Revealing the spatial distribution of a disease while preserving privacy

Shannon C. Wieland[a,b], Christopher A. Cassa[b], Kenneth D. Mandl[b,c,1], and Bonnie Berger[a,d,1]

[a]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307; [b]Children's Hospital Informatics Program at the Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Children's Hospital, Boston, MA 02115; [c]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115; and [d]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139-4307

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved August 19, 2008 (received for review February 1, 2008)

Datasets 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 25–150 times less than aggregation by zip code.

patient privacy | spatial epidemiology | linear programming | data aggregation

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 (1), 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 patients 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; 87% of subjects in one study could be uniquely identified by their gender, zip code and date of birth (2) and low-resolution dot maps of diseases published in several medical journals could be used to trace most patients to single addresses (3).

Although established since the time of Hippocrates (4), 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) (5). 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 10 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 dataset to qualify as non-identifiable. The first specifies that the dataset 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" (5).

The prevailing method for preserving privacy in spatial data is aggregating by predefined 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 (6); for example, the detection of spatial clusters is significantly less sensitive and specific when data are aggregated even by zip code (7). 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 $\frac{1}{1,000}$ 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 $\frac{1}{100}$.

An alternative to aggregation is moving each patient to a new location to ensure privacy (8), formalized by the family of "geographical masks" proposed by Armstrong *et al*. (9). Each is a deterministic or stochastic function of geographical identifiers designed to de-identify patient locations while preserving the approximate spatial distribution of patients. 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 (10). 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 on 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.

Author contributions: S.C.W., C.A.C., K.D.M., and B.B. designed research; S.C.W. performed research; S.C.W. analyzed data; and S.C.W., C.A.C., K.D.M., and B.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: kenneth.mandl@childrens.harvard.edu or bab@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0801021105/DCSupplemental.

In a different application, Machanavajjhala *et al*. (11) ensured a low disclosure risk by generating a de-identified dataset 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 datasets 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 $\mathcal{A}$ of possible original locations must be known in advance; for example, this could be all census block groups in a state or all residential addresses within a city. The actual patient locations to be de-identified must be contained in $\mathcal{A}$. The set $\mathcal{B}$ of possible final locations to which patients may be moved is also defined in advance. This could be a different set than $\mathcal{A}$, such as evenly spaced points on a grid to which patients at exact addresses will be relocated. If $\mathcal{A}$ and $\mathcal{B}$ 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 (12). The decision variables are the transition probabilities P_{ij} of assigning a patient in location $i \in \mathcal{A}$ to a new location $j \in \mathcal{B}$ (see Fig. 1). 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 originates in location $i \in \mathcal{A}$, a new location j is drawn from the set $\mathcal{B}$ 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:

$$0 \le P_{ij} \text{ for all } i \in \mathcal{A} \text{ and } j \in \mathcal{B}. \quad [1]$$

In addition, every case must be moved somewhere, so

$$\sum_j P_{ij} = 1 \text{ for all } i \in \mathcal{A}. \quad [2]$$

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 dataset originated from any specific individual in the underlying population is at most ξ:

$$P_{ij} \cdot \frac{n_i}{N} \le \frac{n_i \cdot \xi}{s} \cdot \sum_{k \in \mathcal{A}} \frac{n_k}{N} \cdot P_{kj} \text{ for all } i \in \mathcal{A} \text{ and } j \in \mathcal{B}. \quad [3]$$

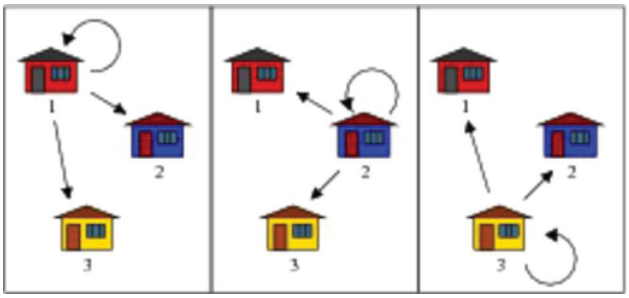

Fig. 1. Schematic of transition probabilities. A patient found at each location in a set $\mathcal{A}$ may transition to any location in a set $\mathcal{B}$. In this example, the sets $\mathcal{A}$ and $\mathcal{B}$ are equivalent for simplicity, each consisting of three locations represented by houses. The nine transition probabilities, represented by arrows, are variables solved by linear programming.

