Fall 2014

Improving disease surveillance : sentinel surveillance network design and novel uses of Wikipedia

Geoffrey Colin Fairchild
University of Iowa

IMPROVING DISEASE SURVEILLANCE: SENTINEL SURVEILLANCE NETWORK DESIGN AND NOVEL USES OF WIKIPEDIA

by

Geoffrey Colin Fairchild

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

December 2014

Thesis Supervisor: Professor Alberto Maria Segre

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Geoffrey Colin Fairchild

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the December 2014 graduation.

Thesis committee: ______________________________
Alberto M. Segre, Thesis Supervisor

____
Philip M. Polgreen

____
Sriram Pemmaraju

____
Ted Herman

____
Gerard Rushton

____
Sara Y. Del Valle

ACKNOWLEDGEMENTS

A Ph.D. is a long and arduous journey. I've been lucky to have been surrounded by a truly wonderful group of people over the last 6 years, and I know for a fact that I wouldn't have made it this far without them.

First and foremost, I want to thank Dr. Alberto Segre and Dr. Phil Polgreen for their patience, guidance, and advice. They were quick to recognize my affinity for medicine and were able to direct me to a research area that I wasn't even aware of prior to grad school. Computational epidemiology, as it turns out, is the perfect union between my skill set and research interests. Both of you have been instrumental in molding me into the computer scientist I am today, and I am forever grateful.

I also want to thank the rest of the CompEpi group at Iowa. Dr. Sriram Pemmaraju, Dr. Ted Herman, and Dr. Gerard Rushton have been especially great at critiquing my work and offering advice when necessary. The rest of my cohort at Iowa was equally important, providing many laughs and thought-provoking discussions.

Dr. Sara Del Valle, I'm not sure you'll ever know how much I appreciate the opportunity you've given me at Los Alamos. I've met some brilliant and inspiring people here, and you've opened up new avenues for me that I didn't even know were available. I hope that one day I can pay it forward.

To my friends outside of academia and work, thanks for keeping me sane and grounded. It's easy to lose track of how to have fun during grad school, but you all ensured that I never fell into that trap.

My mom, sister, dad, and stepmom have provided the kind of unconditional love and support that I so often needed. Mom, I am forever indebted to you for the sacrifices you made in order for Britt and me to have the chances we did. You kept us in good schools, and you've always been encouraging. Britt, you're my best friend. We can talk about anything, we always support one another, and I know you've got my back when I need it. Dad and Linda, I love our daily emails and weekly conversations about life. You've both always shown such an interest in my work, and I really appreciate that. I love you all more than you'll ever know. I'm so lucky to have you in my life.

Finally, I want to thank Tamiko. You supplied motivation when there was none, laughs when I was down, deep discussions, and constant encouragement. Without you, I'm not sure I would've survived grad school.

ABSTRACT

Traditional disease surveillance systems are instrumental in guiding policy-makers' decisions and understanding disease dynamics. The first study in this dissertation looks at sentinel surveillance network design. We consider three location-allocation models: two based on the maximal coverage model (MCM) and one based on the K-median model. The MCM selects sites that maximize the total number of people within a specified distance to the site. The K-median model minimizes the sum of the distances from each individual to the individual's nearest site. Using a ground truth dataset consisting of two million de-identified Medicaid billing records representing eight complete influenza seasons and an evaluation function based on the Huff spatial interaction model, we empirically compare networks against the existing volunteer-based Iowa Department of Public Health influenza-like illness network by simulating the spread of influenza across the state of Iowa. We compare networks on two metrics: outbreak intensity (i.e., disease burden) and outbreak timing (i.e., the start, peak, and end of the epidemic). We show that it is possible to design a network that achieves outbreak intensity performance identical to the status quo network using two fewer sites. We also show that if outbreak timing detection is of primary interest, it is actually possible to create a network that matches the existing network's performance using 42% fewer sites. Finally, in an effort to demonstrate the generic usefulness of these location-allocation models, we examine primary stroke center selection. We describe the ineffectiveness of the current self-initiated approach

and argue for a more organized primary stroke center system.

While these traditional disease surveillance systems are important, they have several downsides. First, due to a complex reporting hierarchy, there is generally a reporting lag; for example, most diseases in the United States experience a reporting lag of approximately 1–2 weeks. Second, many regions of the world lack trustworthy or reliable data. As a result, there has been a surge of research looking at using publicly available data on the internet for disease surveillance purposes. The second and third studies in this dissertation analyze Wikipedia's viability in this sphere.

The first of these two studies looks at Wikipedia access logs. Hourly access logs dating back to December 2007 are available for anyone to download completely free of charge. These logs contain, among other things, the total number of accesses for every article in Wikipedia. Using a linear model and a simple article selection procedure, we show that it is possible to *nowcast* and, in some cases, *forecast* up to the 28 days tested in 8 of the 14 disease-location contexts considered. We also demonstrate that it may be possible in some cases to train a model in one context and use the same model to nowcast or forecast in another context with poor surveillance data.

The second of the Wikipedia studies looked at disease-relevant data found in the article content. A number of disease outbreaks are meticulously tracked on Wikipedia. Case counts, death counts, and hospitalization counts are often provided in the article narrative. Using a dataset created from 14 Wikipedia articles, we trained a named-entity recognizer (NER) to recognize and tag these phrases. The NER achieved an F1 score of 0.753. In addition to these counts in the narrative,

we tested the accuracy of tabular data using the 2014 West African Ebola virus disease epidemic. This article, like a number of other disease articles on Wikipedia, contains granular case counts and deaths counts per country affected by the disease. By computing the root-mean-square error between the Wikipedia time series and a ground truth time series, we show that the Wikipedia time series are both timely and accurate.

PUBLIC ABSTRACT

This dissertation presents three studies that aim to improve the current state of disease surveillance.

The first study looks at designing traditional sentinel surveillance systems, which are instrumental in guiding policy-makers' decisions and understanding disease dynamics. We use several popular location-allocation models to algorithmically design surveillance networks of varying sizes. By simulating the spread of influenza across the state of Iowa, we demonstrate that we are capable of generating smaller networks capable of performance at least as good as the volunteer-based influenza surveillance network used by the Iowa Department of Public Health. We also apply these network design methods to primary stroke center placement.

The second and third studies recognize that while these traditional surveillance systems are important, they have drawbacks, such as reporting lags and untrustworthy, unreliable, or unavailable data in some instances. To help solve these problems, we introduce a novel data source: Wikipedia.

The second study displays how a linear model using time series of article accesses can nowcast and forecast a variety of diseases in a variety of locations.

Finally, the third study demonstrates how disease-related data can be elicited from Wikipedia article content. We show how a named-entity recognizer can be trained to tag case, death, and hospitalization counts in the article narrative. We also analyze tabular time series data and show that they are accurate and timely.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Disease surveillance is a crucial aspect of public health. It guides policymakers' decisions and allows scientists to better understand disease dynamics [119, 173]. Modern disease surveillance efforts have many uses. For example, vaccine allocation decisions (e.g., where and how many doses) can be made based on disease burden levels across the country [126]. Public-health-awareness programs (e.g., hand hygiene awareness campaigns) can also be directed towards specific populations or regions. Scientists use these data to tune simulations (e.g., [55, 80, 133, 151]), which can in turn be used to further aid public health officials' decision making.

Stephen B. Thacker provides a thorough overview of the history of public health and disease surveillance in [173]. Current public health surveillance concepts date back to the Middle Ages. Western European governments sought health protection and health care for their people. In the 17th and 18th centuries, basic disease surveillance data analysis began; Gottfried Wilhelm von Leibniz advocated numerical analysis of mortality data, and John Graunt developed some of the fundamental principles of public health surveillance, including disease-specific death counts, death rates, and the concept of disease patterns. In the mid-19th century, Lemuel Shattuck recommended a decennial census as well as health data collected by age, gender, occupation, socioeconomic status, and locality. Around the same time, William Farr urged the collection of vital statistics as well as reports on these data that would be made available to health authorities and the public.

Thacker then describes the evolution of disease surveillance in the United States (U.S.). In 1850, mortality statistics based on death registration and the decennial census became the first surveillance-related data published by the U.S. federal government. Weekly reporting of prevalent diseases first began in 1874 in Massachusetts. In 1878, Congress authorized the forerunner of the Public Health Service to collect morbidity data related to pestilential diseases such as cholera, smallpox, plague, and yellow fever. Following the 1918-1919 influenza pandemic, 1925 marked the year in which all states in the U.S. participated in national morbidity reporting. In 1952, mortality data were added to the forerunner of the Morbidity and Mortality Weekly Report (MMWR). The Communicable Disease Center, now the Centers for Disease Control and Prevention (CDC), took over responsibility for the MMWR in 1961.

Finally, Thacker demonstrated the importance of a nationwide surveillance program. After the Francis Field Trial of the poliomyelitis vaccine concluded in 1955, a nationwide vaccination program began. Within two weeks of the start of the vaccination program, six cases of paralytic poliomyelitis were reported through the notifiable disease program. An epidemiological study commenced, and 141 more vaccine-related cases were discovered. Investigations showed that a single manufacturer had produced the vaccines used in each case. The manufacturer had produced a single lot of the vaccine that was contaminated with live poliovirus. Were it not for the nationwide surveillance system that subsequently discovered the source contamination, the polio vaccine may have been blamed for the cases.

Traditional disease surveillance systems take the form of a reporting hierarchy [25]. The World Health Organization (WHO) coordinates global health initiatives within the United Nations (UN) system [146]. With little exception, almost every country in the world maintains its own public health department that tracks local disease spread. For example, the CDC coordinates efforts in the U.S. In each country, there are often many smaller public health departments; these exist at the state/province and county/city levels.

In practice, data representing the number of relevant cases of any one of many reportable diseases are prepared at selected primary care facilities (e.g., doctors' offices, clinics, and hospitals) and laboratories. Those data are sent to the local city, county, or state health department. At each level of the hierarchy, data may be aggregated to some level so that patient privacy is preserved. In the U.S., the CDC's National Notifiable Diseases Surveillance System collects data from each state in order to aid nation-wide policy decisions [62]. Members of the UN, such as the U.S., then collaborate with the WHO to combat disease at the global level.

In the U.S., a number of diseases are considered *reportable diseases*, or diseases that, by law, must be reported to state and local health departments, and optionally the CDC, through the appropriate channels because regular, frequent, and timely information regarding individual cases is necessary for the control and prevention of the disease [3]. Examples of reportable diseases include anthrax, cholera, Ebola, smallpox, and rabies. Some non-reportable, but still important, diseases are part of voluntary reporting programs. For example, the CDC coordinates efforts to surveil foodborne illnesses [60] and influenza [63] in more depth than the mandatory reports allow.

As a result of its infectiousness, economic impact, and seasonality, special attention is drawn to influenza by public health departments all over the world. Influenza infects approximately 5–20% of the U.S. population each year, causing, on average, over 200,000 hospitalizations [64]. It is highly infectious due in part to the fact that the disease can be spread 1 day before symptoms appear and up to 5–7 days after becoming sick [61]. Additionally, the virus can be spread through asymptomatic infections. In 1948, the WHO began its Influenza Surveillance Network [168]. It has largely contributed to the influenza vaccine formulation as well as our current understanding of influenza epidemiology. In 2001, the WHO issued a call for proposals to develop a Global Agenda on Influenza Surveillance and Control. In 2002, after consulting with a variety of experts, The Global Agenda was adopted by consensus. The Agenda focuses on *a*) increased virological and epidemiological surveillance, *b*) increased understanding of burden (e.g., economic and health) caused by influenza, *c*) antiviral and vaccine use and effectiveness, and *d*) pandemic preparedness plans.

The CDC's influenza surveillance system relies on five components, which support the WHO's Agenda [63]:

Virological surveillance data Approximately 85 U.S. WHO Collaborating Laboratories and 60 National Respiratory and Enteric Virus Surveillance System laboratories located throughout the U.S. participate in virologic surveillance for influenza.

Outpatient surveillance data Information on patient visits to health care providers for influenza-like illness is collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). ILINet consists of more than 2,700 outpatient healthcare providers in all 50 states, the District of Columbia, and the U.S. Virgin Islands reporting more than 30 million patient visits each year.

Mortality surveillance data 122 cities across the U.S. report the total number of death certificates and the number for which pneumonia or influenza was listed as the underlying or contributing cause of death. Additionally, all influenza-associated deaths in children (age < 18) are reported.

Hospitalization surveillance data Laboratory confirmed influenza-associated hospitalizations in children and adults are monitored through the Influenza Hospitalization Surveillance Network.

Geographic summary surveillance data State health departments report the estimated level of geographic spread of influenza activity in their states each week through the State and Territorial Epidemiologists Reports.

While these efforts have largely been successful, there are still areas where improvements can be made. For example, the local and national outpatient surveillance systems, which support the WHO's Agenda by furthering the understanding of disease burden, are, in many cases, *networks of convenience*; that is, they are comprised of volunteers rather than selected based on some rigorous criteria. We argue how refinement of sentinel surveillance systems can improve our understanding of disease burden. Specifically, we offer methods for designing sentinel surveillance networks that can improve outbreak intensity and timing detection. We then extend the methods used for sentinel surveillance network design by exploring the location of primary stroke centers in order to maximize the number of people that have access to such facilities.

Although the reporting hierarchy used in traditional surveillance systems can produce data that are generally considered accurate and trustworthy, this process requires careful coordination among many different agencies, which often introduces a reporting lag of at least 1–2 weeks [97]. Furthermore, each agency may have slightly different reporting standards, and data reliability may not be uniform. The hierarchy also removes much of the fine-grained details. For example, while the base data at the provider level are very granular (e.g., patient addresses are known), data that become publicly available are generally aggregated to the state or region level. This aggregation, while important for protecting patient privacy, makes fine-grained analysis and simulation difficult. To solve these problems, we study how traditional surveillance systems can be augmented with internet-based data in order to discover what is happening now (i.e., remove the reporting lag) as well as potentially improve the resolution of available data, thereby aiding in the WHO's disease burden and pandemic preparedness plan Global Agenda items.

We introduce a novel data source, Wikipedia, that can be used for a variety of disease surveillance applications. Although a number of internet-based data sources have been suggested for disease surveillance purposes (e.g., news articles, Twitter, search engines queries), no *truly open* data source has been studied. Our analysis of

Wikipedia's viability for disease surveillance purposes is motivated primarily due to its openness; one of the core tenets of Wikipedia is open and free content [71].

The first of our Wikipedia studies looks at publicly available article access logs. We demonstrate how a simple linear model can be used to nowcast and forecast a variety of diseases in a variety of locations. The second Wikipedia study looks at the article content. We show how a named-entity recognizer can be trained to recognize certain "important" phrases related to disease. We also show the viability of granular tabular data.

CHAPTER 2
HOW MANY SUFFICE? A COMPUTATIONAL METHOD FOR SIZING SENTINEL SURVEILLANCE NETWORKS

Abstract

Background

Data from surveillance networks help epidemiologists and public health officials detect emerging diseases, conduct outbreak investigations, manage epidemics, and better understand the mechanics of a particular disease. Surveillance networks are used to determine *outbreak intensity* (i.e., disease burden) and *outbreak timing* (i.e., the start, peak, and end of the epidemic), as well as *outbreak location.* Networks can be tuned to preferentially perform these tasks. Given that resources are limited, careful site selection can save costs while minimizing performance loss.

Methods

We study three different site placement algorithms: two algorithms based on the maximal coverage model and one based on the K-median model. The maximal coverage model chooses sites that maximize the total number of people within a specified distance of a site. The K-median model minimizes the sum of the distances from each individual to the individual's nearest site. Using a ground truth dataset consisting of two million de-identified Medicaid billing records representing eight complete influenza seasons and an evaluation function based on the Huff spatial interaction model, we empirically compare networks against the existing Iowa De-

partment of Public Health influenza-like illness network by simulating the spread of influenza across the state of Iowa.

Results

We show that it is possible to design a network that achieves outbreak intensity performance identical to the status quo network using two fewer sites. We also show that if outbreak timing detection is of primary interest, it is actually possible to create a network that matches the existing network's performance using 42% fewer sites.

Conclusions

By simulating the spread of influenza across the state of Iowa, we show that the networks our tool designs perform better than the status quo in terms of both outbreak intensity and timing. Additionally, our results suggest that network size may only play a minimal role in outbreak timing detection. Finally, we show that it may be possible to reduce the size of a surveillance system without affecting the quality of surveillance information produced.

2.1 Background

Although facilities location algorithms were originally used to help firms decide where to build new retail outlets or distribution centers [43], these algorithms have also been used for decades to help allocate healthcare resources. In the United States (U.S.), for example, the Emergency Medical Services (EMS) Act of 1973 required that 95% of service requests had to be served within 30 minutes in a rural area and within

10 minutes in an urban area [48]. More recently, investigators have studied how to locate EMS facilities to aid in large-scale emergencies such as earthquakes or terrorist attacks [98]. In addition to improving responses to healthcare problems, facilities location algorithms have been used to place preventive healthcare services [182] and also to design healthcare systems in developing countries [159]. In previous work, we have shown how to apply facilities location algorithms to design disease surveillance networks [157] and primary stroke center networks [112].

We focus on outpatient influenza surveillance in this paper. The Centers for Disease Control and Prevention (CDC) currently collects different types of influenza-related information [63]. Although these different systems (see Table 2.1) are in some sense complementary, they were not originally developed to optimize detection of influenza cases in any systematic way (i.e., using an explicit optimization criterion, such as maximizing population coverage or minimizing average distance to population elements). Indeed, these systems were in many cases “networks of convenience”.

Table 2.1: The five categories of ILI surveillance used by the CDC.

Category	**Description**
Viral Surveillance	Approximately 85 U.S. WHO Collaborating Laboratories and 60 National Respiratory and Enteric Virus Surveillance System laboratories located throughout the U.S. participate in virologic surveillance for influenza.
Outpatient Illness Surveillance	Information on patient visits to health care providers for influenza-like illness is collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). ILINet consists of more than 2,700 outpatient healthcare providers in all 50 states, the District of Columbia, and the U.S. Virgin Islands reporting more than 30 million patient visits each year.
Mortality Surveillance	122 cities across the U.S. report the total number of death certificates and the number for which pneumonia or influenza was listed as the underlying or contributing cause of death. Additionally, all influenza-associated deaths in children (age < 18) are reported.
Hospitalization Surveillance	Laboratory confirmed influenza-associated hospitalizations in children and adults are monitored through the Influenza Hospitalization Surveillance Network.
Summary of the Geographic Spread of Influenza	State health departments report the estimated level of geographic spread of influenza activity in their states each week through the State and Territorial Epidemiologists Reports.

Surveillance network design has recently been improved using a data-driven approach, incorporating weekly statewide data, hospitalization data, and Google Flu Trends data [163]. Although such methods may provide for better networks in certain instances, many large and populous regions of the world in critical need of surveillance lack the requisite data for such analysis (e.g., poor/untrustworthy records, lack of reasonable influenza activity estimates, lack of Google Flu Trends data in India, China, and all of Africa). Additionally, Google Flu Trends does not track influenza activity perfectly and can differ dramatically from CDC data [26]. Thus, more traditional approaches based on facilities location algorithms that require only population data are still the method of choice for surveillance network design in many regions of the world.

Surveillance networks are used to determine not just *outbreak location*, but also *outbreak intensity* (i.e., disease burden) and *outbreak timing* (i.e., the start, peak, and end of the epidemic). Using networks to detect these factors of disease spread is not new; however, to our knowledge, no other study has examined the implications of designing networks that are tuned to preferentially perform one of these three tasks. Clearly, if one were primarily interested in outbreak intensity or fine-grained outbreak location information, one would want to incorporate as many sites as possible. But given that resources are inevitably limited, careful site selection can save costs while minimizing performance loss; knowing which detection task is of primary interest is an important first step in designing more efficient and/or effective networks.

In this paper, we examine site placement for an influenza-like illness (ILI) sentinel surveillance network in the state of Iowa. Iowa is a state in the U.S., roughly 310 miles by 199 miles (500 kilometers by 320 kilometers) in area, populated by approximately three million people. In Iowa, ILI is the major form of outpatient surveillance for influenza activity. ILI is a collection of symptoms that indicate a possible influenza infection (e.g., cough, fever, sore throat). Only laboratory tests can confirm actual influenza. The Iowa Department of Public Health (IDPH) maintained 19 ILI sentinel sites in 2007, comprised of primary care facilities and test laboratories selected strictly on a volunteer basis. We analyze and compare several algorithmic surveillance site placement techniques using Iowa as a test environment, specifically in terms of detecting outbreak intensity and outbreak timing. We examine the proportion of cases detected by the different placement methods under explicit probabilistic detection assumptions. We compare these results against the number of cases that would have been detected by the 2007 IDPH network under identical assumptions. We then use statistical correlation as a means to study outbreak timing. We demonstrate how we can dramatically reduce the size of the surveillance network while still successfully detecting the start, peak, and end of the outbreak.

2.2 Methods

2.2.1 Online tool

We have developed a web-based calculator that provides a simple user interface for public health officials to determine the best site placement for every state in the U.S.[1] This web application takes as input a list of possible candidate site locations (by ZIP code — there are 935 in Iowa) and, if the user is extending an existing network, a list of any preselected site locations. The user chooses an algorithm and provides any additional parameters specific to the algorithm as well as the total number of sites required. The application then selects a set of sites and overlays the results on a map. Population coverage statistics are also shown. The calculator is capable of designing networks in every state in the U.S. and currently uses 2010 U.S. Census population data. Iowa population distribution by ZIP code is presented in Figure 2.1.

[1] `http://compepi.cs.uiowa.edu/~gcfairch/siteplacement/`

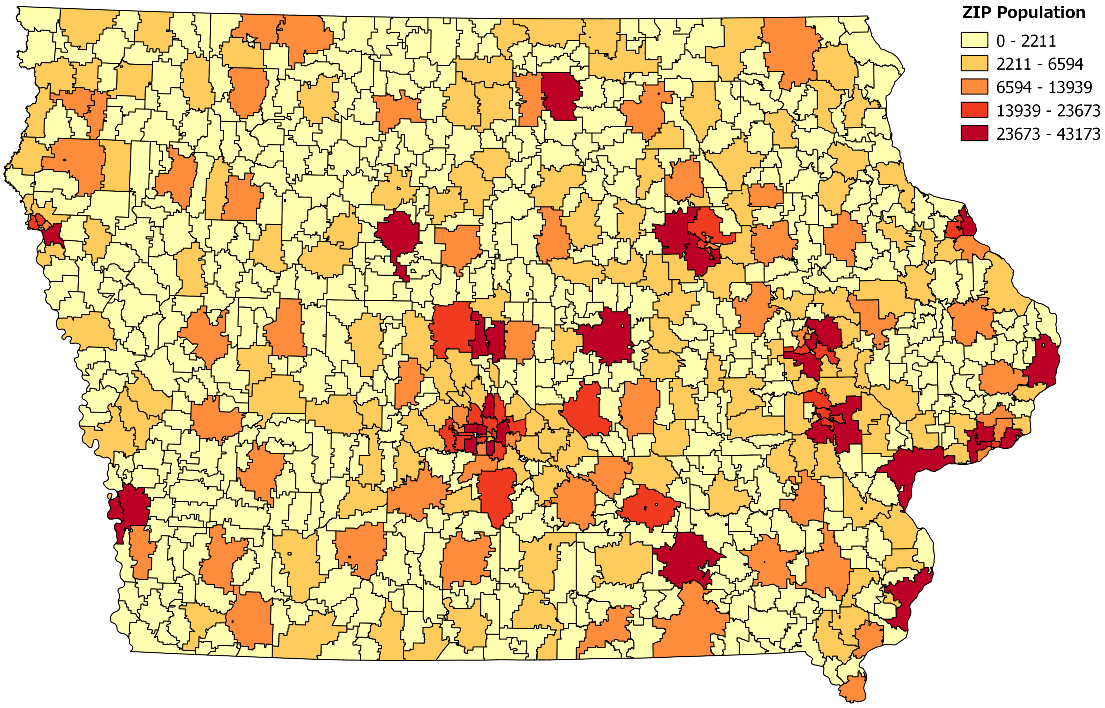


Figure 2.1: Map showing population distribution in Iowa by ZIP code. All 935 ZIP codes in Iowa are shaded by population. Darker colors indicate larger population while lighter colors indicate smaller populations.

The methods in this paper operate at the ZIP code level; each surveillance site is abstracted to the ZIP code in which it resides. Because the ZIP code is an integral part of a site's address, we can determine the location (i.e., latitude and longitude), and also population, without geocoding the address; we simply consult a look-up table. More fine-grained population data may certainly be used (e.g., block- or tract-level), but addresses must be geocoded in those cases to determine location and population. Our abstraction does not preclude network design in the case where multiple sites are located in the same ZIP code.

2.2.2 Algorithms

The web-based calculator supports three different network design algorithms: two algorithms based on the maximal coverage model and one based on the K-median facilities location model.

2.2.2.1 Maximal coverage model

The maximal coverage model (MCM) considers each site as having a fixed coverage radius. For example, given that surveillance sites are typically primary care facilities, it may be reasonable to assume that a site may serve patients who live within a 30-minute driving radius of the site (indeed, this is the radius of coverage we use in our simulations). The resulting optimization problem can be stated informally as follows: given a geographic population distribution and a radius of coverage for each site, we wish to choose the sites that maximize the total number of people within the specified distance of a site [38]. Because the problem is non-deterministic polynomial-time hard (NP-hard) to solve exactly (i.e., it is typically infeasible to compute the optimal solution), we instead implement a greedy approximation algorithm that provides a $(1 - \frac{1}{e})$-approximation of the optimal solution [44]. This approximation algorithm guarantees a rapid solution that is “close enough” to optimal for use in practice.

Note that the standard MCM formulation places no restrictions on the number of cases a site can serve (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 serve. Such site capacity constraints are explicitly modeled in the capacitated MCM formulation where each site is endowed with some intrinsic integer capacity. Each person inside the radius of a site S_i is then uniquely counted against that site's capacity. Once a site's capacity is exhausted, it may become appropriate to place another site S_j near S_i notwithstanding overlapping site radii. For example, using the standard non-capacitated MCM formulation, sites are preferentially placed in very dense urban areas, often with several hundred thousand people within a single site's coverage radius. The capacitated model would instead deploy multiple surveillance sites to high density locations in order to account for each site's intrinsically limited surveillance capacity.

Figure 2.2 shows how 19 sites chosen using the non-capacitated MCM compare against the 19 sites used by the IDPH.

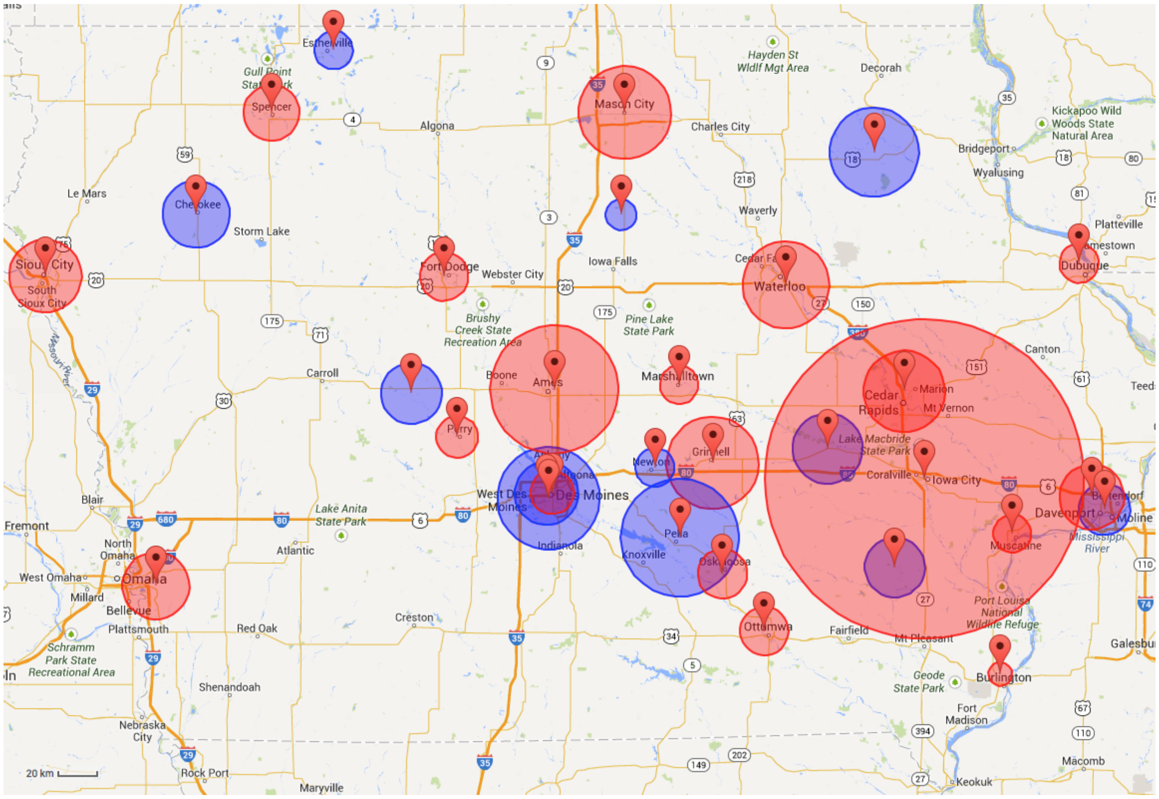


Figure 2.2: Map comparing 19 existing sites against 19 sites chosen using MCM-NC. The 19 existing sites (blue circles) and 19 sites calculated using the MCM-NC (red circles) are shown together. Circles around each marker indicate the average driving distance between patient homes and provider location in the Medicaid dataset. The large red circle around Iowa City represents the University of Iowa Hospitals and Clinics; the average driving distance is 45.47 miles (73.18 kilometers). The MCM-NC tends to choose sites in the more densely populated regions in Iowa. These sites often contain more reputable hospitals and clinics, and as a result, many people are willing to drive further distances to be seen at these locations. The existing network neglects certain populous regions of Iowa (such as Council Bluffs near Omaha) while potentially over-covering other regions (such as Des Moines). Although the Medicaid dataset is used to display average driving distances in this figure, recall that only population data are used to select sites for a network.

