

The effects of the Health Insurance Portability and Accountability Act privacy rule on influenza research using geographical information systems

Norisse Tellman¹, Eric R. Litt², Caprice Knapp^{1,3}, Aaron Eagan¹, Jing Cheng⁴, Lewis J. Radonovich Jr¹

¹National Center for Occupational Health and Infection Control, Veterans Health Administration, Gainesville, FL 32608, USA; ²Rehabilitation Outcomes Research Center, Veterans Health Administration, Gainesville, FL 32608, USA; ³Departments of Epidemiology and Health Policy Research, University of Florida, Gainesville, FL 32608, USA; ⁴Division of Oral Epidemiology and Dental Public Health University of California at San Francisco, San Francisco, CA 94143-1361, USA

Abstract. The Health Insurance Portability and Accountability Act (HIPAA) privacy rule was enacted to protect patients' personal health information from undue disclosure. Despite its intention to protect patients, recent reports suggest that HIPAA restrictions may be negatively impacting health research. Quantitative, visual geographical and statistical analysis of zip code geographical information systems (GIS) mapping, comparing 3-digit HIPAA-compliant and 5-digit HIPAA-non-compliant simulated data, was chosen to identify and describe the type of distortion that may result. It was found that unmitigated HIPAA compliance with HIPAA mapping rules distorted the GIS zip code data by 28% leading to erroneous results. Thus, compliance with HIPAA privacy rule when mapping may lead investigators to publish erroneous GIS maps.

Keywords: Health Insurance Portability and Accountability Act (HIPAA), influenza, infectious disease, outbreak, infectious disease transmission, geographical information systems.

Introduction

The Health Insurance Portability and Accountability Act (HIPAA) privacy rule was enacted, to establish national standards in the United States of America to protect individuals' medical records and other personal health information. The HIPAA privacy rule codified detailed regulations regarding the use and dissemination of information that could be used to identify an individual, often

called protected health information (PHI) (US Department of Human Health Services, the Privacy Rule). Protecting individuals' privacy is an ethical responsibility, which, if done correctly may lead to increased trust between the subject and researcher and greater participation of subjects in health care research. In this context a breach in privacy and noncompliance with HIPAA may lead to negative social, medical or psychological ramifications such as diminished dignity, stigmatization or discrimination at work or school.

In the research community there continues to be debate about the HIPAA privacy rule's effectiveness at protecting individuals' privacy and whether the standards that have been set are hindering research (Ness, 2007; Greene et al., 2008; Gostin and Nass, 2009; Steinberg and Rubin, 2009; Wartenberg and

Corresponding author:

Norisse Tellman

National Center for Occupational Health and Infection Control

Veterans Health Administration

1601 SW Archer Road (151B)

Gainesville, FL 32608, USA

Tel. +1 813 966 3312

E-mail: nmisdary@gmail.com

Thompson, 2010). Accordingly, the National Academies of Sciences' Institute of Medicine (IOM) published a report in 2009 about the HIPAA regulations concluding that "the HIPAA privacy rule does not protect privacy as well as it should, and that, as currently implemented, the privacy rule impedes important health research". The committee found that the privacy rule (i) is not uniformly applicable to all health research; (ii) overstates the ability of informed consent to protect privacy rather than incorporating comprehensive privacy protections; (iii) conflicts with other federal regulations governing health research; (iv) is interpreted differently across institutions; and (v) creates barriers to research and leads to biased research samples, which generate invalid conclusions (IOM, 2009). The IOM recommended that "the committee proposes a bold, innovative, and more uniform approach to the dual challenge of protecting privacy while supporting beneficial and responsible research; second, in the event that policy makers decide that HIPAA was – and continues to be – the most useful model for how to safeguard privacy in health research, the committee proposes a series of detailed proposals to improve the HIPAA privacy rule and associated guidance" (IOM, 2009).

