How Many Suffice? A Computational Method for Sizing Sentinel Surveillance Networks

Data from surveillance networks help epidemiologists and state public health officials detect emerging diseases, conduct outbreak investigations, manage epidemics, and better understand the mechanics of a particular disease. In previous work, we showed how to apply facilities location algorithms to designing disease surveillance networks and primary stroke center networks. In this work, we introduce the notion that surveillance networks are used to determine outbreak intensity (i.e., disease burden), outbreak location, and outbreak timing (i.e., the start, peak, and end of the epidemic). In this work, we primarily consider outbreak intensity and outbreak timing. We focus on outpatient influenza surveillance, although these methods are general and applicable to many more disease surveillance networks.

We developed a web-based calculator that provides a simple user interface for public health officials to determine the best site placement for their state. The web-based calculator supports three different site placement algorithms: two algorithms based on the maximal coverage model and one based on the K-median facilities location model. The maximal coverage model (MCM) considers each site as having a fixed coverage radius. The MCM chooses the sites that maximize the total number of people within the specified distance of a site. Note that the standard MCM formulation places no restrictions on the number of cases a site can service (or in this case, detect). In the real world, however, surveillance sites cannot detect an infinite number of cases, as each site will have some established natural limit, for example, in terms of the number of patients it can service. Such site capacity constraints are explicitly modeled in the capacitated MCM formulation where each site is endowed with some intrinsic integer capacity. The K-median model (sometimes also referred to as the P-median model) minimizes the sum of the distances from each individual to the individual’s nearest site.

Which of these algorithms produces the “best” set of surveillance sites? Are there conditions where the K-median model is “better” than the MCM (or vice versa)? Does either algorithm provide “better” site selection than the existing Iowa Department of Public Health (IDPH) network? We can answer many of these questions empirically by simulating the spread of influenza across the state of Iowa and calculating the probability of each case being detected by any surveillance site. Our dataset consists of two million de-identified Medicaid billing records representing eight complete influenza seasons from July 2000 to June 2008. These records comprise all of the Iowa Medicaid records from this time period that contain any one of 30 pre-specified ILI ICD-9 codes. We treat the Medicaid dataset as the record of all influenza cases that occurred in Iowa between 2000 and 2008. The probability of case detection is determined by the Huff model, a probabilistic model often used in geography literature to analyze and understand aggregate behavior.

We first analyze outbreak intensity and show that networks our algorithms design always outperform the existing IDPH network. In fact, our networks are able to detect the same number of cases that the existing network can detect using many fewer surveillance sites. To analyze outbreak timing, we correlate the percent difference time series for the detected cases and the full dataset of cases. We show that when algorithmically selecting surveillance sites, the size of the network plays only a minimal role in the quality of the outbreak timing detection. Selecting just a few strategically-placed surveillance sites is enough to reliably and accurately determine the onset, peak, and end of the influenza season in the state of Iowa.

The major contribution of this work is the introduction and study of two different metrics on which a surveillance network can be optimized: outbreak intensity and outbreak timing. We show that our methods allow us to design surveillance networks which match or beat the performance of the existing IDPH network using fewer, more strategically-placed, surveillance sites.