2.2.2.2 K-median model

The K-median model (sometimes also referred to as the P-median model, as in [50]) minimizes the sum of the distances from each individual to their nearest site (a more formal specification is found in [37]). Like the maximal coverage problem, the K-median problem is also NP-hard [102], so an approximation algorithm is once again in order. Here, we use a simple greedy algorithm, although there are more complicated approximation algorithms that can generate slightly better solutions (see, e.g., [172, 50]).

Note that there is a fundamental difference between the maximal coverage model and the K-median model: the K-median model has no explicit notion of population coverage; hence no radius of coverage is involved. By definition, every person in the selected geography is "covered", although the "quality" of his or her coverage (in terms of travel distance) will vary (for this reason, our web-based calculator always claims 100% coverage when sites are placed using the K-median model).

2.2.3 Validation

We can evaluate these different methods 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. Because our simulations are based on a historical record of actual influenza-related cases, we can make meaningful comparisons between the performance of algorithmically-derived surveillance networks and the existing IDPH network.

2.2.3.1 Medicaid dataset

We use a dataset consisting of two million de-identified Medicaid billing records representing eight complete influenza seasons from July 2000 to June 2008. Medicaid is a U.S. federal health insurance program for people and families with low incomes. These records comprise *all* of the Iowa Medicaid records from this time period that contain any one of 30 pre-specified ICD-9 codes that have been previously associated with influenza [124]. Note that we use ICD-9-coded data as a proxy measure for influenza activity because laboratory-based influenza were not available for the state of Iowa. A look at a seven-day moving average graph of the dataset in Figure 2.3 clearly shows the well-established seasonal influenza peak that occurs each winter [117].

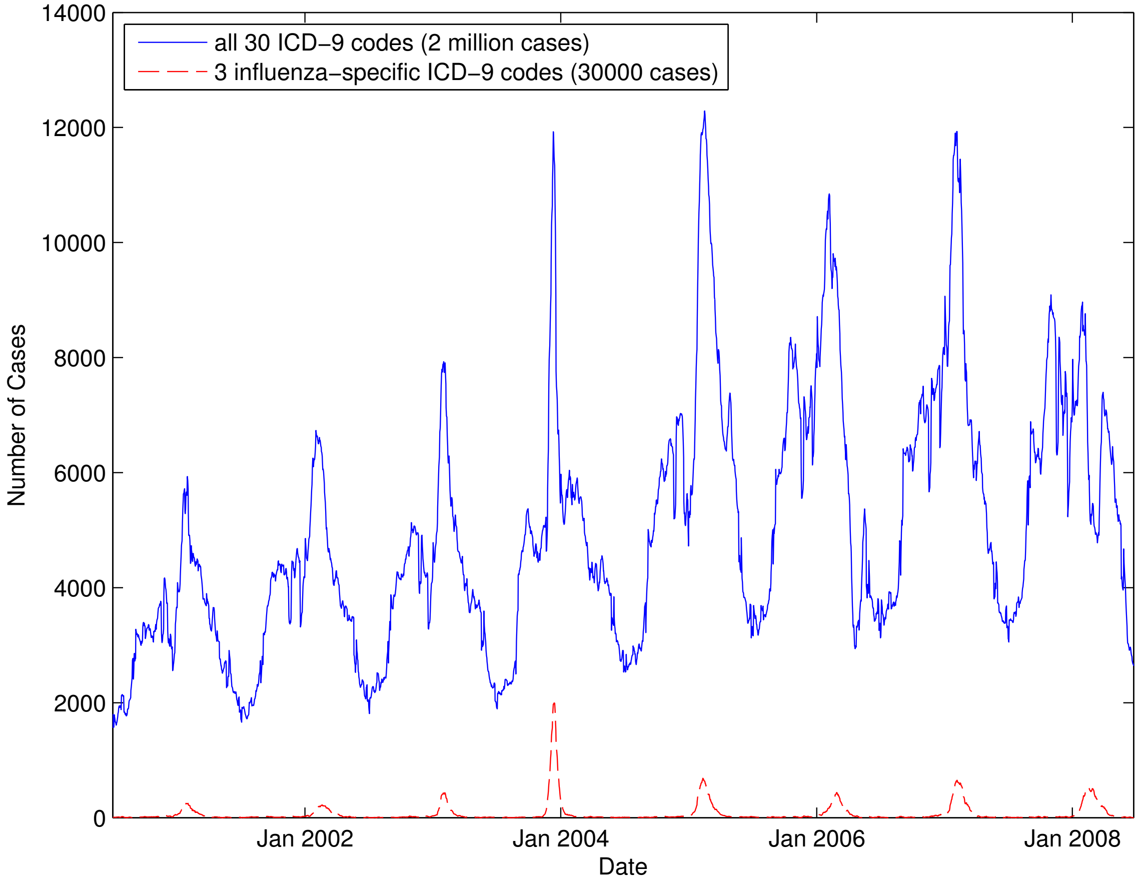


Figure 2.3: Seven-day moving average graph showing the number of cases present in the Medicaid dataset over time. The top line aggregates cases for all 30 ICD-9 codes while the bottom line aggregates cases for three influenza-specific ICD-9 codes (i.e., 487.x). Note that, when using all 30 ICD-9 codes, the case count never goes to zero.

Each record consists of an anonymous unique patient identifier, the ICD-9 diagnosis billing code, the date the case was recorded, the claim type (I=inpatient, O=outpatient, M=medical), the patient ZIP code, age, gender, and provider ZIP code (see Table 2.2). The dataset is very complete; of the two million total entries, only 2500 entries are dropped due to an erroneous or missing field (e.g., a patient ZIP code of 99999). A second influenza-specific subset of the original data can be defined by selecting only three of the original 30 ICD-9 codes that diagnose laboratory-verified influenza (i.e., 487 — influenza, 487.1 — influenza with other respiratory manifestations, and 487.8 — influenza with other manifestations); these three ICD-9 codes constitute approximately 30,000 entries, or about 4,000 per year. When all 30 ICD-9 codes are considered, the disease seems to never disappear (Figure 2.3); even during the summer, there are several thousand cases. This might be attributed to the fact that many of the 30 ICD-9 codes present in our expanded dataset include codes that represent diseases and symptoms seen year-round (e.g., cough and acute nasopharyngitis).

Table 2.2: Ten sample entries in the Medicaid billing dataset.

id	**ICD-9**	**date**	**claimtype**	**patient_zip**	**age**	**gender**	**provider_zip**
2421392	466	12/5/2005	O	50315	43	F	50314
2421392	465.9	1/23/2006	O	50315	43	F	50314
2421392	465.9	2/2/2006	O	50315	43	F	50314
2421392	465.9	3/9/2006	O	50315	43	F	50314
2421392	465.9	11/7/2006	O	50315	44	F	50314
1406011	780.6	11/30/2000	O	50316	37	F	50316
1979061	462	5/16/2001	M	50309	59	F	50315
425531	466	2/2/2004	M	50317	32	F	50316
425531	466	2/12/2004	M	50317	32	F	50313
425531	465.9	8/11/2004	M	50317	32	F	50313

The current diagnosis billing code standard is ICD-10, which provides for more diagnostic granularity than ICD-9. Although our data do not use this new standard, no significant changes would need to be made to the methods used in this paper for validation; only careful selection of ICD-10 codes that correspond to cases of interest is required.

2.2.3.2 Simulation

We treat the Medicaid dataset as a proxy of the record of all ILI 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 consumer behavior [92]; here, we use the Huff model to determine where people might seek care based on distance to the provider and the provider's perceived "attractiveness". More formally, the probability H_{ij} that case i is detected by surveillance site j is given by

$$H_{ij} = \frac{A_j^{\alpha} D_{ij}^{-\beta}}{\sum_{j=1}^{n} A_j^{\alpha} D_{ij}^{-\beta}}, \tag{2.1}$$

where A_j is the attractiveness of site j, D_{ij} is the distance from case i to site j, α is the attractiveness enhancement parameter, β is the distance decay parameter, and n is the total number of surveillance sites.

We use the Huff model because it gives us a way of balancing the "attractiveness" of a site against the distance a patient may be from the site. Although we could use the great-circle distance formula (i.e., geodesic distance on the surface of a sphere) to approximate road distance [19], we instead created a driving distance matrix using Microsoft's Bing Maps API so that our measurements of travel time are as accurate as possible. D_{ij} is measured as driving distance in miles.

The challenge of properly setting appropriate values for the attractiveness, attractiveness enhancement parameter, and distance decay parameter remains. One solution, and the one adopted in this work, is to estimate the attractiveness of a site from the number of cases seen at that site in the Medicaid dataset. Since we have a comprehensive set of Medicaid cases on which we use the Huff model, we can fit appropriate values of α and β from the dataset. Although a number of parameter estimation methods have been proposed (e.g., [16, 86, 90, 87, 138]), we present a method that uses a metaheuristic global optimization algorithm called harmony search (HS) [78] to determine the two parameters. HS has been applied to a variety of problems, including other parameter estimation problems, and it often outperforms other commonly used search algorithms, such as simulated annealing, tabu search, and evolutionary algorithms (e.g., [74, 122, 145, 104, 181, 75, 77, 76]). We treat our parameter estimation problem as a maximization problem, where the goal is to select values of α and β that produce the maximal average number of Medicaid cases "correctly" located using the Huff model; a case is "correctly" located if a number selected at random in the range $[0, 1)$ is less than the Huff probability, H_{ij}. Case count is averaged over 50 replicates.

We use an open source Python implementation of HS called pyHarmonySearch.[2] α and β are both allowed to vary in the range (0, 20]. We set `max_imp` to 100, `hms` to 20, `hmcr` to 0.75, `par` to 0.5, and `mpap` to 0.25. We ran a total of 20 HS iterations. For the full dataset, the best solution gave us a fitness of 1,032,762.2 cases correctly detected (out of two million total cases) with $\alpha = 17.998$ and $\beta = 19.769$. For the influenza-specific dataset, the best solution had a fitness of 15,141.14 cases

[2] `https://github.com/gfairchild/pyHarmonySearch`

(out of 30,000 total cases) with $\alpha = 19.114$ and $\beta = 19.479$.

2.3 Results

We simulate influenza spread considering both the entire dataset and the influenza-specific dataset. Because our simulations are stochastic, results are produced by averaging over 50 replicates. Placement algorithms design networks by selecting sites from an IDPH-provided set of 117 candidate sites spread across the state of Iowa. In addition to the MCM and K-median location-allocation models, our analysis considers surveillance networks designed by selecting sites uniformly at random. Results are reported for each network size by averaging over 50 randomly generated networks.

2.3.1 Outbreak intensity

One way of comparing the quality of two different surveillance networks is to compare the accuracy of their respective measures of outbreak intensity, here the percentage of cases correctly detected by each network using the Huff model. In each graph, the performance of the existing IDPH-selected sites is shown as a single data point at $n = 19$. As seen in Figures 2.4 and 2.5, sites generated by the capacitated and non-capacitated MCM (MCM-C and MCM-NC, respectively) tend to perform best, followed closely by the K-median model. Performance improves as network size grows. Unsurprisingly, selecting sites uniformly at random results in worse outbreak intensity detection than preferentially selecting sites.

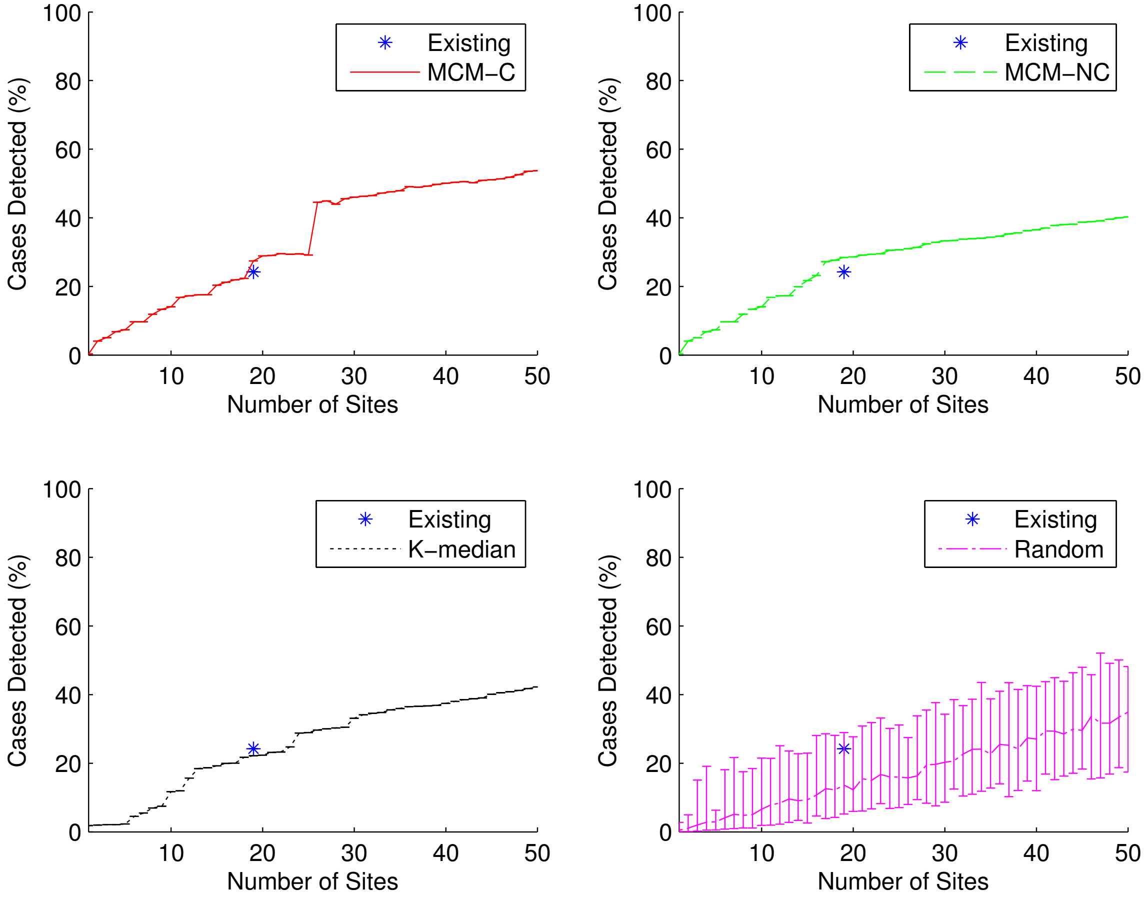


Figure 2.4: Outbreak intensity as a function of network size for all ICD-9 codes. Outbreak intensity is shown for the existing sites (shown as a single data point at $n = 19$) as well as for sites generated using the two MCM variants, K-median model, and randomly selected sites considering all 30 ICD-9 codes. MCM-C is capacitated MCM, and MCM-NC is non-capacitated MCM. Random results at each network size were computed by selecting 50 networks uniformly at random. In all cases, because our simulations are stochastic, results were computed using 50 replicates. Graphs show average results as well as minimum/maximum error bars. In general, the two MCM variants perform best, with the K-median model trailing closely. The two MCM models outperform the existing sites. The existing sites detect approximately 24.2% (±0.02%) of all cases, while it takes only 17 sites for the MCM-NC to accomplish the same level of outbreak intensity detection. At $n = 19$, MCM-NC detects 28.5% (±0.02%) of all cases. As the number of sites grows beyond 20 sites, the capacitated MCM better detects outbreak intensity.

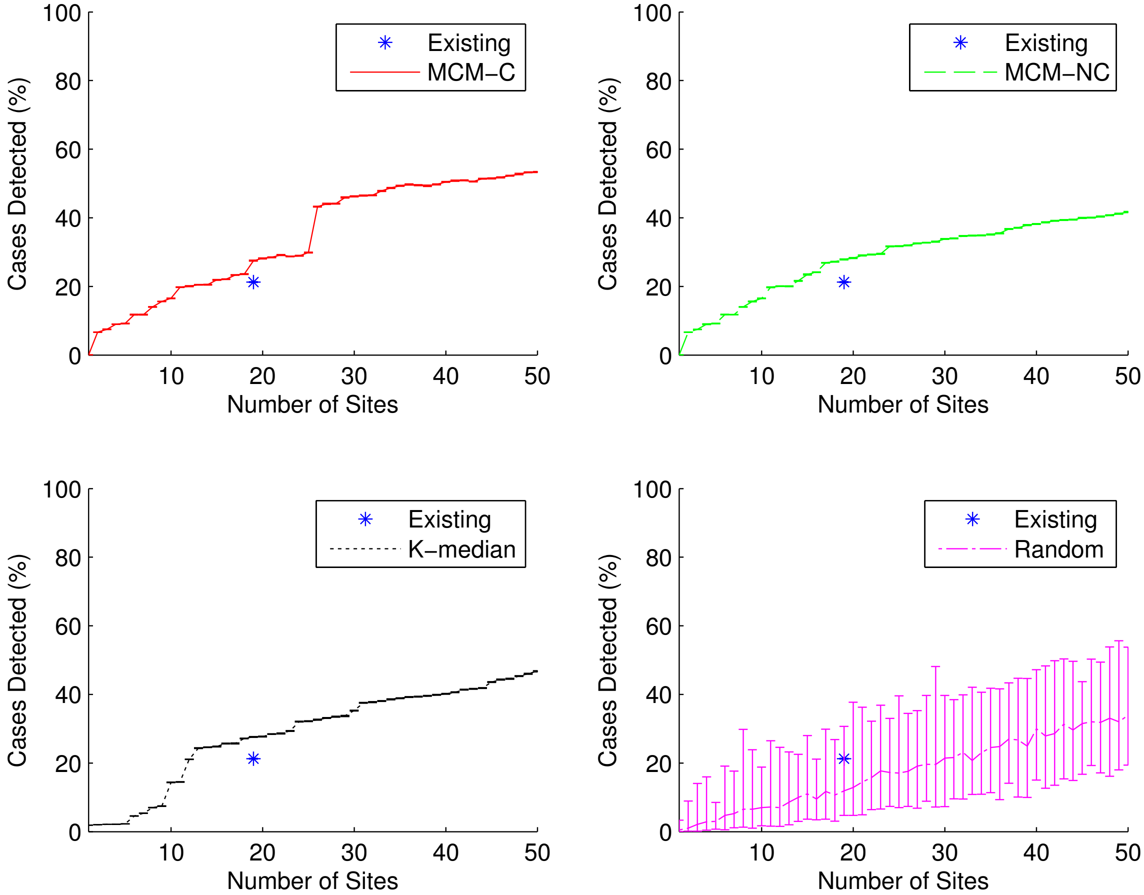


Figure 2.5: Outbreak intensity as a function of network size for the influenza-specific dataset. Outbreak intensity is shown for the existing sites (shown as a single data point at $n = 19$) as well as for sites generated using the two MCM variants, K-median model, and randomly selected sites for the three influenza-specific ICD-9 codes. Here, all three algorithmic variants outperform the existing sites. The existing sites detect approximately 21.2% (±0.1%) of all cases. It takes only 12 sites for the K-median to accomplish the same level of outbreak intensity detection. At $n = 19$, MCM-NC detects 27.9% (±0.2%) of all cases.

It seems particularly appropriate to consider the performance of networks of size 19, since this is the number of surveillance sites in the existing IDPH network. At $n = 19$ for the full dataset, we see that all methods, except K-median and random selection, outperform the existing network. As seen in Figure 2.4, the existing IDPH network detects approximately 24.2% (±0.02%) of all cases using the full dataset. At $n = 19$, MCM-C detects approximately 27.4% (±0.02%) of cases, MCM-NC detects approximately 28.5% (±0.02%), K-median detects approximately 22.2% (±0.01%), while a random network detects 13.7% of cases on average (5.2% lower bound, 28.9% upper bound). MCM-NC is capable of more efficient detection than the existing network with only 17 sites. For the influenza-specific dataset, as seen in Figure 2.5, all three algorithmic site placement methods outperform the existing sites. Here, it only takes 12 sites selected using the K-median model to match the outbreak intensity detection of the existing sites. In other words, in the state of Iowa, a network can be designed that detects outbreak intensity as well as the existing network with two fewer sites when considering the full gamut of possible influenza-related ICD-9 codes. However, if we only consider direct diagnoses of influenza, the network can consist of 37% fewer sites. This practically significant result indicates that preferentially selecting sites can yield more efficient surveillance networks with less overhead cost.

2.3.2 Outbreak timing

In addition to outbreak intensity, a sentinel surveillance network should be able to detect outbreak timing, or the temporal start, peak, and end of a disease season. Intuitively, when attempting to maximize outbreak intensity detection (as well as outbreak location detection), increasing the number of surveillance sites will improve the quality of detection. However, it is not clear that there is an inherent benefit of having more sites when looking at outbreak timing. We would like to explore just how few sites are necessary in order to still accurately detect the timing of a disease season.

A surveillance network will necessarily detect fewer cases than actually occurred among a population; yet, if the surveillance network detects cases temporally in sync with this ground truth, then the disease curve should increase and decrease in proportion with it. We use the Pearson product-moment correlation coefficient (often abbreviated Pearson's r) to correlate each detected time series with the ground truth dataset in order to quantify outbreak timing detection quality [153]. Correlation coefficients range from -1 to 1. Values above 0.5 and below -0.5 are often interpreted to indicate strong positive and negative correlation, respectively, although these limits are not hard and greatly depend on the context [40]. This method for measuring outbreak timing does not require that we explicitly define the start, peak, or end of a disease season; we simply correlate the simulated disease curves with the ground truth disease curves.

Figures 2.6 and 2.7 compare the outbreak timing detection capabilities of the algorithmic placement methods and the existing sites using the full dataset and influenza-specific dataset, respectively. In Figure 2.6, at $n = 19$, we see similar outbreak timing performance among all placement methods, with all networks achieving correlation coefficients of at least 0.98 (indicating very strong positive correlation with ground truth). It only takes 11 algorithmically-placed sites in order to detect outbreak timing at least as well as the existing network, while a network containing only two well-placed sites is capable of achieving a 0.9 correlation coefficient. Even networks with as few as one site are able to achieve correlations of at least 0.67. When the set of ICD-9 codes is restricted to the influenza-specific dataset, as in Figure 2.7, outbreak timing quality is only slightly reduced. It takes 14 sites to match the performance of the existing network, but it only takes six sites to achieve correlation of at least 0.9. These practically significant findings suggest that it may be possible to drastically reduce the size of a network if the metric of primary interest is outbreak timing detection.

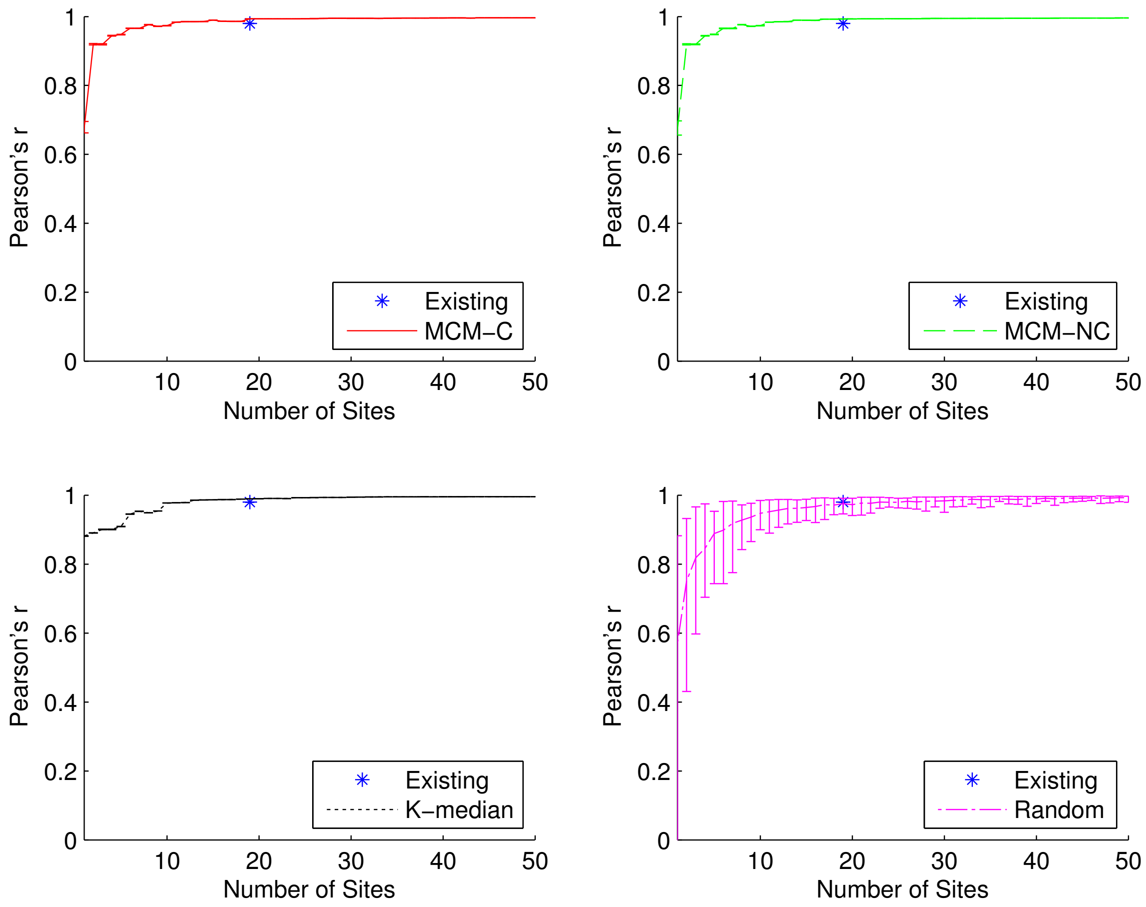


Figure 2.6: Outbreak timing as a function of network size for all ICD-9 codes. Pearson's r for the existing sites as well as for sites generated using the two MCM variants, K-median model, and randomly selected sites for all 30 ICD-9 codes. Recall that Pearson's r is used to quantify the quality of outbreak timing detection for a surveillance network. At $n = 19$, the average Pearson's r for the existing sites is 0.98 (±0.0002), while it is 0.99 (±0.0002) for MCM-C, MCM-NC, and K-median, respectively, and 0.97 (±0.02) for random. All algorithmic methods offer site placements that perform extremely well, achieving at least 0.67 correlation for networks with as few as one site. It only requires two sites selected using either MCM variant to achieve correlation of at least 0.9, and only 11 sites are needed to match the performance of the existing network.