Impact of HIPAA on mapping research

One topic not addressed in the IOM report was health research using geographical information system (GIS) research. "A GIS integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information. GIS allows us to view, understand, question, interpret, and visualize data in many ways that reveal relationships, patterns, and trends in the form of maps, globes, reports, and charts" (<http://www.esri.com/what-is-gis/index.html>). GIS technology is a tool in public health research that can help provide up-to-date graphical information about resource distribution, disease patterns, and loci of events ("hot spots") (Rogers and Randolph, 2003; Waring et al., 2005). For

example, in the context of the recent H1N1 influenza pandemic, GIS could have been used to graphically match resource demand with resource supplies by overlaying a map showing the locations of all known influenza cases with a map of outpatient clinics and hospitals. GIS technology may also help fill knowledge gaps about transmission of diseases, evolution of disease outbreaks and predicting future disease spread using computer-aided stochastic modeling and other analytical techniques. The technology has been used to track the spread of diseases, rapidly identify disease clusters and outbreaks, and examine large scale data for spatio-temporal relationships (Yasnoff and Sondik, 1999; Guerra et al., 2002; Kistermann et al., 2002; Rogers and Randolph, 2003; Waring et al., 2005).

As the applications of GIS have advanced, the need for researchers to publish GIS-related data has become increasingly prevalent and important for academic progress. It is widely understood that certain elements of GIS information may theoretically be used to identify the individual represented by the data. Accordingly, privacy is a concern that is often raised when researchers seek to publish their results. Currently, no specific federal standards exist for publication of maps showing GIS research data (VanWey et al., 2005). According to current HIPAA regulations, protected health information (PHI) must be removed before data collection, analysis and/or publication, unless explicit permission ("informed consent") is granted by the subject(s) from whom the information issues (US Department of Human Health Services, HIPAA Regulations §164.514). In contrast to many other research settings, de-identification of GIS-related information in the context of health research often defeats the primary objective of GIS: to display the data graphically in a way that communicates new information or knowledge.

The effect of HIPAA on publication of GIS data

Specifically, HIPAA states that all geographical subdivisions smaller than state including address, city, county, precinct, zip code and their equivalent

geo-codes (latitude and longitude) must be eliminated from the data to be considered de-identified prior to publication. The only exception is that the initial three digits of a zip code may be published, if the represented geographic area that has a population greater than 20,000 (US Department of Human Health Services, HIPAA Regulations §164.514). Initially it may seem possible to easily meet the HIPAA zip code requirements by simply combining the areas represented by five-digit zip codes into a display showing only the larger area represented by 3-digit zip codes (Fig. 1). To accomplish this manipulation, however, all of the data contained in each of the five-digit areas must be moved to the geographical centroid of the larger corresponding three-digit area. In effect, this manipulation may reduce the level of precision and introduce statistical bias into the results. This issue has been identified as the modifiable area unit problem (Dark and Bram, 2007). Several studies have shown that this type of aggregation can inadvertently introduce bias, blur meaningful variations in data, limit disease clustering detection or artificially shift the geographical location of the results (Armstrong et al., 1999; Kulldorff, 1999; Kwan et al., 2004; Gregorio et al., 2005; Bell et al., 2006).

Despite numerous qualitative examples reported in the scientific literature, relatively little has been done to quantify the extent of error introduced by the HIPAA three-digit zip code rule. One publication (Olsen et al., 2006) that has addressed a similar topic did so by analysing the performance of “spa-

tial cluster detection,” a software programme often utilised by GIS researchers to identify the relative locations of disease clusters. In this study, the authors converted data from an exact address to a regional centroid location, with the aggregation, identification of significant disease clusters erroneously decreased from 73% to 45% (Olsen et al., 2006). Another study aggregated simulated data from an exact location to a census tract, which resulted in an inaccurate display of a pattern of disease risk (Boulos et al., 2006).

The purpose of this paper is to describe and quantify the effects of aggregating data to comply with the three-digit HIPAA zip code rule with an outbreak of contagious infectious disease resembling influenza. This project was undertaken with the intention to provide quantitative information to policy-analysts and decision-makers who may be in a position to make modifications to HIPAA regulation.

