

# Confidentiality and Confidence: Is Data Aggregation a Means to Achieve Both?

NINA H. FEFFERMAN<sup>\*</sup>, EILEEN A. O'NEIL and  
ELENA N. NAUMOVA

## ABSTRACT

The recent adoption of electronic technologies for use in management of personal health data have been accompanied by a commensurate level of concern about privacy. Public health authorities have been able to continue their full access to personal information, while restricting the information given to academic health researchers through the practice of aggregation. Through this band-aid strategy, there is a very real potential that critical pieces of information are missing for the purposes of research. While this might be a logical sacrifice in order to preserve individual privacy, quantitative analysis of the privacy gained through this method of aggregation shows that little, if any, benefit is achieved. If aggregation were the sole available means to reach the aims of both privacy and research, then further analysis of the practice of aggregation would be unnecessary. Yet suitable privacy protection techniques abound, enabling academic research to progress while adding true protection to individual health information.

*Journal of Public Health Policy* (2005) 0, 000–000.  
doi:10.1057/palgrave.jphp.3200029

**Keywords:** privacy, confidentiality, aggregation, HIPAA, epidemiology

## INTRODUCTION

The enlightenment occurring in the public health field as a result of sophisticated analytic technologies for electronic medical records has been accompanied by a vocal demand for protection of privacy for personal information. Other fields have employed a variety of disclosure-limitation methods to shield personal information while supporting research agendas: data masking, aggregation, encryption, and cell suppression, among others. These techniques are well developed for service in the public health domain (1). The time is ripe to begin a discussion on the benefits and shortcomings of approaches to data “de-identification” as they affect public health research. To

<sup>\*</sup>Address for Correspondence: Department of Public Health and Family Medicine, Tufts University school of Medicine, 136 Harrison Avenue, USA. E-mail: nina.fefferman@tufts.edu

provide a background for this problem, we offer a brief historical context for statistical confidence and confidentiality issues. To illustrate how confidence and confidentiality are being set at odds by those in the field of public health, we provide a simple example: conducting academic research into the nature of environmental outbreaks.

Recently, to comply with existing policies on sharing information with outside researchers, the practice of aggregating data has gained momentum as the preferred means of de-identification among those who collect and hold data. But can aggregation maintain privacy of personal information and support the continuation of the research agenda? Data de-identification may be a laudable goal, but the adoption of aggregation as a *de facto* policy or ultimate solution may be premature.

We present evidence that data aggregation, a simplistic approach that has been assumed to protect both scientific advancement and individual privacy, in reality, accomplishes neither. We simulate various scenarios of aggregation and demonstrate the substantially decreased ability of statistical and epidemiological tools to adequately analyze and identify patterns of disease incidence. We demonstrate that not only that the amount of data lost by implementing such tactics can cause clearly distinct courses of disease spread to become indistinguishable, but, additionally, that they result only in minor increases (if any) to the privacy of personal information. We therefore conclude that the recently adopted practice of aggregation of data as a method of protection of privacy is ineffective and poses serious impediments to scientific understanding, accomplishing neither of its assumed goals.

#### THE HIPPOCRATIC BASIS OF PUBLIC HEALTH RESEARCH

The golden rule of research is that understanding cannot be achieved without specificity of data. The importance of maximizing the amount of detail available to researchers becomes readily apparent in a health crisis, when public health departments, researchers and the public at large scramble to understand the nature of the outbreak in order to find a suitable intervention as quickly as possible. Revisiting the research methods of Hippocrates in the 4th century B.C.E., we note that the attention to all detail is the key to the soundness of his

scientific reasoning. For instance, in one treatise, he specifies the factors to consider including the winds of the localities, the water and ground qualities and the habits of the inhabitants. Then he suggests:

From these things [the researcher] must proceed to investigate everything else. For if one knows all these things well, or at least the greater part of them, he cannot miss knowing. And in particular, as the season and the year advances, he can tell what epidemic diseases will attack the city, either in summer or in winter, and what each individual will be in danger of experiencing from the change of regimen (2).

Even in the absence of a crisis that requires immediate response, precise data are indispensable. In another treatise Hippocrates says:

With regard to diseases, the circumstances from which we form a judgment of them are, – by attending to the general nature of all, and the particular nature of each individual, – to the disease, the patient, and the applications... – to the whole constitution of the season, and particularly to the state of the heavens and the nature of each country; – to the patient's habits, regimen, and pursuits.... – from these, and their consequences, we must form our judgment (3).