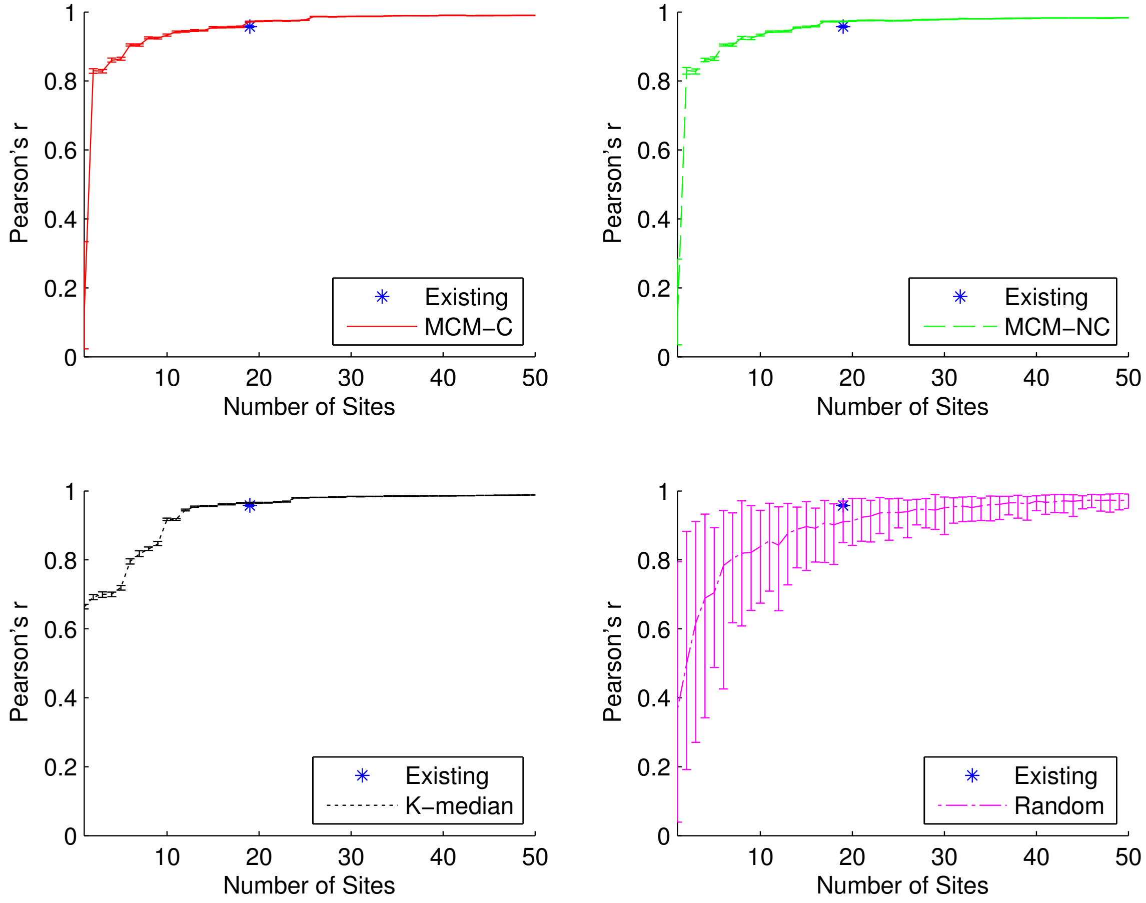


Figure 2.7: Outbreak timing as a function of network size for the influenza-specific dataset. Pearson's r for the existing sites as well as for sites generated using the two MCM variants, K-median model, and randomly selected sites for the three influenza-specific ICD-9 codes. At $n = 19$, the average Pearson's r for the existing sites is 0.96 (±0.01), while it is 0.97 (±0.001) for all three algorithmic placement methods and 0.91 (±0.06) for randomly placed sites. Compared to Figure 2.6, there is a very small reduction in outbreak timing detection capabilities when the ICD-9 codes are restricted. Both MCM variants are capable of 0.9 correlation with as few as six sites, and only 14 sites are required to match the performance of the existing network.

2.4 Conclusions

Disease surveillance is critical in epidemiological studies and in the realm of public health policy. Using a publicly available web-based surveillance site placement calculator and three different algorithmic surveillance site placement methods, we compared the performance of networks generated by the calculator with the volunteer-based network maintained by the IDPH.

The major contribution of this paper is the exploration of two metrics on which a surveillance network can be optimized: *outbreak intensity* and *outbreak timing*. Sites chosen using either MCM variant consistently outperform the baseline IDPH network both in terms of outbreak intensity and timing. Furthermore, we found that preferential selection of sites can yield networks capable of achieving outbreak intensity and timing performance in line with the current IDPH network, requiring, in some cases, only a fraction of the number of sites. We found that, at least in the state of Iowa, the number of sites chosen seems not to matter for outbreak timing detection. This implies that using just a few strategically placed surveillance sites (e.g., Des Moines, Cedar Rapids, Davenport, Sioux City, and Iowa City – the five most populous cities in Iowa) may suffice to reliably and accurately determine the onset, peak, and end of the influenza season in Iowa.

It is important to recognize that although we analyze and compare networks using a dataset of confirmed Medicaid influenza-related cases, network design is accomplished only considering population data. This means that our surveillance network design methods can be used in any location in the world where population data are available.

In practice, surveillance site recruitment, especially in locations where such involvement is voluntary, may prove difficult. This realization opens a new dimension for optimization: cost. Each site brings some inherent cost to the system; the cost may be a monetary value (e.g., incentives), man-hours required for reporting, or some other measure. That is, the real-world optimization problem may actually need to be multi-dimensional. For example, the maximal coverage model may need to be minimal cost, maximal coverage in practice. This direction for future work requires careful consideration when deriving site costs. Additionally, in areas where surveillance site participation is voluntary, a site selected by the methods presented in this paper may decline or hesitate to join the network. The greedy algorithms used here allow for public health officials to rank site importance since, by definition, the most important sites are selected first. This can allow for an adjustment in resource allocation to incentivize important, but unwilling, sites.

In the future, we will look more closely at the problem of selecting the ICD-9 codes worth considering for validation. Here, we only consider two sets of ICD-9 codes: the entire set of all 30 influenza-related ICD-9 codes provided in our Medicaid dataset and an influenza-specific ICD-9 code subset containing only direct diagnoses of influenza (i.e., 487.x ICD-9 codes). One possible approach is to apply machine learning techniques typically used for feature selection to the problem of finding which ICD-9 codes should be used for validation. We will also examine other states exhibiting different population density and geographic characteristics from Iowa, and, eventually, nationwide and worldwide surveillance networks. Ultimately, our goal is to use computational methods to reliably advise public health officials how many surveillance sites suffice and where to place them in order to meet their specific needs.

There are several limitations of our work. First, this work focuses on the state of Iowa in the U.S., which is a relatively simple state geographically and geologically. A more geographically or geologically diverse state such as Colorado with its natural east-west divide in the form of the Rocky Mountains may provide different obstacles in site placement. Second, our placement models ignore demographics, so it is possible the resulting networks are sampling some demographics more than others or possibly missing some demographics altogether. All surveillance networks have difficulty making conclusions about uncovered areas. Third, the Medicaid data used in our simulations represent a particular demographic of Iowa: people and families with low incomes (these data, however, are complete with respect to that particular demographic). Furthermore, all calculations consider the population of a ZIP code to be concentrated at the centroid of that ZIP code. In reality, populations are usually distributed in some fashion across the entire ZIP code region. Additionally, while our simplifying site-as-ZIP code abstraction may be reasonable for less densely populated regions, such as Iowa, it may prove to be problematic in more densely populated regions. A final limitation to our work is that we use administrative data (ICD-9 codes) as a proxy for influenza activity. We would rather use actual ILI data or laboratory-based data, but these data sources were not available across the state.

Our web-based tool can aid public health officials in designing an effective disease surveillance system. We studied two metrics by which a surveillance network may be evaluated: outbreak intensity and outbreak timing. By simulating the spread of influenza across the state of Iowa, we show that the sites our tool selects perform better than the status quo in terms of both metrics. Additionally, we offer new insights that suggest that network size may only play a minimal role in outbreak timing detection. Finally, we show that it may be possible to reduce the size of a surveillance system without affecting the quality of surveillance information the system is able to produce.

2.5 Extension: Primary stroke center placement

The site placement methods described so far have only been used for influenza sentinel surveillance networks. The methods are generic, though, and can be applied to a variety of health-related facilities location problems. This section demonstrates how the maximal coverage model can be used to place primary stroke centers.

2.5.1 Introduction

Outcomes in ischemic stroke depend on both the timely delivery of thrombolytic therapy [111] and the quality of the ancillary care provided [169]. A hospital that can provide adequate stabilization and treatment with rtPA, with or without a dedicated stroke unit, may be certified as a *primary stroke center* (PSC). This certification is a bottom-up voluntary process initiated by interested hospitals. Unfortunately, this self-initiated approach results in insufficient access to PSCs for a large segment of the American population [8]. This is particularly dramatic for non-urban areas [13]. Clearly, additional PSCs are needed to expand the timely access to emergent stroke care for a more reasonable fraction of the population [154]. However, before this shortage can be addressed, it is important to recognize the magnitude of the associated societal cost and any possible approaches that can be made to minimize the cost. Should the process of further PSC certification continue in this bottom-up self-initiated manner, or should an overseeing entity promote and/or direct the placement of additional PSCs? A top-down directed process could be justified if the number of additional centers required is large, or if the directed process significantly increases coverage over the self-initiated process. In this section, we aim to address this question by comparing present and projected PSC population coverage in a traditionally rural state with dispersed population and resources using geographic location-allocation modeling techniques.

2.5.2 Methods

We first identified all the current PSCs in the state of Iowa. Using the 2009 American Hospital Association's Annual Service Database, we determined all the possible additional PSC locations by identifying hospitals meeting the minimum requirements: the availability of a radiology department with staff available to perform head CT, and laboratory services that were available 24 hours a day [7, 114]. We then used the same web-based location-allocation calculator described earlier to place PSCs in Iowa.

We aimed to maximize the number of residents in the state that are within a fixed pre-specified time-distance of at least one of the possible PSC sites. Using our Bing driving distance matrix, which also contains driving time in seconds between ZCTAs, the calculator was used to estimate the current PSC population coverage in the state of Iowa with three different maximum time-distance thresholds (15, 30, and 45 minutes). We then used the calculator to estimate what the hypothetical coverage would be if a MCM had been used to establish the best location for the current number of centers. We also plotted the number of additional PSCs that would be needed to improve the population coverage to a pre-specified threshold. We compared the future improvement in coverage between four different approaches: 1) a random selection approach (i.e., new PSC sites selected uniformly at random), 2) a weighted-random selection approach that mimics the current tendency of PSCs that favor larger hospitals (i.e., sites are selected randomly where the probability of selection is weighted by the population contained within time-distance units of a site independent of the presence of other PSCs), 3) a MCM that builds on the existing 12 PSCs, and 4) a hypothetical MCM that was started *de novo* (i.e., without the existing PSCs).

2.5.3 Results

Of the 126 hospitals in Iowa, 120 reported having the minimum resources required to become a PSC and were included in the analysis. Of those 120 hospitals, 12 (10%) are already PSCs certified through the self-initiation process. The 12 current PSCs, which are located in the largest cities within the state, serve 37.2% of the Iowa population, assuming a time-distance radius of 30 minutes (these same 12 sites cover 21.3% and 60.0% of the population for time-distance radii of 15 and 45 minutes, respectively). The amount of effort that would be required to significantly expand the current coverage of PSCs in Iowa using the MCM is illustrated in Figure 2.8. 31 additional MCM-placed PSCs (dark gray) would be required to augment the 12 existing PSCs (green) in order to cover 75% of the population of Iowa with an assumed 30 minute time-distance radius of coverage. Figures 2.9a, 2.9b, and 2.9c show the relationship between the number of additional MCM-placed PSCs needed and the fraction of the population that would be covered for three different time-distance radii of coverage (15, 30, and 45 minutes, respectively). In these three graphs, calculations were plotted for the four different approaches discussed above. To generate curves for the two random approaches, we averaged the results from 500 replicates.

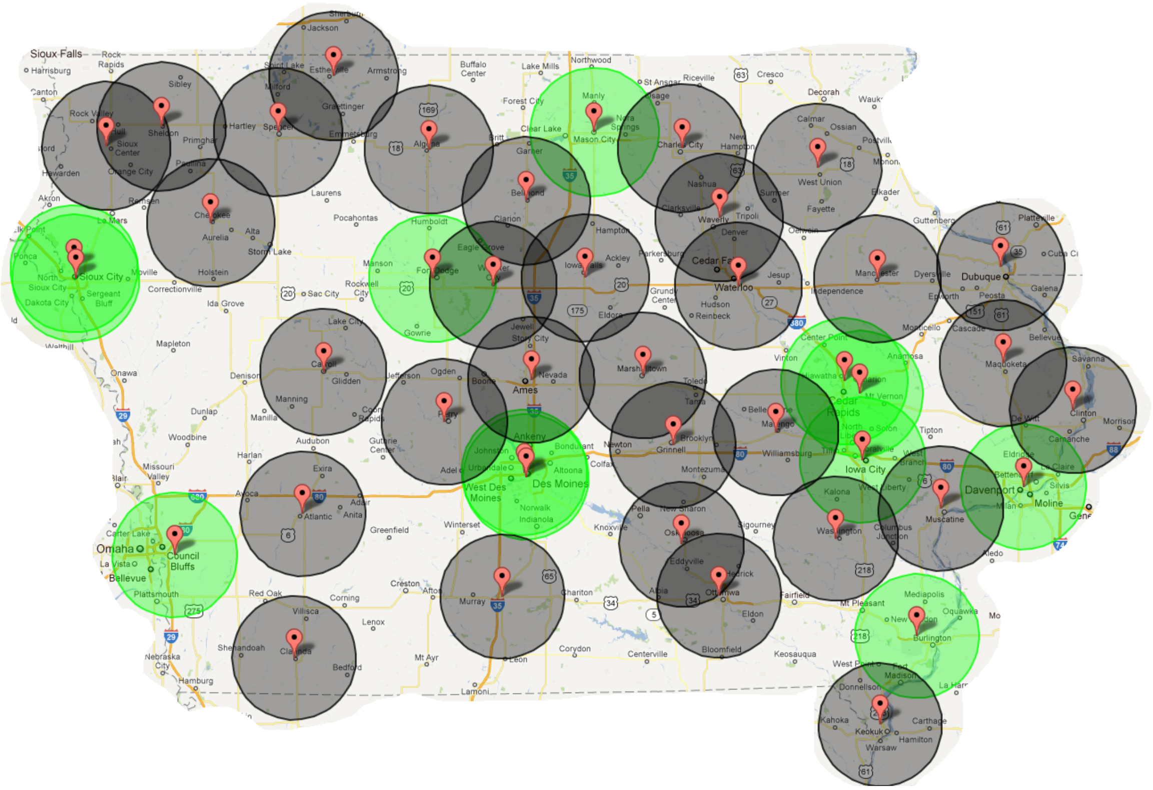
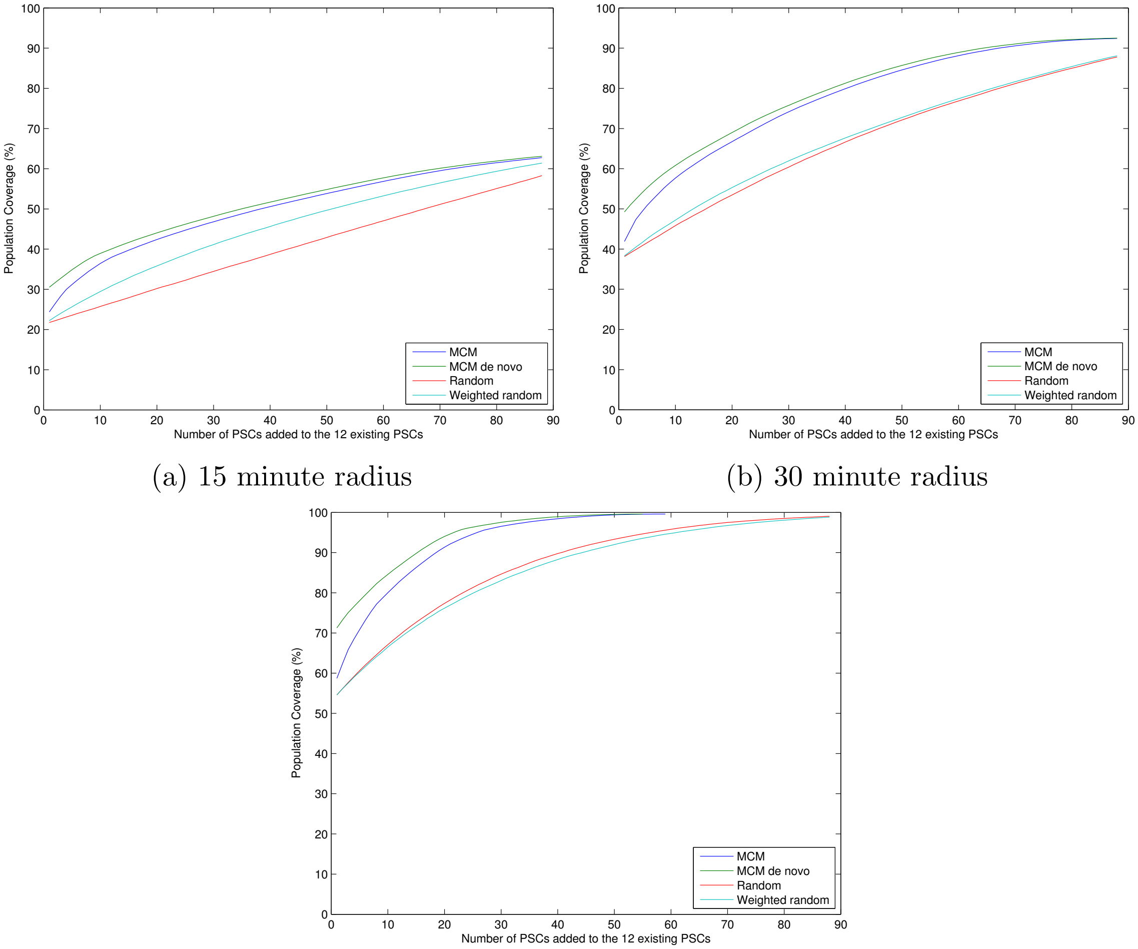


Figure 2.8: Results of a maximum coverage algorithm showing the location of the additional 31 PSCs (dark gray circles) that would be required to cover 75% of the population of Iowa with an assumed 30-minute time-distance radius. The existing 12 PSCs are shown in green circles. Circles with 20-mile radii are shown around each PSC for visualization purposes.



(a) 15 minute radius

(b) 30 minute radius

(c) 45 minute radius

Figure 2.9: Projected population coverage as new PSCs are added to the 12 existing PSCs. The random selection model selects sites uniformly at random. The weighted random selection model simulates the current self-initiation approach by randomly selecting PSCs with probability proportional to the population contained in each site's radius of coverage. Results garnered from both random selection models were averaged from 500 replicates. The MCM line represents a maximal coverage model (MCM) that builds on the existing 12 PSCs. The MCM de novo line represents a hypothetical MCM that was started *de novo* (i.e., without the existing PSCs). Results are shown for 3 different time-distance radii (15, 30, and 45 minutes).

The lack of efficiency of the current self-initiated approach is clearly evidenced by showing that 47.5% of the population could be covered presently within a 30 minute radius had a MCM been used to place the first 12 PSCs. Furthermore, future prediction of coverage using the weighted random selection model to simulate self-initiation in denser population regions shows that 54 additional PSCs would be needed to cover 75% of the population with a 30 minute radius of coverage while only an additional 31 MCM-placed PSCs would be needed to achieve the same degree of population coverage (Figures 2.8 and 2.9b).

2.5.4 Conclusions

The initiative to certify hospitals as PSC by the Joint Commission or by state health departments is an important step in improving stroke care for Americans [4]. Unfortunately, the current system of self-initiation by willing hospitals results in sparse coverage for large areas of the population, particularly in rural states. We have confirmed that the overall percentage of the state population adequately covered by PSCs is minimal, which highlights the critical need to increase the numbers of available centers. While reasonable coverage exists in urban centers of Iowa, the majority of the state's less-densely populated areas have insufficient availability. This reflects the national tendency for larger urban hospitals, typically with an attached stroke unit, to seek PSC certification from the Joint Commission[3]. This trend is at odds with the initial intentions of the PSC initiative to certify hospitals that are able to adequately stabilize and initially treat stroke patients similar to the trauma model [7]. It might also reflect a tendency for PSCs to cluster in a single location

[3] http://www.jointcommission.org/certification/primary_stroke_centers.aspx

due to local competition, a phenomenon described for other critical facilities such as ICUs [88]. The geographic disparity in access to stroke care for a large proportion of the U.S. population can only increase the existing disparity of stroke care between rural and urban areas [113]. Unfortunately, it is clear that the number of PSCs needed to improve the coverage to acceptable levels is quite large.

We recognize that our findings may, at first glance, give the impression of a self-fulfilling prophecy. After all, it should not be a surprise that a MCM produces better results than random selection. But the importance of these findings lies in quantifying the benefit over the current self-initiated approach. We were surprised to learn how inefficient the current self-initiation model of PSCs is to provide population coverage. We realize that in practice, the process of self-initiation is not entirely random. Larger hospitals providing care to denser areas of populations are more likely to seek PSC certification, so our weighted-random approach was designed to test this real-case scenario. Still, the difference with a MCM is significant. The existing PSCs in the state of Iowa are not in the most efficient locations to best serve the population. Since this certification has already been accomplished (and we are by no means advocating the removal of the PSC status for any existing center), we present these findings to improve PSC growth in the future. The ratio between minimal coverage and expenses required should be established based on the available resources. But is clear that no matter what the parameters are determined to be, a MCM would result in fewer required PSCs, and, as a result, lower societal investment. In other words, as in any other resource-limited situation, it is crucial to maximize coverage while minimizing the resources required.

These results raise the ethical and political question of whether the location of future PSCs should be regulated given the important health implications of receiving timely acute stroke care. We believe these findings will be useful to both government-run health care systems and private hospital systems. Governmental operations could stratify their resources to mandate the establishment of PSCs in hospitals located in the most efficient sites. But this would also be true for privately-owned hospital networks that seek to find the ideal location for a PSC within their system based on population distribution. We recognize that a MCM approach would be more difficult to enforce in the private market of small autonomous institutions. In those cases, rather than mandating a location, a MCM could be used by a state institution to decide how many additional centers are needed and in which hospitals they should concentrate educational efforts and incentives to promote PSC certification [143]. A MCM would be the basis for a rational justification by location that can be used to incentivize the process in smaller institutions. This approach might also be useful for other acute stroke care applications such as to find the best locations for remote centers for telemedicine networks by identifying rtPA-ready hospitals as key steps in the regionalization process [129], or identifying the best centers for a spoke-and-hub comprehensive stroke center network [45]. Geographic computerized methods that find the best location to improve access have been proposed for other services such as nephrology services [15] or flu surveillance [157, 163, 57].

We recognize that there might be resistance to further increase in the number of PSCs. Barriers to this process have already been identified [150]. Because of initial self-selection by larger hospitals, any further expansion of the number of PSCs in rural states becomes the burden of the remaining smaller hospitals. Being a small hospital is not per se an impediment for a PSC designation [167]. In fact, most small rural hospitals have adequate resources to become PSCs, suggesting that such certification could be achieved with adequate administrative and financial support [114]. However, it might be difficult to interest additional smaller hospitals in pursuing PSC certification, particularly when they did not volunteer for the process in the years since the PSC initiative has been available. We recognize that becoming a PSC is an expensive and onerous process for small hospitals. For that reason, we propose a central planning incentive to subsidize these costs for critical locations. Alternatively, in cases where optimal locations prove to be infeasible for financial or other reasons, the same methodology can be used to identify second-best locations and the cost incurred in selecting the second best. We hypothesize that institutions in critical uncovered geographical locations might respond to statewide or national incentives to become a PSC.

There are limitations to this research. We recognize that the choice of using PSCs might underestimate the state stroke treatment capabilities because there are hospitals that can give rtPA without being certified as a PSC. We used the current PSC certification to test the optimization model since it is a rigorous initiative widely accepted by the stroke community nationwide as a standard of acute stroke care. Also, there could be uncertainty about what constitutes a “state-wide significant" difference in population coverage. This is analogous to the dilemma between statistical significance and clinical significance. Currently, 10% of the population of the State of Iowa (3,046,355 according to the U.S. Census Bureau 2010) represents 304,635 persons. That means that, if the current PSCs had been placed using a MCM, an additional 300 thousand individuals would have adequate PSC coverage. Again, that is without any additional societal investment, only by locating the PSCs at the most efficient locations. We judge this number to be significant and meaningful for stroke care at a state-wide level. Another limitation is the predictive nature of the model. We have minimized that uncertainty by testing a weighted-random approach that approximates the current process to become a PSC.

There are also operational and implementation limitations to this research. Because ZCTA boundaries are not ZIP code boundaries, we assume that the population is concentrated at the center of the ZCTA. However, although this limitation could affect the magnitude of the effect, is does not affect the outcomes of this model nor the conclusions. We also recognize that there are potential limitations to generalize these results to other states that may have different patterns of population density and resource dispersion; we have used for this research a state that is a moderate example of dispersion for a rural population within the United States.

In summary, the process of becoming a PSC has been one driven by self-initiation. But this approach is in sharp contrast to the systematic planning that normally occurs for other important societal services and has resulted in insufficient access to stroke centers for large segments of the U.S. population. Expanding the number of PSCs is clearly necessary, and such an expansion will likely benefit from the efficiency of directed optimization models. Lastly, and more importantly, such methods would result in minimization of societal investment while assuring some measure of maximal utility.

2.6 Publications

This chapter is derived from two publications:

- G. Fairchild, P. M. Polgreen, E. Foster, G. Rushton, and A. M. Segre, “How many suffice? A computational framework for sizing sentinel surveillance networks,” Int. J. Health Geogr., vol. 12, no. 1, p. 56, Dec. 2013.
- E. C. Leira, G. Fairchild, A. M. Segre, G. Rushton, M. T. Froehler, and P. M. Polgreen, “Primary Stroke Centers Should Be Located Using Maximal Coverage Models for Optimal Access,” Stroke, vol. 43, no. 9, pp. 2417–2422, 2012.

CHAPTER 3
GLOBAL DISEASE MONITORING AND FORECASTING WITH WIKIPEDIA

Abstract

Infectious disease is a leading threat to public health, economic stability, and other key social structures. Efforts to mitigate these impacts depend on accurate and timely monitoring to measure the risk and progress of disease. Traditional, biologically-focused monitoring techniques are accurate but costly and slow; in response, new techniques based on social internet data such as social media and search queries are emerging. These efforts are promising, but important challenges in the areas of scientific peer review, breadth of diseases and countries, and forecasting hamper their operational usefulness.

We examine a freely available, open data source for this use: access logs from the online encyclopedia Wikipedia. Using linear models, language as a proxy for location, and a systematic yet simple article selection procedure, we tested 14 location-disease combinations and demonstrate that these data feasibly support an approach that overcomes these challenges. Specifically, our proof-of-concept yields models with r^2 up to 0.92, forecasting value up to the 28 days tested, and several pairs of models similar enough to suggest that transferring models from one location to another without re-training is feasible.

Based on these preliminary results, we close with a research agenda designed to overcome these challenges and produce a disease monitoring and forecasting sys-

tem that is significantly more effective, robust, and globally comprehensive than the current state of the art.

3.1 Introduction

3.1.1 Motivation and Overview

Infectious disease remains extremely costly in both human and economic terms. For example, the majority of global child mortality is due to conditions such as acute respiratory infection, measles, diarrhea, malaria, and HIV/AIDS [118]. Even in developed countries, infectious disease has great impact; for example, each influenza season costs the United States between 3,000 and 49,000 lives [176] and an average of $87 billion in reduced economic output [136].

Effective and timely disease surveillance — that is, detecting, characterizing, and quantifying the incidence of disease — is a critical component of prevention and mitigation strategies that can save lives, reduce suffering, and minimize impact. Traditionally, such monitoring takes the form of patient interviews and/or laboratory tests followed by a bureaucratic reporting chain; while generally considered accurate, this process is costly and introduces a significant lag between observation and reporting.

