Mortality due to Influenza in the United States—An Annualized Regression Approach Using Multiple-Cause Mortality Data

Jonathan Dushoff1,2, Joshua B. Plotkin3, Cecile Viboud2, David J. D. Earn4, and Lone Simonsen5

1 Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ.
2 Fogarty International Center, National Institutes of Health, Bethesda, MD.
3 Harvard University, Cambridge, MA.
4 Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada.
5 National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD.

Received for publication March 2, 2005; accepted for publication September 14, 2005.

Influenza is an important cause of mortality in temperate countries, but there is substantial controversy as to the total direct and indirect mortality burden imposed by influenza viruses. The authors have extracted multiple-cause death data from public-use data files for the United States from 1979 to 2001. The current research reevaluates attribution of deaths to influenza, by use of an annualized regression approach: comparing measures of excess deaths with measures of influenza virus prevalence by subtype over entire influenza seasons and attributing deaths to influenza by a regression model. This approach is more conservative in its assumptions than is earlier work, which used weekly regression models, or models based on fitting baselines, but it produces results consistent with these other methods, supporting the conclusion that influenza is an important cause of seasonal excess deaths. The regression model attributes an annual average of 41,400 (95% confidence interval: 27,100, 55,700) deaths to influenza over the period 1979–2001. The study also uses regional death data to investigate the effects of cold weather on annualized excess deaths.

cause of death; influenza; linear regression; mortality; seasons; temperature; time series; United States

Abbreviations: NAD, normalized annual death series; $T_{thresh}$, threshold temperature.

In temperate countries, more people die in the winter than in the summer (1). The contribution of influenza to these seasonal excess deaths remains a subject of controversy, with some authors arguing that influenza epidemics trigger the majority of excess deaths (1, 2) and others arguing that they trigger only a small minority (3). Retrospective cohort studies have shown a surprisingly large protective effect of influenza vaccination against deaths from any cause (4–6).

Deaths caused by influenza must be estimated indirectly because most influenza infections are not confirmed virologically (7). Moreover, deaths triggered by influenza may occur as a result of a number of final causes, including pneumonia and a wide variety of respiratory and circulatory causes, and may occur weeks after initial infection (1, 8–11). Serfling (12) developed a method to infer influenza deaths based on seasonal patterns in the deaths attributed to pneumonia and influenza. Serfling’s approach has been further developed by Simonsen et al. (13, 14). Influenza deaths are estimated as the number of deaths (either among all deaths or among pneumonia and influenza deaths) above an epidemic threshold based on a sinusoidal function fitted to summer deaths. Simonsen et al. (13) estimated that influenza caused approximately 21,000 deaths per year in the United States for the period 1972–1992. More recently, Simonsen et al. (14) estimated influenza deaths by use of winter excess mortality, controlling for changes in age.

Correspondence to Dr. Jonathan Dushoff, Department of Ecology and Evolutionary Biology, Princeton University, Guyot Hall, Princeton, NJ 08544 (e-mail: dushoff@eno.princeton.edu).
structure and for dominant viral subtype, and argued that the population-level burden of influenza is not consistent with the high protective values of vaccination found in cohort studies.

In a recent paper, Thompson et al. (11) introduced a new approach for estimating deaths attributable to influenza: They fit a regression model to death time series that combined a sinusoidal function with weekly virologic surveillance data from the United States. They attributed an average of 34,000 deaths per year to influenza for the period 1976–1999, consistent with the 37,500 deaths per year estimated by Serfling-type models for this period (14). Thompson et al. also estimated deaths due to both influenza and respiratory syncytial virus for the period 1990–1999.

This work by Thompson et al. (11) represents an advance over baseline methods, because it quantifies the relationship between deaths attributed to influenza infection and virologic surveillance data. The work also has significant limitations. Both influenza prevalence and mortality are highly seasonal. Therefore, fitting a weekly regression between deaths and influenza virus surveillance has the potential to produce spurious correlations: Even if the annual pattern of deaths were exactly the same every year, unaffected by changes in influenza prevalence, a weekly regression against influenza prevalence would likely indicate a significant correlation. Thompson et al. address this problem by including sinusoidal terms in their regression. Since non-influenza-related mortality is not expected to follow an exact exponentiated sinusoidal pattern, this reduces, but does not eliminate, the potential for spurious correlation and spurious attribution of deaths to influenza.

Another limitation of the weekly approach is the difficulty of correctly adjusting for time delays. It has been shown that peak mortality lags roughly 2 weeks behind the peak of influenza activity (8). This is likely the result of several different time delays from different influenza mortality pathways (1). The model of Thompson et al. (11) does not attempt to correct for these lags.

