STATISTICAL ESTIMATION AND CHANGEPOINT DETECTION METHODS IN PUBLIC HEALTH SURVEILLANCE

A Thesis
Presented to
The Academic Faculty

by

Sue Bath Reynolds

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2015

STATISTICAL ESTIMATION AND CHANGEPOINT DETECTION METHODS IN PUBLIC HEALTH SURVEILLANCE

Approved by:

Dr. David Goldsman, Co-Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kwok-Leung Tsui, Co-Advisor
Department of Systems Engineering and Engineering Management
City University of Hong Kong

Dr. Brani Vidakovic
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Christos Alexopoulos
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. David Shay
Influenza Division, NCIRD
Centers for Disease Control and Prevention

Date Approved: April 1, 2015

To my family:
Jackson, Jim, and Mom

ACKNOWLEDGEMENTS

I would first like to thank all my advisors and committee members: Dr. Christos Alexopoulos, Dr. David Goldsman, Dr. Xiaoming Huo, Dr. David Shay, Dr. Kwok-Leung Tsui, and Dr. Brani Vidakovic. I appreciate all their suggestions and advice throughout this process.

In particular, I would like to thank Dr. Goldsman and Dr. Shay. I could not have completed this thesis without leaning heavily on their guidance, support, encouragement, and wealth of knowledge. I am truly grateful to them both.

I would also like to thank Dr. Tsui for helping me hone in on a thesis topic, and for his guidance and encouragement during a very difficult period. I truly appreciate his patience with me during those long intervals of little contact or progress.

A big thank you also to Dr. Vidakovic for his insight and suggestions on improving my Bayesian modeling methods; to Dr. Huo for suggesting a generalized cross validation method to improve my model fitting with smoothing splines; and to Dr. Alexopoulos for his suggestions on extending the change detection component to incorporate correlated observations. Thank you also to Ms. Pam Morrison for her kindness and encouragement all these years, and to the H. Milton School of Industrial and Systems Engineering for tolerating my highly unusual (and lengthy) student status.

I would also like to thank my family for their unwavering support and encouragement. I am forever indebted to my mother for her constant and unconditional support and kindness. A big thank you to my husband Jim for his support, and for helping me keep things in perspective. And last but not least, a big hug and thank you to my 11-year-old son Jackson, who also has a big graduation coming up soon. Thanks Jack for being the phenomenal person that you are, and for the endless supply of knock-knock jokes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis focuses on assessing and improving statistical methods implemented in two areas of public health research. The first topic involves estimation of national influenza-associated mortality rates via mathematical modeling. The second topic involves the timely detection of infectious disease outbreaks using statistical process control monitoring.

For over fifty years, the Centers for Disease Control and Prevention has been estimating annual rates of U.S. deaths attributable to influenza. These estimates have been used to determine costs and benefits associated with influenza prevention and control strategies. Quantifying the effect of influenza on mortality, however, can be challenging since influenza infections typically are not confirmed virologically nor specified on death certificates. Consequently, a wide range of ecologically based, mathematical modeling approaches have been applied to specify the association between influenza and mortality. To date, all influenza-associated death estimates have been based on mortality data first aggregated at the national level and then modeled. Unfortunately, there are a number of local-level seasonal factors that may confound the association between influenza and mortality – thus suggesting that data be modeled at the local level and then pooled to make national estimates of death.

The first component of the thesis topic involving mortality estimation addresses this issue by introducing and implementing a two-stage hierarchical Bayesian modeling approach. In the first stage, city-level data with varying trends in mortality and weather were modeled using semi-parametric, generalized additive models. In the second stage, the log-relative risk estimates calculated for each city in stage 1 represented the "outcome" variable, and were modeled two ways: (1) assuming spatial independence across cities using a Bayesian generalized linear model, and (2) assuming correlation

among cities using a Bayesian spatial correlation model. Results from these models were compared to those from a more-conventional approach.

The second component of this topic examines the extent to which seasonal confounding and collinearity affect the relationship between influenza and mortality at the local (city) level. Disentangling the effects of temperature, humidity, and other seasonal confounders on the association between influenza and mortality is challenging since these covariates are often temporally collinear with influenza activity. Three modeling strategies with varying representations of background seasonality were compared. Seasonal covariates entered into the model may have been measured (e.g., ambient temperature) or unmeasured (e.g., time-based smoothing splines or Fourier terms). An advantage of modeling background seasonality via time splines is that the amount of seasonal curvature can be controlled by the number of degrees of freedom specified for the spline. A comparison of the effects of influenza activity on mortality based on these varying representations of seasonal confounding is assessed.

The third component of this topic explores the relationship between mortality rates and influenza activity using a flexible, natural cubic spline function to model the influenza term. The conventional approach of fitting influenza-activity terms linearly in regression was found to be too constraining. Results show that the association is best represented nonlinearly.

The second area of focus in this thesis involves infectious disease outbreak detection. A fundamental goal of public health surveillance, particularly syndromic surveillance, is the timely detection of increases in the rate of unusual events. In syndromic surveillance, a significant increase in the incidence of monitored disease outcomes would trigger an alert, possibly prompting the implementation of an intervention strategy. Public health surveillance generally monitors count data (e.g., counts of influenza-like illness, sales of over-the-counter remedies, and number of visits to outpatient clinics). Statistical process control charts, designed for quality control

monitoring in industry, have been widely adapted for use in disease and syndromic surveillance. The behavior of these detection methods on discrete distributions, however, has not been explored in detail.

For this component of the thesis, a simulation study was conducted to compare the CuSum and EWMA methods for detection of increases in negative binomial rates with varying amounts of dispersion. The goal of each method is to detect an increase in the mean number of cases as soon as possible after an upward rate shift has occurred. The performance of the CuSum and EWMA detection methods is evaluated using the conditional expected delay criterion, which is a measure of the detection delay, i.e., the time between the occurrence of a shift and when that shift is detected. Detection capabilities were explored under varying shift sizes and times at which the shifts occurred.

CHAPTER 1

INTRODUCTION

The purpose of this chapter is to provide a brief overview of the two public health research areas covered in this thesis, and to outline the research questions addressed as well as the research contributions made in each chapter.

1.1 Statistical Estimation of Influenza-associated Mortality Rates

For over fifty years, the Centers for Disease Control and Prevention (CDC) has been estimating annual rates of U.S. deaths attributable to influenza [Serfling, 1963; Lui and Kendal, 1987; Simonsen et al., 2000; Thompson et al., 2003, 2009]. These estimates have been used to determine costs and benefits associated with influenza prevention and control strategies [Meltzer et al., 1999; Nichol, 2003; Bridges et al., 2003]. Quantifying the effect of influenza on mortality, however, can be challenging since influenza infections typically are not confirmed virologically nor specified on death certificates [Douglas, 1976; Bisno et al., 1971; Taubenberger and Morens, 2008; Collins & Lehmann, 1951]. Consequently, the CDC has applied a wide range of ecologically based, mathematical modeling approaches to specify the association between influenza and mortality. Some of the more-popular methods include regression, autoregressive integrated moving averages, and rate-differencing [Lui and Kendal, 1987; Thompson et al., 2003; Thompson et al., 2009; Cheng et al., 2014 unpublished]. To date, U.S. influenza-associated death estimates have been based on mortality data first aggregated at the national level and then modeled. Unfortunately, there are a number of local-level seasonal factors that may confound the association between influenza and mortality – thus suggesting that data be modeled at the local level and then pooled to make national estimates of death.

Several recent studies have shown that meteorological factors, e.g., temperature and humidity, may affect influenza-related mortality estimation [Warren-Gash et al., 2011; Yang et al., 2011; Wong et al., 2012; Yang et al., 2012]. Other seasonal viruses with varying local-level circulation, such as respiratory syncytial virus (RSV), can also be influential [Thompson et al., 2003; Mangtani et al., 2006]. Seasonal factors for which daily or weekly measurements are not collected are potential regional/local confounders as well (e.g., differences in predominant influenza viruses, seasonal host health, population density, and age distribution) [Lofgren et al., 2007].

1.2 Modeling Influenza-associated Mortality: Research Questions and Contributions

Chapters 2−5 of this thesis examine in detail the effects of local-level seasonal confounding on the association between influenza activity and mortality rates.

Chapter 2 is a literature review on modeling influenza-associated mortality. The review focuses on the two most-common approaches, generalized linear models (GLMs) and generalized additive models (GAMs), both of which are able to incorporate exogenous information into models and estimates. The review discusses remedial measures used by researchers to address the analytic limitations of modeling seasonal data via GLMs and GAMs such as multicollinearity, serial correlation, and temporal confounding.

Chapter 3 addresses the issue of local-level seasonal confounding in modeling influenza-associated mortality by introducing and implementing a two-stage hierarchical modeling approach. In the first stage, city-level data with varying trends in mortality and weather are modeled using semi-parametric, generalized additive models. In the second stage, the log-relative risk estimates calculated for each city in stage 1 represent the "outcome" variable, and are modeled two ways: (1) assuming spatial independence across cities using a Bayesian generalized linear model, and (2) assuming correlation

among cities using a Bayesian spatial correlation model. Results from these models are compared to those from a more-conventional approach.

Chapter 4 examines the extent to which seasonal confounding and collinearity affect the relationship between influenza and mortality at the local (city) level. Disentangling the effects of temperature, humidity, and other seasonal confounders on the association between influenza and mortality is challenging since these covariates are often temporally collinear with influenza activity. Three modeling strategies with varying representations of background seasonality are compared. Seasonal covariates entered into the model may have been measured (e.g., ambient temperature) or unmeasured (e.g., time-based smoothing splines or Fourier terms). An advantage of modeling background seasonality via time splines is that the amount of seasonal curvature can be controlled by the number of degrees of freedom specified for the spline. A comparison of the effects of influenza activity on mortality based on these varying representations of seasonal confounding is assessed.

Chapter 5 explores the relationship between mortality rates and influenza activity using a flexible, natural cubic spline function to model the influenza term. The conventional approach of fitting influenza-activity terms linearly in regression is found to be too constraining. Results show that the association is best represented nonlinearly.

1.3 Statistical Process Control Methods for Disease Outbreak Detection

A fundamental goal of public health surveillance, particularly syndromic surveillance, is the timely detection of increases in the rate of unusual events. In syndromic surveillance, a significant increase in the incidence of monitored disease outcomes would trigger an alert, possibly prompting the implementation of an intervention strategy. Public health surveillance generally monitors count data (e.g., counts of influenza-like illness, sales of over-the-counter remedies, and number of visits to outpatient clinics).

Statistical process control (SPC) charts, designed for quality control monitoring in industry, have been widely adapted for use in disease and syndromic surveillance [Tsui et al., 2008]. Adaptations of the Cumulative Sum (CuSum) and Exponentially Weighted Moving Average (EWMA) charts have been used to monitor counts of nosocomial infections [Benneyan, 1998; Brown et al., 2002], hospital emergency department visits [Burkom, 2003; Yuan et al., 2004; Ivanov et al., 2003], visits to medical facilities [Burkom, 2003; Yuan et al., 2004; Ivanov et al., 2003; Bradley et al., 2005], prescription drug sales [Chen et al., 2005], and sales of over-the-counter health care products [Burkom, 2003; Hogan et al., 2003; Marx et al., 2006]. When used in syndromic surveillance, a statistically significant increase in observed data demonstrated by a SPC chart might be considered evidence of an emerging outbreak.

1.4 Monitoring Disease Outcomes: Research Question and Contributions

Chapter 6 reviews the current literature on three SPC charts (CuSum, EWMA, and Shewart residual charts) used for detection of rate or count increases in public health surveillance. The behavior of these detection methods on discrete distributions, however, has not been explored in detail. Chapter 7 details a simulation study conducted to compare the CuSum and EWMA methods for detection of increases in negative binomial rates with varying amounts of dispersion. The goal of each method is to detect an increase in the mean number of cases as soon as possible after an upward rate shift occurs. Performance of the CuSum and EWMA detection methods is evaluated using the conditional expected delay criterion which is a measure of the detection delay, i.e., the time between the occurrence of a shift and when that shift is detected. Detection capabilities are explored under varying shift sizes and times at which the shifts occur.

CHAPTER 2

MODELING INFLUENZA-ASSOCIATED MORTALITY: A LITERATURE REVIEW

Estimating deaths associated with influenza using indirect statistical methods dates back to the mid-nineteenth century [Farr, 1847]. Deriving such estimates from ecologic, time-series data continues to be a popular approach, though it also remains challenging for a number of reasons. Since information is collected at the population rather than individual level, results are prone to analytic errors involving temporal ambiguity and misclassification within groups. Because data are collected over time, neighboring observations are often correlated, e.g., exhibiting seasonal or long-term patterns. Validation of estimates is also difficult given that influenza can vary dramatically in timing and magnitude each season. Moreover, any given influenza time series may not contain enough data over time to justify reaching a conclusion or generalizing the results.

To address these and other methodological issues, a wide range of advanced mathematical modeling approaches have been implemented to describe statistically the association between influenza and mortality. Two commonly used approaches are generalized linear models (GLMs) and generalized additive models (GAMs). The main advantage these have over more-traditional time series methods, such as autoregressive integrated moving average (ARIMA) or exponential smoothing (ES) methods, is their capacity to incorporate exogenous information into models. GLMs and GAMs, however, have a number of potential limitations in this context including multicollinearity, issues due to serial correlation, temporal confounding, and incorrect error assumptions. The objective of this paper is to review and discuss approaches used in the literature to address these analytical limitations. This review only compares regression methods that

incorporate external information; therefore, Serfling [1963], ES, and ARIMA modeling methods are not included.

2.1 Mathematical Models

The two modeling methods reviewed in this paper, GLMs and GAMs, are described below.

2.1.1 Generalized Linear Models

GLMs [Neter et al., 1996], commonly used in infectious disease modeling, are nonlinear regression models following the form:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

where Y_i $(i = 1, \dots, n)$ are n independent responses that follow a probability density function belonging to the exponential class of distributions. The term $f(\mathbf{X}_i, \boldsymbol{\gamma})$ is a nonlinear response function where $\mathbf{X}_i$ $(i = 1, \dots, n)$ is the covariate vector, and $\boldsymbol{\gamma}$ is a vector of parameters. The error terms ε_i $(i = 1, \dots, n)$ are generally assumed to be uncorrelated with expectation $\mathrm{E}(\varepsilon_i) = 0$ and constant variance $\mathrm{V}(\varepsilon_i) = \sigma^2$. GLMs, more specifically, have nonlinear response functions that can be linearized by a transformation. In this case,

$$g(\mu_i) = g(\mathrm{E}(Y_i)) = g\big(f(\mathbf{X}_i, \boldsymbol{\gamma})\big) = (\mathbf{X}_i^*)'\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

where $\mu_i = E(Y_i)$, $\mathbf{X}_i^* = [\mathbf{1}_i \quad \mathbf{X}_i]$, and the conventional parameter notation is used, $\boldsymbol{\beta}$ where $\boldsymbol{\beta} = \boldsymbol{\gamma}$. The quantity $(\mathbf{X}_i^*)'\boldsymbol{\beta}$ is the linear predictor based on the independent variables $X_{1,i}, \dots, X_{p,i}$, and g is the link function relating the linear predictor to the mean response $f(\mathbf{X}_i, \boldsymbol{\gamma})$.

2.1.2 Generalized Additive Models

GAMs [Hastie et al., 2001; Wood, 2006; Hastie and Tibshirani, 1990] are a class of regression models that drop the assumption of linearity, thus making them more-flexible

modeling tools compared to GLMs. This additional flexibility allows for better modeling of complex nonlinearity. Additivity across effects is still assumed allowing for interpretation of results similar to that of GLMs. GAMs are defined as:

$$g(\mu_i) = a + f_1(X_{1,i}) + f_2(X_{2,i}) + \cdots + f_p(X_{p,i}),$$

where $\mu_i = E(Y_i)$ is a nonlinear response function, Y_i $(i = 1, \ldots, n)$ are n independent responses that follow a probability density function belonging to the exponential class of distributions, a is a constant, and the nonparametric f_j functions are generally estimated using scatterplot smoothers, such as smoothing splines (see Section 3.3.2). As with GLMs, the link function g relates the response function to an additive function of the predictors.

Semiparametric models incorporate both the linearity of GLMs and flexibility of GAMs, and are of the form:

$$g(\mu_i) = (\mathbf{X_i^*})'\boldsymbol{\beta} + f_1(Z_{1i}) + f_2(Z_{2i}) + \ldots + f_q(Z_{qi})$$

where $\mathbf{X}^*$ is a vector of predictors to be modeled linearly, and the effects of predictors $\mathbf{Z}$ are modeled nonparametrically.

2.2 Modeling Influenza-associated Mortality

This section describes several limitations of GLMs and GAMs in modeling influenza-associated mortality, and reviews how these limitations are accounted for in the literature.

2.2.1 Modeling Correlated Data

Time series mortality data tend to exhibit both long-term trends and seasonal fluctuations. A critical assumption of both GLMs and GAMs, however, is the independence of response variables $Y_i,\ i = 1, \ldots, n$. When error terms in regression models are positively correlated, standard errors may significantly underestimate the true standard deviation [Zeger et al., 2006].

One approach for removing autocorrelation is by adding predictor variables to the model that have time-ordered effects on the response. In modeling influenza-related mortality with GLMs, Fourier terms are often included to account for seasonality, and linear or quadratic time terms included for long-term trends [Thompson et al., 2009; Yu et al., 2013; Charu et al., 2013; Nielsen et al., 2011; Newall et al., 2008]. Recently, cubic spline functions, also known as regression splines or parametric splines, have been used in GLMs to better account for seasonality, thus reducing residual correlation [Goldstein et al., 2012]. With the additional flexibility of smoothing splines, GAMs can model nonlinear associations more easily and with fewer parametric constraints than GLMs with regression splines. Muscatello et al. [2013] use a smoothing spline of time to account for seasonal and time-varying patterns in non-influenza associated mortality. Yang et al. [2009, 2011] chose the effective degrees of freedom (edf) of smoothing splines by assessing the partial autocorrelation function plots of the residuals. The edf were chosen when the residuals fell within +/- 0.1 variability around zero with no discernible pattern.

Using time splines to account for non-influenza related seasonality in regression models, however, can be challenging since influenza terms and seasonal terms tend to be correlated temporally. For instance, several recent studies have shown that meteorological factors, e.g., temperature and humidity, may affect influenza-related mortality estimation [Warren-Gash et al., 2011; Yang et al., 2011; Wong et al., 2012; Yang et al., 2012]. Other seasonal viruses with varying local-level circulation, such as respiratory syncytial virus (RSV), may also affect estimates [Thompson et al., 2003; Mangtani et al., 2006]. Additional city-specific factors that may be influential include population density, age distributions, and differences in predominant influenza viruses [Lofgren et al., 2007]. When regression splines are used to model these confounding terms, the appropriate number of degrees of freedom (df) can be difficult to determine. With increasing df, splines may inadvertently dampen influenza-associated effects by overfitting data.

Even with careful adjustment of covariates, there may remain some amount of temporal residual correlation. Yet very few studies mention an assessment of residuals. Yang et al. [2009] assessed autocorrelation plots of residuals for serial patterns. Nielson et al. [2011] noted checking correlograms of the residuals and finding no autocorrelation. Muscatello et al. [2013] checked the normality of residuals using quantile-quantile plots. If residual correlation remains, certain remedial measures may be applicable. Goldstein et al. [2012] found residual correlation after fitting linear models. They first determined the underlying stochastic structure of the residuals which was found to be first-order autoregressive, AR(1). A bootstrapped estimate of the AR(1) autoregressive parameter was then used to obtain a sample of regression parameters from which confidence intervals (CIs) were obtained.

Nonparametric bootstrapping is another popular approach for estimating precision given complex or unknown error structure (Hastie et al., 2001). Yang et al. [2011] derived 95% CI estimates by bootstrapping the scaled Pearson residuals, $r_{P_s,i} = \frac{r_{P_i}}{\sqrt{\hat{\phi}(1-h_{ii})}}$ where $r_{P_i}, i = 1, \dots, n$ is the Pearson residual (raw residual divided by the square root of the variance), $\hat{\phi}$ is an estimate of the dispersion parameter, and h_{ii} denotes the i^{th} diagonal of the hat matrix. Wu et al. [2012] also used a bootstrapping approach to calculate 95% CIs, though details of the bootstrapping method were not given.

Some studies have implemented regression models with autoregressive error structures when modeling mortality time series. Theoretically, such models should be able to simultaneously account for both external information and the underlying autocorrelative structure within the series. Warren-Gash et al. [2011] examined partial autocorrelation functions for correlation within the residuals. Autocorrelation was found at a lag of one week; thus, the authors fit models with a term for residuals lagged by 1 week. Yang et al. [2011] added autoregressive terms to remove significant autocorrelation of residuals in the first four weeks.

Estimation from regression models containing a lagged dependent variable, however, presents challenges because y_{t-1} is not strictly exogenous [Cryer, 1986]. Lagging the model one period means that $y_{t-1} = \delta + \phi_1 y_{t-2} + \theta_0 x_t + u_{t-1}$, so y_{t-1} is clearly correlated with u_{t-1}. Further, if the error structure u_{t-1} is not white noise, then it may be the case that $cov(u_t, u_{t-1}) \neq 0$ and y_{t-1} and u_t are also correlated. In this case, the lagged dependent variable is not even weakly exogenous and the ordinary least squares estimator becomes biased and inconsistent.

For large sample sizes a transformed response variable may help eliminate the autocorrelation. If a first-order autoregressive model can be assumed, the transformation $Y_i' = Y_t - \rho Y_{t-1}$ may be applicable [Goldstein et al., 2012]. Several methods may be utilized to estimate ρ including the Cochrane-Orcutt, Hildreth-Lu, and first-differencing procedures [Wooldridge, 2003].

2.2.2 Defining Error Structure

Mortality data are discrete and non-negative, and like many other types of count data, tend to be right-skewed and heteroskedastic. Extremely high weekly (or daily) death counts are uncommon occurrences, and variability tends to increase as the mean number of deaths increases (e.g., during winter months). Because of these characteristics, mortality counts are often assumed to follow a Poisson distribution [Yang et al., 2009, 2011a, 2011b, 2012; Thompson et al., 2003; Lemaitre et al., 2012; Wong et al., 2004, 2012; Newall et al., 2010]. The Poisson model, however, requires equality of mean and variance. Since overdispersion with respect to the Poisson is often observed, the negative binomial distribution is instead assumed [Yu et al., 2013; Feng et al., 2012]. Though the mean and variance of the negative binomial distribution are not completely independent, the distribution's two parameters offer greater modeling flexibility than the Poisson. In other studies, a quasi-likelihood method (e.g., quasi-Poisson) is used to model mortality

[Charu et al., 2013; Warren-Gash et al., 2011]. This approach does not require an explicit specification of the underlying distribution (or log-likelihood function). Instead, parameter estimates are based entirely on the sample mean and variance of the model observations. For overdispersed Poisson data, the quasi-likelihood method scales the variance using a constant multiplier.

Parameter estimates from the negative binomial and quasi-Poisson models are often very close [Ver Hoef & Boveng, 2007]. Differences may arise with smaller sample sizes. The variance of a quasi-Poisson model is a linear function of the mean, while that of a negative binomial model is a quadratic function of the mean. These variance relationships affect the weights in the iteratively weighted least-squares algorithm which is used to find the maximum likelihood estimates in GLMs. Because the variance is a function of the mean, large and small counts are weighted differently. The negative binomial gives smaller sample means more weight relative to the quasi-Poisson; thus, smaller sample means have a greater effect on maximum likelihood parameter estimates in negative binomial regression.

In some instances, a normal distribution is assumed for modeling mortality counts [Nielsen et al., 2011; Newall et al., 2008; Goldstein et al., 2012; Muscatello et al., 2013; Wu et al., 2012]. While both the Poisson and negative binomial distributions converge asymptotically to a normal distribution, the mean value at which the normality approximation holds will vary depending on the spread of the data. With smaller counts of data, the assumption may not be appropriate. Linear regression also assumes homoskedastic errors and continuous data, neither of which describe the mortality data modeled in this context. Diagnostics should be performed, particularly for heteroskedasticity and serial correlation.

2.2.3 Adjustment for Unmeasured Seasonal Covariates

In modeling influenza-associated mortality, careful adjustment of seasonal covariates exhibiting temporal patterns similar to influenza is necessary to avoid misrepresentation of the influenza effect. Underrepresenting background seasonality tends to inflate influenza parameter estimates, while overrepresenting background seasonality leads to dampened or erratic influenza parameter estimates. Seasonal covariates entered into regression models may be measured or unmeasured. Measured seasonal covariates might include temperature, humidity, and proxies for other circulating viruses (e.g., respiratory syncytial virus, or RSV). Unmeasured seasonal covariates are generally represented via mathematical functions. Three approaches often used to account for unmeasured seasonality over time include: (1) Fourier terms; (2) parametric cubic splines; and (3) nonparametric smoothing splines. A brief review of parametric and nonparametric splines is given next.

2.2.3.1 Parametric Cubic Splines

Splines are piecewise polynomials used to model complex curvature [Hastie et al., 2001]. They serve as an alternative to sinusoidal terms or global polynomial terms, both of which can lead to a more-constrained fit of the data. For a particular covariate, X, the range of X is partitioned into k+1 intervals by k points, $\{x_1, \dots, x_k\}$, referred to as knots. A separate cubic polynomial is fitted to each interval of data.

Cubic splines are defined as follows: Given a set of knots, $x_1 < x_2 < \cdots < x_k$, contained within interval (*a, b*), a cubic spline is a continuous function f such that (*i*) f is a cubic polynomial over $(x_1, x_2), (x_2, x_3), \dots, (x_{k-1}, x_k)$ and (*ii*) f has continuous first and second derivatives at all knots. In general, an Mth-order spline is a piecewise $M-1$ degree polynomial with $M-2$ continuous derivatives at the knots. Once the number

and location of knots is decided, splines are entered directly into the covariate matrix via their basis functions.

Cubic splines ($M = 4$) are the lowest-order splines for which the discontinuity at the knots is not noticeable. Continuous first and second derivatives ensure smoothness across intervals. As such, cubic splines are one of the most-commonly used splines in practice. More-flexible curves are obtained by increasing the degree of the spline and/or by adding more knots. But there is a tradeoff. Too few knots or lower order may result in a function that is too restrictive (low bias, high variance). Too many knots or higher-order polynomials may overfit the data (high bias, low variance).

2.2.3.2 Nonparametric Smoothing Splines

Smoothing splines control the complexity of fit by regularization [Hastie et al., 2001; Wood, 2006]. Among all twice differentiable functions f, a smoothing spline is one that minimizes the penalized residual sum of squares:

$$PRSS(f, \lambda) = \sum_{i=1}^{N} \{y_i - f(x_i)\}^2 + \lambda \int_{x_{\min}}^{x_{\max}} \{f''(x)\}^2 \, dx,$$

where λ is a fixed smoothing parameter. The first term is the residual sum of squares, while the second term is a roughness penalty that penalizes curvature. Note that when $f(x)$ is a linear function, the penalty term is 0. The parameter λ controls the trade-off between fitting the data and penalizing curvature. Thus, λ controls the amount of smoothness in the fitted function. When $\lambda = 0$, there is no smoothing and the solution is the interpolation function. When $\lambda \to \infty$, $PRSS(f, \lambda)$ is minimized only when $\int \{f''(x)\}^2 dx = 0$. In this case, f converges to the linear regression estimator.

It has been shown that the function $\hat{f}(x)$ that minimizes $PRSS(f, \lambda)$ is a natural cubic spline with knots at all distinct observed values of x. While it might seem as if the function $\hat{f}$ may be overparameterized given that the natural cubic spline introduces up to

$x_{\max}$ df (since there are up to $x_{\max}$ knots), the smoothing parameter imposes constraints on $\hat{f}$ that translate to a penalty on the spline parameters –- shrinking them toward a linear fit (i.e., effectively decreasing the df).

2.2.3.3 Accounting for Background Seasonality

Although simple and easy to implement, generalized linear models tend to model poorly effects characterized by complex nonlinearity. Global polynomials may account for curvature well in one area of the range, but often at the expense of another area [Hastie et al., 2001]. Fourier terms, while able to capture regular seasonality, assume a constant baseline across seasons. Regression splines address these issues by modeling lower-order polynomials piecewise over the full range of the variable. Goldstein et al. [2012] use parametric, periodic cubic splines in GLMs to account for unmeasured seasonality in mortality time series. Depending on the number of knots, parametric splines can significantly increase dimensionality of the covariate matrix. Collinearity is also a potential problem if certain basis sets are used (e.g., a truncated power basis).

A number of studies use smoothing splines in GAMs to account for unmeasured seasonality [Muscatello et al., 2013; Yang et al., 2009; Yang, Chen, He et al., 2011]. Unlike regression splines, smoothing splines are nonparametric and thus do not add terms to the covariate matrix. This helps reduce the amount of correlation across temporally similar terms. An advantage of modeling background seasonality via time splines is that complex curvature can be easily controlled by the number of df or edf specified for the spline. Because of this, however, under- or overfitting background seasonality poses a potential problem since either can lead to a biased influenza-associated mortality effect.

2.2.4 Multicollinearity

High temporal correlation among covariate terms is a major limitation in modeling mortality via regression. Temperature, humidity, influenza, and RSV are all seasonal

factors associated with mortality and, in the U.S., all tend to have temporally consistent peaks. Multicollinearity, i.e., redundancy in covariate information, has two negative effects: (1) the parameter estimates are unstable due to inflated standard errors resulting from near-singularity in the variance-covariance matrix, and (2) interpreting parameter estimates becomes difficult since the effects of highly correlated covariates cannot be disentangled. Thus, an assessment of collinearity among covariates should be conducted. Most studies did not note checking for multicollinearity, though some mention excluding certain covariates to avoid collinearity [Yang et al., 2011b].

A common approach used to diagnose collinearity is by assessing the covariate correlation matrix [Neter et al., 1996]. High correlations between predictor variables suggest collinearity. This method, however, can only detect pairwise correlation. There is also no standard cut-off criterion for indicating collinearity. A second method for detecting multicollinearity uses Variance Inflation Factors (VIFs) [Neter et al., 1996]. VIFs are obtained by regressing each predictor on all the other predictors, and then estimating an R-square value for each. A third approach examines all the predictor variables together using a principle components analysis [Jolliffe, 2002]. Once all orthogonal components are determined, conditional indices are computed as ratios of the variables between two components. Conditional indices greater than 30 suggest multicollinearity.

An advantage of GAM smoothing splines is that they are a type of regularization method designed to control multicollinearity. On the other hand, regression splines may introduce correlated basis functions to the covariate matrix depending on which basis sets are utilized.

2.3 Summary

GLMs and GAMs are popular regression methods used to model influenza-associated mortality. Unlike traditional time series methods such as ARIMA or exponential

smoothing, both have the advantage of incorporating exogenous variables into the model. Because both are regression methods, they also have a number of shared limitations. GLMs and GAMs have difficulty addressing serial correlation and multicollinearity. Both approaches can be made more flexible via the use of spline functions. The increased flexibility in modeling complex curvature helps remove residual autocorrelation while better adjusting for unmeasured seasonal covariates. GAMs have the added advantage of offering nonparametric adjustment of covariates which should help reduce multicollinearity.