These problems have motivated new surveillance techniques based upon internet data sources such as search queries and social media posts. Essentially, these methods use large-scale data mining techniques to identify health-related activity traces within the data streams, extract them, and transform them into some useful metric. The basic approach is to train a statistical estimation model against ground truth data, such as ministry of health disease incidence records, and then apply the model to generate estimates when the true data are not available, e.g., when forecasting or when the true data have not yet been published. This has proven effective and has spawned operational systems such as Google Flu Trends (`http://www.google.org/flutrends/`). However, four key challenges remain before internet-based disease surveillance models can be reliably integrated into an decision-making toolkit:

1. Models should afford review, replication, improvement, and deployment by third parties. This guarantees a high-quality scientific basis, continuity of operations, and broad applicability. These requirements imply that model algorithms — in the form of source code, not research papers — must be generally available, and they also imply that complete input data must be available. The latter is the key obstacle, as terms are dictated by the data owner rather than the data user; this motivated our exploration of Wikipedia access logs. To our knowledge, no models exist that use both open data and open algorithms.

2. Dozens of diseases in hundreds of countries have sufficient impact to merit surveillance; however, adapting a model from one *disease-location context* to another can be costly, and resources are often, if not usually, unavailable to do so. Thus, models should be cheaply adaptable, ideally by simply entering new incidence data for training. While most published models afford this flexibility in principle, few have been expressly tested for this purpose.

3. Many contexts have insufficient reliable incidence data to train a model (for example, the relevant ministry of health might not track the disease of interest), and in fact these are the contexts where new approaches are of the greatest urgency. Thus, trained models should be translatable to new contexts using alternate, non-incidence data such as a bilingual dictionary or census demographics. To our knowledge, no such models exist.

4. Effective disease response depends not only on the current state of an outbreak but also its future course. That is, models should provide not only estimates of the current state of the world — *nowcasts* — but also *forecasts* of its future state.

   While recent work in disease forecasting has made significant strides in accuracy, forecasting the future of an outbreak is still a complex affair that is sharply limited in contexts with insufficient data or insufficient understanding of the biological processes and parameters underpinning the outbreak. In these contexts, a simpler statistical approach based on leading indicators in internet data streams may improve forecast availability, quality, and time horizon. Prior evaluations of such approaches have yielded conflicting results and to our knowledge have not been performed at time granularity finer than one week.

In order to address these challenges, we propose a new approach based on freely available Wikipedia article access logs. In the current proof of concept, we use language as a proxy for location, but we hope that access data explicitly aggregated by geography will become available in the future. (Our implementation is available as open source software: `http://github.com/reidpr/quac`.) To demonstrate the feasibility of techniques built upon this data stream, we built linear models mapping daily access counts of encyclopedia articles to case counts for 7 diseases in 9 countries, for a total of 14 contexts. Even a simple article selection method was successful in 8 of the 14 contexts, yielding models with r^2 up to 0.89 in nowcasting and 0.92 in forecasting, with most of the successful contexts having forecast value up to the tested limit of 28 days. Specifically, we argue that approaches based on this data source can overcome the four challenges as follows:

C1. Anyone with relatively modest computing resources can download the complete Wikipedia dataset and keep it up to date. The data can also be freely shared with others.

C2. In cases where estimation is practical, our approach can be adapted to a new context by simply supplying a reliable incidence time series and selecting input articles. We demonstrate this by computing effective models for several different contexts even with a simple article selection procedure. Future, more powerful article selection procedures will increase the adaptability of the approach.

C3. In several instances, our models for the same disease in different locations are very similar; i.e., correlations between different language versions of the same article and the corresponding local disease incidence are similar. This suggests that simple techniques based on inter-language article mappings or other readily available data can be used to translate models from one context to another without re-training.

C4. Even our simple models show usefully high r^2 when forecasting a few days or weeks into the future. This suggests that the general approach can be used to build short-term forecasts with reasonably tight confidence intervals.

In short, this paper offers two key arguments. First, we evaluate the potential of an emerging data source, Wikipedia access logs, for global disease surveillance and forecasting in more detail than is previously available, and we argue that the openness and other properties of these data have important scientific and operational benefits. Second, using simple proof-of-concept experiments, we demonstrate that statistical techniques effective for estimating disease incidence using previous internet data are likely to also be effective using Wikipedia access logs.

We turn next to a more thorough discussion of prior work, both to set the stage for the current work as well as outline in greater detail the state of the art's relationship to the challenges above. Following that, we cover our methods and data sources, results, and a discussion of implications and future work.

3.1.2 Related Work

Our paper draws upon prior scholarly and practical work in three areas: traditional patient- and laboratory-based disease surveillance, Wikipedia-based measurement of the real world, and internet-based disease surveillance.

3.1.2.1 Traditional Disease Surveillance

Traditional forms of disease surveillance are based upon direct patient contact or biological tests taking place in clinics, hospitals, and laboratories. The majority of current systems rely on syndromic surveillance data (i.e., about symptoms) including clinical diagnoses, chief complaints, school and work absenteeism, illness-related 911 calls, and emergency room admissions [105].

For example, a well-established measure for influenza surveillance is the *fraction of patients with influenza-like illness*, abbreviated simply ILI. A network of outpatient providers report the total number of patients seen and the number who present with symptoms consistent with influenza that have no other identifiable cause [30]. Similarly, other electronic resources have emerged, such as the Electronic Surveillance System for the Early Notification of Community Based Epidemics (ESSENCE), based on real-time data from the Department of Defense Military Health System [20] and BioSense, based on data from the Department of Veterans Affairs, the Department of Defense, retail pharmacies, and Laboratory Corporation of America [18]. These systems are designed to facilitate early detection of disease outbreaks as well as response to harmful health effects, exposure to disease, or related hazardous conditions.

Clinical labs play a critical role in surveillance of infectious diseases. For example, the Laboratory Response Network (LRN), consisting of over 120 biological laboratories, provides active surveillance of a number of disease agents in humans ranging from mild (e.g., non-pathogenic *E. coli* and *Staphylococcus aureus*) to severe (e.g., Ebola and Marburg), based on clinical or environmental samples [105]. Other systems monitor non-traditional public health indicators such as school absenteeism rates, over-the-counter medication sales, 911 calls, veterinary data, and ambulance run data. For example, the Early Aberration Reporting System (EARS) provides national, state, and local health departments alternative detection approaches for syndromic surveillance [96].

The main value of these systems is their accuracy. However, they have a number of disadvantages, notably cost and timeliness: for example, each ILI datum requires a practitioner visit, and ILI data are published only after a delay of 1–2 weeks [30].

3.1.2.2 Wikipedia

Wikipedia is an online encyclopedia that has, since its founding in 2001, grown to contain approximately 30 million articles in 287 languages [185]. In recent years, it has consistently ranked as a top-10 website; as of this writing, it is the 6th most visited website in the world and the most visited site that is not a search engine or social network [9], serving roughly 850 million article requests per day [184]. For numerous search engine queries, a Wikipedia article is the top result.

Wikipedia contrasts with traditional encyclopedias on two key dimensions: it is free of charge to read, and anyone can make changes that are published immediately — review is performed by the community *after* publication. (This is true for the vast majority of articles. Particularly controversial articles, such as “George W. Bush” or “Abortion”, have varying levels of edit protection.) While this surprising inversion of the traditional review-publish cycle would seem to invite all manner of abuse and misinformation, Wikipedia has developed effective measures to deal with these problems and is of similar accuracy to traditional encyclopedias such as

Britannica [81].

Wikipedia article access logs have been used for a modest variety of research. The most common application is detection and measurement of popular news topics or events [5, 39, 91, 149, 10]. The data have also been used to study the dynamics of Wikipedia itself [158, 174, 180]. Social applications include evaluating toponym importance in order to make type size decisions for maps [23], measuring the flow of concepts across the world [177], and estimating the popularity of politicians and political parties [193]. Finally, economic applications include attempts to forecast movie ticket sales [128] and stock prices [135]. The latter two applications are of particular interest because they include a forecasting component, as the present work does.

In the context of health information, the most prominent research direction focuses on assessing the quality of Wikipedia as a health information source for the public, e.g., with respect to cancer [115, 160], carpal tunnel syndrome [120], drug information [106], and kidney conditions [175]. To our knowledge, only four health studies exist that make use of Wikipedia access logs. Tausczik et al. examined public “anxiety and information seeking” during the 2009 H1N1 pandemic, in part by measuring traffic to H1N1-related Wikipedia articles [171]. Laurent and Vickers evaluated Wikipedia article traffic for disease-related seasonality and in relation to news coverage of health issues, finding significant effects in both cases [110]. Aitken et al. found a correlation between drug sales and Wikipedia traffic for a selection of approximately 5,000 health-related articles [6]. None of these propose a time-series model mapping article traffic to disease metrics.

The fourth study is a recent article by McIver and Brownstein, which uses statistical techniques to estimate the influenza rate in the United States from Wikipedia access logs [125]. In the next section, we compare and contrast this article with the present work in the context of a broader discussion of such techniques.

In summary, use of Wikipedia access logs to measure real-world quantities is beginning to emerge, as is interest in Wikipedia for health purposes. However, to our knowledge, use of the encyclopedia for quantitative disease surveillance remains at the earliest stages.

3.1.2.3 Internet-based Disease Surveillance

Recently, new forms of surveillance based upon the social internet have emerged; these data streams are appealing in large part because of their real-time nature and the low cost of information extraction, properties complementary to traditional methods. The basic insight is that people leave traces of their online activity related to health observations, and these traces can be captured and used to derive actionable information. Two main classes of trace exist: *sharing* such as social media mentions of face mask use [134] and *health-seeking behavior* such as Web searches for health-related topics [82]. (In fact, there is evidence that the volume of internet-based health-seeking behavior dwarfs traditional avenues [161, 72].)

In this section, we focus on the surveillance work most closely related to our efforts, specifically, that which uses existing single-source internet data feeds to estimate some scalar disease-related metric. For example, we exclude from detailed analysis work that provides only alerts [41, 198], measures public perception of a disease [162], includes disease dynamics in its model [165], evaluates a third-party method [144], uses non-single-source data feeds [41, 73], or crowd-sources health-related data ("participatory disease surveillance") [34, 36]. We also focus on work that estimates biologically-rooted metrics. For example, we exclude metrics based on seasonality [14, 164] and over-the-counter drug sales volume, itself a proxy [116].

These activity traces are embedded in search queries [11, 21, 27, 31, 33, 42, 51, 53, 54, 56, 82, 85, 94, 95, 93, 99, 101, 142, 155, 156, 183, 187, 191, 190, 192, 194, 196, 197], social media messages [1, 2, 12, 22, 35, 47, 52, 83, 89, 103, 107, 109, 108, 137, 152, 166], and web server access logs [100, 125, 190]. At a basic level, traces are extracted by counting query strings, words or phrases, or web page URLs that are related to the metric of interest, forming a time series of occurrences for each item. A statistical model is then created that maps these input time series to a time series estimating the metric's changing value. This model is trained on time period(s) when both the internet data and the true metric values are available and then applied to estimate the metric value over time period(s) when it is not available, i.e., *forecasting* the future, *nowcasting* the present, and *anti-forecasting* the past (the latter two being useful in cases where true metric availability lags real time).

Typically, this model is linear, e.g.:

$$M = \sum_{j=1}^{J} \alpha_j x_j \tag{3.1}$$

where x_j is the count of some item, J is the total number of possible items (i.e., vocabulary size), M is the estimated metric value, and α_j are selected by linear regression or similar methods. When appropriately trained, these methods can be quite accurate; for example, many of the cited models can produce near real-time estimates of case counts with correlations upwards of $r = 0.95$.

The collection of disease surveillance work cited above has estimated incidence for a wide variety of infectious and non-infectious conditions: avian influenza [27], cancer [42], chicken pox [155], cholera [35], dengue [11, 31, 83], dysentery [197], gastroenteritis [51, 94, 155], gonorrhea [99], hand foot and mouth disease (HFMD) [190], HIV/AIDS [196, 197], influenza [1, 2, 12, 22, 33, 47, 52, 53, 56, 82, 89, 95, 93, 100, 101, 103, 107, 109, 108, 125, 137, 155, 152, 156, 166, 191, 194], kidney stones [21], listeriosis [187], malaria [142], methicillin-resistant *Staphylococcus aureus* (MRSA) [54], pertussis [137], pneumonia [156], respiratory syncytial virus (RSV) [27], scarlet fever [197], stroke [183], suicide [85, 192], tuberculosis [197], and West Nile virus [27].

Closely related to the present work is an independent, simultaneous effort by McIver & Brownstein to measure influenza in the United States using Wikipedia access logs [125]. This study used Poisson models fitted with LASSO regression to estimate ILI over a 5-year period. The results, Pearson's r of 0.94 to 0.99 against official data, depending on model variation, compare quite favorably to prior work that tries to replicate official data. More generally, this article's statistical methods are more sophisticated than those employed in the present study. However, we offer several key improvements:

- We evaluate 14 location-disease contexts around the globe, rather than just one.

In doing so, we test the use of language as a location proxy, which was noted briefly as future work in McIver & Brownstein. (However, as we detail below, we suspect this is not a reliable geo-location method for the long term.)

- We test our models for forecasting value, which was again mentioned briefly as future work in McIver & Brownstein.
- We evaluate models for translatability from one location to another.
- We present negative results and use these to begin exploring when internet-based disease surveillance methods might and might not work.
- We offer a systematic, well-specified, and simple procedure to select articles for model inclusion.
- We normalize article traffic by total language traffic rather than using a few specific articles as a proxy for it.
- Our software is open source and has only freely available dependencies, while the McIver & Brownstein code is not available and depends on proprietary components (Stata).

Finally, the goals of the two studies differ. McIver & Brownstein wanted to "develop a statistical model to provide near-time estimates of ILI activity in the US using freely available data gathered from the online encyclopedia Wikipedia" [125, p. 2]. Our goals are to assess the applicability of these data to global disease surveillance for operational public health purposes and to lay out a research agenda for achieving this end.

These methods are the basis for at least one deployed, widely used surveillance system. Based upon search query data, Google Flu Trends offers near-real-time estimates of influenza activity in 29 countries across the world (15 at the province level); another facet of the same system, Google Dengue Trends (`http://www.google.org/denguetrends/`) estimates dengue activity in 9 countries (2 at the province level) in Asia and Latin America.

Having laid out the space of quantitative internet disease surveillance as it exists to the best of our knowledge, we now consider this prior work in the context of our four challenges:

C1. **Openness.** Deep access to search queries from Baidu, a Chinese-language search engine serving mostly the Chinese market (`http://www.baidu.com`) [99, 194, 197]; Google [11, 21, 27, 31, 33, 51, 53, 54, 56, 82, 85, 101, 142, 155, 183, 187, 192, 191, 190, 196]; Yahoo [42, 156]; and Yandex, a search engine serving mostly Russia and Slavic countries in Russian (`http://www.yandex.ru`), English (`http://www.yandex.com`), and Turkish (`http://www.yandex.com.tr`) [196], as well as purpose-built health website search queries [94, 95, 93] and access logs [100, 190] are available only to those within the organizations, upon payment of an often-substantial fee, or by some other special arrangement. While tools such as Baidu Index (`http://index.baidu.com`), Google Trends (`http://www.google.com/trends/`), Google Correlate (`http://www.google.com/trends/correlate/`), and Yandex's WordStat (`http://wordstat.yandex.com`) provide a limited view into specific search queries and/or time periods, as do occasional lower-level data dumps offered for research, neither affords the large-scale, broad data analysis that drives the most effective models.

The situation is only somewhat better for surveillance efforts based upon Twitter [1, 2, 12, 22, 35, 47, 52, 83, 89, 103, 107, 109, 108, 137, 152, 166]. While a small portion of the real-time message stream (1%, or 10% for certain grandfathered users) is available outside the company without substantial fees, terms of use prohibit sharing historical data needed for calibration between researchers. Access rules are similar or significantly more restrictive for alternative social media platforms such as Sina Weibo, the leading Chinese microblogging site (`http://weibo.com`), and Facebook. Consistent with this, we were able to find no research meeting our inclusion criteria based on either of these extremely popular systems.

We identified only one prior effort making use of open data, McIver & Brownstein with Wikipedia access logs [125]. Open algorithms in this field of inquiry are also very limited. Of the works cited above, again only one, Althouse et al. [11], claims general availability of their algorithms in the form of open source code.

Finally, we highlight the quite successful Google Flu and Dengue Trends as a case study in the problems of closed data and algorithms. First, because their data and algorithms are proprietary, there is little opportunity for the wider community of expertise to offer peer review or improvements (for example, the list of search terms used by Dengue Trends has never been published, even in summary form); the importance of these opportunities is highlighted by the system's well-publicized estimation failures during the 2012–2013 flu season [26] as well as more comprehensive scholarly criticisms [144]. Second, only Google can choose the level of resources to spend on Trends; no one else, regardless of their available resources, can add new contexts or take on operational responsibility should Google choose to discontinue the project.

C2. **Breadth.** While in principle these surveillance approaches are highly generalizable, nearly all extant efforts address a small set of diseases in a small set of countries, without testing specific methods to expand these sets.

The key exception is Paul & Dredze [152], which proposes a content-based method, *ailment topic aspect model* (ATAM), to automatically discover a theoretically unbounded set of medical conditions mentioned in Twitter messages. This unsupervised machine learning algorithm, similarly to latent Dirichlet allocation (LDA) [17], accumulates co-occurring words into probabilistic *topics*. Lists of health-related lay keywords, as well as the text of health articles written for a lay audience, are used to ensure that the algorithm builds topics related to medical issues. A test of the method discovered 15 coherent condition topics including infectious diseases such as influenza, non-infectious diseases such as cancer, and non-specific conditions such as aches and pains. The influenza topic's time series correlated very well with ILI data in the United States.

However, we identify three drawbacks of this approach. First, significant curated text input data in the target language are required; second, output topics require expert interpretation; and third, the ATAM algorithm has several parameters that require expert tuning. That is, in order to adapt the algorithm to a new location and/or language, expertise in both machine learning as well as the target language are required.

In summary, to our knowledge, no disease measurement algorithms have been proposed that are extensible to new disease-location contexts solely by adding examples of desired output. We propose a path to such algorithms.

C3. **Transferability.** To our knowledge, no prior work offers trained models that can be translated from one context to another. We propose using the inter-language article links provided in Wikipedia to accomplish this translation.

C4. **Forecasting.** A substantial minority of the efforts in this space test some kind of forecasting method. (Note that many papers use the term *predict*, and some even misuse *forecast*, to indicate nowcasting.) In addition to forecasting models that incorporate disease dynamics (recall that these are out of scope for the current paper), two basic classes of forecasting exist: *lag analysis*, where the internet data are simply time-shifted in order to capture leading signals, and statistical forecast models such as linear regression.

Lag analysis has shown mixed results in prior work. Johnson et al. [100], Pelat et al. [155], and Jia-xing et al. [99] identified no reliable leading signals. On the other hand, Polgreen et al. [156] used lag analysis with a shift granularity of one week to forecast positive influenza cultures as well as influenza and pneumonia mortality with a horizon of 5 weeks or more (though these indicators may trail the onset of symptoms significantly). Similarly, Xu et al. [190] reported evidence that lag analysis may be able to forecast HFMD by up to two months, and Yang et al. [192] used lag analysis with a granularity of one month to identify search queries that lead suicide incidence by up to two months.

The more complex method of statistical forecast models appears potentially fruitful as well. Dugas et al. tested several statistical methods using positive influenza tests and Google Flu Trends to make 1-week forecasts [53], and Kim et al. used linear regression to forecast influenza on a horizon of 1 month [103].

In summary, while forecasts based upon models that include disease dynamics are clearly useful, sometimes this is not possible because important disease parameters are insufficiently known. Therefore, it is still important to pursue simple methods. The simplest is lag analysis; our contribution is to evaluate leading information more quantitatively than previously attempted. Specifically, we are unaware of previous analysis with shift granularity less than one week; in contrast, our analysis tests daily shifting even if official data are less granular, and each shift is an independently computed model; thus, our ±28-day evaluation results in 57 separate models for each context.

In summary, significant gaps remain with respect to the challenges blocking a path to an open, deployable, quantitative internet-based disease surveillance system. In this paper, we propose a path to overcoming these challenges and offer evidence demonstrating that this path is plausible.

3.2 Methods

We used two data sources, Wikipedia article access logs and official disease incidence reports, and built linear models to analyze approximately 3 years of data for each of 14 disease-location contexts. This section details the nature, acquisition, and processing of these data as well as how we computed the estimation models and evaluated their output.

3.2.1 Wikipedia Article Access Logs

Access logs for all Wikipedia articles are available in summary form to anyone who wishes to use them. We used the complete logs available at `http://dumps.wikimedia.org/other/pagecounts-raw/`. Web interfaces offering a limited view into the logs, such as `http://stats.grok.se`, are also available. These data are referred to using a variety of terms, including *article views*, *article visits*, *pagecount files*, *page views*, *pageviews*, *page view logs*, and *request logs*.

These summary files contain, for each hour from December 9, 2007 to present and updated in real time, a compressed text file listing the number of requests for every article in every language, except that articles with no requests are omitted. (This request count differs from the true number of human views due to automated requests, proxies, pre-fetching, people not reading the article they loaded, and other factors. However, this commonly used proxy for human views is the best available.) We analyzed data from March 7, 2010 through February 1, 2014 inclusive, a total of 1,428 days. This dataset contains roughly 34,000 data files totaling 2.7TB. 266

hours or 0.8% of the data are missing, with the largest gap being 85 hours. These missing data were treated as zero; because they were few, this has minimal effect on our analyses.

We normalized these request counts by language. This yielded, for each article, a time series containing the number of requests for that article during each hour, expressed as a fraction of the hour's total requests for articles in the language. This normalization also compensates for periods of request undercounting, when up to 20% fewer requests were counted than served [195]. Finally, we transposed the data using Map-Reduce [49] to produce files from which the request count time series of any article can be retrieved efficiently.

3.2.2 Disease Incidence Data

Our goal was to evaluate a broad selection of diseases in a variety of countries across the world, in order to test the global applicability and disease agnosticism of our proposed technique. For example, we sought diseases with diverse modes of transmission (e.g., airborne droplet, vector, sexual, and fecal-oral), biology (virus, bacteria, protozoa), types of symptoms, length of incubation period, seasonality, and prevalence. Similarly, we sought locations in both the developed and developing world in various climates. Finally, we wanted to test each disease in multiple countries, to provide an opportunity for comparison.

These comprehensive desiderata were tempered by the realities of data availability. First, we needed reliable data establishing incidence ground truth for specific diseases in specific countries and at high temporal granularity; such official data are frequently not available for locations and diseases of interest. We used official epidemiological reports available on websites of government public health agencies as well as the World Health Organization (WHO).

Second, we needed article access counts for specific countries. This information is not present in the Wikipedia article access logs (i.e., request counts are global totals). However, a proxy is sometimes available in that certain languages are mostly limited to one country of interest; for example, a strong majority of Thai speakers are in Thailand, and the only English-speaking country where plague appears is the United States. In contrast, Spanish is spoken all over the world and thus largely unsuitable for this purpose.

Third, the language edition needs to have articles related to the disease of interest that are mature enough to evaluate and generate sufficient traffic to provide a reasonable signal.

With these constraints in mind, we used our professional judgement to select diseases and countries. The resulting list of 14 disease-location contexts, which is designed to be informative rather than comprehensive, is enumerated in Table 3.1.

Table 3.1: Diseases-location contexts analyzed, with data sources.

Disease	**Country**	**Language**	**Start**	**End**	**Resolution**	**Sources**
Cholera	Haiti	French	2010-12-05	2013-12-05	daily	[130]
Dengue	Brazil	Portuguese	2010-03-07	2013-03-16	weekly	[132]
	Thailand	Thai	2011-01-01	2014-01-31	monthly	[24]
Ebola	Uganda/DRC	English	2011-01-01	2013-12-31	daily	[188, 189, 131]
HIV/AIDS	China (PRC)	Chinese	2011-01-01	2013-12-31	monthly	[32]
	Japan	Japanese	2010-10-09	2013-10-18	weekly	[139]
Influenza	Japan	Japanese	2010-06-26	2013-07-05	weekly	[139]
	Poland	Polish	2010-10-17	2013-10-23	weekly	[140]
	Thailand	Thai	2011-01-23	2014-02-01	weekly	[24]
	United States	English	2011-01-01	2014-01-10	weekly	[28]
Plague	United States	English	2011-01-22	2014-01-31	weekly	[29]
Tuberculosis	China (PRC)	Chinese	2010-12-01	2013-12-31	monthly	[32]
	Norway	Norwegian	2010-12-01	2013-12-31	monthly	[127]
	Thailand	Thai	2010-12-01	2013-12-31	monthly	[24]

This table lists the 7 diseases in 9 locations analyzed, for a total of 14 disease-location contexts. For each context, we list the language used as a location proxy, the inclusive start and end dates of analysis, the resolution of the disease incidence data, and one or more citations for those data.

These incidence data take two basic forms: (a) tabular files such as spreadsheets mapping days, weeks, or months to new case counts or the total number of infected persons or (b) graphs presenting the same mapping. In the latter case, we used plot digitizing software (Plot Digitizer, `http://plotdigitizer.sourceforge.net`) to extract a tabular form. We then translated these diverse tabular forms to a consistent spreadsheet format, yielding for each disease-location context a time series of disease incidence.

3.2.3 Article Selection

The goal of our models is to create a linear mapping from the access counts of some set of Wikipedia articles to a scalar disease incidence for some disease-location context. To do so, a procedure for selecting these articles is needed; for the current proof-of-concept work, we used the following:

1. Examine the English-language Wikipedia article for the disease and enumerate the linked articles. Select for analysis the disease article itself along with linked articles on relevant symptoms, syndromes, pathogens, conditions, treatments, biological processes, and epidemiology. For example, the articles selected for influenza include “Influenza”, “Amantadine”, and “Swine influenza”, but not “2009 flu pandemic”.
2. Identify the corresponding article in each target language by following the inter-language wiki link; these appear at the lower left of Wikipedia articles under the heading “Languages”. For example, the Polish articles selected for influenza include “Grypa”, “Amantadyna”, and “Świńska grypa”, but not “Pandemia grypy A/H1N1 w latach 2009-2010”, respectively.
3. Translate each article title into the form that appears in the logs. Specifically, encode the article’s Unicode title using UTF-8, percent-encode the result, and replace spaces with underscores. For example, the Polish article “Choroby zakaźne” becomes `Choroby_zaka%C5%BAne`. This procedure is accomplished by simply copying the article’s URL from the web browser address bar.

This procedure has two potential complications. First, an article may not exist in the target language; in this case, we simply omit it. Second, Wikipedia contains null articles called *redirects* that merely point to another article, called the *target* of the redirect. These are created to catch synonyms or common misspellings of an article. For example, in English, the article “Flu” is a redirect to “Influenza”. When a user visits `http://en.wikipedia.org/wiki/Flu`, the content served by Wikipedia is actually that of the “Influenza” article; the server does not issue an HTTP 301 response nor require the reader to manually click through to the redirect target.

This complicates our analysis because this arrangement causes the redirect itself (“Flu”), not the target (“Influenza”), to appear in the access log. While in principle we could sum redirect requests into the target article’s total, reliably mapping redirects to targets is a non-trivial problem because this mapping changes over time, and in fact Wikipedia’s history for redirect changes is not complete [186]. Therefore, we have elected to leave this issue for future work; this choice is supported by our observation below that when target and redirect are reversed, traffic to “Dengue fever” in Thai follows the target.

If we encounter a redirect during the above procedure, we use the target article.

3.2.4 Building and Evaluating Each Model

Our goal was to understand how well traffic for a simple selection of articles can nowcast and forecast disease incidence. Accordingly, we implemented the following procedure in Python to build and evaluate a model for each disease-location context.

1. Align the hourly article access counts with the daily, weekly, or monthly disease incidence data by summing the hourly counts for each day, week, or month in the incidence time series. This yields article and disease time series with the same frequency, making them comparable. (We ignore time zone in this

procedure. Because Wikipedia data are in UTC and incidence data are in unspecified, likely local time zones, this leads to a temporal offset error of up to 23 hours, a relatively small error at the scale of our analysis. Therefore, we ignore this issue for simplicity.)

