessary. This Gaussian randomization is used in concert with population-density and age-based multipliers in the anonymization algorithm described in the following section.

E. Anonymization Algorithm

To achieve a similar *k*-anonymity between high- and low-density population areas, the amount a specific patient in a spatial data set is skewed should be inversely related to the local population density – patients in rural areas need to be moved a greater distance than those in cities. Additionally, age-based adjustments were integrated to compensate for spatial age-group population density variations, as regions may have markedly different age distribution patterns. To do this, we create multipliers reflecting the relative magnitude needed to move a specific point from its original location.

First, we calculate the average population density for all US Census Blocks in the region of interest, both for Census Block Group age ranges and for the total population density. Next, we calculate multipliers for each case that vary with the inverse of the population density in the census block group, below:

Anonymization Multipliers and Factors:

$$Age\_based\ pop.density\ multiplier = \frac{average\ age\_group\ population\ density}{patient's\ block\ group\ age\ density}$$

$$Total\ pop.density\ multiplier = \frac{average\ total\ population\ density}{patient's\ block\ group\ population\ density}$$

These multipliers allow the anonymization system to move patients with large population multipliers farther than those with smaller multipliers on average, in a data set.

Age Population Density vs. Total Population Density

$$\begin{aligned} &Combined\ multiplier \\ &\qquad = (\boldsymbol{Age\ parameter}) \cdot (Age\_based\ population\ density\ multiplier) \\ &\qquad + (\boldsymbol{1 - Age\ parameter}) \cdot (Total\ population\ density\ multiplier) \end{aligned}$$

Additionally, users may wish to control the relative importance of the age-based population density multiplier in comparison with the total population density multiplier. The age parameter allows this, and ranges from 0 to 1 where a value of 1 considers only age-based population density and 0 considers only the total population density when choosing the anonymization magnitude.

The desire to directly control the skew level of this skew algorithm can be achieved using a parameter or multiplier, which we now describe. This overall anonymization parameter is not in terms of actual anonymity afforded (which will be discussed later in this article), which should vary quadratically with this parameter.

Overall Anonymization Parameter:

$$\begin{aligned} &Overall\ multiplier \\ &\qquad = [\boldsymbol{c}] \cdot [(Age\ parameter) \\ &\qquad \cdot (Age\_based\ population\ density\ multiplier) \\ &\qquad + \ (1 - Age\ parameter) \cdot (Total\ population\ density\ multiplier)] \end{aligned}$$

The additional parameter c is a scaling factor that easily adjusts the magnitude of the overall skew applied to a specific latitude-longitude pair. The overall degree of anonymization is altered by changing this value, although it should be noted that the relationship between the degree of anonymization and the anonymization multiplier is non-linear.

F. Test data sets

To determine whether spatial detection performance is adversely affected by transformation of a data set using this anonymization algorithm, we created a set of test data sets that varied with several parameters. Five separate weeks of ED visit data were categorized into syndrome using chief complaint and ICD-9-CM diagnosis codes as previously described, [118] to identify visits for respiratory illness. Each week of this respiratory visit data set was injected with 252 artificially-generated clusters [119, 120] to create 1,260 data sets with one week of encounter data and one artificial cluster per data set. The 252 clusters contain 10, 25, or 40 extra points placed randomly within circles with a radius of 250, 500, 1000, or 3000m. These data sets were located 8.05, 24.14, or 80.47 km (5, 15, 50 mi) away from a center point (the hospital location) at 7 evenly-spaced angles. Each of the 1,260 data sets was then processed using the anonymization algorithm at ten different anonymization skew levels (magnitudes of anonymization), creating a total of 12,600 test data sets (Figure 15). Non-injected patient data are assumed to have no existing clustering, however this is a conservative assumption. If this assumption is false, it will likely lower the number of false positives that are identified.

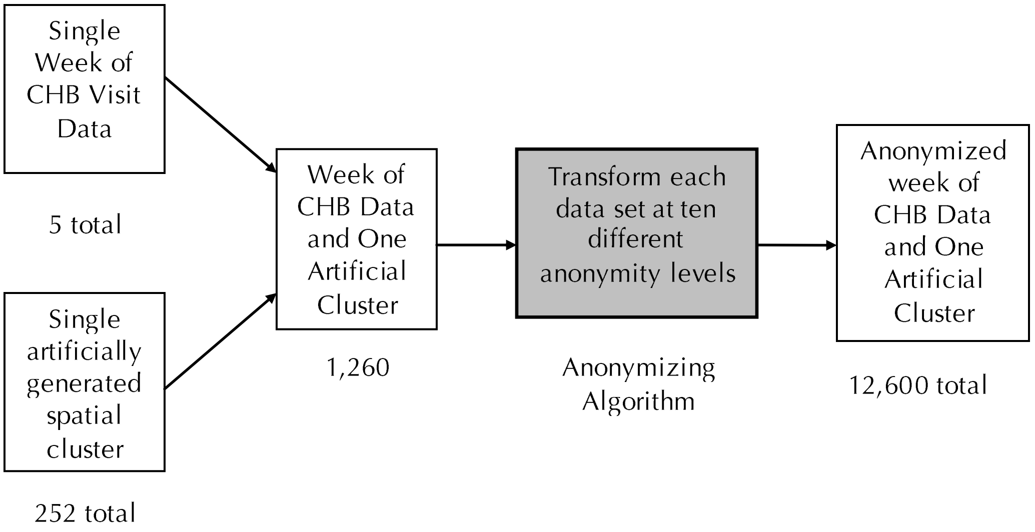

Single Week of CHB Visit Data
5 total
Single artificially generated spatial cluster
252 total
Week of CHB Data and One Artificial Cluster
1,260
Transform each data set at ten different anonymity levels
Anonymizing Algorithm
Anonymized week of CHB Data and One Artificial Cluster
12,600 total

Figure 15: Experiment description: five weeks of Children's Hospital Boston visit data are each individually combined with 252 different artificially-generated spatial clusters. Each of the resulting 1,260 data sets was then anonymized at ten different levels for a total of 12,600 experimental data sets.

G. Measuring Clustering Detection Performance

The method used to measure clustering was the SaTScan Spatial Bernoulli Model scanning algorithm using 999 Monte-Carlo replications [112, 115]. After the test data sets were created, each was analyzed using the aforementioned SaTScan purely spatial scanning statistic to find the p-value of the most likely cluster. Because these data sets each contained an artificially-generated cluster of patients, we used SaTScan to determine whether at least 50% of the artificially-injected cluster-points were identified with a p-value ≤ 0.05. This cutoff was selected because we wanted to only consider our cluster detection effort a success if at least a majority of the artificially generated case points were identified, as there would be limited utility in an anonymization algorithm if it could only allow cluster detection algorithms to identify one case in a cluster in practice. To clarify, this cutoff threshold makes our ability to detect these clusters harder in the analysis that follows, because we specify that a majority of the cases must be identified, even if clusters with very low p-values are detected with 40% of the artificially injected cluster points, for example. If the cluster was identified, we also recorded what proportion of the total identified cluster points were from the artificial cluster.

H. Estimate of k-Anonymity

It is possible to estimate the expected level of *k*-anonymity for an individual skewed case by multiplying the local population density [(population)(area^{-1})] by a circular ring area approximation of the Gaussian probability distribution function (Figure 16.) Because 68.26% of patients should fall, on average, within the first standard deviation, σ miles in radius from where they were originally located, we can multiply the local population density by the area, $\pi\sigma^2$ and by the probability that the patient would have been moved into that region, 0.6826. We can add to this the next ring's population density multiplied by its area and its probability that a patient would be transplanted into that area, 0.2718. Finally we can add the area of the last ring multiplied by its local population density by its probability density, 0.0428. The sum of these three numbers provides a computationally-tractable expectation of *k*-anonymity achieved for a specific case in a data set.

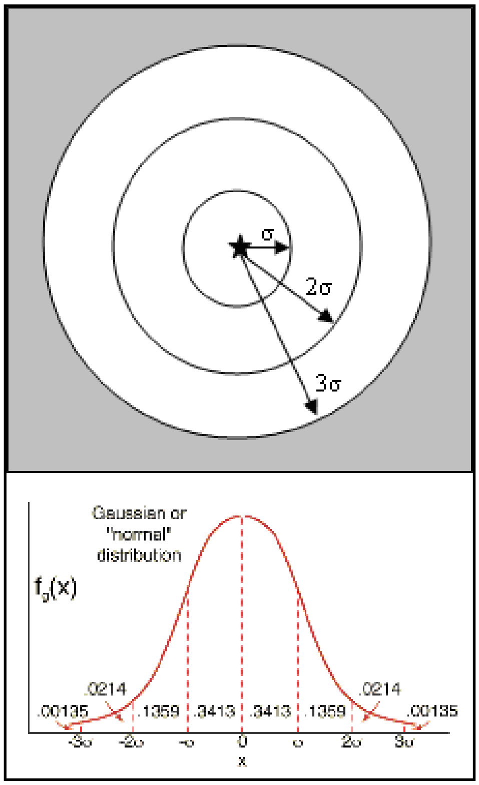


Figure 16: Estimating expected *k*-anonymity: using the data set standard deviation of the distance each patient is moved in the anonymization, •, an estimate of achieved *k*-anonymity is calculated, assuming no other external knowledge of specific patient information. The local population density [people/km^2] is multiplied by each area [km^2] and then multiplied by the probability that the patient would have been in that area, from the Gaussian probability distribution function.

The circular areas used in these calculations may contain several census block groups, so estimate accuracy can be increased by multiplying the fraction of area comprised by each census block by the population density of that block. The sum of those partitioned values can then be multiplied by the above probability distribution values. This estimate of *k*-anonymity relies on the probability density distribution of the 2D Gaussian. Sufficient numbers of patients are needed to statistically ensure that the central limit theorem has been satisfied, a reasonable assumption given the size of most public health surveillance data sets.

I. Outlier Assessment and Percentage of Points Meeting Anonymity Thresholds

To determine whether a subset of patients (those potentially in rural areas) might not have attained anonymity at the level specified by the user, the skew distance cumulative distribution functions for different user-specified *k*-anonymity values can be inspected to easily determine the quantity of cases in a large data set that have not been sufficiently individually de-identified. In aggregate form, most of these data are still sufficiently anonymized from a user with no external information; however some rural cases may still pose risk of information disclosure. An outlier analysis allows users to determine which cases in a specific data set should be re-anonymized or excluded and what fraction of cases have been successfully anonymized.

J. Client Tool and GUI for Remote De-Identification of Data

The source code and binary installation toolkits have been made available in an open source repository at http://sourceforge.net/projects/patientanon/ . This stand-alone toolkit implements the de-identification algorithm explored in this thesis. Data sets are accepted in either a CSV or XML format, and the anonymization toolkit allows the user to specify the order of the required variables to suit almost any previously created data set. Special care was taken to make this anonymization system deployable as a standalone application by extracting all of the necessary census block group data and storing it in a local database. In the standalone version, this information is stored as a set of local xml files to remove complexity from the setup of the program, so that no database software or connections are necessary to anonymize patient data. For better performance, we allow users to load their choice of state census block group data into memory. Hash tables are also used to improve lookup speed for identifying a subset of candidate census block groups for each patient record.

Results

A. Distribution of Location Skew

The distance from the original address to the transformed address for each patient was calculated (Figure 17) for four sample anonymized data sets with different skew magnitudes. This illustrates empirical anonymization distributions with respect to skew level. The normal probability distribution function has the greatest density centered about the mean value, where the mean value represents no positive or negative linear skew. Nearly all cases were moved at least some distance due to the bivariate nature of this Gaussian blurring algorithm. As expected, only a small portion of patients were moved a large distance from their original addresses.

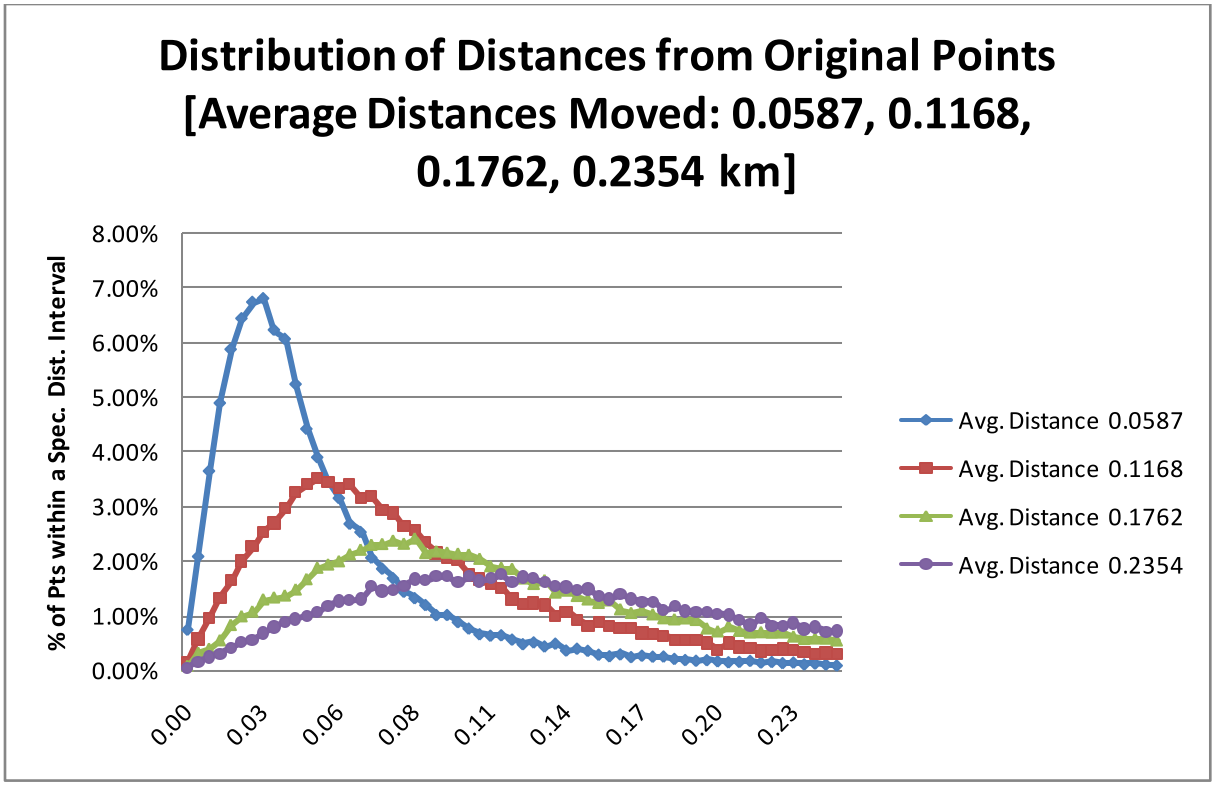


Figure 17: Distribution of distance from original location: Each case was moved from an original home address to a new de-identified location. Each data series represents the percentage of patients that were displaced plotted against distance [km] displaced from original location.

B. Average Distance Moved vs. Estimate of k-anonymity

Using the population-density estimate of *k*-anonymity described above, the average *k*-anonymity for each anonymized data set was calculated (Figure 18). As the magnitude of anonymization increases (as the average distance from original points in the data set increases), the *k*-anonymity increases quadratically. The method to estimate *k*-anonymity in these data sets uses the area around each patient circumscribed by a radius that is the standard deviation of distance from original address in each data set. These areas may contain multiple census block groups, each with a different population density, so we chose to use a conservative estimate, using the smallest population density in the relevant area. As these standard deviations increase linearly (as the magnitude of Gaussian blurring that is applied to each data set increases), the area enclosed by the radius around the patients increases as a second order polynomial. An average distance value of 0.25 km corresponds to an average *k*-anonymity value of 250, such that in this sample data set, a patient is not reversely identifiable among a group of 250 people.

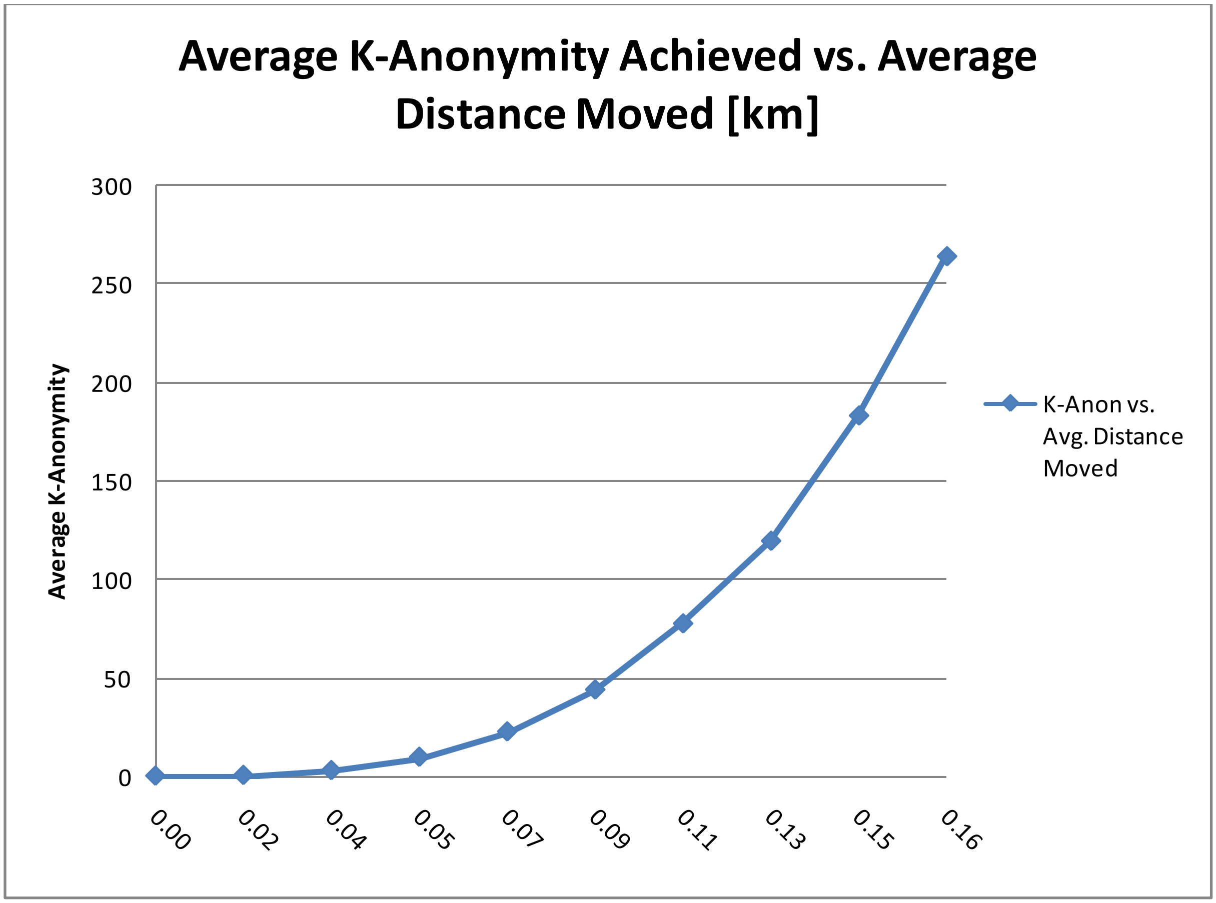


Figure 18: Average *k*-anonymity achieved vs. Average Distance Moved: As the average distance [km] moved in a given data set increases, the anonymity achieved also increases in a quadratic fashion.

C. Sensitivity of Spatial Clustering Detection

The SaTScan purely spatial Bernoulli model was used to identify whether at least 50% of artificially-injected test clusters points were identified in 12,600 spatial data sets in a cluster with a p-value of less than or equal to 0.05. This cutoff is a conservative one that makes it more difficult to call a cluster that is identified with some artificially-injected cluster points a 'success', and requires that at least the majority of the artificially-injected points must be detected. The following sensitivity and specificity analysis is based on cluster points that are part of clusters that are either successfully 'detected' (true positives) or 'not detected' (false negatives).

As the magnitude of the spatial skew increased (as the average distance from original point increased), the rate of spatial detection performance decreased (Figure 19). The average sensitivity and average specificity are graphed for each skew magnitude. The sensitivity and specificity values are defined for each cluster with artificially-injected cases counted as true positives and non-injected patients counted as false positives. De-identification with a data set average distance to original point of 0.25km lowers the spatial cluster detection sensitivity less than 4% and lowers detection specificity less than 1%. This result demonstrates that this approach has a minimal negative effect on spatial clustering detection sensitivity and specificity.