Materials and methods

Simulated patient-level influenza data was created to represent the number of influenza cases during a 12-month period in Florida. This type of simulated data was used in lieu of real influenza case-data because of the privacy limitations posed by HIPAA. Accordingly, no institutional review board (IRB) approval was obtained for this project. One intention of the simulated data was to have its graphic appearance resemble the spatial distribution pattern of an ordinary influenza outbreak in Florida during

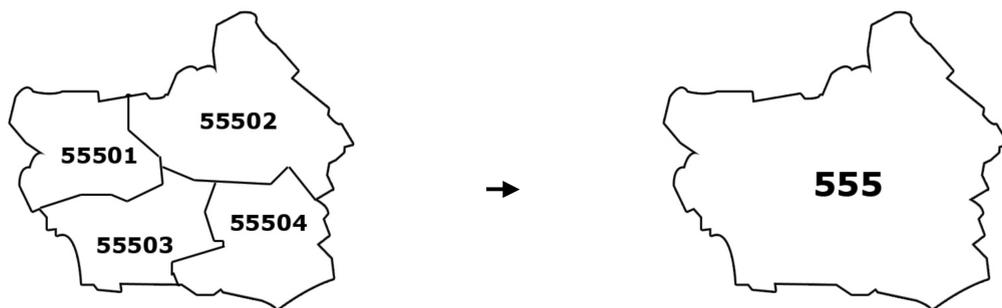


Fig. 1. Transforming zip code data for publication: initial data collected at the precision level of five-digit zip codes censored to show a precision level of three-digit zip codes.

the winter months. A close graphical approximation of an influenza outbreak was achieved using publicly available influenza data posted on websites operated by the State of Florida (Florida Department of Health, 2009). To create the simulated data representing an influenza outbreak, the following steps were performed 100 times:

- (i) *Assigning prevalence (P_i) in each five-digit zip code.* Commensurate with the variability, typically experienced with influenza prevalence during an outbreak (Florida Department of Health, 2009), each of the 914 geographical areas represented by five-digit zip codes within the State of Florida were assigned a random percentage between 5% and 15% (Centers for Disease Control and Prevention, 2009). Each prevalence rate, called true prevalence, was represented by P_i , $i = 1, \dots, 914$, the true prevalence per identified zip-code.
- (ii) *Determining case burden (C_i) in each five-digit zip code.* The total number of influenza cases in each zip code denoted as C_i , $i = 1, \dots, 914$, was computed by multiplying the prevalence of influenza in the corresponding five-digit zip code by the population in the corresponding five-digit zip code (ESRI population data - 1999-2009, ArcGIS 9.3 Redlands, CA, USA).
- (iii) *Assigning density (D_i) in each five-digit zip code.* The flu density, denoted as D_i , $i = 1, \dots, 914$, in each five-digit zip code, was calculated by dividing the number of influenza cases in each five-digit zip code in the geographical area of the corresponding five-digit zip code in km^2 (ESRI area data - ArcGIS 9.3 Redlands, CA, USA).
- (iv) *Computing estimated prevalence (\hat{P}_i) with three-digit zip code data.* In practice, we were not able to know the true prevalence in each five-digit zip code area (P_i) but could only estimate the prevalence in each five-digit zip code area \hat{P}_i with the three-digit zip code data available to public. The estimated prevalence in each five-digit zip area (\hat{P}_i) was computed by summing all the cases that fell in the three-digit zip code area and dividing by the sum by the total

population in the corresponding three-digit zip code area. The five-digit zip code areas within the same three-digit zip code area would have the same estimated prevalence \hat{P}_i .

- (v) *Computing estimated density (\hat{D}_i) with three-digit zip code data.* Similarly, we can estimate the five-digit zip code flu density (\hat{D}_i) with the three-digit zip code area data as $(\hat{P}_i \times \text{the population size in the five-digit zip code area}) / (\text{area in } \text{km}^2)$.