In this equation, ξ is a user-specified privacy bound between zero and one. The parameter $s \ge 1$ is the number of patients in the particular dataset to be de-identified; for example, this could be the number of patients enrolled in a case-control study of a certain disease. The variable n_i is the number of people in region i, and $N = \sum_{r \in \mathcal{A}} n_r$ is the population summed over all possible original locations. For instance, if the regions are census block groups, then the constants $\{n_i\}_{i \in \mathcal{A}}$ 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 and N is the number of possible original addresses. Any randomly or methodically chosen member of the population is guaranteed to belong to the dataset with probability at most ξ. Consequently, given the de-identified list, one could expect to search through at least $\frac{1}{\xi}$ members of the population by any method before encountering one person on the list. Derivation of this constraint is found in the Appendix.

We wish to move patients as little as possible subject to the constraints above. For each $i \in \mathcal{A}$ and $j \in \mathcal{B}$, 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 dataset, a patient originates in region i with probability $\frac{n_i}{N}$. Hence, the expected distance that a patient is moved, which is the objective function to be minimized, is

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

Several standard linear programming techniques to solve LP models, such as that specified by Eqs. **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 N available individual addresses, then no strategy to de-identify $s \le N$ patients by reassigning new addresses can achieve a risk of re-identification below $\frac{s}{N}$. 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. 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.

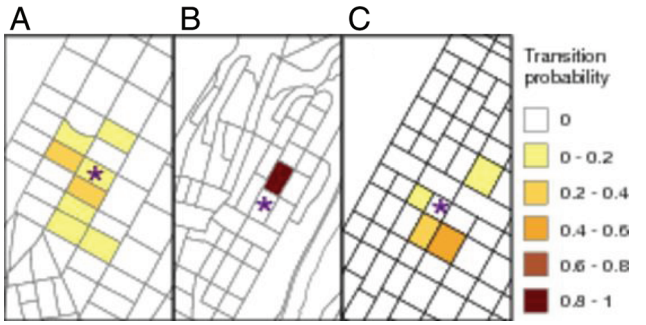

Fig. 2. Transition probabilities for the optimal strategy to de-identify $s \leq 20,000$ patients from New York County, NY, with a maximum re-identification probability of $\frac{s}{20,000}$. Transition probabilities from 3 of the 988 census blocks are shown, illustrating a few of the many possible transition distributions. The shading in region j represents the value of the probability P_{ij} of transitions into the region. (*A*) Patients in one census block (asterisk) may remain there or 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 reassigned from one block (asterisk) to one of four nearby census blocks. No patients are reassigned to the original census block (i.e., $P_{ii} = 0$).

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 $I_j = \{i : P_{ij} = 1\}$, if nonempty, has the property that $\sum_{i \in I_j} n_i \geq \frac{s}{\xi}$. 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 $\frac{s}{\xi}$. 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.

Application

We determine optimal strategies 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 ≈1,500 people (13). According to the 2000 census, the 988 census blocks in New York County contain between 0 and 15,112 people. We devise the optimal strategy to de-identify a set of $1 \leq s \leq 20,000$ patients with a maximum re-identification probability of $\frac{s}{20,000}$. Transitions from any census block were restricted to its nearest 100 neighbors. The LP model was solved using CPLEX LP software (14), 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 Fig. 2. 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.

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 $\frac{s}{1,696,038}$, or $s \cdot 0.00000059$. The expected transition distance is 6.4 km. The least populated nonempty census block group contains only one individual, so the strategy of reassigning 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 Fig. 3.