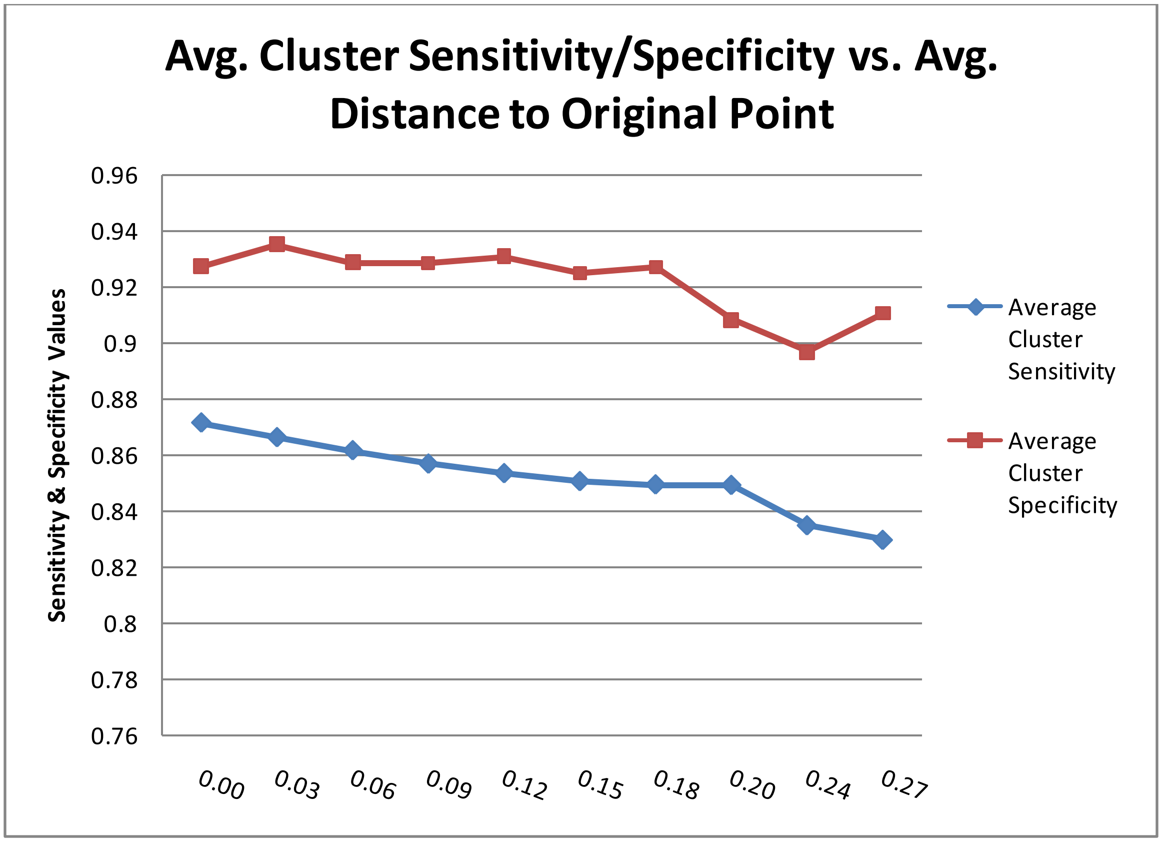


Figure 19: Average Cluster Sensitivity/Specificity vs. Average Distance to Original Point [Average Distance Increases as Anonymization Level Increases]: The average sensitivity and specificity of spatial detection (using SaTScan Bernoulli Spatial Model with p-value • 0.05) of artificially-injected clusters of patients is displayed with respect to the average distance that patients in a de-identified data set are moved with respect to their original home addresses. Sensitivity and specificity are calculated using cases from the cluster and control data that were or were not identified properly.

D. Outlier Assessment and Percentage of Points Not Meeting Anonymity Thresholds

We describe the *k*-anonymity of results in our anonymization experiments using the average *k*-anonymity achieved in aggregate transformed data sets. To determine whether a subset of patients (those potentially in rural areas) might not have attained adequate anonymity, the cumulative distribution functions for user-specific *k*-anonymity values are presented with respect to average distance from original address (Figure 20). As the average data set distance from original point increases, the percentage of points that do not achieve a given *k*-anonymity value decreases. In this example, it is possible to calculate that a *k*-anonymity value of 20 has been reached in 99% of all patients in this sample data set when the average distance to original point is 0.25km. This is because the estimate of *k*-anonymity is separately generated for each point based on the distance it would be expected to have been moved in an anonymizing skew, so a small fraction of points (in this case < 1%) would have been moved a distance so small that they would not achieve a *k*-anonymity level of at least 20. It still would not be possible to determine which points in an anonymized dataset were these points, however. Points that do not meet a user-specified threshold can either be removed from a data set, or they can be re-anonymized. It is important to note, however, that re-anonymization of a subset of points will alter the characteristic output described above.

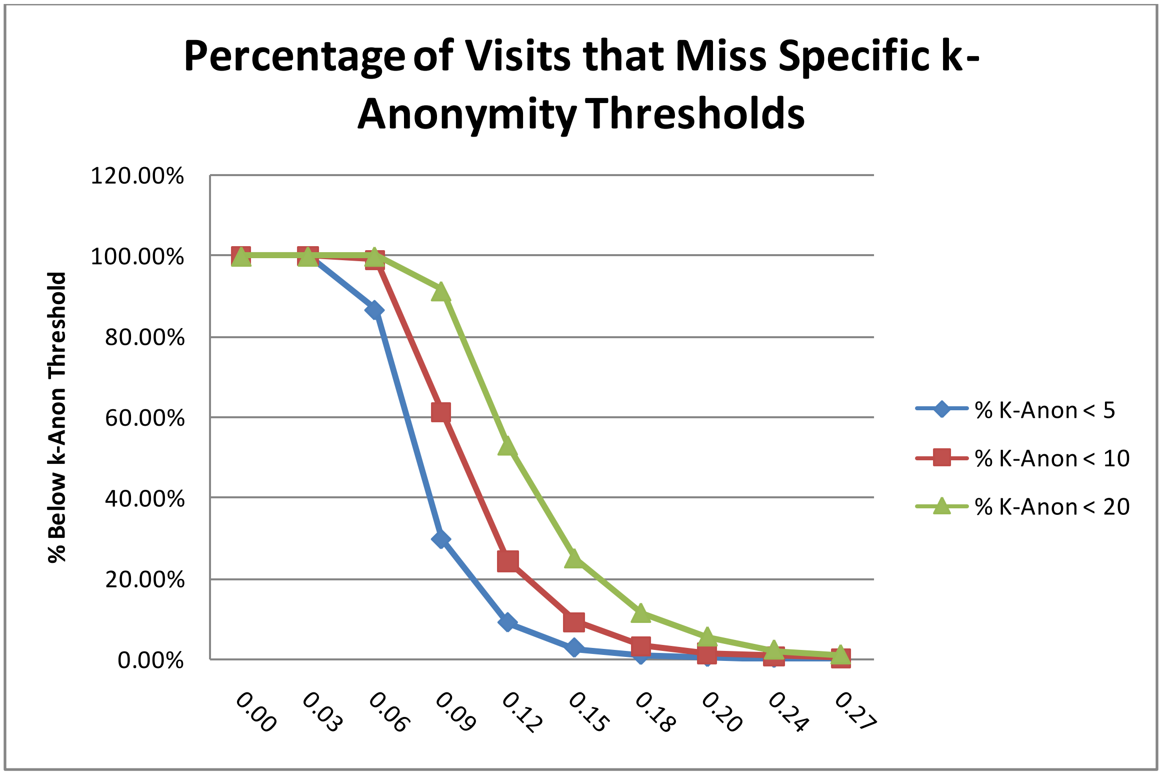


Figure 20: Percentage of visits that meet specific k-anonymity thresholds: For different user-specified k-anonymity minimum thresholds, the percentage of visits in a data set with a k-anonymity value below the minimum threshold (and not sufficiently de-identified) decreases quickly as the average distance moved increases. For over 99% of the visits in all test data sets, a minimum k-anonymity value of 20 could be achieved with an average distance moved of 0.25km.

Discussion

Population-based Gaussian skew represents a novel anonymization method that can provide a user-defined level of *k*-anonymity. Further, this method can readily anonymize public health surveillance data sets containing identifiable, protected health information with minimal impact on the performance of an outbreak detection system. We have explored the use of population density and age-based population density data for de-identification in this manuscript, but we do believe the principles explored in this paper are generally applicable to other types of patient and demographic data.

We propose a public health use case for this anonymization system. The data exchanged, for example, between a hospital and a public health authority for use in a syndromic surveillance system can contain skewed locations. As the anonymization system is completely abstracted from the spatial detection systems that utilize it, there is no need to align the use of this algorithm with a specific toolkit for cluster detection. If clustering is detected and an outbreak investigation is required, the fully identified data could be subsequently exchanged according to the HIPAA regulations as applied to public health.

One approach that might be considered is a web-services paradigm, where a client wishing to anonymize spatial data might send a data set containing only spatial data and possible de-identification requirements, such as minimum *k*-anonymity or average *k*-anonymity, to a de-identification server. The client could then reunite a returned data set with other data that had been stored about those patients without having transferred linked spatial data over the internet.

Moving forward, it will be necessary to determine what degree of skew will provide sufficient anonymity for distribution of a patient data set to permit different levels of data exchange. Determining what level of anonymity is required for HIPAA compliance using an anonymization system is a challenging and complex issue. A policy could be envisioned under which patients volunteering their information for use by public health agencies might be able to specify the desired *k*- anonymity.

The skew method described here readily achieves far higher degrees of k anonymity than are generally considered acceptable for public health data sets. It is important to be aware, however, that *k*-anonymity can vary from case to case within a data set. Consider the example of a data set containing one case which is located in a rural town of 50 residents. Consider further that the desired *k*-anonymity is 100. It is difficult to achieve this de-identification level without increasing the magnitude of anonymization for all cases in the data set to a high level. Hence, a tradeoff arises between keeping the difficult-to-anonymize cases (maintaining the integrity of the data set) versus discarding them as outliers, and thereby enabling lower intensity anonymization for the other cases. Cases may need to be removed from data sets to ensure that *k*-anonymity thresholds are met for every patient in a specific data set. This suggests that better results would likely be achieved by allowing the clustering algorithm and the anonymization algorithm to interact in some way.

This algorithm randomizes the magnitude of the address skew for each patient using randomly selected seed parameters that inversely vary with the underlying population density values. Those seed values are then used to select a random x and random y offset based on a Gaussian probability distribution. Knowledge or disclosure of all of the randomly selected seed and offset values could aid a nefarious agent in reversely identifying patients by lowering the data set anonymity achieved, however the seed and offset values are calculated separately and are not stored anywhere in this de-identification process.

The main limitation of this study is that measurements were made in only one geographic area and only one approach to detecting spatial clustering was investigated. However, the urban setting is a common one for intensive public health surveillance (such as syndromic surveillance) and SatScan is a widely employed method. Additionally, we have explored the use of population density and age-based population density data for de-identification in this manuscript, but we do believe the principles explored in this paper are generally applicable to other types of patient and demographic data. De-identification that attempts to accurately estimate *k*-anonymity is a function of all of the fields contained in a data set – for anonymity to be achieved, it must be adequately achieved across all combinations of attributes of a data set. For public health surveillance data sets, this objective is tenable as the number and types of data fields contained in these data sets are limited.

Conclusion

We present experimental results demonstrating that a population-density based Gaussian spatial blurring markedly decreases the ability to identify individuals in a data set while only slightly decreasing the performance of a standard outbreak detection tool--SaTScan. These findings suggest new approaches to anonymizing data for the real-world application of spatial epidemiology in public health practice.

Optimal discrete anonymization using linear programming techniques

This section represents joint work with Dr. Shannon Wieland and Dr. Bonnie Berger from MIT, and Dr. Kenneth Mandl, and has been described in a PNAS manuscript, currently in review.

Abstract

Data sets describing the health status of individuals are important for medical research, but must be used cautiously to protect patient privacy. For patient data containing geographical identifiers, the conventional solution is to aggregate the data by large areas. This method often preserves privacy but suffers from substantial information loss, which degrades the quality of subsequent disease mapping or cluster detection studies. Other heuristic methods for de-identifying spatial patient information do not quantify the risk to individual privacy. We develop an optimal method based on linear programming to add noise to individual locations that preserves the distribution of a disease. The method ensures a small, quantitative risk of individual re-identification. Because the amount of noise added is minimal for the desired degree of privacy protection, the de-identified set is ideal for spatial epidemiological studies. We apply the method to patients in New York County, New York, showing that privacy is guaranteed while moving patients 25 to 150 times less than aggregation by zip code.

Background

Since the publication of the first disease dot map more than 200 years ago revealed the locations of yellow fever patients in New York City [121], a collection of methods to analyze health characteristics and location have coalesced to comprise the field of spatial epidemiology. Disease mapping; assessing the tendency of cases to cluster in space; detecting localized clusters of diseases; and testing for clustering around a putative environmental point source are all distinct activities within the field. Although spatial analyses of geographical identifiers such as zip codes, street addresses, and locations on maps may ultimately improve medical care and public health, the identifiers themselves are protected health information that pose a threat to patient privacy if disclosed. Even common identifiers can be linked to individuals; eighty-seven percent of subjects in one study could be uniquely identified by their gender, zip code and date of birth [82] and low-resolution dot maps of diseases published in several medical journals could be used to trace most cases to single addresses [122].

Although established since the time of Hippocrates [123], the professional responsibility to protect patient privacy has been newly formalized with the passage of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [124]. Effective since 2003, HIPAA details specific information disclosures that violate privacy. Noncompliance may result in fines of up to $250,000, and imprisonment for up to ten years. The rule defines a category of “non-identifiable data sets,” whose dissemination is not restricted; this is desirable from a research perspective because it allows analysis by the entire scientific community, and makes independent verification of results possible. Either of two criteria must be met for a data set to qualify as non-identifiable. The first specifies that the data set must not include any of 18 specific identifiers, including five-digit zip codes. The first three digits of a zip code may be included, provided that at least 20,000 people share the same first three digits. The second criterion specifies that a qualified individual determines "that there is a very small risk that the information could be used by others to identify a subject of the information" [124].

The prevailing method for preserving privacy in spatial data is aggregation by pre-defined administrative regions, such as counties or census enumeration districts. These areas must be larger than the zip code level to comply with HIPAA. However, aggregation may compromise subsequent research by erasing useful spatial information [125]; for example, the detection of spatial clusters is significantly less sensitive and specific when data are aggregated even by zip code [126]. Furthermore, the level of privacy protection depends on the number of patient records. For example, if it is revealed that 20 patients having a certain disease reside in a region containing 20,000 people, then there is a 1/1000 chance that a randomly selected individual from the region is one of the patients. However, if 200 patients with the disease live in the region, then the probability that a random individual from the region is among the set of patients increases to 1/100.

An alternative to aggregation is moving each patient to a new location to ensure privacy [127], formalized by the family of "geographical masks" proposed by Armstrong et al. [128]. Each is a deterministic or stochastic function of geographical identifiers designed to de-identify patient locations, while preserving the approximate spatial distribution of cases. They encompass previous approaches such as aggregation and translation by fixed distances, as well as affine transformations, adding independent noise, and random perturbations adjusted for population density [129]. Although these techniques represent a significant advance over aggregation, they apply the same transformation independent of the local geography, the number of patient records, and, in several cases, the underlying population counts. Consequently, the probability that any of the de-identified records originated from a single individual depends upon all of these variables. For example, consider a geographical mask that moves each record to a new location with uniform probability inside a circle of radius *r* centered at the record. Given a masked case location, it is obvious that its original location must lie within the circle of radius *r* centered at the masked location. If part of this circular region intersects a body of water or other uninhabited region, then the area from which the case originated is narrowed, conceivably to a tiny fraction of the map. In the general case, quantifying the re-identification probability may be extremely difficult. However, a quantitative measure of privacy protection is essential to ensure that the standard of “very small risk” specified by HIPAA is met.

In a different application, Machanavajjhala et al. [130] ensured a low disclosure risk by generating a de-identified data set from a model of the original data. This approach is sensitive to the user’s belief about the data, reflected in the choice of model. Furthermore, in order to preserve essential data features needed for subsequent analysis, these features must be captured by the model. If the data are sparse, or if the essential features are unknown in advance, this may not be possible.

We present a principled approach to de-identifying patient locations based on linear programming that allows the user to specify the maximum probability of associating any of the transformed locations with any individual in the population. The solution is optimal in that it guarantees that patients are moved the minimum distance for the level of privacy protection offered. The method has the advantage that it does not move patients to unrealistic locations, such as lakes and rivers. It may be used to create de-identified data sets that can be shared without restriction for spatial epidemiological investigations. Application of the method to de-identifying patients in several counties shows that a high level of privacy can be achieved while preserving clusters and moving patients relatively short distances.

LP De-identification

Given the locations of a set of patients, the aim is to randomly assign new, de-identified locations that can be associated with the original patients with very low risk. The distance between the original and new locations should be minimized. The original locations may be any discrete geographical identifiers. We assume that the data are purely spatial, containing no other identifying information such as age or sex. The set A of available original locations, which contains the actual locations in the data set, must be known in advance; for example, these could be all the census block groups in a state, or all the residential addresses within a city. The set B of possible final locations to which patients may be moved is also defined in advance. These may be different sets, such as evenly spaced points on a grid to which patients at exact addresses will be relocated. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are disjoint, then no case will be assigned to the original location of any other case.

This problem can be captured by a linear programming (LP) model, a simple type of mathematical model that consists of a set of decision variables, constraint equations, and an objective function [131]. The decision variables are the transition probabilities P_{ij} of assigning a patient in location $i \in \boldsymbol{A}$ to a new location $j \in \boldsymbol{B}$ (see Figure 21). Once values have been assigned to the decision variables, each of s patients in a list of original locations is moved to a new location independently of the other patients. If a patient is originally in location $i \in \boldsymbol{A}$, a new location j is drawn from the set $\boldsymbol{B}$ using a multinomial distribution with probabilities P_{ij}. The goal is thus to assign a value to each decision variable P_{ij} so that this procedure ensures privacy and minimizes patient movement. Constraint equations specify conditions that must be satisfied by the decision variables P_{ij}. Because the decision variables are probabilities, each must be nonnegative:

$$0 \leq P_{ij} \ for \ all \ i \ \epsilon \ \boldsymbol{A} \ and \ j \ \epsilon \ \boldsymbol{B} \qquad [1]$$

In addition, every case must be moved somewhere, so

$$\sum_j P_{ij} = 1 \ for \ all \ i \ \epsilon \ \boldsymbol{A} \qquad [2]$$

A final constraint guarantees that the risk of linking any randomized location with any original patient is small. In formal terms, we specify that the probability that any location from the randomized data set originated from any specific individual in the underlying population is at most ξ :

$$P_{ij} \cdot \frac{n_i}{N} \leq \frac{n_i \cdot \xi}{s} \cdot \sum_{k \epsilon A} \frac{n_k}{N} \cdot P_{kj} \quad for\ all\ i\ \epsilon\ \boldsymbol{A}\ and\ j\ \epsilon\ \boldsymbol{B} \qquad [3]$$

In this equation, ξ is a user-specified privacy bound between zero and one, $s \geq 1$ is the number of patients in the data set to be de-identified, n_i is the number of people in region *i*, and $N = \sum_{r \epsilon A} n_r$. For example, if the regions are census block groups, then the constants $\{n_i\}_{i \in A}$ may be corresponding populations drawn from the same census. If the regions are exact addresses, then n_i is assumed to be 1 for each *i*. Any randomly or methodically chosen member of the population is guaranteed to belong to the data set with probability at most ξ . Consequently, given the de-identified list, one could expect to search through at least $1/\xi$ members of the population by any method before encountering one person on the list. Derivation of this constraint is found below at the end of this section.

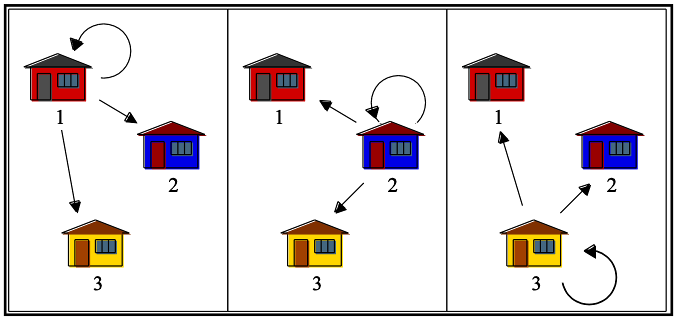


Figure 21: Schematic of transition probabilities. A patient found at each location in a set A may transition to any location in a set B. In this example, the sets A and B are equivalent for simplicity, each consisting of three locations represented by houses. The nine transition probabilities, represented by arrows, are variables solved by linear programming.