2. For each candidate article in the target language, compute Pearson's correlation r against the disease incidence time series for the target country.
3. Order the candidates by decreasing $|r|$ and select the best 10 articles.
4. Use ordinary least squares to build a linear multiple regression model mapping accesses to these 10 articles to the disease time series. No other variables were incorporated into the model. Below, we report r^2 for the multi-article models as well as a qualitative evaluation of success or failure.

In order to test forecasting potential, we repeat the above with the article time series time-shifted from 28 days forward to 28 days backward in 1-day increments. For example, to build a 4-day forecasting model — that is, a model that estimates disease incidence 4 days in the future — we would shift the article time series *later* by 4 days so that article request counts for a given day are matched against disease incidence 4 days in the future. The choice of ±28 days for lag analysis is based upon our *a priori* hypothesis that these statistical models are likely effective for a few weeks of forecasting.

We refer to models that estimate current (i.e., same-day) disease incidence as *nowcasting* models and those that estimate past disease incidence as *anti-forecasting* models; for example, a model that estimates disease incidence 7 days ago is a 7-day anti-forecasting model. (While useless at first glance, effective anti-forecasting models that give results sooner than official data can still reduce the lead time for action. Also, it is valuable for understanding the mechanism of internet-based models to know the temporal location of predictive information.) We report r^2 for each time-shifted multi-article model.

Finally, to evaluate whether translating models from one location to another is feasible, we compute a metric r_{t} for each pair of locations tested on the same disease. This meta-correlation is simply the Pearson's r computed between the correlation scores r of each article found in both languages; the intent is to give a gross notion of similarity between models computed for the same disease in two different languages. A value of 1 means that the two models are identical, 0 means they have no relationship, and -1 means they are opposite. We ignore articles found in only one language because the goal is to obtain a sense of feasibility: given favorable conditions, could one train a model in one location and apply it to another? Table 3.2 illustrates an example.

Table 3.2: Transferability $\boldsymbol{r}_\mathrm{t}$ example.

Article	**Japanese**	**Thai**
Fever	0.23	0.21
Chills	0.59	
Headache	-0.10	0.15
Influenza	0.85	0.77

This table shows simplified models for influenza in two locations: Japan, where Japanese is spoken, and Thailand, where Thai is spoken. The Japanese model yielded correlations for Japanese versions of the articles "Fever", "Chills", "Headache", and "Influenza" of 0.23, 0.59, -0.10, and 0.85, respectively. The Thai model yielded correlations of 0.21, 0.15, and 0.77 for "Fever", "Headache", and "Influenza", respectively. Note that the article "Chills" is not currently present in the Thai Wikipedia. Therefore, the correlation vectors are $\{0.23, -0.10, 0.85\}$ and $\{0.21, 0.15, 0.77\}$ for the two languages. The meta-correlation, r_t, for these two vectors, which provides a gross estimate of how similar the models are, is 0.97. Extending this computation to the full models yields $r_\mathrm{t} = 0.81$, as noted below in Table 3.4.

3.3 Results

Among the 14 disease-location contexts we analyzed, we found three broad classes of results. In 8 cases, the model succeeded, i.e., there was a usefully close match between the model's estimate and the official data. In 3 cases, the model failed, apparently because patterns in the official data were too subtle to capture, and in a further 3, the model failed apparently because the signal-to-noise ratio (SNR) in the Wikipedia data was too subtle to capture. Recall that this success/failure classification is based on subjective judgement; that is, in our exploration, we discovered that r^2 is insufficient to completely evaluate a model's goodness of fit, and a complementary qualitative evaluation was necessary.

Below, we discuss the successful and failed nowcasting models, followed by a summary and evaluation of transferability. (No models failed at nowcasting but succeeded at forecasting, so we omit a detailed forecasting discussion for brevity.)

3.3.1 Successful Nowcasting

Model and official data time series for selected successful contexts are illustrated in Figure 3.1. The method's good performance on dengue and influenza is consistent with voluminous prior work on these diseases; this offers evidence for the feasibility of Wikipedia access as a data source.

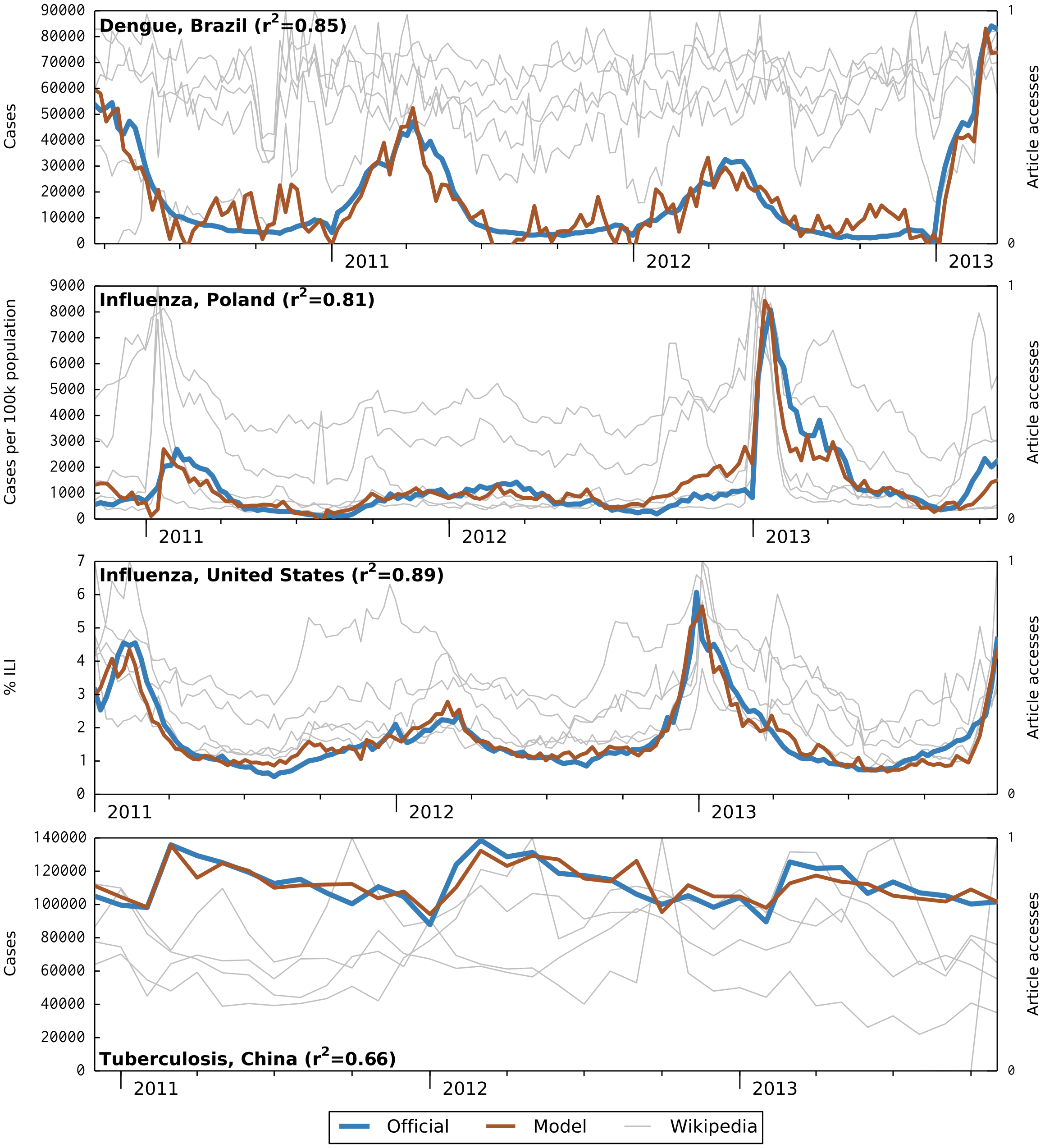


Figure 3.1: Selected successful model nowcasts. These graphs show official epidemiological data and nowcast model estimate (left Y axis) with traffic to the five most-correlated Wikipedia articles (right Y axis) over the 3 year study periods. The Wikipedia time series are individually self-normalized.

Success in the United States is somewhat surprising. Given the widespread use of English across the globe, we expected that language would be a poor location proxy for the United States. We speculate that the good influenza model performance is due to the high levels of internet use in United States, perhaps coupled with similar flu seasons in other Northern Hemisphere countries. Similarly, in addition to Brazil, Portuguese is spoken in Portugal and several other former colonies, yet problems again failed to arise. In this case, we suspect a different explanation: the simple absence of dengue from other Portuguese-speaking countries.

The case of dengue in Brazil is further interesting because it highlights the noise inherent in social data, a property shared by many other internet data sources. That is, noise in the input articles is carried forward into the model's estimate. We speculate that this problem could be mitigated by building a model on a larger, more carefully selected set of articles rather than just 10.

Finally, we highlight tuberculosis in China as an example of a marginally successful model. Despite the apparently low r^2 of 0.66, we judged this model successful because it captured the high baseline disease level excellently and the three modest peaks well. However, it is not clear that the model provides useful information at the time scale analyzed. This result suggests that additional quantitative evaluation metrics may be needed, such as root mean squared error (RMSE) or a more complex analysis considering peaks, valleys, slope changes, and related properties.

Forecasting and anti-forecasting performance of the four selected contexts is illustrated in Figure 3.2. In the case of dengue and influenza, the models contain significant forecast value through the limit of our 28-day analysis, often with the maximally effective lag comprising a forecast. We offer three possible reasons for this. First, both diseases are seasonal, so readers may simply be interested in the syndrome for this reason; however, the fact that models were able to correctly estimate seasons of widely varying severity provides counterevidence for this theory. Second, readers may be interested due to indirect reasons such as news coverage. Prior work disagrees on the impact of such influences; for example, Dukic et al. found that adding news coverage to their methicillin-resistant *Staphylococcus aureus* (MRSA) model had a limited effect [54], but recent Google Flu Trends failures appear to be caused in part by media activity [26]. Finally, both diseases have a relatively short incubation period (influenza at 1–4 days and dengue at 3–14); soon-to-be-ill readers may be observing the illness of their infectors or those who are a small number of degrees removed. It is the third hypothesis that is most interesting for forecasting purposes, and evidence to distinguish among them might be obtained from studies using simulated media and internet data, as suggested by Culotta [47].

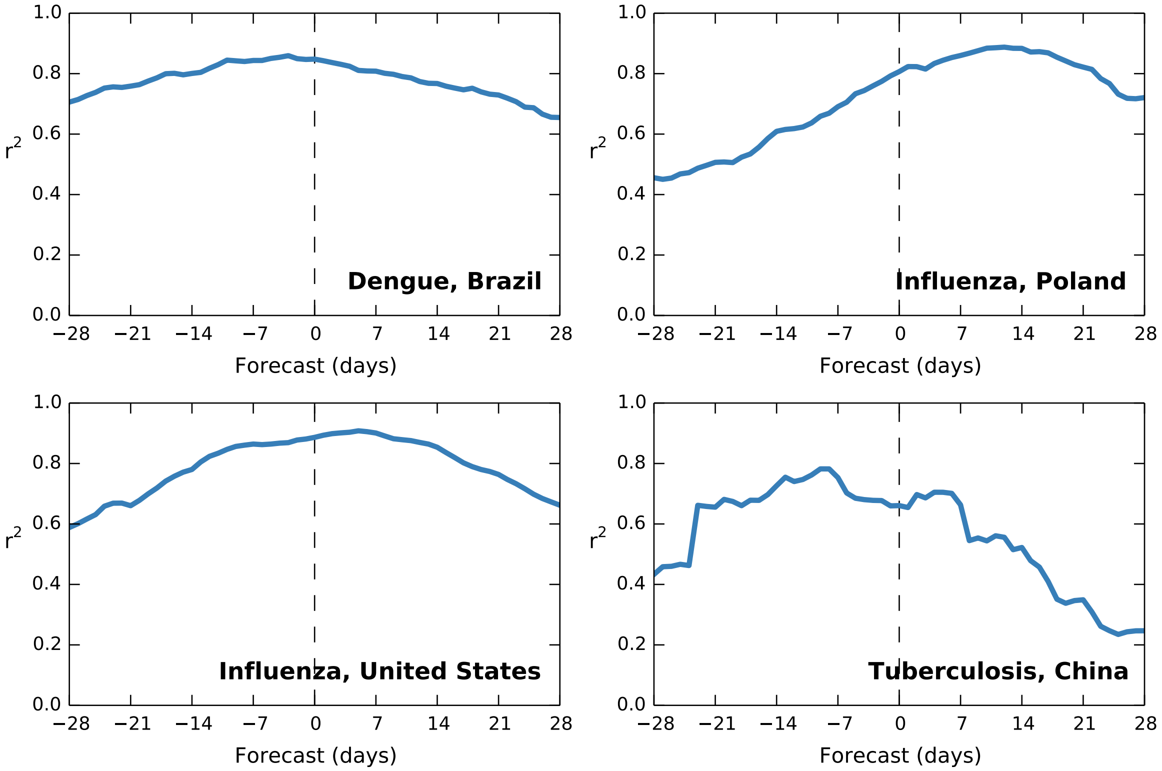


Figure 3.2: Forecasting effectiveness for selected successful models. This figure shows model r^2 compared to temporal offset in days: positive offsets are forecasting, zero is nowcasting (marked with a dotted line), and negative offsets are anti-forecasting.

Tuberculosis in China is another story. In this case, the model's effectiveness is poorer as the forecast interval increases; we speculate that this is because seasonality is absent and the incubation period of 2–12 weeks is longer, diluting the effect of the above two mechanisms.

3.3.2 Failed Nowcasting

Figure 3.3 illustrates the three contexts where the model was not effective because, we suspect, it was not able to discern meaningful patterns in the official data. These suggest a few patterns that models might have difficulty with:

1. **Noise.** True patterns in data may be obscured by noise. For example, in the case of HIV/AIDS in China, the official data vary by a factor of 2 or more throughout the graph, and the model captures this fairly well, but the pattern seems epidemiologically strange and thus we suspect it may be merely noise. The other two contexts appear to also contain significant noise.

   (Note that we distinguish noisy official data from an unfavorable signal-to-noise ratio, which is discussed below.)

2. **Too slow.** Disease incidence may be changing too slowly to be evident in the chosen analysis period. In all three contexts shown in Figure 3.3, the trend of the official data is essentially flat, with HIV/AIDS in Japan especially so. The models have captured this flat trend fairly well, but even doing so excellently provides little actionable value over traditional surveillance.

   Both HIV/AIDS and tuberculosis infections progress quite slowly. A period of analysis longer than three years might reveal meaningful patterns that could be captured by this class of models. However, the social internet is young and turbulent; for example, even 3 years consumes most of the active life of some languages of Wikipedia. This complicates longitudinal analyses.

3. **Too fast.** Finally, incidence may be changing too quickly for the model to capture. We did not identify this case in the contexts we tested; however, it is clearly plausible. For example, quarterly influenza data would be hard to model meaningfully using these techniques.

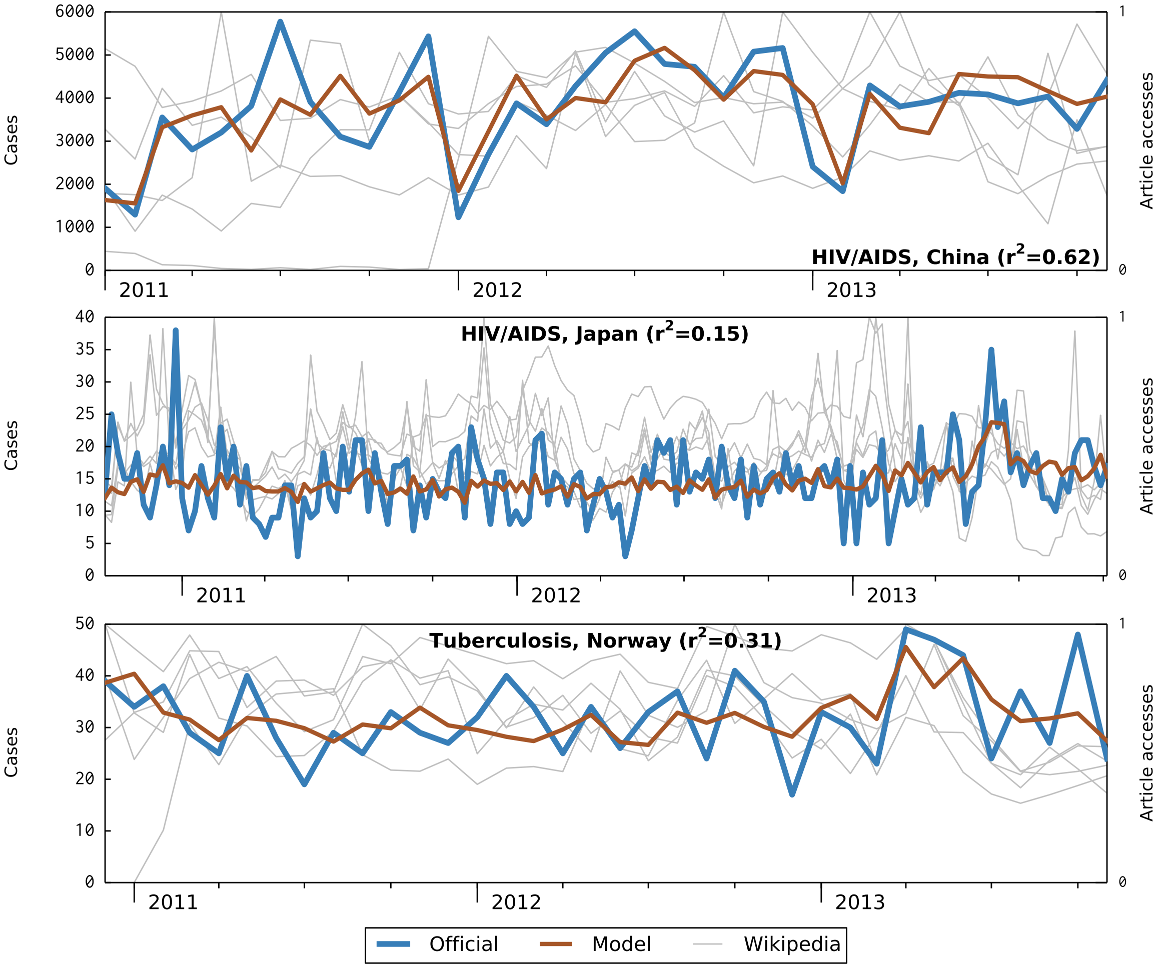


Figure 3.3: Nowcast attempts where the model was unable to capture a meaningful pattern in official data.

In all three patterns, improvements such as non-linear models or better regression techniques could lead to better results, suggesting that this is a useful direction for future work. In particular, noise suppression techniques as well as models tuned for the expected variation in a particular disease may prove fruitful.

Figure 3.4 illustrates the three contexts where we suspect the model failed due to a signal-to-noise ratio (SNR) in the Wikipedia data that was too low. That is, the number of Wikipedia accesses due to actual observations of infection is drowned out by accesses due to other causes.

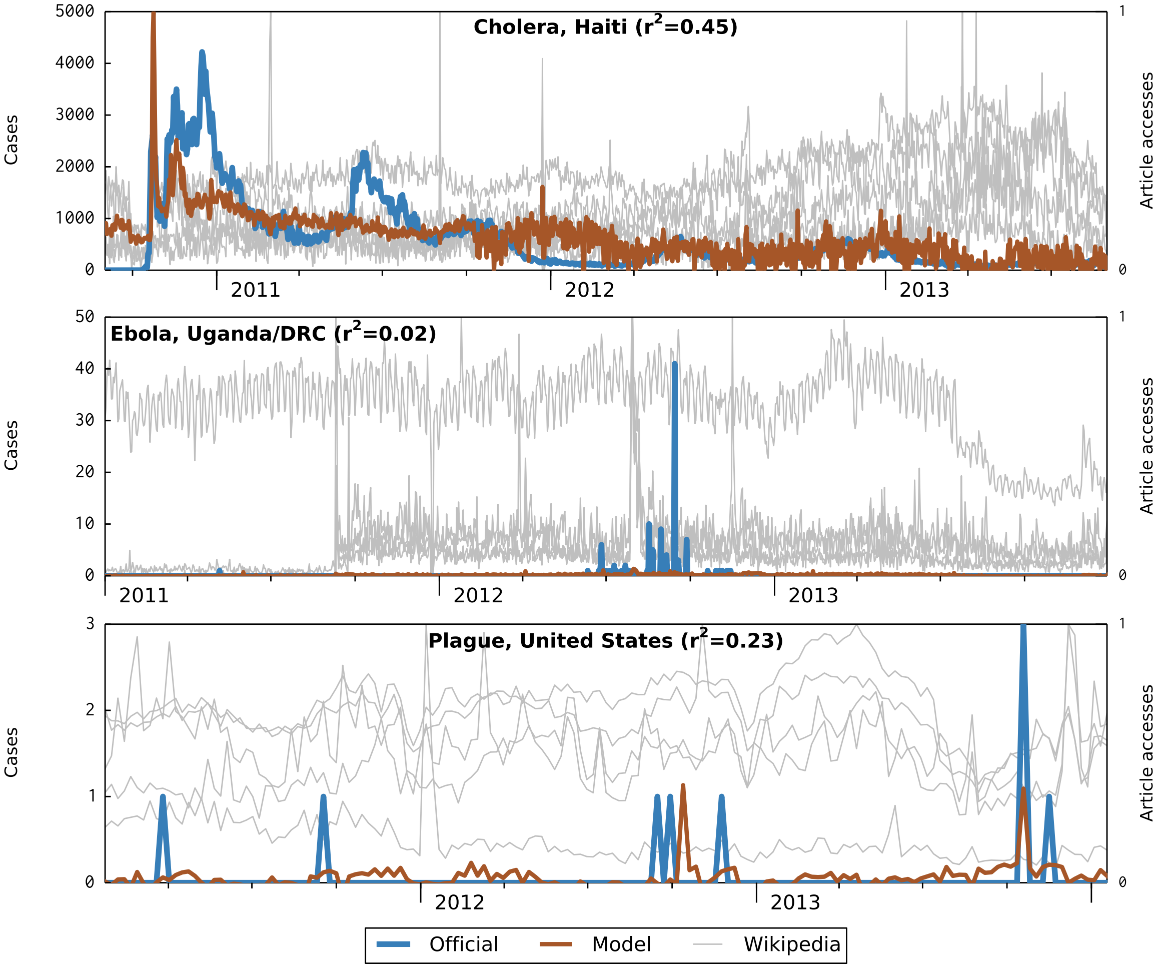


Figure 3.4: Nowcast attempts with poor performance due to unfavorable signal-to-noise ratio.

In the case of Ebola, there are relatively few direct observations (a major outbreak has tens of cases), and the path to these observations becoming internet traces is hampered by poor connectivity in the sub-Saharan countries where the disease is active. On the other hand, the disease is one of general ongoing interest; in fact, one can observe on the graph a pattern of weekly variation (higher in midweek, lower on the weekend), which is common in online activity. In combination, these yield a completely useless model.

The United States has good internet connectivity, but plague has even lower incidence (the peak on the graph is three cases) and this disease is also popularly interesting, resulting in essentially the same effect. The cholera outbreak in Haiti differs in that the number of cases is quite large (the peak of the graph is 4,200 cases in one day). However, internet connectivity in Haiti was already poor even before the earthquake, and the outbreak was a major world news story, increasing noise, so the signal was again lost.

3.3.3 Performance Summary

Table 3.3 summarizes the performance of our models in the 14 disease-location contexts tested. Of these, we classified 8 as successful, producing useful estimates for both nowcasting and forecasting, and 6 as unsuccessful. Performance roughly broke down along disease lines: all influenza and dengue models were successful, while two of the three tuberculosis models were, and cholera, ebola, HIV/AIDS, and plague proved unsuccessful. Given the relatively simple model building technique used, this suggests that our Wikipedia-based approach is sufficiently promising to explore in more detail. (Another hypothesis is that model performance is related to popularity of the corresponding Wikipedia language edition. However, we found no relationship between r^2 and either a language's total number of articles or total traffic.)

Table 3.3: Model performance summary.

			r^2 at forecast				**Best forec.**	
Disease	**Location**	**Result**	**0**	**7**	**14**	**28**	**Days**	**r^2**
Cholera	Haiti	Failure (SNR)	0.45	0.39	0.41	0.48	26	0.50
Dengue	Brazil	Success	0.85	0.81	0.77	0.65	-3	0.86
	Thailand	Success	0.55	0.54	0.57	0.74	28	0.74
Ebola	Uganda/DRC	Failure (SNR)	0.02	0.01	0.02	0.02	5	0.14
HIV/AIDS	China (PRC)	Failure (Official data)	0.62	0.48	0.34	0.31	-1	0.63
	Japan	Failure (Official data)	0.15	0.19	0.15	0.05	9	0.22
Influenza	Japan	Success	0.82	0.92	0.86	0.52	8	0.92
	Poland	Success	0.81	0.86	0.88	0.72	12	0.89
	Thailand	Success	0.79	0.76	0.67	0.48	-2	0.80
	United States	Success	0.89	0.90	0.85	0.66	5	0.91
Plague	United States	Failure (SNR)	0.23	0.03	0.05	0.07	0	0.23
Tuberculosis	China (PRC)	Success	0.66	0.66	0.52	0.25	-9	0.78
	Norway	Failure (Official data)	0.31	0.41	0.40	0.42	20	0.48
	Thailand	Success	0.68	0.68	0.69	0.69	9	0.69

This table summarizes the performance of our estimation models. For each disease and location, we list the subjective success/failure classification as well as model r^2 at nowcasting (0-day forecast) and 7-, 14-, and 28-day forecasts. We also list the temporal offset in days of the best model (again, a positive offset indicates forecasting) along with that model's r^2.

At a higher level, we posit that a successful estimation model based on Wikipedia access logs or other social internet data requires two key elements. First, it must be sensitive enough to capture the true variation in disease incidence data. Second, it must be sensitive enough to distinguish between activity traces due to health-related observations and those due to other causes. In both cases, further research on modeling techniques is likely to yield sensitivity improvements. In particular, a broader article selection procedure — for example, using big data methods to test *all* nontrivial article time series for correlation, as Ginsberg et al. did for search queries [82]

— is likely to prove fruitful, as might a non-linear statistical mapping.

3.3.4 Transferability

Table 3.4 lists the transferability scores r_{t} for each pair of countries tested on the same disease. Because this paper is concerned with establishing feasibility, we focus on the highest scores. These are encouraging: in the case of influenza, both Japan/Thailand and Thailand/United States are promising. That is, it seems plausible that careful source model selection and training techniques may yield useful models in contexts where no training data are available (e.g., official data are unavailable or unreliable). These early results suggest that research to quantitatively test methods for translating models from one disease-location context to another should be pursued.

Table 3.4: Transferability scores $\boldsymbol{r}_{\mathrm{t}}$ for paired models.

Disease	**Location 1**	**Location 2**	$\boldsymbol{r}_{\mathrm{t}}$
Dengue	Brazil	Thailand	0.39
HIV/AIDS	China (PRC)	Japan	-0.06
Influenza	Japan	Poland	0.45
	Japan	Thailand	0.81
	Japan	United States	0.62
	Poland	Thailand	0.48
	Poland	United States	0.44
	Thailand	United States	0.76
Tuberculosis	China (PRC)	Norway	0.19
	China (PRC)	Thailand	-0.20
	Norway	Thailand	n/a

This table lists the transferability scores r_{t} for each tested pair of countries within a disease. Countries that did not share enough articles to compute a meaningful r_{t} are indicated with *n/a*.

3.4 Discussion

Human activity on the internet leaves voluminous traces that contain real and useful evidence of disease dynamics. Above, we pose four challenges currently preventing these traces from informing operational disease surveillance activities, and we argue that Wikipedia data are one of the few social internet data sources that can meet all four challenges. Specifically:

C1. **Openness.** Open data and algorithms are required, in order to offer reliable science as well as a flexible and robust operational capability. Wikipedia access logs are freely available to anyone.

C2. **Breadth.** Thousands of disease-location contexts, not dozens, are needed to fully understand the global disease threat. We tested simple disease estimation models on 14 contexts around the world; in 8 of these, the models were successful with r^2 up to 0.92, suggesting that Wikipedia data are useful in this regard.

C3. **Transferability.** The greatest promise of novel disease surveillance methods is the possibility of use in contexts where traditional surveillance is poor or nonexistent. Our analysis uncovered pairs of same-disease, different-location models with similarity up to 0.81, suggesting that translation of trained models using Wikipedia's mappings of one language to another may be possible.