Following this rule of understanding, in the last decades medical and public health research has grown enormously as the result of integration of precise data and sophisticated analytical tools. In fact, what was infeasible only a few years ago is now routine in the promotion of health of the public. The benefit gained by all through public health research is clear: a complete picture of the prevalence, incidence and etiology of diseases, and from this knowledge the public health community can devise effective interventions to protect individuals, resulting in an overall promotion of health.

The achievements in public health depend on the success of breaching gaps between basic science and public health. This means that public health officials need to provide outside scientists with access to detailed information on key elements in order to ultimately accomplish protection of the public. Person, time and place are the fundamentals of environmental outbreak investigation. These key elements of research also form the core of personal health

information, requiring the most care to ensure confidentiality, and therefore became the first and the main subjects for aggregation as a practical implication of the implementation of privacy protection policies.

#### THE PUBLIC'S PUSH FOR PRIVACY

##### *The Notion and Significance of Personal Privacy is Not New (4)*

In 1890, Louis Brandeis, later a U.S. Supreme Court justice, co-authored an article in the Harvard Law Review in which a definition of privacy was crafted that still holds sway in many law and ethics discussions – a person has a basic “right to be let alone” (5). Yet, each country, and the various cultures within countries, differ in how purely private information is distinguished from that which is to be shared communally. Within cultures, as well, the boundary between what is private and what is public knowledge shifts over time as a result of changing social norms. Referring back to the Hippocratic treatises *On Airs, Waters and Places* and *Of the Epidemics* (2), we find no evidence of concern about the privacy of personal health information. His treatises contain references to individual names, residences, dates of illness and specific symptoms, among other data that many modern cultures would consider intrusive. Even in the modern era of heightened concern for personal privacy, we find that when the level of alarm about risks to health rises, our quest for understanding what is happening so as to find an effective intervention trumps the desire for confidentiality. The recent anthrax outbreak of 2001 clearly demonstrated these phenomena (6,7).

In general, however, the growth of research efforts and the prevalence of electronically available data have spurred a major push for enhancing the protection of confidentiality and individual privacy (8). In the United States, for example, the recent adoption of the Health Insurance Portability and Accountability Act (HIPAA) regulatory protections stem from people's concern about the misuse of health information and its potential for serious economic consequences, including health insurance exclusions and employment discrimination (9). As such, the HIPAA regulatory intent is geared more toward restricting disclosures of specific medical information, including limits on how much data researchers outside

of public health agencies can access, rather than carefully balancing the need for privacy with protecting the academic research (10).

#### THE CREATION OF AN UNNECESSARY CONFLICT

Identifiable data are essential for academic health research (11) and statisticians are also looking for analytical techniques for preserving both data integrity and confidentiality (12,13). One widespread tactic adopted by health authorities around the globe to address the demand for personal privacy is to restrict the level of data access and to provide researchers with data aggregated across both time and space. Looking at the United States as one example, recent regulations under HIPAA have restricted the release of information by requiring data to be “de-identified”. The means to achieve this aim of de-identification have been simplified to the technique of aggregation, even absent a demonstration that these restrictions provide any additional privacy. By imposing artificial restrictions on data, such as aggregation of location identifiers or time of disease onset, the researcher’s ability to be precise and draw solid conclusions is seriously jeopardized.

#### ANALYTICAL EXERCISE FOR AGGREGATION

Our research explores the logical grounding for aggregation as a measure of protection. We use a simple example from the field of environmental health, specifically an environmental outbreak study. In order to be able to compare the effective increase in privacy with the resultant loss of data sensitivity, we suggest two hypothetical scenarios of disease spread across three hypothetical populations with different demographic distributions. By examining the corresponding increase in difficulty involved in isolating individuals with particular traits caused by aggregating over these three populations, we are able to abstract and understand the actual benefit achieved by this method of de-identification of data. This sort of quantitative decision-making model must be employed when evaluating any de-identification implementation to consider both potential harm to the public through loss of understanding achieved by researchers and protection of the privacy of personal information.