We wish to move patients as little as possible subject to the constraints above. For each $i \in \boldsymbol{A}$ and $j \in \boldsymbol{B}$, we define d_{ij} to be the distance between region *i* and region *j*. Assuming that each individual in the study area is equally likely to be in the data set, a patient originates in region i with probability n_i/N. Hence the expected distance that a patient is moved, which is the objective function to be minimized, is

$$\frac{\sum_{i\epsilon A} \sum_{j\epsilon B} d_{ij} \cdot n_i \cdot P_{ij}}{N}. \qquad [4]$$

Several standard linear programming techniques to solve LP models, such as that specified by equations 1-4, have been developed. When applied to an LP model, they either locate an optimal solution that minimizes the objective function, or they prove that no solution exists. The latter happens if *no* probabilistic de-identification strategy has a risk of re-identification of at most ξ . For example, if there are *m* available individual addresses, then no strategy to de-identify $s \leq m$ patients by reassigning new addresses can achieve a risk of re-identification below s / m. If no strategy exists, then a larger re-identification risk can be specified (if acceptable for privacy protection), or the set of available locations can be expanded.

Simple variations of the linear program make it possible to capture other objective functions, constraint equations, or decision variable constraints. Instead of minimizing the expected distance, the expected squared distance may be used to penalize long distance moves more heavily than short moves. In fact, any objective function that is a linear combination of the decision variables P_{ij} may be used without complicating the analysis.

If a deterministic strategy is preferred to a randomized strategy, the LP model may be converted into a binary integer program. This specifies that only the values 0 or 1 may be assigned to the decision variables. For a fixed *j*, the set $\boldsymbol{I}_j = \{i : P_{ij} = 1\}$, if nonempty, has the property that $\sum_{i\epsilon I_j} n_i \geq \frac{s}{\xi}$. In other words, the patients are binned into a subset of the locations, the number and positions of the bins minimize the expected transition distance, and the total population assigned to each bin is at least s / ξ. In general, the optimal deterministic strategy moves patients farther than the optimal randomized strategy because the set of deterministic strategies is contained by the set of randomized strategies.

It is also simple to add additional linear constraints to the problem. For example, if $\boldsymbol{A} = \boldsymbol{B}$ it is possible to guarantee that no case is assigned to its original location by specifying in the LP model that $P_{ii} = 0$ for every $i \in \boldsymbol{A}$. Although this would not increase the level of privacy, it may assuage fears that original locations may be released. It is also possible to limit the number of outgoing transitions from any position to its *k* nearest neighbors, for a fixed *k*. In general, additional constraints increase the optimal value of the objective function.

Derivation of the constraint in Equation 3. The constraint in Equation 3 guarantees that the probability that any location from a de-identified data set originated from any specific individual in the underlying population is at most ξ.

Consider the probability of re-identifying a set of *s* cases that have been randomized to new locations. Given the set $\boldsymbol{A}$ of possible original locations and $\boldsymbol{B}$ of possible final locations, let P_{ij} denote the probability of transition from location $i \in \boldsymbol{A}$ to location $j \in \boldsymbol{B}$. Given the set of *s* locations comprising the de-identified data set, we require the probability that any one of these derived from one specific individual to be at most ξ. This is guaranteed if the probability that a location from the randomized data set originated from an arbitrary specific individual is required to be at most ξ / s. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the original and transformed locations, respectively. This condition is formally expressed as:

$$p(patient\ q|Y = j) \leq \frac{\xi}{s} \qquad [5]$$

for every individual *q* in the population and every location $j \in \boldsymbol{B}$. The left hand side of this inequality is equivalent to

$$p(patient\ q \cap X = L(q)|Y = j) \qquad [6]$$

where *L(q)* is the location of individual *q*, or

$$p(patient\ q|X = L(q)) \cdot p(X = L(q)|Y = j) \qquad [7]$$

by the definition of conditional probability. Assuming that all individuals in location *L(q)* have an equal chance of having the disease, we have

$$p(patient\ q|X = L(q)) = \frac{1}{n_{L(q)}} \qquad [8]$$

where $n_{L(q)}$ is the number of people in location *L(q)*. Hence the condition expressed by equation 5 is

$$p(X = L(q)|Y = j) \leq n_{L(q)} \cdot \frac{\xi}{s} \qquad [9]$$

for every individual *q* and location $j \in B$. Because the location of *q*, *L(q)*, may only take on values in A, this is equivalent to

$$p(X = i|Y = j) \leq n_i \cdot \frac{\xi}{s} \qquad [10]$$

for every $i \in A$ and $j \in B$. After multiplying both sides of equation 10 by *p(Y = j)*, the left hand side becomes *p(X = i ∩ Y = j)*, or *p(Y = j|X = i) · p(X = i)*. Furthermore, *p(Y = j|X = i)* is simply the transition probabilityfrom location *i* to location *j*, so it is equivalent to the decision variable P_{ij}. Hence equation 10 is equivalent to

$$P_{ij} \cdot p(X = i) \leq n_i \cdot \frac{\xi}{s} \cdot \sum_{k \in A} P_{kj} \cdot p(X = k) \qquad [11]$$

for all $i \in A$ and $j \in B$. Assuming that all individuals in the population have an equal prior probability of belonging to the original data set, we have

$$p(X = i) = \frac{n_i}{N} \qquad [12]$$

for all $i \in A$, where $N = \sum_{r \epsilon A} n_r$ is the total population. Hence, we obtain

$$P_{ij} \cdot \frac{n_i}{N} \leq \frac{n_i \cdot \xi}{s} \cdot \sum_{k \in A} \frac{n_k}{N} P_{kj} \;\; for\, all\, i \in A\, and\, j \in B. \qquad [13]$$

Equation 13 is incorporated into the LP model as a set of constraint equations. Thus the final set of transition probabilities P_{ij} satisfy this equation for all $i \in A$ and $j \in B$.

Following the proof backwards from equation 3, this means that the probability that a location from the de-identified data set originated from an arbitrary specific individual is less than or equal to ξ / s for every location. Because the probability of the union of events is bounded above by the sum of the probability of events, the probability that any specific individual is represented in the final data set is at most ξ.

Application

We determine an optimal strategy to randomize patients in New York County for a range of maximum re-identification risks. The strategy moves patients much shorter distances than aggregation by zip code or aggregation by the first three digits of zip code, and it preserves disease clusters in the data to a greater degree than either aggregation method. The method also compares favorably to aggregation for other counties having a range of population densities.

Stringent De-identification of Locations

We consider de-identifying case locations in New York County, NY grouped by census blocks. A census block is a small geographical unit typically containing approximately 1500 people [116]. According to the 2000 census, the 988 census blocks in New York County contain between 0 and 15112 people. We devise the optimal strategy to de-identify a set of $1 \leq s \leq 20000$ patients with a maximum probability of $s / 20000$. Transitions from any census block were restricted to its nearest 100 neighbors. The LP model was solved using CPLEX LP software [132], resulting in a 988 × 988 matrix of transition probabilities.

Under the optimal strategy, the expected distance between a patient's original and de-identified location is only 265 m. Three of the 988 matrix rows are illustrated in Figure 22. These show three possible configurations: patients are re-assigned to the same census block group or one of a few neighboring census block groups; patients are re-assigned to a single nearby census block group; and patients are moved to one of several possible census block groups which do not include the original location. Even from this limited subset, it is clear that the optimal strategy would be difficult to devise by hand. In particular, the optimal transition probabilities are not a monotonic or regular function of the distance between census block groups, such as a Gaussian function.

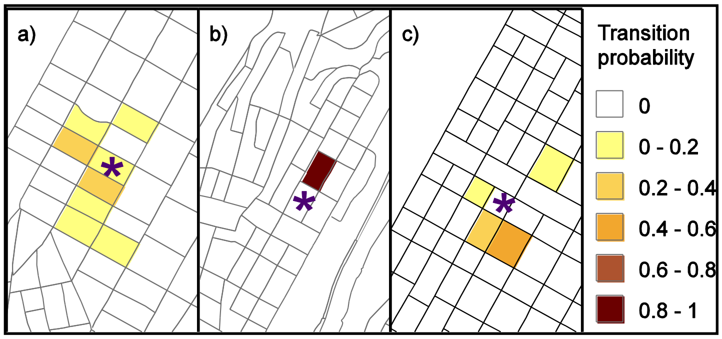


Figure 22: Transition probabilities for the optimal strategy to de-identify s ≤ 20, 000 patients from New York County, New York with a maximum re-identification probability of s / 20000. Transition probabilities from three of the 988 census blocks are shown, illustrating a few of the many possible transition distributions. The shading in region j represents the value of the probability Pij of transitions into the region. a) Patients in one census block (asterisk) may remain there, or they may transition to one of several nearby blocks. b) All patients originally in one census block (asterisk) are assigned to one neighboring block. c) Patients are re-assigned from one block (asterisk) to one of four nearby census blocks. No patients are re-assigned to the original census block (i.e. Pii = 0).

Comparison to Aggregation

To examine the relationship between the re-identification probability and the expected distance moved by a patient, we calculated the optimal de-identification strategies for a range of re-identification bounds. Because the total population summed over all census block groups is 1,696,038, the minimum achievable re-identification probability, corresponding to complete randomization, is s / 1696038, or s · 0.00000059. The expected transition distance is 6.4 kilometers (km). The least populated non-empty census block group contains only one individual, so the strategy of re-assigning patients to their original locations has a re-identification probability of 1 (which would be realized if one patient in a "de-identified" set originated from that census block group) and an expected transition distance of 0 km. The optimal strategies for de-identifying patients were calculated for a range of re-identification probabilities between these two extremes, and the expected distance moved by each patient is shown in Figure 23.

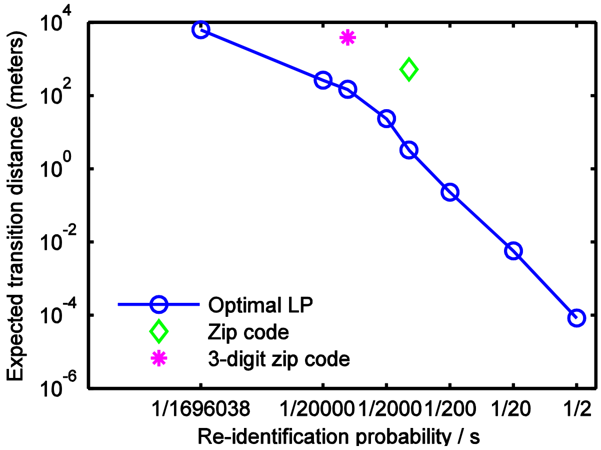


Figure 23: Relationship between the re-identification probability, the number s of patients, and the expected transition distance for the optimal LP strategy to de-identify patients by census block group in New York County, New York. As the level of privacy protection decreases (from left to right along the x-axis), patients are moved a smaller distance in expectation. Aggregation by zip code (green diamond) and first three zip code digits (magenta asterisk) are suboptimal strategies yielding larger distance movements than the optimal LP strategy at the same re-identification probability. Note that log scales are used, so the expected transition distance increases 100-fold between tick marks on the y-axis.

The optimal LP strategies move patients much less than aggregation when the level of privacy protection is held constant. Aggregation by zip code moves patients an expected 519 m. The least populated zip code contains 884 people (excluding empty zip codes and one zip code containing only one person), so there is a maximum re-identification probability of $s/884$ for a set of $s \leq 884$ patients under this strategy. The optimal LP strategy at the same re-identification probability moves patients by only 3.3 m. Aggregating by the first three digits of zip code moves patients an expected 3.9 km, and has a maximum re-identification probability of $s/8188$. At this probability of re-identification, the optimal LP strategy moves an average patient a much smaller distance of 149 m. Thus for the same level of privacy protection, aggregation moves patients 25 to 150 times farther than the optimal LP strategy (Figure 23).

Cluster Detection

To determine the degree to which LP de-identification preserves spatial clusters in data, we applied a standard cluster detection algorithm to simulated case-control data that had been de-identified using the LP method or aggregation. We constructed 1000 data sets, each consisting of 100 controls and 100 cases. Cases and controls were randomly placed in census block groups to reflect the underlying population density, with an excess number of cases within a randomly placed circular region of radius 1 km to simulate a disease cluster. Each set of cases and controls was de-identified using the LP method for a range of re-identification probabilities from 0.005 to 0.5. Each set was also de-identified using aggregation by zip code and by the first three zip code digits. SaTScan circular cluster detection software [112, 133] was applied to each de-identified set, and the p-value of the most significant cluster found was recorded (Figure 24).

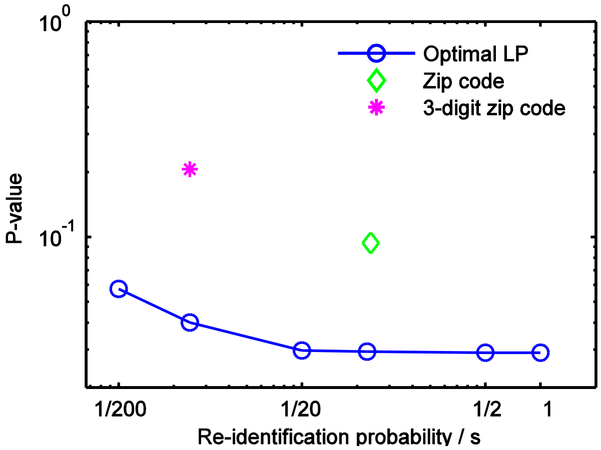


Figure 24: Detection of clusters in case-control data sets. One thousand sets of controls and cases containing a cluster were de-identified using the LP method (blue line), aggregation by zip code (green diamond), or aggregation by the first three zip code digits (magenta asterisk). The x-axis shows the re-identification probability, which ranged from 0.005 to 1 (original data set). The y axis shows the mean p-

value of the most likely cluster averaged over all data sets. Clusters de-identified using the LP method were detected with greater fidelity (i.e. lower p-value) than those de-identified using aggregation.

The mean p-value of the most likely cluster in the original data sets was 0.029. De-identification using the LP method prior to applying SaTScan resulted in clusters that were slightly harder to detect; under the most stringent strategy with a re-identification probability of 0.005, the mean p-value of the most likely cluster was 0.057. Aggregation decreased the detectability to a greater extent, while offering less privacy protection. Aggregation by zip code, corresponding to a maximum re-identification probability of 0.11, increased the mean p-value of the most likely cluster to 0.094. Aggregation by the first three zip code digits had a maximum re-identfcation probability of 0.012, and increased the mean p-value to 0.21.

Effect of Underlying Population Density

In order to generalize our results to less densely populated regions, we compared the LP method to aggregation for three other counties having a range of population densities. For data sets in Franklin, Plymouth and Middlesex Counties in Massachusetts, we calculated the expected transition distance under the optimal LP strategy for one data point with re-identification probabilities from 0.1 to 0.0001. We also calculated the re-identification probability and expected transition distance under aggregation by zip code and the first three zip code digits (Table 9). The LP method performed favorably relative to aggregation for all of the counties. For example, in Plymouth County, which is about one hundredth less dense than New York County, the LP strategy with re-identification probability 0.0001 is expected to move a data point 1.9 km, while aggregation by zip code moves points a farther distance of 3.1 km and has a five-fold greater disclosure risk.

Table 9: Re-identification probability and expected distance moved for LP strategy and aggregation in counties having a range of population densities.

County name	*ρ* *	LP method *d* †				Zip 5 ‡		Zip 3 §	
		ξ/s¶ = 10E-1	ξ/s = 10E-2	ξ/s = 10E-3	ξ/s = 10E-4	ξ/s	d (m)	ξ/s	d (m)
Franklin County, MA	39.3	0.00	0.00	89.6	4736	2.80E-03	2640	1.20E-05	13226
Plymouth County, MA	276.2	0.00	0.00	62.0	1908	5.50E-04	3123	8.90E-06	19115
Middlesex County, MA	687.1	0.00	0.02	31.5	1105	6.50E-04	1770	4.90E-06	10793
New York County, NY	25846	0.00	0.08	4.3	172	1.10E-03	519	1.20E-04	3866

∗ρ = population density expressed in people per square kilometer

†d = expected distance for strategy in meters

‡Zip 5 = aggregation by five-digit zip code

§Zip 3 = aggregation by first three digits of zip code

¶ξ = re-identification probability, s = number of records in the data set

Discussion

In the current climate of public concern for patient privacy and legislation imposing strict controls on the dissemination of patient-identifiable data, new strategies for de-identifying individual-level data sets while preserving information for disease surveillance and epidemiology are needed. It is imperative that strategies quantify the level of disclosure risk.

For tabular data, such as small area tabulations of demographic, financial and social categories, there is a sophisticated body of research techniques for de-identification. These primarily consist of suppressing of certain cells, aggregating rows or columns, and rounding or adding noise to cells [127, 134-136]. These methods were developed for a different kind of data and problem, and straightforward application to our individual-level *x-y* coordinate data results in previously explored or suboptimal approaches. The binary integer version of our LP method, which is suboptimal to the non-binary method as discussed in the LP De-identification section, is very similar in principle to tabular aggregation methods, while having the advantage of taking the underlying population into account. Tabular methods that round or perturb data, either naively or to preserve features in the data, guarantee that a cell value cannot be known with certainty up to a range of values. These methods do not incorporate geography or population data not contained in the table, and are thus similar to previous perturbation techniques for individual level data. Like those techniques, they would not guarantee privacy in this setting because the risk of re-identifying a permuted location depends on the local geography and population density.

The flexible LP technique presented here for de-identifying spatial data offers a mathematically well-defined re-identification risk, which is simply the maximum probability that any patient in the de-identified data set corresponds to any single individual in the population. This probability holds even if the complete set of transition probabilities $\{P_{ij}\}$ is known to the data recipients.

The strategy ensures that patients are moved as little as possible to guarantee privacy. In both densely and sparsely populated areas, the LP strategy can be expected to move patients a smaller distance than the common practice of aggregating by pre-defined regions. In fact, it moves patients a smaller distance, on average, than every other possible strategy, either deterministic or random, obeying the same re-identification bound that can be expressed as a matrix of transition probabilities.

We illustrated the improved accuracy of the method compared to aggregation for cluster detection for synthetic circular clusters using a circular scan statistic. Like this statistic, most methods in spatial epidemiology consider two point processes, cases and controls. This allows the spatial structure of the disease to be compared with variations in the underlying population. It is important to note that prior to applying any statistical method, both the cases and the controls must be de-identified using exactly the same strategy, even if the control locations do not represent a threat to privacy. If only the cases are moved, then spurious clusters may be formed by relocating dispersed individuals to the same or nearby locations.

The accuracy of the re-identification bound depends on a few assumptions. First, the underlying population size at each location must be known in advance, although the method appears to be robust to small inaccuracies (see supporting information). Second, the data recipient must not have knowledge to suggest that membership in the data set is not completely random; otherwise it may be possible to apply a de-noising technique to reveal deterministic structure in the data. This is a limitation of the method because the user may guess that membership is not random from the de-identified data set itself. Devising such a de-noising technique, however, would be difficult in general because the noise added by the LP model depends on the original data in a complicated way [137]. Third, we assume that no other information is available to help identify individuals. Ensuring privacy in the face of existing or future additional information is a highly nontrivial problem that has not been adequately addressed by existing methods for individual-level exact location data [134, 138], although progress has been made for other types of data [139-141]. In the simplest case, a coarse discrete identifier can be incorporated into the de-identification procedure. For example, if the final version of the data set is to contain both the location and the sex of each patient, then a de-identification strategy may be developed independently for each sex represented. This is not always possible because stratified population data may not be available, and it becomes intractable for finely-grained identifiers, or multiple identifiers having many possible combinations of values.

For individual addresses, we recommend using a population size of 1 for each address in the LP model. This limits the probability of associating any household with a case to the re-identification probability. Because the public may not feel comfortable with any addresses released in a de-identified set, even if the probability that an individual at each address has the disease is very small, the set $\boldsymbol{B}$ of final locations should be grid points or small administrative units instead of addresses.

The measure of privacy protection proposed here captures what is essentially important to a patient: "Will I be identified as having a disease as a result of the disclosure?" Several other measures of confidentiality have also been proposed. These include Spruill's measure for business data [142], equivalent in the spatial context to the proportion of records in the de-identified set that lie closer to their original location than to all other locations in the original set. The value of the measure for our LP strategy depends not only on the privacy bound ξ, but also on the number and locations of original records and on the particular values for destination locations drawn from the multinomial distribution. However, Spruill's measure does not always capture intuition about privacy. For example, creating a de-identified set by shuffling the exact locations of all patients in the original set measures well by Spruill, but is clearly unacceptable for privacy protection. Conversely, assigning completely random locations to de-identify a data set of two patients measures poorly by Spruill, but would certainly preserve privacy.