The absolute error (AE) and relative error (RE) of the estimated prevalence values and flu density values in each five-digit zip code which were calculated with the three-digit zip code data were computed for each simulated data set as the absolute difference between the estimated prevalence (density) and true prevalence (density) and the percentage of the absolute difference relative to the true prevalence (density):

$$\text{AE}_i = | \hat{P}_i - P_i | \text{ for prevalence, } = | \hat{D}_i - D_i | \\ \text{for flu density, } i = 1, \dots, 914$$

$$\text{RE}_i = 100\% \times | \hat{P}_i - P_i | / P_i = 100\% \times | \hat{D}_i - D_i | / D_i, \\ i = 1, \dots, 914$$

where AE is the absolute difference between the estimated values computed with the three-digit data and the true values we simulated for each five-digit and RE is the ratio (percentage) of AE to the true value for the five digit.

Finally average AE and RE were computed over 100 simulated data sets for each five-digit zip code in Florida. A random data set from the 100 simulated data sets was mapped to visually display differences in the two different mapping techniques (using ArcGIS 9.3, 1999-2009). Two maps were developed: (i) a map of the cases assigned to each five-digit zip code area; and (ii) a map of the cases assigned to each three-digit zip code area, such that the five-digit zip code values were placed in aggregate at the centroid of the corresponding three-digit area.

Results

Figures 2a and 2b display the results of both maps described in the previous section. When mapped at

the precision level of five-digit zip codes on this map, the cases tend to cluster at the coastal areas and areas that correspond to large urban cities. Cases tend to be sparser in the panhandle and the lower central portion of the state. The lowest density can be seen in the central lower portion of the state.

When mapped at the three-digit level, cases tend to be spread out more evenly over the entire state. There is less of a distinction in case-count between the panhandle and the remainder of the state. There is an area with case clustering in the south-eastern portion of that state. In both Figure 2a and 2b there is a portion of the state where no cases can occur. This location can be seen in the south-central por-

tion of the state and represents large national parks and preserves. The difference between the two maps can be seen by the location of case clustering. In Figure 2a there are distinct areas of case clustering and areas in which there is a sparse number of cases as well as no cases. In Figure 2b this clustering is less identifiable and cases are more even distributed throughout the state.

Tables 1 and 2 show the average AE and RE of the estimated prevalence and influenza density with the three-digit zip data for Florida across the 914 zip codes. The mean absolute difference between the estimated prevalence with three-digit data and true prevalence simulated for the five-digit data was

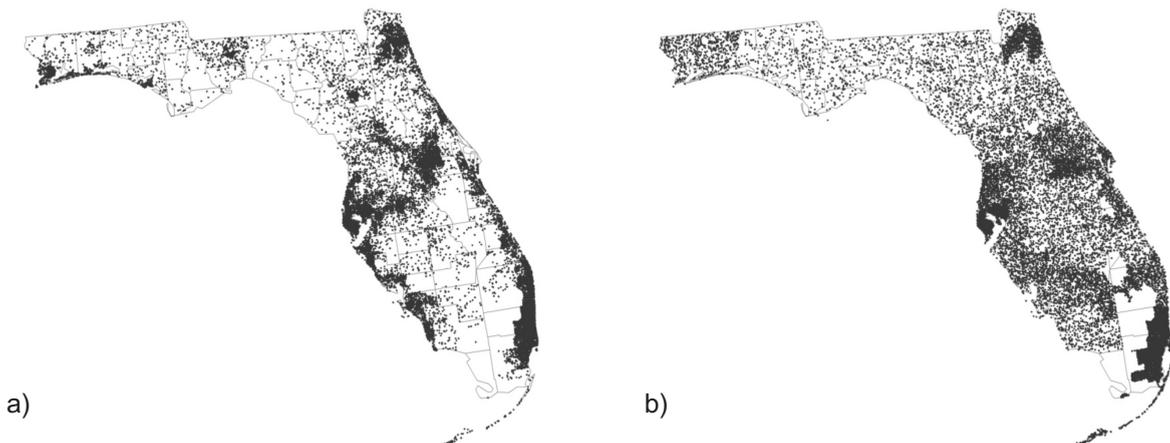


Fig. 2. GIS mapping of identical datasets showing a simulated influenza outbreak (each dot represents one case). 2a) Influenza data at the precision level of five-digit zip codes; 2b) influenza data at the precision level of three-digit zip codes.

Table 1. Average absolute error (AE) and relative error (RE) of estimated prevalence and influenza density in State of Florida with three-digit zip data.