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 $\frac{s}{884}$ 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 $\frac{s}{8,188}$. 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 25–150 times farther than the optimal LP strategy (Fig. 3).

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.

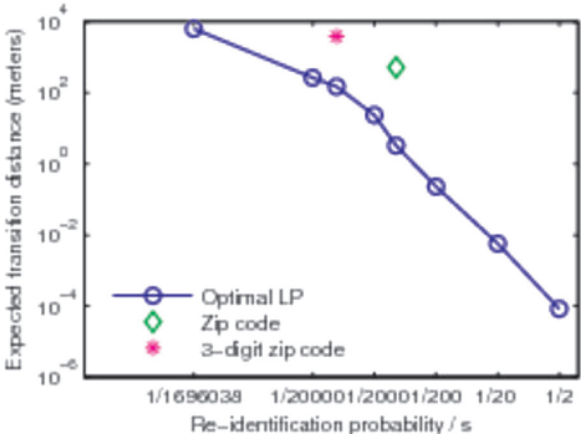

Fig. 3. Relationship between the re-identification probability, the number s of patients, and the expected transition distance for the optimal LP strategy to de-identify patients by census block group in New York County, NY. As the level of privacy protection decreases (from left to right along the x-axis), 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.

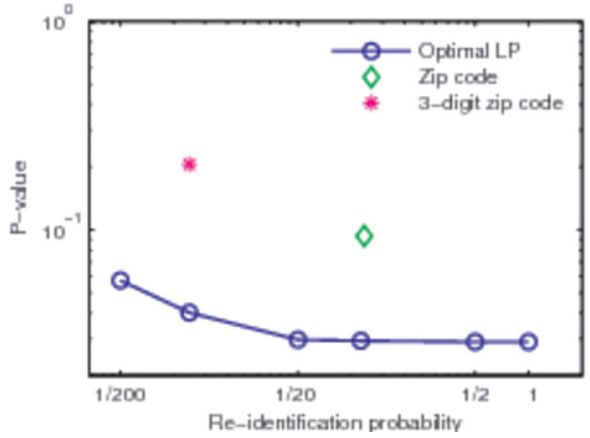

Fig. 4. Detection of clusters in case-control datasets. One thousand sets of controls and cases containing a cluster were de-identified using the LP method (blue line), aggregation by zip code (green diamond), or aggregation by the first three zip code digits (magenta asterisk). The *x*-axis shows the re-identification probability, which ranged from 0.005 to 1 (original dataset). The *y*-axis shows the mean *p*-value of the most likely cluster averaged over all datasets. Clusters de-identified using the LP method were detected with greater fidelity (i.e. lower *p*-value) than those de-identified using aggregation.

We constructed 1,000 datasets of 100 patients representing cases and 100 patients representing controls. All patients 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 patients 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 (15, 16) was applied to each de-identified set, and the *p*-value of the most significant cluster found was recorded (Fig. 4).

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-identification 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 1). The LP method performed favorably relative to aggregation for all of the counties. For example, in Plymouth County, which is about one-hundredth as dense than New York County, the LP strategy with re-identification probability 0.0001 is expected to move a data point 1.9 km, whereas aggregation by zip code moves points a farther distance of 3.1 km and has a 5-fold greater disclosure risk.

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 datasets 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 certain cells, aggregating rows or columns, and rounding or adding noise to cells (8, 17–19). 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 nonbinary method as discussed under *LP De-identification*, 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 dataset corresponds to any single individual in

Table 1. Re-identification probability and expected distance moved for LP strategy and aggregation in counties having a range of population densities