Armstrong et al. also proposed four other measures of confidentiality. The first of these is a qualitative measure of vulnerability to geographical knowledges, under which our LP strategy has no disclosure risk. The second measures the ability to infer from the de-identified set regions within the map having a high disease risk. Like aggregation and random perturbation, our LP method may reveal regions of high disease risk. However, this is both a strength and a liability of the method because the de-identified set may be used to assess spatial variation in the disease risk. The third measures the ability to re-identify all the patients, given the identity of some of the patients, and the final confidentiality measure is the minimum number of unlabeled locations from the original data set that can be used to compromise the entire de-identified set. As with aggregation, there is minimal risk under our LP strategy by these measures. If one patient is re-identified in a data set of s patients created using the LP method with disclosure risk ξ, then the problem of re-identifying a different patient is equivalent to the problem of re-identification starting from a data set created with a slightly lower risk of disclosure $\xi \cdot (s - 1 / s)$, but in which one of the census numbers n_i had been overestimated by one in the model. This is likely to have little effect on the disclosure risk. Please see the supporting information for further discussion of inaccurate census estimates.

Chapter IV: Reverse Identification Potential of Authentic and Anonymized Geographical Data

In this section, we explore two distinct threats models of reverse identification of geographical data. The first model demonstrates that geographical data may be mined from low resolution maps that contain points representing cases, commonly used in journals and in public health practice. The second threat model deals with obfuscative and cryptographic algorithms and how they may be susceptible to weakening when it is possible for an adversary to produce output from the algorithm according to adversary-provided input. Under this model, an adversary could use an anonymized data provisioning system to request patient data from a RHIO or other health network several different times. This use case is not uncommon—often disease surveillance systems request the patient visits for the previous week each day, in a sliding window of requests. If the anonymized results are produced each time they are requested, it is possible to average the visit geocodes (which actually only anonymity afforded by the algorithm may be reduced.

Exploiting Repeatedly Non-deterministically Anonymized Spatial Data to Re-identify Individuals: A Vulnerability and Proposed Solutions

The work in this section is described in a manuscript entitled Re-identification of home addresses from spatial locations anonymized by Gaussian skew in the International Journal of Health Geographics with Dr. Shannon Wieland from the Harvard Medical School and Dr. Kenneth Mandl from the Children's Hospital Informatics Program and the Harvard-MIT Division of Health Sciences and Technology.

Abstract

Knowledge of the geographical locations of individuals is valuable in the practice of spatial epidemiology, yet poses a substantial risk to privacy. One approach to preserving privacy is the use of algorithms that de-identify spatial data by blurring location information. We investigate whether such algorithms may be weakened when an adversary can access multiple non-deterministically anonymized versions of the original data set. We are able to more accurately re-identify patient addresses when multiple anonymized copies are shared, in close alignment with theoretically expected values. With only 10 anonymized copies of an original data set, we find that the average distance to original addresses decreased from 0.7 km to 0.2 km using both uniform skew and Gaussian skew anonymization methods, and with 50 anonymized copies of an original data set, we find that the average distance decreases from 0.7 km to 0.1 km. We demonstrate that multiple anonymized versions of the same data set can be used to ascertain original geographical

locations, and present a privacy risk. We explore solutions to this problem that include infrastructure to support the safe disclosure of anonymized medical data to prevent inference or re-identification of original address data, and the use of a Markov-process based algorithm to mitigate this risk.

Background

To develop broadly integrated national healthcare information infrastructure, the utility of sharing personally identifiable data for clinical care, public health and research must always be weighed against privacy concerns. For example, automated outbreak detection systems for surveillance of influenza and bioterrorism, use data from a variety of sources (hospitals, clinics, laboratories) for aggregation, analysis and investigation [73, 74, 143]. For the detection of spatial clustering among disease cases, these aggregation systems achieve optimal detection sensitivity and specificity when using the most complete, accurate patient location data [126].

We have previously described a spatial de-identification algorithm that blurs precise point locations for patients, moving them a randomized distance according to a 2-dimensional Gaussian distribution with variance inversely proportional to the square of the underlying population density [144]. Other spatial anonymization approaches that have been employed include random skews, affine transformations, data aggregation techniques, and the use of software agents to preserve confidentiality [128, 145]. Anonymization of patient address data by reassignment of geographic coordinates allows privacy preservation while sharing data for disease surveillance or biomedical research [144]. As the volume of personally-identifiable health data that is electronically transmitted and published has consistently increased [146], so has the magnitude of the threat to privacy. Geographical information is particularly identifying; we have demonstrated that it is possible to correctly identify most home addresses even from low resolution point-maps commonly published in journal articles [122].

We specifically explore whether de-identification algorithms that use spatial blurring—a non-deterministic process—may be susceptible to weakening when an adversary can access multiple anonymized versions of the same original data set [147]. For example, if data anonymized by a Gaussian blurring function were available upon request from a data source, the adversary could request anonymized patient data repeatedly. Since the data are non-deterministically anonymized, the results vary each time they are requested. By averaging the geocoded values for each visit, the anonymity afforded by the blurring algorithm may be reduced (Figure 25 illustrates the effect of averaging locations across the repeated anonymization passes to increase resolution for re-identification).

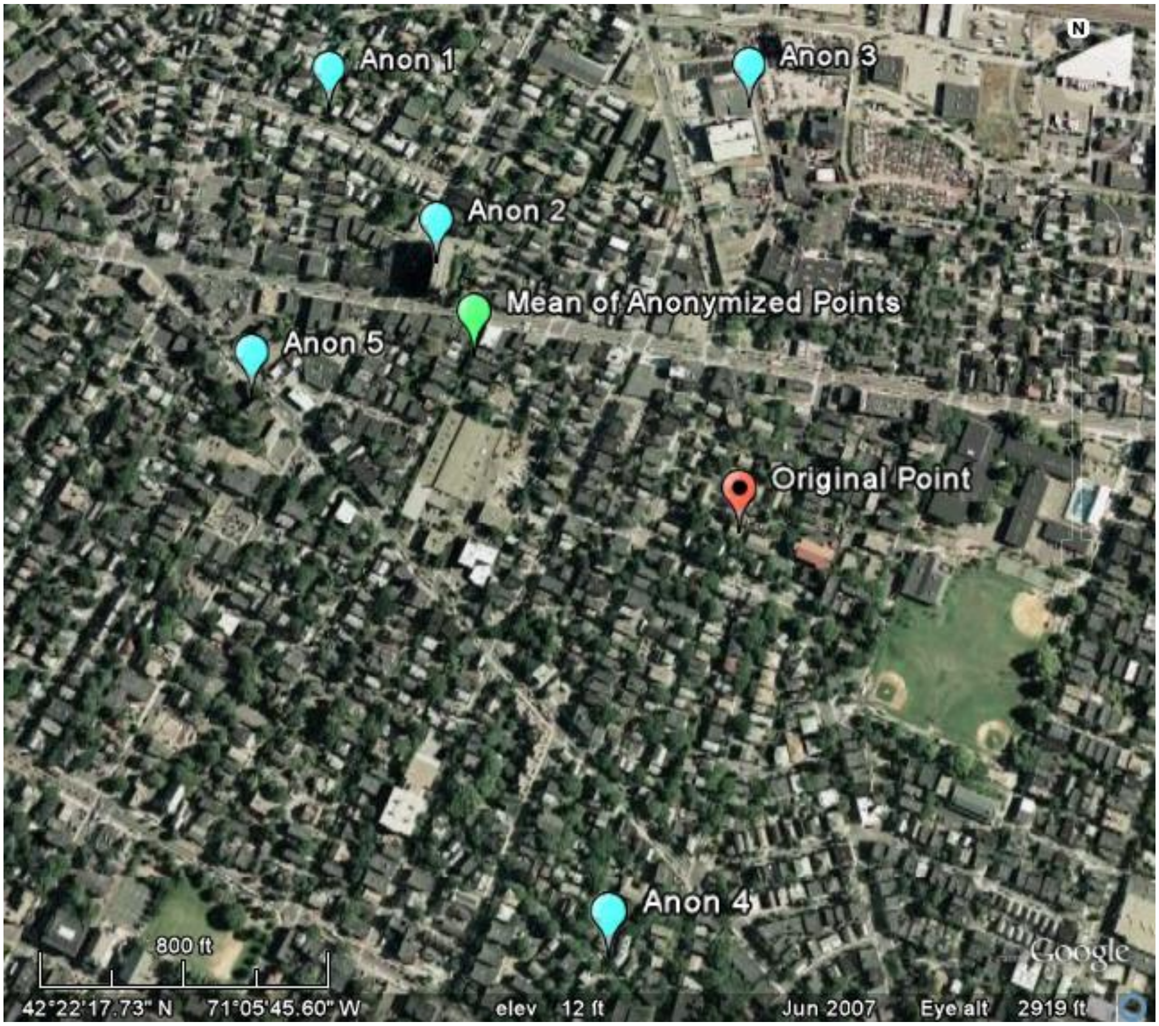


Figure 25: Example of anonymized points that have been averaged. An original data point (red) was anonymized using a population-density adjusted Gaussian skew algorithm five times (light blue points). Those points were averaged and the average coordinate value is plotted (green). The average of the anonymized points is nearer to the original point than each of the anonymized points. (Courtesy Google Earth.)

Disease and outbreak detection systems often transmit data from a variety of sources (hospitals, clinics, laboratories) to public health departments for aggregate review, analysis and investigation. These aggregation systems are designed to improve the sensitivity and specificity of outbreak detection by evaluating clustering on a more complete set of data, avoiding referral bias from catchment areas. If these systems are provided with the ability to request anonymized data, or are provided this data in a sliding temporal window (say, the last six weeks of data) these systems may anonymize all records anew. This may be a sensible strategy to prevent disclosure of information about visit dates of cases, but if cases are anonymized anew for each new window, there will be a steep decline in data set privacy.

Here, we quantitatively demonstrate this vulnerability in two common anonymization approaches. We produce multiple anonymized data sets using a single set of addresses and then progressively average the anonymized results related to each address, characterizing the steep decline in distance of the re-identified point to the original location, (and the reduction in privacy) at each stage. Next, we propose and discuss two solutions to this specific class of vulnerabilities. The first tightly couples anonymization to a distributed health infrastructure that exchanges the data, so that it can control the number of copies distributed to any one party. The second is an extension to the spatial anonymization process employing a Markov process for increasing the anonymity of a 2-dimensional data set anonymized by Gaussian skew.

Methods

Geographical test data sets

A data set containing artificially-generated geocoded values for 10,000 sample patients was created using a spatial cluster creation tool [148, 149]. All points were uniformly distributed within a circle of radius 800m centered in Boston, MA, and assigned a unique numeric identifier for tracking. Each of the geocoded addresses was then anonymized using a Gaussian 2-dimensional spatial blur skew that was adjusted for population density [144], fifty separate times. A second anonymization approach, a uniform skew, was used to create a second group of 50 anonymized data sets. Each geocode that was anonymized using the uniform skew method was moved a distance, in meters, ranging from [-λ,λ], independently in each dimension. Figure 26 describes the 2-dimensional probability distribution function for both of these anonymization algorithms.

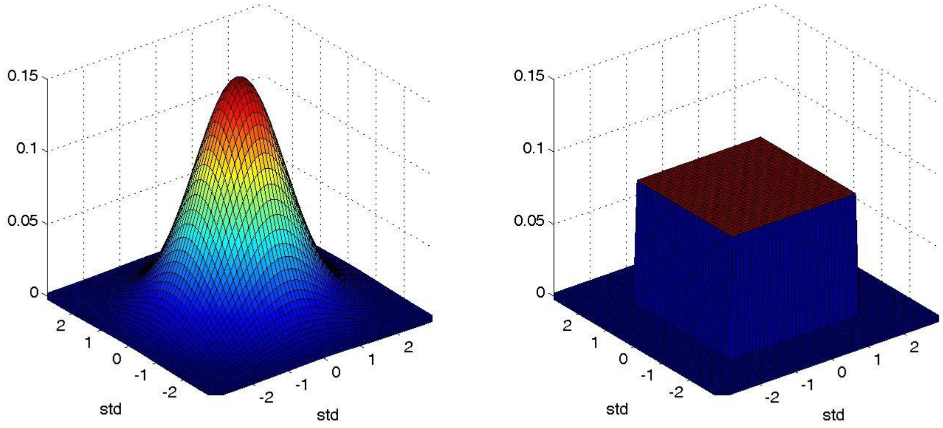


Figure 26: Anonymization algorithm translation probability density functions. Probability distribution functions for the two anonymization methods, 2-dimensional Gaussian skew (left) and uniform skew (right).

Population-adjusted 2-dimensional Gaussian skew

In the simplest case, the Gaussian skew anonymization procedure is a probabilistic strategy that reassigns an original point, with coordinates $\binom{x_0}{y_0}$, to a new location based on two Gaussian probability density functions

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{\frac{-(x-x_0)^2}{2\sigma_x^{\ 2}}}, \quad f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{\frac{-(y-y_0)^2}{2\sigma_y^{\ 2}}}. \qquad (1)$$

These are simply 1-dimensional Gaussians with means equal to the original coordinates x_0 and y_0, respectively, and standard deviations σ_x and σ_y. The parameters σ_x and σ_y are proportional to the desired level of anonymity *k*, and are inversely proportional to the population density at $\binom{x_0}{y_0}$. In other words, the greater the anonymity desired, or the lower the underlying population density, the farther points are moved on average.

Re-Identification through averaging

With each subsequent anonymized version, the geocoded points that referred to the same individual address were averaged to estimate the original address. For re-identification inference number *n*, the anonymized versions of the same address were averaged from data sets [1, *n*], as shown in Figure 27. For example, the second re-identification inference data set included the averages of addresses from anonymized data sets 1 and 2, the third inference data set included anonymized data from data sets 1, 2, and 3, and so on. After each pass, the distance between the average anonymized point and the original address was calculated.

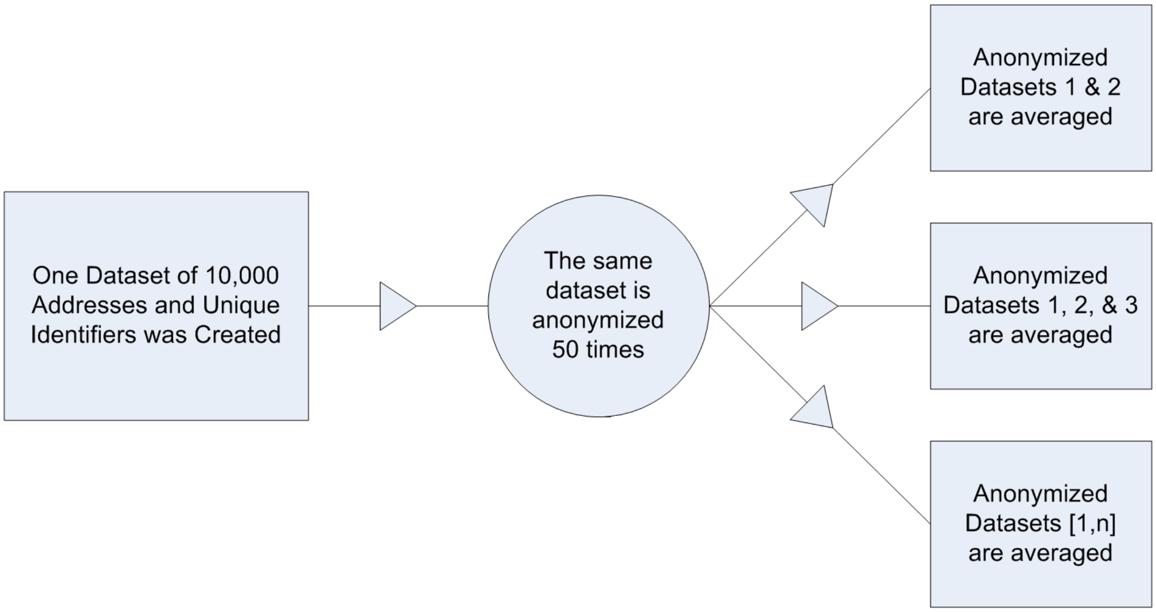

One Dataset of 10,000 Addresses and Unique Identifiers was Created
The same dataset is anonymized 50 times
Anonymized Datasets 1 & 2 are averaged
Anonymized Datasets 1, 2, & 3 are averaged
Anonymized Datasets [1,n] are averaged

Figure 27: Experimental methods design. One data set of 10,000 artificially generated case locations and unique identifiers were created. The data set was anonymized 50 times using a 2-dimensional Gaussian-based skew, and 50 times using a 2-dimensional uniform skew.

Results

Re-Identification of points anonymized using Gaussian and randomized skew

Additional information was ascertained from multiple anonymized copies of one original set of point locations, significantly weakening the anonymization used. The average distance to the original addresses after one anonymization pass, which represents the previously described [144] use of an anonymizing algorithm, was 0.69 km. After each point was inferred using the average of fifty Gaussian skew anonymization passes, the mean distance from the average of all of the anonymized points to the original point in the data set was reduced to 0.1 km. Similarly, when the anonymizing algorithm is a uniform skew (a random skew that involves moving a point randomly within a square), re-identification attempts using the average of several anonymized data sets also reduced data set privacy markedly. The average distance to the original addresses after one anonymization pass, the traditional use of such algorithms, was set at 0.69 km, to match the level of skew used in the 2-dimensional Gaussian data sets. As in the case of the 2-dimensional Gaussian skew, the average distance to the original point was also reduced to just under 0.1 km after averaging 50 anonymized data sets.

The average distance to the original address is plotted as a function of the number of separate anonymization passes used in the re-identification inference, for both anonymization methods in Figure 28. Attempts at inferring the original addresses using multiple anonymization passes, show that the average distance inversely varies with the square root of the number of anonymized data sets used in the inference. There is a sharp decrease in the average distance to the original address with 10 anonymization passes and thus a sharp decrease in data set anonymity.

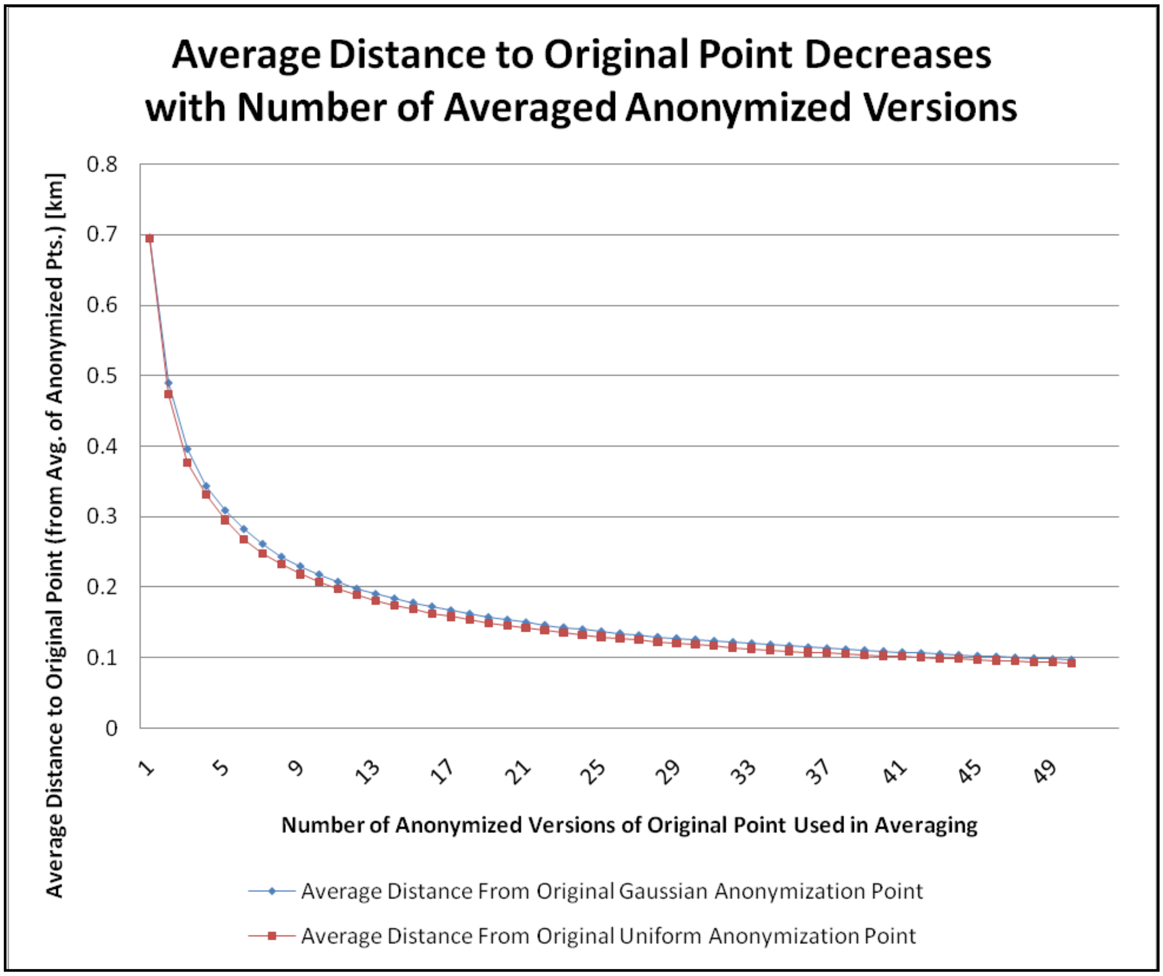


Figure 28: Average distance to original point vs. number of anonymization versions. The average distance to original point [km] vs. number of anonymization versions used in averaging is plotted for both Gaussian and uniform skew.