CHAPTER 3

POOLING INFLUENZA-ASSOCIATED MORTALITY RISKS ACROSS LOCATIONS AND QUANTIFYING SPATIAL HETEROGENEITY

For over fifty years, the Centers for Disease Control and Prevention (CDC) has been estimating annual rates of U.S. deaths attributable to influenza [Serfling, 1963; Lui and Kendal, 1987; Simonsen et al., 2000; Thompson et al., 2003, 2009]. These estimates have been used to determine costs and benefits associated with influenza prevention and control strategies [Meltzer et al., 1999; Nichol, 2003; Bridges et al., 2003]. Quantifying the effect of influenza on mortality, however, can be challenging since influenza infections typically are not confirmed virologically nor specified on death certificates [Douglas, 1976; Bisno et al., 1971; Taubenberger and Morens, 2008; Collins & Lehmann, 1951]. The CDC has applied a wide range of mathematical modeling approaches to specify the association between influenza and mortality [Lui and Kendal, 1987; Thompson et al., 2003; Thompson et al., 2009; Cheng et al., 2014 unpublished]. U.S. influenza-associated death estimates have been based usually on mortality data first aggregated at the national level and then modeled. Unfortunately, there are a number of local-level seasonal factors that may confound the association between influenza and mortality – thus suggesting that data be modeled at the local level and then pooled to make national estimates of death.

Several recent studies have shown that meteorological factors, e.g., temperature and humidity, may affect influenza-related mortality estimation [Warren-Gash et al., 2011; Yang et al., 2011; Wong et al., 2012; Yang et al., 2012]. Other seasonal viruses with varying local-level circulation, such as respiratory syncytial virus (RSV), can also be influential [Thompson et al., 2003; Mangtani et al., 2006]. Seasonal factors for which daily or weekly measurements are not collected are potential regional/local confounders

as well (e.g., differences in predominant influenza viruses, seasonal host health, population density, and age distribution) [Lofgren et al., 2007].

The current study has two objectives: 1) to determine the extent to which the association between influenza activity and mortality varies by city when accounting for local-level seasonal covariates, and 2) to pool relative risk estimates across cities and compare this pooled estimate to one obtained using a more-traditional approach where data are first pooled nationally and then modeled.

The chapter is organized as follows: Section 3.1 describes the data used for all analyses. Section 3.2 explains the modeling approaches including the two-stage hierarchical Bayesian modeling method and the more-traditional GLM method. It also details the generalized additive modeling method used for Level 1 of the hierarchical approach. Section 3.3 gives city-level results from the hierarchical method for the association between influenza activity and mortality. It also compares results of the pooled relative risk estimates from each modeling method. Section 3.4 discusses the differences in these results, their implications in interpretation, and several limitations in the methodology.

3.1 Mortality and Influenza Surveillance Data

The data used in this study were downloaded from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) website http://www.ihapss.jhsph.edu/data/NMMAPS/R/. These data were chosen for our analyses because they contain temporal information on mortality counts and covariates of interest (e.g., temperature, pollution, etc.) from 105 U.S. cities. Details on data collection and processing methods can be found at the Internet-based Health and Air Pollution Surveillance System (iHAPSS) website http://www.ihapss.jhsph.edu. Data modeled in this study spanned a 10-year period from January 1, 1991 to December 31, 2000 because of widespread circulation of influenza A(H3N2), the influenza subtype associated with

the highest mortality rates. Daily counts of mortality, and daily measures of mean temperature and mean dew point temperature were obtained for each of 88 U.S. cities with the largest populations of persons aged 65 or older as of the year 2000. Mortality data include daily counts of respiratory and circulatory deaths among persons aged 65 or older.

Weekly numbers of total respiratory specimens tested for influenza and positive-influenza isolates by virus type and subtype {A(H1N1), A(H3N2), and B} were obtained from surveillance data collected from 50 to 75 World Health Organization (WHO) collaborating laboratories in the U.S. Weekly numbers of total respiratory specimens tested for respiratory syncytial virus (RSV) and positive-RSV isolates were obtained from the National Respiratory and Enteric Virus Surveillance System. We used regional (North East, Midwest, South, West) proportions of respiratory specimens testing positive for influenza or RSV, referred to as 'percent-positive' viral activity, as proxies for true viral activity. Local or state data were too sparse to use directly in models. Daily percent-positive data for each influenza subtype and RSV were imputed linearly from weekly percent-positive data.

U.S. population estimates by year, age group (65+ years), and city were obtained from the U.S. Census Bureau. Daily city-level population estimates were imputed via a step function (i.e., the annual estimate was used for each day of that calendar year).

3.2 Two-stage Hierarchical Bayesian Model

We used a two-stage hierarchical, Bayesian modeling approach to compute an overall national relative risk estimate of the association between influenza activity and mortality rates. At the first stage, 88 U.S. cities with varying trends in mortality and weather were modeled using semi-parametric regression. At the second stage, the log-relative risk estimates calculated for each city in stage 1 represented the "outcome" variable, and were modeled two ways: (1) assuming spatial independence across cities using a Bayesian

generalized linear model, and (2) assuming correlation among cities using a Bayesian spatial correlation model.

3.2.1 Stage 1: Local-level Semi-parametric Regression

Eighty-eight U.S. cities with varying trends in mortality and weather were modeled using generalized additive models (GAMs), a semi-parametric regression method described in detail in Section 3.2.2. Each city was modeled separately. The outcome variable for all models was daily mortality due to underlying respiratory or circulatory (R&C) causes among persons aged 65 or older. Annual city-level population estimates of persons aged ≥65 years were used as the offset term to account for changes in population size over time. The percent-positive influenza subtypes (i.e., A(H1N1), A(H3N2), and B) were entered into all models as linear terms. A long-term trend was accounted for in each of the models by a natural cubic spline. A day-of-week effect was also accounted for using six indicator variables (Sunday was the referent group).

Local-level seasonal covariates included in this study were daily mean temperature, daily mean dew point temperature, and daily percent-positive RSV activity. Percent-positive RSV was entered into all models linearly. Daily ambient temperature and dew point temperature (a proxy for humidity) were entered into models as either current or lagged (backshifted a specified number of days) terms. These two covariates were entered into the models via smoothing splines. Estimates for smoothing parameters were calculated using the generalized cross validation (gcv) metric in the 'mgcv' package for R software. The Akaike information criterion was used for model comparisons and selection. Ambient temperature was assumed to be a proxy for other unknown seasonal covariates.

3.2.2 Generalized Additive Models

GAMs are a class of regression models which drop the assumption of linearity, thus making them more-flexible modeling tools compared to GLMs [Hastie et al., 2001; Hastie and Tibshirani, 1990; Wood, 2006]. This additional flexibility allows for better modeling of complex nonlinearity. Additivity across effects is still assumed allowing for interpretation of results similar to that of GLMs. GAMs are defined as:

$$g(\mu_i) = a + f_1(X_{1,i}) + f_2(X_{2,i}) + \cdots + f_p(X_{p,i}) \quad (1)$$

where $\mu_i = E(Y_i)$ is a nonlinear response function, Y_i $(i = 1, \ldots, n)$ are n independent responses that follow a probability density function belonging to the exponential class of distributions, a is a constant, the link function g relates the response function to an additive function of the predictors, and the nonparametric f_j functions are estimated using smoothing splines which are a type of scatterplot smoother that control the complexity of fit by regularization. Among all twice-differentiable functions f, a smoothing spline is one that minimizes the penalized residual sum of squares:

$$PRSS(f, \lambda) = \sum_{i=1}^{N}\{y_i - f(x_i)\}^2 + \lambda \int_{x_{\min}}^{x_{\max}}\{f''(x)\}^2\, dx \quad (2)$$

where λ is a fixed smoothing parameter. The first term is the residual sum of squares, while the second term is a roughness penalty that penalizes curvature. Note that when $f(x)$ is a linear function, the penalty term is 0. The parameter λ controls the trade-off between fitting the data and penalizing curvature. Thus, λ controls the amount of smoothness in the fitted function. When $\lambda = 0$, there is no smoothing and the solution is the interpolation function. When $\lambda \to \infty$, $PRSS(f, \lambda)$ is minimized only when $\int\{f''(x)\}^2 dx = 0$. In this case, f converges to the linear regression estimator. It has been shown that the function $\hat{f}(x)$ that minimizes $PRSS(f, \lambda)$ is a natural cubic spline with knots at all distinct observed values of x [Hastie et al., 2001]. The smoothing

parameter imposes constraints on $\hat{f}$ that translate to a penalty on the spline parameters –- shrinking them toward a linear fit (i.e., effectively decreasing the df). The smoothing parameter λ is often selected by minimizing prediction error estimates from a cross-validation method. In this study, the generalized cross-validation method defined in the 'mgcv' R package [Wood, 2006] was used to automate selection of the parameter.

Additive models are fit using a backfitting algorithm. The following criterion is iteratively solved [Hastie et al., 2001]:

$$PRSS(a, f_1, f_2, \dots, f_p) = \sum_{i=1}^{N}\{y - a - \sum_{j=1}^{p} f_j(x_{ij})\}^2 + \sum_{j=1}^{p} \lambda_j \int\{f_j''(x_j)\}^2 dx_j \quad (3)$$

where the $\lambda_j \geq 0$ are smoothing parameters. First, $\hat{a} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is fixed. Then, a cubic smoothing spline S_k is applied to the terms $\{y_i - \hat{a} - \sum_{k\neq j} \hat{f}_j(x_{ij})\}_{i=1}^{N}$ as a function of x_{ik}, thereby giving a new estimate for $\hat{f}_k$. This procedure is repeated for each predictor X_j in turn, using the most-current estimates of all other functions $\hat{f}_j$, until all functions $\hat{f}_j$ have stabilized. The iterative algorithm is summarized as follows [Hastie et al., 2001]:

(1) Initialize: $\hat{a} = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad \hat{f}_j \equiv 0, \quad \forall i, j$

(2) Cycle: $j = 1,2, \dots, p, 1,2, \dots, p, \dots$

$$\hat{f}_j \leftarrow S_j\left[\{y_i - \hat{a} - \sum_{k\neq j} \hat{f}_k(x_{ik})\}_{i=1}^{N}\right],$$

$$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N}\sum_{i=0}^{N} \hat{f}_j(x_{ij})$$

until the functions $\hat{f}_j$ change less than a prespecified threshold. For GAMs, the appropriate criterion is a penalized log-likelihood which is maximized using a backfitting procedure with a likelihood maximizer.

We implemented the following semi-parametric GAM model:

$$\log(Y_i) = \log(a) + \beta_0 + \sum_{j=1}^{6} \beta_j[I_{ij}] + \beta_7[\mathrm{A(H1N1)}_i] + \beta_8[\mathrm{A(H3N2)}_i] + \beta_9[\mathrm{B}_i] + \beta_{10}[\mathrm{RSV}_i] + ns(t_i) + s(\mathrm{Temp}_{li}), \quad (4)$$

where $Y_i \sim$ Negative Binomial (NB) represents the number of deaths on day i; a is the offset term equal to the annual population size; β_0 represents the model intercept; β_1 through β_6 are coefficients for day-of-week indicators, I_{ij} (Sunday = referent); β_7 through β_{10} are coefficients for percent-positive viral terms; $ns(t_i)$ is a parametric, natural cubic spline representing a long-term time trend; and $s(\text{Temp}_{li})$ is a nonparametric smoothing spline representing a temperature effect lagged l days, $l \in \{2,3,\ldots,6\}$.

3.2.3 Stage 2: Pooling Log-Relative Risks

In the first stage, a semi-parametric, negative binomial GAM was used to model each of the 88 cities. The influenza parameter estimates in equation (4), ($\hat{\beta}_7, \hat{\beta}_8$, and $\hat{\beta}_9$), are the log-relative risks associated with percent-positive influenza activity and deaths among persons 65+ years. The coefficients for all adjustment variables, including the splines in the semi-parametric model, are considered nuisance parameters.

In the second stage, the log-relative risks estimated for each influenza type/subtype are pooled across all cities, and become the outcome variables modeled in two ways: (1) assuming spatial independence across city estimates, and (2) assuming spatial correlation among city estimates.

3.2.3.1 Spatial Independence Model

In the spatial independence model, linear regression is used to model the city-level, log-relative risk estimates relating to each influenza virus type/subtype. Explanatory variables compiled at the city level are included to characterize geographic location, and to explain some of the geographical heterogeneity of the log-relative risk parameter estimates. More formally,

$$\beta^c|\alpha_0, \alpha_1, \ldots, \alpha_p, \tau^2 \sim N(\alpha_0 + \textstyle\sum_{j=1}^{p} \alpha_j W_j^c, \tau^2), \qquad (5)$$

where β^c is the log-relative risk of death among persons aged 65+ years associated with the percent-positive influenza type/subtype activity (either A(H1N1), A(H3N2), or B) in city c, the W_j^c are the p city-level explanatory variables, and τ^2 represents the 'between-cities' variance.

A hierarchical Bayesian modeling approach can be used to pool these estimates across cities [Dominici, 2000]. With this approach, the joint posterior distribution of all parameters included in the model (influenza and nuisance parameters) is simulated, and then integrated over all nuisance parameters to obtain the marginal posterior distributions of interest (i.e., the marginal posteriors for log-relative rates by type/subtype). In this context − with a high-dimensional parameter space, large city-level sample sizes, and a large number of cities included in the first stage of analysis − the computational expense for a full Bayesian approach becomes time-consuming and impractical.

Le Cam and Yang [1990] showed that with large sample sizes, a good approximation of the posterior distribution can be obtained by using a normal approximation of the likelihood. Daniels and Kass [1998] have shown that a hierarchical modeling approach based on the normal approximation of the likelihood leads to a two-stage model that well approximates the exact Bayesian model with more-efficient simulation from the posterior. If the likelihood function of the influenza parameters and the nuisance parameters can be approximated by a multivariate normal distribution with mean equal to the maximum likelihood estimates $\widehat{\boldsymbol{\beta}}^{\boldsymbol{c}}$ and variance-covariance matrix $\widehat{\mathbf{V}}$, then by definition, the marginal likelihood of any component of $\widehat{\boldsymbol{\beta}}^{\boldsymbol{c}}$ has a normal distribution with mean $\hat{\beta}^c$ and variance $\hat{v}^2$. With this assumption, the first stage of the model was replaced with the MLE-based normal approximation to the likelihood function with means and variances equal to the maximum likelihood estimates obtained by fitting the GAM models described above. Thus, the first-stage computation was simplified by assuming

$$\hat{\beta}^c \mid \beta^c \sim N(\beta^c, (\hat{v}^2)^c) \quad (6)$$

where $\hat{\beta}^c$ and $(\hat{v}^2)^c$ are the MLEs of β^c and its variance obtained by fitting the GAMs as described in stage 1.

City-level predictor variables included at the second stage were geographic region, city latitude, city longitude, percentage of residents living in poverty, percentage of residents age 65 or older, and percentage of homeowners age 65 or older. When these predictor variables are centered at their means, a simple pooled estimate of the effect of influenza activity on mortality can be obtained by setting all covariates to 0. In this scenario, the intercept α_0 can be interpreted as the average log-relative risk of the association between percent-positive influenza activity and mortality outcome given mean-centered, adjustment of location-specific predictors. In other words, the intercept parameter α_0 represents the pooled, overall average of the true log-relative risks. The independent regression parameters α_j measure the change in true log-relative risk of mortality associated with a unit change in the corresponding location-specific variable W_j^c.

Sources of variation in the estimation of α_0 are specified in the two stages of the hierarchical model. The variation of $\hat{\beta}^c$ around β^c describes the within-location variance $(v^2)^c$, while the variation of β^c around α_0 describes the between-location variance τ^2. The within-location variance depends on the predictive power of the Stage 1 regression models, while the between-location variance measures the heterogeneity across geographical locations that is unexplained by the covariates W_j^c in Stage 2.

The Bayesian hierarchical model was completed by specifying vague prior distributions for the parameters of the stage 2 model. *A priori*, the joint prior distribution was assumed to be the product of the marginals for $\boldsymbol{\alpha}$ and τ^2. The following prior distributions were assumed: for the intercept $\alpha_0 \sim N(0,100)$; for all explanatory variable parameters $\boldsymbol{\alpha_v} \sim N(0,100)$; and for the variance $\tau^{-2} \sim \Gamma(0.001,0.001)$. Sensitivity

analyses were conducted by varying the prior distributions. To compute the marginal posterior distributions of all second-stage parameters, a computational algorithm developed by Everson and Morris [2000] utilizing Markov chain Monte Carlo methods was implemented.

3.2.3.2 Spatial Correlation Model

In the spatial independence hierarchical model described above, we assume that, for any two locations (c, c') regardless of geographical distance, the log-relative risks β^c and $\beta^{c'}$ are independent. Though geographical region is accounted for and is a significant covariate in the spatial independence model, cities classified in the same geographic region, but far apart in terms of geographical distance, are considered "more similar" than two closer locations belonging to separate geographical categories. To overcome this limitation, the assumption of independence can be relaxed to allow for possible spatial correlation among the β^c. In the spatial correlation model, it is assumed that each city-specific relative risk is shrunk towards the average relative risk in neighboring cities which are defined as such based on geographical distance [Peng and Dominici, 2008].

The degree of similarity of the log-relative rates in locations c and c' can be defined as a function of the Euclidean distance, notated $d(c, c')$, between the cities. Euclidean distance was defined in terms of longitude and latitude coordinates. More specifically, it is assumed that $corr(\beta^c, \beta^{c'}) = exp\{-\phi d(c, c')\}$. In words, the correlation between β^c and $\beta^{c'}$ decays as the distance between the two cities c and c' increases. The parameter ϕ, represents the rate of decay to zero.

The spatial correlation model is specified by assigning prior distributions to the Stage 2 parameters. For $\boldsymbol{a}$ and τ^2 we choose the same priors specified for the hierarchical model assuming independence across cities. For the parameter ϕ, a uniform distribution in the range $[\phi_{\min}, \phi_{\max}]$ was chosen [Peng and Dominici, 2008]. The values $\phi_{\min}$ and $\phi_{\max}$ were selected so that, when $\phi = \phi_{\min}$, the correlation between the two

relative risks was within the range 0.01 (maximum distance apart) to 0.8 (minimum distance apart). When $\phi = \phi_{\max}$, the correlation between the two relative risks ranged from 0 (maximum distance) to 0.5 (minimum distance).

The correlation function specified above was used to define the spatial structure, and the model was fit using the 'spatialgibbs' function in the 'tsModel' package for R software [Peng and Dominici, 2008]. The hierarchical Bayesian spatial correlation model was fit using different values of ϕ to control for the strength of spatial correlation between cities. Weak, moderate, and strong spatial correlation can be represented by values of ϕ equal to 1, 0.1, and 0.01, respectively. When $\phi = 1$, the Stage 2 model is close to the spatial independence model, and when $\phi = 0.01$, even distant cities maintain some spatial correlation. For example, when using $\phi = 0.01$, the two cities furthest apart in this dataset have a correlation of approximately 0.5.

3.2.4 Conventional Modeling Approach for Comparison

We compared the hierarchical modeling approach to a widely used regression method for estimating influenza-associated mortality. First, mortality data was compiled by day across all 88 cities. Then, a negative binomial generalized linear model with sinusoidal terms representing unmeasured seasonal confounders was used to model influenza-associated morbidity and mortality [Thompson et al., 2003; Thompson et al., 2009; Warren-Gash et al., 2011; Liao et al., 2009; Newall et al., 2010]. The full model is described as follows:

$$\begin{aligned}\log(Y_i) = \log(a_i) + \beta_0 + \sum_{j=1}^{6} \beta_j [I_{ij}] + \beta_7 t_i + \beta_8 t_i^2 \\ + \beta_9[\sin(2\pi t_i/365.25)] + \beta_{10}[\cos(2\pi t_i/365.25)] \\ + \beta_{11}[\mathrm{A(H1N1)}_i] + \beta_{12}[\mathrm{A(H3N2)}_i] + \beta_{13}[\mathrm{B}_i] + \beta_{14}[\mathrm{RSV}_i]\end{aligned} \quad (7)$$

where $Y_i \sim$NegBin, represents the number of deaths on day i; a_i is the offset term equal to the annual population size; β_0 is the model intercept; β_1 through β_6 are coefficients for

day-of-week indicators, I_{ij} (Sunday = referent); β_7 and β_8 are coefficients for the long-term, non-linear time trend; β_9 and β_{10} are coefficients for the seasonal trend; and β_{11} through β_{14} are coefficients for the percent-positive viral terms.

3.3 Pooled Relative Risk Results

Figure 2.1 is a map of the 88 cities modeled in this study. The five cities with the largest year 2000 populations aged 65 or older were Los Angeles, New York, Chicago, Dallas/Fort Worth, and Houston. The five smallest cities were Jackson, Lexington, Huntsville, Corpus Christi, and Fort Wayne. Figure 3.2 shows daily time series of respiratory and circulatory deaths, mean temperature, mean dew point temperature, and regional RSV activity for four U.S. cities, one in each of four regions (Northeast: New York; South: Miami; Midwest: Chicago; West: Los Angeles). Because temperature and dew point temperature were found to be highly collinear in all 88 cities, the two terms were not modeled together. The temperature term tended to be more-highly associated with mortality; therefore, dew point temperature was dropped from all models. An adjusted mean dew point temperature was also modeled by first regressing mean dew point temperature on mean temperature, and then modeling the residuals. This term, however, did not substantially affect any of the virus parameter estimates and therefore was dropped from all models. Figure 3.3 shows daily time series of respiratory and circulatory deaths and regional circulation of the three influenza subtypes [A(H3N2), A(H1N1), and B] for the four U.S. cities. Note that A(H3N2) was the dominant influenza strain during this 10-year period. The A(H1N1) subtype shows very little activity during this period.

Figure 3.4 shows the daily mortality counts plotted with fitted values for the four cities. Correlograms of residuals from all 88 city-level models (not shown) revealed a mild level of autocorrelation ($\rho < 0.25$), up to 14 lags, for 3 cities: Los Angeles, New York, and Chicago. For the remaining cities, $\rho < 0.08$ at all lags. These results suggest

that the GAM method removed a significant amount of serial dependence from city-level series.

Figures 3.5−3.7 show the log-relative risk estimates for all 88 U.S. cities during the 10-year study period by influenza sub/type. A(H3N2) displayed wide variability across cities ranging from $\hat{\beta} = 0.3$ in Colorado Springs, CO to $\hat{\beta} = 1.3$ in Jacksonville, FL. A significant association between A(H3N2) activity and mortality was observed for all but one city (Colorado Springs, CO). A(H1N1) influenza activity showed significant positive association with mortality in only 16 of the 88 cities. Log-relative risk estimates for type B influenza were found to be the least variable across cities. While most cities trended towards a positive association between mortality and type B influenza activity, 48 of 88 estimates were not statistically significant.

Table 3.1 summarizes results of the pooled analyses from the three models: the independent-observations hierarchical model; the spatial-correlation hierarchical model; and the traditional GLM model. Means and confidence intervals were obtained from the marginal posterior distributions of the overall effect $(a_{0,v}),\ v \in \{\mathrm{A(H1N1)}, \mathrm{A(H3N2)}, \mathrm{B}\}$. Results from the independent-observations hierarchical model show that with a 10% increase in percent-positive A(H3N2) influenza activity (from baseline, or no measured activity), the risk of respiratory or circulatory deaths among persons 65+ years was increased by 8% (95% CI: 8−9). At 20% and 30% levels of percent-positive A(H3N2) activity, the risk of death among persons 65+ years increases to 17% (95% CI: 16−18) and 27% (95% CI: 26−29) above baseline, respectively. The spatial-correlation hierarchical model yielded results similar to the independent-observations hierarchical model. Given the traditional GLM model, a 10% increase in A(H3N2) influenza activity is associated with a 9% (95% CI: 9−10) increase in risk of death among persons aged 65+ years. At 20% and 30% increases in A(H3N2) activity, the risk of death for this age group increases to 19% (95% CI: 18−19) and 30% (95% CI: 29−31), respectively.

At 10% and 20% increases in A(H1N1) activity, both the independent-observations hierarchical model and the spatial-correlations hierarchical model yielded no significant increase in deaths among persons 65+ years, likely due to low A(H1N1) activity during the observed period. The traditional GLM model, however, showed a 4% (95% CI: 3−4) increase in deaths with a 10% increase in A(H1N1) activity, and a 7% (95% CI: 6−9) increase in deaths at 20% increase in A(H1N1) activity compared to baseline. Measured percent-positive activity of A(H1N1) during the observed study period did not reach 30%; therefore, a relative risk for this level of estimated circulation was not calculated.

All three methods calculated similar results for deaths associated with type B influenza circulation. At 10% circulation, deaths among 65+ was 4% (95% CI: 3−4), 5% (95% CI: 4−6), and 3% (95% CI: 3−4) higher than baseline based on the independent-observations hierarchical, spatial-correlation hierarchical, and traditional GLM models, respectively. At 20% circulation, deaths increased to 7% (95% CI: 6−8), 9% (95% CI: 8−11), and 6% (95% CI: 6−7) relative to baseline levels, respectively. Measured circulation of type B influenza during the observed study period did not reach 30% in all regions; therefore, a relative risk for this level of circulation was not calculated.

3.4 Conclusions

Results from this study show considerable variability at the local level with respect to the association between influenza activity and mortality for subtypes A(H1N1) and A(H3N2); the corresponding heterogeneity metrics are $I^2 = 0.62$ and $I^2 = 0.75$ respectively. These findings suggest that ambient temperature and other local-level seasonal factors significantly affect the relationship between influenza and deaths among persons aged 65 or older. To better account for these city-level, time-dependent confounders, hierarchical modeling methods should be considered.

Results from two hierarchical models (modeled at the city-level, then pooled) were compared to those from a traditional modeling approach (pooled at the national level, then modeled). Assuming a 30% increase in A(H3N2) influenza activity from baseline, the traditional GLM method estimated the risk of death to be approximately 3% higher than the other two methods, a statistically significant difference. This difference may be attributable to the greater influence of counts from large cities when modeled via the traditional method. Based on U.S. Census Bureau year 2000 population counts, the age 65+ years population in three cities (New York, Los Angeles, and Chicago) accounted for approximately 25% of the entire study population of persons age 65 and older. The hierarchical approaches essentially average all city-level parameter estimates taking into account estimated city-level variance, while the traditional method places more weight on outcomes of larger cities by pooling all data before modeling.

Results from the A(H1N1) influenza-associated mortality show the hierarchical and traditional approaches leading to very different conclusions (Table 3.1). With A(H1N1), the relative risk is null from both hierarchical models, which is not surprising given the very low activity of A(H1N1) during the study period. The traditional GLM model, however, showed a significant increase in deaths with 10% and 20% increases in A(H1N1) activity. Again, this may be attributed to the high log-RR values of a few large cities.

Applying the traditional method to obtain an overall relative risk of deaths associated with influenza may bias the estimate toward the city-level outcomes of a few large cities. Another possible explanation for the upward shift in relative risk via the traditional approach may be due to the 'pooling' of seasonal covariates across cities. By pooling death counts across cities, the assumption is that background seasonal factors are also combined at the national level. Non-influenza seasonal factors, however, vary greatly by city. For instance, generally the magnitude and timing of peak cold temperatures in southern cities are less pronounced and occur later in the year relative to northern cities.

When background seasonal covariates from all cities are pooled, the result is a national-level seasonal covariate with a tempered, less-pronounced signal that does not clearly represent seasonality from any particular region. Modeling a tempered seasonal background factor along with a strong seasonal influenza component may inadvertently attribute some percentage of non-influenza-associated deaths to influenza.

For type B influenza activity, the three modeling approaches yielded similar results, suggesting that viral activity did not vary much geographically and/or other seasonal factors were not temporally collinear with type B influenza. In this case, any of the three modeling approaches may be suitable for modeling.

There are several limitations to this study. First, hierarchical modeling is a novel approach to modeling influenza-associated deaths. It requires further in-depth investigation. For instance, the assumption of a multivariate normal approximation for the likelihood function of each city-level model needs further examination, and an appropriate estimate for the correlation parameter in the spatial-correlation hierarchical model should be determined. Second, residuals from the hierarchical models still show slight serial dependence in the larger cities. The city-level models are not able to capture the high winter-time mortality peaks of the largest cities, leaving this correlation in the residuals. The amount of correlation in the hierarchical models is, however, less than that observed using the traditional GLM modeling approach. Other modeling approaches at the city-level should be considered to better account for this residual correlation, e.g., incorporating additional confounders in GAMs, using ARIMAX models which include exogenous information, or even modeling traditional GLMs at local rather than national levels. Third, validation of results is difficult for any time series modeling approach. This is particularly difficult when modeling influenza-associated deaths with regression methods since there are a number of temporally collinear variates. Proper simulation studies which adequately represent mortality series with non-regression derived

components of influenza-associated and non-influenza-associated deaths are needed to properly compare regression modeling methods.

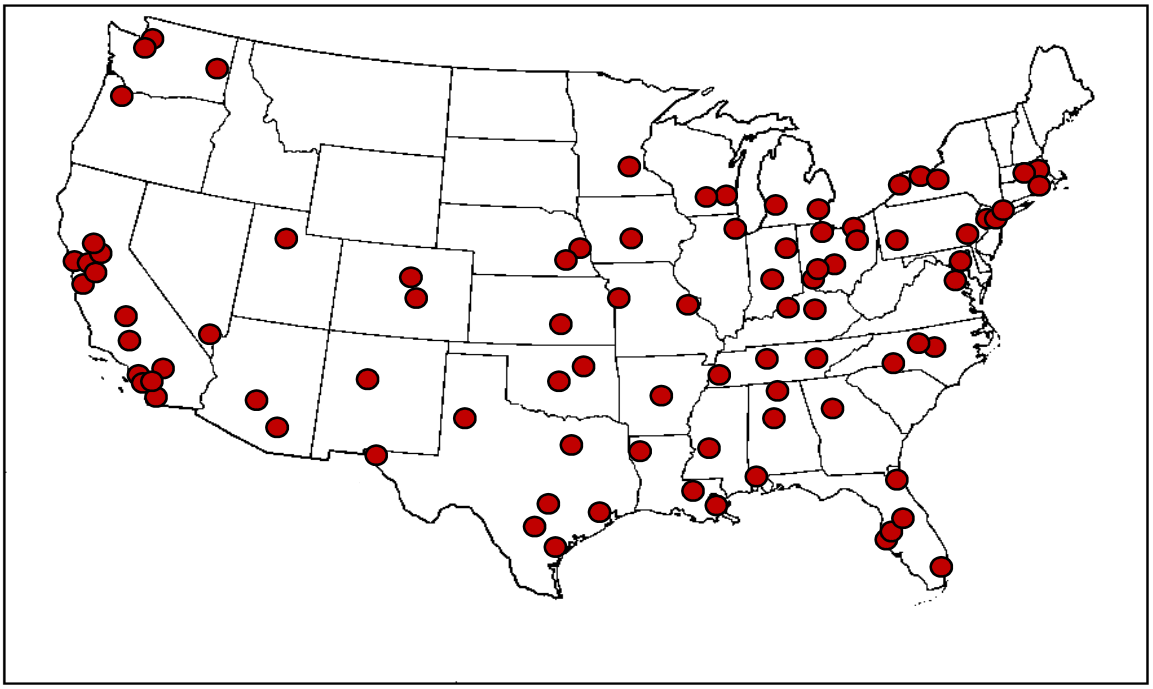

Figure 3.1: Map of 88 cities included in hierarchical analyses.