TABLE 1: Excerpts from the 1990 Census of MA. STF 3 Standard Extract Report – basic tables

<i>General population</i>	<i>Malden, MA Geocode 02148 used as Town A</i>		<i>Medford, MA Geocode 02155 used as Town B</i>		<i>Revere, MA Geocode 02151 used as Town C</i>	
Total persons/percent sampled	54,121	11.5%	57,353	12.7%	42,761	11.6%
Females	28,538	52.7%	30,588	53.3%	21,973	51.4%
<i>Persons by race</i>						
White	48,669	89.9%	53,479	93.2%	39,845	93.2%
Black	2,250	4.2%	2,341	4.1%	664	1.6%
Asian & Pacific Islander	2,670	4.9%	1,229	2.1%	1,534	3.6%
American Indian, ESK., ALEUT	116	0.2%	95	0.2%	77	0.2%
Hispanic (any race)	1,567	2.9%	960	1.7%	1,644	3.8%
<i>Persons by age</i>						
0-4/PCT/CUM PCT	3,674	6.8	3,161	5.5	2,560	6.0
5- 9	2,638	4.9	2,712	4.7	2,344	5.5
10-13	1,938	3.6	2,010	3.5	1,455	3.4
14-17	2,335	4.3	2,106	3.7	1,599	3.7
18-24	5,994	11.1	7,792	13.6	4,617	10.8

TABLE 1 (continued)

<i>General population</i>	<i>Malden, MA Geocode 02148 used as Town A</i>		<i>Medford, MA Geocode 02155 used as Town B</i>		<i>Revere, MA Geocode 02151 used as Town C</i>	
25-34	11,971	22.1	11,258	19.6	8,009	18.7
35-44	7,661	14.2	7,780	13.6	5,809	13.6
45-54	4,818	8.9	5,303	9.2	4,414	10.3
55-59	2,266	4.2	2,850	5.0	2,240	5.2
60-64	2,541	4.7	2,720	4.7	2,380	5.6
65-74	4,466	8.3	5,304	9.2	4,164	9.7
75-84	2,861	5.3	3,218	5.6	2,410	5.6
85 and over	958	1.8%	1,139	2.0%	760	1.8%
Under 20	11,886	22.0%	12,326	21.5%	8,919	20.9%
20-39	21,104	39.0%	20,859	36.4%	14,622	34.2%
40-64	12,846	23.7%	14,507	25.3%	11,886	27.8%
65 and over	8,285	15.3%	9,661	16.8%	7,334	17.1%
Median age	33.7		34.7		36.3	

PPL-JPHF3200029

## METHODS

We used census data from three existing zip codes in the outskirts of Boston, MA from 1990 in order to inform the demographics for three hypothetically adjacent populations: Towns A, B and C with populations of 55,000, 45,000 and 25,000 respectively (see Table 1) (14). We introduced a hypothetical environmentally introduced disease with a 7-day latent period and no secondary transmission into these towns via two separate scenarios. In the first scenario, pathogen exposure was introduced separately in each of the towns on consecutive days and was present in each town for exactly 1 day. In the second, pathogen exposure was present on the same day in all three towns and the exposure lasted for exactly 1 day. We assumed the same probability of exposure (when present) and subsequent infection in all towns. Note that this can easily be relaxed without altering the result. To estimate disease incidence per day, we used a hypothetical distribution  $S(T)$ , the probability that, having been exposed and infected at time  $T=0$ , symptoms begin to manifest at time  $T$  (see Table 2). This distribution incorporated the 7-day latent period, so no disease incidence is expected until 1 week subsequent to initial exposure. In each of these two scenarios, we then deidentified the data by aggregating the reported disease incidence across towns and over time in weeks. So that the shape of the total incidence curve can be seen, including the slope from prior to first onset, even in the graphs aggregated by week, we incorporated the latent period.

De-identification by aggregation of data achieves increased protection of privacy for the individual by creating a larger group of possible individuals whom a particular disease report could represent, thereby providing greater anonymity as part of a larger crowd. This anonymity is the privacy achieved and an increase in anonymity is considered equivalent to an increase in privacy. As a result, in order to estimate the level of privacy provided by these aggregation steps, a “least likely” hypothetical individual was created from the demographic information for Town C (the smallest population), causing that individual to be the most identifiable since the group from which he was drawn is then the smallest possible, providing the fewest alternatives with whom he can be confused and therefore the least amount of privacy. In other words,  $s$  = percent of

TABLE 2: Distribution over time of the probability of developing symptoms given infection on day  $T=0$ 

<i>Day</i>	<i>S(T)</i>
1	0.00
2	0.00
3	0.00
4	0.00
5	0.00
6	0.00
7	0.00
8	0.10
9	0.15
10	0.25
11	0.22
12	0.15
13	0.08
14	0.04
15	0.01

Note: This distribution function does not need to sum to 1 if not all those infected develop symptoms, however, in this case, we have assumed that it does.

the general population with the same sex as the individual in question,  $r$  = percent of the general population with the same race, and  $a$  = percent of the general population in the same age category were chosen in order to minimize  $X$  defined as  $X = s \times r \times a$ . This person will therefore have the least privacy due to the minimal population of same description.