Discussion

Re-Identification of data anonymized with Gaussian and randomized skew

Re-anonymizing a single patient located at $\binom{x_0}{y_0}$ n different times using Gaussian skew is equivalent to observing a sequence L_1, L_2, …, L_n of independent, identically distributed two-dimensional Gaussian random variables (all having the same probability density function). The average of these n observations

$$\frac{\sum_{i=1}^{n} L_i}{n}, \qquad (2)$$

is itself a two-dimensional Gaussian random variable with mean $\binom{x_0}{y_0}$ and covariance matrix

$$\begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{n} \end{bmatrix}.$$

In other words, the x- and y-coordinates are independent Gaussian random variables, each having a standard deviation of $\sigma/\sqrt{n}$. Hence, by taking the average of the anonymization passes, one can obtain the equivalent of a single anonymization pass under a less stringent Gaussian skew anonymization strategy with standard deviation $\sigma/\sqrt{n}$; for 100 passes, reducing the skewing standard deviation along each axis by a factor of 10.

In the uniform skew anonymization procedure, a patient at $\binom{x_0}{y_0}$ is moved with equal probability to any position in the square $[x_0 - \lambda, x_0 + \lambda] \cdot [y_0 - \lambda, y_0 + \lambda]$. The new position is thus a two-dimensional uniform random variable, with mean $\binom{x_0}{y_0}$ and covariance

$$\begin{bmatrix} \frac{\lambda^2}{3} & 0 \\ 0 & \frac{\lambda^2}{3} \end{bmatrix}.$$

By the central limit theorem,

$$\frac{\sum_{i=1}^{n} L_i}{n}$$

is approximately normally distributed with mean $\binom{x_0}{y_0}$, and covariance matrix

$$\begin{bmatrix} \frac{\lambda^2}{3n} & 0 \\ 0 & \frac{\lambda^2}{3n} \end{bmatrix}.$$

Hence as the number of observations increases, the average of the observations tends to fall nearer to the original point.

It is important to note that while the difference in change to the covariance matrices would appear to make the Gaussian and uniform anonymization skews similar in their ability to protect privacy, this is not necessarily the case. The quantifiable estimate of anonymity, *k*-anonymity (a metric for data set privacy where *k* is the number of people among whom a specific individual cannot be distinguished [82]), that is achieved by each method is different – the Gaussian skew yields higher levels of *k*-anonymity than the randomized skew does with respect to the average distance moved for each case in a data set. We previously described a method for estimating spatial k-anonymity [144].

One might fear that an adversary could do even better, devising a novel strategy that uses the sequence L_1, L_2, ..., L_n to get even closer to the original point $\binom{x_0}{y_0}$ than a Gaussian with this reduced variance; however, this is not possible without additional data. Stein showed that given *n* observations of a two-dimensional

Gaussian random variable, the most efficient estimator of the mean of the Gaussian is simply the average of the points [150]. Although this seems intuitive for two dimensions, it is surprisingly not the case for three and higher dimensions [150, 151].

Anonymizing within a distributed network or health information exchange

We believe that these results make a compelling case for infrastructure to control disclosure of anonymized data, so that the risk of this vulnerability is reduced. In Figure 29, we show an infrastructural solution for integrating anonymization into a distributed network that transmits health data [152, 153]. Ideally, data sources -- and even patients -- would be able to set a preferred level of data disclosure for a number of different purposes including research studies that integrate their clinical data, outcomes and public health surveillance. A data provisioning system could then distribute data to consumers at a variety of anonymized levels, under a clear set of policies and authorization requirements.

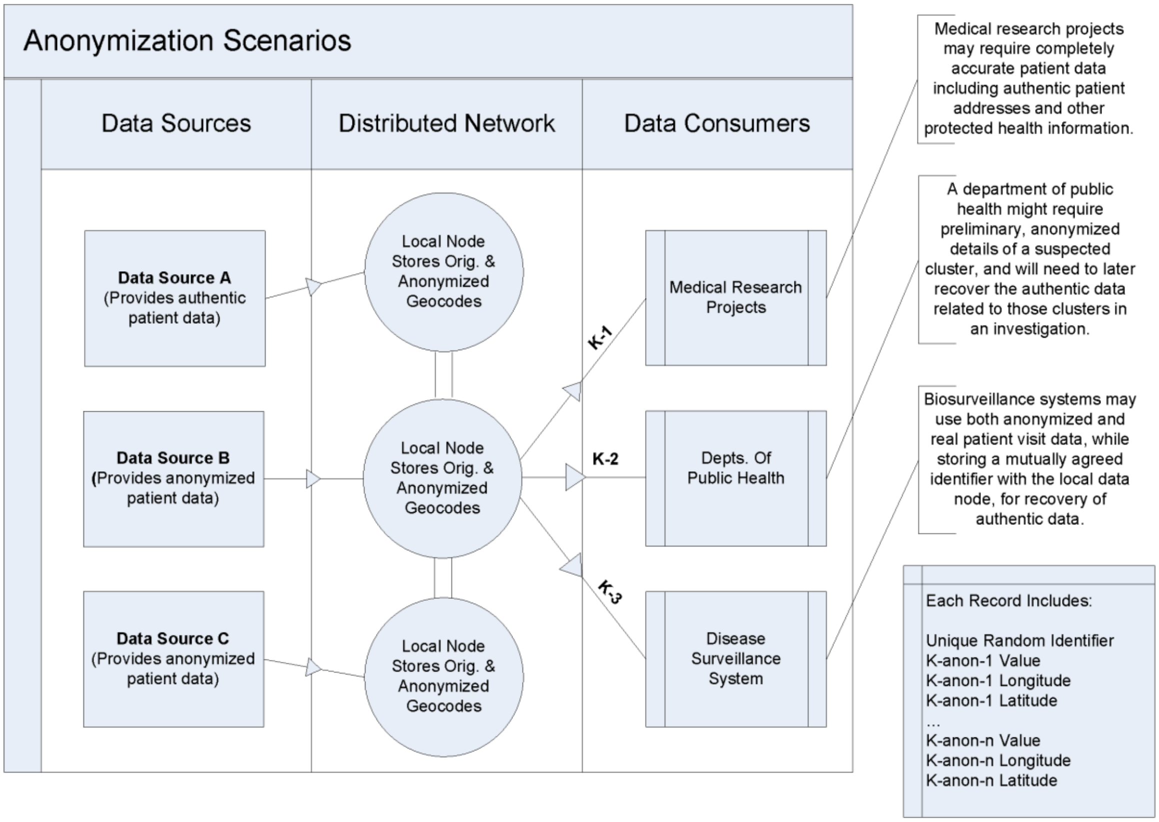

Anonymization Scenarios
Data Sources
Distributed Network
Data Consumers
Data Source A (Provides authentic patient data)
Data Source B (Provides anonymized patient data)
Data Source C (Provides anonymized patient data)
Local Node Stores Orig. & Anonymized Geocodes
Local Node Stores Orig. & Anonymized Geocodes
Local Node Stores Orig. & Anonymized Geocodes
K-1
K-2
K-3
Medical Research Projects
Depts. Of Public Health
Disease Surveillance System
Medical research projects may require completely accurate patient data including authentic patient addresses and other protected health information.
A department of public health might require preliminary, anonymized details of a suspected cluster, and will need to later recover the authentic data related to those clusters in an investigation.
Biosurveillance systems may use both anonymized and real patient visit data, while storing a mutually agreed identifier with the local data node, for recovery of authentic data.
Each Record Includes:
Unique Random Identifier
K-anon-1 Value
K-anon-1 Longitude
K-anon-1 Latitude
...
K-anon-n Value
K-anon-n Longitude
K-anon-n Latitude

Figure 29: Integration of anonymization within distributed EMR infrastructure. Integration with a distributed electronic medical record infrastructure: a distributed data provisioning system provides anonymized spatial address data to three data consumers at three distinct *k*-anonymity privacy levels.

Removing other identifying information from data sets to avoid re-linking

The vulnerability described in this paper relies on the ability to link anonymized data sets together using additional identifiers, or other demographic or clinical data. One possible solution is to swap the addresses in a given data set so that they are effectively unlinked with any unique clinical fields or identifiers, such as a medical record number. This unlinking of spatial data from unique identifiers, however, poses additional challenges: unlinking from any demographic identifiers could reduce the ability to conduct informative disease surveillance, or worse, could make it difficult to actually uncover the addresses of clustered cases when necessary, which is certainly a priority for a public health investigation.

This can be mitigated through the use of randomly generated identifiers for each anonymized instance of a specific record, stored for use in re-linking anonymized data with original data. Additionally, when attempting to determine correlates or predictors of disease, these additional fields may prove important for group stratification. With knowledge of the specific anonymization algorithm and background knowledge such as regional demographic data, it may be possible to further weaken some anonymization algorithms even without repeated attempts.

When considering only two dimensional geographical data, the best way to estimate original locations from several anonymized versions of the same original data set is to average the anonymized longitudes and latitudes. However, there are even more advanced re-identification techniques that can be used to improve the resolution of cases in practice, using data sets with three or more dimensions. If additional fields or identifiers are included in the data set, and those fields are in any way not randomly distributed (anything other than a randomly generated identifier), their presence has the potential to help achieve a higher resolution on the spatial coordinates, even if they do not contain geographical information. This is because there may be additional implicit information linked with spatial addresses in the other dimensions (or fields) that can lend intuition about the distribution of the anonymized spatial coordinates, using approaches that are more advanced than averaging of the fields to estimate the original location [151].

Increasing anonymity using an algorithm based on a Markov process

As shown in Figure 29, one possible anonymization scenario is the sharing of data at a variety of privacy levels with different data consumers. To prevent privacy degradation by averaging when sharing data at multiple levels of *k*-anonymity [82], a Markov state process can be used to successively generate increasingly anonymized versions of the data set. The Markov property guarantees that several versions anonymized this way cannot be used to infer additional information about a patient's location. One example might be the need to provide multiple versions at two anonymized levels, one at *k*=50 and another at *k*=100. If the anonymization process is restricted to increasing the anonymization level to *k*=100 by increasing the skew level from the *k*=50 data set, and *not* from the original data set, there will be no way to decrease the privacy below the *k*=50 level, simply by averaging the two data sets. This is illustrated in a Markov process model in Figure 30.

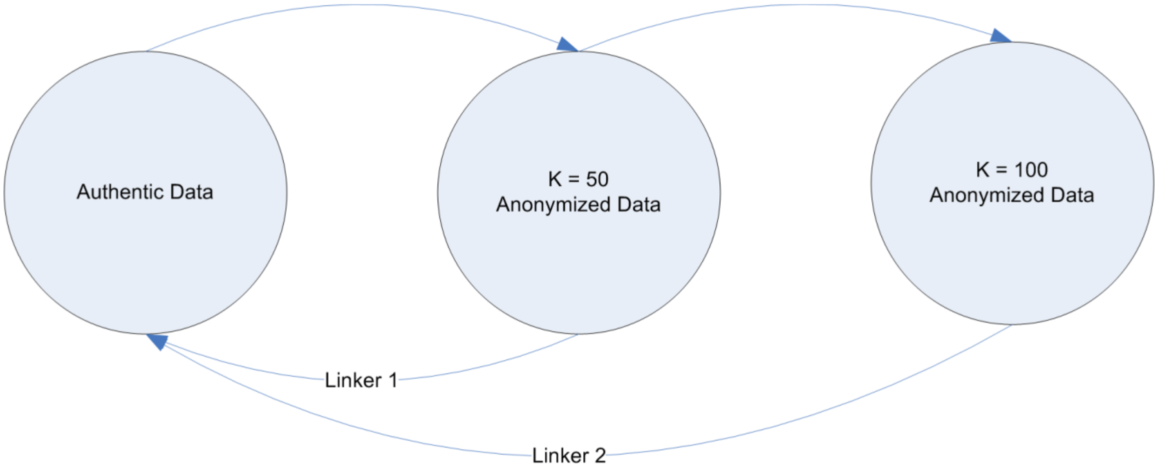


Figure 30: Markov anonymization process to increase data set anonymity. Markov processes to increase the anonymity level in a data set: an increase in the anonymity level of a data set, for example, increasing from k=50 to k=100, could be achieved by increasing the skew level of the k=50 data set without knowledge of the authentic data. If increases are done in this way, the risk of a reverse identification attempt using averaging can be avoided.

While infrastructure for controlled exchange of anonymized health data protects against some vulnerabilities, there are still other methods that could reduce the privacy level of a data set. For example, it is still possible to gain insight into the actual distribution of cases anonymized with knowledge of physical boundaries, highly constrained patient distributions, or other clinical or demographic information about cases. Further study is needed to adequately constrain the anonymized geographical distributions of cases such that this risk is minimized.

Conclusions

In order to protect privacy when using spatial skew algorithms, the number of distinct anonymization results or passes that represent the same data must be controlled. Limiting the generation or disclosure of more than one version will avoid re-identification through averaging. Alternative approaches include integration of anonymization into data provisioning systems to achieve such a restricted data release, or the use of a Markov process to generate multiple anonymized data sets of the same records. These approaches avoid running the algorithm anew with each request, reducing the variation that is at the root of the vulnerability.

An unsupervised classification method for inferring original case locations from low-resolution disease maps

Preface

This section is comprised of joint work with Dr. John Brownstein, Assistant Professor at Harvard Medical School and faculty at the Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology. This research was published as a letter to the medical research community at the New England Journal of Medicine, and as a technical paper with experimental data in the International Journal of Health Geographics.

Background

Geocoding patient data – translating the plaintext addresses of patients into longitudes and latitudes – has become routine and enables display and analysis of disease patterns. Many public health surveillance systems and academic investigations rely on specific case locations for identifying patterns, correlates, and predictors of disease [71, 154, 155]. Maps that display such geocoded health data are frequently presented publicly and published electronically and in print.

However, publishing patient address locations on maps also creates a risk of re-identification of individuals [122, 156-158]. We recently reported an inadvertent breach of privacy across five major medical journals, identifying 19 articles from 1994-2004 that include maps with patient addresses plotted as individual dots or symbols [122, 156]. From these publications, over 19,000 patient addresses are plotted on map figures. We demonstrated through a process of reverse identification that the home addresses of many of these patients could be discovered, despite the low resolution of the disease maps.

Here, we provide the details of that method. We rely on unsupervised classification of the spectral properties of the map image to identify case locations. Because we do not have available to us the original addresses of the patients represented in the published maps, we devised an indirect approach relying on simulation.

Methods

We sought to quantify the degree of re-identifiability of patient home addresses from published maps. To accomplish this, a hypothetical low-resolution map of geocoded patient addresses is produced and then the accuracy with which patient addresses can be resolved (reversely identified) through a five step process is measured. First, an original, prototypical patient map for an urban metropolitan area in Boston, MA was produced (Figure 31). Using building parcel outlines for the city of Boston [159], we generated a synthetic or hypothetical set of patient addresses by randomly selecting buildings. Cases were assigned by a stratified sampling design of building parcels to achieve a distribution representative of all building and population densities in the city. Buildings were selected with equal spacing of 0.02 degrees. A total of 550 addresses were randomly selected. Centers of the selected building were then calculated and plotted on a county map of Boston to represent patient addresses [160]. One important issue is that our use of the building footprint for geocoding does not mirror the reduced accuracy obtained from geocoding addresses. Address geocoding will have a series of associated errors that may be related to the underlying structure of a geographic area (e.g.: road length, parcel size, housing density) [161].

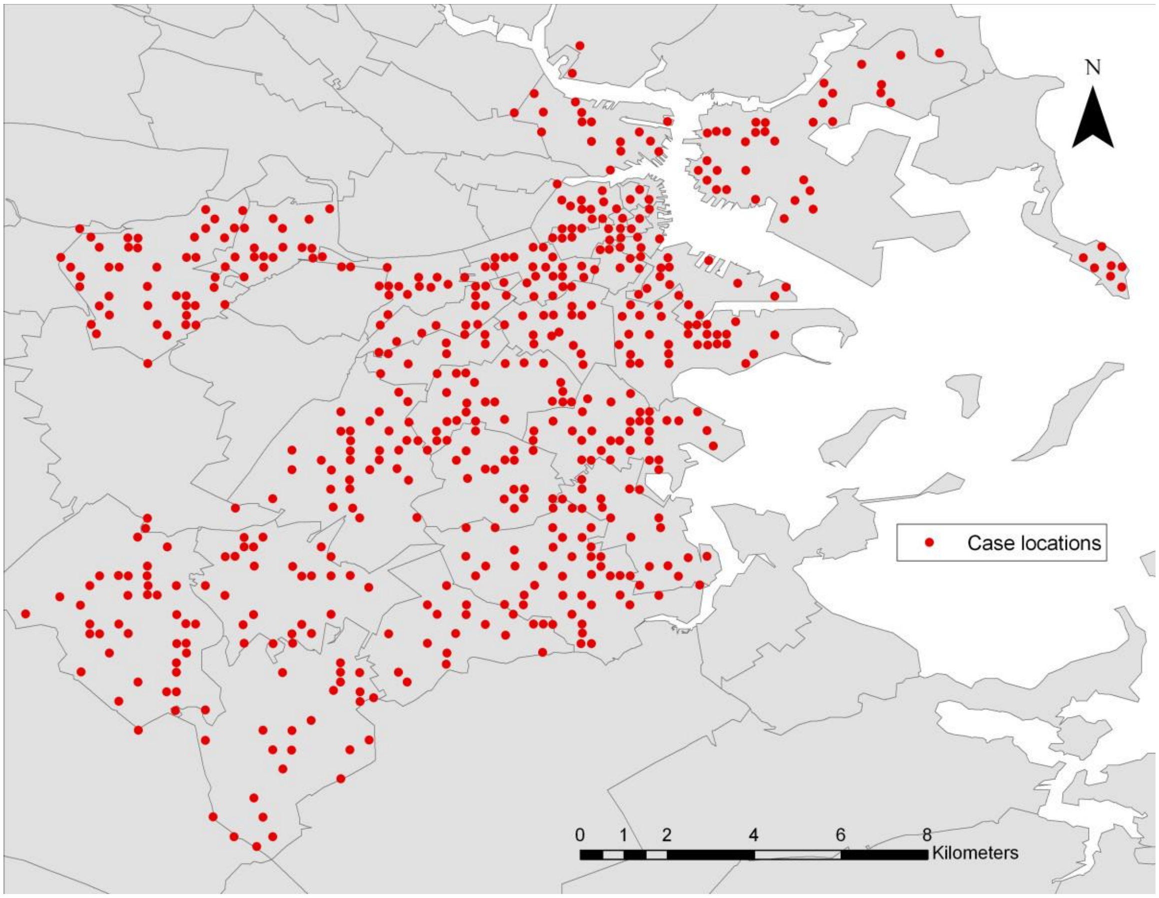


Figure 31: Prototypical patient map for Boston, Massachusetts. The image displays 550 addresses selected by stratified random sampling design. The original JPEG image used in the analysis had a resolution of 50 dots per inch (550x400 pixels), a file size of 129kb and a scale of 1:100,000. This would be a typical output for web display and usually lower resolution than would be shown in a slide presentation or in a peer-reviewed publication.

We created a JPEG image with a resolution of 50 dots per inch (dpi), 550x400 pixels, a file size of 129kb and a scale of 1:100,000. This low resolution is typical for web display and is lower than generally used in slide presentations. Also the re-identification of patient addresses was evaluated using a higher-resolution map (266 dpi, 2926x2261 pixels, 712kb, 1:100,000), often the minimum resolution for peer-reviewed publications.