		LP method $d^{\dagger}$				Zip 5[‡]		Zip 3[§]	
County name	ρ^{*}	$\frac{\xi}{s}^{\P} = 10^{-1}$	$\frac{\xi}{s} = 10^{-2}$	$\frac{\xi}{s} = 10^{-3}$	$\frac{\xi}{s} = 10^{-4}$	$\frac{\xi}{s}$	d, m	$\frac{\xi}{s}$	d, m
Franklin County, MA	39.3	0.00	0.00	89.6	4,736	$2.8 \cdot 10^{-3}$	2,640	$1.2 \cdot 10^{-5}$	13,226
Plymouth County, MA	276.2	0.00	0.00	62.0	1,908	$5.5 \cdot 10^{-4}$	3,123	$8.9 \cdot 10^{-6}$	19,115
Middlesex County, MA	687.1	0.00	0.02	31.5	1,105	$6.5 \cdot 10^{-4}$	1,770	$4.9 \cdot 10^{-6}$	10,793
New York County, NY	25,846	0.00	0.08	4.3	172	$1.1 \cdot 10^{-3}$	519	$1.2 \cdot 10^{-4}$	3,866

[*] 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 dataset.

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 predefined 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 datasets in which each patient is labeled as a case or a control. 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. If the control locations do not represent a threat to privacy, or if they are selected by the end-user, they may be independently de-identified by the end-user with the matrix $\{P_{ij}\}$. 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 (*SI*) *Text*). Second, the data recipient must not have knowledge to suggest that membership in the dataset is not completely random; otherwise it may be possible to apply a denoising 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 dataset itself. Devising such a denoising technique, however, would be difficult in general because the noise added by the LP model depends on the original data in a complicated way (20). 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 (17, 21), although progress has been made for other types of data (22–24). In the simplest case, a coarse discrete identifier can be incorporated into the de-identification procedure. For example, if the final version of the dataset 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 $\mathcal{B}$ of final locations should be grid points or small administrative units instead of addresses. To minimize the likelihood that distinct original locations are moved to common final locations, $\mathcal{B}$ should chosen to satisfy the condition that each point in $\mathcal{A}$ has a distinct nearest neighbor in $\mathcal{B}$.

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 for individual spatial data have also been proposed. These include Spruill's measure for business data (25), 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 permuting the order of the exact locations of all patients in the original set measures well by Spruill but is clearly unacceptable for privacy protection because it reveals all the locations. Conversely, assigning completely random locations to de-identify a dataset of two patients measures poorly by Spruill but would certainly preserve privacy.

Armstrong *et al*. (9) also proposed four other measures of confidentiality. The first of these is a qualitative measure of vulnerability to geographical knowledge, 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 dataset 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 dataset 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 dataset created with a slightly lower risk of disclosure $\xi \cdot \frac{s-1}{s}$, but in which one of the census numbers n_i has been overestimated by one in the model. This is likely to have little effect on the disclosure risk. Please see *SI Text* for further discussion of inaccurate census estimates.

Appendix

Here, we derive Eq. **3**, which guarantees that the probability that any location from a de-identified dataset originated from any specific individual in the underlying population is at most ξ. Consider the probability of re-identifying a set of s patients that have been randomized to new locations. Given the set $\mathcal{A}$ of possible original locations and $\mathcal{B}$ of possible final locations, let P_{ij} denote the probability of transition from location $i \in \mathcal{A}$ to location $j \in \mathcal{B}$. Given the set of s locations comprising the de-identified dataset, 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 dataset originated from an arbitrary specific individual is required to be at most $\frac{\xi}{s}$. Let X and Y denote the original and transformed locations, respectively. This condition is formally expressed as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Eq. **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 $i \in \mathcal{A}$ and $j \in \mathcal{B}$. Following the proof backwards from Eq. **13**, this means that the probability that a location from the de-identified dataset originated from an arbitrary specific individual is less than or equal to $\frac{\xi}{s}$ for every location. Since 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.

ACKNOWLEDGMENTS. We thank Clark Friefield, Gopal Ramachandran, Lucy Hadden, Peter Szolovits, Gil Alterovitz, Michael Baym, and Ronald Rivest for helpful discussions. This work was supported by National Library of Medicine. Grant LM007677-03S1.