C4. **Forecasting.** Effective response to disease depends on knowing not only what is happening now but also what will happen in the future. Traditional mechanistic forecasting models often cannot be applied due to missing parameters, motivating the use of simpler statistical models. We show that such statistical models based on Wikipedia data have forecasting value through our maximum tested horizon of 28 days.

This preliminary study has several important limitations. These comprise an agenda for future research work:

1. The methods need to be tested in many more contexts in order to draw lessons about when and why this class of methods is likely to work.
2. A better article selection procedure is needed. In the current paper, we tried a simple manual process yielding at most a few dozen candidate articles in order to establish feasibility. However, real techniques should use a comprehensive process that evaluates thousands, millions, or all plausible articles for inclusion in the model. This will also facilitate content analysis studies that evaluate which types of articles are predictive of disease incidence.
3. Better geo-location is needed. While language as a location proxy works well in some cases, as we have demonstrated, it is inherently weak. In particular, it is implausible for use at a finer scale than country-level. What is needed is a hierarchical geographic aggregation of article traffic. The Wikimedia Foundation, operators of Wikipedia and several related projects, could do this using IP addresses to infer location before the aggregated data are released to the public. For most epidemiologically-useful granularities, this will still preserve reader privacy.
4. Statistical estimation maps from article traffic to disease incidence should be

more sophisticated. Here, we tried simple linear models mapping a single interval's Wikipedia traffic to a single interval's disease incidence. Future directions include testing non-linear and multi-interval models.

5. Wikipedia data have a variety of instabilities that need to be understood and compensated for. For example, Wikipedia shares many of the problems of other internet data, such as highly variable interest-driven traffic caused by news reporting and other sources.

   Wikipedia has its own data peculiarities that can also cause difficulty. For example, during preliminary exploration for this paper in spring 2013, we used the inter-language link on the English article "Dengue fever" to locate the Thai version, "ไข้เลือดออกเด็งกี" (roughly, "dengue hemorrhagic fever"); article access logs indicated several hundred accesses per day for this article in the month of June 2013. When we repeated the same process in March 2014, the inter-language link led to a page with the same content, but a different title, "ไข้เด็งกี" (roughly, "dengue fever"). As none of the authors are Thai speakers, and Google Translate renders both versions as "dengue fever", we did not notice that the title of the Thai article had changed and were alarmed to discover that the article's traffic in June 2013 was essentially zero.

   The explanation is that before July 23, 2013, "ไข้เด็งกี" was a redirect to "ไข้เลือดออกเด็งกี"; on that day, the direction of the redirect was reversed, and almost all accesses moved over to the new redirect target over a period of a few days. That is, the article was the same all along, but the URL under which its accesses were recorded changed.

Possible techniques for compensation include article selection procedures that exclude such articles or maintaining a time-aware redirect graph so that different aliases of the same article can be merged. Indeed, when we tried the latter approach by manually summing the two URLs' time series, it improved nowcast r^2 from 0.55 to 0.65. However, the first technique is likely to discard useful information, and the second may not be reliable because complete history for this type of article transformation is not available [186].

In general, ongoing, time-aware re-training of models will likely be helpful, and limitations of the compensation techniques can be evaluated with simulation studies that inject data problems.

6. We have not explored the full richness of the Wikipedia data. For example, complete histories of each language edition are available, which include editing metadata (timestamps, editor identity, and comments), the text of each version, and conversations about the articles; these would facilitate analysis of edit activity as well as the articles' changing text. Also, health-related articles are often mapped to existing ontologies such as the International Statistical Classification of Diseases and Related Health Problems (ICD-9 or ICD-10).
7. Transferability of models should be tested using more realistic techniques, such as simply building a model in one context and testing its performance in another.

Finally, it is important to recognize the biases inherent in Wikipedia and other social internet data sources. Most importantly, the data strongly over-represent people and places with good internet access and technology skills; demographic biases such as age, gender, and education also play a role. These biases are sometimes quantified (e.g., with survey results) and sometimes completely unknown. As noted above, simulation studies using synthetic internet data can quantify the impact and limitations of these biases.

Despite these limitations, we have established the utility of Wikipedia access logs for global disease monitoring and forecasting, and we have outlined a plausible path to a reliable, scientifically sound, operational disease surveillance system. We look forward to collaborating with the scientific and technical community to make this vision a reality.

3.5 Publications

This chapter is derived from one publication:

- N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky, "Global Disease Monitoring and Forecasting with Wikipedia," PLOS Comput. Biol., vol. 10, no. 11, p. e1003892, Nov. 2014.

CHAPTER 4
ELICITING DISEASE DATA FROM WIKIPEDIA ARTICLES

Abstract

Traditional disease surveillance systems suffer from several disadvantages, including reporting lags and antiquated systems, that have caused a movement towards internet-based disease surveillance systems. Internet systems are particularly attractive for disease outbreaks because they can provide data in near real-time and can be verified by individuals around the globe. However, most existing systems have focused on disease monitoring and do not provide a data repository for policy makers or researchers. In order to fill this gap, we analyzed Wikipedia article content.

We demonstrate how a named-entity recognizer can be trained to tag case counts, death counts, and hospitalization counts in the article narrative that achieves an F1 score of 0.753. We also show, using the the 2014 West African Ebola virus disease epidemic article, that there are detailed time series data that are consistently updated and closely align with ground truth data.

We argue that Wikipedia can be used to create the first community-driven open-source emerging disease detection, monitoring, and repository system.

4.1 Introduction

Most traditional disease surveillance systems rely on data from patient visits or lab records [119, 25, 3]. These systems, while generally recognized to contain accurate information, rely on a hierarchy of public health systems that causes reporting lags of up to 1–2 weeks in many cases [25]. Additionally, many regions of the world lack the infrastructure necessary for these systems to produce reliable and trustworthy data. Recently, in an effort to overcome these issues, timely global approaches to disease surveillance have been devised using internet-based data. Data sources such as search engine queries (e.g., [156, 82]), Twitter (e.g., [46, 166]), and Wikipedia access logs (e.g., [125, 79]) have been shown to be effective in this arena.

A notably different internet-based disease surveillance tool is HealthMap [73]. HealthMap analyzes, in real-time, data from a variety of sources (e.g., ProMED-mail [121], Google News, the World Health Organization) in order to allow simple querying, filtering, and visualization of outbreaks past and present. During emerging outbreaks, HealthMap is often used to understand the current state (e.g., incidence and death counts, outbreak locations). For example, HealthMap was able to detect the 2014 Ebola epidemic nine days before the World Health Organization (WHO) officially announced it [84].

While HealthMap has certainly been influential in the digital disease detection sphere, it has some drawbacks. First and foremost, it runs on source code that is not open and relies on certain data sources that are not freely available in their entirety (e.g., Moreover Newsdesk[1]). Generous et al. argue that there is a genuine need for open source code and open data in order to validate, replicate, and improve existing systems [79]. They argue that while certain closed source services, such as HealthMap and Google Flu Trends [82], are popular and useful to the public, there is no way

[1] `http://www.moreover.com/`

for the public to contribute to the service or continue the service, should the owners decide to shut it down. For example, Google offers a companion site to Google Flu Trends, Google Dengue Trends[2]. However, since Google's source code and data are closed, it is not possible for anyone outside of Google to create similar systems for other diseases, e.g., Google Ebola Trends. Additionally, it is not possible for anyone outside of the HealthMap development team to add new features or data sources to HealthMap. For these reasons, Generous et al. argue for the use of Wikipedia access logs coupled with open source code for digital disease surveillance.

Much richer Wikipedia data are available, however, than just access logs. The entire Wikipedia article content and edit histories are available, complete with edit history metadata (e.g., timestamps of edits and IP addresses of anonymous editors). A plethora of open media—audio, images, and video—are also available. This study presents a novel use of Wikipedia article content and edit history in which disease data are elicited in a timely fashion. Using standard natural language processing (NLP) techniques, we demonstrate our methods by capturing case counts, death counts, and hospitalization counts. We also show there are valuable data present in the tables found in certain disease articles. We argue that Wikipedia data can not only be used for monitoring and forecasting diseases but also as a centralized repository system for collecting disease-related data in near real-time.

[2] http://www.google.org/denguetrends/

4.2 Methods

Disease-related information can be found in a number of places on Wikipedia. We demonstrate how two aspects of Wikipedia article content, historical changes to article text and tabular content, can be harvested for disease surveillance purposes. We first show how a named-entity recognizer can be trained to elicit "important" phrases from outbreak articles, and we then study the accuracy of tabular time series data found in certain articles using the 2014 West African Ebola epidemic as a case study.

4.2.1 Wikipedia data

Wikipedia is an open collaborative encyclopedia consisting of approximately 30 million articles across 287 languages [185]. The English edition of Wikipedia is by far the largest and most active edition; it alone contains approximately 4.7 million articles, while the next largest Wikipedia edition (Swedish) contains only 1.9 million articles [69]. The textual content of the current revision of each English Wikipedia article totals approximately 10 gigabytes [67].

One of Wikipedia's primary attractions to researchers is its openness. All of the historical article content, dating back to Wikipedia's inception in 2001, is available to anyone free of charge. Wikipedia content can be acquired through two means: *a*) Wikipedia's official web API[3] or *b*) downloadable database dumps[4]. Al-

[3] `http://www.mediawiki.org/wiki/API:Main_page`

[4] `http://dumps.wikimedia.org/enwiki/latest/`

though the analysis in this study could have been done offline using the downloadable database dumps, this option is in practice difficult, as the database dumps containing all historical English article revisions are very large (multiple terabytes when uncompressed) [70]. We therefore decided to use Wikipedia's web API, caching content when appropriate.

Wikipedia contains a plethora of articles on specific disease outbreaks and epidemics (e.g., the 2014 West Africa Ebola epidemic[5] and the 2012 Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV) outbreak[6]). We identified two key aspects of Wikipedia disease outbreak articles that can aid disease surveillance efforts: *a*) key phrases in the article text and *b*) tabular content. Most outbreak articles we surveyed contained: dates, locations, case counts, death counts, case fatality rates, demographics, and hospitalization counts in the text. These data are, in general, swiftly updated as new data become available. Perhaps most importantly, sources are often provided so that external review can occur. The following two excerpts came from the articles on the 2012 MERS-CoV outbreak and 2014 Ebola epidemic, respectively:

> On 16 April 2014, Malaysia reported its first MERS-COV related death.[34] The person was a 54 year-old man who had traveled to Jeddah, Saudi Arabia, together with pilgrimage group composed of 18 people, from 15–

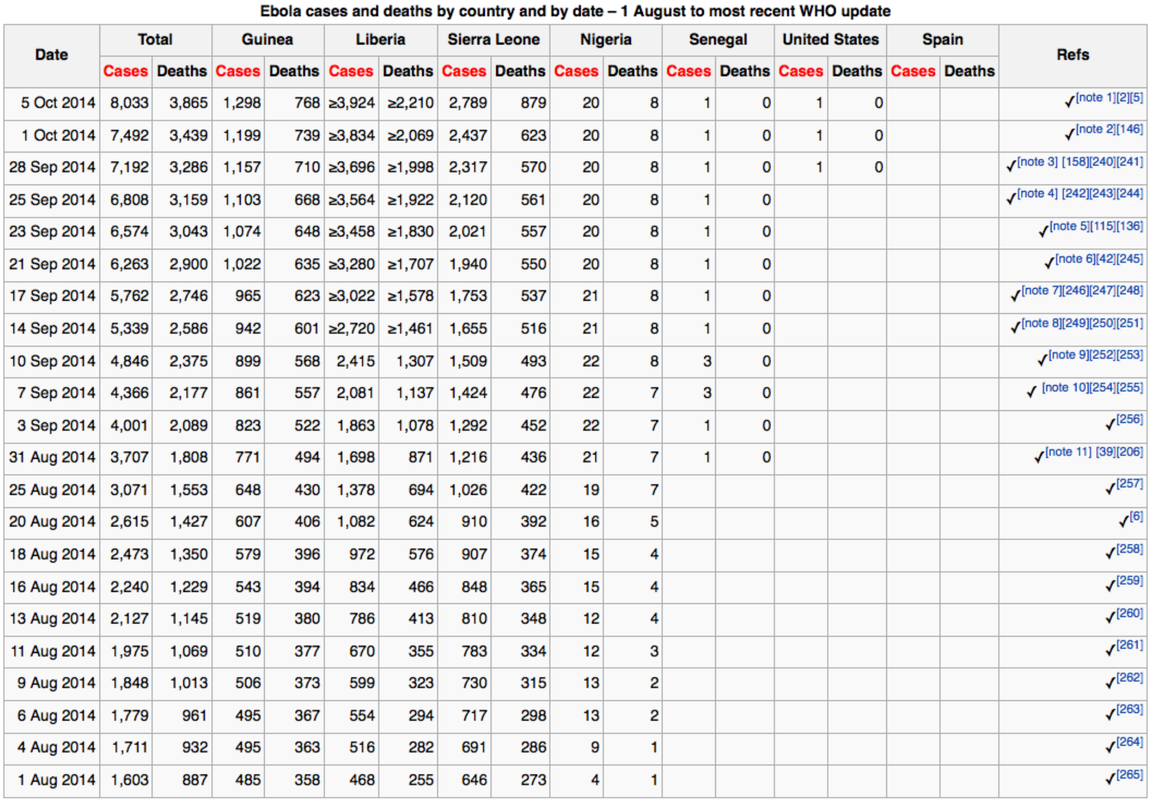
[5] `http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa`

[6] `http://en.wikipedia.org/wiki/2012_Middle_East_respiratory_syndrome_coronavirus_outbreak`

> 28 March 2014. He became ill by 4 April, and sought remedy at a clinic in Johor on 7 April. He was hospitalized by 9 April and died on 13 April.[35] [65]

> On 31 March, the U.S. Centers for Disease Control and Prevention (CDC) sent a five-person team to assist Guinea's Ministry of Health and the WHO to lead an international response to the Ebola outbreak. On that date, the WHO reported 112 suspected and confirmed cases including 70 deaths. Two cases were reported from Liberia of people who had recently traveled to Guinea, and suspected cases in Liberia and Sierra Leone were being investigated.[24] On 30 April, Guinea's Ministry of Health reported 221 suspected and confirmed cases including 146 deaths. The cases included 25 health care workers with 16 deaths. By late May, the outbreak had spread to Conakry, Guinea's capital, a city of about two million inhabitants.[24] On 28 May, the total cases reported had reached 281 with 186 deaths.[24] [66]

Although most outbreak articles contain content similar to the above examples, not all outbreak articles on Wikipedia contain tabular data. The tabular data that do exist, though, are often consistently updated and easily parseable. For example, Figure 4.1 presents a screenshot of a table taken from the 2014 Ebola epidemic article. This table contains case counts and death counts for all regions of the world affected by the epidemic, complete with references for the source data. The time granularity is irregular, but updated counts are consistently provided every 2–5 days.



Ebola cases and deaths by country and by date – 1 August to most recent WHO update

Date	Total		Guinea		Liberia		Sierra Leone		Nigeria		Senegal		United States		Spain		Refs
	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	
5 Oct 2014	8,033	3,865	1,298	768	≥3,924	≥2,210	2,789	879	20	8	1	0	1	0			✓[note 1][2][5]
1 Oct 2014	7,492	3,439	1,199	739	≥3,834	≥2,069	2,437	623	20	8	1	0	1	0			✓[note 2][146]
28 Sep 2014	7,192	3,286	1,157	710	≥3,696	≥1,998	2,317	570	20	8	1	0	1	0			✓[note 3] [158][240][241]
25 Sep 2014	6,808	3,159	1,103	668	≥3,564	≥1,922	2,120	561	20	8	1	0					✓[note 4] [242][243][244]
23 Sep 2014	6,574	3,043	1,074	648	≥3,458	≥1,830	2,021	557	20	8	1	0					✓[note 5][115][136]
21 Sep 2014	6,263	2,900	1,022	635	≥3,280	≥1,707	1,940	550	20	8	1	0					✓[note 6][42][245]
17 Sep 2014	5,762	2,746	965	623	≥3,022	≥1,578	1,753	537	21	8	1	0					✓[note 7][246][247][248]
14 Sep 2014	5,339	2,586	942	601	≥2,720	≥1,461	1,655	516	21	8	1	0					✓[note 8][249][250][251]
10 Sep 2014	4,846	2,375	899	568	2,415	1,307	1,509	493	22	8	3	0					✓[note 9][252][253]
7 Sep 2014	4,366	2,177	861	557	2,081	1,137	1,424	476	22	7	3	0					✓ [note 10][254][255]
3 Sep 2014	4,001	2,089	823	522	1,863	1,078	1,292	452	22	7	1	0					✓[256]
31 Aug 2014	3,707	1,808	771	494	1,698	871	1,216	436	21	7	1	0					✓[note 11] [39][206]
25 Aug 2014	3,071	1,553	648	430	1,378	694	1,026	422	19	7							✓[257]
20 Aug 2014	2,615	1,427	607	406	1,082	624	910	392	16	5							✓[6]
18 Aug 2014	2,473	1,350	579	396	972	576	907	374	15	4							✓[258]
16 Aug 2014	2,240	1,229	543	394	834	466	848	365	15	4							✓[259]
13 Aug 2014	2,127	1,145	519	380	786	413	810	348	12	4							✓[260]
11 Aug 2014	1,975	1,069	510	377	670	355	783	334	12	3							✓[261]
9 Aug 2014	1,848	1,013	506	373	599	323	730	315	13	2							✓[262]
6 Aug 2014	1,779	961	495	367	554	294	717	298	13	2							✓[263]
4 Aug 2014	1,711	932	495	363	516	282	691	286	9	1							✓[264]
1 Aug 2014	1,603	887	485	358	468	255	646	273	4	1							✓[265]

Figure 4.1: Table containing updated worldwide Ebola case counts and death counts. This is a screenshot taken directly from the 2014 Ebola epidemic Wikipedia article [66]. Time granularity is irregular but is in general every 2–5 days. References are also provided for all data points.

While there are certainly other aspects of Wikipedia article content that can be leveraged for disease surveillance purposes, these are the two we focus on in this study. The following sections detail the data extraction methods we use.

4.2.2 Named-entity recognition

In order to recognize certain key phrases in the Wikipedia article narrative, we trained a *named-entity recognizer* (NER). Named-entity recognition is a task commonly used in natural language processing (NLP) to identify and categorize certain key phrases in text (e.g., names, locations, dates, organizations). NERs are *sequence labelers*; that is, they label sequences of words. Consider the following example [68]:

Jim bought 300 shares of Acme Corp. in 2006.

Entities in this example could be named as follows:

[Jim]$_{\texttt{PERSON}}$ bought 300 shares of [Acme Corp.]$_{\texttt{ORGANIZATION}}$ in [2006]$_{\texttt{TIME}}$.

This study specifically uses Stanford's NER [58][7]. The Stanford NER is an implementation of a conditional random field (CRF) model [170]. CRFs are probabilistic statistical models that are the discriminative analog of hidden Markov models (HMMs). Generative models, such as HMMs, learn the joint probability $p(x, y)$, while discriminative models, such as CRFs, learn the conditional probability $p(y \mid x)$. In practice, this means that generative models like HMMs classify by modeling the actual distribution of each class, while discriminative models like CRFs classify by modeling the boundaries between classes. In most cases, discriminative models outperform generative models [141].

While Stanford's NER includes models capable of recognizing common named entities, such as `PERSON`, `ORGANIZATION`, and `LOCATION`, it also provides the capability for us to train our own model so that we can capture new types of named entities we are interested in. For this specific task, we were interested in automatically identifying three entity types: *a*) `DEATHS` *b*) `INFECTIONS`, and *c*) `HOSPITALIZATIONS`. Our trained

[7] `http://nlp.stanford.edu/software/CRF-NER.shtml`

model should therefore be able to automatically tag phrases that correspond to these three entities in the text documents it receives as input.

A number of steps were required to prepare the data for annotation so that the NER could be trained:

1. We first queried Wikipedia's API in order to get the complete revision history for the articles used in our training set.
2. We cleaned each revision by stripping all MediaWiki markup from the text, as well as removing tables.
3. We computed the diff between successive pairs of articles. This provided lines deleted and added between the two article revisions. We retained a list of all the line additions across all article revisions.
4. Many lines in this resulting list were similar to one another (e.g., "There are 45 new cases." → "There are 56 new cases."). For the purposes of training the NER, it is not necessary to retain highly similar or identical lines. We therefore split each line into sentences and removed similar sentences by computing the Jaccard similarity between each sentence using trigrams as the constituent parts in the Jaccard equation. The Jaccard similarity equation for measuring the similarity between two sets A and B, defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, is commonly used for near-duplicate detection [123]. We only kept sentences for which the similarity with all the distinct sentences retained so far was no greater than 0.75.
5. We split each line into tokens in order to create a tab-separated value file that

is compatible with Stanford's NER.

6. Finally, we used Stanford's part-of-speech (POS) tagger [179][8] to add a POS feature to each token.

In order to train the NER, we annotated a dataset derived from the following 14 Wikipedia articles generated according to the above methodology: *a*) Ebola virus epidemic in West Africa[9], *b*) Haiti cholera outbreak[10], *c*) 2012 Middle East respiratory syndrome coronavirus outbreak[11], *d*) New England Compounding Center meningitis outbreak[12], *e*) Influenza A virus subtype H7N9[13], *f*) 2013–14 chikungunya outbreak[14], *g*) Chikungunya outbreaks[15], *h*) Dengue fever outbreaks[16], *i*) 2013 dengue outbreak in Singapore[17], *j*) 2011 dengue outbreak in Pakistan[18], *k*) 2009–10 West African menin-

[8] `http://nlp.stanford.edu/software/tagger.shtml`

[9] `http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa`

[10] `http://en.wikipedia.org/wiki/Haiti_cholera_outbreak`

[11] `http://en.wikipedia.org/wiki/2012_Middle_East_respiratory_syndrome_coronavirus_outbreak`

[12] `http://en.wikipedia.org/wiki/New_England_Compounding_Center_meningitis_outbreak`

[13] `http://en.wikipedia.org/wiki/Influenza_A_virus_subtype_H7N9`

[14] `http://en.wikipedia.org/wiki/2013%E2%80%9314_chikungunya_outbreak`

[15] `http://en.wikipedia.org/wiki/Chikungunya_outbreaks`

[16] `http://en.wikipedia.org/wiki/Dengue_fever_outbreaks`

[17] `http://en.wikipedia.org/wiki/2013_dengue_outbreak_in_Singapore`

[18] `http://en.wikipedia.org/wiki/2011_dengue_outbreak_in_Pakistan`

gitis outbreak[19], *l*) Mumps outbreaks in the 21st century[20], *m*) Zimbabwean cholera outbreak[21], and *n*) 2006 dengue outbreak in India[22]. The entire cleaned and annotated dataset contained approximately 55,000 tokens. The inside-outside-beginning (IOB) scheme, popularized in part by the CoNLL-2003 shared task on language-independent named-entity recognition [178], was used to tag each token. The IOB scheme offers the ability to tie together sequences of tokens that make up an entity.

The annotation task was split between two annotators. In order to tune inter-annotator agreement, the annotators each annotated three sets of 5,000 tokens. After each set of annotations, differences were identified, and clarifications to the annotation rules were made. The third set resulted in a Cohen's kappa coefficient of 0.937, indicating high agreement between the annotators.

4.2.3 Tabular data

To understand the viability of tabular data in Wikipedia, we concentrate on the Ebola virus epidemic in West Africa article[23]. As seen in Figure 4.1, this article contains detailed tables of case counts and death counts by country. The article is regularly updated by the Wikipedia community (see Figure 4.2); over the 165-day period analyzed, the article averaged approximately 31 revisions per day.

[19] `http://en.wikipedia.org/wiki/2009%E2%80%9310_West_African_meningitis_outbreak`

[20] `http://en.wikipedia.org/wiki/Mumps_outbreaks_in_the_21st_century`

[21] `http://en.wikipedia.org/wiki/Zimbabwean_cholera_outbreak`

[22] `http://en.wikipedia.org/wiki/2006_dengue_outbreak_in_India`

[23] `http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa`

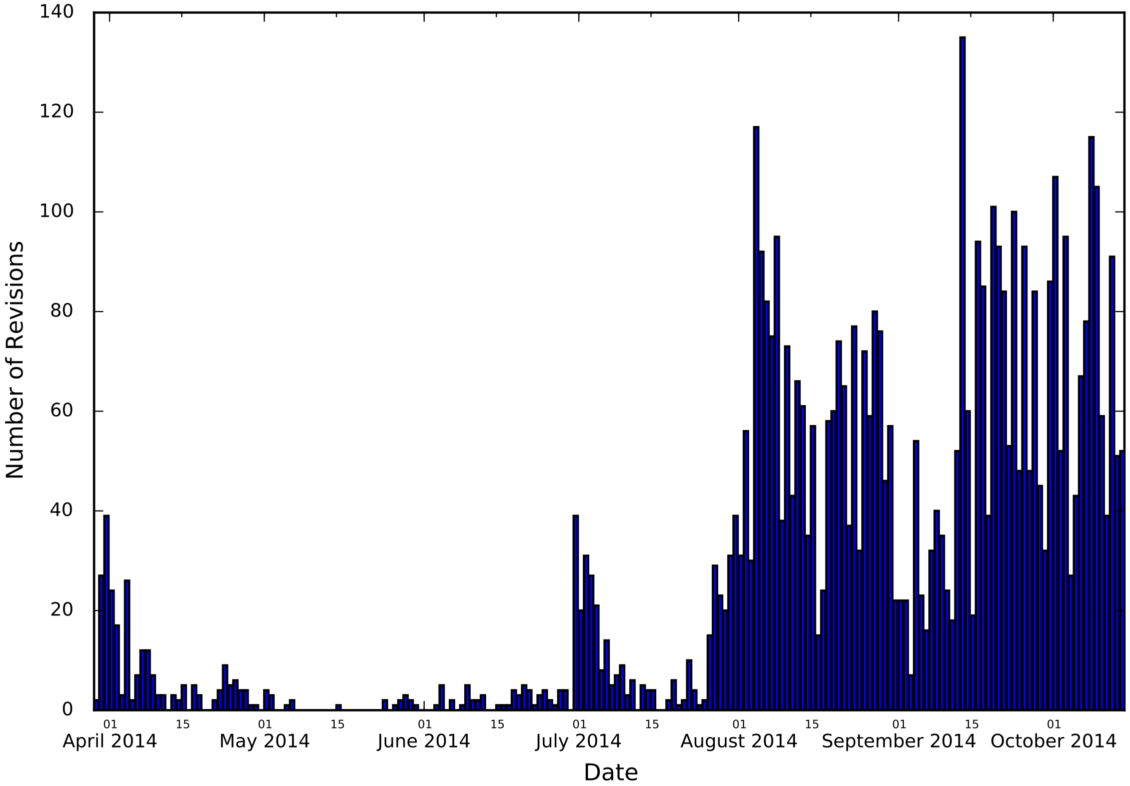


Figure 4.2: The number of revisions made each day to the 2014 Ebola virus epidemic in West Africa Wikipedia article (`http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa`). A total of 5,137 revisions were made over the 165-day period analyzed.

Ebola is a rare but deadly virus that first appeared in 1976 simultaneously in two different remote villages in Africa. Outbreaks of Ebola virus disease (EVD), previously known as Ebola hemorrhagic fever (EHF), are sporadic and generally short-lived. The average case fatality rate is 50%, but it has varied between 25% and 90% in previous outbreaks. EVD is transmitted to humans from animals (most commonly, bats, apes, and monkeys) and also from other humans through direct contact with blood and body fluids. Signs and symptoms appear within 2–21 days of exposure (average 8–10 days) and include fever, severe headache, muscle pain, weakness, diarrhea, vomiting, abdominal pain, and unexplained bleeding or bruising. Although there is currently no known cure, treatment in the form of aggressive rehydration seems to improve survival rates [147, 59].

The West African EVD epidemic was officially announced by the WHO in March 2014 [148]. The disease spread rapidly and has proven difficult to contain in several regions of Africa. At the time of this writing, it has spread to 7 different countries (including two outside of Africa): Guinea, Liberia, Sierra Leone, Nigeria, Senegal, United States, and Spain.