We assumed all identifiers to be independent (e.g. males and females are equally as likely to be white, and equally as likely to be 45 years old, etc.). Note that this causes the privacy benefit to be increased, and is therefore a conservative assumption, since it maximizes the size of each population sharing similar descriptions. We then compared the size of this “concurrent” population to those generated by each of the aggregation steps in order to demonstrate comparative privacy benefits. In order to examine the “most likely” relative privacy benefits achieved by aggregation, these same steps were repeated for a second hypothetical individual having in each category the most common characteristics as falling into the majority

TABLE 3: Descriptive properties used as hypothetical individuals used for Scenarios 1 and 2

	<i>'Least likely' individual</i>	<i>'Most likely' individual</i>
Sex ( <i>s</i> )	Male	Female
Race ( <i>r</i> )	American Indian, Esk., Aleut	White
Age Bin ( <i>a</i> )	85 and over	25–34

for each census category. Mathematically speaking, the selection of *s*, *r* and *a* maximize *X*.

These two cases of “least likely” and “most likely” represent the two extremes of a progressive spectrum of anonymity and all other cases will fall between the two, so the privacy benefit gained by the aggregation will as well. These computations are the result of the simple multiplication of mathematical probabilities dictated by population demography. By examining the differences between the incidence curves in the two scenarios from the three separate towns and those in the aggregated curves, and comparing the relative increases in privacy resulting from that aggregation, we can estimate the protective benefits of the aggregation policy and the costs to our ability to differentiate among spread scenarios (Table 3).

Q1

## RESULTS

The actual results from the two different spread scenarios when left completely un-aggregated were distinguishable from one another (see Figure 1a and b). In Town A, because neither the timing nor level of exposure differ between scenarios, neither daily nor temporally aggregated reported case numbers allowed differentiation between patterns of disease spread (see Figures 2a and 4a), the resulting numbers are identical and therefore indistinguishable. In Towns B and C, the two different disease spread scenarios were clearly distinct when numbers were counted daily (see Figure 2b and c). The patterns were identical, merely shifted in terms of timing, so there is no room for confusion between the scenarios. When new cases are aggregated across all three towns, it is no longer possible to tell whether disease is slowly spreading across all three towns or is present in all at once at a lower level of either exposure or pathogenicity or even reporting.