	Mean	Minimum	Median	Maximum
AE of \hat{P}_i (%)	2.46	1.96	2.46	2.89
AE of \hat{D}_i (cases/km ²)	18.57	0.02	9.79	202.91
RE of \hat{P}_i and \hat{D}_i (%)	28.3%	21.4%	28.2%	37.3%

Table 2. Average absolute error (AE) and relative error (RE) of estimated prevalence and case density in State of Florida using three-digit data in simulated data.

	AE	RE
	Median (1 st quartile, 3 rd quartile)	Median (1 st quartile, 3 rd quartile)
\hat{P}_i (%)	2.46 (2.36, 2.57)	28.2% (26.6%, 30.0%)
\hat{D}_i (cases/km ²)	9.79 (1.49, 28.45)	

approximately 2%. The mean absolute difference between the estimated density with the three-digit data and true density for the five-digit data was approximately 19 cases per km². Similarly, the mean RE of the estimated values with the three-digit data compared to the true values for the five-digit data was approximately 28%.

Discussion

The analyses presented in this paper demonstrate that full compliance with zip code constraints posed by HIPAA privacy regulations may substantially distort GIS mapping data to the extent that the results may be misrepresented and/or misinterpreted. Compliance with these regulations while publishing the type of incidence data, typically produced in an influenza outbreak, may introduce errors approaching 30%. Errors of this magnitude stand to have short-term and long-term effects. In the short term, censoring may prompt decision-makers to inadvertently misappropriate funds. In the long term, systematic blurring and publication of data to make it less precise carries the risk of directing researchers and the wider public health community down an erroneous pathway chasing false results. In this context, it is conceivable that the consequences of data censoring, performed in the name of privacy protection may, in fact, be worse than a breach in privacy.

An important finding of this project is that this type of data censoring may result in a disproportionate application to low population or rural geographic areas compared to urban or large population because this aggregation skews the data (Figs. 2a and 2b). This can be seen when the findings are applied to an example of influenza cases and needed medical supplies for a city population of 50,000 compared to a city with a population of 10,000. If a coastal city on the Gulf of Mexico with a population of ~50,000, were to purchase a supply of N95 respirators for healthcare workers based on 15% of the population developing an influenza infection (Centers for Disease Control and Prevention, 2009),

approximately 16,600 respirators would be needed according to a predictive model developed by one of the authors (Lewis, 2009). If the number of respirators ordered erroneously fell by 28%, an order would be placed for 14,500 respirators (a shortage of 2,100 respirators). Conversely, for an inland city with a population of ~10,500 and an estimated 15% of the population infected annually by influenza (Centers for Disease Control and Prevention, 2009), there would need to be about 3,500 respirators ordered for healthcare workers to treat influenza patients. With a 28% greater amount of cases identified about 3,900 respirators would be ordered due to the error (an excess of about 400). Although these numbers may appear small, they become massive when applied across a large population. This error occurs with the use of the three-digit data map (Fig. 3) where cases are now more evenly spread over the state. There are more cases introduced into rural areas where low population and disease rate occur and with the error introduction a surplus of supplies might be sent. On the other hand, in large urban city where cases have been decreased and spread to outer areas, fewer cases would be represented per population and therefore a deficient of supplies might be assigned.

With such a wide variety of applications for the use of GIS and public health research changes in policy may allow for more accurate and vital information to be published. In addition to the recommendations made by the IOM in 2009, we propose that new policies should seek to better ensure protection of individuals' privacy and exhibit sufficient flexibility to take into account unique research methods such as GIS or unforeseen technologies to be developed in the future.

Note

The views expressed in his manuscript do not necessarily reflect the views of the United States Government, the United States Department of Veterans Affairs, the University of Florida or the University of California at San Francisco.