1. Seaman V (1798) An inquiry into the cause of the prevalence of the yellow fever in New York. *Med Repository* 1:315–372.
2. Sweeney L (2002) k-Anonymity: A model for protecting privacy. *Int J Uncertainty Fuzziness Knowl Based Syst* 10:557–570.
3. Brownstein J, Cassa C, Mandl K (2006) No place to hide: Reverse identification of patients from published maps. *N Engl J Med* 355:1741–1742.
4. Moskop JC, Marco CA, Larkin GL, Geiderman JM, Derse AR (2005) From Hippocrates to HIPAA: Privacy and confidentiality in emergency medicine. Part I: Conceptual, moral, and legal foundations. *Ann Emerg Med* 45:53–59.
5. Copyright Office, Library of Congress (2002) US Federal Register 67:53182–53273.
6. Fefferman NH, O'Neil EA, Naumova EN (2005) Confidentiality and confidence: Is data aggregation a means to achieve both? *J Public Health Policy* 26:430–449.
7. Olson KL, Grannis SJ, Mandl KD (2006) Privacy protection versus cluster detection in spatial epidemiology. *Am J Public Health* 96:2002–2008.
8. Cox LH (1996) Protecting confidentiality in small population health and environmental statistics. *Stat Med* 15:1895–1905.
9. Armstrong MP, Rushton G, Zimmerman DL (1999) Geographically masking health data to preserve confidentiality. *Stat Med* 18:497–525.
10. Cassa CA, Grannis SJ, Overhage M, Mandl KD (2006) A context-sensitive approach to anonymizing spatial surveillance data: Impact on outbreak detection. *J Am Med Inform Assoc* 13:160–165.
11. Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L (2008) Privacy: Theory meets practice on the map. *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7–12 2008* (IEEE, Washington D.C.), pp 277–286.
12. Strayer JK (1989) *Linear Programming and Its Applications* (Springer, New York).
13. US Census Bureau (2000) *Census 2000: Census Block Groups Cartographic Boundary Files Descriptions and Metadata*. Available at www.census.gov/geo/www/cob/bg_metadata.html. Accessed 12/1/07.
14. ILOG, Inc (1999) ILOG CPLEX (ILOG, Gentilly, France), Ver 10.010. Available at www.ilog.com/products/optimization/archive.cfm.
15. Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: Detection and inference. *Stat Med* 14:799–810.
16. Kulldorff M (1997) A spatial scan statistic. *Commun Stat Theory Methods* 26:1481–1496.
17. VanWey LK, Rindfuss RR, Gutmann MP, Entwisle B, Balk DL (2005) Confidentiality and spatially explicit data: Concerns and challenges. *Proc Natl Acad Sci USA* 102:15337–15342.
18. Salazar-Gonzalez J (2005) Protecting tables with cell perturbation, Working Paper 25. *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* (UNECE/Eurostat, New York).
19. Duncan G, Fienberg S (1998) Obtaining information while preserving privacy: A Markov perturbation method for tabular data. *Proc Stat Data Protection Conf* (Eurostat, Lisbon), pp 351–362.
20. Kargupta H, Datta S, Wang Q, Sivakumar K (2005) Random-data perturbation techniques and privacy-preserving data mining. *Knowl Inf Syst* 7:387–414.
21. Gutmann MP, Stern PC (2007) *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* (National Research Council of the National Academies Press, Washington, DC).
22. Lakshmanan L, Ng R, Ramesh G (2005) To do or not to do: The dilemma of disclosing anonymized data. *Proc ACM SIGMOD Conf* (Association for Computing Machinery, New York), pp 61–72.
23. Ganta S, Acharya R (2008) On breaching enterprise data privacy through adversarial information fusion. *Proc Workshop on Information Integration Methods, Architecture, and Systems, at the 24th IEEE International Conference on Data Engineering* (IEEE, Washington D.C.), pp 246–249.
24. Aggarwal CC, Pei J, Zhang B (2006) On privacy preservation against adversarial data mining. *Proc 12th ACM SIGKDD Conf* (Association for Computing Machinery, New York), pp 510–516.
25. Spruill NL (1984) The confidentiality and analytic usefulness of masked business microdata. *Review of Public Data Use*, pp 307–314.