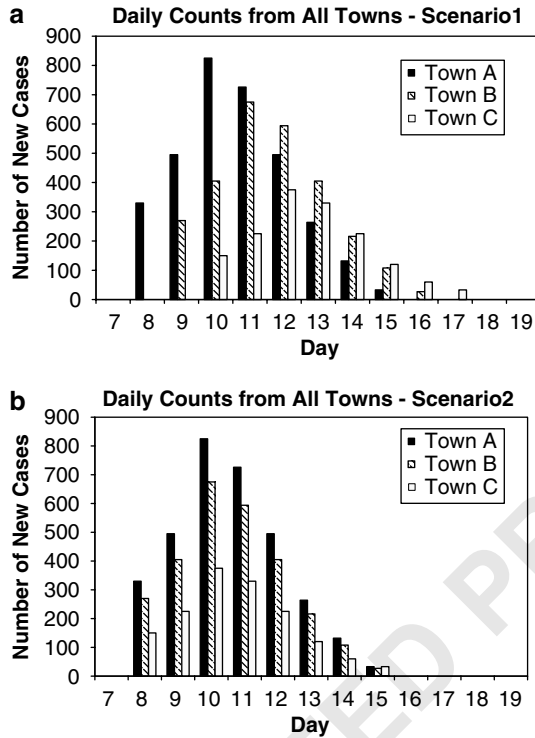
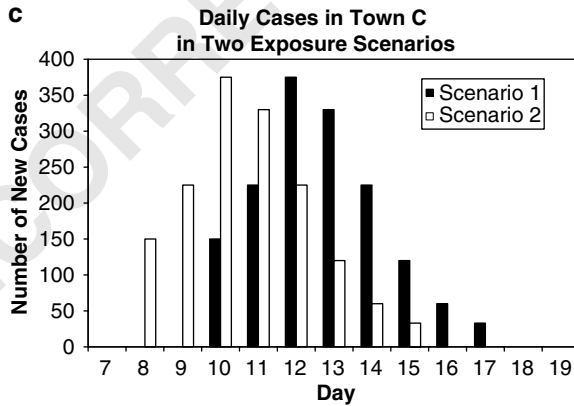
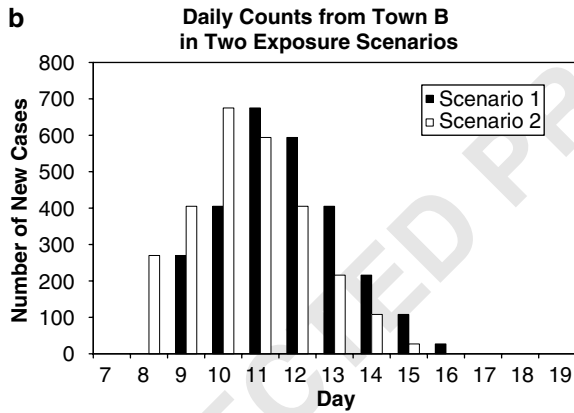
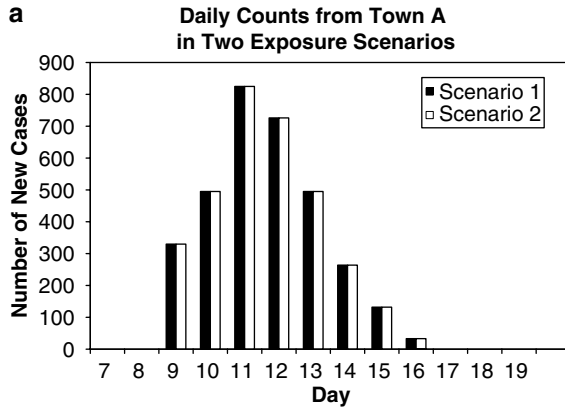


FIGURE 1  
The actual numbers of cases in each town per day in the two exposure scenarios

In this scenario, it is not until the 11th day following the first instance of exposure that the two scenarios may be distinguished from each other (see Figure 3). Even on the 11th day, differentiation is only possible because in one scenario the disease incidence continues to rise and in the other, it begins to fall. If instead of aggregating spatially, but allowing the reporting of daily counts, we allow each town to report separately but aggregate temporally, grouping by week, differentiating the two scenarios, especially if there is uncertainty as to level of exposure or pathogenicity, is impossible until the end of the third week following exposure. Up until that point, there are no differences in the progressing of changes of reported level and by that point, in this example, all incidence of disease stemming from the exposure in question has already occurred (see Figure 4a–c). If both spatial and temporal aggregation is



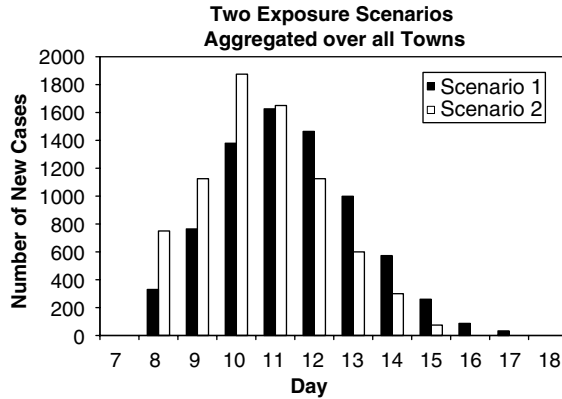


FIGURE 3

Once the reported case numbers are aggregated across all three towns, during the first 3 days of reported cases (days 8 through 10), the shape of the incidence curves are nearly identical, although of different magnitudes. This makes it impossible to tell the two scenarios apart since the exposure spread pattern in Scenario 2 with a slightly lessened level of exposure would produce nearly the same numbers as those seen here in Scenario 1. It isn't until the 11th day after exposure was first present (the fourth day of reported cases) that the two scenarios become clearly differentiable

employed, the two scenarios of disease spread become completely indistinguishable (i.e. the shape of the curves representing disease incidence and the magnitude of those curves are nearly identical), even with complete knowledge after all incidence of disease has been reported and analyzed (see Figure 5).