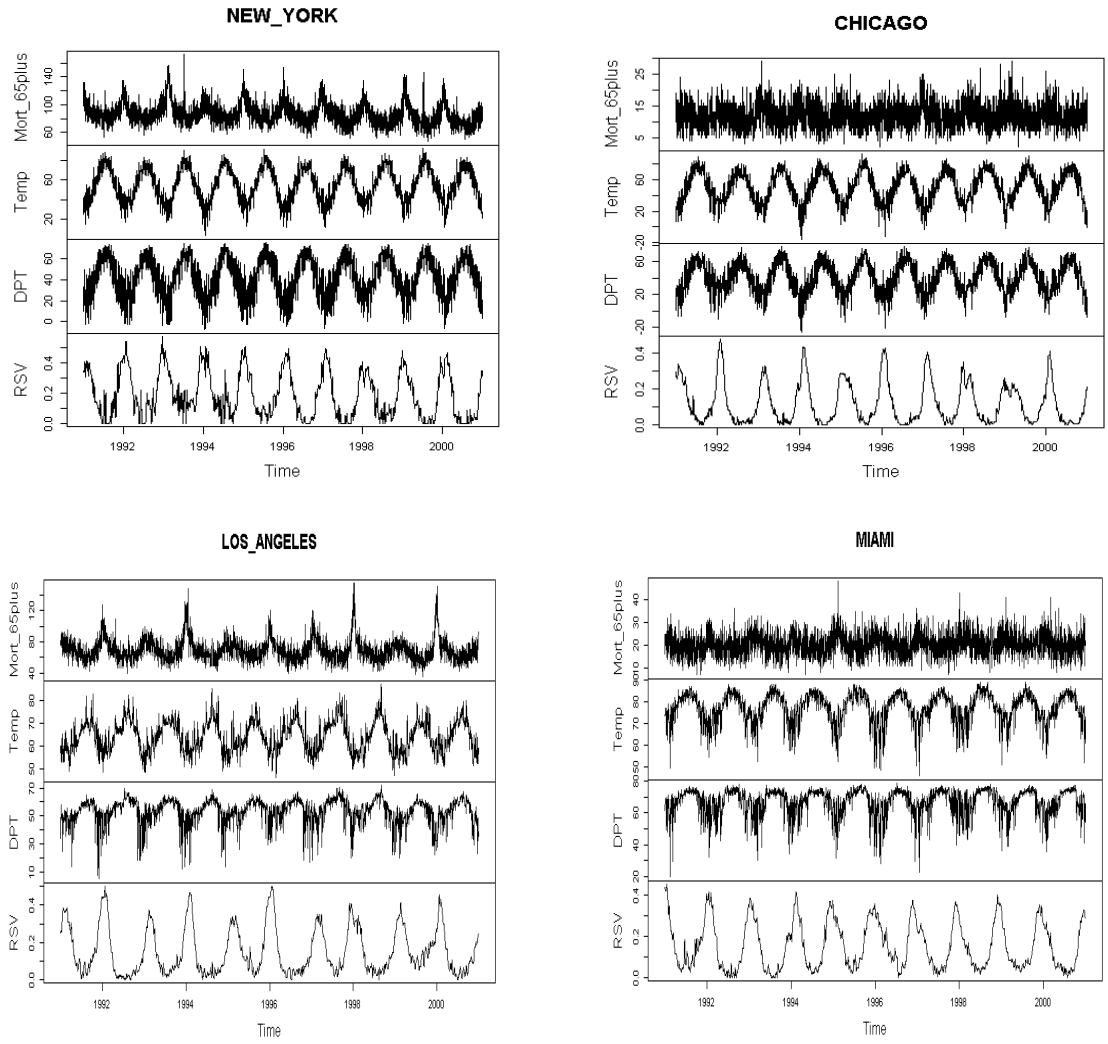

Figure 3.2: Daily mortality counts, mean temperature, mean dew point temperature, and regional RSV activity by city, 1991−2000.

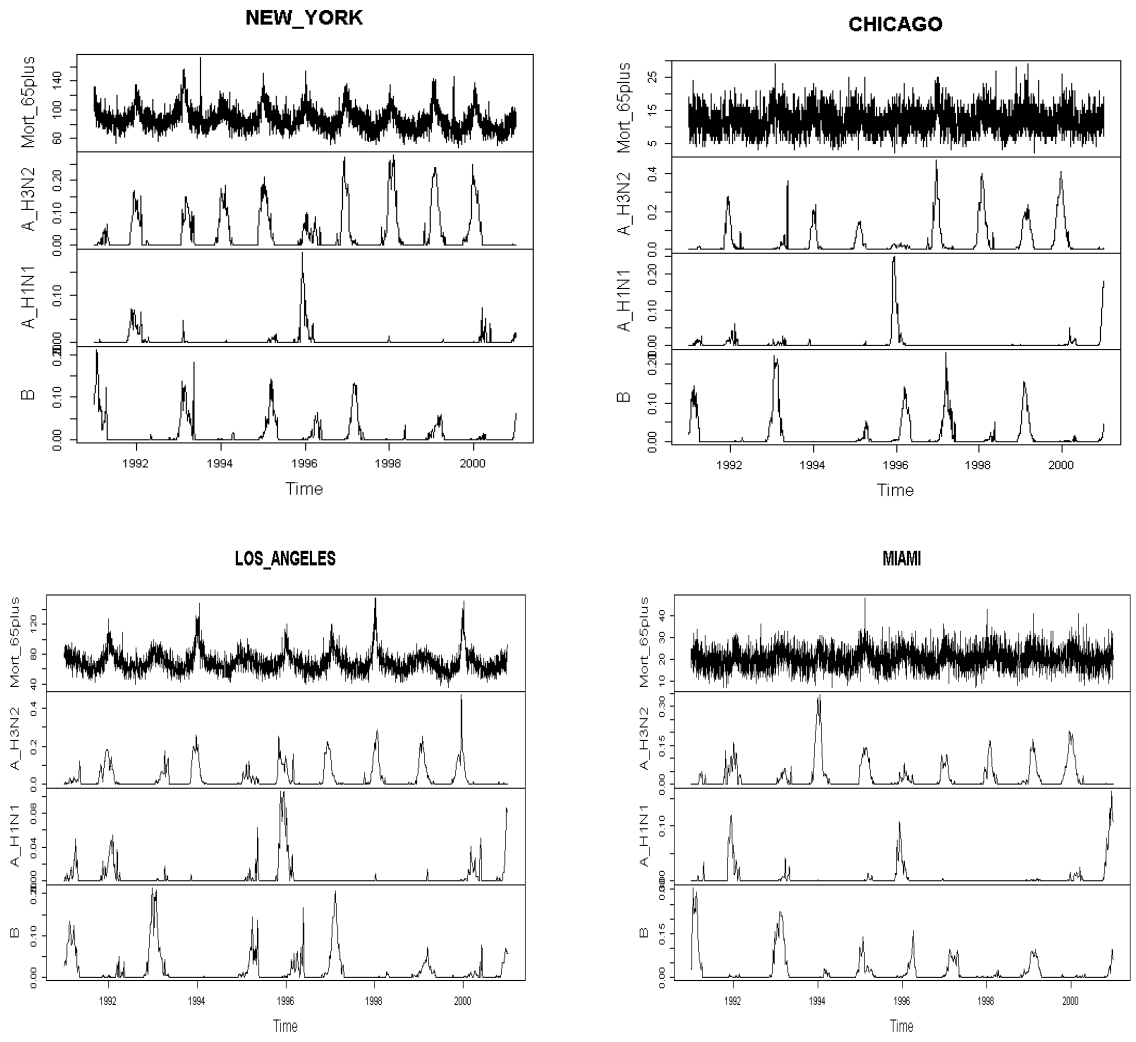

Figure 3.3: Daily mortality counts and regional influenza (AH3N2, AH1N1, B) activity by city, 1991−2000.

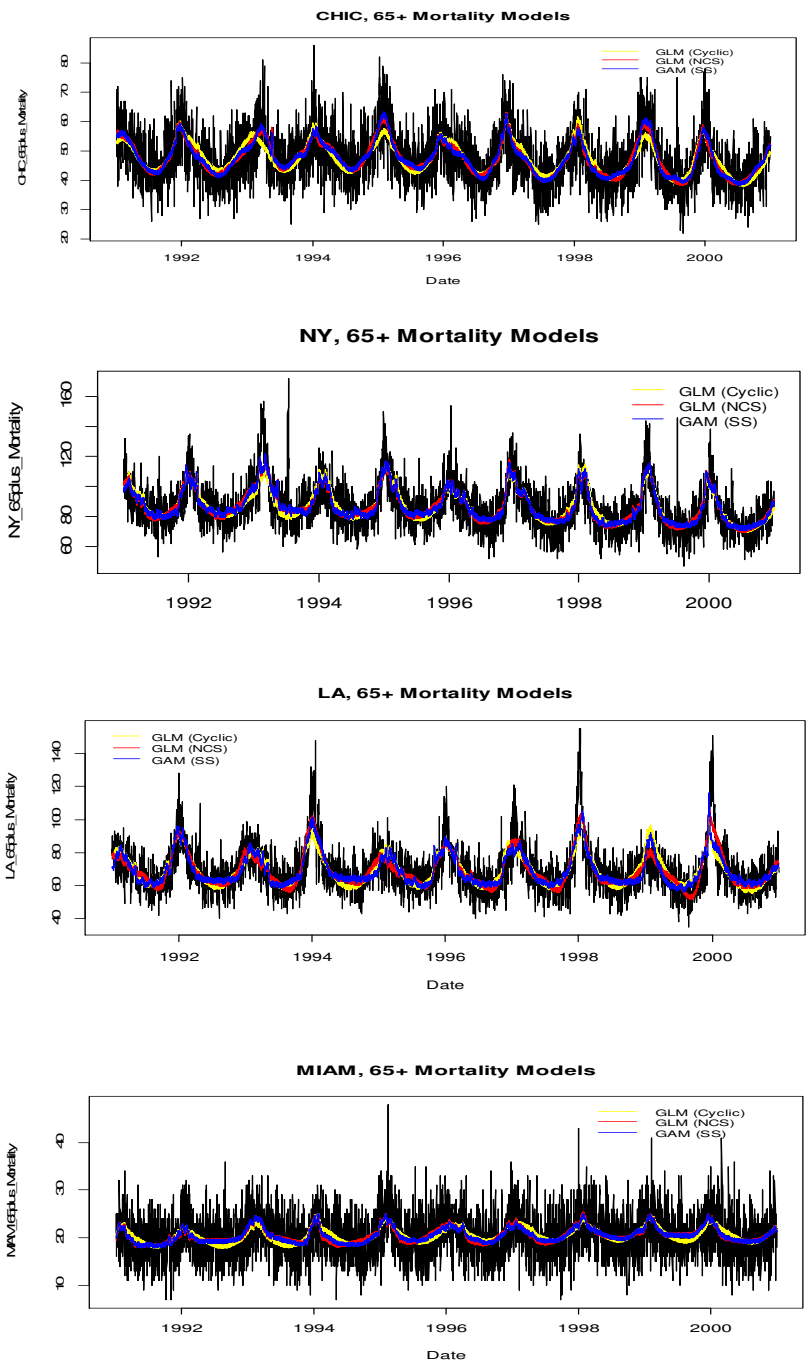

Figure 3.4: Observed and fitted mortality data for 4 U.S. cities. Results from the generalized additive models implemented in this study are represented by the blue, fitted lines.

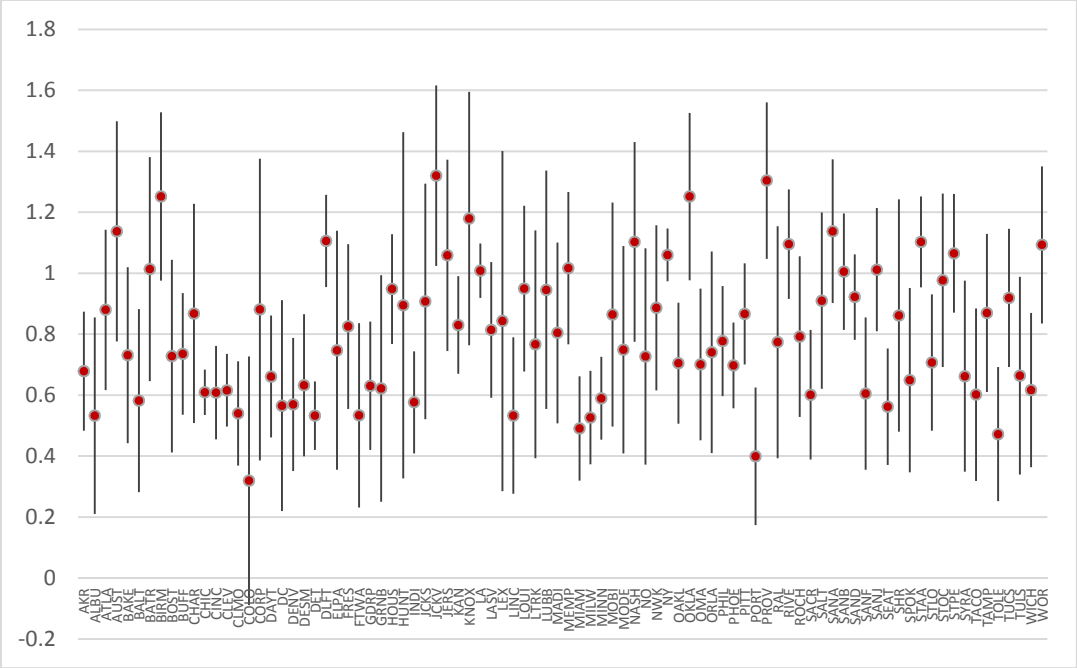

Figure 3.5: Log-relative risks of death among persons 65+ years of age and percent-positive A(H3N2) influenza activity during the 10-year period, 1991−2000, by city.

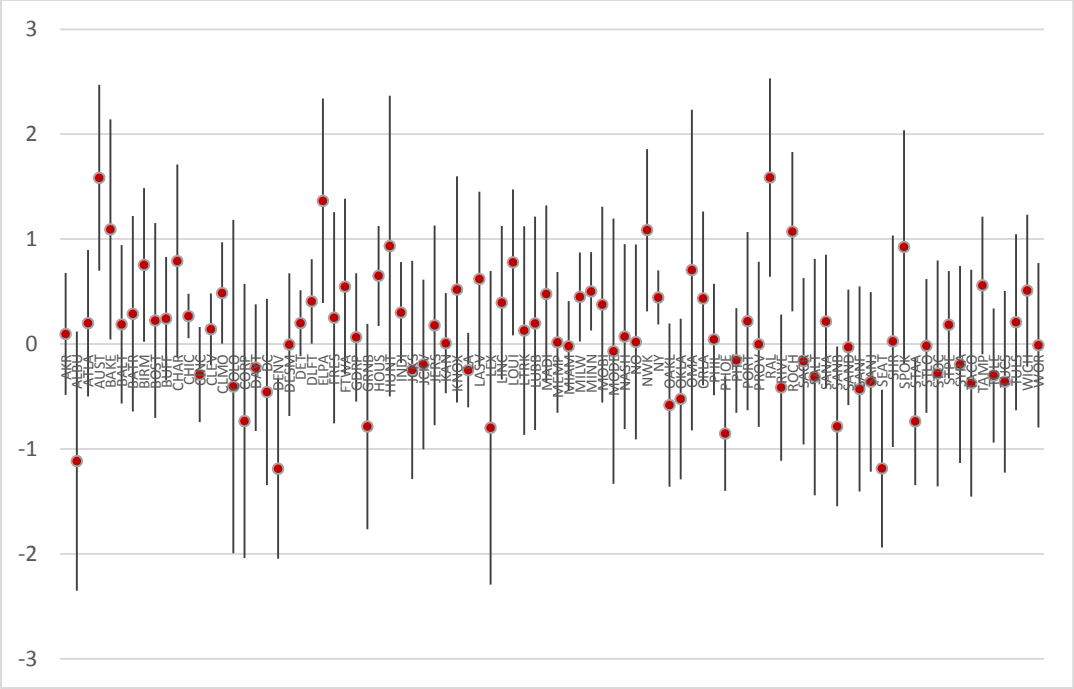

Figure 3.6: Log-relative risks of death among persons 65+ years of age and percent-positive A(H1N1) influenza activity during the 10-year period, 1991−2000, by city.

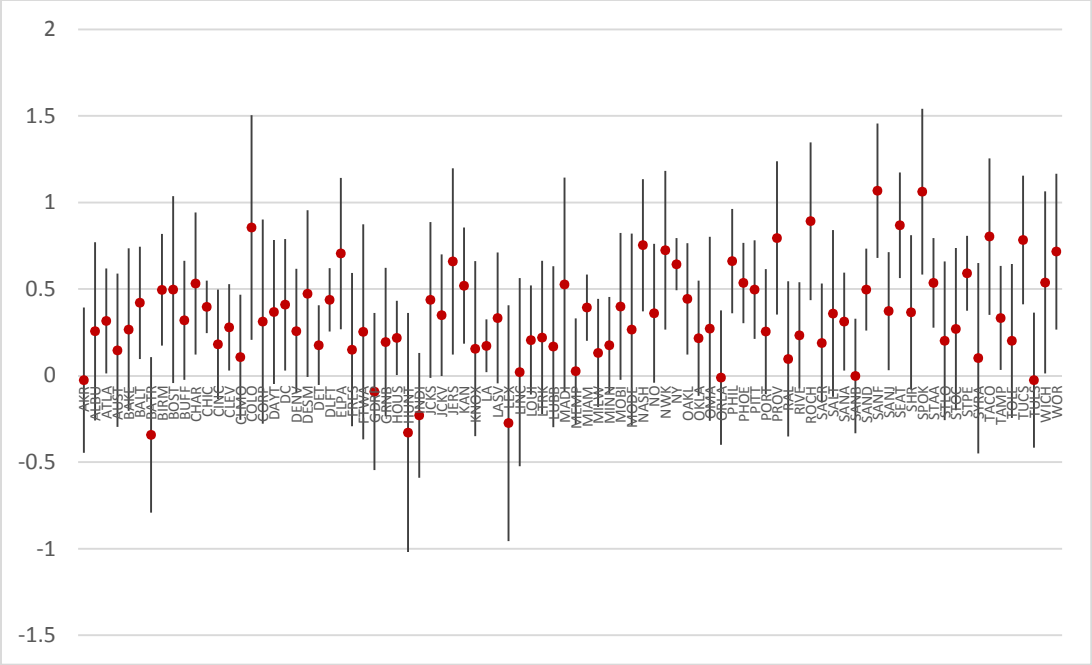

Figure 3.7: Log-relative risks of death among persons 65+ years of age and percent-positive type B influenza activity during the 10-year period, 1991−2000, by city.

Table 3.1: Relative risk of deaths among persons 65+ years of age and percent-positive influenza circulation during the 10-year period 1991−2000, by type/subtype and modeling approach.

Virus	Percent increase in viral activity	Independent-Observations Hierarchical Model	Spatial-Correlation Hierarchical Model	Traditional GLM Model
A(H3N2)	10%	1.083 (1.079, 1.088)	1.082 (1.078, 1.086)	1.092 (1.089, 1.095)
	20%	1.174 (1.163, 1.184)	1.171 (1.162, 1.180)	1.192 (1.186, 1.199)
	30%	1.271 (1.255, 1.288)	1.267 (1.253, 1.282)	1.302 (1.292, 1.312)
A(H1N1)	10%	1.010 (1.000, 1.020)	1.004 (0.968, 1.041)	1.035 (1.027, 1.042)
	20%	1.020 (1.000, 1.041)	1.008 (0.938, 1.084)	1.070 (1.055, 1.086)
B	10%	1.036 (1.030, 1.041)	1.046 (1.037, 1.055)	1.031 (1.027, 1.036)
	20%	1.073 (1.061, 1.084)	1.094 (1.075, 1.113)	1.064 (1.055, 1.073)

CHAPTER 4

MODELING INFLUENZA-ASSOCIATED MORTALITY WITH MEASURED AND UNMEASURED BACKGROUND SEASONALITY

Time-series regression is a common modeling approach used for estimating influenza-associated mortality. Deaths associated with a particular virus are estimated as the difference between the sum of predicted values from a fitted model containing all viral terms and the sum of predicted values from the same model excluding viral activity for a particular virus [Thompson et al., 2009]. Because this approach relies on unbiased parameter estimates of the viral terms in estimating attributable deaths, any confounding or collinearity among model terms should be understood and addressed. Collinearity among seasonal covariates can lead to unstable influenza-associated death estimates and inflated standard errors.

This chapter examines the effects of varying measures of seasonality on the association between influenza and mortality using regression-based modeling. Data from four U.S. cities with varying trends in mortality and weather were modeled (New York, Chicago, Miami, and Los Angeles). These data were described in detail in Chapter 3. In this study, background seasonality refers to the set of seasonal covariates included in a model that measurably affect the influenza-mortality relationship. These covariates include data-driven as well as mathematically modeled proxies for seasonality.

Section 4.1 describes in detail the three statistical models used to compare death estimates based on various representations of background seasonality. This section also includes a review of natural cubic spline functions. Results comparing influenza-associated death estimates across the three modeling methods are given in Section 4.2. Conclusions are given in Section 4.3 along with suggestions for improvements on modeling background seasonality in this context.

4.1 Statistical Models and Seasonality

To assess the impact of influential seasonal confounders, we compared parameter estimates of the viral terms using three modeling approaches which vary in their representations of background seasonality. Robust or stable parameter estimates across models would suggest minimal collinearity, while estimates that fluctuate with the inclusion of different seasonal covariates might suggest that the viral and seasonal effects are highly correlated and cannot be separated or assessed independently.

The three regression-based models that were implemented to quantify the association between influenza and mortality while accounting for background seasonality were: (*i*) GLM with negative binomial error structure and sinusoidal terms representing background seasonality, (*ii*) GLM with negative binomial error structure and natural cubic splines representing background seasonality, and (*iii*) GAMs with negative binomial error structure and smoothing splines representing background seasonality.

Four U.S. cities with varying trends in mortality and weather (New York, Chicago, Miami, and Los Angeles) were modeled. Each city was modeled separately. The outcome variable for all models was daily mortality due to underlying respiratory or circulatory (R&C) causes among persons aged 65 or older. Annual city-level population estimates of persons aged $\geq$65 years were used as the offset term to account for changes in population size over time. The percent-positive influenza subtypes {i.e., A(H1N1), A(H3N2), and B} were entered into all models as linear terms. A long-term trend was accounted for in each of the models by including linear and quadratic time terms. A day-of-week effect was also accounted for using six indicator variables (Sunday was the referent group).

To represent local-level patterns of background seasonality, several approaches were taken. Measured covariates such as city-level temperature were included directly into models. Functions of calendar time (e.g., splines) were used as proxies for time-

dependent, unmeasured covariates such as local patterns of non-influenza virus circulation. Measured seasonal covariates included in this study were temperature, dew point temperature, and percent-positive RSV. Percent-positive RSV was entered into all models linearly. Temperature and dew point temperature, a proxy for humidity, were entered into models as either current or lagged (backshifted a specified number of days) terms. These two covariates were entered into the models either linearly, via natural cubic splines, or as smoothing splines.

Unmeasured background seasonality was modeled by including time in days represented through natural cubic splines, smoothing splines, or Fourier functions. One advantage of modeling background seasonality via time splines is that the amount of seasonal curvature can be controlled by the number of df specified for the spline. This technique allows for an assessment of the change in influenza parameter estimates by degree of control for background seasonality. Fewer df translates to very little adjustment for background seasonality and vice versa. A review of natural cubic splines is given in Section 4.1.2, and a review of smoothing splines is given in Section 3.2.2.

4.1.1 Model 1: Poisson GLM with Sinusoidal Seasonality

Poisson generalized linear models with sinusoidal terms representing unmeasured seasonal confounders are widely used to model influenza-associated morbidity and mortality [Thompson et al., 2003; Thompson et al., 2009; Warren-Gash et al., 2011; Liao et al., 2009; Newall et al., 2010]. Because this modeling strategy is considered a fairly conventional approach, we use it as our referent model for comparison. The full model is described as follows:

$$\log(Y_i) = \log(a_i) + \beta_0 + \sum_{j=1}^{6} \beta_j [I_{ij}] + \beta_7 t_i + \beta_8 t_i^2 + \beta_9[\sin(2\pi t_i/365.25)] + \beta_{10}[\cos(2\pi t_i/365.25)] + \beta_{11}[\mathrm{A(H1N1)}_i] + \beta_{12}[\mathrm{A(H3N2)}_i] + \beta_{13}[\mathrm{B}_i] + \beta_{14}[\mathrm{RSV}_i]$$

where $Y_i \sim$Poisson, represents number of deaths on day i; α is the offset term equal to the annual population size; β_0 is the model intercept; $\beta_1$through β_6 are coefficients for day-of-week indicators, I_{ij} (Sunday = referent); β_7 and β_8 are coefficients for the long-term, nonlinear time trend; β_9 and β_{10} are coefficients for the seasonal trend; and β_{11} through β_{14} are coefficients for the percent positive viral terms.

4.1.2 Model 2: Negative Binomial GLM with Cubic Splines

Before specifying the generalized linear model with negative binomial error structure (GLM-NB), a review of natural cubic splines follows.

4.1.2.1 Natural Cubic Splines

Often in regression modeling, global polynomial functions are used to model curvature found in the association between the outcome and a particular predictor variable. Such functions, however, tend to fit data poorly near the boundaries of the outcome. Increasing the order of the polynomial to better fit a particular region often leads to erratic fit in another region. Splines serve as an alternative to sinusoidal terms or global polynomial terms, offering greater flexibility in curve-fitting. Splines are piecewise polynomials used to model complex curvature [Hastie et al., 2001]. For a particular covariate, X, the range of X is partitioned into $k+1$ intervals by k points, $\{x_1,\ldots,x_k\}$, referred to as knots. A separate cubic polynomial is fitted to each interval of data. Cubic splines are defined as follows:

Given a set of knots, $x_1 < x_2 < \cdots < x_k$, contained within interval $[a, b]$, a cubic spline is a function *f* such that (i) *f* is a cubic polynomial over each of $(x_1, x_2), (x_2, x_3), \ldots, (x_{k-1}, x_k)$ and (ii) *f* has continuous first and second derivatives at all knots. In general, an Mth-order spline is a piecewise $M-1$ degree polynomial with $M-2$ continuous derivatives at the knots.

Cubic splines ($M = 4$) are the lowest-order spline for which the discontinuity at the knots is not noticeable. Continuous first and second derivatives ensure smoothness across intervals. As such, cubic splines are one of the most-commonly used splines in practice. More-flexible curves are obtained by increasing the degree of the spline and/or by adding more knots. There is, however, a tradeoff. Too few knots or lower-order polynomials may result in a function that is too restrictive (high bias, low variance). Too many knots or higher-order polynomials may overfit the data (low bias, high variance).

Cubic splines, as defined above, tend to poorly fit the boundary intervals, $[a, x_1]$ and $[x_k, b]$, since no conditions are placed on the boundary points. Natural cubic splines address this problem by adding an additional constraint to cubic splines, that is, the cubic spline must be linear beyond the boundaries of the data. Though this constraint introduces some bias near the boundaries, the trade-off is preferable to a spurious outcome. A natural cubic spline can be defined as follows:

Let $a = x_1 < x_2 < \cdots < x_{k+1} = b$ be a partition of $[a, b]$. Note here that, unlike the cubic splines definition above, boundary points are defined as knots. On each subinterval $[x_i, x_{i+1}]$, $i = 1, \ldots, k$, we fit a cubic polynomial

$$f_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3.$$

The natural cubic spline F has the following properties:

(1) $F(x) = f_i(x), \quad x \in [x_i, x_{i+1}]$ for $i = 1, \ldots, k$

(2) $f_i(x_{i+1}) = f_{i+1}(x_{i+1}),$ for $i = 1, \ldots, k-1$

(3) $f_i'(x_{i+1}) = f_{i+1}'(x_{i+1}),$ for $i = 1, \ldots, k-1$

(4) $f_i''(x_{i+1}) = f_{i+1}''(x_{i+1}),$ for $i = 1, \ldots, k-1$

(5) $F''(x_1) = F''(x_{k+1}) = 0.$

To calculate the number of df for a natural cubic spline, i.e., the number of basis functions added to the covariate matrix, note that F introduces $4k$ parameters to the

covariate matrix. Equations (2)−(4) impose $3 \times (k-1)$ constraints on the internal knots. Equation (5) imposes 1 constraint on each boundary knot. Thus,

$$\text{df} = 4k - (3k-3) - 2 = k + 1.$$

For natural cubic splines, the number of basis functions equals the total number of knots defining the spline.

A basis is defined as a set of orthogonal functions spanning a particular vector or function space. In other words, a linear combination of a set of basis functions can define any vector or function within the spanned space. Many sets of basis functions exist for a particular function space, and many can be used to define natural cubic splines. The B-spline basis is generally used to define natural cubic splines. The B-spline basis is defined as follows:

Let $a = x_1 < x_2 < \cdots < x_{k+1} = b$ be a partition of $[a, b]$. A new knot sequence τ is defined such that

$$\tau_1 \leq \tau_2 \leq \cdots \leq \tau_M \leq x_1,$$

$$\tau_{j+M} = x_j,\ j = 1, \ldots, k,$$

$$x_{k+1} \leq \tau_{k+M+1} \leq \tau_{k+M+2} \leq \cdots \leq \tau_{k+2M}.$$

The values of the knots augmenting the original set are arbitrary. Typically, these new knots are set to equal x_1 for all lower-valued knots and x_{k+1} for all higher-valued knots. Denoting $B_{i,m}(x)$ as the ith basis function of order m for the sequence τ where $m = 1, \ldots, M$, the B-spline basis functions are then defined recursively as:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x \leq \tau_{i+1} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \ldots, k + 2M - 1,$$

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x) \quad \text{for } i = 1, \ldots, k + 2M - m.$$

Since the basis functions $\{B_{i,m}\}$ for natural cubic splines are fixed and orthogonal, they are entered and fitted in linear or generalized linear models via ordinary least squares or maximum likelihood, respectively, along with all other covariates in the model.

4.1.2.2 Details of Model 2

Negative binomial regression models are often used to account for potential Poisson under- or over-dispersion [Hilbe, 2008]. The distribution can be formulated in terms of the mean and a dispersion parameter, therefore offering greater flexibility than the Poisson distribution in model fitting [Hilbe, 2008]. The general model can be described as follows:

$$\log(Y_i) = \log(a_i) + \beta_0 + \sum_{j=1}^{6} \beta_j \left[I_{ij}\right] + \beta_7[\mathrm{A(H3N2)}_i] + \beta_8[\mathrm{A(H1N1)}_i] + \beta_9[\mathrm{B}_i] + \beta_{10}[\mathrm{RSV}_i] + \sum_{k=11}^{m} \beta_k[f(t_{ik})] + \sum_{p=m+1}^{r} \beta_p\left[f(\mathrm{Temp}_{lpi})\right],$$

where $Y_i \sim$ Negative Binomial, represents the number of deaths on day i; α is the offset term equal to annual population size; β_0 represents the model intercept; $\beta_1$through β_6 are coefficients for day-of-week indicators, I_{ij} (Sunday = referent); $\beta_7$through β_{10} are coefficients for percent positive viral terms; β_{11} through β_m are coefficients for basis functions of a natural cubic spline representing either long-term or seasonal time trends; and β_{m+1} through β_r are coefficients for basis functions of a natural cubic spline representing a temperature effect lagged l days.