There are several steps involved in reversely identifying a patient address. First, the sample map is scanned or imported into GIS software as an image file [162]. Second, the imported map is georeferenced. The cartographic projection of the map is used to set the coordinate system. Generally, the projection of a published map would be unknown and the correct projection would need to be found by manually matching the image of the map to an image of a correctly georegistered map of the same area. In this case, we have a priori knowledge of the map projection. In either case, ground control points are selected on the image using a corresponding vector outline of the map area to re-project the image file of patient locations and reference it to a coordinate system. In this example, an outline of counties around Boston provided by the US Census Bureau to set the ground control points [160]. The process of scanning and georeferencing the disease map parallels the methodology detailed by Curtis et al [158]. Third, using image analysis software [163], unsupervised classification of the georeferenced map is performed. Given the spectral properties of the image file, pixels are classified so that pixels representing the patient points are aggregated together. Fourth, a reclassified raster map (an image composed of individual pixel elements arranged in a grid) that only contains patient points is extracted and converted to a vector file. Finally, Coordinates of the patient points are then calculated.

Accuracy of reverse geocoding was measured as (a) the number of correctly identified patient addresses (b) the distance between the reversely identified address coordinate and the boundary of the building of the patient home address and (c) the number of buildings in which the patient could reside, given the reversely geocoded address. To calculate (c), we estimated the minimum buffer size from the predicted location needed to contain the centroid of the correct address. Accuracy in this case is therefore defined as the number of incorrect addresses within this buffer.

Results

Our reverse identification method correctly identified 26% (144/550) of patient addresses precisely, from a sample map with low-resolution GIS output. We observed increased detection with the higher-resolution publication quality output to 79% (432/550) of patient addresses identified exactly.

For the low resolution presentation quality map, reversely geocoded locations were on average within 28.9 meters (95% CI, 27.4-30.4) of the correct original address (Figure 32a). On average, correct patient address was identified within eight buildings (95% CI, 7.0-8.3). Overall, 51.6% of addresses were identified as being at any of five buildings, 70.7% at any of ten and 93% at any of 20 (Figure 32b). For the higher resolution publication quality map, all addresses were predicted within 14m of the correct address. This distance is well within the footprint of most apartment buildings and even many single family residences. While most addresses (79%) could be identified to a single building, the maximum number of buildings in which the patient could reside, given the reversely geocoded addresses was 11 buildings.

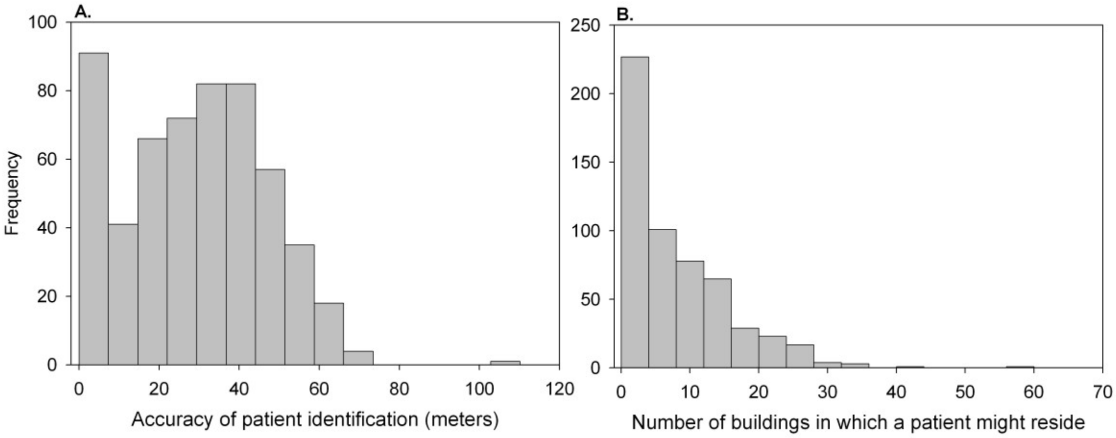


Figure 32: Accuracy of reversely identifying patient location from a hypothetical low-resolution patient map in Boston, Massachusetts. The accuracy of the reverse identification was determined by (A) the distance between the reversely identified and the original addresses and (B) the number of buildings in which the patient could reside, given the reversely geocoded address. The reversely geocoded location was on average within 28.9 meters (95% CI, 27.4-30.4) of the correct address. The mean number of buildings in which the patient might reside was 7.7 (95% CI, 7.0-8.3).

Predictions of patient location were accurate in both densely-populated urban settings as well as suburban regions, as illustrated in Figure 33. Among those addresses precisely identified, there was no observed effect of housing density on the rate of patient addresses re-identification. However, given the variation in number of individuals per housing unit, we expect that the anonymity of patients in suburban single family houses would be significantly reduced compared to urban areas. Locales with a high probability of living in large apartment buildings afford greater anonymity. In this study, we essentially controlled for the variability of geocoding accuracy by using building footprint data rather than address data. Previous research has shown that housing density may have substantial impact on address geocoding accuracy [161].

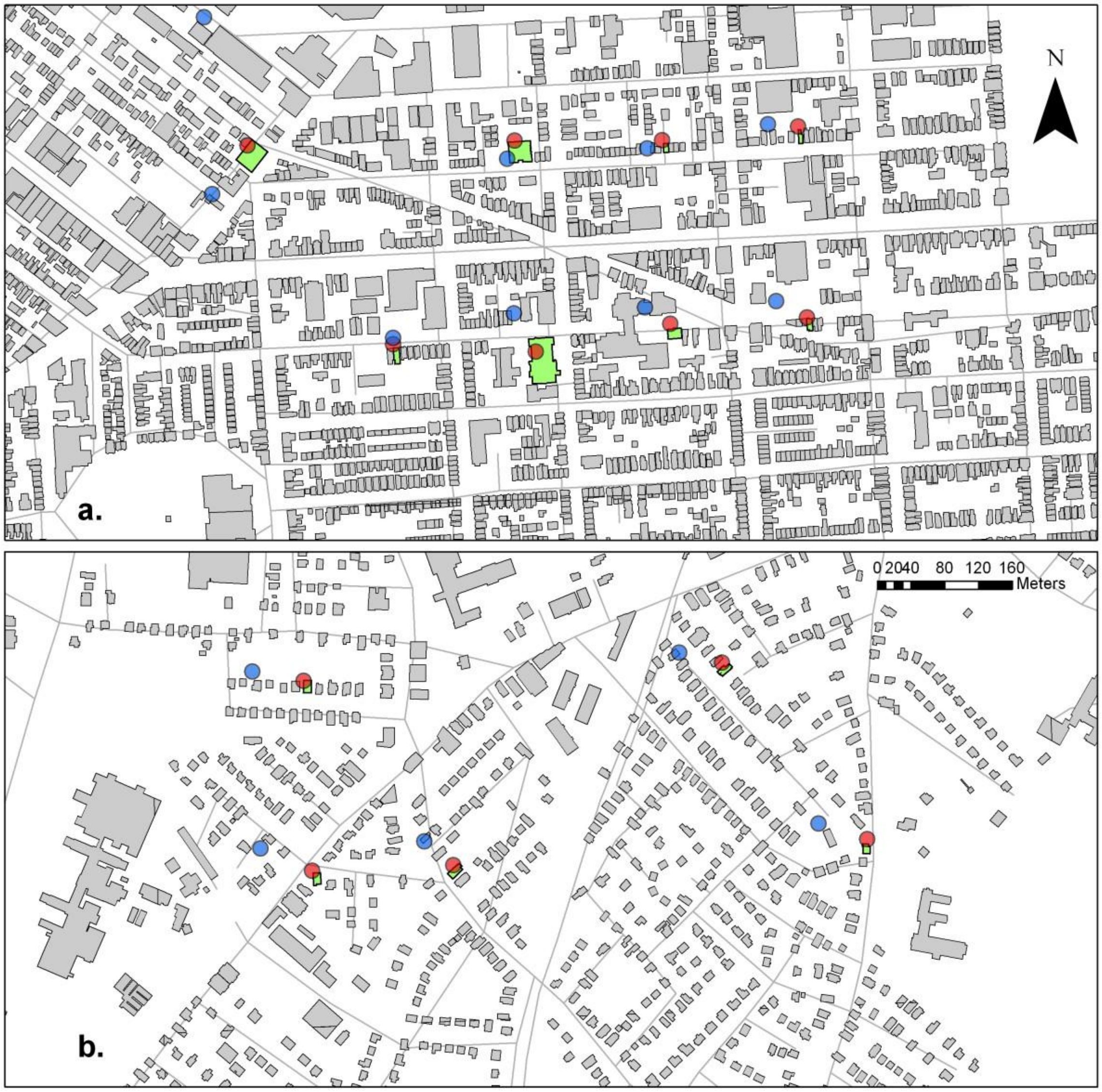


Figure 33: Results of reversely identifying patient addresses in Boston, Massachusetts. The green buildings are the randomly selected patient locations. The blue points are the predicted locations of the cases from the presentation quality map (50 dpi) and red points are predictions from the publication quality map (266 dpi). Proximities of the predicted to the actual location are displayed for both (A) a high density urban area and (B) a low density suburban area.

Discussion

Our results demonstrate that even lowering the resolution of a map displaying geocoded patient addresses does not sufficiently protect patient addresses from re-identification. Despite the low quality of output sources, these images – based on high precision input sources – preserve positional accuracy. Using a low quality map that would serve the purpose of web or presentation display, we were able to precisely identify more than one quarter of all randomly selected home addresses and on average patients could be identified to a city block or within one of eight buildings. Using a map with minimum resolution for peer-reviewed publication, we could identify almost all patient addresses and on average patients could be identified within 14m.

The ultimate accuracy of the patient re-identification will no doubt depend on the number of individuals residing at these addresses. In the case of multi-family apartment dwellings, address identification may still afford a certain level of privacy protection. In the case of single family dwellings, re-identification becomes much more likely. However, even in the best case scenario of an urban area multi-family apartment building, an additional concern is that individuals at these addresses can be fully re-identified when linked with other data sets or by using other characteristics supplied in the publication [82]. Previous research has shown that combinations of seemingly innocuous data are adequate to uniquely identify individuals with a high level of reliability [164]. For example, an experiment using

1990 U.S. Census summary data surprised the public health community by showing that data sets previously thought to be adequately de-identified, containing only 5-digit ZIP code, gender and date of birth, could be linked with other publicly available data (e.g., voting records) and used to uniquely identify 87% of the population of the United States [165]. Low-resolution maps of patient locations pose an additional risk to individual privacy—allowing considerably more precision in re-identification than might be expected. Although the Health Insurance Portability and Accountability Act Privacy Rule (Section 164.514) does not explicitly address the publication of such maps, certain formats of geographic data display most likely violate the spirit of that rule.

Curtis et al have also recently described a method to re-identify patients from published maps through manual outlining of case markers [158]. Though the vector-based approach of heads-up digitizing can be more accurate than raster-based unsupervised classification in certain circumstances, in this case, it may be difficult to find the true border of the case markers from a scanned paper-based maps (such as the newspaper article described by Curtis et al) or even low-resolution digital images. If the marker is not digitized accurately, then it follows that the centroid of this polygon will also less accurately reflect the original geocoded location. Our approach differs from the manual approach in that we rely on analyzing the spectral properties of the map image through unsupervised classification to automatically identify patient locations. The raster-based method based on the spectral properties of the image can provide a reliable means of recreating the original vector file and systematically obtaining the center point of a low-resolution marker. This comparison, however, warrants further evaluation. Nonetheless, the results of the two papers are very similar in that they show that maps containing point data are vulnerable to patient address re-identification. These studies and our previous publication on this topic [122] should be viewed together informing policy around the display of geographic data.

The main question that should be asked by both authors and editors is what are the benefits and risks of point localization of patients? Is it necessary to publish maps of point locations, for the presentation of relevant results of research or are they presented merely for illustrative purposes? The answer to these questions should guide decisions on how to report disease maps [79]. If just for illustrative purposes, there are techniques available to visualize spatial data without revealing patient information [145]. For instance, a common approach to de-identifying such data has been to use ZIP or postal code rather than home address to protect anonymity. While usually appropriate for the reporting of study results, aggregation of data to an administrative unit poses constraints on the analysis and visualization of disease patterns [126, 145, 166]. Other approaches are available for masking geographic data, such as spatial masking of cases by randomly relocating cases within a given distance of their true location [87, 128, 167, 168] or the population-density adjusted 2D Gaussian blurring approach which results in only a small reduction in sensitivity to detect clustering patterns [129]. These methods avoid these visualization constraints of data aggregation and afford sufficient privacy for publication without substantial loss to visual display. Masking methods provide more systematic and reliable means of de-identification rather than simply reducing map resolution. Spruill developed a measure of privacy protection for any mask, analogous to our measure of number of addresses within which the patient could reside [142]. Such a measure could be used by journal editors as a rule for not publishing maps of individual cases unless a certain value of anonymity was attained. This measure, often referred to as K-anonymity, could help to establish guidelines for the safe publication of disease maps [82, 129].

Our approach relies on simulation, rather than attempting to re-identify patients from published maps. We chose this approach to avoid propagating any prior inadvertent disclosures of patient identity, and to avoid impugning particular authors or journals. An advantage of our approach is that because we know the value of the original plotted location, we can precisely measure the accuracy of re-identification. Our analysis also does not address the geocoding method. Accuracy of re-identification will also be dependent on the method for geocoding patient address. Use of a global positioning system (GPS) will provide greater accuracy then that of an address geocoder (automatic conversion from home address text to latitude and longitude using interpolation along street line data). When a geocoder is applied, the input data source will affect the accuracy of the estimate address coordinate. Many US-based studies rely on the freely available US Census TIGER line file as input to assign coordinates to addresses. Although TIGER line files differ in accuracy across the US, they rarely, if ever, approach the geometric accuracy of

GPS coordinates or even more detailed commercial data sets. In fact, geocoding based on the free Census data available to most health researchers increases patient anonymity as the proportional placement of the address location can greatly affect geocoding accuracy [161, 169]. Outside the US, street level data may not be available for address geocoding. Therefore, spatial analysis studies in these areas would rely on the more accurate GPS measures. By extension, greater positional accuracy is revealed in these studies. Our findings may therefore be highly pertinent for GIS-based studies in developing countries.

The issues we raise here have, of course, much wider implications than for just health data, including crime data, housing data (e.g.: Section 8 units, shelters for abused women, etc.), and other administrative data sets [128, 170, 171]. New spatial data standards that protect confidentiality while still effectively communicating information about spatial patterns require immediate evaluation [172].

Conclusions

The publication of low-resolution disease maps poses an inherent jeopardy to patient privacy. Because the appropriate use of the patient address level data can bring real benefit to many areas of public health research that deal with spatial analysis, accidental disclosure of patient information from such maps may lead to constraints on obtaining geographically referenced health data. Thus, guidelines for the display or publication of health data are needed to guarantee privacy protection. Further, the editors of journals and textbooks should consider implementing policies to ensure the safe reporting of spatial data.

Chapter V: Future Directions and Conclusions

There are many areas that are of interest to pursue based on the findings in this thesis.

Disclosure Control Mechanisms that Incorporate Quantitative Estimates

Finding closed-form solutions that adequately quantify the amount of information in different patient genomic data sets will help add the necessary clarity to patient decision making. In this thesis, work is described that helps quantify how readily individuals can be identified or re-identified using certain types of demographic and genomic data, leading to a discussion of whether a data set is suitable for public disclosure for research purposes.

For research purposes, it is conceivable that a subject would want to limit her identifiable genomic or demographic disclosures to a subset of SNPs, or to a reasonable, blurred resolution, that would not be directly re-identifiable. Providing subjects with different levels of genomic and demographic anonymity based on their protected health data, along with the probability of re-identification for each of those anonymity levels, will allow patients to select a comfortable level of altruistic sharing [105].

Additionally, it may be beneficial for researchers to request a set of SNP values from a patient's medical record that contains a list of requested and required SNPs in order to participate in a research study. The patient could then use a utility—a "risk engine" to reduce the amount of uniquely identifying information by first removing the most personally identifying requested SNPs while maintaining the required ones.

Using the four risk models we have earlier identified, this risk engine would provide internal answers about the information content and identifiability of genomic medical record data. With such complex data, these user interfaces would need to clearly display these clinical and research-based scenarios for a variety of skill and education levels. Certainly, these models can be mathematically and biologically complex, extending beyond the reach of even well-educated individuals, so a layered approach where clear but simple information is available for each of these clinical or logical scenarios as needed in a patient interaction, but increasingly complex and complete information could be made available if a user requested it.

Information Theoretic Approaches and Multi-Locus Measures

In this thesis, we have evaluated risk of disclosure thoroughly, as well as the implicit information that can be derived from family members; however, we have not conducted a traditional information theoretic analysis for genomic data types. Risk modeling for genomic privacy requires understanding of not only what is unique in a data set but also how informative a patient's genotypic values are at common loci of variation (SNPs in our studies). SNP linkage disequilibrium dependency is an important component of this, as well as population-specific SNP frequency values. For mutation data, it will also be important to distinguish any sequence data that is available that is distinct from a reference sequence.

The further study of genomic information content may include information theoretic approaches. Entropy describes how many bits are needed, on average, to encode a sequence of values based on the frequencies of those values. The information theoretic measure of Shannon's entropy, *H(X)*, defined for a random variable *X*, and *p(x)*, the probability that random variable *X* takes the value at a given x,

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

can be used to help quantify the information content in a set of multiple SNP loci in linkage disequilibrium [70]. Shannon's entropy has also proven useful in finding the edges of haplotype blocks in sets of SNPs, which is a similar and important SNP problem, as it pertains to measuring how informative specific SNPs are with respect to one another in a data set. For example, a multi-locus measure of linkage disequilibrium was created for calculating Shannon's entropy for a set of SNP haplotypes, where the value of each $p_i(x)$ used to calculate *H(X)*, was the population frequency of each possible haplotype sequence, *i*. Research effort should be spent to investigate the efficacy of information theoretic approaches including Shannon's joint entropy measure as well as the information distance for both genomic medicine and privacy purposes.

It will be important to incorporate the population-specific frequency values of the patient's sequenced set of *n* SNPs into the model, as well as the (*n**(*n*-1))/2 possible linkage disequilibrium values, to more accurately ascertain the information content in that data set.

This may be approached from two angles:

(1) explore the use of information theory findings such as fast measures of Shannon's Joint Entropy that compare the amount of information in the joint distribution of the linkage disequilibrium values of the SNPs among one another. This will involve calculating how informative a set of SNPs is using the joint entropy measure, *H(X,Y)* where random variables X and Y are jointly distributed according to the probability distribution *p(X,Y)*

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

We calculate the average minimum number of bits needed to encode a joint distribution of random variables *X* and *Y*. Instead of calculating the average minimum number of bits, or questions one would have to ask to re-identify a given probability distribution, we instead pose our calculation as how informative a set of SNPs are, on average, by measuring how many 'effective SNPs'—the number of perceived independent SNPs—there are, given the overlapping, joint distribution of the SNP linkage disequilibrium values. Because the SNPs in a data set selected for sharing or publishing will likely be related to one another, they are likely to be dependent, or in linkage disequilibrium with one another. Accurate privacy and re-identification models must include this 'discount factor' on how informative a specific set of SNPs is, reducing the information from a set of *n*, somewhat mutually dependent SNPs to a number *m,* smaller than *n*.

Here we have approached the problem of how to appropriately discount the measure of how informative a set of SNPs is by placing the *n choose 2* mutual SNP linkage disequilibrium r^2 values into a normalized *p(x, y)* distribution and then calculating *H(X,Y)*. Genomic privacy decision making systems can then incorporate this discount into estimates for how revealing a set of SNPs is for a patient, allowing the patient to share the right number of SNPs (depending on what population she comes from), what SNPs she would like to share, and what her sequenced values are at those SNP loci.

(2) use linear algebra and graph theoretic approaches to address the SNP 'information overlap' issue through network nodes that represent SNPs and edges to those network graphs that represent the amount of overlap between those two SNPs in the form of a linkage disequilibrium value. The goal of this project would be to quantify the total overlap in moderately-sized, shareable sets of SNPs, in a more computationally efficient manner, as it is important to consider that much privacy analysis for records may be conducted on web application servers.