Using an American Indian, 85-year-old male as a hypothetical “least likely” individual, on average, less than one individual in any possible aggregated population was likely to report disease. Therefore, no studied scenario of aggregation provided greater anonymity, or privacy, within the population. Using a white, 29-year-old female as a hypothetical “most likely” individual, spatial aggregation did not alter the order of magnitude of the population of identical demographic description likely to report illness during a daily count.



FIGURE 2

(a) Two different disease exposure scenarios for Town A reported daily. Since the two different scenarios did not differ in the exposure for Town A, the numbers of new cases they produce are the same. (b and c) Two different disease exposure scenarios for Towns B and C, respectively, reported daily. In this case, the two scenarios are distinguishable from one another by the end of the first day of reported cases in each town. The daily reporting makes the time shift obvious

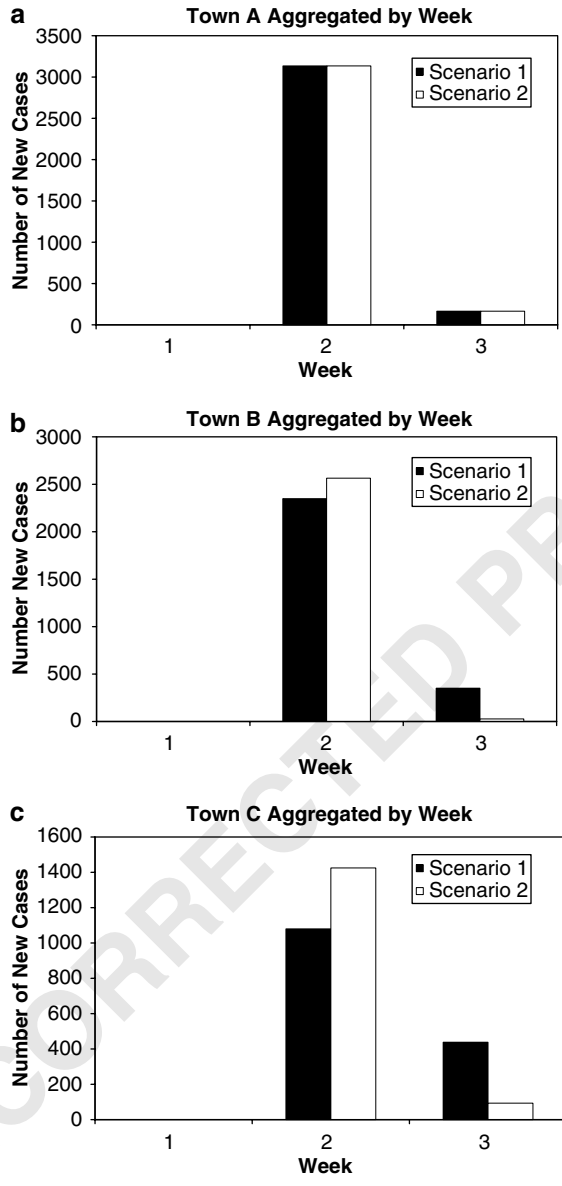


FIGURE 4

Aggregating temporally by collecting weekly data in each town separately does not allow differentiation between the two scenarios before the end of the third week after exposure

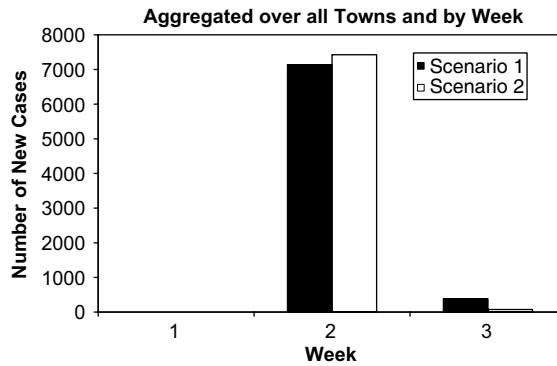


FIGURE 5

Aggregating reported cases both spatially over all three towns and temporally by week, all ability to discriminate between the two scenarios of disease spread is lost

TABLE 4: The size of the population sharing the same demography as a hypothetical individual over different aggregations and how many from that population are likely to report illness during a given period of time