All natural cubic splines are entered into the model linearly through their basis functions. The number of basis functions depends on the df (or knots) defining the spline. To represent long-term trends, knots were placed every 2 to 5 years (determined by lowest Akaike Information Criterion value). Note that if time splines were used to represent seasonality, then no long-term trend covariate was included in the model since this trend would have been automatically incorporated into the seasonal time spline.

To represent seasonality, knots placed every 3 to 6 months would have been an intuitive choice based on *a priori* knowledge of general influenza circulation in the U.S. We modeled time splines one knot at a time (up to 60 knots) to assess the effect of low vs. high representations of background seasonality on the influenza parameter estimates.

Underrepresenting background seasonality (too few knots) should inflate influenza parameter estimates and result in some amount of autocorrelation in the residuals. Overrepresenting background seasonality (too many knots) should lead to deflated or erratic influenza parameter estimates, and uncorrelated residuals. In other words, underestimating background seasonality should lead to an overestimation of influenza-related mortality, and vice versa. The goal is to find an optimal number of knots to represent background seasonality so that the influenza signal is not amplified – or lost – in modeling.

4.1.3 Model 3: Negative Binomial GAM with Smoothing Splines

Although simple and easy to implement, GLMs tend to poorly model effects characterized by complex nonlinearity. Natural cubic splines address this issue by modeling lower-order polynomials piecewise over the full range of the variable. GAMs are another class of regression models which drop the assumption of linearity, thus making them more-flexible modeling tools compared to GLMs [Hastie, 1990; Hastie, 2001; Wood, 2006]. With this additional flexibility, nonlinear associations may be better modeled or revealed using GAMs. Additivity across effects is still assumed allowing for interpretation of results similar to that of GLMs. GAMs are defined as:

$$g[\mu(X)] = a + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) + \varepsilon,$$

where the error term has mean 0, and the nonparametric f_i functions are estimated using a scatterplot smoother such as a smoothing spline. GAMs and smoothing splines are described in detail in Chapter 2. In this study, we are using smoothing splines to approximate the effects of background seasonality on the influenza-mortality association. We do not use a formal method of selecting the smoothing parameter in order to observe how the effective degrees of freedom (edf) − ranging from 1 to 60 − of the time-splines affects the influenza parameters.

The following semi-parametric GAM model was implemented:

$$\log(Y_i) = \log(a) + \beta_0 + \sum_{j=1}^{6} \beta_j [I_{ij}] + \beta_7[\mathrm{A(H3N2)}_i] + \beta_8[\mathrm{A(H1N1)}_i]$$

$$+ \beta_9[\mathrm{B}_i] + \beta_{10}[\mathrm{RSV}_i] + f(t_i) + f(\mathrm{Temp}_{li}),$$

where $Y_i \sim$ NegBin, represents the number of deaths on day i; α is the offset term equal to annual population size; β_0 represents the model intercept; β_1 through β_6 are coefficients for day-of-week indicators, I_{ij} (Sunday = referent); β_7 through β_{10} are coefficients for percent positive viral terms; $f(t_i)$ is a smoothing spline representing either long-term or seasonal time trends; and $f(\mathrm{Temp}_{li})$ is a smoothing spline representing a temperature effect lagged l days.

4.2 Comparing Influenza-associated Death Estimates

Figure 3.2 from Chapter 3 shows daily time series of respiratory and circulatory deaths, temperature, dew point temperature, and regional RSV circulation by city. Across all cities, temperature and dew point temperature were found to be highly collinear (correlation coefficient ranging from $r = 0.61$ in Los Angeles to $r = 0.95$ in New York). As such, the two terms could not be modeled together. The temperature term tended to be more-highly associated with mortality; therefore, dew point temperature was dropped from all models. We also tried modeling an adjusted dew point temperature by first regressing dew point temperature on temperature, and then modeling the residuals. This term, however, did not substantially influence any of the virus parameter estimates and therefore was dropped from all models. Figure 3.3 shows daily time series of respiratory and circulatory deaths and regional percent-positive viral activity of the three influenza types/subtypes (A(H1N1), A(H3N2), and B) by city. Note that A(H3N2) was the dominant influenza strain during this 10-year period. The A(H1N1) subtype shows very little activity during this period.

Figure 4.1 contains four plots for New York City. The *y*-axis on all plots is the parameter estimate for the specified viral terms: A(H1N1), A(H3N2), B, or RSV. The *x*-

axis shows the df used to represent each spline. For smoothing splines, this value represents the edf determined by the amount of penalization due to the smoothing parameter. Results from two models are plotted: GLM-NB with natural cubic splines (black) and GAM-NB with smoothing splines (red). Figures 4.2, 4.3, and 4.4 show the same plots for Chicago, Los Angeles, and Miami, respectively. For all plots, df = edf = 10 translates to 1 knot per year, while df = edf = 20 describes 1 knot every 6 months, and df = edf = 40 describes 1 knot every 3 months.

Using the GLM-NB model, the A(H3N2) parameter estimate was stable up to 20 knots in all four cities. At 20 knots, there is a slight drop in A(H3N2) parameter estimates. Beyond 20 knots, the parameter estimates erratically decline. Based on the GAM results, the parameter estimates exhibit a smooth but similar pattern up to edf = 20, after which the estimates smoothly decline.

For RSV, a pattern similar to A(H3N2) estimates is observed. The dip in parameter estimates at 20 df is more substantial for the colder cities, New York and Chicago. There is a minimal drop at df = 20 for Los Angeles and a minimal increase in estimates for Miami. It was discovered when modeling the RSV term with seasonal covariates, particularly the ambient temperature and the Fourier terms, that the terms are very highly correlated (e.g., $r = 0.95$ for RSV with the cosine term in the NYC mortality model, and $r = 0.83$ for RSV modeled with temperature).

Because there was so little circulation of A(H1N1) during this 10-year period, the confidence intervals are much wider for this term across all four cities relative to the other viral terms. Given the GLM-NB model, New York and Chicago have erratic parameter estimates even with <20 df. Los Angeles and Miami are more stable up to 20 knots using either method.

For influenza type B, there is a decline in parameter estimates with increasing spline df or edf for all 4 cities. There appears to be slight stability in the B estimates, up to 10 df (or edf) for New York, Chicago, and Miami, but a steady drop beyond that

measure. This downward trend is not as pronounced in Los Angeles. At 10 df, the GLM-NB estimates increase slightly and remain fairly stable up to 20 df.

Tables 4.1−4.4 list results from 10 plausible models for each of the 4 cities. Shown are viral parameter estimates, and covariate terms included in the model to represent long-term or seasonal trends. Of note, terms modeled with temperature data yielded nearly identical results whether the temperature term was included linearly or via splines. In all cities, lagging the temperature series by a few days led to better fits than entering the current day's temperature (based on AIC results). Lagging temperature, however, led to decreased estimates for the influenza terms. A(H3N2) yielded the most-stable estimates across models and various representations of background seasonality. RSV terms were found to be highly correlated with the cosine function ($r = 0.7$ to $r = 0.95$), and to a lesser but still significant degree, with the temperature series.

Figures 4.5 through 4.8 show the fitted functions of 3 models on each of the 4 cities' mortality data. Table 4.5 is a summary of influenza-related death estimates by city and model. Estimated deaths attributable to the A(H3N2) subtype tend to be highest using the GAM models and lowest using the GLM-Poisson model. Across all models, the fewest deaths were attributed to A(H1N1) in New York and Chicago, while no or very few deaths were attributed to A(H1N1) in Los Angeles or Miami. Deaths associated with subtype B were the most-variable across all models for Los Angeles. For RSV, the GAM model attributed the highest number of deaths to influenza, while the GLM-Poisson model attributed the fewest.

4.3 Conclusions and Modeling Recommendations

Disentangling the effects of seasonal confounders on the association between influenza and mortality is challenging given the level of temporal confounding and collinearity among modeled covariates. Representing background seasonality appropriately is critical in approximating the influenza-mortality association. Overestimating background

seasonality will underestimate the impact of influenza on mortality. Underestimating background seasonality, however, will attribute too many deaths to influenza.

Based on results from this study, background seasonality affects estimates of influenza-associated deaths to varying degrees depending on viral subtype, city, and climate. Estimates of A(H3N2) deaths are the most robust, with very little evidence of seasonal confounding across model types or cities in the data used. During this 10-year period, there was not enough A(H1N1) activity for valid statistical estimation in Los Angeles and Miami. For New York and Chicago, A(H1N1) parameter estimates decrease significantly at df = 20 suggesting that the A(H1N1) subtype may be highly correlated with seasonal covariates. Influenza type B does not appear to be highly associated with mortality, showing stable estimates only up to 10 df (roughly 1 knot per year).

The significance of df = 10 and df = 20 in this context should be noted. Ten df roughly means that a single cubic function is used to model background seasonality for a full year of mortality data. In the context of parametric cubic splines, one knot per year is not enough to appropriately model the distribution ('rise' and 'fall') over time of a seasonal covariate. One knot per year, in this context, translates to a roughly linear or long-term adjustment rather than seasonal adjustment of background cofactors. Since type B parameter estimates are only stable with splines containing 1 df per year, it suggests that the association between B and mortality is not strong when other seasonal factors are accounted for. This is in stark contrast to A(H3N2) parameter estimates which show a strong association with mortality despite adjustment of other strong seasonal factors. Twenty df translates to roughly 1 knot every six months. In this context, this means that two cubic polynomials were used to model annual background seasonality; one for the period from January to June, and another for July to December (see Figures 3.5−3.8). Generally, in the U.S., only one wintertime peak in mortality is observed annually. Thus, findings from these analyses suggest that natural cubic splines

containing >20 knots overfit seasonal background, leading to erratic influenza parameter estimates.

In our data, RSV was highly correlated with seasonality. Its parameter estimates are unstable particularly in the two colder cities. Based on results from this study, the RSV term should not be modeled with sinusoidal terms representing background seasonality. RSV and the cosine function were found to be highly collinear ($r = 0.91$). The RSV series was, in fact, so closely associated with temperature and seasonal effects in New York and Chicago that RSV-associated death estimates from these cities were indeterminate. It should be noted, however, that these highly unstable RSV estimates do not imply that RSV activity does not affect mortality. One possible approach to break up this collinearity would be to employ an alternative proxy for the RSV term. Rather than using percent-positive RSV activity, a proxy based on both viral circulation and influenza-like illness (ILI) activity may help reduce the near-perfect collinearity between the virus' seasonal behavior and the cosine function.

On modeling influenza mortality with parameter splines, the use of forty knots (1 cubic polynomial modeled every 3 months) to model background seasonality clearly overfits the data, losing the influenza effect on mortality and leading to unstable parameter estimates. For influenza type B and subtype A(H3N2), knots at roughly every six months lead to a slight drop in the influenza parameter estimates compared to the GAM influenza parameter estimates. Because natural cubic splines are entered into the covariate matrix with influenza terms and other seasonal confounders, including such splines with too many knots suggests that collinearity may be introduced. With the GAM models, however, this drop does not occur. Regression splines are sensitive to the number of knots and their placement. Smoothing splines appear to be less sensitive to collinearity likely because they are not included directly in the covariate matrix. They are instead calculated via a regularization method which helps reduce collinearity in seasonal parameter estimates.

Based on results from this study, temperature may serve as a reasonable proxy for seasonality. Influenza parameter estimates based on temperature as a proxy for background seasonality were very similar to those from splines with df = 20 or edf = 20. Further, temperature data cannot be inadvertently overmodeled via spline functions in the way mathematical proxies for seasonality often are.

Finally, there seems to be no one-size-fits-all model. A number of plausible models should be assessed to determine appropriate and reasonable parameter estimates. Season-specific influenza virus data could be used to estimate the length in weeks of each influenza season. This information would further help in determining reasonable ways to account for non-influenza seasonality. Simulation models with "known" influenza effects are also needed to help assess the validity of a variety of theoretically plausible model types.

4.4 Limitations

There are several limitations to this study. First, the influenza proxy used here may be an inadequate representation of influenza activity. Influenza proxies based on percent-positive influenza activity and influenza-like illness data have been suggested as alternatives [Goldstein et al., 2012]. Second, for smaller cities, mortality counts at the city level may be too small to be properly modeled. For these situations, a zero-inflated negative binomial model may be more appropriate for modeling. Residuals from regression models still have slight periodicity in some models. As such, confidence intervals may need to be bootstrapped rather than estimated parametrically.

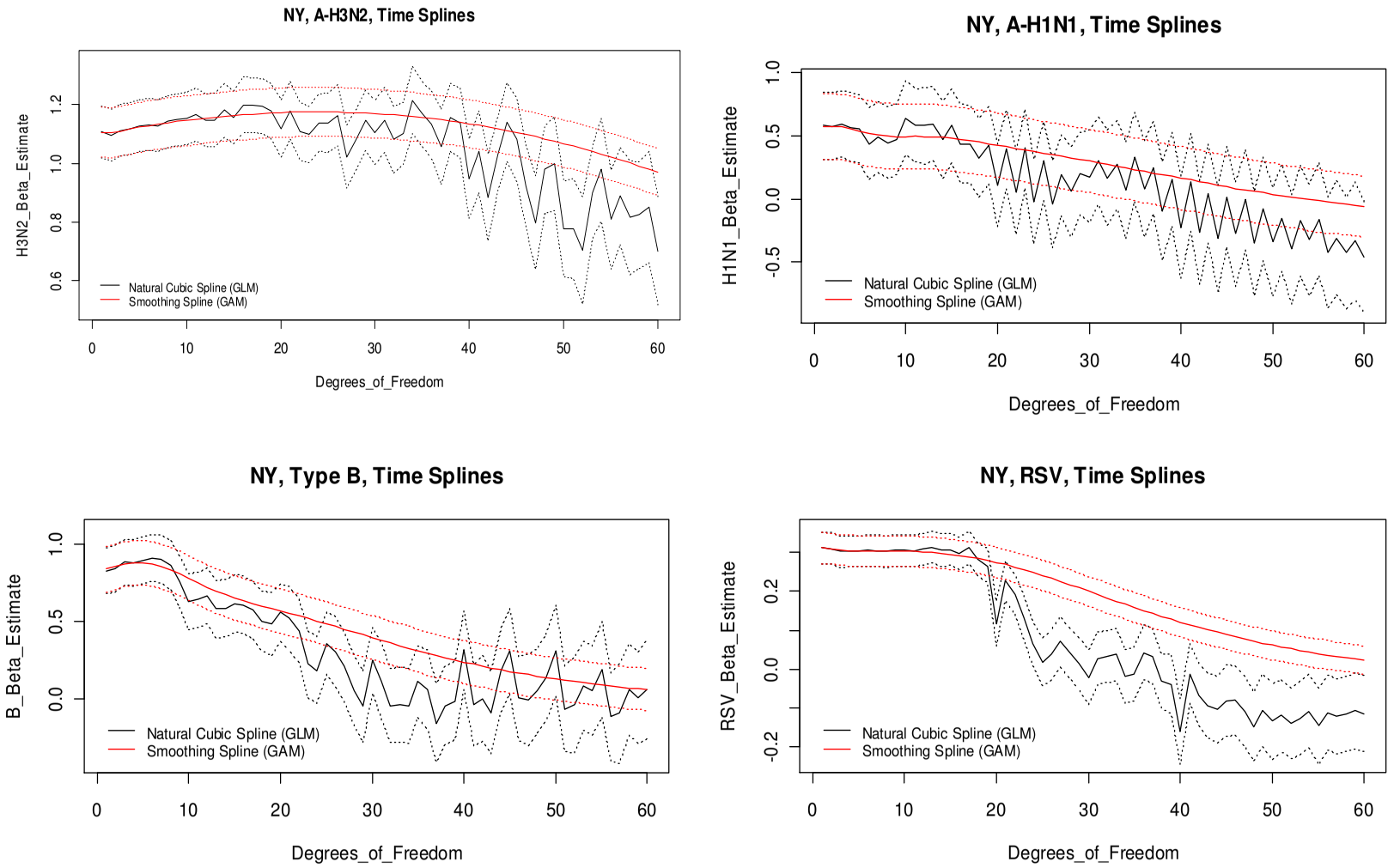

Figure 4.1: Plots of viral parameter estimates (A-H3N2, A-H1N1, B, RSV) by degrees of freedom for natural cubic splines in GLM models (black), and by effective degrees of freedom for smoothing splines from GAM models (red), New York City.

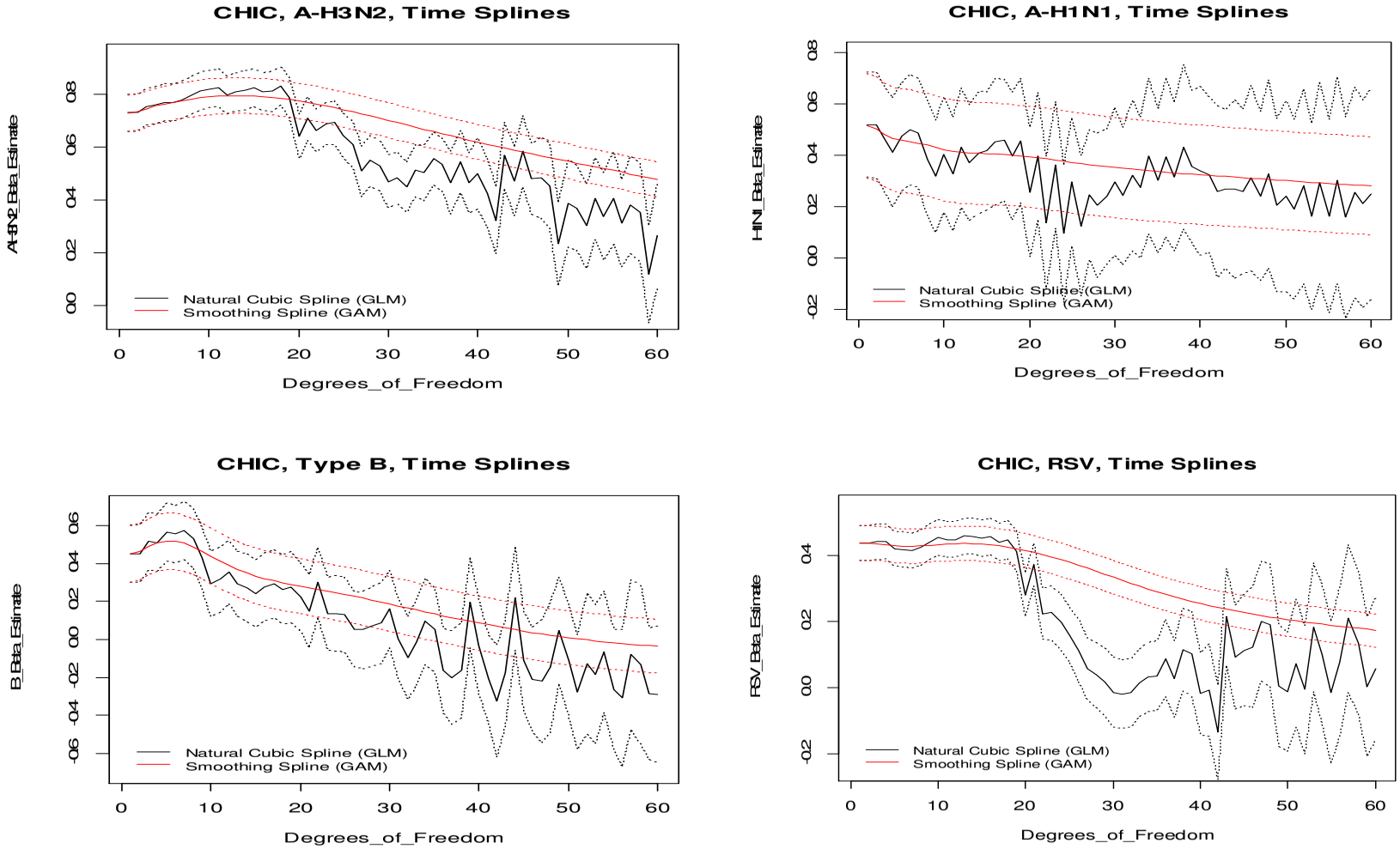

Figure 4.2: Plots of viral parameter estimates (A-H3N2, A-H1N1, B, RSV) by degrees of freedom for natural cubic splines in GLM models (black), and by effective degrees of freedom for smoothing splines from GAM models (red), Chicago.

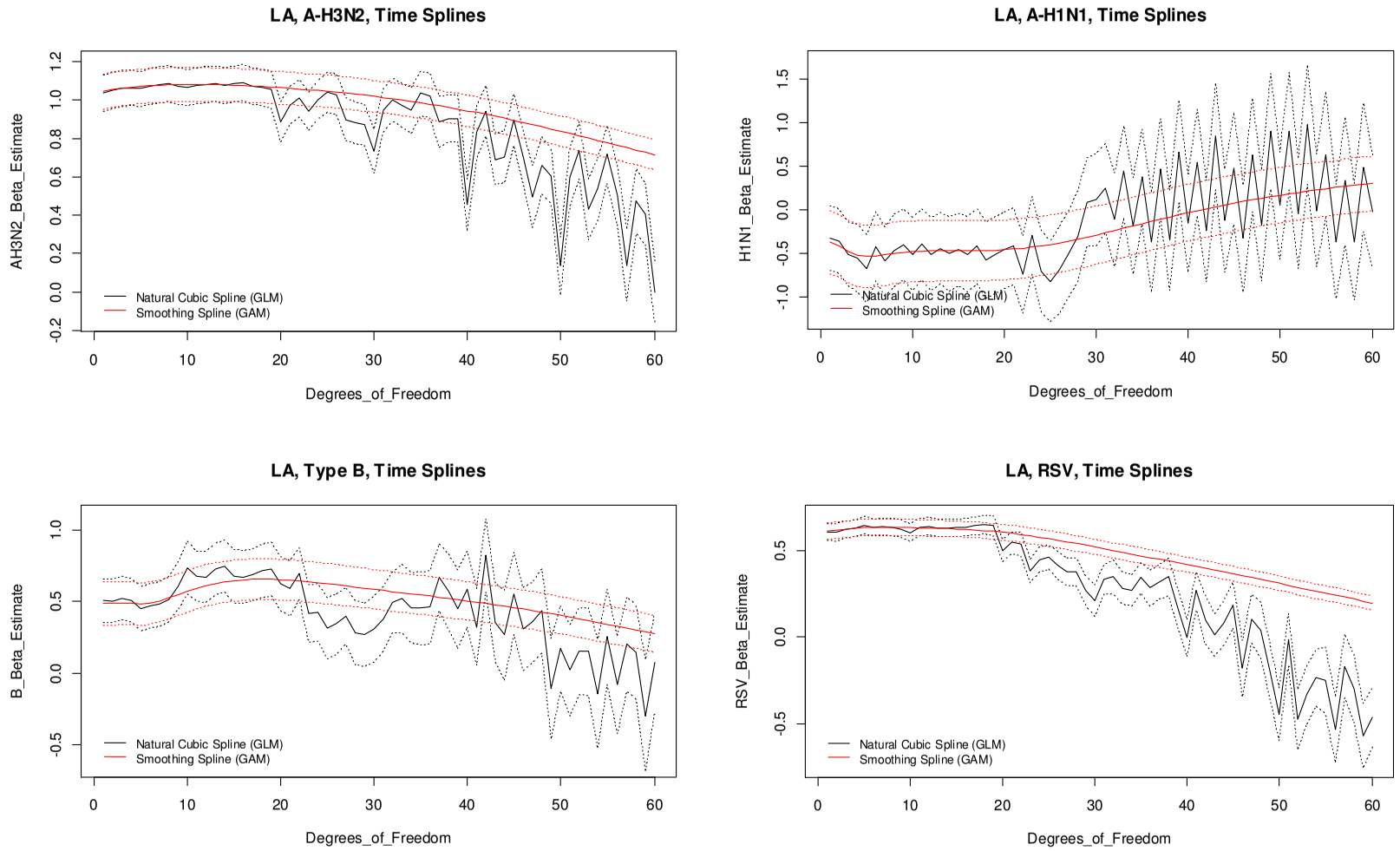

Figure 4.3: Plots of viral parameter estimates (A-H3N2, A-H1N1, B, RSV) by degrees of freedom for natural cubic splines in GLM models (black), and by effective degrees of freedom for smoothing splines from GAM models (red), Los Angeles.

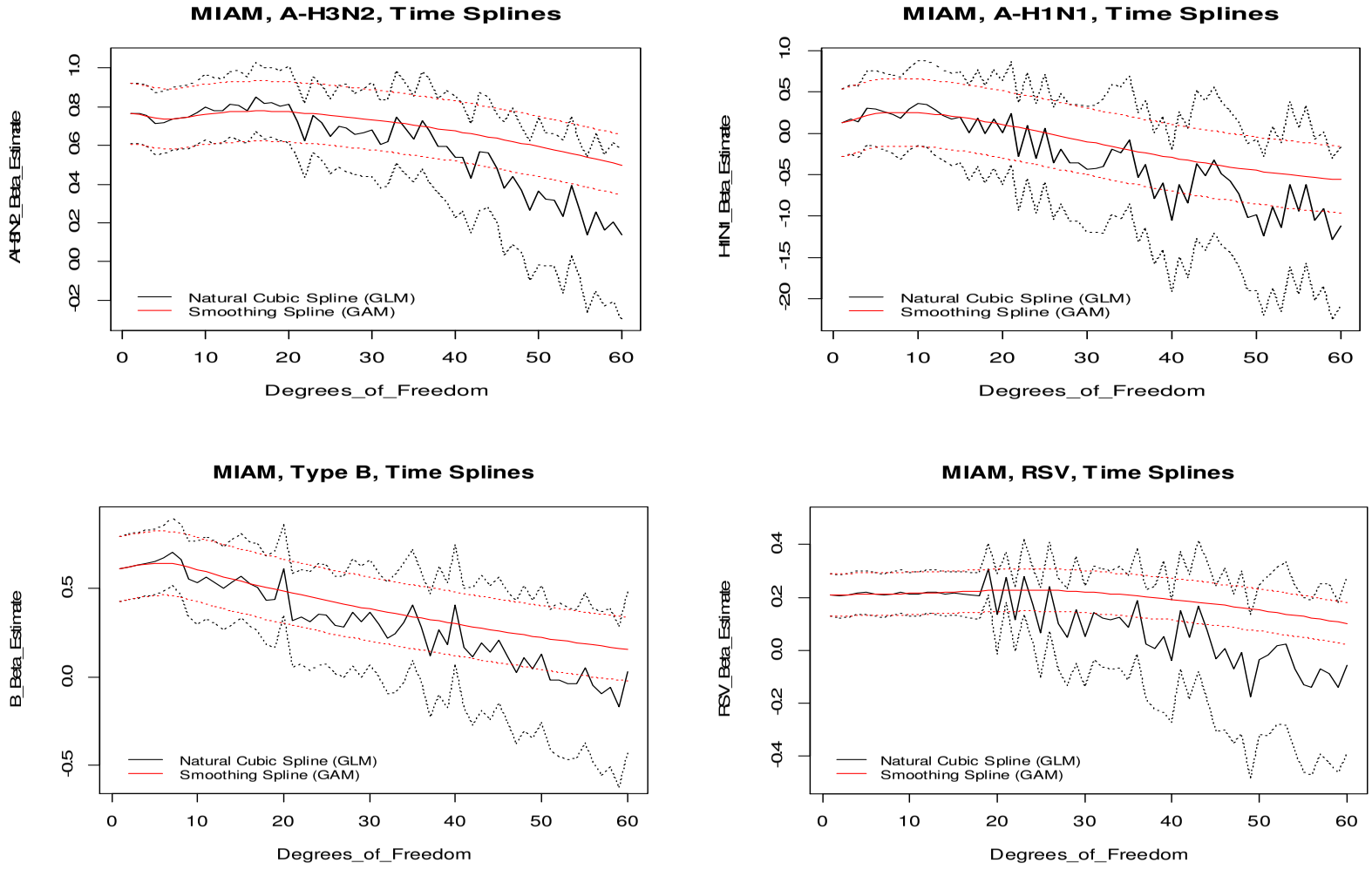

Figure 4.4: Plots of viral parameter estimates (A-H3N2, A-H1N1, B, RSV) by degrees of freedom for natural cubic splines in GLM models (black), and by effective degrees of freedom for smoothing splines from GAM models (red), Miami.