We may begin by exploring the use of matrix composition

$$[L]^T W[L]\vec{x} = \vec{f}$$

with *L* as the triangular multi-locus linkage disequilibrium matrix of r^2 values, *W* a diagonal weighting matrix perhaps to describe clinical importance to a set of SNPs in question, *x* a selection vector, and *f,* a solution vector, which would hold multi-locus linkage scores. There are several other related techniques that we will explore to inform our approach. These techniques will be borrowed from the disciplines of graph theory and linear algebra [173]. Specifically, we may study problem variants of final weighting problems for graph systems, minimizing cost in graph networks, among others. It is possible to explore similar mathematic problems in this domain that can be applied to information theory analysis for large biological data sets.

To evaluate whether these risk models and disclosure risk estimates are appropriate, one could test randomly-selected sets of SNPs from contiguous segments that are derived from the HapMap study in the thoroughly mapped HapMap regions. Then, it is possible to compare whether estimates from these test data sets differ appreciably from previous information theoretic estimates using Shannon's univariate entropy, *H(X)*, as described earlier,

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

as well as estimates based on multi-locus LD metrics. Those sets of multiple SNP loci with higher multi-locus LD values should be less informative, so investigating whether this is the trend, on average, for patients in a study would be useful. This would also be useful in investigating the boundaries that risk models place on disclosure control, including whether they are appropriate for clinical use, by evaluating whether common genetic screening analyses that use SNPs in contiguously sequenced HapMap segments could be conducted in an anonymous fashion for the sequenced patients.

Geographical Anonymization and Privacy

The transmission of geographic protected health information will continue to expand, whether through the use of integrated healthcare information networks or via AHIC standards, including consumer empowerment, chronic care, biosurveillance and EMR messaging scenarios [174]. The efficacy of disease surveillance efforts and identification of disease correlates that could improve care rely on precise patient locations to enhance spatial clustering performance, and are better than aggregate data [126]. The methodologies that have been discussed in this thesis for spatially anonymizing geographic data using underlying demographic factors can be extended with development of new methods.

One new area of research in this field would be to create a new class of anonymization algorithms for spatial data that preserve complex geographical and demographic relationships among cases. Enhancing existing anonymization algorithms to allow for common geographic and demographic constraints, and developing methodologies to manage a wider variety of spatial disease patterns, it would be possible to provide greater privacy for different patterns of patient data. This might also allow the transmission of selected pieces of additional epidemiological and demographic data while balancing identifiability. For example, aggregate descriptive data might be made available by census regions while point data are separately made available, e.g. percentage of age 65+, gender male, moderate income, etc. by region before and after anonymization.

Certainly another important component that will further adoption of anonymization algorithms is a thorough evaluation of efficacy; these algorithms should be tested using large surveillance data sets, measuring the degree of re-identifiability of individuals and the impact of anonymization on cluster detection. To advance the adoption and widespread use, it will be necessary to evaluate and validate use for a diverse array of patient data to ascertain the limits of robustness of different anonymization technologies. This might include a team that is dedicated to external review of algorithms that attempts to recover personally identifying information or a public contest.

Anonymization Type Standards and Meta-Data

The level of anonymity in data shared in or released from health networks should be determined by a number of potentially diverse interests:

- the utility afforded by sharing authentic data, for public health or clinical use
- the rights and wishes of the patients whose data would be shared
- the interests and policies of the hospitals sharing protected data on the network

To enable such granularity of authorization, type and messaging meta-data standards should be developed that will allow a health network to respond to requests, maintain records of who has requested which data, and maintain linking identifiers to recover authentic data. Supporting such patient and hospital-specifiable levels of anonymity requires that health data networks store several different levels of anonymized data values so that user and hospital specifications can be changed as appropriate.

Other meta-data that may be stored about anonymized points would include the anonymizing algorithm as well as the demographic composite of the data and the level of anonymity achieved.

Availability of Anonymization Modules

To increase the use of privacy preserving algorithms, the public availability of these algorithms and standards is necessary. This involves making available a set of the core anonymizing algorithms and sample implementations publicly available as open source projects that can be downloaded and rapidly integrated into other programs. All of the algorithms should require low storage and processing time overhead and include advanced features including authentic data randomized linkers.

Development of a cryptographically secured anonymization web service

We believe it is worthwhile to investigate whether a publicly available, cryptographically secured anonymization web service would adequately serve the needs of small hospitals to anonymize their patient data. Small hospitals may not have the capability to adequately join a health data network. A web service could automate the process of both anonymizing patient data while simultaneously pushing that anonymized data onto the health network. Of course, if such a service were offered commercially (not through one institutional research provider) it would need to enter into a HIPAA-defined business relationship with the covered entities, because this service would be handling highly identifying PHI while anonymizing it.

Improve anonymization toolkits to include secured upload, integrated geocoding, and interchangeable algorithm type

To foster anonymization efforts, a publicly-available, open source client has been created that allows users to anonymize data sets at their site producing flexible output data sets containing only anonymized data [175]. The anonymization algorithm built into the client is a census block group population-density adjusted 2D Gaussian randomized skew. The anonymization client currently allows users to select an average data set k-anonymity level as well as a basement filtering threshold that removes records that fail to meet a specified minimum value. Enhancements to this tool could include several additional features:

- secured upload to health networks using the aforementioned standards
- integrated, user-specifiable geocoding service, such that users can select a locally networked service such as ESRI or a Geo-Coder-US installation
- interchangeable anonymization algorithm type, such that anonymization algorithms that implement a standardized interface can be selected for use, including versioning and hash-lookup for veracity of algorithm
- storage of local randomized linking identifier data for authentic data lookup

- enhanced HL7 data support for data stream processing
- visualizations and anonymization statistics that enable users to better understand the anonymization that has taken place

Describing quantitative anonymity estimates to users and explaining how to set exclusion criteria from transmissions

Effort should be dedicated to clearly explaining anonymity estimates and algorithm abilities to users of all anonymization systems. This will be an especially difficult task for users of code components that can be integrated within existing infrastructure, as there is often no user interface for meaningful feedback.

Constrained anonymization techniques

The collection of spatial health data is generally conducted with the purpose of discovering geographical clustering or identifying a correlate, predictor, or other association between geographical locations and a specific finding or disease. This is precisely why the precision of geographical data is valuable in these studies, and why clustering in data should potentially be protected in an anonymization process. Some spatial anonymization algorithms have shown that it is possible to blur geographic patient data in a way that increases privacy while retaining the ability to identify clustering patterns among cases. Such systems have encountered difficulty in certain conditions, such as with small cluster sizes or high underlying background noise. We have taken a set of artificially generated clusters that would ordinarily be difficult to detect after anonymization and detected those clusters before anonymization, and then used clustering data to inform the anonymization process to maintain a clustering distribution among the involved cases.

Here, we would alter the anonymization of the subset of points identified in a cluster, separately anonymizing the geocodes in the cluster by replacing each address with another randomly placed address within a circle that will provide the desired output k-anonymity.

Mutli-Factor Authentication using Contents from Disparate EHRs

The development of integrated systems to provide patient-centric health records requires granting access to medical records that have been created at a set of disparate institutions. Consider the case where a patient has decided to create a comprehensive health history that includes visits and lab results from previous providers that no longer have an operational relationship with the patient. This certainly poses a complex authentication problem if these data are to be made available to patients electronically.

We propose a multi-factor authentication framework that allows a trusted intermediate authority to use the contents of potentially matching medical records to generate secondary authentication questions and to manage authorization of appropriate access. This helps to disambiguate between similar records from different patients, as well as ensure that the patient is who she purports to be.

There are several methods to directly authenticate a user to a medical document. One is the use a shared secret key such as a previously assigned or agreed password, PIN or biometric hash. This could be created and recorded by the user at point of care or via provider portal. Another method might provide a patient ticket that links her to a records (or location containing subsequent records.) Such a ticket could be provided at a clinical visit, or afterward via email. These methods rely on previously authenticating users (either in person or through email).

Oftentimes, a user may not be able to authenticate in person or electronically. To create a complete health record, it is desirable to integrate clinical data from all institutional relationships, including those that are terminated, or no longer in physical proximity. Additionally, the time and effort required to link to disparate data feeds for patients may not be reasonable, as it is dependent on the number and types of points of care. A suitable alternative that removes the direct authentication burden from patients and providers would solve this issue.

Secondary authentication questions could solve this problem well, though they bring new requirements. One is standardized, structured data so that useful polling information can be gathered. The information value of answering such questions also requires knowledge of the probability of candidate choice events, for example, a visit to a clinic for respiratory infection may be very common, so if a patient selects that choice, it is not very 'informative'.

We propose a mutli-factor authentication process similar to those now required in finance [176] to authenticate users with two distinct layers (Figure 1). First, a user accesses her health record portal, and provides identifiers and demographic data. The portal makes a request on her behalf to access health records from disparate institutions by polling a record locator service with her SSN or MRN. The record locator service then polls all participating hospitals to gather a set of candidate records (through a trusted, secure relationship). The record locator service then generates secondary authentication questions which are presented to the user by the record portal. The patient provides answers through the portal, which are transmitted to the record locator service. If the answers are correct, the record locator service will authorize access to a set of records. This service would certainly need to be part of a trusted authority, regardless of whether it is a business or government entity.

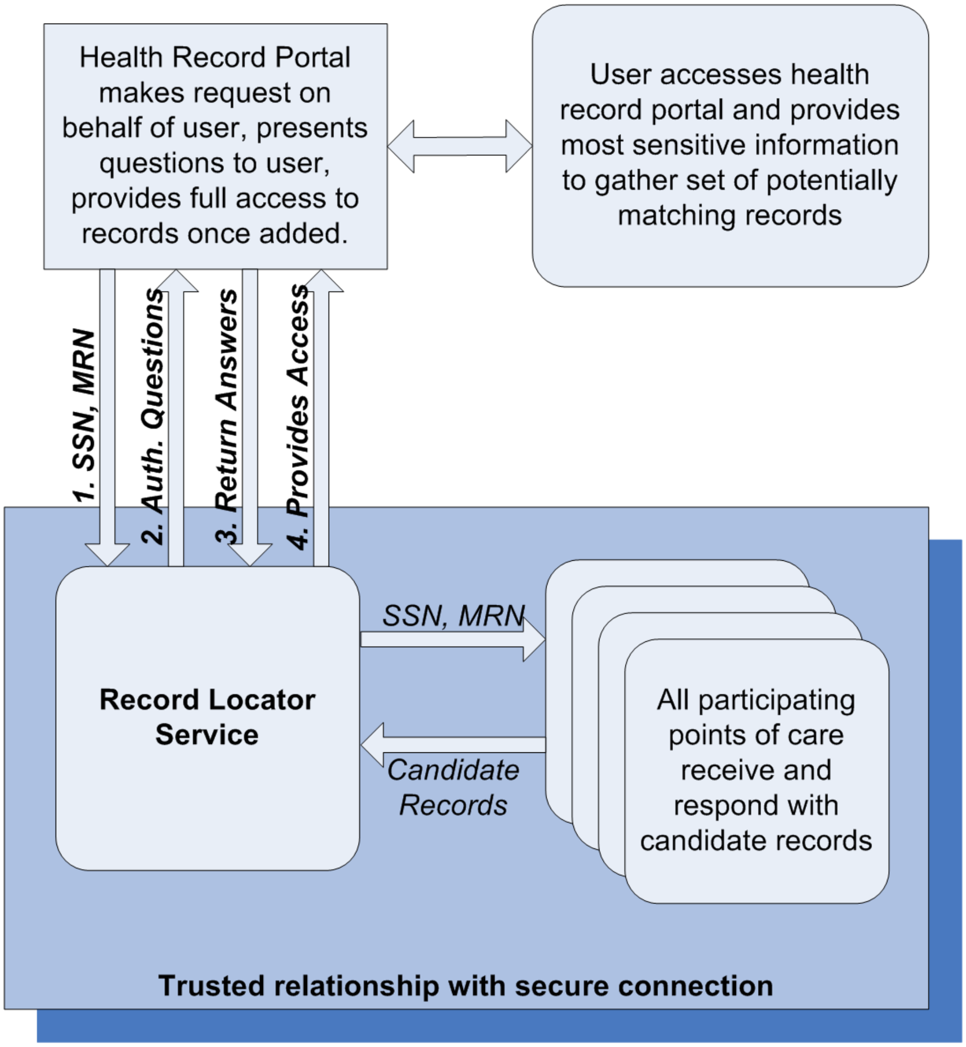

Health Record Portal makes request on behalf of user, presents questions to user, provides full access to records once added.
User accesses health record portal and provides most sensitive information to gather set of potentially matching records
1. SSN, MRN
2. Auth. Questions
3. Return Answers
4. Provides Access
Record Locator Service
SSN, MRN
Candidate Records
All participating points of care receive and respond with candidate records
Trusted relationship with secure connection

Figure 34: A proposed mutli-factor authentication framework for the retrieval of patient medical records from a set of disparate points of care. In this example, a user begins by accessing a health record portal, and over a secure network connection, provides highly identifying information to his or her trusted portal provider. The health record portal then makes a network query to a record locator service, using that highly identifying data, such as a social security number, date of birth, gender, home address, other medical record numbers. The record locator service then uses the matching and potentially matching set of medical records to generate a set of authentication questions for the user. Once a sufficient number of those questions have been answered correctly, authorization is provided to access that individual's records.

A mutli-factor authentication process could help solve the complex process of authorizing a user to a set of correct records from all previous points of care.

Conclusion

The amount of electronically transmitted protected health information will surely increase in the coming years, and scenarios where patients might share their demographic, clinical and genomic data with researchers, specialists and public health practitioners will become increasingly common. Some of these transmissions will inevitably reveal highly identifying information unless techniques to protect the privacy of individually identifying genomic and clinical data advance dramatically. Because of this, we believe that the quantitative modeling we offer in this thesis will enable patients to make more informed decisions, fully able to consider the implications of their PHI disclosures on themselves and family members. The personalized medicine research movement promises to advance medical practice and science, but will certainly require many micro (personal choices by individuals) and macro (legislation and broad health information protections) balances between the costs and benefits of the new complex data that will be shared with investigators.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. **HapMap Sample Populations** [http://hapmap.org/hapmappopulations.html.en]
3. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
4. **Genome-Wide Association Studies** [http://www.genome.gov/20019523]
5. Benjamin EJ, Dupuis J, Larson MG, Lunetta KL, Booth SL, Govindaraju DR, Kathiresan S, Keaney JF, Jr., Keyes MJ, Lin JP, et al: **Genome-wide association with select biomarker traits in the Framingham Heart Study.** *BMC Med Genet* 2007, **8 Suppl 1:**S11.
6. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nat Rev Genet* 2005, **6:**95-108.
7. Morton NE: **Into the post-HapMap era.** *Adv Genet* 2008, **60:**727-742.
8. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, et al: **Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base.** *Pharmacogenomics J* 2001, **1:**167-170.
9. Watson JD, Cook-Deegan RM: **The human genome project and international health.** *JAMA* 1990, **263:**3322-3324.
10. Collins F, Galas D: **A new five-year plan for the U.S. Human Genome Project.** *Science* 1993, **262:**43-46.
11. Baird PA: **Genetics and health care: a paradigm shift.** *Perspect Biol Med* 1990, **33:**203-213.
12. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309:**1728-1732.
13. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals.** *Nat Rev Genet* 2004, **5:**335-344.
14. **Beyond genetic testing: Personal DNA analysis and research for health, family, ancestry, and genealogy** [https://www.23andme.com/]
15. **deCODEme - Empowering prevention. Calculate genetic risk for diseases, DNA research for personal and family health and ancestry** [http://www.decodeme.com/]
16. **Navigencis - Your genes offer a road map to optimal health.** [http://www.navigenics.com/]
17. Church GM: **The personal genome project.** *Mol Syst Biol* 2005, **1:**2005 0030.
18. **This time it's personal.** *Nature* 2008, **453:**697.
19. ten Kate LP: **Editorial.** *Community Genet* 2001, **4:**1.
20. Wideroff L, Freedman AN, Olson L, Klabunde CN, Davis W, Srinath KP, Croyle RT, Ballard-Barbash R: **Physician use of genetic testing for cancer susceptibility: results of a national survey.** *Cancer Epidemiol Biomarkers Prev* 2003, **12:**295-303.

21. ten Kate LP: **Carrier screening for cystic fibrosis and other autosomal recessive diseases.** *Am J Hum Genet* 1990, **47:**359-361.

22. ten Kate LP, Tijmstra T: **Carrier screening for Tay-Sachs disease and cystic fibrosis.** *Lancet* 1990, **335:**1527-1528.

23. Juengst ET: **Can enhancement be distinguished from prevention in genetic medicine?** *J Med Philos* 1997, **22:**125-142.

24. McGee G, Juengst ET: **Genetic enhancement.** *Med Ethics* 1999:6-7.

25. Juengst ET, Binstock RH, Mehlman MJ, Post SG: **Aging. Antiaging research and the need for public dialogue.** *Science* 2003, **299:**1323.

26. Holden C: **Genetic discrimination. Long-awaited genetic nondiscrimination bill headed for easy passage.** *Science* 2007, **316:**676.

27. Hudson KL, Holohan MK, Collins FS: **Keeping pace with the times--the Genetic Information Nondiscrimination Act of 2008.** *N Engl J Med* 2008, **358:**2661-2663.

28. Bieber FR, Brenner CH, Lazer D: **Human genetics. Finding criminals through DNA of their relatives.** *Science* 2006, **312:**1315-1316.

29. Bieber FR, Lazer D: **Guilt by association: should the law be able to use one person's DNA to carry out surveillance on their family? Not without a public debate.** *New Sci* 2004, **184:**20.

30. **Freedom of Information Act** *5 USC 552* 1996.

31. Kohut K, Manno M, Gallinger S, Esplen MJ: **Should healthcare providers have a duty to warn family members of individuals with an HNPCC-causing mutation? A survey of patients from the Ontario Familial Colon Cancer Registry.** *J Med Genet* 2007, **44:**404-407.

32. Offit K, Groeger E, Turner S, Wadsworth EA, Weiser MA: **The "duty to warn" a patient's family members about hereditary disease risks.** *Jama* 2004, **292:**1469-1473.

33. Yoon PW, Chen B, Faucett A, Clyne M, Gwinn M, Lubin IM, Burke W, Khoury MJ: **Public health impact of genetic tests at the end of the 20th century.** *Genet Med* 2001, **3:**405-410.

34. Hall WD, Morley KI, Lucke JC: **The prediction of disease risk in genomic medicine.** *EMBO Rep* 2004, **5 Spec No:**S22-26.

35. Lin K, Lipsitz R, Miller T, Janakiraman S: **Benefits and harms of prostate-specific antigen screening for prostate cancer: an evidence update for the U.S. Preventive Services Task Force.** *Ann Intern Med* 2008, **149:**192-199.

36. **Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement.** *Ann Intern Med* 2008, **149:**185-191.

37. **Summaries for patients. Screening for prostate cancer with prostate-specific antigen testing: U.S. Preventive Services Task Force recommendations.** *Ann Intern Med* 2008, **149:**I37.

38. Kruer M: **The incidentalome.** *JAMA* 2006, **296:**2801; author reply 2801-2802.

39. Wolf SM, Kahn JP, Lawrenz FP, Nelson CA: **The incidentalome.** *JAMA* 2006, **296:**2800-2801; author reply 2801-2802.

40. Kohane IS, Masys DR, Altman RB: **The incidentalome: a threat to genomic medicine.** *JAMA* 2006, **296:**212-215.

41. Zondervan KT, Cardon LR: **The complex interplay among factors that influence allelic association.** *Nat Rev Genet* 2004, **5:**89-100.
42. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies.** *Genet Med* 2002, **4:**45-61.
43. Colhoun HM, McKeigue PM, Davey Smith G: **Problems of reporting genetic associations with complex outcomes.** *Lancet* 2003, **361:**865-872.
44. Cappuccio FP, Oakeshott P, Strazzullo P, Kerry SM: **Application of Framingham risk estimates to ethnic minorities in United Kingdom and implications for primary prevention of heart disease in general practice: cross sectional population based study.** *Bmj* 2002, **325:**1271.
45. Colditz GA, Coakley E: **Weight, weight gain, activity, and major illnesses: the Nurses' Health Study.** *Int J Sports Med* 1997, **18 Suppl 3:**S162-170.
46. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, Ruidavets JB, Haas B, Yarnell J, Bingham A, Amouyel P, Dallongeville J: **Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study.** *Eur Heart J* 2003, **24:**1903-1911.
47. Halamka JD, Mandl KD, Tang PC: **Early experiences with personal health records.** *J Am Med Inform Assoc* 2008, **15:**1-7.
48. Mandl KD, Simons WW, Crawford WC, Abbett JM: **Indivo: a personally controlled health record for health information exchange and communication.** *BMC Med Inform Decis Mak* 2007, **7:**25.
49. Simons WW, Halamka JD, Kohane IS, Nigrin D, Finstein N, Mandl KD: **Integration of the personally controlled electronic medical record into regional inter-regional data exchanges: a national demonstration.** *AMIA Annu Symp Proc* 2006:1099.
50. Simons WW, Mandl KD, Kohane IS: **The PING personally controlled electronic medical record system: technical architecture.** *J Am Med Inform Assoc* 2005, **12:**47-54.
51. Riva A, Mandl KD, Oh DH, Nigrin DJ, Butte A, Szolovits P, Kohane IS: **The personal internetworked notary and guardian.** *Int J Med Inform* 2001, **62:**27-40.
52. Weingart SN, Rind D, Tofias Z, Sands DZ: **Who uses the patient internet portal? The PatientSite experience.** *J Am Med Inform Assoc* 2006, **13:**91-95.
53. Kohane IS, Mandl KD, Taylor PL, Holm IA, Nigrin DJ, Kunkel LM: **Medicine. Reestablishing the researcher-patient compact.** *Science* 2007, **316:**836-837.
54. McMurry AJ, Gilbert CA, Reis BY, Chueh HC, Kohane IS, Mandl KD: **A self-scaling, distributed information architecture for public health, research, and clinical care.** *J Am Med Inform Assoc* 2007, **14:**527-533.
55. Mandl KD, Kohane IS: **Tectonic shifts in the health information economy.** *N Engl J Med* 2008, **358:**1732-1737.
56. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449:**851-861.
57. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, et al: **Blocks of limited haplotype diversity**

revealed by high-resolution scanning of human chromosome 21. *Science* 2001, **294:**1719-1723.