	<i>'Least likely' individual population size</i>	<i>Reporting illness during this period</i>	<i>'Average' individual population size</i>	<i>Reporting illness during this period</i>
Town A – Day 14	0.0104	0.00003	5759	14
Town B – Day 14	0.0084	0.00004	4381	21
Town C – Day 14	0.0049	0.00004	2240	20
Town A – Week 2	0.0104	0.00057	5759	305
Town B – Week 2	0.0084	0.00042	4381	229
Town C – Week 2	0.0049	0.00020	2240	105
All Towns – Day 14	0.0226	0.00010	12161	56
All Towns – Week 2	0.0226	0.00129	12161	694

While temporal aggregation did achieve an increase of privacy by an order of magnitude in the sub-population likely to report illness, it did not alter the magnitude of the population of identical demographic description. Aggregation across both time and space achieved an increase of an order of magnitude in the anonymity of the “most likely” individual (from  $10^4$  to  $10^5$ ) (see Table 4).

## DISCUSSION

Fostering an expansion of knowledge through public health research and protecting individual privacy are both foci of public concern. While promoting the highest level of privacy for the public when sensitive information is involved, the same public depends equally on the work of public health officials and academic researchers for safeguarding health status. Attention must be paid to the costs and benefits inherent in any implementation strategy proposed in order to ensure both goals with a minimum of compromise. The use of any one trade-off strategy between the two, in the absence of quantitative analysis and understanding, can lead to well-meaning but ineffectual outcomes that neither protect the privacy of the individual nor benefit the society as a whole through the support of basic research. By examining the progressive levels of aggregation applied to the different scenarios of disease spread and analyzing their implications both with regard to personal privacy and sensitivity of monitoring, we are able to provide just such a quantitative analysis and understand the actual effects of data aggregation. We are therefore able to conclude that aggregation is an unsatisfactory method of implementation of public policy.

Spatial aggregation of disease reporting can significantly decrease a researcher's ability to assess the shape of a disease outbreak signature (as can be seen in the progression of aggregation from Figures 1–3). This leads to the possible mis-estimation of the magnitude of a coming outbreak by either underestimating the pathogenicity of the disease in question, exposure rates in a given population, or the size of population(s) affected. Each of these factors is crucial to the understanding of disease spread, for both public health officials and academic researchers, in order to make the most effective policy decisions such as those designed to increase awareness among healthcare providers or to issue public health advisories.

Temporal aggregation deals an even more detrimental blow to the amounts of information that can be drawn from incidence numbers. Even when still spatially discrete, grouping reported case numbers to weekly (rather than daily) rates increases the length of data collection necessary to differentiate between even two simple scenarios, by an entire week (see Figure 4a–c). Given the incubation

period of our particular hypothetical illness, this delays any possible understanding until after the last cases of infection from that exposure are already reported. Finally, aggregating over both time and space, we are able to distinguish nothing, even after all cases have been reported (see Figure 5).

Clearly, the process of data aggregation involves the loss of information but this approach might still be considered of value if aggregation confers commensurate privacy benefits to individuals seeking medical care. However, by examining the size of the sub-population sharing exactly the same demographic characteristics as our hypothetical patients, we see that the increase in privacy is minimal. In practice, privacy is most unlikely to be maintained in the case of the most unusual individuals since, by definition, they have some qualities that easily separate them from the norm. In the populations of the hypothetical towns studied, even when aggregating both temporally and spatially, suppressing the information most crucial to researchers and public health workers, the gain to privacy was insufficient to increase the size of the sub-population in which our “least likely” individual could effectively “hide” by an order of magnitude (see Table 4). In fact, given the extreme minority of our “least likely” population, it was insufficient even to increase the number of likely patients fitting the demographic description to two. However, this is a function of our population size and model demography. It is possible to imagine both a larger population and a more even distribution of race and age class, which would lead to a likely ill sub-population of more than 1. It should be noted that any population can be evaluated for individual vulnerability to identification in the way here presented: finding both a minimal and maximal  $X$  given all pertinent characteristics.

Examining the benefits to the average individual (those least in need of privacy protection since their anonymity is already greater than individuals in any other group because they belong to the most likely demographic bin for each characteristic) leads us to conclude that in this case also, the benefits are minimal. The magnitude of the population sharing the same demographic characteristics remained unchanged by both temporal and spatial aggregation separately. Only by aggregating over both time and space, our most information-costly course of action, were we able to see any actual benefit (see Table 4). In practice, even this benefit is likely to be useless to

anyone wishing to track a particular individual given their demographic information; the least privacy afforded, given our population size and demography, was already in the thousands. It may not be unreasonable to assume that this is already a prohibitively large population from which to find a particular individual, even for the most dedicated.