We parsed the Ebola article's tables in several steps:

1. We first queried Wikipedia's API to get the complete revision history for the West African EVD epidemic article. Our initial dataset contained 5,137 revisions from March 29, 2014 to October 14, 2014.
2. We then parsed each revision to pull out case count and death count time series for each revision. To parse the tables, we first used pandoc[24] to convert the MediaWiki markup to consistently-formatted HTML and then used BeautifulSoup[25] to parse the HTML. Because the Wikipedia time series contain a number of missing data points prior to June 30, 2014, we use this date for the beginning of our analysis; time series data prior to June 30, 2014 are not used in this study. This resulting dataset contained 3,803 time series.

[24] `http://johnmacfarlane.net/pandoc/`

[25] `http://www.crummy.com/software/BeautifulSoup/`

3. As Figure 4.1 shows, there are non-regular gaps in the Wikipedia time series; these gaps range from 2–5 days. We linearly interpolated to fill in missing data points where necessary so that we have daily time series. Daily time series data simplify comparisons with ground truth data (described later).
4. Recognizing that the tables will not necessarily change between article revisions (i.e., an article revision might contain edits to only the text of the article, not to a table in the article), we then removed identical time series. This final dataset contained 39 time series.

4.3 Results

4.3.1 Named-entity recognition

To test the classifier's performance, we averaged precision, recall, and F1 score results from 10-fold cross-validation. Table 4.1 demonstrates a typical confusion matrix used to bin cross-validation results, which are then used to compute precision, recall, and the F1 score. Precision asks, "Out of all the examples the classifier labeled, what fraction were correct?" and is computed as $\frac{\text{TP}}{\text{TP}+\text{FP}}$. Recall asks, "Out of all labeled examples, what fraction did the classifier recognize?" and is computed as $\frac{\text{TP}}{\text{TP}+\text{FN}}$. The F1 score is the harmonic mean of precision and recall: $2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$. All 3 scores range from 0 to 1, where 0 is the worst score possible and 1 is the best score possible.

Table 4.1: Typical classifier confusion matrix.

	Ground truth positive	**Ground truth negative**
Test positive	True positive (TP)	False positive (FP)
Test negative	False negative (FN)	True negative (TN)

Table 4.2 shows these results as we varied the `maxNGramLeng` option (Stanford's default value is 6). The `maxNGramLeng` option determines sequence length when training. We were somewhat surprised to discover that larger `maxNGramLeng` values did not improve the performance of the classifier, indicating that more training data are likely necessary to further improve the classifier. Furthermore, roughly maximal performance is achieved with `maxNGramLeng` = 4; there is no tangible benefit to larger sequences (despite this, we concentrate on the `maxNGramLeng` = 6 case since it is the default). Our 14-article training set achieved precision of 0.812 and recall of 0.710, giving us an F1 score of 0.753 for `maxNGramLeng` = 6.

Table 4.2: Classifier performance determined from 10-fold cross-validation.

maxNGramLeng	**Precision**	**Recall**	**F1 score**
1	0.820	0.693	0.747
2	0.810	0.690	0.740
3	0.815	0.702	0.750
4	0.814	0.709	0.753
5	0.813	0.709	0.753
6	0.812	0.710	0.753
7	0.812	0.706	0.751
8	0.814	0.708	0.753
9	0.815	0.707	0.753
10	0.815	0.708	0.753
11	0.813	0.708	0.753
12	0.811	0.709	0.752

For maxNGramLeng = 6, Table 4.3 shows the average precision, recall, and F1 scores for each of the named entities we annotated (DEATHS, INFECTIONS, and HOSPITALIZATIONS). There were a total of 264 DEATHS, 633 INFECTIONS, and 16 HOSPITALIZATIONS entities annotated across the entire training dataset. Recall that we used the IOB scheme for annotating sequences; this is reflected in Table 4.3, with B-* indicating the beginning of a sequence and I-* indicating the inside of a sequence. It is generally the case that identifying the beginning of a sequence is easier than identifying all of the inside words of a sequence; the only exception to this is HOSPITALIZATIONS, but we speculate that the identical beginning and inside results for this entity are due to the relatively small sample size.

Table 4.3: Classifier performance for each of the entities we used in our annotations.

Named entity	**Precision**	**Recall**	**F1 score**
`B-Deaths`	0.888	0.744	0.802
`I-Deaths`	0.821	0.730	0.764
`B-Infections`	0.812	0.719	0.756
`I-Infections`	0.762	0.714	0.730
`B-Hospitalizations`	0.933	0.833	0.853
`I-Hospitalizations`	0.933	0.833	0.853

4.3.2 Tabular data

To compute the accuracy of the Wikipedia West African EVD epidemic time series, we used Caitlin Rivers' crowdsourced Ebola data[26]. Her country-level data come from official WHO data and reports. As with the Wikipedia time series, we linearly interpolated to fill in missing data where necessary so that the ground truth data are specified daily; this ensured that the Wikipedia and ground truth time series were specified at the same granularity. Note that time granularity of the WHO-based ground truth dataset is generally finer than the Wikipedia data; the gaps in the ground truth time series were not the same as those in the Wikipedia time series. In many cases, the ground truth data were updated every 1–2 days.

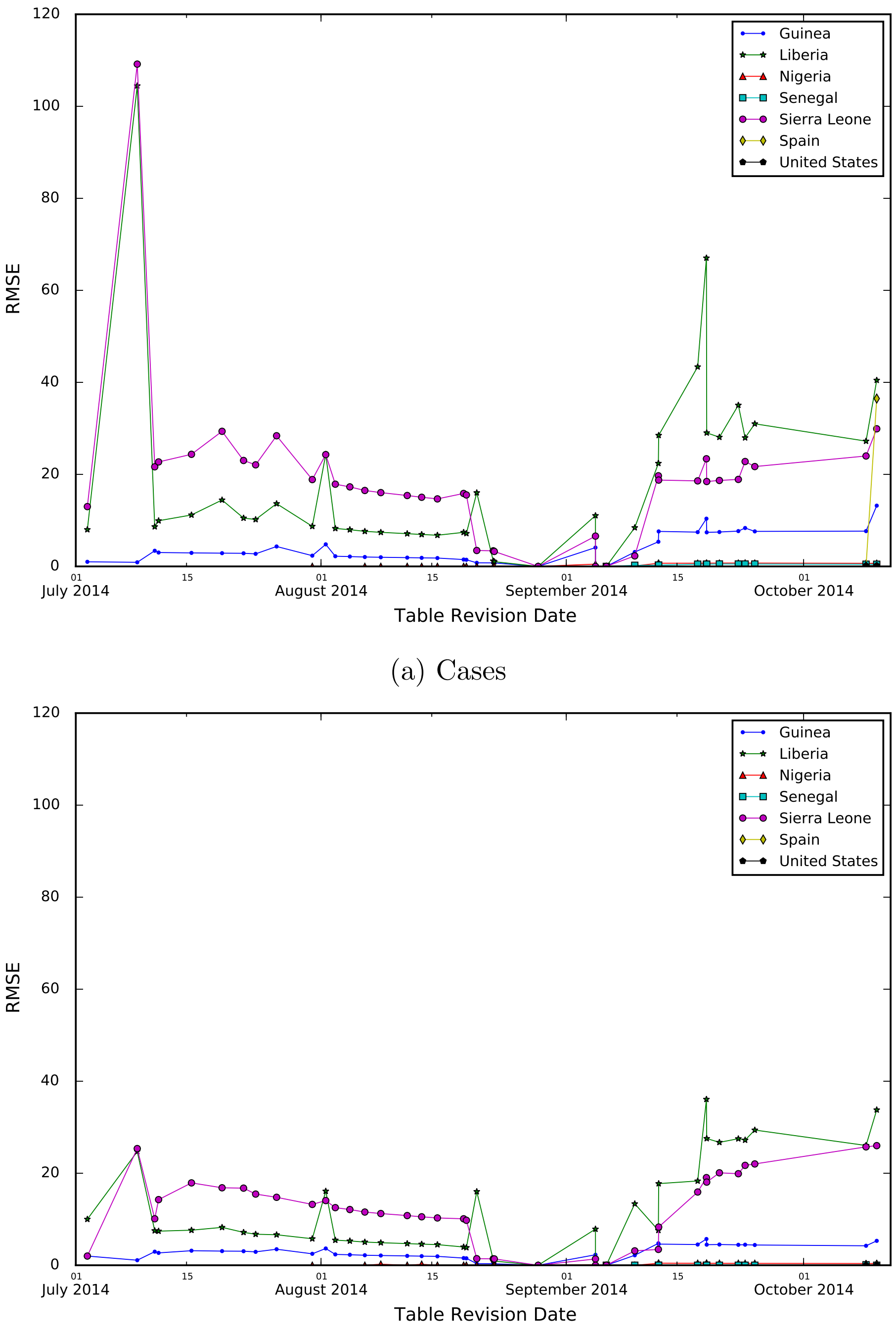
We compared the 39 Wikipedia epidemic time series to the ground truth data by computing the root-mean-square error (RMSE). We use the RMSE rather than the mean-square error (MSE) because the testing and ground truth time series both have the same units (cases or deaths); when they have the same units, the computed

[26] `https://github.com/cmrivers/ebola`

RMSE also has the same unit, which makes it easily interpretable. The RMSE,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{Y_i} - Y_i)^2}, \tag{4.1}$$

computes the average number of cases or deaths difference between a Wikipedia epidemic time series ($\hat{Y}$) and the ground truth time series (Y). Figure 4.3 shows how the case time series and death time series RMSE changes with each table revision for each country. Of particular interest is the large spike in Figure 4.3a on July 8, 2014 in Liberia and Sierra Leone. Shortly after the 6:27pm spike, an edit from a different user at 8:16pm the same day with edit summary “correct numbers in wrong country columns” corrected the error.



(a) Cases

(b) Deaths

Figure 4.3: Root-mean-square error (RMSE) values for the cases and deaths time series are shown for each revision where the tables changed. The RMSE spikes on July 8, 2014 (Liberia and Sierra Leone) and August 20, 2014 (Liberia) in 4.3a were due to Wikipedia contributor errors and were fixed shortly after they were made. Most RMSE spikes are quickly followed by a decrease; this is due to updated WHO data or contributor error detection.

The average RMSE values for each country's time series are listed in Table 4.4. Even in the worst case, the average deviation between the Wikipedia time series and the ground truth is approximately 19 cases and 12 deaths. Considering the magnitude of the number of cases (e.g., approximately 1,500 in Liberia and 3,500 in Sierra Leone during the time period considered) and deaths (e.g., approximately 850 in Liberia and 1,200 in Sierra Leone), the Wikipedia time series are generally within 1–2% of the ground truth data.

Table 4.4: Average cases and deaths RMSE across all table revisions.

Country	**Mean Cases RMSE**	**Mean Deaths RMSE**
Guinea	3.790	2.701
Liberia	18.168	11.983
Nigeria	0.310	0.189
Senegal	0.403	0.008
Sierra Leone	18.847	12.015
Spain	18.243	0.050
United States	0.174	0.000

4.4 Conclusions

Internet data are becoming increasingly important for disease surveillance. Internet data can be used to overcome the reporting lags inherent in traditional disease surveillance data, and they can also be used to detect and monitor emerging diseases. Additionally, internet data can simplify global disease data collection. Collecting global disease data is a formidable task that often requires browsing websites written in an unfamiliar language, and data are specified in a number of formats ranging from well-formed spreadsheets to unparseable PDF files containing low resolution images of tables. Although several popular internet-based systems exist to help overcome some of these traditional disease surveillance system weaknesses, most notably HealthMap [73] and Google Flu Trends [82], no such system exists that relies solely on open data and runs using 100% open source code.

Previous work explored Wikipedia access logs to tackle some of the disadvantages traditional disease surveillance systems face [125, 79]. This study explores a new facet of Wikipedia: the content of disease-related articles. We present methods on how to elicit data that can potentially be used for near-real-time disease surveillance purposes. We argue that in some instances, Wikipedia may be viewed as a centralized crowdsourced data repository.

First, we demonstrate using a named-entity recognizer (NER) how case counts, death counts, and hospitalization counts can be tagged in the article narrative. Our NER, trained on a dataset derived from 14 Wikipedia articles on disease outbreaks/epidemics, achieved an F1 score of 0.753, evidence that this method is fully capable of recognizing these entities in text. Second, we analyzed the quality of tabular data available in the 2014 West Africa Ebola virus disease article. By computing the root-mean-square error (RMSE), we show that the Wikipedia time series very closely align with WHO-based ground truth data.

There are many future directions for this work. First and foremost, more training data are necessary for an operational system in order to improve precision and recall. There are many more disease- and outbreak-related Wikipedia articles that can be annotated. Additionally, other open data sources, such as ProMED-mail, might be used to enhance the model. Second, a thorough analysis of the quality and correctness of the entities tagged by the NER is needed. This study presents the methods by which disease-related named entities can be recognized, but we have not throughly studied the correctness and timeliness of the data. Third, our analysis of tabular data consisted of a single article. A more rigorous study looking at the quality of tabular data in more articles is necessary. Finally, the work presented here considers only the English Wikipedia. NERs are capable of tagging entities in a variety of other languages; more work is needed to understand the quality of data available in the 286 non-English Wikipedias.

There are several limitations to this work. First, the ground truth time series we used to compute RMSEs is static, while the Wikipedia time series vary over time. Because the relatively recent static ground truth time series may contain corrections for reporting errors made earlier in the epidemic, the RMSE values may be artificially inflated in some instances. Second, we are ignoring the user-provided edit summary. This edit summary provides information about why the edit was made. The edit summary identifies article vandalism (and subsequent vandalism reversion) as well as content corrections and updates. Taking these edit summaries into account can further improve model performance (e.g., processing edit summaries would allow us to disregard the erroneous edit that caused the July 8, 2014 spike in Figure 4.3a).

Ultimately, we envision this work being incorporated into a community-driven open-source emerging disease detection and monitoring system. Wikipedia access log time series gauge public interest and, in many cases, correlate very well with disease incidence. Demonstrated in this study, Wikipedia article content can be used to dynamically understand the global state of an outbreak. ProMED-mail could be incorporated as well in order to inform the system which disease articles should be monitored. These three data sources could provide the backbone of a truly open internet-based disease surveillance system. A community-driven effort to improve global disease surveillance data is imminent, and Wikipedia can play a crucial role in realizing this need.

CHAPTER 5
CONCLUSIONS

Traditional disease surveillance systems are instrumental in guiding policymakers' decisions and understanding disease dynamics. These systems generally rely on data gathered from primary care physicians or laboratory reports. The first study in this dissertation analyzed sentinel surveillance network design. Many sentinel surveillance networks in the U.S. are "networks of convenience"; that is, the primary care facilities and laboratories that make up the networks were not chosen in any sort of systematic way that would methodically optimize disease detection. We recognize that surveillance networks are used to determine *outbreak location*, *outbreak intensity*, and *outbreak timing*. We explored how network size and design affects the quality of outbreak intensity and outbreak timing data the network is capable of measuring.

We studied three facilities location models for placing sites: two based on a maximal coverage model (MCM) and one based on a K-median model. The MCM selects sites that maximize the total number of people within a specified distance to the site. Recognizing that sites can only realistically see a limited number of patients during any given time period, we introduced a capacitated MCM, where each site has some intrinsic integer capacity. In practice, the capacitated MCM allows for multiple sites in the same general area with overlapping radii of coverage. The K-median model minimizes the sum of the distances from each individual to the individual's nearest site (i.e., it minimizes the average "travel" distance for a population).

To measure performance, we simulated the spread of influenza across the state of Iowa using the Huff spatial interaction model and a dataset consisting of two million de-identified Medicaid billing records representing eight complete influenza seasons. We compared hypothetical surveillance networks constructed using the MCM and K-median models against the existing Iowa Department of Public Health (IDPH) influenza-like illness surveillance network (consisting of 19 sites). We showed that both MCM models were capable generating networks that can better detect outbreak intensity than the IDPH network. Furthermore, we showed that it is possible to algorithmically create a smaller network capable of the same outbreak intensity performance as the existing IDPH network. We then analyzed outbreak timing by computing Pearson's product-moment correlation coefficient (Pearson's r). All algorithmic placement schemes are capable of generating networks that outperform the IDPH's network. Perhaps most importantly, though, we discovered that it may be possible to create drastically smaller networks ($n = 2$) that are capable of very good outbreak timing detection ($r \geq 0.9$).

We then demonstrated that the location-allocation models used for influenza surveillance site placement are general and can be used for other purposes. We showed how primary stroke centers can be algorithmically selected. We also demonstrated the ineffectiveness of the current self-initiated approach to primary stroke center certification and argue that a more systematic approach would be preferable.

While these traditional disease surveillance systems are important, they have several disadvantages. First, due to a complex reporting hierarchy, there is a reporting lag for most diseases in the U.S. of around 1–2 weeks. Second, many regions of the world lack trustworthy and reliable data. As a result, there has been a surge of research looking at using publicly available data on the internet for disease surveillance purposes. The second and third studies in this dissertation analyze Wikipedia's viability in this sphere.

The first of these two studies analyzed Wikipedia access logs. Hourly access logs dating back to December 2007 are available for anyone to download completely free of charge. These logs contain, among other things, the total number of accesses for every article in Wikipedia. Using a linear model and a simple article selection procedure, we were able to correctly *nowcast* and, in some cases, *forecast* up to the 28 days tested in 8 of the 14 disease-location contexts considered. We also demonstrated that it may be possible in some cases to train a model in one context and use the same model to nowcast or forecast in another context with poor surveillance data.

The second of the Wikipedia studies analyzed disease-relevant data found in the article content. A number of disease outbreaks are meticulously tracked on Wikipedia. Case counts, death counts, and hospitalization counts are often provided in the article narrative. Using a dataset created from 14 Wikipedia articles, we trained a named-entity recognizer (NER) to recognize and tag these phrases. The NER achieved an F1 score of 0.753. In addition to these counts in the narrative, we tested the accuracy of tabular data using the 2014 West African Ebola virus disease outbreak. This article, like a number of other outbreak articles on Wikipedia, contains granular case counts and deaths counts per country affected by the disease. By computing the root-mean-square error between the Wikipedia time series and a ground truth time series, we show that the Wikipedia time series are both timely and accurate.

Ultimately, we envision the Wikipedia work presented here being incorporated into the first community-driven open-source emerging disease detection and monitoring system. The primary draw to Wikipedia for us is its openness; a fundamental tenet of the Wikimedia Foundation is free and open knowledge. We feel that it is this dedication to openness that can direct the next generation internet-based disease surveillance systems.

REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using Twitter data. In *Proc. Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707. IEEE, 2011.

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Twitter improves seasonal influenza prediction. In *Proc. Health Informatics (HEALTHINF)*, page 61–70, 2012.

[3] Deborah A Adams, Kathleen M Gallagher, Ruth Ann Jajosky, Jeffrey Kriseman, Pearl Sharp, Willie J Anderson, Aaron E Aranas, Michelle Mayes, Michael S Wodajo, Diana H Onweh, and John P Abellera. Summary of Notifiable Diseases, United States, 2011. Technical Report 53, Centers for Disease Control and Prevention, Atlanta, Georgia, July 2013.

[4] Robert Adams, Joe Acker, Mark Alberts, Liz Andrews, Richard Atkinson, Kathy Fenelon, Anthony Furlan, Meighan Girgus, Katie Horton, Richard Hughes, Walter Koroshetz, Richard Latchaw, Ellen Magnis, Marc Mayberg, Arthur Pancioli, Rose Marie Robertson, Tim Shephard, Renee Smith, Sidney C Smith, Jr, Suzanne Smith, Steven K Stranne, Edgar J Kenton, III, Gil Bashe, Altagracia Chavez, Larry B Goldstein, Richard Hodosh, Cindy Keitel, Margaret Kelly-Hayes, Anne Leonard, Lewis Morgenstern, and Jack Owen Wood. Recommendations for Improving the Quality of Care Through Stroke Centers and Systems: An Examination of Stroke Center Identification Options: Multidisciplinary Consensus Recommendations from the Advisory Working Group on Stroke Center Identification Optio. *Stroke*, 33(1):e1–e7, January 2002.

[5] Byung Gyu Ahn, Benjamin Van Durme, and Chris Callison-Burch. WikiTopics: What is popular on Wikipedia and why. In *Proc. Workshop on Automatic Summarization for Different Genres, Media, and Languages (WASDGML)*, page 33–40. Association for Computational Linguistics, 2011.

[6] Murray Aitken, Thomas Altmann, and Daniel Rosen. Engaging patients through social media. Tech report, IMS Institute for Healthcare Informatics, 2014.

[7] Mark J Alberts, George Hademenos, Richard E Latchaw, Andrew Jagoda, John R Marler, Marc R Mayberg, Rodman D Starke, Harold W Todd, Kenneth M Viste, Meighan Girgus, Tim Shephard, Marian Emr, Patti Shwayder, and Michael D Walker. Recommendations for the Establishment of Primary Stroke Centers. *JAMA*, 283(23):3102–3109, June 2000.

[8] Karen C Albright, Charles C Branas, Brett C Meyer, Dawn E Matherne-Meyer, Justin A Zivin, Patrick D Lyden, and Brendan G Carr. ACCESS: Acute Cerebrovascular Care in Emergency Stroke Systems. *Archives of Neurology*, 67(10):1210–1218, October 2010.

[9] Alexa Internet, Inc. Alexa top 500 global sites, 2013. Accessed December 23, 2013.

[10] Tim Althoff, Damian Borth, Jörn Hees, and Andreas Dengel. Analysis and forecasting of trending topics in online media streams. In *Proc. Multimedia*, page 907–916. ACM, 2013.

[11] Benjamin M. Althouse, Yih Yng Ng, and Derek A. T. Cummings. Prediction of dengue incidence using search query surveillance. *PLOS Neglected Tropical Diseases*, 5(8):e1258, August 2011.

[12] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, page 1568–1576. Association for Computational Linguistics, 2011.

[13] Andrew W Asimos, Dianne Enright, Sara L Huston, and Laurie H Mettam. Drive-time proximity to Joint Commission Primary Stroke Centers among North Carolina residents who died of stroke. *North Carolina Medical Journal*, 71(5):413–420, 2010.

[14] John W. Ayers, Benjamin M. Althouse, Jon-Patrick Allem, J. Niels Rosenquist, and Daniel E. Ford. Seasonality in seeking mental health information on Google. *American Journal of Preventive Medicine*, 44(5):520–525, May 2013.

[15] Bharati Ayyalasomayajula, Natasha Wiebe, Brenda R Hemmelgarn, Aminu Bello, Braden Manns, Scott Klarenbach, and Marcello Tonelli. A Novel Technique to Optimize Facility Locations of New Nephrology Services for Remote Areas. *Clinical Journal of the American Society of Nephrology*, 6(9):2157–2164, September 2011.

[16] M Batty and S Mackie. The calibration of gravity, entropy, and related models of spatial interaction. *Environment and Planning*, 4(2):205–233, 1972.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Machine Learning Research*, 3:993–1022, 2003.

[18] Stephanie M. Borchardt, Kathleen A. Ritger, and Mark S. Dworkin. Categoriza-

tion, prioritization, and surveillance of potential bioterrorism agents. *Infectious Disease Clinics of North America*, 20(2):213–225, June 2006.

[19] Francis P Boscoe, Kevin A Henry, and Michael S Zdeb. A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals. *The Professional Geographer*, 64(2):188–196, May 2012.

[20] Dena M. Bravata, Kathryn M. McDonald, Wendy M. Smith, Chara Rydzak, Herbert Szeto, David L. Buckeridge, Corinna Haberland, and Douglas K. Owens. Systematic review: Surveillance systems for early detection of bioterrorism-related diseases. *Annals of Internal Medicine*, 140(11):910–922, June 2004.

[21] Benjamin N. Breyer, Saunak Sen, David S. Aaronson, Marshall L. Stoller, Bradley A. Erickson, and Michael L. Eisenberg. Use of Google Insights for Search to track seasonal and geographic kidney stone incidence in the United States. *Urology*, 78(2):267–271, August 2011.

[22] David A. Broniatowski, Michael J. Paul, and Mark Dredze. National and local influenza surveillance through Twitter: An analysis of the 2012–2013 influenza epidemic. *PLOS ONE*, 8(12):e83672, December 2013.

[23] Jan Burdziej and Piotr Gawrysiak. Using Web mining for discovering spatial patterns and hot spots for spatial generalization. In Li Chen, Alexander Felfernig, Jiming Liu, and Zbigniew W. Raś, editors, *Foundations of Intelligent Systems*, pages 172–181. Springer, January 2012.

[24] Bureau of Epidemiology. Weekly epidemiological surveillance report, Thailand. Accessed March 25, 2014.

[25] Guthrie S Burkhead and Christopher M Maylahn. State and Local Public Health Surveillance. In Steven M Teutsch and R Elliot Churchill, editors, *Principles and Practice of Public Health Surveillance*, chapter 12, pages 253–286. Oxford University Press, New York, 2nd edition, 2000.

[26] Declan Butler. When Google got flu wrong. *Nature*, 494(7436):155–156, 2013.

[27] Herman Anthony Carneiro and Eleftherios Mylonakis. Google Trends: A Web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564, November 2009. PMID: 19845471.

[28] Centers for Disease Control and Prevention (CDC). FluView. Accessed March 25, 2014.

[29] Centers for Disease Control and Prevention (CDC). Morbidity and mortality weekly report (MMWR) tables. Accessed March 25, 2014.

[30] Centers for Disease Control and Prevention (CDC). Overview of influenza surveillance in the United States, October 2012.

[31] Emily H. Chan, Vikram Sahai, Corrie Conrad, and John S. Brownstein. Using Web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLOS Neglected Tropical Diseases*, 5(5):e1206, May 2011.

[32] Chinese Center for Disease Control and Prevention. Notifiable infectious diseases statistic data. Accessed January 23, 2014.

[33] Sungjin Cho, Chang Hwan Sohn, Min Woo Jo, Soo-Yong Shin, Jae Ho Lee, Seoung Mok Ryoo, Won Young Kim, and Dong-Woo Seo. Correlation between national influenza surveillance data and Google Trends in South Korea. *PLOS ONE*, 8(12):e81422, December 2013.

[34] Rumi Chunara, Susan Aman, Mark Smolinski, and John S. Brownstein. Flu Near You: An online self-reported influenza surveillance system in the USA. *Online Journal of Public Health Informatics*, 5(1), March 2013.

[35] Rumi Chunara, Jason R Andrews, and John S Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, January 2012.

[36] Rumi Chunara, Vina Chhaya, Sunetra Bane, Sumiko R. Mekaru, Emily H. Chan, Clark C. Freifeld, and John S. Brownstein. Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010–2011. *Malaria Journal*, 11(1):43, February 2012.

[37] Richard Church. Location modelling and GIS. In Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind, editors, *Geographical Information Systems: Principles, Techniques, Management and Applications*, pages 293–303. John Wiley & Sons, Inc., New York, 2 edition, 1999.

[38] Richard Church and Charles ReVelle. The maximal covering location problem. *Papers in Regional Science*, 32(1):101–118, 1974.

[39] Marek Ciglan and Kjetil Nørvåg. WikiPop: Personalized event detection system based on Wikipedia page view statistics. In *Proc. Information and Knowledge Management (CIKM)*, page 1931–1932. ACM, 2010.

[40] Jacob Cohen. The Significance of a Product Moment r. In *Statistical Power Analysis for the Behavioral Sciences*, pages 75–107. Lawrence Erlbaum Associates, Inc., 2 edition, 1988.

[41] Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. BioCaster: Detecting public health rumors with a Web-based text mining system. *Bioinformatics*, 24(24):2940–2941, December 2008. PMID: 18922806.

[42] Crystale Purvis Cooper, Kenneth P Mallon, Steven Leadbetter, Lori A Pollack, and Lucy A Peipins. Cancer internet search activity on a major search engine, United States 2001–2003. *Journal of Medical Internet Research*, 7(3), July 2005. PMID: 15998627 PMCID: PMC1550657.

[43] Leon Cooper. Location-Allocation Problems. *Operations Research*, 11(3):331–343, May 1963.

[44] Gerard Cornuejols, Marshall L Fisher, and George L Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23(8):789–810, 1977.

[45] Steven C Cramer, Dana Stradling, David M Brown, Ignacio M Carrillo-Nunez, Anthony Ciabarra, Michael Cummings, Richard Dauben, David L Lombardi, Nirav Patel, Elizabeth N Traynor, Stephen Waldman, Ken Miller, and Samuel J Stratton. Organization of a United States County System for Comprehensive Acute Stroke Care. *Stroke*, 43(4):1089–1093, April 2012.