58. Malin BA, Sweeney LA: **Inferring genotype from clinical phenotype through a knowledge based algorithm.** *Pac Symp Biocomput* 2002:41-52.

59. Lin Z, Owen AB, Altman RB: **Genetics. Genomic research and human subject privacy.** *Science* 2004, **305:**183.

60. Henneman L, Timmermans DR, van der Wal G: **Public experiences, knowledge and expectations about medical genetics and the use of genetic information.** *Community Genet* 2004, **7:**33-43.

61. Levitt DM: **Let the consumer decide? The regulation of commercial genetic testing.** *J Med Ethics* 2001, **27:**398-403.

62. Miller SM, Fleisher L, Roussi P, Buzaglo JS, Schnoll R, Slater E, Raysor S, Popa-Mabe M: **Facilitating informed decision making about breast cancer risk and genetic counseling among women calling the NCI's Cancer Information Service.** *J Health Commun* 2005, **10 Suppl 1:**119-136.

63. Mouchawar J, Hensley-Alford S, Laurion S, Ellis J, Kulchak-Rahm A, Finucane ML, Meenan R, Axell L, Pollack R, Ritzwoller D: **Impact of direct-to-consumer advertising for hereditary breast cancer testing on genetic services at a managed care organization: a naturally-occurring experiment.** *Genet Med* 2005, **7:**191-197.

64. Mouchawar J, Laurion S, Ritzwoller DP, Ellis J, Kulchak-Rahm A, Hensley-Alford S: **Assessing controversial direct-to-consumer advertising for hereditary breast cancer testing: reactions from women and their physicians in a managed care organization.** *Am J Manag Care* 2005, **11:**601-608.

65. Malin BA: **An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future.** *J Am Med Inform Assoc* 2005, **12:**28-34.

66. Malin B, Sweeney L: **How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems.** *J Biomed Inform* 2004, **37:**179-192.

67. Lin Z, Hewett M, Altman RB: **Using binning to maintain confidentiality of medical data.** *Proc AMIA Symp* 2002:454-458.

68. **Facts About Genome Sequencing** [http://www.ornl.gov/sci/techresources/Human_Genome/faq/seqfacts.shtml#whose]

69. Malin BA: **Protecting genomic sequence anonymity with generalization lattices.** *Methods Inf Med* 2005, **44:**687-692.

70. Nothnagel M, Furst R, Rohde K: **Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks.** *Hum Hered* 2002, **54:**186-198.

71. Croner CM, Sperling J, Broome FR: **Geographic information systems (GIS): new perspectives in understanding human health and environmental relationships.** *Stat Med* 1996, **15:**1961-1977.

72. Pickle LW, Waller LA, Lawson AB: **Current practices in cancer spatial data analysis: a call for guidance.** *Int J Health Geogr* 2005, **4:**3.

73. Reis BY, Kirby C, Hadden LE, Olson K, McMurry AJ, Daniel JB, Mandl KD: **AEGIS: a robust and scalable real-time public health surveillance system.** *J Am Med Inform Assoc* 2007, **14:**581-588.

74. Bradley CA, Rolka H, Walker D, Loonsk J: **BioSense: implementation of a National Early Event Detection and Situational Awareness System.** *MMWR Morb Mortal Wkly Rep* 2005, **54 Suppl:**11-19.

75. Brownstein JS, Freifeld CC, Reis BY, Mandl KD: **Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project.** *PLoS Med* 2008, **5:**e151.

76. Freifeld CC, Mandl KD, Reis BY, Brownstein JS: **HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports.** *J Am Med Inform Assoc* 2008, **15:**150-157.

77. Brownstein JS, Freifeld CC: **HealthMap: the development of automated real-time internet surveillance for epidemic intelligence.** *Euro Surveill* 2007, **12:**E071129 071125.

78. **EpiSPIDER** [http://www.epispider.org/]

79. Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL: **Geocoding in cancer research: a review.** *Am J Prev Med* 2006, **30:**S16-24.

80. Cassa CA, Grannis SJ, Overhage JM, Mandl KD: **A Novel, Context-Sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection.** *J Am Med Inform Assoc* 2005.

81. Sweeney L: **Uniqueness of Simple 'Demographics in the U.S. Population, LIDAP-WP4.** In *Forthcoming book entitled, The Identifiability of Data.* 2000

82. Sweeney L: **k-anonymity: A model for protecting privacy.** *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* 2002, **10:**557-570.

83. Sweeney L: **Replacing Personally-Identifying Information in Medical Records, the Scrub System.** *Proc AMIA Annu Fall Symp* 1996:333-337.

84. Uzuner O, Luo Y, Szolovits P: **Evaluating the state-of-the-art in automatic de-identification.** *J Am Med Inform Assoc* 2007, **14:**550-563.

85. Uzuner O, Sibanda TC, Luo Y, Szolovits P: **A de-identifier for medical discharge summaries.** *Artif Intell Med* 2008, **42:**13-35.

86. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD: **Automated De-Identification of Free-Text Medical Records.** *BMC Med Inform Decis Mak* 2008, **8:**32.

87. Ohno-Machado L, Silviera SP, Vinterbo S: **Protecting patient privacy by quantifiable control of disclosures in disseminated databases.** *Int J Med Inform* 2004, **73:**599-606.

88. Vinterbo SA, Ohno-Machado L, Dreiseitl S: **Hiding information by cell suppression.** *Proc AMIA Symp* 2001:726-730.

89. Ohno-Machado L SP, Vinterbo S.: **Protecting patient privacy by quantifiable control of disclosures in disseminated databases.** *Int J Med Inform* 2004, **73:**599-606.

90. Armstrong MP RG, Zimmerman DL.: **Geographically masking health data to preserve confidentiality.** *Statistics in Medicine* 1999, **18:**497-525.

91. Cassa CA, Schmidt BW, Kohane IS, Mandl KD: **My sister's keeper?: genomic research and the identifiability of siblings.** *BMC Med Genomics* 2008, **1:**32.

92. Adida B, Kohane IS: **GenePING: secure, scalable management of personal genomic data.** *BMC Genomics* 2006, **7:**93.

93. Hoffman MA: **The genome-enabled electronic medical record.** *J Biomed Inform* 2007, **40:**44-46.

94. Kaiser J: **Genomic databases. NIH goes after whole genome in search of disease genes.** *Science* 2006, **311:**933.

95. Thomas DC: **Are we ready for genome-wide association studies?** *Cancer Epidemiol Biomarkers Prev* 2006, **15:**595-598.

96. Lowrance WW, Collins FS: **Ethics. Identifiability in genomic research.** *Science* 2007, **317:**600-602.

97. Brenner CH, Weir BS: **Issues and strategies in the DNA identification of World Trade Center victims.** *Theor Popul Biol* 2003, **63:**173-178.

98. **Violence Against Women and Department of Justice Reauthorization Act of 2005.** *HR 3402; Public Law 109-162* 2005.

99. **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.

100. Olivier M: **A haplotype map of the human genome.** *Physiol Genomics* 2003, **13:**3-9.

101. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH: **Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia.** *Nat Genet* 2005, **37:**435-440.

102. Slaughter LM: **The Genetic Information Nondiscrimination Act: Why Your Personal Genetics are Still Vulnerable to Discrimination.** *Surg Clin North Am* 2008, **88:**723-738.

103. Korobkin R, Rajkumar R: **The Genetic Information Nondiscrimination Act--a half-step toward risk sharing.** *N Engl J Med* 2008, **359:**335-337.

104. Dugan RB, Wiesner GL, Juengst ET, O'Riordan M, Matthews AL, Robin NH: **Duty to warn at-risk relatives for genetic disease: genetic counselors' clinical experience.** *Am J Med Genet C Semin Med Genet* 2003, **119C:**27-34.

105. Kohane IS, Altman RB: **Health-information altruists--a potentially critical resource.** *N Engl J Med* 2005, **353:**2074-2077.

106. **Population Mutation Databases** [http://snpnet.jst.go.jp/link/Population_e.html]

107. **Ethnic & National Variation Databases** [http://www.hgvs.org/dblist/deth.html]

108. Krawczak M, Ball EV, Cooper DN: **Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes.** *Am J Hum Genet* 1998, **63:**474-488.

109. **Tables that summarize the results of a meta-analysis of single base-pair substitutions logged in the Human Gene Mutation Database.** [http://www.hgmd.cf.ac.uk/docs/msajhg1.txt]

110. Olson KL BM, Pagano M, Mandl KD: **Real time spatial cluster detection using interpoint distances among precise patient locations.** *BMC Med Inform Decis Mak* 2005, **5:**19.

111. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW, Project B: **Algorithms for rapid outbreak detection: a research synthesis.** *Journal of Biomedical Informatics* 2005, **38:**99-113.

112. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F: **A space-time permutation scan statistic for disease outbreak detection.** *Plos Medicine* 2005, **2:**216-224.

113. Wieland SC, Brownstein JS, Berger B, Mandl KD: **Density-equalizing Euclidean minimum spanning trees for the detection of all disease cluster shapes.** *Proc Natl Acad Sci U S A* 2007, **104:**9404-9409.

114. Cassa CA OK, Mandl KM: **System To Generate Semisynthetic Data Sets of Outbreak Clusters for Evaluation of Outbreak-Detection Performance.** *MMWR Morb Mortal Wkly Rep* 2004, **53:**231.

115. Kulldorff M, Nagarwalla N: **Spatial Disease Clusters - Detection and Inference.** *Statistics in Medicine* 1995, **14:**799-810.

116. Bureau USC: **Census Block Groups Cartographic Boundary Files Descriptions and Metadata - U.S. Census Bureau.** 2005.

117. Documentation SJ: **Random Class: nextGaussian() Method Documentation.** 2005.

118. Beitel AJ, Olson KL, Reis BY, Mandl KD: **Use of Emergency Department Chief Complaint and Diagnostic Codes for Identifying Respiratory Illness in a Pediatric Population.** *Pediatr Emerg Care* 2004, **20:**355-360.

119. Mandl KD RB, Cassa C.: **Measuring outbreak-detection performance by using controlled feature set simulations.** *MMWR Morb Mortal Wkly Rep* 2004, **53:**130-136.

120. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, Pavlin JA, Gesteland PH, Treadwell T, Koski E, et al: **Implementing syndromic surveillance: A practical guide informed by the early experience.** *Journal of the American Medical Informatics Association* 2004, **11:**141-150.

121. Seaman V: **An inquiry into the cause of the prevalence of the yellow fever in New York.** *The Medical Repository* 1798, **1:**315-372.

122. Brownstein JS, Cassa CA, Mandl KD: **No place to hide--reverse identification of patients from published maps.** *N Engl J Med* 2006, **355:**1741-1742.

123. Moskop JC, Marco CA, Larkin GL, Geiderman JM, Derse AR: **From Hippocrates to HIPAA: privacy and confidentiality in emergency medicine--Part I: conceptual, moral, and legal foundations.** *Ann Emerg Med* 2005, **45:**53-59.

124. **Federal Register.** (Office C ed., vol. 67. pp. 53182-53273: Library of Congress; 2002:53182-53273.

125. Fefferman NH, O'Neil EA, Naumova EN: **Confidentiality and confidence: is data aggregation a means to achieve both?** *J Public Health Policy* 2005, **26:**430-449.

126. Olson KL, Grannis SJ, Mandl KD: **Privacy protection versus cluster detection in spatial epidemiology.** *Am J Public Health* 2006, **96:**2002-2008.

127. Cox LH: **Protecting confidentiality in small population health and environmental statistics.** *Stat Med* 1996, **15:**1895-1905.

128. Armstrong MP, Rushton G, Zimmerman DL: **Geographically masking health data to preserve confidentiality.** *Stat Med* 1999, **18:**497-525.

129. Cassa CA, Grannis SJ, Overhage JM, Mandl KD: **A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection.** *J Am Med Inform Assoc* 2006, **13:**160-165.

130. Machanavajjhala A KD, Abowd J, Gehrke J, Vilhuber L: **Privacy: Theory meets practice on the map.** In *International Conference on Data Engineering (ICDE)* 2008

131. Strayer J: *Linear Programming and Its Applications.* New York: Springer-Verlag; 1989.

132. ILOG I: **ILOG CPLEX 10.010** Gentilly, France; 1999.

133. Kulldorff M, Nagarwalla N: **Spatial disease clusters: detection and inference.** *Stat Med* 1995, **14:**799-810.

134. VanWey LK, Rindfuss RR, Gutmann MP, Entwisle B, Balk DL: **Confidentiality and spatially explicit data: concerns and challenges.** *Proc Natl Acad Sci U S A* 2005, **102:**15337-15342.

135. Salazar-Gonzalez J: **Protecting tables with cell perturbation.** In *Proceedings of the Joint UN-ECE/Eurostat Work Session on Statistical Data Confidentiality*. 2005

136. Duncan G, Fienberg, S **Obtaining information while preserving privacy: a Markov perturbation method for tabular data.** In *Proceedings of the Statistical Data Protection Conference*. 1998: 351-362.

137. Kargupta H, Datta, S, Wang, Q, Sivakumar, K: **Random-data perturbation techniques and privacy-preserving data mining.** *Knowledge and Information Systems* 2005, **7:**387-414.

138. Gutmann M, Stern, PC **Putting people on the map: Protecting confidentiality with linked social-spatial data. Panel on confidentiality issues arising from the integration of remotely sensed and self-identifying data.** In *National Research Council of the National Academies Press* Washington D.C.; 2007.

139. Lakshmanan L, Ng, R, Ramesh, G: **To do or not to do: the dilemma of disclosing anonymized data.** In *Proc ACM SIGMOD Conference*. 2005: 61–72.

140. Aggarwal C, Pei, J, Zhang, B: **On privacy preservation against adversarial data mining.** In *12th ACM SIGKDD Conference*. 2006: 510-516.

141. Ganta S, Acharya, R: **On breaching enterprise data privacy through adversarial information fusion.** In *Proc Workshop on Information Integration Methods, Architecture, and Systems, at the 24th IEEE International Conference on Data Engineering*. 2008: 246-249.

142. Spruill NL: **The confidentiality and analytic usefulness of masked business microdata.** *Proceedings of the American Statistical Association Section on Survey Research Methods* 1983**:**602-607.

143. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, Pavlin JA, Gesteland PH, Treadwell T, Koski E, et al: **Implementing syndromic**

surveillance: a practical guide informed by the early experience. *J Am Med Inform Assoc* 2004, **11:**141-150.

144. Cassa CA GS, Overhage JM, Mandl KD: **A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection.** *J Am Med Inform Assoc* 2006, **13:**160-165.
145. Kamel Boulos MN, Cai Q, Padget JA, Rushton G: **Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses.** *J Biomed Inform* 2006, **39:**160-170.
146. Statistics NCfH: **Health, United States.** pp. p.370, 429, 434; 2006:p.370, 429, 434.
147. Bleichenbacher D: **Chosen ciphertext attacks against protocols based on the RSA encryption standard PKCS #1.** In *Advances in Cryptology*. 1998: 1-12.
148. Cassa CA IK, Olson KL, Mandl KD: **A software tool for creating simulated outbreaks to benchmark surveillance systems.** *BMC Med Inform Decis Mak* 2005, **5:**22-28.
149. **AEGIS-CCT** [http://www.sourceforge.net/projects/chipcluster/]
150. Stein C: **Inadmissibility of the usual estimator for the mean of a multivariate distribution.** In *Proc Third Berkeley Symp Math Statist Prob* 1956: 197-206.
151. Baranchik AJ: **A family of minimax estimators of the mean of a multivariate normal distribution.** *Ann Math Statist* 1970, **41:**642-645.
152. **HHS Awards Contracts to Develop Nationwide Health Information Network** [http://www.hhs.gov/news/press/2005pres/20051110.html]
153. **American Health Information Community (the Community)**
154. Pickle LW: **Spatial analysis of disease.** *Cancer Treat Res* 2002, **113:**113-150.
155. Elliot P, Wakefiled JC, Best NG, Briggs DJ: *Spatial epidemiology: methods and applications.* Oxford: Oxford University Press; 2000.
156. Brownstein JS, Cassa CA, Kohane IS, Mandl KD: **Reverse geocoding: Concerns about patient confidentiality in the display of geospatial health data.** *AMIA Annu Symp Proc* 2005**:**905.
157. Curtis A, Mills JW, Leitner M: **Keeping an eye on privacy issues with geospatial data.** *Nature* 2006, **441:**150.
158. Curtis AJ, Mills JW, Leitner M: **Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina.** *Int J Health Geogr* 2006, **5:**44.
159. Boston Water and Sewer Commission: **Boston planimetric and topographic data.** 1996.
160. US Census Bureau: **Redistricting Census 2000 TIGER/Line Files [machine-readable data files].** Washington, DC; 2000.
161. Cayo MR, Talbot TO: **Positional error in automated geocoding of residential addresses.** *Int J Health Geogr* 2003, **2:**10.
162. ESRI: **ArcGIS.** 8.1 edition. Redlands, CA: Environmental Systems Institute Inc.; 2001.
163. ERDAS: **IMAGINE.** 8.5 edition. Atlanta, GA: ERDAS; 2001.
164. Guha R: *Object Coidentification on the Semantic Web.* IBM Research, Almaden. New York: http://tap.stanford.edu/CoIdent.pdf; 2004.

165. Sweeney L: *Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4.* Pittsburgh, PA: Carnegie Mellon University, Laboratory for International Data Privacy; 2000.

166. Gregorio DI, Dechello LM, Samociuk H, Kulldorff M: **Lumping or splitting: seeking the preferred areal unit for health geography studies.** *Int J Health Geogr* 2005, **4:**6.

167. Leitner M, Curtis A: **Cartographic guidelines for geographically masking the locations of confidential point data** *Cartographic Perspectives* 2004, **49:**22-39.

168. Leitner M, Curtis A: **A First Step Towards a Framework for Presenting the Location of Confidential Point Data on Maps - Results of an Empirical Perceptual Study.** *International Journal of Geographical Information Science* 2006, **20:**797-811.

169. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW: **On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research.** *Am J Public Health* 2001, **91:**1114-1116.

170. Monmonier M: *Spying with Maps: Surveillance Technologies and the Future of Privacy.* Chicago, IL: University of Chicago Press.; 2002.

171. Wartell J, McEwen JT: *Privacy in the Information Age: A Guide for Sharing Crime Maps and Spatial Data.* U.S. Department of Justice, Office of Justice Programs: http://www.ncjrs.org/pdffiles1/nij/188739.pdf; 2001.

172. Pickle LW, Szczur M, Lewis DR, Stinchcomb DG: **The crossroads of GIS and health information: a workshop on developing a research agenda to improve cancer control.** *Int J Health Geogr* 2006, **5:**51.

173. Strang G: *Introduction to Applied Mathematics.* 1 edn: Wellesley Cambridge Press; 1986.

174. **American Health Information Community: Breakthroughs** [http://www.hhs.gov/healthit/breakthroughs.pdf]

175. **Patient Record Anonymization GUI** [http://www.sourceforge.net/projects/patientanon/]

176. **Authentication in an Internet Banking Environment.** [http://www.ffiec.gov/pdf/authentication_guidance.pdf]