With reliance on aggregation as the practical method of implementing public health privacy policies, extremes are replaced with averages, links are obscured, and paths for further investigation are blocked. In short, solid research conclusions may be impossible. While public health workers are able to access more private information than their academic counterparts, the fundamental limitation on their interactions with colleagues performing research which could better inform policy decisions again compromises their protective role. This becomes especially costly since there is no substantial benefit to privacy achieved by the method of “protection.” After careful examination, it is apparent that the aggregation method greatly compromises the ability of researchers to analyze reported data and communicate their results to the public, while also conferring no true benefit of privacy to individual members of the public. This can be particularly troubling when the outbreak of disease is a “public” event, by virtue of its media attention and potential threat to individuals in a context of crisis.

We therefore recommend that the strategy of aggregation for implementing the policy of protecting privacy be employed only as a last resort until another method, one that, under quantitative investigation, is able to provide the desired increase in personal privacy *and* maintain an appropriate level of detail for public health officials and academic researchers. There is already a plethora of easily implementable tools that have been developed by theoretical computer scientists and applied mathematicians that can protect privacy without losing specificity of pertinent information (15). These cryptographic methods were designed to protect the anonymity of users engaging in business transactions over the Internet, but can be easily co-opted for use in public health. They are constantly being revised and reviewed to ensure that the privacy they promise is actually achieved and that the distinct information they preserve remains reliable. Studies are published in readily available and highly respected journals (e.g. *Journal of Cryptology*), reviewed at interna-

tional conferences (e.g. Crypto, EuroCrypt, AsiaCrypt, Financial Cryptography), discussed in industry (e.g. crypto groups at AT&T, IBM, Microsoft, Matsushita), academia (e.g. MIT, Stanford, University of Waterloo, ETH, Technion) and many governmental agencies in the USA (e.g. NSA, NIST) and nearly every other country already. We recommend that it is to these methods that researchers and public officials turn for more effective strategies. More fundamentally, however, we recommend that no strategy for the implementation of public health policy be put to use without first weighing both its costs to basic research and (more importantly) its efficacy. Without such study, it is easy for sub-optimal or, indeed, detrimental techniques to become the standard method in use.

## REFERENCES

1. Statistical Policy Working Paper 22 – Report on Statistical Disclosure Limitation Methodology; <http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=1>.
- Q2 2. Hippocrates, *On airs, waters, and places II*, trans. Francis Adams.
- Q3 3. Hippocrates, *Of the epidemics, Book I, Section III*, trans. Francis Adams.
4. Freeman P, Robbins A. The U.S. health data privacy debate. Will there be comprehension before closure? *Int J Technol Assess Health Care*. 1999;15(2):316–31.
5. Warren SD, Brandeis LD. The right to privacy. *Harvard Law Rev*. 1890;4(5):193–220.
- Q4 6. Washington Post, The, 2002, postal inspector has symptoms but no confirmation of anthrax, Washington Post. January 10, p. B2.
7. Steinhauer J. Two new anthrax infections found; previous cases share same strain. *The New York Times*. 2001, October 20.
- Q5 8. Rotenberg M. Fair information practices and the architecture of privacy. *Stanford Technol Law Rev*. 2001:1.
9. Annas G. HIPAA regulations – a new era of medical record privacy? *N Engl J Med*. 2003;348(15):1486–90.
- Q6 10. 45 CFR Part 160.
11. Black N. Secondary use of personal data for health and health services research: why identifiable data are essential. *J Health Services Res Policy*. 2003;51:36–40.
12. Korn D. Medical information privacy and the conduct of biomedical research. *Acad Med*. 2000;75(10):963–8.

13. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. *Stat Med.* 1999;18(5):497–525.
14. 1990 Census of MA. *STF 3 Standard Extract Report – Basic Tables*: <http://www.census.gov/>.
15. Schneier B. *Applied cryptography: protocols, algorithms, and source code in C*, 2nd edition. New York: Wiley & Sons; 1996.
16. Cox LH. Protecting confidentiality in small population health and environmental statistics. *Stat Med.* 1996;15(17–18):1895–905.

Q7

UNCORRECTED PROOF