[46] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, pages 115–122, Washington, DC, 2010. ACM Press.

[47] Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, 47(1):217–238, March 2013.

[48] Mark S Daskin and Latoya K Dean. Location of health care facilities. In Margaret L Brandeau, François Sainfort, and William P Pierskalla, editors, *Operations Research and Health Care*, volume 70, pages 43–76. Springer US, 2005.

[49] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. *Communications of the ACM (CACM)*, 51(1):107–113, January 2008.

[50] Paul J Densham and Gerard Rushton. A more efficient heuristic for solving large p-median problems. *Papers in Regional Science*, 71(3):307–329, 1992.

[51] Rishi Desai, Aron J. Hall, Benjamin A. Lopman, Yair Shimshoni, Marcus Rennick, Niv Efron, Yossi Matias, Manish M. Patel, and Umesh D. Parashar. Norovirus disease surveillance using Google internet query share data. *Clinical Infectious Diseases*, 55(8):e75–e78, October 2012. PMID: 22715172.

[52] S. Doan, L. Ohno-Machado, and N. Collier. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In *Proc. Healthcare Informatics, Imaging and Systems Biology (HISB)*, pages 62–71. IEEE, 2012.

[53] Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, and Richard E. Rothman. Influenza forecasting with Google Flu Trends. *PLOS ONE*, 8(2):e56176, February 2013.

[54] Vanja M. Dukic, Michael Z. David, and Diane S. Lauderdale. Internet queries and methicillin-resistant Staphylococcus aureus surveillance. *Emerging Infectious Diseases*, 17(6):1068–1070, June 2011.

[55] Stephen Eubank, Hasan Guclu, V S Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltán Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[56] Gunther Eysenbach. Infodemiology: Tracking flu-related searches on the Web for syndromic surveillance. *AMIA Annual Symposium*, 2006:244–248, 2006. PMID: 17238340 PMCID: PMC1839505.

[57] Geoffrey Fairchild, Philip M Polgreen, Eric Foster, Gerard Rushton, and Alberto M Segre. How many suffice? A computational framework for sizing sentinel surveillance networks. *International Journal of Health Geographics*, 12(1):56, December 2013.

[58] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, number June, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[59] Centers for Disease Control and Prevention. Ebola (Ebola Virus Disease). `http://www.cdc.gov/vhf/ebola/`. Accessed: 2014-10-27.

[60] Centers for Disease Control and Prevention. Foodborne Illness Surveillance, Response, and Data Systems. http://www.cdc.gov/foodborneburden/surveillance-systems.html. Accessed: 2014-04-09.

[61] Centers for Disease Control and Prevention. How Flu Spreads. http://www.cdc.gov/flu/about/disease/spread.htm. Accessed: 2014-11-26.

[62] Centers for Disease Control and Prevention. National Notifiable Diseases Surveillance System (NNDSS). http://wwwn.cdc.gov/nndss/. Accessed: 2014-04-10.

[63] Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States. http://www.cdc.gov/flu/weekly/overview.htm. Accessed: 2014-04-09.

[64] Centers for Disease Control and Prevention. Seasonal Influenza Q&A. http://www.cdc.gov/flu/about/qa/disease.htm. Accessed: 2014-04-10.

[65] Wikimedia Foundation. 2012 Middle East respiratory syndrome coronavirus outbreak. http://en.wikipedia.org/w/index.php?title=2012_Middle_East_respiratory_syndrome_coronavirus_outbreak&oldid=628796140. Accessed: 2014-10-10.

[66] Wikimedia Foundation. Ebola virus epidemic in West Africa. http://en.wikipedia.org/w/index.php?title=Ebola_virus_epidemic_in_West_Africa&oldid=629094432. Accessed: 2014-10-10.

[67] Wikimedia Foundation. English Wikipedia. http://en.wikipedia.org/w/index.php?title=English_Wikipedia&oldid=627512912. Accessed: 2014-10-07.

[68] Wikimedia Foundation. Named-entity recognition. http://en.wikipedia.org/w/index.php?title=Named-entity_recognition&oldid=627138157. Accessed: 2014-10-11.

[69] Wikimedia Foundation. Wikipedia Statistics. http://stats.wikimedia.org/EN/Sitemap.htm. Accessed: 2014-10-07.

[70] Wikimedia Foundation. Wikipedia:Database download. http://en.wikipedia.org/w/index.php?title=Wikipedia:Database_download&oldid=627253774. Accessed: 2014-10-08.

[71] Wikimedia Foundation. Wikipedia:Five pillars. https://en.wikipedia.org/

w/index.php?title=Wikipedia:Five_pillars&oldid=632392090. Accessed: 2014-11-05.

[72] Susannah Fox. Online health search 2006. Technical report, Pew Research Center, October 2006.

[73] Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, 2008.

[74] Zong Woo Geem. Optimal cost design of water distribution networks using harmony search. *Engineering Optimization*, 38(3):259–277, April 2006.

[75] Zong Woo Geem. Harmony Search Algorithm for Solving Sudoku. In Bruno Apolloni, Robert J Howlett, and Lakhmi Jain, editors, *Knowledge-Based Intelligent Information and Engineering Systems*, pages 371–378. Springer Berlin Heidelberg, 2007.

[76] Zong Woo Geem. Optimal Scheduling of Multiple Dam System Using Harmony Search Algorithm. In Francisco Sandoval, Alberto Prieto, Joan Cabestany, and Manuel Graña, editors, *Computational and Ambient Intelligence*, pages 316–323. Springer Berlin Heidelberg, 2007.

[77] Zong Woo Geem and Jeong-Yoon Choi. Music Composition Using Harmony Search Algorithm. In Mario Giacobini, editor, *Applications of Evolutionary Computing*, pages 593–600. Springer Berlin Heidelberg, 2007.

[78] Zong Woo Geem, Joong Hoon Kim, and G V Loganathan. A New Heuristic Optimization Algorithm: Harmony Search. *Simulation*, 76(2):60–68, February 2001.

[79] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, and Reid Priedhorsky. Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*, 10(11):e1003892, November 2014.

[80] Timothy C Germann, Kai Kadau, Ira M Longini, Jr, and Catherine A Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5935–5940, April 2006.

[81] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[82] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.

[83] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proc. Web Science Conference (WebSci)*, page 1–8. ACM, 2011.

[84] Larry Greenemeier. Smart Machines Join Humans in Tracking Africa Ebola Outbreak. *Scientific American*, September 2014.

[85] Akihito Hagihara, Shogo Miyazaki, and Takeru Abe. Internet suicide searches and the incidence of suicide in young people in Japan. *European Archives of Psychiatry and Clinical Neuroscience*, 262(1):39–46, February 2012.

[86] George H Haines, Jr, Leonard S Simon, and Marcus Alexis. Maximum Likelihood Estimation of Central-City Food Trading Areas. *Journal of Marketing Research*, 9(2):154–159, 1972.

[87] R P Haining. Estimating spatial-interaction models. *Environment and Planning A*, 10(3):305–320, 1978.

[88] J Churchill Hindes. *Resource Allocation and Cardiovascular Disease: Location of Hospital Cardiovascular Services in the Iowa Region.* Ph.d., University of Iowa, 1977.

[89] H. Hirose and Liangliang Wang. Prediction of infectious disease spread using Twitter: A case of influenza. In *Proc. Parallel Architectures, Algorithms and Programming (PAAP)*, pages 100–105. IEEE, 2012.

[90] M J Hodgson. Toward more realistic allocation in location - allocation models: an interaction approach. *Environment and Planning A*, 10(11):1273–1285, 1978.

[91] Martin Rudi Holaker and Eirik Emanuelsen. *Event Detection using Wikipedia.* Master's thesis, Institutt for datateknikk og informasjonsvitenskap, 2013.

[92] David L Huff. A probabilistic analysis of shopping center trade areas. *Land Economics*, 39(1):81–90, 1963.

[93] A Hulth and G Rydevik. Web query-based surveillance in Sweden during the influenza A(H1N1)2009 pandemic, April 2009 to February 2010. *Euro Surveillance*, 16(18), 2011. PMID: 21586265.

[94] Anette Hulth, Yvonne Andersson, Kjell-Olof Hedlund, and Mikael Andersson. Eye-opening approach to norovirus surveillance. *Emerging Infectious Diseases*, 16(8):1319–1321, August 2010. PMID: 20678337 PMCID: PMC3298324.

[95] Anette Hulth, Gustaf Rydevik, and Annika Linde. Web queries as a source for syndromic surveillance. *PLOS ONE*, 4(2):e4378, February 2009.

[96] Lori Hutwagner, William Thompson, G. Matthew Seeman, and Tracee Treadwell. The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health*, 80(Suppl 1):i89–i96, March 2003. PMID: 12791783 PMCID: PMC3456557.

[97] Ruth Ann Jajosky and Samuel L Groseclose. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health*, 4:29, July 2004.

[98] Hongzhong Jia, Fernando Ordóñez, and Maged Dessouky. A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions*, 39(1):41–55, January 2007.

[99] Bao Jia-xing, Lv Ben-fu, Peng Geng, and Li Na. Gonorrhea incidence forecasting research based on Baidu search data. In *Proc. Management Science and Engineering (ICMSE)*, pages 36–42. IEEE, 2013.

[100] Heather A Johnson, Michael M Wagner, William R Hogan, Wendy Chapman, Robert T Olszewski, John Dowling, and Gary Barnas. Analysis of Web access logs for surveillance of influenza. *Studies in Health Technology and Informatics*, 107(2):1202–1206, 2004. PMID: 15361003.

[101] Min Kang, Haojie Zhong, Jianfeng He, Shannon Rutherford, and Fen Yang. Using Google Trends for influenza surveillance in South China. *PLOS ONE*, 8(1):e55205, January 2013.

[102] O Kariv and Seifollah L Hakimi. An algorithmic approach to network location problems. II: The p-medians. *SIAM Journal on Applied Mathematics*, 37(3):539–560, 1979.

[103] Eui-Ki Kim, Jong Hyeon Seok, Jang Seok Oh, Hyong Woo Lee, and Kyung Hyun Kim. Use of Hangeul Twitter to track and predict human influenza infection. *PLOS ONE*, 8(7):e69305, July 2013.

[104] Joong Hoon Kim, Zong Woo Geem, and Eung Seok Kim. Parameter Estimation of the Nonlinear Muskingum Model Using Harmony Search. *Journal of the American Water Resources Association*, 37(5):1131–1138, October 2001.

[105] Nicholas E. Kman and Daniel J. Bachmann. Biosurveillance: A review and update. *Advances in Preventive Medicine*, 2012, January 2012.

[106] Natalie Kupferberg and Bridget McCrate Protus. Accuracy and completeness of drug information in Wikipedia: An assessment. *Journal of the Medical Library Association*, 99(4):310–313, October 2011. PMID: 22022226 PMCID: PMC3193353.

[107] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proc. Human Language Technologies (NAACL-HLT)*, page 789–795. North American Chapter of the Association for Computational Linguistics, 2013.

[108] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social Web. In *Proc. Cognitive Information Processing (CIP)*, pages 411–416. IEEE, June 2010.

[109] Vasileios Lampos and Nello Cristianini. Nowcasting events from the social Web with statistical learning. *Transactions on Intelligent Systems and Technology*, 3(4):72:1–72:22, September 2012.

[110] Michaël R. Laurent and Tim J. Vickers. Seeking health information online: Does Wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, July 2009.

[111] Kennedy R Lees, Erich Bluhmki, Rüdiger von Kummer, Thomas G Brott, Danilo Toni, James C Grotta, Gregory W Albers, Markku Kaste, John R Marler, Scott A Hamilton, Barbara C Tilley, Stephen M Davis, Geoffrey A Donnan, and Werner Hacke. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. *Lancet*, 375(9727):1695–1703, May 2010.

[112] Enrique C Leira, Geoffrey Fairchild, Alberto M Segre, Gerard Rushton, Michael T Froehler, and Philip M Polgreen. Primary Stroke Centers Should Be Located Using Maximal Coverage Models for Optimal Access. *Stroke*, 43(9):2417–2422, 2012.

[113] Enrique C Leira, David C Hess, James C Torner, and Harold P Adams Jr. Rural-Urban Differences in Acute Stroke Management Practices: A Modifiable Disparity. *Archives of Neurology*, 65(7):887–891, July 2008.

[114] Enrique C Leira, Jennifer K Pary, Patricia H Davis, Karla J Grimsman, and Harold P Adams. Slow Progressive Acceptance of Intravenous Thrombolysis for

Patients With Stroke by Rural Primary Care Physicians. *Archives of Neurology*, 64(4):518–521, April 2007.

[115] Andreas Leithner, Werner Maurer-Ertl, Mathias Glehr, Joerg Friesenbichler, Katharina Leithner, and Reinhard Windhager. Wikipedia and osteosarcoma: A trustworthy patients' information? *Journal of the American Medical Informatics Association*, 17(4):373–374, July 2010. PMID: 20595302.

[116] Johan Lindh, Måns Magnusson, Maria Grünewald, and Anette Hulth. Head lice surveillance on a deregulated OTC-sales market: A study using Web query data. *PLOS ONE*, 7(11):e48666, November 2012.

[117] Eric Lofgren, Nina H Fefferman, Yuri N Naumov, Jack Gorski, and Elena N Naumova. Influenza Seasonality: Underlying Causes and Modeling Theories. *Journal of Virology*, 81(11):5429–5436, June 2007.

[118] Alan D Lopez, Colin D Mathers, Majid Ezzati, Dean T Jamison, and Christopher JL Murray. Global and regional burden of disease and risk factors, 2001: Systematic analysis of population health data. *The Lancet*, 367(9524):1747–1757, May 2006.

[119] Joseph Z Losos. Routine and sentinel surveillance methods. *Eastern Mediterranean Health Journal*, 2(1):46–50, 1996.

[120] Kevin Lutsky, Joseph Bernstein, and Pedro Beredjiklian. Quality of information on the internet about carpal tunnel syndrome: An update. *Orthopedics*, 36(8):e1038–e1041, 2013.

[121] Lawrence C Madoff. ProMED-mail: An Early Warning System for Emerging Diseases. *Clinical Infectious Diseases*, 39(2):227–232, July 2004.

[122] M Mahdavi, M Fesanghary, and E Damangir. An improved harmony search algorithm for solving optimization problems. *Applied Mathematics and Computation*, 188(2):1567–1579, May 2007.

[123] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Number c. Cambridge University Press, Cambridge, England, 2009.

[124] Nicola Marsden-Haug, Virginia B Foster, Philip L Gould, Eugene Elbert, Hailiang Wang, and Julie A Pavlin. Code-based Syndromic Surveillance for Influenzalike Illness by International Classification of Diseases, Ninth Revision. *Emerging Infectious Diseases*, 13(2):207–216, February 2007.

[125] David J McIver and John S Brownstein. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLOS Computational Biology*, 10(4):e1003581, April 2014.

[126] Jan Medlock and Alison P Galvani. Optimizing Influenza Vaccine Distribution. *Science*, 325(5948):1705–1708, September 2009.

[127] Meldingssystem for Smittsomme Sykdommer. MSIS statistikk. Accessed January 28, 2014.

[128] Márton Mestyán, Taha Yasseri, and János Kertész. Early prediction of movie box office success based on Wikipedia activity big data. *PLOS ONE*, 8(8):e71226, August 2013.

[129] Madeline L Miley, Bart M Demaerschalk, Nicole L Olmstead, Terri-Ellen J Kiernan, Doren A Corday, Vatsal Chikani, and Bentley J Bobrow. The State of Emergency Stroke Resources and Care in Rural Arizona: A Platform for Telemedicine. *Telemedicine and e-Health*, 15(7):691–699, September 2009.

[130] Ministère de la Santé Publique et de la Population. Centere de documentation. Accessed January 23, 2014.

[131] Ministère de la Santé Publique, République Démocratique du Congo. Fièvre hémorragique à virus Ebola dans le district de haut uele, October 2012. Accessed March 25, 2014.

[132] Ministério da Saúde. Portal da saúde. Accessed September 26, 2013.

[133] Susan M Mniszewski, Sara Y Del Valle, Phillip D Stroud, Jane M Riese, and Stephen J Sydoriak. Pandemic simulation of antivirals + school closures: buying time until strain-specific vaccine is available. *Computational and Mathematical Organization Theory*, 14(3):209–221, April 2008.

[134] Susan M. Mniszewski, Sara Y. Del Valle, Reid Priedhorsky, James M. Hyman, and Kyle S. Hickman. Understanding the impact of face mask usage through epidemic simulation of large social networks. In Vahid Dabbaghian and Vijay Kumar Mago, editors, *Theories and Simulations of Complex Social Systems*, pages 97–115. Springer, January 2014.

[135] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, 3, May 2013.

[136] Noelle-Angelique M. Molinari, Ismael R. Ortega-Sanchez, Mark L. Messonnier, William W. Thompson, Pascale M. Wortley, Eric Weintraub, and Carolyn B. Bridges. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, June 2007.

[137] Anna C Nagel, Ming-Hsiang Tsou, Brian H Spitzberg, Li An, J Mark Gawron, Dipak K Gupta, Jiue-An Yang, Su Han, K Michael Peddecord, Suzanne Lindsay, and Mark H Sawyer. The complex relationship of realspace events and messages in cyberspace: Case study of influenza and pertussis using Tweets. *Journal of Medical Internet Research*, 15(10), October 2013. PMID: 24158773 PMCID: PMC3841359.

[138] Masao Nakanishi and Lee G Cooper. Parameter Estimation for a Multiplicative Competitive Interaction Model: Least Squares Approach. *Journal of Marketing Research*, 11(3):303–311, 1974.

[139] National Institute of Infectious Diseases, Japan. Infectious diseases weekly report. Accessed January 24, 2013.

[140] National Institute of Public Health. Influenza in Poland. Accessed September 24, 2013.

[141] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In Thomas Glen Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Proceedings of the 2001 Neural Information Processing Systems Conference*, pages 841–848, British Columbia, Canada, 2002. MIT Press.

[142] Alex J. Ocampo, Rumi Chunara, and John S. Brownstein. Using search queries for malaria surveillance, Thailand. *Malaria Journal*, 12(1):390, November 2013. PMID: 24188069.

[143] Nicholas J Okon, Crystelle C Fogle, Michael J McNamara, Carrie S Oser, Dennis W Dietrich, Dorothy Gohdes, Todd S Harwell, Daniel V Rodriguez, and Steven D Helgerson. Statewide Efforts to Narrow the Rural–Urban Gap in Acute Stroke Care. *American Journal of Preventive Medicine*, 39(4):329–333, October 2010.

[144] Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLOS Computational Biology*, 9(10):e1003256, October 2013.

[145] Mahamed G H Omran and Mehrdad Mahdavi. Global-best harmony search. *Applied Mathematics and Computation*, 198(2):643–656, May 2008.

[146] World Health Organization. About WHO. http://www.who.int/about/en/. Accessed: 2014-04-09.

[147] World Health Organization. Ebola virus disease. http://www.who.int/mediacentre/factsheets/fs103/en/. Accessed: 2014-10-27.

[148] World Health Organization. Ebola virus disease in Guinea (Situation as of 25 March 2014). http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4065-ebola-virus-disease-in-guinea-25-march-2014.html. Accessed: 2014-12-01.

[149] Miles Osborne, Sasa Petrovic, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using Twitter and Wikipedia. In *Proc. SIGIR Workshop on Time-aware Information Access (TAIA)*. ACM, 2012.

[150] Laurence J O'Toole, Catherine P Slade, Gene A Brewer, and Lauren N Gase. Barriers and Facilitators to Implementing Primary Stroke Center Policy in the United States: Results From 4 Case Study States. *American Journal of Public Health*, 101(3):561–566, March 2011.

[151] Arsalan Paleshi, Gerald W Evans, Sunderesh S Heragu, and Kamran S Moghaddam. Simulation of mitigation strategies for a pandemic influenza. In *Proceedings of the 2011 Winter Simulation Conference*, pages 1340–1348, 2011.

[152] Michael J. Paul and Mark Dredze. You are what you Tweet: Analyzing Twitter for public health. In *Proc. Weblogs and Social Media (ICWSM)*. AAAI, 2011.

[153] Karl Pearson. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318, January 1896.

[154] Ashley S Pedigo and Agricola Odoi. Investigation of Disparities in Geographic Accessibility to Emergency Stroke and Myocardial Infarction Care in East Tennessee Using Geographic Information Systems and Network Analysis. *Annals of Epidemiology*, 20(12):924–930, December 2010.

[155] Camille Pelat, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. More diseases tracked by using Google Trends. *Emerging*

Infectious Diseases, 15(8):1327–1328, August 2009. PMID: 19751610 PMCID: PMC2815981.

[156] Philip M Polgreen, Yiling Chen, David M Pennock, and Forrest D Nelson. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, December 2008.

[157] Philip M Polgreen, Zunqui Chen, Alberto M Segre, Meghan L Harris, Michael A Pentella, and Gerard Rushton. Optimizing Influenza Sentinel Surveillance at the State Level. *American Journal of Epidemiology*, 170(10):1300–1306, November 2009.

[158] Reid Priedhorsky, Jilin Chen, S. K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. Supporting Group Work (GROUP)*. ACM, 2007.

[159] Shams-ur Rahman and David K Smith. Use of location-allocation models in health service development planning in developing nations. *European Journal of Operational Research*, 123(3):437–452, June 2000.

[160] Malolan S. Rajagopalan, Vineet K. Khanna, Yaacov Leiter, Meghan Stott, Timothy N. Showalter, Adam P. Dicker, and Yaacov R. Lawrence. Patient-oriented cancer information on the internet: A comparison of Wikipedia and a professionally maintained database. *Journal of Oncology Practice*, 7(5):319–323, September 2011. PMID: 22211130.

[161] Ronald E. Rice. Influences, usage, and outcomes of internet health information searching: Multivariate results from the Pew surveys. *International Journal of Medical Informatics*, 75(1):8–28, January 2006.

[162] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In *Proc. 1st International Workshop on Mining Social Media*, 2009.

[163] Samuel V Scarpino, Nedialko B Dimitrov, and Lauren Ancel Meyers. Optimizing Provider Recruitment for Influenza Surveillance Networks. *PLOS Computational Biology*, 8(4):e1002472, January 2012.

[164] Ari Seifter, Alison Schwarzwalder, Kate Geis, and John Aucott. The utility of “Google Trends” for epidemiological research: Lyme disease as an example. *Geospatial Health*, 4(2):135–137, May 2010.

[165] Jeffrey Shaman, Alicia Karspeck, Wan Yang, James Tamerius, and Marc Lip-

sitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4, December 2013.

[166] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS ONE*, 6(5):e19467, January 2011.

[167] Eric E Smith, Paul Dreyer, Janet Prvu-Bettger, Abdul R Abdullah, Gail Palmeri, Louise Goyette, Cathleen McElligott, and Lee H Schwamm. Stroke Center Designation Can be Achieved by Small Hospitals: The Massachusetts Experience. *Critical Pathways in Cardiology*, 7(3):173–177, September 2008.

[168] Klaus Stöhr. The Global Agenda on Influenza Surveillance and Control. *Vaccine*, 21(16):1744–1748, May 2003.

[169] Stroke Unit Trialists' Collaboration. Organised inpatient (stroke unit) care for stroke. *Cochrane Database of Systematic Reviews*, 9(CD000197), January 2013.

[170] Charles Sutton. An Introduction to Conditional Random Fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, November 2011.

[171] Yla Tausczik, Kate Faasse, James W. Pennebaker, and Keith J. Petrie. Public anxiety and information seeking following the H1N1 outbreak: Blogs, newspaper articles, and Wikipedia visits. *Health Communication*, 27(2):179–185, 2012. PMID: 21827326.

[172] Michael B Teitz and Polly Bart. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research*, 16(5):955–961, 1968.

[173] Stephen B Thacker. Historical Development. In Steven M Teutsch and R Elliot Churchill, editors, *Principles and Practice of Public Health Surveillance*, chapter 1, pages 1–16. Oxford University Press, New York, 2nd edition, 2000.

[174] Marijn ten Thij, Yana Volkovich, David Laniado, and Andreas Kaltenbrunner. Modeling page-view dynamics on Wikipedia. *arXiv:1212.5943 [physics]*, December 2012.

[175] Garry R. Thomas, Lawson Eng, Jacob F. de Wolff, and Samir C. Grover. An evaluation of Wikipedia as a resource for patient education in nephrology. *Seminars in Dialysis*, 26(2):159–163, 2013.

[176] M.G. Thompson, D.K. Shay, H. Zhou, C.B. Bridges, P.Y. Cheng, E. Burns, J.S. Bresee, and N.J. Cox. Estimates of deaths associated with seasonal influenza — United States, 1976–2007. *Morbidity and Mortality Weekly Report*, 59(33), August 2010.

[177] Ramine Tinati, Thanassis Tiropanis, and Lesie Carr. An approach for using Wikipedia to measure the flow of trends across countries. In *Proc. World Wide Web (WWW) Companion*, page 1373–1378. ACM, 2013.

[178] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, volume 4, pages 142–147, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[179] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, volume 1, pages 173–180, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[180] Khoi-Nguyen Tran and Peter Christen. Cross language prediction of vandalism on Wikipedia using article views and revisions. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 268–279. Springer, January 2013.

[181] A Vasebi, M Fesanghary, and S M T Bathaee. Combined heat and power economic dispatch by harmony search algorithm. *International Journal of Electrical Power & Energy Systems*, 29(10):713–719, December 2007.

[182] Vedat Verter and Sophie D Lapierre. Location of preventive health care facilities. *Annals of Operations Research*, 110(1):123–132, 2002.

[183] Brian P. Walcott, Brian V. Nahed, Kristopher T. Kahle, Navid Redjal, and Jean-Valery Coumans. Determination of geographic variance in stroke prevalence using internet search engine analytics. *Journal of Neurosurgery*, 115(6):E19, June 2011.

[184] Wikimedia Foundation. Page views for Wikipedia, all platforms, normalized, December 2013. Accessed December 23, 2013.

[185] Wikipedia editors. Wikipedia, December 2013.

[186] Wikipedia editors. Wikipedia:Moving a page, April 2014. Section "Moving over a redirect".

[187] Kumanan Wilson and John S. Brownstein. Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal*, 180(8):829–831, April 2009.

[188] World Health Organization (WHO). Ebola in Uganda, May 2011. Accessed March 25, 2014.

[189] World Health Organization (WHO). Ebola outbreak in Democratic Republic of Congo – Update, October 2012. Accessed March 25, 2014.

[190] Danqing Xu, Yiqun Liu, Min Zhang, Shaoping Ma, Anqi Cui, and Liyun Ru. Predicting epidemic tendency through search behavior analysis. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, page 2361–2366. AAAI, 2011.

[191] Wei Xu, Zhen-Wen Han, and Jian Ma. A neural netwok [sic] based approach to detect influenza epidemics using search engine query data. In *Proc. Machine Learning and Cybernetics (ICMLC)*, pages 1408–1412. IEEE, 2010.

[192] Albert C. Yang, Shi-Jen Tsai, Norden E. Huang, and Chung-Kang Peng. Association of internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. *Journal of Affective Disorders*, 132(1–2):179–184, July 2011.

[193] Taha Yasseri and Jonathan Bright. Can electoral popularity be predicted using socially generated big data? *arXiv:1312.2818 [physics]*, December 2013.

[194] Qingyu Yuan, Elaine O. Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S. Brownstein. Monitoring influenza epidemics in China with search query from Baidu. *PLOS ONE*, 8(5):e64323, May 2013.

[195] Erik Zachte. readme.txt, May 2012. Accessed December 24, 2013.

[196] Andrey Zheluk, Casey Quinn, Daniel Hercz, and James A Gillespie. Internet search patterns of human immunodeficiency virus and the digital divide in the Russian Federation: Infoveillance study. *Journal of Medical Internet Research*, 15(11), November 2013. PMID: 24220250 PMCID: PMC3841350.

[197] Xi-chuan Zhou and Hai-bin Shen. Notifiable infectious disease surveillance with data collected by search engine. *Journal of Zhejiang University SCIENCE C*, 11(4):241–248, April 2010.

[198] Xichuan Zhou, Qin Li, Zhenglin Zhu, Han Zhao, Hao Tang, and Yujie Feng. Monitoring epidemic alert levels by analyzing internet search volume. *IEEE Transactions on Biomedical Engineering*, 60(2):446–452, 2013.