Table 4.1: Viral parameter estimates based on 11 models with varying representations of background seaonality: New York City

	Covariates			Influenza and RSV Parameter Estimates†			
	Measured	Unmeasured					
Regression Model†	Weather (Temp or DPT)	Long-term Trend	Seasonal Trend	AH3N2 (p)	AH1N1 (p)	B (p)	RSV (p)
GLM - NB		t (NS,df=5)		1.13 ***	0.55 ***	0.90 ***	0.30 ***
GLM - Pois		$t + t^2$	sin + cos	0.98 ***	0.35 **	0.69 ***	0.07 **
GLM - NB	Temp	t (NS,df=5)		1.12 ***	0.53 ***	0.87 ***	0.28 ***
GLM - NB	Temp (NS,df=20)	t (NS,df=5)		1.11 ***	0.48 ***	0.85 ***	0.26 ***
GLM - NB	Lag3Temp	t (NS,df=5)		1.05 ***	0.40 **	0.73 ***	0.15 ***
GLM - NB	Lag3Temp (NS,df=20)	t (NS,df=5)		1.04 ***	0.37 **	0.72 ***	0.13 ***
GLM - NB	Lag3Temp (NS,df=20)		t (NS,df=20)	1.04 ***	0.06 ns	0.40 ***	0.03 ns
GLM - NB			t (NS,df=20)	1.12 ***	0.11 ns	0.56 ***	0.12 **
GAM - QP	Temp (SS,df=20)	t (SS,df=5)		1.09 ***	0.45 ***	0.83 ***	0.26 ***
GAM - QP	Lag3Temp (SS,df=20)	t (SS,df=5)		1.04 ***	0.34 **	0.70 ***	0.13 ***
GAM - QP			t (SS,df=20)	1.18 ***	0.42 **	0.58 ***	0.27 ***

†All models contain DOW and population offset term

†† P-values: * < .05, ** < 0.01, *** < 0.001, ns = not significant

Table 4.2: Viral parameter estimates based on 11 models with varying representations of background seaonality: Chicago

	Covariates			Influenza and RSV Parameter Estimates†			
	Measured	Unmeasured					
Regression Model†	Weather (Temp or DPT)	Long-term Trend	Seasonal Trend	AH3N2 (p)	AH1N1 (p)	B (p)	RSV (p)
GLM - NB		t (NS,df=4)		0.76 ***	0.41 ***	0.51 ***	0.44 ***
GLM - Pois		$t + t^2$	sin + cos	0.48 ***	0.28 **	0.17 *	0.00 ns
GLM - NB	Temp	t (NS,df=4)		0.66 ***	0.27 *	0.40 ***	0.29 ***
GLM - NB	Temp (NS,df=20)	t (NS,df=4)		0.63 ***	0.25 *	0.37 ***	0.27 ***
GLM - NB	Lag3Temp	t (NS,df=4)		0.60 ***	0.18 ns	0.32 ***	0.19 ***
GLM - NB	Lag3Temp (NS,df=20)	t (NS,df=4)		0.58 ***	0.18 ns	0.29 ***	0.17 ***
GLM - NB	Lag3Temp (NS,df=20)		t (NS,df=20)	0.58 ***	0.22 ns	0.10 ns	0.15 ***
GLM - NB			t (NS,df=20)	0.64 ***	0.26 *	0.22 *	0.28 ***
GAM - QP	Temp (SS,df=20)	t (SS,df=4)		0.63 ***	0.28 **	0.36 ***	0.25 ***
GAM - QP	Lag3Temp (SS,df=20)	t (SS,df=4)		0.58 ***	0.22 *	0.30 ***	0.15 ***
GAM - QP			t (SS,df=20)	0.77 ***	0.39 ***	0.28 ***	0.42 ***

†All models contain DOW and population offset term

†† P-values: * < .05, ** < 0.01, *** < 0.001, ns = not significant

Table 4.3: Viral parameter estimates based on 11 models with varying representations of background seaonality: Los Angeles

	Covariates			Influenza and RSV Parameter Estimates†			
	Measured	Unmeasured					
Regression Model†	Weather (Temp or DPT)	Long-term Trend	Seasonal Trend	AH3N2 (p)	AH1N1 (p)	B (p)	RSV (p)
GLM - NB		t (NS,df=5)		1.06 ***	-0.68 ns	0.45 ***	0.64 ***
GLM - Pois		$t + t^2$	sin + cos	0.84 ***	-0.20 ns	0.04 ns	0.23 ***
GLM - NB	Temp	t (NS,df=5)		0.99 ***	-0.65 ***	0.30 **	0.56 ***
GLM - NB	Temp (NS,df=20)	t (NS,df=5)		0.98 ***	-0.66 ***	0.21 **	0.53 ***
GLM - NB	Lag3Temp	t (NS,df=5)		0.95 ***	-0.59 **	0.15 ns	0.49 ***
GLM - NB	Lag3Temp (NS,df=20)	t (NS,df=5)		0.93 ***	-0.61 **	0.08 ns	0.47 ***
GLM - NB	Lag3Temp (NS,df=20)		t (NS,df=20)	0.85 ***	-0.52 *	0.40 ***	0.40 ***
GLM - NB			t (NS,df=20)	0.89 ***	-0.46 *	0.63 ***	0.50 ***
GAM - QP	Temp (SS,df=20)	t (SS,df=5)		0.99 ***	-0.51 **	0.28 ***	0.52 ***
GAM - QP	Lag3Temp (SS,df=20)	t (SS,df=5)		0.95 ***	-0.50 **	0.19 *	0.48 ***
GAM - QP			t (SS,df=20)	1.08 ***	-0.52 **	0.68 ***	0.61 ***

†All models contain DOW and population offset term

†† P-values: * < .05, ** < 0.01, *** < 0.001, ns = not significant

Table 4.4: Viral parameter estimates based on 11 models with varying representations of background seaonality: Miami

	Covariates			Influenza and RSV Parameter Estimates†			
	Measured	Unmeasured					
Regression Model†	Weather (Temp or DPT)	Long-term Trend	Seasonal Trend	AH3N2 (p)	AH1N1 (p)	B (p)	RSV (p)
GLM - NB		t (NS,df=2)		0.77 ***	0.17 ns	0.62 ***	0.20 ***
GLM - Pois		$t + t^2$	sin + cos	0.51 ***	0.12 ns	0.36 ***	0.04 ns
GLM - NB	Temp	t (NS,df=2)		0.66 ***	0.08 ns	0.55 ***	0.14 **
GLM - NB	Temp (NS,df=20)	t (NS,df=2)		0.63 ***	0.04 ns	0.53 ***	0.14 **
GLM - NB	Lag3Temp	t (NS,df=2)		0.51 ***	0.00 ns	0.45 ***	0.07 ns
GLM - NB	Lag3Temp (NS,df=20)	t (NS,df=2)		0.45 ***	-0.04 ns	0.41 ***	0.08 ns
GLM - NB	Lag3Temp (NS,df=20)		t (NS,df=20)	0.57 ***	-0.05 ns	0.39 **	0.12 ns
GLM - NB			t (NS,df=20)	0.81 ***	0.01 ns	0.61 ***	0.14 ns
GAM - QP	Temp (SS,df=20)	t (SS,df=2)		0.63 ***	0.03 ns	0.53 ***	0.14 **
GAM - QP	Lag3Temp (SS,df=20)	t (SS,df=2)		0.46 ***	-0.08 ns	0.41 ***	0.08 ns
GAM - QP			t (SS,df=20)	0.77 ***	0.11 ns	0.48 ***	0.23 ***

†All models contain DOW and population offset term

†† P-values: * < .05, ** < 0.01, *** < 0.001, ns = not significant

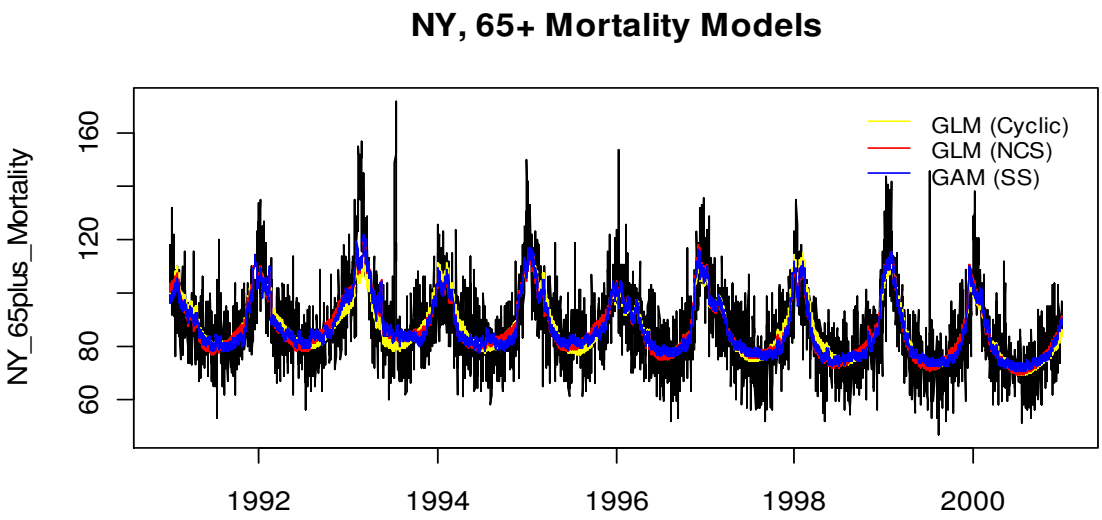

Figure 4.5: New York City – Observed daily mortality counts for population 65+ years of age, and fitted mortality derived from Model 1: GLM-Poisson with Fourier terms for background seasonality (yellow); Model 2: GLM-NB with natural cubic spline for background seasonality (red); and Model 3: GAM-NB with smoothing spline for background seasonality (blue).

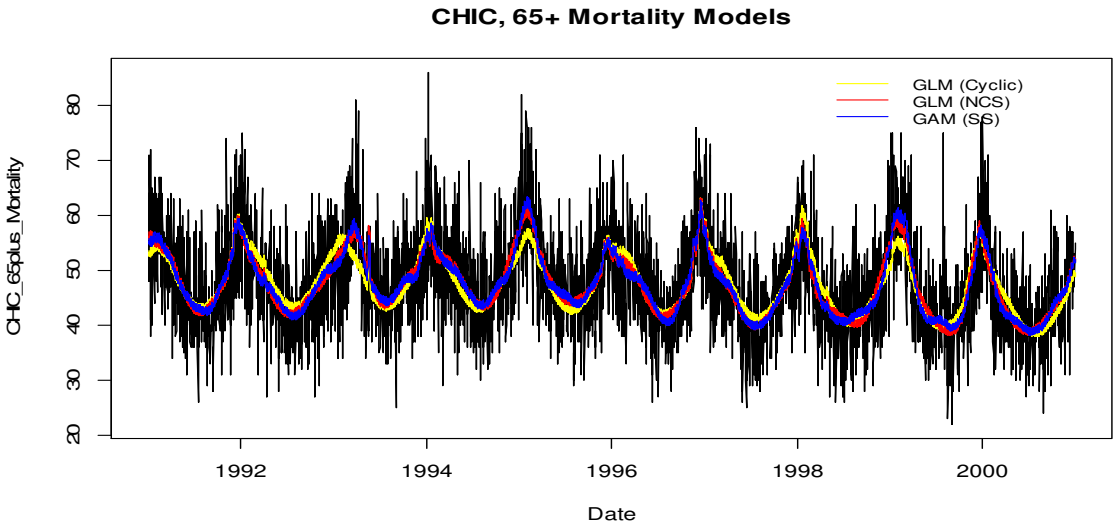

Figure 4.6: Chicago – Observed daily mortality counts for population 65+ years of age, and fitted mortality derived from Model 1: GLM-Poisson with Fourier terms for background seasonality (yellow); Model 2: GLM-NB with natural cubic spline for background seasonality (red); and Model 3: GAM-NB with smoothing spline for background seasonality (blue).

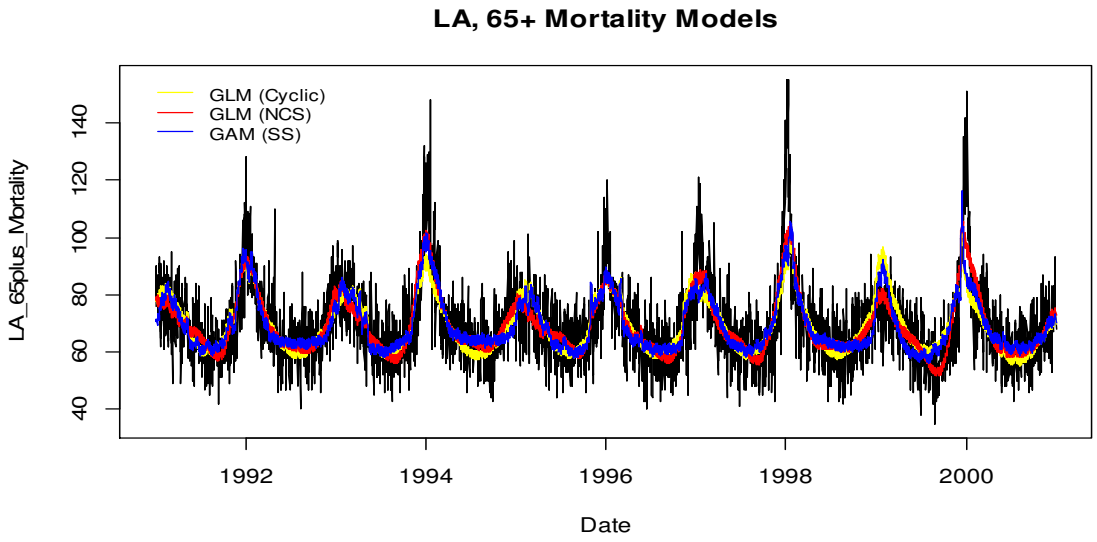

Figure 4.7: Los Angeles - Observed daily mortality counts for population 65+ years of age, and fitted mortality derived from Model 1: GLM-Poisson with Fourier terms for background seasonality (yellow); Model 2: GLM-NB with natural cubic spline for background seasonality (red); and Model 3: GAM-NB with smoothing spline for background seasonality (blue).

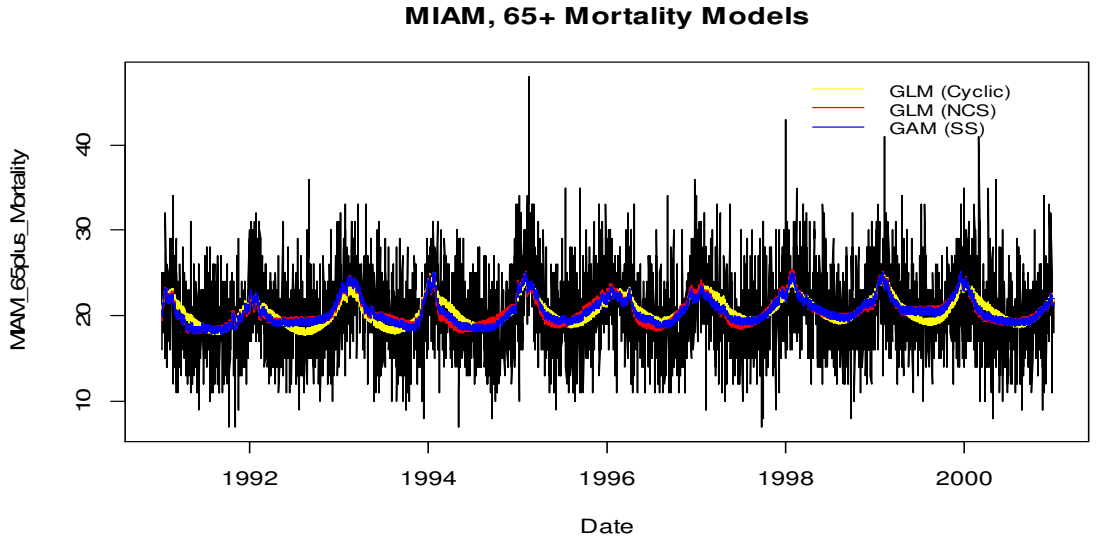

Figure 4.8: Miami - Observed daily mortality counts for population 65+ years of age, and fitted mortality derived from Model 1: GLM-Poisson with Fourier terms for background seasonality (yellow); Model 2: GLM-NB with natural cubic spline for background seasonality (red); and Model 3: GAM-NB with smoothing spline for background seasonality (blue).

Table 4.5: Number of deaths attributable to influenza type/subtype, by city and model, 1991-2000.

		Influenza-associated Deaths			
City	**Model**	**A-H3N2**	**A-H1N1**	**B**	**RSV**
NY	**GLM - Pois**	11361	512	2921	4541
	GLM - Lag3Temp	12121	591	3064	9549
	GLM - Lag3Temp (NS, df=20)	12087	543	3009	8352
	GLM - Time (NS, df=20)	12910	158	2378	7286
	GAM - Time (SS, df=20)	13483	619	2417	16778
CHICAGO	**GLM - Pois**	3239	287	463	0
	GLM - Lag3Temp	3971	189	859	4172
	GLM - Lag3Temp (NS, df=20)	3879	186	787	3692
	GLM - Time (NS, df=20)	4232	264	603	6034
	GAM - Time (SS, df=20)	5045	402	756	8794
LA	**GLM - Pois**	7336	0	162	9559
	GLM - Lag3Temp	8319	0	871	20672
	GLM - Lag3Temp (NS, df=20)	8093	0	478	19412
	GLM - Time (NS, df=20)	7714	0	2553	19949
	GAM - Time (SS, df=20)	9136	0	2653	23823
MIAMI	**GLM - Pois**	967	47	504	523
	GLM - Lag3Temp	977	1	625	806
	GLM - Lag3Temp (NS, df=20)	867	0	565	914
	GLM - Time (NS, df=20)	1519	2	830	1600
	GAM - Time (SS, df=20)	1447	43	665	2607

CHAPTER 5

MODELING THE NONLINEAR ASSOCIATION BETWEEN INFLUENZA AND MORTALITY WITH LOCAL-LEVEL ADJUSTMENT FOR SEASONAL CONFOUNDING

Quantifying the effects of influenza on mortality rates can be particularly challenging when estimates are derived from ecological studies. Because data are collected at the population rather than individual level, such studies are prone to analytic errors involving misclassification within groups, temporal ambiguity, collinearity, and inadequate control of confounding [Morganstern, 1995]. To address these issues, a wide range of mathematical modeling approaches have been developed and implemented to statistically describe the association between influenza and mortality. Some of these methods include time series regression, autoregressive integrated moving average modeling, and rate-differencing [Lui and Kendal, 1985; Thompson et al., 2003; Thompson et al., 2005; Cheng et al. (in review)].

U.S. influenza-associated death rates estimated by the Centers for Disease Control and Prevention (CDC) have generally been based on ecological study designs where mortality data are first aggregated at the national level and then modeled. There are, however, a number of seasonal factors that may confound the association between influenza and mortality at the local (e.g., city) level, thus suggesting that data be modeled locally first and then pooled to make national estimates of death. Several recent studies have shown that meteorological factors, e.g., temperature and humidity, may affect influenza-related mortality estimation [Warren-Gash et al., 2011; Yang et al., 2011; Wong et al., 2012; Yang et al., 2012]. Other seasonal viruses with varying local-level circulation, such as respiratory syncytial virus (RSV), may also affect estimates [Thompson et al., 2003; Mangtani et al., 2006]. Additional city-specific factors that may

be influential include population density, age distributions, and differences in predominant influenza viruses [Lofgren et al., 2007].

In addition to local seasonal confounding, the relationship between influenza and mortality may change over the course of a season. In most time series regression analyses, a linear relationship between influenza and log-transformed mortality is assumed [Thompson et al., 2003; Thompson et al., 2005; Cheng et al. (in review)]. In these models, a unit increase in the influenza proxy leads to a multiplicative increase in mortality. Some studies, however, have modeled influenza and mortality assuming an additive association [Goldstein et al., 2012]. In this chapter, we explore the relationship between influenza and mortality by using natural cubic spline functions on variates representing influenza circulation. Natural cubic splines model lower-order polynomials piecewise over the full range of a variable. As such, they offer greater flexibility in modeling curvature relative to global polynomials or sinusoidal terms.

The objectives of this study are two-fold: (*i*) to assess and compare rates of influenza-associated mortality across 10 U.S. cities with varying climates and population demographics, and (*ii*) to determine if the relationship between influenza and mortality is better represented nonlinearly rather than linearly. This chapter is structured as follows: Section 5.1 describes in detail the surveillance data and study population. Section 5.2 details the three modeling approaches used to estimate influenza-associated death rates. In Section 5.3, the nature of the influenza-mortality relationship is assessed. Local-level death rate estimates attributable to influenza by city and modeling method are given in Section 5.4. Conclusions based on this study are given in Section 5.5, and several limitations of the modeling approaches are given in Section 5.6.

5.1 Surveillance Data and Study Population

Data modeled in this study spanned a 10-year period from January 1, 1991 to December 31, 2000. Daily counts of mortality and daily measures of temperature and dew point

temperature were obtained for each of ten U.S. cities. Mortality data include daily respiratory and circulatory deaths among persons aged 65 or older. These data were downloaded from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) website http://www.ihapss.jhsph.edu/data/NMMAPS/R/. Details on data collection and processing methods can be found at the Internet-based Health and Air Pollution Surveillance System (iHAPSS) website http://www.ihapss.jhsph.edu.

Weekly numbers of total respiratory specimens tested for influenza and positive-influenza isolates by virus type and subtype (A(H1N1), A(H3N2), and B) were obtained from surveillance data maintained by the Influenza Division of the CDC. Weekly RSV data were obtained from the National Respiratory and Enteric Virus Surveillance System maintained by the CDC's Division of Viral Diseases. We used regional (Northeast, Midwest, South, West) proportions of respiratory specimens testing positive for influenza or RSV, referred to as 'percent-positive' viral activity, as proxies for viral activity. Local or state data were too sparse to use directly in models. Daily percent-positive data for each influenza subtype and RSV were imputed linearly from weekly percent-positive data.

U.S. population estimates by year, age group (65+), and city were obtained from the U.S. Census Bureau. Daily city-level population estimates were imputed via a step function (i.e., the annual estimate was used for each day of that calendar year).

5.2 Modeling the Nonlinearity of Influenza-Mortality Association

Three time series regression models (described below) were used to quantify and compare the association between influenza and mortality while accounting for seasonal confounding. Each model assumed a negative binomial error structure which allowed for potential under- or over-dispersion of variance relative to a Poisson distribution [Hilbe, 2008]. The negative binomial distribution was assumed to represent daily, city-level mortality because (i) local-level, daily mortality counts tended to be left-skewed (right-

tailed) particularly among the smaller cities, and (ii) the variance of daily mortality counts was not equal to the mean throughout all seasons (thus a Poisson distribution would have been inappropriate). All analyses in this study were conducted using R version 3.1.2 (www.r-project.org).

Ten U.S. cities with varying trends in mortality and weather were modeled. Each city was modeled separately. The outcome variable for all models was daily mortality due to underlying respiratory or circulatory (R&C) causes among persons aged 65 or older. Annual city-level population estimates of persons aged $\geq$ 65 years were used as the offset term to account for changes in population size over time. Regional percent-positive influenza subtypes (i.e., A(H1N1), A(H3N2), and B) and percent-positive RSV were used to represent local viral circulation. A long-term trend in mortality was accounted for in each of the models using spline functions. A day-of-week effect was also accounted for using six indicator variables (Sunday was the referent group).

Deaths associated with a particular virus were estimated as the difference between the sum of predicted values from a fitted model containing all viral terms and the sum of predicted values from the same model excluding viral activity for that particular virus [Thompson et al., 2009]. Annual, city-level influenza-associated mortality rates were calculated by dividing the average annual number of deaths attributable to a particular virus by the average annual population aged 65 and older.

Residual analyses were conducted to assess model fit using Q-Q plots, autocorrelation functions, and plotting residuals over time. Confidence intervals were calculated for the expected value of total influenza-associated mortality counts from each model as follows by: 1) obtaining fitted values of daily mortality counts from the estimated model substituting in the upper and lower 95% limits of β-estimates for the influenza terms, 2) subtracting the upper and lower fitted values from the baseline estimates, and 3) summing all upper-limit terms, and summing all lower-limit terms.

5.2.1 Modeling with Parametric Splines: A Short Review

All splines incorporated in the regression models (for time, temperature, and viral terms) are natural cubic splines. A detailed review of natural cubic splines is given in Section 4.1.2. All natural cubic splines are entered into the model linearly through their basis functions. The number of basis functions depends on the degrees of freedom (df) defining the spline. We use the 'splines' package for R to specify all spline functions. We use the following notation, taken from the 'splines' package, to represent all spline functions:

$$ns(var, k),$$

where var is the variate represented via a natural cubic spline and k is the df. For example, $ns(temp, 4)$ defines a natural cubic spline function with 4 df (i.e., 3 knots) representing the ambient temperature variate. Four basis functions (dimensions) for the temperature term would be included in the covariate matrix, i.e.,

$$ns(temp, 4) = \sum_{i=1}^{4} \beta_i f_i(temp),$$

where β is the regression parameter estimate and f is the associated spline basis function.

To represent long-term trends, knots were evenly spaced every 2 to 5 years. The number of optimal knots by city was determined by the lowest Akaike Information Criterion (AIC) value. Measured seasonal covariates included in this study were temperature and dew point temperature (a proxy for humidity). Temperature was included in the models by first lagging the series by a certain number of days (the optimal lag by city was determined by the lowest AIC value) and then using a spline function with 3 to 5 evenly spaced knots to account for the nonlinear relationship between mortality and temperature. Again the number of knots (or df) was determined by the lowest AIC value. Because dew point temperature was found to be highly collinear with ambient temperature, it was dropped from all models. For comparison, a more-

conventional approach using sinusoidal functions to model seasonal confounding was also implemented.

All viral terms were included linearly in Models 1 and 2, and via splines in Model 3 (models described below). The boundary knots for all viral terms were the lowest and highest percent-positive values determined from the combined set of all ten seasons. This approach was taken to avoid a singularity in the covariate matrix which occurred when the lower boundary point was set to zero. Since no viral circulation was assumed during summer months, having a zero lower bound for viral splines led to at least one basis function containing all zero values, thus creating a singular matrix. To avoid potential bias due to subjective judgement, knots were evenly spaced by quantiles within these defined boundaries. To avoid overfitting, no more than 4 df (i.e., 3 knots) were considered for each viral term. Thus, three knots divided the viral terms into quartiles (again, excluding the zero values).

The statistical significance of a particular viral spline was assessed via a likelihood ratio test comparing models with and without the associated viral spline basis functions. If spline functions were not significant, the spline was remodeled with one less knot. If a viral spline containing only 1 knot was not significant in the model, a linear function was used to represent the viral term. If the linear function was not significant, the linear viral term remained in the model based on the *a priori* decision that all three influenza terms be modeled together for overall seasonal adjustment. Death rates, however, were not calculated for any non-significant viral terms.

5.2.2 Model 1: Sinusoidal Seasonality and Linear Viral Terms

Generalized linear models with sinusoidal terms used to control for seasonal confounding are widely used to model influenza-associated morbidity and mortality [Thompson et al., 2003; Thompson et al., 2009; Warren-Gash et al., 2011; Newall et al., 2010; Liao et al.,

2009]. Since it is considered a conventional approach, we use this as our referent model for comparison. The full model is described as follows:

$$\log(Y_i) = \log(c_i) + \beta_0 + \sum_{j=1}^{6} \beta_j[I_{ij}] + ns(t,k) + \beta_7[\mathrm{A(H1N1)}_i] + \beta_8[\mathrm{A(H3N2)}_i] + \beta_9[\mathrm{B}_i] + \beta_{10}[\mathrm{RSV}_i] + \beta_{11}[\sin(2\pi t_i/365.25)] + \beta_{12}[\cos(2\pi t_i/365.25)]$$

where $Y_i \sim$ NegBin, represents the number of deaths on day i; c_i is the offset term equal to the annual population size; β_0 is the model intercept; β_1 through β_6 are coefficients for day-of-week indicators, I_{ij} (Sunday = referent); $ns(t,k)$ is a natural cubic spline with $df = k \in \{2,3,4,5\}$ representing long-term time trends; β_9 and β_{10} are coefficients for the percent-positive viral terms; and β_{11} and β_{12} are coefficients for the seasonal trend.

5.2.3 Model 2: Temperature Splines and Linear Viral Terms

Instead of using a mathematical proxy for seasonal confounding (e.g., a cosine function), Model 2 represents local seasonal confounding via ambient temperature, the assumption being that a measureable proxy for seasonal confounding may be more valid than an unmeasured, mathematical proxy. Models 1 and 2 allow a direct comparison of these two approaches used for modeling confounding seasonality. The general model can now be described as follows:

$$\log(Y_i) = \log(c_i) + \beta_0 + \sum_{j=1}^{6} \beta_j[I_{ij}] + ns(t,k) + \beta_7[\mathrm{A(H1N1)}_i] + \beta_8[\mathrm{A(H3N2)}_i] + \beta_9[\mathrm{B}_i] + \beta_{10}[\mathrm{RSV}_i] + ns(\mathrm{Lag}(Temp_i, l), k),$$

where $ns(\mathrm{Lag}(Temp_i, l), k)$ is a natural cubic spline with $df = k \in \{3,4,5\}$ representing the average daily temperature lagged $l \in \{3,4,\ldots,9\}$ number of days. All other terms in this model are defined in Model 1 (Section 5.2.2).

5.2.4 Model 3: Temperature and Viral Splines

To determine if influenza has a nonlinear association with mortality, Model 3 enters all influenza terms, as well as the RSV term, into the model via natural cubic spline functions. In every other regard, Model 3 is identical to Model 2. This allows for a direct comparison of results by modeling viral terms linearly versus nonlinearly. To avoid overfitting, only a few knots were used to specify viral splines ($df \in \{2,3,4\}$). Viral splines were optimized separately by city. The following model was implemented:

$$\log(Y_i) = \log(c_i) + \beta_0 + \sum_{j=1}^{6} \beta_j [I_{ij}] + ns(t,k) + ns(\text{Lag}(Temp_i, l), k) + ns(\text{A(H1N1)}_i, k) + ns(\text{A(H3N2)}_i, k) + ns(\text{B}_i, k) + ns(\text{RSV}_i, k),$$

where $ns(\text{A(H1N1)}_i, k)$ is a spline with $df = k \in \{2,3,4\}$ for the percent-positive A(H1N1) term; $ns(\text{A(H3N2)}_i, k)$ is a spline with $df = k \in \{2,3,4\}$ for the percent-positive A(H3N2) term; $ns(\text{B}_i, k)$ is a spline with $df = k \in \{2,3,4\}$ for the percent-positive B term; and $ns(\text{RSV}_i, k)$ is a spline with $df = k \in \{2,3,4\}$ for the percent-positive RSV term. All other terms in this model are defined in Models 1 and 2 (Sections 5.2.2 and 5.2.3).

5.3 Results: Linear versus Nonlinear Model Fit

Table 5.1 shows the ten U.S. cities modeled in this study along with their high, low, and average temperatures on January 1 and July 1 from the ten-year study period (1991−2000). Also given for each city are regional grouping and average population during this period for persons aged 65 or older. Region specifies which percent-positive regional viral series (Figure 2.3) was used to approximate each city's viral activity. The city with the highest population of persons 65+ years of age was New York, and that with the lowest population was Denver. Measured on January 1, the city with the coldest

average temperature was Minneapolis and the warmest was Miami. Measured on July 1, the hottest average temperature occurred in Phoenix, and the coolest in Seattle.

Figure 3.2 displays daily time series of R&C deaths, temperature, dew point temperature, and regional A(H3N2) circulation for four U.S. cities. Across all cities, temperature and dew point temperature were found to be highly collinear (correlation coefficient ranging from $r = 0.61$ in Los Angeles to $r = 0.95$ in New York). Thus, the two terms were not modeled together. The temperature term tended to be more-highly associated with mortality; therefore, dew point temperature was dropped from all models. We also modeled an adjusted dew point temperature by first regressing dew point temperature on temperature, and then modeling the residuals. This term, however, did not substantially affect any of the viral parameter estimates and therefore was dropped from all models.

Figure 3.3 shows daily time series of the percent-positive viral activity proxies for the three influenza subtypes (A(H1N1), A(H3N2), and B) and RSV for four regions. A(H3N2) was the predominant influenza strain during this ten-year period. Very little A(H1N1) circulation was observed, and type B influenza was not as prevalent as A(H3N2).

Figures 5.1−5.4 depict the univariate association between mortality and each of the four viral proxies for four U.S. cities. The green line is the fitted linear function of log-transformed mortality regressed on each viral term, while the red line is the fitted spline function. To avoid overfitting, it was determined that all influenza and RSV splines should have no more than 2 df, i.e., one or zero knots (0 knots collapses the spline to a linear function). For all modeled cities, A(H3N2) influenza was found to have a nonlinear relationship with mortality. Compared to the spline fit, the linear fit tended to underestimate the association during lower viral circulation (e.g., early outbreak) and overestimate during peak influenza season. The influenza splines revealed a tapering-off

effect toward peak season. RSV on the other hand tended to have a more clearly defined linear, or near-linear, relationship with mortality in all ten cities.