Nicholson (15) included viral and clinical surveillance data in a model of mortality in England and Wales using a 4-week timescale. This model, built on earlier models by Tillett et al. (16), used dummy variables for some time periods (but not others) to control for seasonal trends. Interpretation of the results is complicated by the complexity of the model (27 independent variables overall).

Sophisticated time-series methods attempt to separate seasonal components, by use of autoregressive integrated moving average models, for example (17), before applying regression. Carrat and Valleron (18) fit such models to time series of noninfluenza deaths and modeled the residuals against influenza deaths, in an attempt to attribute the total number of deaths caused by influenza in France. The usefulness of this approach is limited by the unreliability of attribution of influenza deaths on death certificates, however (19). Recently, increasingly sophisticated methods including generalized additive modeling (20), hierarchical modeling (21), and distributed-lag modeling (22) have been applied to mortality time series to assess the impact of air pollution on deaths (21, 22), with the work of Braga et al. (20) explicitly attempting to control for epidemics of respiratory disease. Removing seasonal confounding remains a complex issue, however (22). Further, influenza surveillance data for this time period at daily, weekly, or even monthly scales are not available in the public domain.

In this paper, we take a robust, conservative approach to the problem of seasonal confounding by investigating the relationship between influenza virus surveillance data and mortality at an annual timescale. We aggregate excess deaths and prevalence measures over an entire influenza season and estimate the number of deaths attributable to influenza infection in the United States. Additionally, we applied our method to two regions of the United States to investigate the contribution of cold temperatures to seasonal excess mortality on an annualized basis.

**MATERIALS AND METHODS**

Using public-use mortality data files for the United States, obtained from the Centers for Disease Control and Prevention’s National Center for Health Statistics, we created a database of all US deaths from 1979 to 2001 that can be queried by age, gender, underlying cause of death, and contributing cause of death, among other variables. Using this database, we generated several monthly time series of US mortality (available at http://mortality.princeton.edu/annualfu/) by use of different disease categories from the International Classification of Diseases, Ninth Revision and Tenth Revision, that were based on either the underlying cause of death or the underlying and contributing cause of death (table 1). We divide the calendar into “flu years,” each containing a single influenza season, that run from July to June. Deaths were corrected multiplicatively to a standard month length of 1,461/48 ≈ 30.44 days.

**TABLE 1. Mean monthly deaths from 1979 to 2001 for the US mortality time series investigated in this study**

<table>
<thead>
<tr>
<th>Name</th>
<th>All deaths</th>
<th>Definition</th>
<th>Deaths/month</th>
</tr>
</thead>
<tbody>
<tr>
<td>All deaths</td>
<td>All deaths</td>
<td>Any ICD-9* cause 480–487, any ICD-10* cause J10–J18</td>
<td>181,721</td>
</tr>
<tr>
<td>Pneumonia and influenza deaths</td>
<td>Any ICD-9 cause 390–519, any ICD-10 category I or J</td>
<td>16,405</td>
<td></td>
</tr>
<tr>
<td>Respiratory and circulatory deaths</td>
<td>Underlying ICD-9 cause 480–487, underlying ICD-10 cause J10–J18</td>
<td>129,275</td>
<td></td>
</tr>
<tr>
<td>Underlying pneumonia and influenza deaths</td>
<td>Underlying ICD-9 cause 390–519, underlying ICD-10 category I or J</td>
<td>5,849</td>
<td></td>
</tr>
<tr>
<td>Underlying respiratory and circulatory deaths</td>
<td>Underlying ICD-9 cause 390–519, underlying ICD-10 category I or J</td>
<td>95,313</td>
<td></td>
</tr>
</tbody>
</table>

Estimates based on mortality data depend on the consistency of mortality reports. This study takes a novel approach within the influenza epidemiology literature to the categorization of causes of death. Instead of using only the listed “underlying” cause of death, we tabulate all deaths with a specific cause—either pneumonia and influenza causes or all respiratory and circulatory causes—listed anywhere on the death certificate. We expect these “multiple-cause” data series to be more robust, both because more deaths are included and because considering all causes on the death certificate should make these series less sensitive to coding and nosologic changes (figure 1).

Annual virologic surveillance data (aggregated by influenza season) for the United States were obtained from Thompson et al. (11) and Morbidity and Mortality Weekly Report (23, 24). Following Thompson et al., we use the proportion of tested samples positive as a proxy for prevalence of each influenza subtype. Although this proportion is an imperfect measure of prevalence, we feel that it is the best influenza-specific measure available. The model of Thompson et al. makes use of weekly virus surveillance data, but these data have not been made available in the public domain.

