

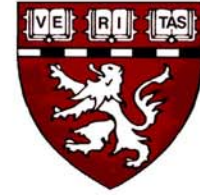


# Bioterrorism Syndromic Cluster Creation Tool: A system to generate sets of cluster coordinates

Christopher A. Cassa,<sup>1,2,3</sup> Karen L Olson, PhD,<sup>2,3,4</sup> Kenneth D Mandl, MD MPH<sup>2,3,4</sup>

<sup>1</sup>Massachusetts Institute of Technology, LCS Clinical Decision Making Group; <sup>2</sup>Children's Hospital Informatics Program; <sup>3</sup>Center for Biopreparedness at Children's Hospital Boston; <sup>4</sup>Harvard Medical School

This work was supported by the National Institutes of Health through R01LM07677-01 and R01LM007970-01 from the National Library of Medicine and by grant 2002-12-1 from the Alfred P. Sloan Foundation.



## Objective

Create a software program that enables easy generation of semisynthetic datasets for outbreak detection benchmarking.

## Background

- Syndromic surveillance algorithms can be measured by their ability to detect signal (disease outbreak) against a noisy background (normally varying baseline disease in the region).
- Such benchmarking requires training and validation datasets. Because few people have been infected with biological warfare agents, such data are virtually nonexistent.
- One approach to simulation is to use semisynthetic datasets—real background noise spiked with an artificial signal—to measure the performance characteristics of detection algorithms.

## Methods

### Single cluster creation

A tool was created that generates single artificial patient clusters using the user-entered parameters below.

### Variable parameters

- Cluster ID number
- Number of points in the cluster
- "Hospital" GIS location
- Maximum cluster radius
- "Angle" from the hospital
- Distance from the hospital
- Numbers of days cluster should span
- Date algorithm to use
- Description, and output filenames

## Multiple Cluster Creation

The cluster generation tool can also create series of patient clusters that vary over a given range for specific parameters. Each cluster in the series will share the same parameter values except for one varied parameter. Users enter the range of that varied parameter in the user interface.

### Parameters that vary over sets of clusters

- Number of points in the cluster
- Maximum distance from cluster center point
- Number of days of cluster duration
- Angle around the hospital
- Distance from hospital

### GIS data points and data engine

- Patient addresses are converted to latitude/longitude pairs for output in the datasets.
- Cluster creation tool uses geospatial data engine to create large sets of randomized datapoints that meet specific parameter criteria.



Figure at left: A single linear time-growth cluster north of MIT. Legend enumerates the relative date number and the corresponding color that is used in the map. Figure at right: Creation of a series of four clusters about MIT (with the angle varied.)

### Temporal data simulation

Cluster generating program models disease growth that follows different time patterns: a random spread, a linear spread, and an exponential spread.

### Output files from cluster creation tool

Each cluster that is created has its own file that contains the cluster point ID, the longitude and latitude of the cluster point, and the relative date of the cluster point.

### Results

- The cluster generator was thoroughly tested by creating single clusters and series of clusters.
- Results clusters were entered into a GIS mapping visualization program for validation and testing.
- All cluster points fell within the specified boundaries requested in the client interface and all output files were successfully created.

## Conclusion

This cluster creation tool can be used to thoroughly test outbreak detection algorithms over many parameters to uncover strengths and weaknesses of each algorithm in different patient cluster situations.

The screenshot shows a web-based interface for generating clusters. It is divided into two main sections: 'Generate Single Cluster' and 'Generate Series of Clusters'. The 'Generate Single Cluster' section includes input fields for 'Cluster ID', 'Number of points in the cluster', 'Hospital location (Longitude and Latitude)', 'Maximum cluster radius', 'Angle from the hospital', 'Distance from the hospital', and 'Number of days cluster should span'. It also has a dropdown for 'Date algorithm to use' (linear, exponential) and a 'Save Cluster Data' button. The 'Generate Series of Clusters' section includes input fields for 'Number of clusters', 'Parameter to vary', and 'Parameter Minimum Value'. It has a dropdown for 'Number of points in the cluster' and a 'Create Clusters' button. A note at the bottom states: 'Note: All other parameters (including the file name and the cluster ID) will be taken from the upper portion of the screen.'

Figure. The graphical user interface for the cluster generator. The top portion includes parameters for creation of single patient clusters. The bottom portion includes parameters to vary over a set of clusters and range values to be used.