5.4 Results: Deaths Attributable to Influenza

Tables 5.2 and 5.3 show the estimated number of deaths attributable to influenza (by type/subtype and overall) and RSV during the ten-year study period by city and model. The df used for each viral spline in Model 3 are also included. Values in red are considered non-significant as they are based on model parameter estimates that were not found to be significant at the $\alpha = 0.05$ level. Values in black were based on significant parameter estimates ($\alpha < 0.05$), and thus are considered statistically significant death estimates. Tables 5.4 and 5.5 give respective population rates of death attributable by viral type/subtype for each city and model. Rates are not estimated for attributable death counts not found to be statistically significant in Tables 5.2 and 5.3.

5.4.1 Attributable Deaths by Influenza Type/Subtype

For all ten cities, mortality was significantly associated with A(H3N2) (Table 5.2). Comparing Models 1 and 2, which varied only by seasonal proxies (Model 1 used sinusoidal terms; Model 2 used daily average temperature terms), more deaths were attributable to A(H3N2) with Model 2. Comparing Models 2 and 3, which varied only by viral representation in models (linearly versus nonlinearly), five cities had higher death estimates based on Model 2, while the other five cities had higher death estimates based on Model 3.

Based on results from all three models, only three of the ten cities had statistically significant A(H1N1) attributable death estimates (New York, Chicago, Minneapolis/St.Paul). Only two of these (New York and Chicago) revealed a nonlinear association between A(H1N1) and mortality. For most cities, robust statistically

significant estimates for A(H1N1) attributable deaths could not be made – likely due to the fact that there was very little A(H1N1) influenza activity during the study period.

Type B influenza circulation during the study period was less prevalent compared to A(H3N2). Fewer deaths were attributable to B relative to A(H3N2). The two smallest cities (Denver and Minneapolis) did not have significant numbers of deaths attributable to type B, likely due to small daily mortality counts coupled with low percent-positive B activity in most seasons. It should also be noted that for most cities, as expected [CDC, 2008; CDC, 2013], the type B influenza peak occurred several weeks after the 65+ mortality peak for most seasons (Figure 3.3). Of the eight cities with significant B deaths, seven revealed a significant nonlinear relationship between influenza and mortality. Type B spline terms were not significant in Los Angeles; however, the type B term entered linearly was significant.

Rates of deaths attributable to influenza sub/types and RSV are given in Table 5.4. Only rates based on statistically significant counts from Table 5.2 are given. For A(H3N2), rates varied greatly by city. New York and Los Angeles, the two cities with the largest 65+ populations, had the highest rates of mortality attributable to A(H3N2). A(H3N2) rates of death tended to be highest in cities with large populations and colder January temperatures. A(H1N1) was found significant, by all three models, in the three coldest cities (New York, Chicago, Minneapolis) regardless of population size. Rates for B were not as widely spread by city compared to A(H3N2) rates. They also did not appear to vary consistently by population size or average January temperature.

5.4.2 Deaths Attributable to RSV

Respiratory syncytial virus was included in all models as a confounder to avoid attributing RSV-associated deaths to influenza. Though the focus of this study is on influenza and mortality, the high correlation between RSV and seasonal confounding bears mentioning. Despite high RSV activity during the ten-year study period, only two

cities revealed statistically significant deaths attributable to RSV. For eight of the ten cities, RSV estimates based on Model 1 were either non-significant or negative. An assessment of the correlation between RSV and the sinusoidal terms found extremely high collinearity between RSV and the cosine function. This correlation ranged from $r = 0.65$ to $r = 0.95$. Also noted was the high correlation between the temperature terms and RSV (ranging from $r = 0.72$ to $r = 0.96$). For the cities in which the temperature terms were not as highly correlated with RSV, rates for attributable deaths were calculated (Table 5.4); and like A(H3N2), results revealed a wide range of attributable deaths by city. Unlike A(H3N2), RSV was better modeled linearly than quadratically.

5.4.3 Deaths Attributable to Influenza Modeled as a Singe Covariate

A(H1N1)-associated deaths were not statistically significant for most cities and models, and deaths attributable to type B were only significant in three cities. Because of this, all percent-positive influenza sub/types were collapsed into one variable containing all influenza. Results from the 'All Influenza' model (Table 5.3) revealed influenza attributable death totals near those of the combined totals from the separate 'type/subtype' models. Rates of total influenza deaths again varied greatly by city (Table 5.5). A general trend revealed more influenza-attributable deaths in larger and colder cities.

Despite the more-parsimonious model, deaths attributable to RSV were still not found to be statistically significant in most cities. This was not surprising given the high collinearity between RSV and the seasonal confounding terms, specifically, the cosine function and the lagged temperature terms. RSV death rates varied greatly by city with no recognizable pattern (Table 5.5).

5.5 Conclusions

5.5.1 Local-Level Influenza-Associated Death Rates

We found that influenza-associated death rates vary significantly by city. A(H3N2) death rates tended to be higher among larger and colder cities. The same pattern was found with pooled influenza death rates. These findings suggest that modeling influenza-attributable deaths should be conducted at the local level. Estimates by city could then be pooled for a national estimate that is less likely to be biased by confounding due to temporal misalignment.

Due to low A(H1N1) activity measured during the study period, death estimates attributable to A(H1N1) were not possible for most cities. Type B influenza activity was also low, and when coupled with small population estimates, death rates were not significant. There was no apparent pattern between B-attributable deaths and climate or population density.

5.5.2 Modeling Influenza-attributable Mortality

A significant finding from this study is the apparent nonlinear association between influenza and mortality. Given log-transformed mortality, we expected to find a nonlinear association between influenza and mortality that increased multiplicatively. Results based on linear viral terms (green fitted lines from Figures 5.1−5.4) showed no multiplicative increase, and instead showed only a linear association (despite the logged mortality term). The spline functions found that the association is indeed nonlinear; however, instead of mortality increasing multiplicatively with increasing influenza activity, it tends to taper off toward peak influenza season. Figures 5.1−5.4 show that relative to the spline fit, the linear function consistently underestimated the association during off-peak periods and overestimated during peak periods. Based on the observed shape of the association, either spline functions, quadratic terms, or logged viral terms

might offer a better fit for the influenza viral terms. These modeling approaches should be further explored and compared to the conventional, linear representation of influenza.

Within each city, comparing results from the three models revealed that Model 1 gave consistently lower attributable-death estimates. The effect of aligning Fourier terms with the mortality peaks tended to dampen the association between A(H3N2) and mortality. Using temperature as a proxy for seasonality may offer greater reliability since it is based on measured data rather than mathematical terms chosen for optimal fit. Although differences in death estimates between Models 1 and 2 appear relatively small at the city level, the cumulative estimate of attributable deaths summed across many cities may reveal a sizeable overall difference in rates between the two models.

5.5.3 Modeling RSV-attributable Mortality

Two notable findings were obtained in modeling RSV and mortality. First, RSV tended to be better modeled linearly. Unlike the influenza terms, a tapering-off effect was not observed. Second, the association between RSV and the seasonal background terms was found to be highly collinear. When modeled with the cosine function in particular, the RSV signal was either lost or significantly dampened. Lagging the temperature terms to better fit the mortality data inadvertently increased the correlation between RSV and temperature. Because RSV was found to be so highly collinear with seasonal confounders, robust statistical estimates for RSV attributable deaths were incalculable by most models. The association between RSV, mortality, and the variates used to control for seasonal confounding should be investigated further.

5.6 Limitations

First, as noted earlier, results from ecological studies should be interpreted cautiously since such studies are susceptible to design and modeling errors stemming from data collected and aggregated at population levels. Second, residuals from regression models

still have slight periodicity in some models. As such, confidence intervals may need to be bootstrapped rather than estimated parametrically. Third, each of these models needs validating through simulation. Plausible simulated data sets are needed to better compare and contrast tested models.

Table 5.1: Temperature statistics and average population aged 65 or older for 10 U.S. cities, 01/01/1991 – 12/31/2000

City	Region	Temperature (F°)						Average Population Age 65+
		January 1			July 1			
		Average	High	Low	Average	High	Low	
Chicago	Midwest	14	42	28	66	81	73	642,233
Minn / St. Paul	Midwest	-7	32	19	59	78	71	182,230
Philadelphia	Northeast	24	46	34	73	82	77	222,063
New York	Northeast	20	46	33	72	79	74	1,068,846
Miami	South	59	78	72	80	88	84	292,764
Dallas / Ft. W	South	33	63	47	77	92	85	303,250
Denver	West	18	50	35	58	80	71	125,458
Seattle	West	30	51	43	57	72	63	177,696
Phoenix	West	52	64	57	88	100	94	320,656
Los Angeles	West	52	64	58	66	75	70	925,910

Table 5.2: Estimated influenza and RSV associated death counts by city, viral type, and model, 1991−2000 cumulative total. Degrees of freedom for each viral spline in Model 3 are also given.

City	Model	A(H3N2)	df	A(H1N1)	df	B	df	RSV	df
Chicago									
	1	3282		252		617		0	
	2	3740		250		739		2523	
	3	4693	2	262	1	1260	2	3560	2
Dallas/Ft.Worth									
	1	1966		216		436		0	
	2	2534		281		746		0	
	3	2293	2	246	1	1147	2	375	1
Denver									
	1	504		0		0		474	
	2	603		0		134		1750	
	3	603	2	0	1	142	1	1574	2
Los Angeles									
	1	7738		0		319		7994	
	2	8715		0		779		15540	
	3	8148	2	0	1	1085	1	9235	2
Miami									
	1	959		42		442		705	
	2	1006		6		583		1408	
	3	1353	2	0	1	725	2	1321	1
Minneapolis									
	1	884		133		118		0	
	2	983		130		119		0	
	3	1032	2	158	2	127	1	0	1
New York									
	1	11736		569		2863		1172	
	2	12285		651		2622		7251	
	3	10878	2	986	2	3213	2	5359	2
Philadelphia									
	1	1532		39		295		0	
	2	1587		0		469		810	
	3	1487	2	13	1	565	2	908	1
Phoenix									
	1	1897		0		423		2453	
	2	1854		0		678		4851	
	3	2143	2	0	1	1207	2	1709	2
Seattle									
	1	796		0		444		0	
	2	807		0		615		749	
	3	986	2	0	1	771	2	548	1

Table 5.3: Estimated influenza and RSV associated death counts by city and model, 1991−2000 cumulative total. Degrees of freedom for each viral spline in Model 3 are also given.

City	Model	Influenza*	df	RSV	df
Chicago					
	1	5078		0	
	2	5778		1679	
	3	6168	2	1379	1
Dallas/Ft.Worth					
	1	2608		0	
	2	3406		0	
	3	3692	2	1235	2
Denver					
	1	504		295	
	2	712		1510	
	3	700	2	1555	2
Los Angeles					
	1	9407		8210	
	2	10994		14816	
	3	9879	2	8637	2
Miami					
	1	1476		313	
	2	1748		1230	
	3	2200	2	1019	1
Minneapolis					
	1	1262		0	
	2	1386		0	
	3	1368	2	0	1
New York					
	1	15722		1582	
	2	16224		7867	
	3	12055	2	7621	2
Philadelphia					
	1	2046		292	
	2	2174		935	
	3	1824	2	1147	1
Phoenix					
	1	2396		1943	
	2	2423		4238	
	3	2715	2	2090	2
Seattle					
	1	1118		0	
	2	1202		291	
	3	1342	2	201	1

* Influenza includes all A(H1N1), A(H3N2), and B types/subtypes

Table 5.4: Estimated average annual rates* of influenza and RSV attributable deaths by city, viral type, and model, 1991−2000.

City	Model	A(H3N2)	A(H1N1)	B	RSV
Chicago					
	1	51.10	3.92	9.61	ns**
	2	58.23	3.89	11.51	39.28
	3	73.07	4.08	19.62	55.43
Dallas/Ft.Worth					
	1	64.83	7.12	14.38	ns
	2	83.56	ns	24.60	ns
	3	75.61	ns	37.82	ns
Denver					
	1	40.17	ns	ns	ns
	2	48.06	ns	ns	139.49
	3	48.06	ns	ns	125.46
Los Angeles					
	1	83.57	ns	ns	86.34
	2	94.12	ns	8.41	167.83
	3	88.00	ns	11.72	99.74
Miami					
	1	32.76	ns	15.10	ns
	2	34.36	ns	19.91	48.09
	3	46.21	ns	24.76	45.12
Minneapolis					
	1	48.51	7.30	ns	ns
	2	53.94	7.13	ns	ns
	3	56.63	8.67	ns	ns
New York					
	1	109.80	5.32	26.79	ns
	2	114.94	6.09	24.53	67.84
	3	101.77	9.22	30.06	50.14
Philadelphia					
	1	68.99	ns	13.28	ns
	2	71.47	ns	21.12	ns
	3	66.96	ns	25.44	ns
Phoenix					
	1	59.16	ns	13.19	76.50
	2	57.82	ns	21.14	151.28
	3	66.83	ns	37.64	53.30
Seattle					
	1	44.80	ns	24.99	ns
	2	45.41	ns	34.61	ns
	3	55.49	ns	43.39	ns

* Attributable deaths rates are given only for statistically significant mortality counts listed in Table 5.2.

** ns = not statistically significant

Table 5.5: Estimated average annual rates* of influenza and RSV attributable deaths by city and model, 1991−2000.

City	Model	All Flu	RSV*
Chicago			
	1	79.07	ns**
	2	89.97	26.14
	3	96.04	ns
Dallas/Ft.Worth			
	1	86.00	ns
	2	112.32	ns
	3	121.75	40.73
Denver			
	1	40.17	ns
	2	56.75	120.36
	3	55.80	123.95
Los Angeles			
	1	101.60	88.67
	2	118.74	160.02
	3	106.70	93.28
Miami			
	1	50.42	ns
	2	59.71	42.01
	3	75.15	34.81
Minneapolis			
	1	69.25	ns
	2	76.06	ns
	3	75.07	ns
New York			
	1	147.09	ns
	2	151.79	73.60
	3	112.79	71.30
Philadelphia			
	1	92.14	ns
	2	97.90	ns
	3	82.14	ns
Phoenix			
	1	74.72	60.59
	2	75.56	132.17
	3	84.67	65.18
Seattle			
	1	62.92	ns
	2	67.64	ns
	3	75.52	ns

* For RSV, attributable death rates are given only for statistically significant RSV-attributable death counts from Table 5.3.

** ns = not statistically significant

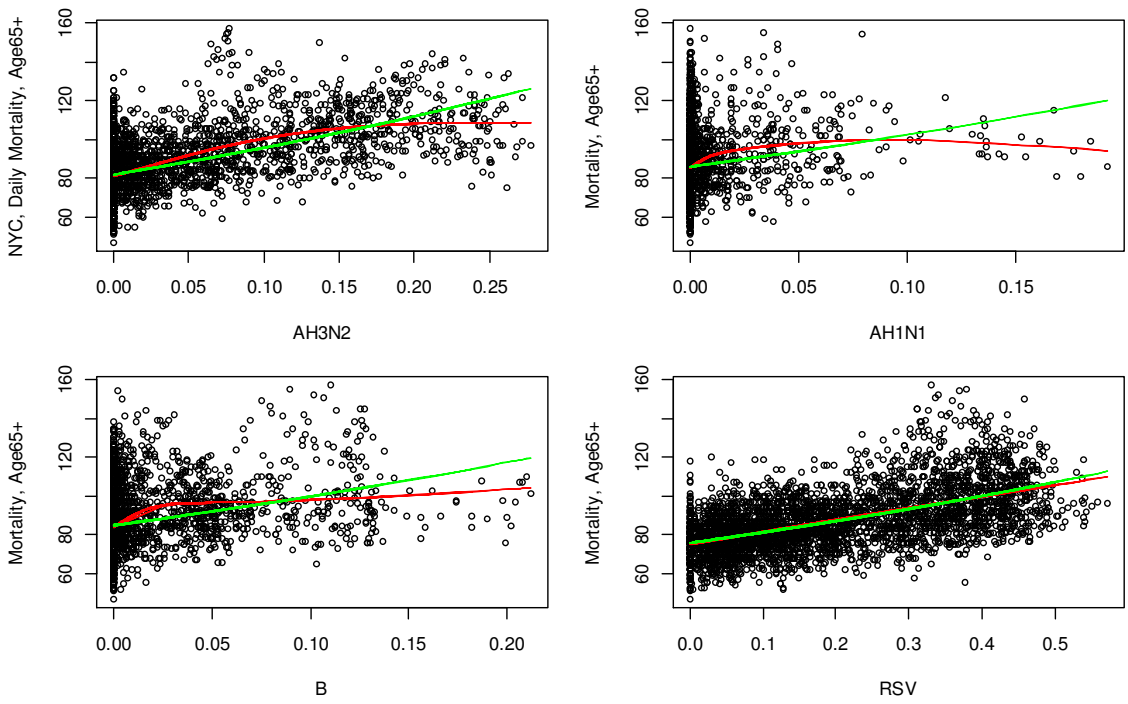

Figure 5.1: Association between daily mortality count for persons age 65+ years and percent-positive viral type, New York City, combined years 1991−2000. Association modeled using spline function (red) and linear function (green).

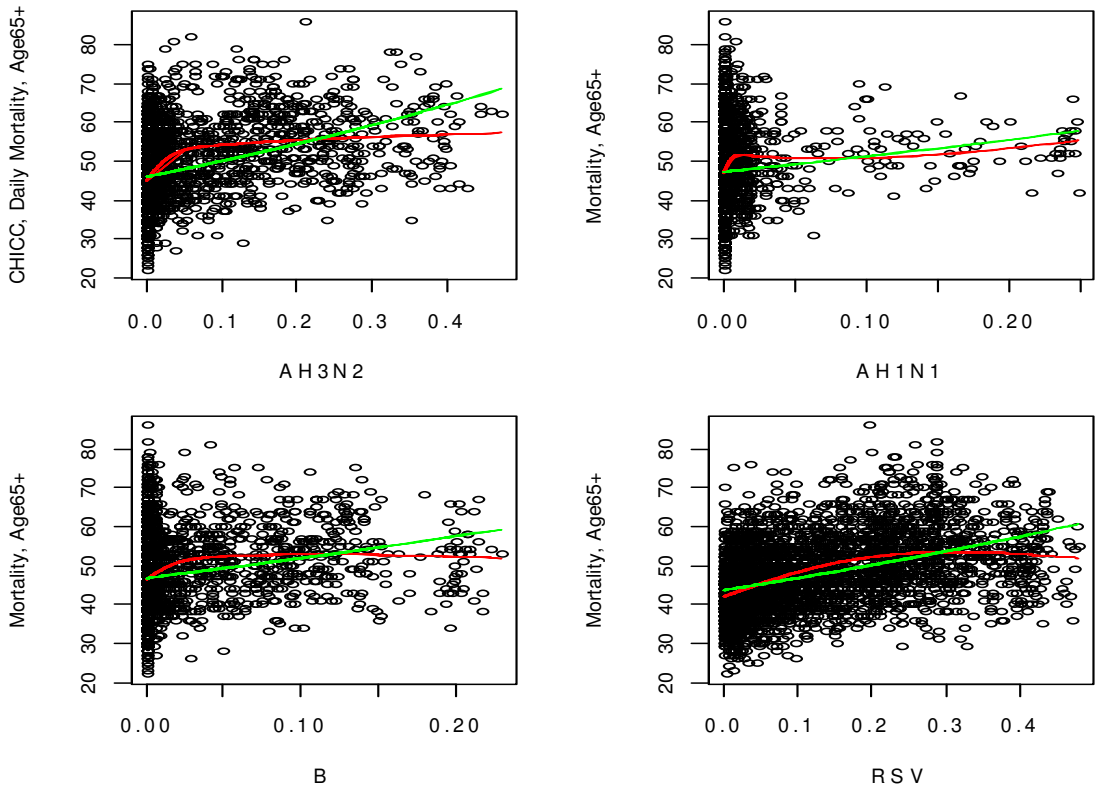

Figure 5.2: Association between daily mortality count for persons age 65+ years and percent-positive viral type, Chicago, combined years 1991−2000. Association modeled using spline function (red) and linear function (green).

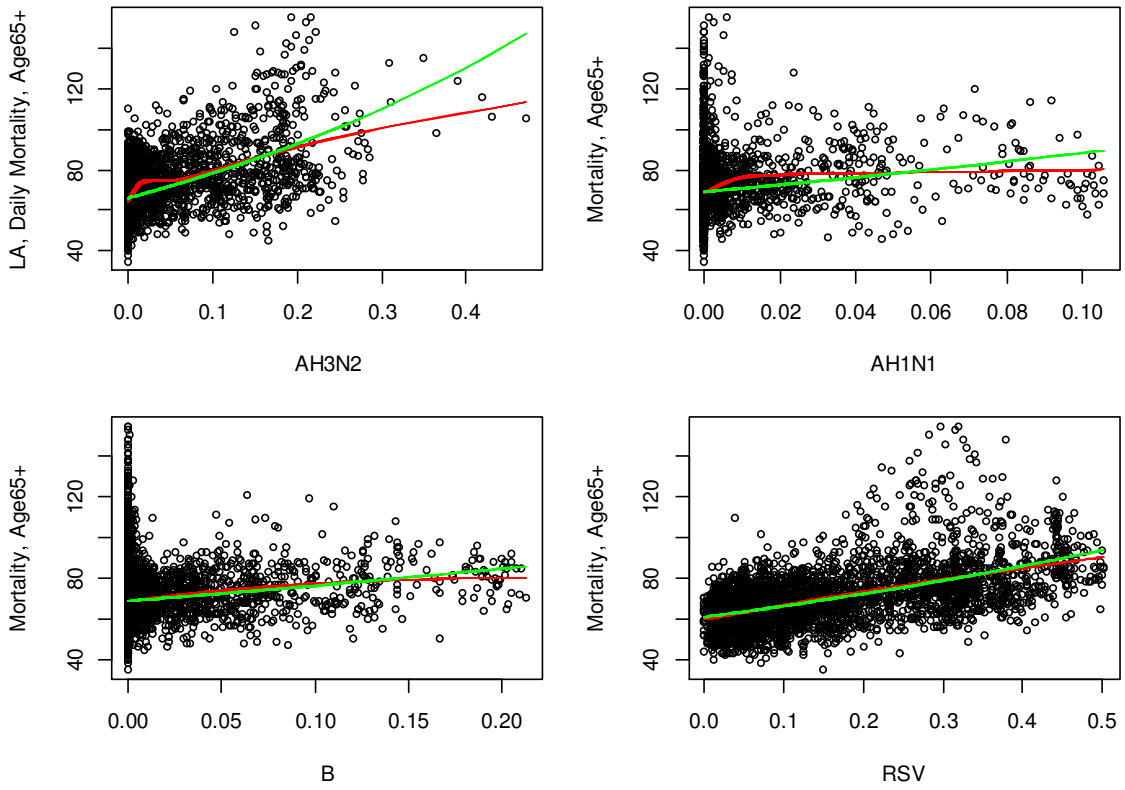

Figure 5.3: Association between daily mortality count for persons age 65+ years and percent-positive viral type, Los Angeles, combined years 1991−2000. Association modeled using spline function (red) and linear function (green).

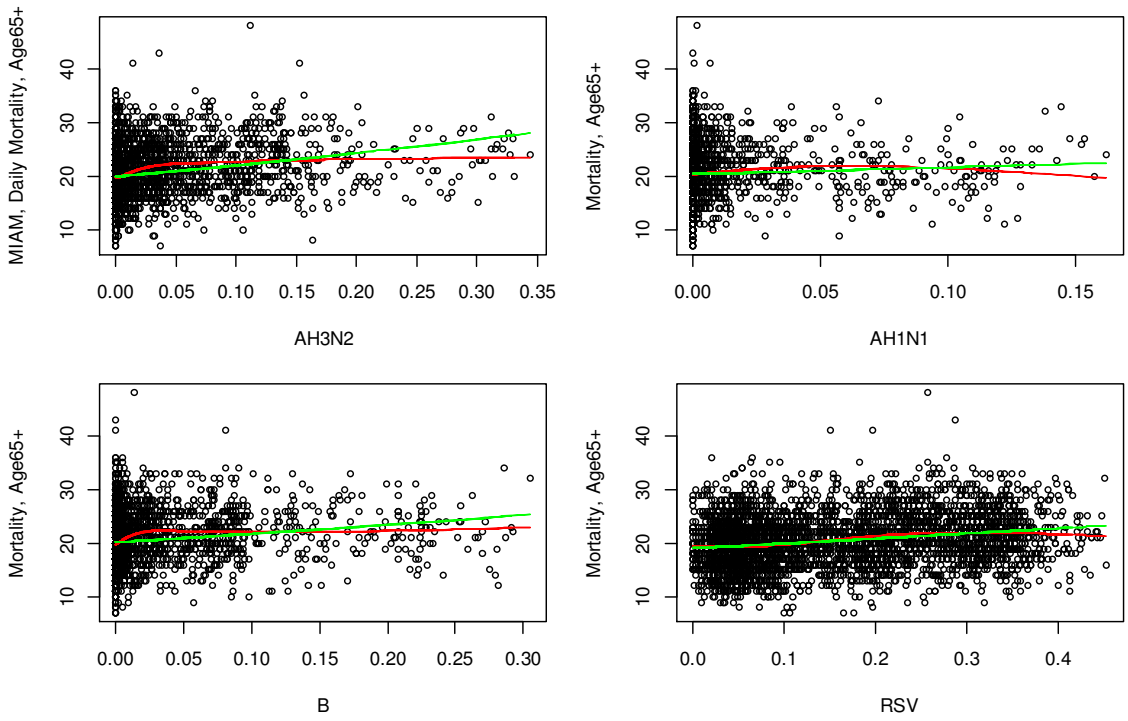

Figure 5.4: Association between daily mortality count for persons age 65+ years and percent-positive viral type, Miami, combined years 1991−2000. Association modeled using spline function (red) and linear function (green).

CHAPTER 6

LITERATURE REVIEW: STATISTICAL PROCESS CONTROL CHARTS FOR OUTBREAK DETECTION IN SYNDROMIC SURVEILLANCE

Statistical process control (SPC) charts, designed for quality control monitoring in industry, have been widely adapted for use in disease and syndromic surveillance. Adaptations of the Cumulative Sum (CuSum) and Exponentially Weighted Moving Average (EWMA) charts have been used to monitor counts of nosocomial infections [Benneyan, 1998; Brown et al., 2002], hospital emergency department visits [Burkom, 2003; Yuan et al., 2004; Ivanov et al., 2003], visits to medical facilities [Burkom, 2003; Yuan et al., 2004; Ivanov et al., 2003; Bradley et al., 2005], prescription drug sales [Chen et al., 2005], and sales of over-the-counter health care products [Burkom, 2003; Hogan et al., 2003; Marx et al., 2006]. Tsui et al. [2008] give a detailed review of popular SPC methods and performance measures used in public health and syndromic surveillance. When used in syndromic surveillance, a statistically significant increase in observed data demonstrated by an SPC chart might be considered evidence of an emerging outbreak.

This chapter reviews the current literature on three SPC charts (CuSum, EWMA, and Shewart residual charts) used for detection of rate or count increases in public health surveillance.

6.1 Statistical Process Control Charts

The typical control chart is a graphical representation of information collected from a monitored process over time. It displays sample measurements of a quality characteristic, either a variable (quantitative, e.g., length of product) or an attribute (qualitative, e.g.,

defective/conforming). The purpose of such a chart is to allow for the detection of an unusual occurrence that may reflect an actual process change.

The control chart is essentially a set of sequential tests of the hypothesis that the process is in a state of statistical control. If a point plots within the control limits, the hypothesis would not be rejected. If the point plots beyond the control limits, the statistical control hypothesis would be rejected. Thus, specifying the control limits is a critical decision. By widening the control limits, the risk of a Type I error is decreased, i.e., the risk of a point falling outside the limits when the process is still in control (or no assignable cause is present) is decreased. However, the risk of a Type II error, i.e., a point falling within the limits when the process is actually out of control, is increased. The opposite effect occurs when the control limits are brought closer to the center line. In that case, the Type I risk decreases while the Type II risk increases.

The average run length (ARL) is the average number of points plotted on the chart before a point indicates an out-of-control condition. For uncorrelated process observations, the ARL for any Shewhart chart can be calculated as ARL = 1/*p,* where *p* is the probability that any point falls beyond the control limits [Montgomery, 2005]. This method is often used to evaluate the performance of the control chart. For an in-control process, the ARL (onward denoted ARL_0) represents Type I error, or a "false alarm." As such, this value should be large. For an out-of-control process, on the other hand, the ARL (onward denoted ARL_1) should be very small for timely detection. In general, ARL_0 represents the control chart's reliability while ARL_1 measures how sensitive the chart is to process excursions.

Numerous types of control charts have been developed to monitor various quality characteristics. Three popular variable charts used widely in business and industry include: Shewhart charts, CuSum charts, and EWMA charts. CuSum and EWMA charts utilize information collected from prior observations and therefore are able to detect smaller shift changes more efficiently than Shewhart charts.

6.2 Cumulative Sum Charts in Public Health Surveillance

Let μ_0 be the target value for the process mean, and $\bar{x}_j$ be the average of the j^{th} sample. Then $(\bar{x}_j - \mu_0)$ is the deviation of the j^{th} sample mean from the target value. The CuSum chart plots the following:

$$C_i = \sum_{j=1}^{i} (\bar{x}_j - \mu_0).$$

The quantity C_i is the cumulative sum of deviations up to and including the i^{th} sample. Because they combine information from all prior samples, CuSum charts are more effective than Shewhart charts for detecting small process shifts.

If the process is in control at the target value μ_0, the cumulative sum is a random walk with mean zero. If the process mean shifts upward, however, then an upward (or positive) drift develops and is depicted on the chart. Similarly, if the mean shifts downward, then a negative drift develops. If such trends are depicted on the CuSum chart, the process is considered out-of-control and causes for variation should be determined.

The theoretical properties of CuSum charts have been widely investigated [e.g., Page, 1954; Shiryayev, 1963; Lorden, 1971; Pollak, 1985; Lai, 2001]. Lucas [1985] gives a detailed examination of the run length of Poisson CuSum charts. Brook et al. [1972] utilized the CuSum chart for detection of a shift in mean rates given a Poisson error structure. White et al. [1996] approximated the target threshold for an in-control average run length (i.e., ARL_0) using a Markov chain algorithm. In biosurveillance research, Lee et al. [2014] examined the performance of analytically approximated control limits in multivariate CuSum charts used to detect emerging disease clusters. Hill et al. [1968] and Weatherall et al. [1976] utilized CuSum charts for the surveillance of congenital abnormalities. Cowling et al. [2006] monitored influenza sentinel surveillance data using an upper CuSum chart that incorporated a seven-week buffer period designed

to avoid inadvertently averaging out a gradual increase in number of cases. Jackson et al. [2007] compared the performance of an EWMA chart, a Shewart chart, and a generalized linear modeling method using daily count data of emergency department visits related to respiratory illnesses. Joner et al. [2008] and Fricker [2008] introduced modifications to traditional CuSum charts to account for the spatiotemporal characteristics of biosurveillance data. Woodall [2006] and Tsui et al. [2008] reviewed applications of CuSum and other SPC methods in health surveillance.