References

- Armstrong MP, Rushton G, Zimmerman DL, 1999. Geographically masking health data to preserve confidentiality. *Stat Med* 18, 497-525.
- Bell S, Hoskins R, Pickle L, Wartenberg D, 2006. Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and public. *Int J Health Geogr* 5, 49.
- Boulos M, Cai Q, Padgett J, Rushton G, 2006. Using software agents to preserve individual health data confidentiality in micro-scale geographical analyses. *J Biomed Inform* 39, 160-170.
- Centers for Disease Control and Prevention, 2009. Influenza. <http://www.cdc.gov/flu/keyfacts.htm> (accessed: June 2009).
- Dark S, Bram D, 2007. The modifiable areal unit problem (MAUP) in physical geography. *Prog Phys Geog* 31, 471-479.
- Environmental System Research Institute (ESRI), 2009. What is GIS? Geographic Information System. <http://www.esri.com/what-is-gis/index.html> (accessed: June 2009).
- Florida Department of Health, 2009. Influenza surveillance reports 2009. http://www.doh.state.fl.us/disease_ctrl/epi/htopics/flu/reports.htm (accessed: June 2009).
- Gostin L, Nass S, 2009. Reforming the HIPAA privacy rule: safeguarding privacy and promoting research. *JAMA* 301, 1373-1375.
- Greene S, Bennet S, Kirilin B, Oliver K, Pardee R, Wagner E, 2008. Impact of the HIPAA privacy rule in the HMO research network. The National Academy of Science.
- Gregorio D, DeChello L, Samociuk H, Kulldorff M, 2005. Lumping or splitting: seeking the preferred areal unit for health geography studies. *Int J Health Geogr* 4, 6.
- Guerra M, Walker E, Jones C, Paskewitz S, Cortinas MR, Stancil A, Beck L, Bobo M, Kitron U, 2002. Predicting the risk of Lyme disease: habitat suitability for *Ixodes scapularis* in the North Central United States. *Emerg Infect Dis* 8, 289-297.
- IOM (Institute of Medicine), 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. The National Academies Press, Washington, DC, USA.
- Kistermann T, Munzinger A, Dangendorf F, 2002. Spatial patterns of tuberculosis incidence in Cologne (Germany). *Soc Sci Med* 55, 7-19.
- Kulldorff M, 1999. Geographic information system (GIS) and community health: some statistical issues. *J Public Health Manag Pract* 5, 100-106.
- Kwan M, Casas I, Schmitz B, 2004. Protection of geoprivacy and accuracy of spatial information: how effective are geographical mask? *Cartographica* 39, 15-28.
- Ness R, 2007. Influence of the HIPAA privacy rule on health research. *JAMA* 298, 2164-2170.
- Olsen K, Grannis S, Mandl K, 2006. Privacy protection versus detection in spatial epidemiology. *Am J Public Health* 96, 2002-2008.
- Radonovich LJ, Magalian PD, Hollingsworth MK, Baracco G, 2009. Stockpiling supplies for the next influenza pandemic. *Emerg Infect Dis* [serial on the Internet], <http://www.cdc.gov/EID/content/15/6/el.htm> (accessed: June 2009).
- Rogers DJ, Randolph SE, 2003. Studying the global distribution of infectious diseases using GIS and RS. *Nat Rev Microbiol* 1, 231-237.
- Steinberg M, Rubin E, 2009. The HIPAA privacy rule: lacks patient benefit impedes research growth. Association of Academic Health Centers, Washington DC, USA.
- US Department of Health and Human Services. HIPAA Regulations §164.514(a) and (b). http://edocket.access.gpo.gov/cfr_2004/octqtr/pdf/45cfr164.514.pdf (accessed: June 2009).
- US Department of Health and Human Services. The privacy rule. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/index.html> (accessed: June 2009).
- VanWey L, 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.
- Waring S, Zakos-Feliberti A, Wood R, Stone M, Padgett P, Arafat R, 2005. The utility of geographic information systems (GIS) in rapid epidemiological assessments following weather-related disasters: methodological issues based on the tropical storm Allison experience. *Int J Hyg Environ Health* 208, 109-116.
- Wartenberg D, Thompson WD, 2010. Privacy versus public health: the impact of current confidentiality rules. *Am J Public Health* 100, 407-411.
- Yasnoff W, Sondik E, 1999. Geographic information systems (GIS) in public health practice in the new millennium. *J Public Health Manag Pract* 5, 9-12.