Reverse Geocoding: Concerns about Patient Confidentiality in the Display of Geospatial Health Data

John S. Brownstein, PhD [1,2,3], Christopher Cassa, MEng[1,2]
Isaac S, Kohane MD PhD[1,3] and Kenneth D. Mandl MD MPH[1,2,3]
[1]Children's Hospital Informatics Program, Children's Hospital Boston, Boston, MA
[2]Division of Emergency Medicine, Children's Hospital Boston, Boston, MA
[3]Department of Pediatrics, Harvard Medical School, Boston, MA

Abstract

Widespread availability geographic information systems (GIS) software has facilitated the use health mapping in both academia and government. Maps that display patients as points are often exchanged in public forums (journals, meetings, web). However, even these low resolution maps may reveal confidential patient location information. In this report, we describe a method to test whether privacy is being breached. We reverse geocode from maps with cases and describe the accuracy with which patient addresses can be extracted.

Methodology

We created a prototypical patient map for an urban metropolitan area (Figure 1). The image displays 550 randomly selected patients at a resolution of 50 dots per inch and a scale of 1:100,000. The reverse geocoding process involves georeferencing the raster image file followed by conversion to vector data to obtain estimated patient address centroids. Our ability to identify patient address was assessed by measuring the distance from the predicted location to the known location. Using digitized building outlines for the city, we estimated the minimum buffer size needed to contain the correct address. Accuracy can then be defined as the number of incorrect addresses within this buffer. To test a possible remedy, the original dataset was de-identified using a spatial anonymization package and detection results were compared with those from the original dataset.

Evaluation Results

We directly identified 26% (144/550) of the addresses with low quality GIS output of patient location. The reverse geocoded location was on average within 21.0 meters (SD, 31.8) of the correct address. Overall, 99.8% of the addresses were within 70 meters. The average number of building needed to identify the correct address was 8.8 (SD, 11.9). Overall, 51.6% of addresses were identified within five buildings, 70.7% within ten buildings and 93% within twenty buildings. A significant decrease in accuracy of detection was obtained after applying anonymization.

Conclusions

Our results signify that, even from low resolution maps, patient addresses can be re-identified. Thus, the release of geospatial data information on the web, at meetings and in publications may be in direct violation patient confidentiality. This result serves as a warning to individuals that use GIS in the context of medical research. New spatial data standards that protect confidentiality while still effectively communicating information about spatial pattern are needed.

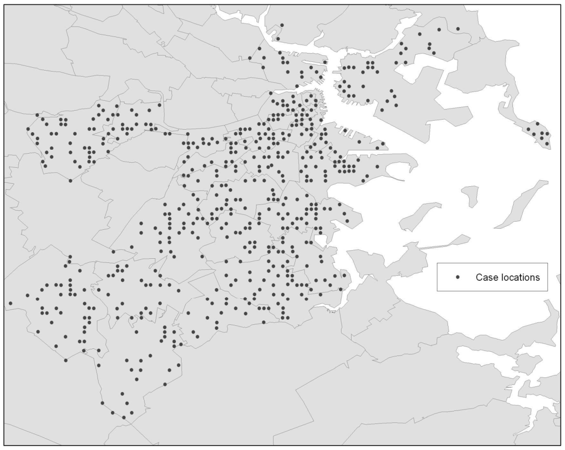

Figure 1. Prototypical patient map for Boston, Massachusetts. Image displays 550 randomly selected patient addresses.