6.3 Exponentially Weighted Moving Average Charts in Public Health Surveillance

Like the CuSum chart, the EWMA control chart is highly effective in detecting small shifts in the process mean. The exponentially weighted moving average is defined as follows:

$$z_i = \lambda x_i + (1-\lambda) z_{i-1},$$

where $0 < \lambda < 1$ is a constant. The starting value is generally either the process target $z_0 = \mu_0$, or the average of preliminary data ($z_0 = \bar{x}$). The EWMA z_i is a weighted average of all prior sample means.

$$\begin{aligned} z_i &= \lambda x_i + (1-\lambda) z_{i-1} \\ &= \lambda x_i + (1-\lambda)[\lambda x_{i-1} + (1-\lambda) z_{i-2}] \\ &= \lambda x_i + \lambda(1-\lambda) x_{i-1} + (1-\lambda)^2 z_{i-2} \end{aligned}$$

Substituting recursively for $z_{i-j}, j = 2,3, \ldots t$ gives

$$z_i = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j x_{i-j} + (1-\lambda)^i z_o.$$

The weights $\lambda(1-\lambda)^j$ decrease geometrically and sum to unity since

$$\lambda \sum_{j=0}^{i-1} (1-\lambda)^j = \lambda \left[\frac{1-(1-\lambda)^i}{1-(1-\lambda)} \right] = 1-(1-\lambda)^i.$$

If the process observations $x_i, i = 1,2, \ldots, n$ are independent random variables with variance σ^2, then the variance of z_i is

$$\sigma_{z_i}^2 = \sigma^2 \left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right].$$

The EWMA control chart is then constructed as follows:

$$\text{UCL} = \mu_0 + L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]}$$

$$\text{CL} = \mu_0$$

$$\text{LCL} = \mu_0 - L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]}$$

where $L\sigma$ is the distance from the center line to the control limit and UCL, CL, and LCL are the upper confidence limit, center line, and lower confidence limit, respectively. The z_i values are plotted on this chart and any deviations beyond the control limits indicate an out-of-control process.

Two parameters of the EWMA chart are the multiple of sigma used in the control limits (L) and the value of λ in the EWMA equation. These parameters can be chosen so that the ARL performance of the EWMA control chart closely approximates the CuSum ARL performance for detecting small shifts.

The optimal design procedure using the EWMA chart starts by specifying the desired ARL_0 and ARL_1 values as well as the anticipated magnitude of the process shift. The combination of λ and L should then be selected based on these specifications. Typically, smaller values of λ are used to detect smaller shifts. Smaller λ values place more emphasis on prior information.

The theoretical properties of two-sided EWMA charts have been widely investigated [Hunter, 1986; Crowder, 1987; Crowder, 1989; Ng and Case, 1989; Lucas and Saccucci, 1990]. Performance of the two-sided EWMA is limited when λ is small. If a dramatic shift in the mean occurs toward the opposite side of the center line from the

current EWMA value, it could take several periods before the shift is detected. With smaller λ, the newer data would not be as heavily weighted. A comparison of one- and two-sided EWMA charts can be found in Shu et al. [2007]. One-sided EWMA control limits were also examined in Robinson et al. [1978]. Borror et al. [1992, 1998] approximated the in-control ARL of a Poisson EWMA chart using Markov chain simulation methods. Gan [1990] monitored the mean shift of a Poisson process using a modified EWMA chart. Joner et al. [2008] introduced a one-sided multivariate EWMA (MEWMA) to monitor Poisson counts via a T^2 statistic.

Many studies have compared the performance of CuSum and EWMA chart methods under continuous distributions. Srivastava and Wu [1993] have shown that under stationary conditions, the EWMA chart is less efficient than the CuSum chart. They also showed that the performance of a two-sided EWMA detection method is comparable to that of a CuSum [1997]. Lucas [1985] and Yashchin [1993] showed that the CuSum method slightly outperformed the EWMA when a mean shift size is equal to its standard deviation. Joner et al. [2008] compared a CuSum and a scan statistic method on Bernoulli observations and found that the CuSum outperformed the scan statistic given a steady-state ARL_0.

6.4 Residual Charts in Public Health Surveillance

An important assumption in control chart usage is that the data generated by the in-control process are independently distributed with mean μ and standard deviation σ. Both parameters are considered fixed and unknown. A shift in either μ or σ to a different value would lead to an out-of-control condition. When the process is in-control, the quality characteristic at time t, x_t, can be represented by the model $x_t = \mu + \varepsilon_t$, where $\varepsilon_t \sim (0, \sigma)$. Under these assumptions, conventional control charts can be used to draw conclusions about the statistical control state of the process. Though a Gaussian error structure is assumed, these charts still work reasonably well even when the normality

assumption is slightly to moderately violated [Montgomery, 2005]. Chen et al. [in review] investigated the in-control and out-of-control sensitivities to the third and fourth standardized moments of the symmetric $\bar{X}$ chart.

The independence of observations is a critical assumption regarding control chart usage. Conventional control charts perform poorly even with low levels of autocorrelation over time. If successive observations are positively correlated, even to a small degree such as $r = 0.25$, where r is Pearson's correlation coefficient, the number of false alarm signals will substantially increase [Montgomery, 2005]. Since the assumption of independence is sometimes not satisfied in practice, autocorrelation is an important issue to consider in control chart implementation.

An approach often used in dealing with autocorrelated data is to remove the autocorrelative structure with an appropriate model, and then to apply a conventional control chart to the residuals. In this case, the residuals would be approximately normal and independently distributed with mean zero and constant variance. Any unusual pattern in the sequence of residuals on the chart would then imply that the original variable was out of control.

It has been noted that residual control charts are not very sensitive to small process shifts. The CuSum or EWMA charts could be applied to residual data instead of Shewhart charts to help improve sensitivity. Tseng and Adams [1994] have found that since the EWMA is not an optimal forecasting scheme for most processes, it will not completely account for all autocorrelation. This, in turn, can affect the statistical performance of control charts that are based on EWMA residuals. Montgomery and Mastrangelo [1991] show that the use of supplementary procedures called tracking signals**,** combined with residual control charts, considerably enhance the performance of such charts. Various forecasting methods have been used to remove the autocorrelation in process data for use in residual control charts. Three widely used approaches include regression modeling, ARIMA modeling, and EWMA methods.

6.4.1 Preconditioning Data with Regression Models

The NYC Department of Health [Das et al., 2005] monitors sales of over-the-counter (OTC) medications by tracking residuals from a cyclic linear regression model. Brillman et al. [2005] developed a loglinear regression model based on the Serfling cyclic regression method for each of seven chief complaint categories. They tracked emergency department (ED) chief complaints data via a residual chart. Burkom et al. [2007] used a residual chart to compare a nonadaptive, loglinear regression model using a long historical baseline and an adaptive regression model with a shorter, sliding baseline. Lewis et al. [2002], Jackson et al. [2007], and Burr et al. [2006] used Poisson regression to model seasonal data, then tracked residuals using a Shewart method.

6.4.2 Preconditioning Data with ARIMA Methods

Autoregressive Integrated Moving Average (ARIMA) methods are also commonly used to model seasonal effects in syndromic surveillance data. ESSENCE II, a syndromic surveillance system developed by the Department of Defense Global Emerging Infections System and the Johns Hopkins University Applied Physics Laboratory, is a regression-based behavior modeling method with an ARIMA error structure [Burkom, 2003]. Mandl et al. [2004] used a hybrid method of ARIMA and cyclic regression with good predictive ability. Lewis et al. [2002] described the ESSENCE outbreak detection system in the greater Washington, DC area. Baseline levels of three of seven syndrome groups were established through a regression-ARIMA model. Miller et al. [2004] used a hybrid model to predict daily counts of influenza-like illness given a three-year historical period. Ozonoff et al. [2004] compared a regression-ARIMA model, a spatial statistic, and a bivariate test statistic to predict upper respiratory infection counts in a major healthcare provider setting in eastern Massachusetts. Wang et al. [2005] proposed an automated outbreak detection system for syndromic surveillance which utilized an autoregressive periodic model (ARP) to describe daily ED visits relating to respiratory syndromes. Reis

and Mandl [2003] preconditioned data using a trimmed-mean seasonal model based on historical averages to estimate expected counts. To account for autocorrelation in the residuals, the authors fit an ARIMA(2,0,1) to the overall ED volume counts, and an ARIMA(1,0,1) to the respiratory-related ED volume. The same modeling strategy was used in Reis et al. [2003] which assessed the use of multi-day temporal filters for outbreak detection. Buckeridge et al. [2005] simulated outbreaks based on inhalational anthrax exposure which were superimposed onto real, baseline data. The authors first smoothed the data using a procedure described by Reis et al. [2003], then accounted for autocorrelation in the residuals using a seasonal ARIMA model.

6.4.3 Preconditioning Data using Exponential Smoothing Methods

A third preconditioning approach for residual chart usage is Holt-Winters exponential smoothing. The Holt-Winters exponential smoothing algorithm, a variant of simple exponential smoothing, is often used for forecasting series that exhibit seasonality or trend, characteristics inherent to most syndromic surveillance series. Burkom et al. [2007] compared the performance of the multiplicative Holt-Winters procedure to adaptive and nonadaptive regression methods and found that it outperformed both procedures in modeling the original series. Murphy and Burkom [2008] coupled six forecasting methods (including Holt-Winters and adaptive regression) with six anomaly detection measures, 36 pairs total, to compare detection performance and again found the Holt-Winters method to be superior to other approaches.

6.5 Summary

SPC charts have been widely adapted for use in disease and syndromic surveillance. These charts have been used to monitor counts of nosocomial infections, hospital emergency room visits, prescription drug sales, etc. An unusual change, usually an increase, in disease counts could trigger an alarm from an SPC chart before the increase

is observed using more-traditional health and disease surveillance methods. CuSum and EWMA charts utilize information collected from prior observations and therefore are able to detect smaller shift changes more efficiently than Shewhart charts. If data are correlated over time, CuSum and EWMA charts can be used on residual data where a preconditioning method was used to remove autocorrelation from the original series.

CHAPTER 7

COMPARISON OF CUSUM AND EWMA CHARTS FOR DETECTION OF INCREASES IN NEGATIVE BINOMIAL COUNTS

A fundamental goal of public health surveillance, particularly syndromic surveillance, is the timely detection of increases in the rate of unusual events. In syndromic surveillance, a significant increase in the incidence of monitored prodromal covariates would trigger an alert, possibly prompting the implementation of an intervention strategy after further investigation of a possible illness outbreak. Public health surveillance generally monitors count data (e.g., counts of influenza-like illness, sales of over-the-counter remedies, and number of visits to outpatient clinics). These data, observed sequentially, are often assumed to follow Poisson dynamics. When an outbreak occurs, a shift in the baseline Poisson rate would occur. In many cases, however, the Poisson distribution is not an appropriate choice due to the assumption of mean and variance equality. Public health data are often overdispersed with respect to the Poisson distribution. To this end, the negative binomial distribution can be useful in describing discrete data where the variance exceeds the mean. Since the negative binomial distribution has two parameters, the second parameter can be used to adjust the variance independently of the mean.

Popular methods for monitoring and detecting public health surveillance data involve the use of CuSum and EWMA statistical process control charts [Montgomery, 2005; Hawkins and Olwell, 1998]. A detailed review of these methods as well as their applications in public health surveillance is given in Chapter 6. This chapter compares the CuSum and Exponentially Weighted Moving Average (EWMA) methods for detection of increases in negative binomial rates under independent and identically distributed (iid) conditions. As noted by Han et al. [2010], the behavior of these detection methods on discrete distributions has not been explored in detail. Several

studies have assessed the performance of such methods under Bernoulli or Poisson dynamics [Han et al., 2010; Joner et al., 2008]. Performance of the CuSum and EWMA detection methods will be evaluated using the conditional expected delay (CED) criterion under different shift sizes and different times at which the shift occurs.

7.1 Negative Binomial Distribution

There are several parameterizations of the negative binomial distribution [Hilbe, 2008; Hogg, 1995]. One widely used parameterization is as follows. Let the random variable X denote the total number of failures before the r^{th} success in a sequence of iid Bernoulli trials. Here, r is a positive, fixed integer and the parameter p denotes the probability of success. In this case, the probability mass function (pmf) for the negative binomial distribution is:

$$\Pr(X = x) = \binom{r + x - 1}{x} p^r (1-p)^x, \qquad x = 0,1,2, \ldots$$

$$= 0 \qquad \text{elsewhere}$$

with mean $\mu = \frac{r(1-p)}{p}$ and variance $\sigma^2 = \frac{r(1-p)}{p^2}$. Writing parameters r and p in terms of the mean μ and variance σ^2 gives $r = \frac{\mu^2}{\sigma^2 - \mu}$ and $p = \frac{\mu}{\sigma^2}$. From these equations, we have

$$\sigma^2 = \mu + \frac{1}{r}\mu^2.$$

In words, the variance is larger than the mean for the negative binomial, and it approaches the mean as r gets larger. Based on this relation, smaller values of r correspond to greater dispersion, and as such, r is sometimes referred to as the 'inverse dispersion parameter' (with $\alpha = 1/r$ referred to as the 'dispersion parameter').

An alternative formulation of the negative binomial pmf is parameterized with the mean μ and dispersion factor r since $p = \frac{r}{r+\mu}$. In this version, the negative binomial pmf can be written:

$$\Pr(X = x) = \frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)}\left(1+\frac{\mu}{r}\right)^{-r}\left(\frac{\mu}{\mu+r}\right)^{x}, \qquad \mu, r > 0, \quad x = 0.1.2 \ldots.$$

$$= 0 \qquad \text{elsewhere}$$

where $\Gamma(x)$ is the gamma function. In this formulation, the combinatorial roots for defining the negative binomial distribution are essentially ignored, and instead the distribution is viewed as a general-purpose discrete distribution that can be used to model nonnegative, integer-valued data. It should be noted that r can now be any non-negative, real number since the binomial coefficient in the pmf definition is replaced by equivalent gamma expressions. Further, the second formulation can be viewed as a generalization of the Poisson. Consider:

$$\lim_{r\to\infty}[\Pr(X = x)] = \lim_{r\to\infty}\frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)}\left(1+\frac{\mu}{r}\right)^{-r}\left(\frac{\mu}{\mu+r}\right)^{x}$$

$$= \lim_{r\to\infty}\frac{\mu^x}{x!}\frac{\Gamma(x+r)}{\Gamma(r)(\mu+r)^x}\left(1+\frac{\mu}{r}\right)^{-r}$$

$$= \frac{\mu^x e^{-\mu}}{x!}$$

In words, the negative binomial distribution converges to the Poisson distribution with parameter r controlling the deviation from the Poisson. Again, this makes the negative binomial distribution a more–robust alternative to the Poisson. The negative binomial approaches the Poisson for large r, but has a larger variance than Poisson for small r. In this study, the negative binomial parameterization with mean μ and dispersion parameter r will be utilized for the CuSum log-likelihood ratio since it directly incorporates the in-control and out-of-control parameters, μ_0 and μ_1, being tested.

7.2 Detection Methods

7.2.1 Negative Binomial Cumulative Sum Chart

The cumulative sum (CuSum) technique can be viewed as a sequential hypothesis test. For the negative binomial distribution with fixed dispersion factor r, the null hypothesis

specifies a target, in-control location parameter (μ_0) while the alternative hypothesis specifies a target, out-of-control location parameter (μ_1). The method monitors the statistic C_t where

$$C_t = \max\{0, C_{t-1} + L_t\}$$

and where the increment L_t is the log-likelihood ratio:

$$L_t = \ln \frac{f(x_t; r, \mu_1)}{f(x_t; r, \mu_0)}.$$

By convention, $C_0 = 0$. The method stops when an alarm is triggered, which occurs when $C_t > h$, where h is a threshold level determined by the in-control ARL. Further, once an alarm is triggered, the test concludes that a shift in the location parameter has occurred from μ_0 to μ_1.

If f_0 and f_1 are negative binomial pmf's with fixed dispersion factor *r*, the log-likelihood ratio simplifies to:

$$\begin{aligned}
L(x) &= \ln \left[\frac{\frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)} \left(1 + \frac{\mu_1}{r}\right)^{-r} \left(\frac{\mu_1}{\mu_1 + r}\right)^{x}}{\frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)} \left(1 + \frac{\mu_0}{r}\right)^{-r} \left(\frac{\mu_0}{\mu_0 + r}\right)^{x}} \right] \\
&= \ln \left[\left(1 + \frac{\mu_1}{r}\right)^{-r} \left(1 + \frac{\mu_0}{r}\right)^{r} \left(\frac{\mu_1}{\mu_1 + r}\right)^{x} \left(\frac{\mu_0}{\mu_0 + r}\right)^{-x} \right] \\
&= r \ln \left[\frac{r + \mu_0}{r + \mu_1} \right] + x \ln \left[\frac{\mu_1(\mu_0 + r)}{\mu_0(\mu_1 + r)} \right] \\
&= x \ln \left[\frac{\mu_1(\mu_0 + r)}{\mu_0(\mu_1 + r)} \right] - r \ln \left[\frac{r + \mu_0}{r + \mu_1} \right]^{-1}
\end{aligned}$$

According to Hawkins and Olwell [1998], if the probability distribution for the CuSum statistic is a member of the single–parameter exponential family, then the optimal CuSum design is completely specified by the selection of the in-control parameter, the out-of-

control parameter, and the in-control ARL. Further, in such cases, a one-sided, upper CuSum statistic C_t can be recursively calculated by the equation

$$C_t = \max\{0, C_{t-1} + X_t - k\},$$

where k is known as the reference value.

In this case, since r is considered fixed, the negative binomial is a member of the exponential family, and with the log-likelihood ratio above, the reference value k becomes:

$$k = \frac{r\ln\left[\frac{r+\mu_1}{r+\mu_0}\right]}{\ln\left[\frac{\mu_1(\mu_0+r)}{\mu_0(\mu_1+r)}\right]}.$$

7.2.2 Exponentially Weighted Moving Average Chart

The EWMA statistic, E_t, is recursively calculated by the equation

$$E_t = \alpha X_t + (1-\alpha)E_{t-1},$$

where $0 < \alpha \leq 1$ and $E_0 = E(X)$. A straightforward, thorough explanation of the EWMA statistic and control chart is given by Montgomery [2005]. The one-sided EWMA is used in cases where only an increase (or decrease) in the process mean is of interest. In this case, the monitored statistic is

$$E_t' = \max\{\mu_0, E_t\}.$$

For both the conventional EWMA and the one-sided EWMA, the parameter α is generally optimized through a grid search of α values (e.g., $\alpha = 0.1, 0.2, \ldots, 1.0$), where the optimal α value is determined by minimizing a function of the forecast residuals.

7.3 Simulation Study

7.3.1 Study Design and Parameter Selection

A simulation study was conducted to investigate the detection capabilities of the EWMA and CuSum methods with negative binomial count data. Three negative binomial data

distributions were used to represent disease incidence, all with an in-control parameter mean, $\mu_0 = 1.4$. Three levels of the dispersion parameter, r, were used to depict three levels of variance: (1) $r = 19.6$, (2) $r = 0.7$, and (3) $r = 0.156$. A low level of variance, where the variance is approximately equal to the in-control mean ($\mu_0 = 1.4, \sigma^2 = 1.5$), corresponded to dispersion factor $r = 19.6$. Note that when $r = 19.6$, the negative binomial distribution is a close approximation of the Poisson distribution with parameters $\mu_0 = \sigma^2 = 1.4$. The second dispersion value, $r = 0.7$, was used to describe a mid-level of variance, $\sigma^2 = 4.2$, which is three times the in-control mean value, $\mu_0 = 1.4$. The third dispersion factor, $r = 0.156$, corresponds to a high level of variance, $\sigma^2 = 14$, ten times the in-control mean, $\mu_0 = 1.4$.

For each of these three data sets, two upward shifts from the target in-control mean ($\mu_1 = 1.4$) were to be detected: a small shift defined as a 25% increase in the mean number of cases ($\mu_1 = 1.75$), and a large shift defined as a 75% increase in the mean number of cases ($\mu_1 = 2.45$). Table 7.1 shows the parameter values and shifts for each of the six scenarios. Each of these simulated data sets was used to compare the performance of the CuSum and EWMA monitoring methods.

For the simulation study, optimal threshold levels were obtained. For each of the methods, ARL_0 was set as close as possible to 1,500 without going below this threshold. In other words, the threshold was set to generate one false alarm per 1,500 time periods. The threshold level corresponding to an ARL_0 = 1,500 was determined based on 160,000 simulations.

The threshold values for ARL_0 were determined using an approach similar to that taken by Han et al. [2010]. For the CuSum simulation, Table 7.2 displays the set of parameters, thresholds, and corresponding ARL_0. For the EWMA approach, the parameters and thresholds determined by simulation are given in Table 7.3.

7.3.2 Conditional Expected Delay

The goal of each method is to detect an increase in the mean number of cases as soon as possible after an upward shift (μ_1, where $\mu_1 > \mu_0$) has occurred. Generally, the performance of a detection method is evaluated using two criteria: (1) the false alarm rate during the time of the in-control state, and (2) the detection delay in the out-of-control state. As previously stated, the false alarm rate is fixed at 1 alert per 1,500 time periods. The conditional expected delay criterion, $CED(v, \mu_1)$, is a measure of the detection delay, i.e., the time between the occurrence of a shift and when that shift is detected. The variable *v* represents the time at which the true shift occurred. The variable μ_1 represents the size of the shift. Thus, the criterion $CED(v, \mu_1)$ is used to compare the two methods at various shift times *v*, as well as a range of shift sizes (μ_1). Each calculated $CED(v, \mu_1)$ value was based on 15,000 simulations.

7.3.3 $CED(v, \mu_1)$ results under fixed size of shift (μ_1) and varying time of shift (v)

The two detection methods were compared by first considering different points of time, *v*, at which a shift occurred. The value μ_1 represents the shift size, either $\mu_1 = 1.75$ (25% upward shift) or $\mu_1 = 2.45$ (75% upward shift). Figures 7.1*a,c,e* depict $CED(v, \mu_1)$ results for the small shift size ($\mu_1 = 1.75$). Figure 7.1*a* shows results for small variance ($r = 19.6$), Figure 7.1*c* for mid-level variance ($r = 0.7$), and Figure 7.1*e* for high variance ($r = 0.156$). Based on these three figures, the EWMA approach outperformed the CuSum approach in detecting the smaller shift (using the criterion $CED(v, \mu_1 = 1.75)$). The greater the variance, the better the EWMA method performed relative to the CuSum. Further, for all three variance levels, the EWMA *CED* values appear to be more robust relative to the CuSum approach. In other words, for the EWMA method, *CED* values for a shift occurring at an early time period are very close to a *CED* value for a shift occurring at a later time period.

Figures 7.1*b,d,f* illustrate *CED* results for the larger shift size, ($\mu_1 = 2.45$). The two methods are compared given negative binomial data with small variance (Figure 7.1*b*), mid-level variance (Figure 7.1*d*), and high variance (Figure 7.1*f*). Figure 7.1*b* shows that, after a short start-up period, the CuSum outperformed the conventional EWMA in detecting the larger shift. This result is consistent with findings by Han et al. [2010] for Poisson data. Montgomery [2005] also notes that the CuSum method detects higher shifts more quickly than does the EWMA method. The greater amount of variability in the data again makes both approaches less efficient in detecting a shift. For the mid-level variation (Figure 7.1*d*), the CuSum performed worse than the EWMA method when the shift occurred at an early time period. However, when the shift occurred later ($v \approx 33$ or later), the CuSum performed as well as the EWMA. For large variance (Figure 7.1*f*), the EWMA method outperformed the CuSum regardless of the time when shift occurred.

7.3.4 $CED(v, \mu_1)$ results under fixed time of shift (v) and varying size of shift (μ_1)

In the previous subsection, the two methods were compared using the *CED* criterion when the true shift size was known, but the time of the shift was unknown. In this section, the $CED(v, \mu_1)$ values are assessed when the true shift size μ_1 is unknown. The time of a true shift occurrence is held constant at $v = 50$ so that the methods have time to adjust to the in-control, background data prior to the shift. To investigate the condition where μ_1 is unknown, a target shift size (μ_1^*) can be set and the pattern of *CEDs* under different true shift sizes at some fixed point in time $v = v^*$ can be assessed.

Figure 7.2 depicts the results for the three data distributions: low variance (Figures 7.2*a,b*), mid-level variance (Figures 7.2*c,d*) and high variance (Figures 7.2*e,f*). Figures 7.2*a,c,e* illustrate results with a small target shift at $\mu_1^* = 1.75$. Figures 7.2*b,d,f* give the results for a higher target shift at $\mu_1^* = 2.45$. The true shift sizes μ_1 (*x*-axis) range from 1.45 to 3.50.

Regardless of target shift size, the *CED* values of both methods increase with increasing variance, and take longer to converge with increasing variance. Among all three charts with smaller target value (Figures 7.2*a,c,e*), smaller true shift size is detected more quickly than among charts with larger target shift size (Figures 7.2*b,d,f*). With the smaller target shift size (Figures 7.2*a,c,e*), the EWMA method performed moderately better than the CuSum for low- and mid-variance data, and performed only slightly better given high-variance data. The two methods converged when the true and target shifts were equal (i.e, when $\mu_1^* = \mu_1 = 1.75$). Given the higher target shift size and the smallest variance case (Figure 7.2*b*), the EWMA outperformed the CuSum for smaller true shift sizes with both methods converging near $\mu_1 = 1.75$. The same outcome was observed for the higher target shift and mid-level variance case (Figure 7.2*d*). For the high target shift and high-level variance case (Figure 7.2*f*), the EWMA chart only slightly outperformed the CuSum given smaller true time shifts.

7.4 Conclusions

This study evaluated the performance of the CuSum and EWMA monitoring methods with negative binomial data observations. These results show that when the variance is larger than the mean, *CED* is larger. The greater variability in the data makes it more difficult for the two methods to detect shifts in general.

With the smaller shift size, the conventional EWMA method detected the shift more quickly than the CuSum at all levels of variance. With larger shift sizes, if the variance is relatively small (in this case approximately equal to the mean), the CuSum method outperformed the EWMA after a short start-up period. For large shifts with data having mid-level variance, the CuSum performed as well as the EWMA method after a longer start-up period.

When the two methods are monitoring data targeted for large shifts, true smaller shifts will either go unnoticed or will be extremely difficult to detect. On the other hand,

when the methods are monitoring for a smaller shift size, larger shift sizes also become easier to detect. In both scenarios, given low- to mid-level variance, the EWMA performed slightly better in detecting smaller true shifts. With high variance, however, both methods performed comparably with EWMA having only a slight edge in detecting smaller true shifts.

Results from this study should be helpful in deciding which chart to use for monitoring and detecting changes in rates of rare diseases with overdispersed variance. The negative binomial distribution may better approximate the underlying distribution of events over time compared to the Poisson. Further, based on these results, the detection of small shifts in disease rates would be quicker with the EWMA chart compared to the CuSum. For detection of larger shifts, either approach may be used.

Table 7.1 Negative binomial parameter values for the simulation study assessing the performance of the CuSum and EWMA monitoring methods, $\mu_0 = 1.4$.

μ_1	σ^2	r
1.75	1.5	19.6
1.75	4.2	0.7
1.75	14	0.156
2.45	1.5	19.6
2.45	4.2	0.7
2.45	14	0.156

Table 7.2 CuSum thresholds (h) determined via simulation (target $ARL_0 = 1{,}500$)

r	$\mu_1\ (v = 1)$	μ_1^*	h	ARL_0 (s.e.)
19.6	1.75	1.75	18.150	1526.22 (3.54)
19.6	2.45	2.45	9.905	1522.79 (3.61)
0.7	1.75	1.75	41.200	1523.13 (3.57)
0.7	2.45	2.45	26.850	1523.67 (3.58)
0.156	1.75	1.75	95.0	1522.93 (3.65)
0.156	2.45	2.45	70.2	1531.15 (3.61)

Table 7.3 EWMA thresholds (h) determined via simulation (target $ARL_0 = 1{,}500$).

r	$\mu_1\ (v = 1)$	a	h	ARL_0 (s.e.)
19.6	1.75	0.02	1.715	1519.21 (3.59)
19.6	2.45	0.09	2.274	1526.40 (3.62)
0.7	1.75	0.01	1.731	1517.64 (3.56)
0.7	2.45	0.03	2.160	1528.21 (3.63)
0.156	1.75	0.01	2.033	1532.71 (3.56)
0.156	2.45	0.01	2.033	1532.71 (3.56)

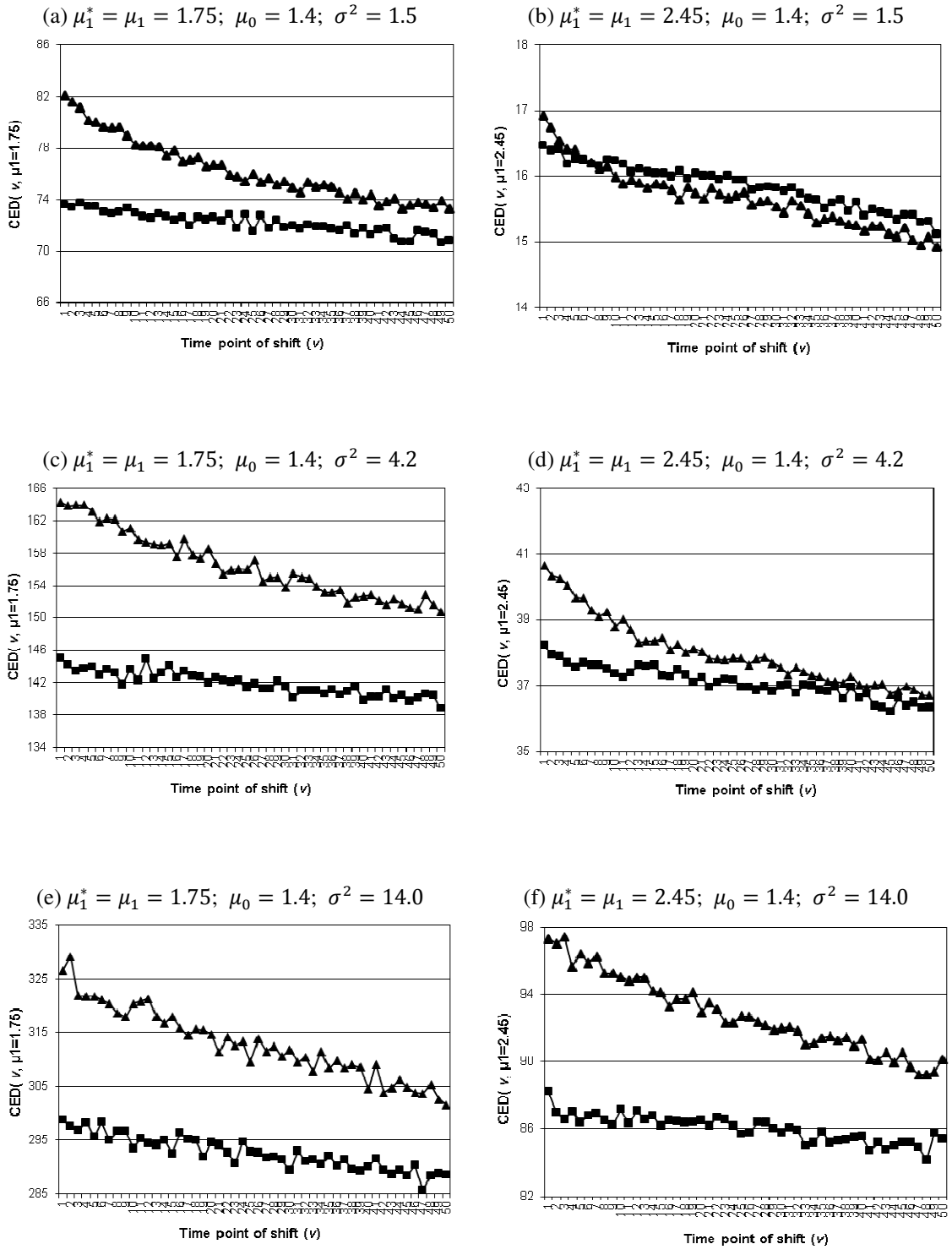

Figure 7.1 Conditional Expected Delay (*CED*) comparisons of CuSum (triangles) and EWMA (squares) by time of true shift (v). Figures *a,c,e* have true shift $\mu_1 = 1.75$ and target shift $\mu_1^* = 1.75$. Figures *b,d,f* have true shift $\mu_1 = 2.45$ and target shift $\mu_1^* = 2.45$. Detection methods are compared across three simulated negative binomial time series with parameters $\mu_0 = 1.4$ and variances $\sigma^2 = 1.5$ (Figures *a,b*); $\sigma^2 = 4.2$ (Figures *c,d*); and $\sigma^2 = 14$ (Figures *e,f*).

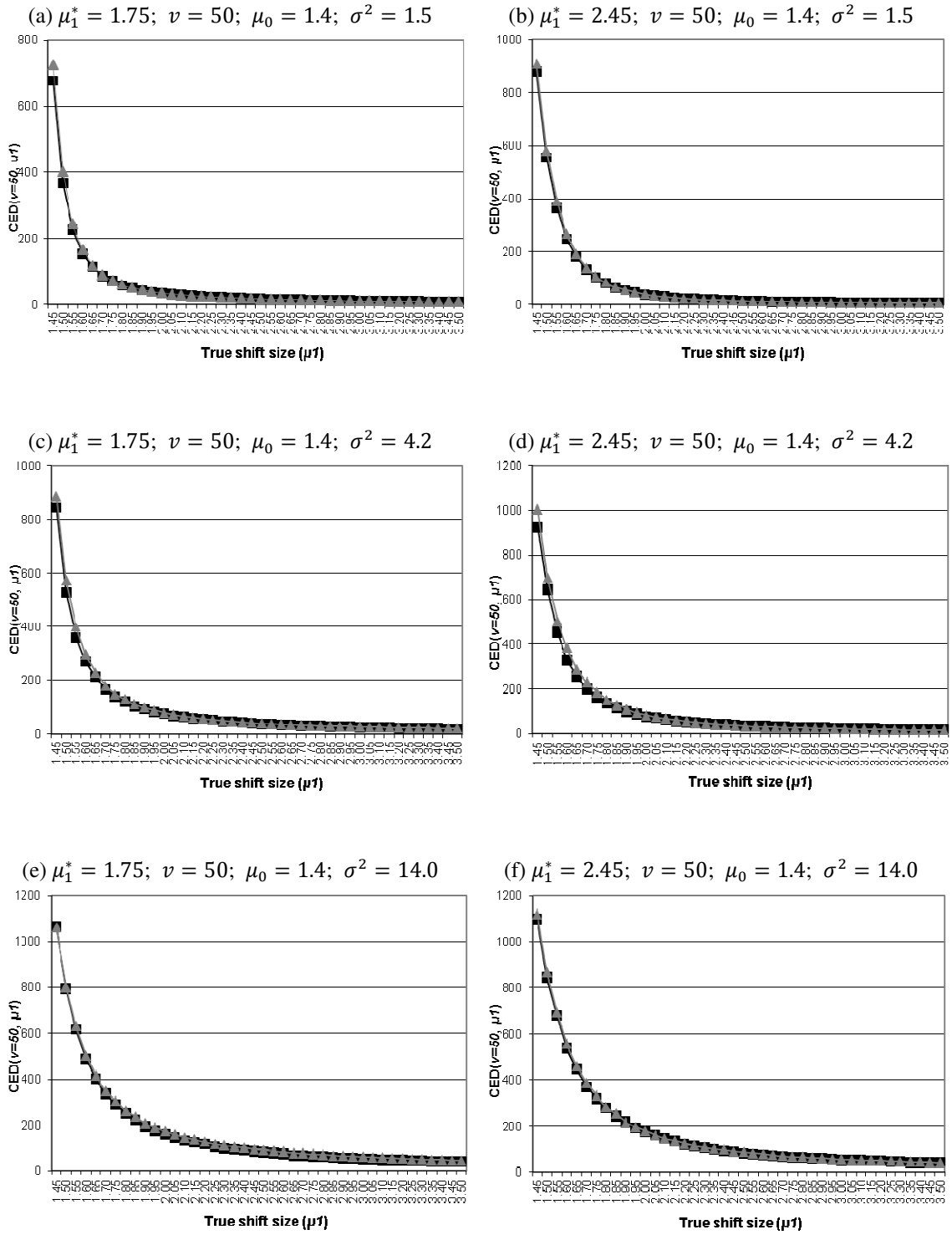

Figure 7.2 *CED* comparisons of CuSum (grey triangles) and EWMA (black squares) detection methods across different true mean shift sizes (μ_1) at fixed time of shift ($v = 50$). Figures *a,c,e* have target shift $\mu_1^* = 1.75$. Figures *b,d,f* have target shift $\mu_1^* = 2.45$. Detection methods are compared across three simulated negative binomial time series with parameters $\mu_0 = 1.4$ and variances $\sigma^2 = 1.5$ (Figures *a,b*); $\sigma^2 = 4.2$ (Figures *c,d*); and $\sigma^2 = 14$ (Figures *e,f*).

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This chapter summarizes the major results and conclusions from this thesis. Suggestions for future research in these areas of public health are also given.

8.1 Summary of Results and Conclusions

8.1.1 Pooling Influenza-associated Mortality Risks Across Locations

This topic addressed the issue of local-level seasonal confounding in modeling national-level, influenza-associated mortality rates by introducing and implementing a two-stage hierarchical Bayesian modeling approach. Results from this study showed considerable variability at the local level with respect to the association between influenza activity and mortality for subtypes A(H1N1) and A(H3N2). These findings suggest that ambient temperature and other local-level seasonal factors significantly affect the relationship between influenza and deaths among persons aged 65 or older. To better account for these city-level, time-dependent confounders, hierarchical modeling methods should be considered in future studies.

Results from two hierarchical models were compared to those from a traditional modeling approach. For A(H3N2) activity, the traditional GLM method estimated the risk of death to be significantly higher than the estimates from the two hierarchical models. This difference may be attributable to the greater influence of counts from large cities when modeled via the traditional method. The hierarchical approaches average all city-level parameter estimates taking into account estimated city-level variance, while the traditional method places more weight on outcomes of larger cities by pooling all data before modeling.

Results from the A(H1N1) influenza-associated mortality showed the hierarchical and traditional approaches leading to very different conclusions. With A(H1N1), the relative risk is null from both hierarchical models, which is not surprising given the very low activity of A(H1N1) during the study period. The traditional GLM model, however, showed a significant increase in deaths given A(H1N1) activity. Again, this may be attributed to the high log-relative risk values of a few large cities. For type B influenza activity, the three modeling approaches yielded similar results, suggesting that viral activity did not vary much geographically and/or other seasonal factors were not temporally collinear with type B influenza. In this case, any of the three modeling approaches may be suitable for modeling.

8.1.2 Modeling Measured and Unmeasured Background Seasonality

This topic examined the effect of varying measures of seasonality on the association between influenza and mortality. Three regression modeling approaches with distinct representations of background seasonality were chosen for comparison. The covariates included data-driven as well as mathematically modeled proxies for seasonality. Based on results from this study, background seasonality affects influenza-associated death estimates to varying degrees depending on viral subtype, city, and climate. Estimates of A(H3N2) deaths are the most robust across all model types and seasonal representations. Findings from this study suggest, given U.S. seasonal influenza patterns, that spline functions with 20 knots are good approximations for data-driven seasonality variables such as ambient temperature.

8.1.3 The Nonlinear Association Between Influenza and Mortality

This topic investigated the relationship between influenza and mortality by modeling influenza terms using a flexible, natural cubic spline function instead of the conventional linear approach. A significant finding from this study is the apparent nonlinear

association between influenza and mortality. Given log-transformed mortality, we expected to find a nonlinear association between influenza and mortality that increased multiplicatively. Results based on linearly modeled viral terms showed no multiplicative increase, and instead showed only a linear association (despite the logged mortality term). The influenza spline functions showed that the association is indeed nonlinear; however, instead of mortality increasing multiplicatively with increasing influenza activity, the association tended to taper off towards peak influenza season. This study showed that in this study that relative to the spline fit, the linear function consistently underestimated the association during off-peak periods and overestimated during peak periods. Based on the observed shape of the association, either spline functions, quadratic functions, or logged viral terms might offer a better fit for the influenza viral terms.

8.1.4 Comparison of CuSum and EWMA Charts for Outbreak Detection

This study evaluated the performance of the CuSum and EWMA monitoring methods with negative binomial data observations. With the smaller shift size, the conventional EWMA method detected the shift more quickly than the CuSum at all levels of variance. With larger shift sizes, if the variance is relatively small (in this case approximately equal to the mean), the CuSum method outperformed the EWMA after a short start-up period. However, as variability increased, the EWMA again outperformed the CuSum. Results from this study should be helpful in deciding which chart to use for monitoring and detecting changes in rates of diseases with overdispersed variance. The negative binomial distribution may better approximate the underlying distribution of events over time compared to the Poisson. Further, based on these results, the detection of small shifts in disease rates may occur more quickly with the EWMA chart compared to the CuSum. For detection of larger shifts, either approach may be used.

8.2 Future Research

Based on results and conclusions from this work, the following suggestions are made for future research.

(1) Validation of results is particularly difficult when modeling influenza-associated deaths via regression methods since there are a number of temporally collinear variates. Proper simulation studies which adequately represent mortality series with non-regression derived components of influenza-associated and non-influenza-associated deaths are needed to validate and compare results from these modeling methods.

(2) Bayesian hierarchical modeling is a novel approach to modeling influenza-associated deaths. It requires further investigation. For instance, the assumption of a normal approximation for the likelihood function of each city-level model needs further examination, and an appropriate approach for estimating the correlation parameter in the spatial-correlation hierarchical model should be determined.

(3) The percent-positive, influenza activity proxy may need further vetting. An alternative proxy for influenza activity that incorporates both estimated percent-positive activity as well as estimated percentage of influenza-like illness medical visits might better capture the true level of influenza activity in a particular region.

(4) The percent-positive influenza proxy was also found to be nonlinearly associated with mortality. Based on the observed shape of the association, either spline functions, quadratic terms, or logged viral terms might offer better fits for the influenza viral terms in future studies.

(5) Syndromic surveillance data tends to contain long-term and seasonal trends, and generally are correlated over time. A logical extension of this part of the thesis

would involve monitoring data with seasonal and other long-term patterns. One approach might be to first precondition the data to remove time-dependent trends, and then compare SPC methods that monitor the residuals for detection of unusual activity.

REFERENCES

Benneyan JC. Statistical quality control methods in infection control and hospital epidemiology, part II: Chart use, statistical properties, and research issues. Infection Control and Hospital Epidemiology 1998;19(4):265−283.

Bisno AL, Griffin JP, Van Epps KA, Niell HB, Rytel MW. Pneumonia and Hong Kong influenza: a prospective study of the 1968-1969 epidemic. Am J Med Sci 1971; 261:251−263.

Borror CM, Rigdon SE, Champ CW. An exponentially weighted moving average control chart for Poisson data. *Proceedings of the 23rd Annual Modeling and Simulation Conference*, Pittsburgh, PA, 1992; 1775−1782.

Borror CM, Champ CW, Rigdon SE. Poisson EWMA control chart. *Journal of Quality Technology* 1998; **30**:352−361.

Bradley CA, Rolka H, Walker D, Loonsk J. BioSense: Implementation of a National Early Event Detection and Situational Awareness System. Morbidity and Mortality Weekly Report 2005;54(Suppl):11−19.

Bridges CB, Harper SA, Fukuda K, Uyeki TM, Cox NJ, Singleton JA. Prevention and control of influenza. Recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep 2003; 52(RR-8):1−34.

Brillman J, Burr T, Forslund D, Joyce E, Picard R, Umland E. Modeling emergency department visit patterns for infectious disease complaints: results and application to disease surveillance. BMC Medical Informatics and Decision Making 2005;5:4. Available at www.biomedcentral.com/1472-6947/5/4.

Brook D, Evan DA. An approach to the probability distribution of CUSUM run length. *Biometrika* 1972; **59**:539−549.

Brown SM, Benneyan JC, Theobald DA, Sands K, Hahn MT, Potter-Bynoe GA, et al. Binary cumulative sums and moving averages in nosocomial infection cluster detection. Emerging Infectious Disease 2002;8(12):1426−1432.

Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics 2005;38:99−113.

Burkom H. Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance. Johns Hopkins APL Technical Digest 2003;24(4): 335−42.

Burkom HS, Murphy SP, Shmueli G. Automated Time Series Forecasting for Biosurveillance. Statistics in Medicine 2007;26(22):4202−18.

Burr T, Graves T, Klamann R, Michalak S, Picard R, Hengartner N. Accounting for seasonal patterns in syndromic surveillance data for outbreak detection. BMC Medical Informatics and Decision Making 2006;6:40. Available at www.biomedcentral.com/1472-6947/6/40.

Centers for Disease Control and Prevention. Influenza Activity -- United States, 2012–13 Season and Composition of the 2013–14 Influenza Vaccine. MMWR 2013; 62(23);473−479. In text citation (CDC, 2013).

Charu V, Simonsen L, Lustig R, Steiner C, Viboud C. Mortality burden of the 2009-10 influenza pandemic in the United States: Improving the timeliness of influenza severity estimates using inpatient mortality records. Influenza Other Respir Viruses 2013; 7(5):863-71. doi: 10.1111/irv.12096. Epub 2013 Feb 19.

Chen H, Goldsman D, Schmeiser BW, Tsui K-L. Symmetric $\bar{X}$ Charts: Sensitivity to Nonnormality and Control-Limit Estimation. Communications in Statistics (in review).

Chen J-H, Schmit K, Chang H, Herlihy E, Miller J, Smith P. Use of Medicaid Prescription Data for Syndromic Surveillance – New York. Morbidity and Mortality Weekly Review 2005;54(Suppl):31−34.

Cheng P-Y, Thompson W, Dhara R, Ozonoff A, Brammer L, Weintraub E, Blanton L, Shay D. Estimating the weekly number of influenza-associated deaths using the 122 Cities Mortality Reporting System. PlosOne (in review).

Collins SD, Lehmann J. Trends and epidemics of influenza and pneumonia: 1918−1951. Public Health Rep 1951; 66:1487−1516.

Cowling BJ, Wong IOL, Ho LM, Riley S, Leung GM. Methods for monitoring influenza surveillance data. International Journal of Epidemiology 2006; **35**:1314−1321.

Crowder SV. A simple method for studying run length distributions of exponentially weighted moving average control charts. Technometrics 1987; **29**:401−407.

Crowder SV. Design of exponentially weighted moving average schemes. Journal of Quality Technology 1989; **21**:155−162.

Cryer JD. *Time Series Analysis*. 1986 Boston: Duxbury Press.

Daniels MJ, Kass RE. A note on first-stage approximation in two-stage hierarchical models. The Indian Journal of Statistics Series B 1998; 60:19−30.

Das D, Metzger K, Heffernan R, Balter S, Weiss D, Mostashari F. Monitoring Over-The-Counter Medication Sales for Early Detection of Disease Outbreaks -- New York City. Morbidity and Mortality Weekly Report 2005;54(Suppl):41−46.

Dominici F, Samet JM, Zeger SL. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. Journal of the Royal Statistical Society Series A 2000; 163:263−302.

Douglas RG Jr. Influenza: the disease and its complications. Hosp Pract 1976; 11:43−50.

Everson PJ, Morris CN. Inference for multivariate normal hierarchical models. Journal of the Royal Statistical Society Series B 2000; 62:399−412.

Farr W. Annual Report of the Registrar General. In: HMSO, ed. London, 1847.

Feng L, Shay DK, Jiang Y, Zhou H, Chen X, Zheng Y, Jiang L, Zhang Q, Lin H, Wang S, Ying Y, Xu Y, Wang N, Feng Z, Viboud C, Yang W, Yu H. Influenza-associated mortality in temperate and subtropical Chinese cities, 2003-2008. Bull World Health Organ. 2012 Apr 1;90(4):279−288B. doi: 10.2471/BLT.11.096958.

Fricker RD Jr. Syndromic surveillance. In *Encyclopedia of Quantitative Risk Assessment*, Melnick E, Everitt B (eds). Wiley: New York, 2008;1743−1752.

Gan FF. Monitoring Poisson observations using modified exponentially weighted moving average control charts. Communications in Statistics—Simulation and Computation 1990; **19**:103−124.

Goldstein E, Viboud C, Charu V, Lipsitch M. Improving the estimation of influenza-related mortality over a seasonal baseline. Epidemiology. 2012 Nov;23(6):829−38. doi: 10.1097/EDE.0b013e31826c2dda.

Han SW, Tsui K-L, Ariyajanya B, Kim SB. A Comparison of CUSUM, EWMA, and Temporal Scan Statistics for Detection of Increases in Poisson Rates. Quality and Reliability Engineering International 2010; 26:279−289.

Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. 1990 London: Chapman and Hall.

Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning, Data Mining, Inference, and Prediction*. 2001 New York: Springer-Verlag.

Hawkins DM, Olwell DH. *Cumulative Sum Charts and Charting for Quality Improvement.* 1998 New York: Springer-Verlag.

Hilbe J. *Negative Binomial Regression.* 2008 New York: Cambridge University Press.

Hill GB, Spicer CC, Weatherall JAC. The computer surveillance of congenital malformations. British Medical Journal 1968; **24**:215−218.

Hogan WR, Tsui F-C, Ivanov O, Gesteland PH, Grannis S, Overhage M, Robinson M, Wagner MM. Detection of Pediatric Respiratory and Diarrheal Outbreaks from Sales of Over-the-Counter Electrolyte Products. Journal of the American Medical Informatics Association 2003;10:555−562.

Hogg RV, Allen TC. *Mathematical Statistics.* 1995 Englewood Cliffs, NJ: Prentice-Hall, Inc.

Hunter JS. The exponentially weighted moving average. Journal of Quality Technology 1986; **18**:203−210.

Ivanov O, Gesteland PH, Hogan W, Mundorff MB, Wagner MM. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. AMIA Annual Symposium Proceedings Archive 2003:318−322.

Jackson ML, Baer A, Painter I, Duchin J. A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Medical Informatics and Decision Making 2007; **7**:6.

Jolliffe IT. *Principal Component Analysis*. Second Edition, 2002 New York: Springer-Verlag.

Joner MD, Woodall WH, Reynolds MR, Fricker RD. A one-sided MEWMA chart for health surveillance. Quality and Reliability Engineering International 2008; **24**:503−519.

Joner MD, Woodall WH, Reynolds MR. Detecting a rate increase using a Bernoulli scan statistic. Statistics in Medicine 2008; **27**:2555−2575.

Lai TL. Sequential analysis: some classical problems and new challenges. Statistica Sinica 2001; **11**:303−408.

Le Cam L, Yang GL. *Asymptotics in statistics: some basic concepts.* 1990 New York: Springer-Verlag.

Lee ML, Goldsman D, Kim S-H, Tsui K-L. Spatiotemporal biosurveillance with spatial clusters: Control limit approximation and impact of spatial correlation. IIE Transactions 2014;46:813−827.

Lemaitre M, Carrat F, Rey G, Miller M, Simonsen L, Viboud C. Mortality burden of the 2009 A/H1N1 influenza pandemic in France: comparison to seasonal influenza and the A/H3N2 pandemic. PLoS One. 2012;7(9):e45051. doi: 10.1371/journal.pone.0045051. Epub 2012 Sep 20.

Lewis M, Pavlin J, Mansfield J, O'Brian S, Boomsma L, Elbert Y, Kelley P. Disease Outbreak Detection System Using Syndromic Data in the Greater Washington DC Area. American Journal of Preventive Medicine 2002;23(3):180−186.

Liao C-M, Chang S-Y, Chen S-C, Chio C-P. Influenza-associated morbidity in subtropical Taiwan. International Journal of Infectious Diseases 2009; 13:589−599.

Lofgren E, Fefferman NH, Naumov YN, Gorski J, Naumova EN. Influenza seasonality: Underlying causes and modeling theories. Journal of Virology 2007; 81:5429−5436.

Lorden G. Procedures for reacting to a change in distribution. Annals of Mathematical Statistics 1971; **42**:1897−1908.

Lucas JM. Counted data CUSUMs. Technometrics 1985; **27**:129−144.

Lucas JM, Saccucci MS. Exponentially weighted moving average control schemes: properties and enhancements. Technometrics 1990; **32**:1−12.

Lui KJ, Kendal AP. Impact of influenza epidemics on mortality in the United States from October 1972 to May 1985. Am J Public Health 1987; 77:712−716.

Mandl KD, Overhage M, Wagner MM, Lober WB, Sebastiani P, Mostashari F, et al. Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience. Journal of the American Medical Informatics Association 2004;11(2):141−150.

Mangtani P, Hajat S, Kovats S, Wilkinson P, Armstrong B. The association of respiratory syncytial virus infection and influenza with emergency admissions for respiratory disease in London: An analysis of routine surveillance data. Clinical Infectious Diseases 2006; 42:640−646.

Marx MA, Rodriguez CV, Greenko J, Das D, Heffernan R, Karpati AM, Mostashari F, Balter S, Layton M, Weiss D. Diarrheal Illness Detected Through Syndromic Surveillance After a Massive Power Outage: New York City, August 2003. American Journal of Public Health 2006;96(3): 547−553.

Meltzer MI, Cox NJ, Fukuda K. The economic impact of pandemic influenza in the United States: priorities for intervention. Emerg Infect Dis 1999; 5:659−671.

Miller B, Kassenborg H, Dunsmuir W, Griffith J, Hadidi M, Nordin JD, Danila R. Syndromic Surveillance for Influenzalike Illness in an Ambulatory Care Network. Emerging Infectious Diseases 2004;10(10):1806−1811.

Montgomery DC. *Statistical Quality Control*. Fifth Edition, 2005: Hoboken, NJ: John Wiley and Sons, Inc.

Montgomery DC and Mastrangelo CM. Some statistical process control methods for autocorrelated data. Journal of Quality Technology 1991;23(3):179−204.

Morgenstern H. Ecologic Studies in Epidemiology: Concepts, Principles, and Methods. Annual Review of Public Health 1995; Vol. 16: 61−81.

Murphy SP, Burkom H. Recombinant Temporal Aberration Detection Algorithms for Enhanced Biosurveillance. Journal of the American Medical Association 2008;15(1):77−86.

Muscatello DJ, Newall A, Dwyer DE, Macintyre CR. Mortality attributable to seasonal and pandemic influenza, Australia, 2003 to 2009, using a novel time series smoothing approach. PLoS One. 2013 Jun 3;8(6):e64734. doi: 10.1371/journal.pone.0064734. Print 2013.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models.* Fourth edition, 1996 Chicago: Irwin.

Newall AT, Wood JG, Macintyre CR. Influenza-related hospitalisation and death in Australians aged 50 years and older. Vaccine. 2008 Apr 16;26(17):2135−41. doi: 10.1016/j.vaccine.2008.01.051. Epub 2008 Feb 15.

Newall AT, Viboud C, Wood JG. Influenza-attributable mortality in Australians aged more than 50 years: a comparison of different modelling approaches. Epidemiol Infect 2010; 138:836-42. doi: 10.1017/S095026880999118X. Epub 2009 Nov 27.

Ng CH, Case KE. Development and evaluation of control charts using exponentially weighted moving averages. Journal of Quality Technology. 1989; **21**:242−250.

Nichol KL. The efficacy, effectiveness, and cost-effectiveness of inactivated influenza virus vaccines. Vaccine 2003; 21:1769−1775.

Nielsen J, Mazick A, Glismann S, Mølbak K. Excess mortality related to seasonal influenza and extreme temperatures in Denmark, 1994-2010. BMC Infect Dis. 2011 Dec 16;11:350. doi: 10.1186/1471-2334-11-350.

Ozonoff A, Forsberg L, Bonetti M, Pagano M. Research Methods: Bivariate Method for Spatio-Temporal Syndromic Surveillance. Morbidity and Mortality Weekly Report 2004;53(Suppl):59−66.

Page ES. Continuous inspection schemes. Biometrika 1954; **41**:100−115.

Peng RD, Dominici F. Statistical methods for environmental epidemiology with R. 2008 New York: Springer.

Pollak M. Optimal detection of a change in distribution. Annals of Statistics 1985; **13**:206−227.

Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. PNAS 2003;100(4):1961−1965.

Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Medical Informatics and Decision Making 2003;3:2. Available at www.biomedcentral.com/1472-6947/3/2

Robinson PB, Ho TY. Average run lengths of geometric moving average charts by numerical methods. Technometrics 1978; **20**:85−93.

Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Rep 1963; 78:494−505.

Shiryayev AN. On optimum methods in quickest detection problems. Theory of Probability and its Applications 1963; **8**:22−46.

Shu L, Jiang W, Wu S. A one-sided EWMA control chart for monitoring process means. Communications in Statistics—Simulation and Computation 2007; **36**:901−920.

Simonsen L, Fukuda K, Schonberger LB, Cox NJ. The impact of influenza epidemics on hospitalizations. J Infect Dis 2000; 181:831−837.

Srivastava MS, Wu YH. Comparison of EWMA, CUSUM and Shiryayev–Roberts procedures for detecting a shift in the mean. Annals of Statistics 1993; **21**:645−670.

Srivastava MS, Wu YH. Evaluation of optimum weights and average run lengths in EWMA control schemes. Communications in Statistics—Theory and Methods 1997; **26**:1253−1267.

Taubenberger JK, Morens DM. The pathology of influenza virus infections. Annu Rev Pathol 2008; 3:499−522.

Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ, Fukuda K. Mortality associated with influenza and respiratory syncytial virus in the United States. JAMA 2003; 289:179−186.

Thompson WW, Weintraub E, Dhankhar P, Cheng P-Y, Brammer L, Meltzer MI, Bresee JS, Shay DK. Estimates of US influenza-associated deaths made using four different methods. Influenza and Other Respiratory Viruses 2009; 3:37−49.

Tseng S, Adams BM. Monitoring Autocorrelated Processes with an Exponentially Weighted Moving Average Forecast. Journal of Statistical Computation and Simulation 1994;**50**(3): 187−195.

Tsui K-L, Chiu W, Gierlich P, Goldsman D, Liu X, Maschek T. A review of healthcare, public health, and syndromic surveillance. Quality Engineering 2008; **20**:435−450.

Ver Hoef JM, Boveng PL. Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? Ecology. 2007 Nov;88(11):2766−72.

Wang L, Ramoni MF, Mandl KD, Sebastiani P. (2005). "Factors affecting automated syndromic surveillance." Artificial Intelligence in Medicine 34(3):269−78.

Warren-Gash C, Bhaskaran K, Hayward A, Leung GM, Lo S-V, Wong C-M, Ellis J, Pebody R, Smeeth L, Cowling B. Circulating influenza virus, climatic factors, and acute myocardial infarction: A time series study in England and Wales and Hong Kong. Journal of Infectious Diseases 2011; 203:1710−1718.

Weatherall JAC, Haskey JC. Surveillance of malformations. British Medical Journal 1976; **32**:39−44.

White CH, Keats JB. ARLs and higher order run length moments for Poisson CUSUM. Journal of Quality Technology 1996; **28**:363−369.

Wong CM, Chan KP, Hedley AJ, Peiris JS. Influenza-associated mortality in Hong Kong. Clin Infect Dis. 2004 Dec 1;39(11):1611−17. Epub 2004 Nov 10.

Wong CM, Peiris JSM, Yang L, Chan KP, Thach TQ, Lai HK, Lim WWL, Hedley AJ, He J, Chen P, Ou C, Deng A, Zhang X, Zhou D, Ma S, Chow A. Effect of influenza on cardiorespiratory and all-cause mortality in Hong Kong, Singapore and Guangzhou. Hong Kong Med Journal 2012; 18(Suppl 2):S8−11.

Wood SN. *Generalized Additive Models: An introduction with R*. 2006 Boca Raton: CRC Press.

Woodall WH. The use of control charts in health-care and public health surveillance. Journal of Quality Technology. 2006;38:89−104.

Wooldridge JM. *Introductory Econometrics: A Modern Approach*. Second Edition, 2003 Ohio: South-Western.

Wu P, Goldstein E, Ho LM, Yang L, Nishiura H, Wu JT, Ip DK, Chuang SK, Tsang T, Cowling BJ. Excess mortality associated with influenza A and B virus in Hong Kong, 1998-2009. J Infect Dis. 2012 Dec 15;206(12):1862-71. doi: 10.1093/infdis/jis628. Epub 2012 Oct 8.

Yang L, Wong CM, Chan KP, Chau PY, Ou CQ, Chan KH, Peiris JS. Seasonal effects of influenza on mortality in a subtropical city. BMC Infect Dis. 2009 Aug 22;9:133. doi: 10.1186/1471-2334-9-133.

Yang L, Stefan M, Chen PY, He JF, Chan KP, Chow A, Ou CQ, Deng AP, Hedley AJ, Wong CM, Peiris JSM. Influenza associated mortality in the subtropics and tropics: Results from three Asian cities. Vaccine 2011a; 29:8909−8914.

Yang L, Chen PY, He JF, Chan KP, Ou CQ, Deng AP, Malik Peiris JS, Wong CM. Effect modification of environmental factors on influenza-associated mortality: a

time-series study in two Chinese cities. BMC Infect Dis. 2011b Dec 14;11:342. doi: 10.1186/1471-2334-11-342.

Yang L, Chan KP, Cowling BJ, Chiu SS, Chan KH, Peiris JSM, Wong CM. Excess mortality associated with the 2009 pandemic of influenza A(H1N1) in Hong Kong. Epidemiol Infect 2012; 140:1542−1550.

Yashchin E. Statistical control schemes: methods, applications and generalizations. International Statistical Review 1993; **61**:41−66.

Yu H, Feng L, Viboud CG, Shay DK, Jiang Y, Zhou H, Zhou M, Xu Z, Hu N, Yang W, Nie S. Regional variation in mortality impact of the 2009 A(H1N1) influenza pandemic in China. Influenza Other Respir Viruses 2013; 7(6):1350−60. doi: 10.1111/irv.12121. Epub 2013 May 13.

Yuan, C.M., Love, S., Wilson, M. Syndromic Surveillance at Hospital Emergency Departments – Southeastern Virginia. Morbidity and Mortality Weekly 2004;53(Suppl):56−58.

Zeger SL, Irizarry R, Peng R. On time series analysis of public health and biomedical data. Ann Rev Public Health 2006; 27:57−79.