To remove long-term trends from our mortality data, we fit a least-squares quadratic model to the time series of all deaths in each age category studied and divided the number of deaths in each category by this quadratic fit to all deaths to obtain “normalized” deaths by month and category. This form of normalization is an alternative to dividing by population estimates, and it has the advantages of being robust to changes in population estimation methods and of allowing us to conveniently divide our population by age and region.

Annual excess deaths were defined as the total number of deaths in a selected set of “influenza months” above the monthly average of a disjoint set of “baseline months.” For the main analysis, we used November to April as the flu months and the remaining 6 months as the baseline months (figure 2).

The main statistical model used was as follows:

$$NAD = \alpha + \beta_H H3 + \beta_{H1} H1 + \beta_B B,$$

where “NAD” is any normalized annual death series; “H3,” “H1,” and “B” refer to the proportions of samples positive for influenza A subtypes H3N2 and H1N1 and influenza B; $\alpha$ is the intercept; and the $\beta$s are regression coefficients.

Regional studies used deaths from 16 large counties in the New York metropolitan area, linked to temperature data from Newark International Airport, and deaths from the states of Illinois and Indiana, linked to temperature data from O’Hare International Airport. We chose regions that are densely populated, were expected to have similar weather across the region, and had complete monthly temperature series available at the International Research Institute for Climate Prediction (25). Cold was measured on a monthly basis as the average number of degrees below a threshold temperature ($T_{\text{thresh}}$) or cold ($C = \min(T_{\text{thresh}} - T, 0)$, where $T$ is the mean monthly temperature. Cold was measured for each flu year as the average of monthly cold values from November through April. The statistical model used was as follows:

$$NAD = \alpha + \beta_H H3 + \beta_{H1} H1 + \beta_B B + \beta_C C.$$

Because virus surveillance data were not available with age structure or regional breakdowns, we used national virus surveillance data for all regressions involving surveillance data.
Statistical analyses were performed using the statistical package R (26). Deaths attributed to an independent variable in a regression model were calculated by the difference between the number of normalized deaths predicted by the regression model and the number predicted when the independent variable is set to zero. Normalized deaths attributed in each year were then multiplied by the appropriate normalization factor. Because this method assumes that each cofactor has the same effect each year (e.g., no strain of a given subtype is more virulent than any other), estimates of attributable deaths in a particular season are less reliable than the estimates of average annual deaths over the entire time period. R scripts used in the analysis are available at http://mortality.princeton.edu/annualflu/.

RESULTS

We find positive correlations between excess deaths and the annual prevalence of each of the three influenza subtypes, for each category of deaths that we study. Using these correlations, we can estimate the average number of deaths in each category caused by influenza over a 20-year period (table 2) (Web table 1). (This latter tabular material is described in the first of seven supplementary tables; each is referred to as “Web table” in the text and is posted on the Journal’s website (http://aje.oxfordjournals.org/).) The regression model attributes an annual average of 41,400 (95 percent confidence interval: 27,100, 55,700) deaths to influenza over the period 1979–2001.

To investigate the effects of cold, we compared death records for two regions of the United States with winter temperature records from these regions. We found associations between normalized excess regional pneumonia and influenza deaths and the national prevalence of H3 and B viruses in both regions (table 3). Pneumonia and influenza deaths were used with these smaller samples because they showed more robust correlations with surveillance data in all analyses. No consistent relations were found between various combinations of monthly mean temperatures and normalized excess deaths, however, whether or not nationwide influenza virus surveillance data were used in the regression models (the example of cold over the whole winter season is shown in table 3).

Confidence intervals on the number of deaths attributed to cold weather are large, so we cannot conclude that influenza is a more important cause of winter mortality on an annual timescale than is cold weather. The result is suggestive, however, because the weather data used are more localized and more precise than the influenza surveillance data, so we

FIGURE 2. A normalized, detrended annual time series of excess deaths from monthly mortality data, United States, 1979–2001. Respiratory and circulatory (R&C) deaths for the whole population are used in this example. In A, a quadratic is fit to all deaths for a given age class (or the whole population); in B, deaths from a given cause are normalized by dividing by this quadratic fit; in C and D, normalized annualized winter deaths are defined as normalized deaths in “influenza” months above the monthly average for “baseline” months in the same season.
might have expected more statistical power in weather analyses. Results in table 3 are for $T_{\text{thresh}} = 10^\circ C$. Results setting $T_{\text{thresh}}$ to the mean temperature for each month, or to $20^\circ C$, are similar, but with much smaller and much larger confidence intervals, respectively.

We performed a sensitivity analysis to test our expectation that attributing deaths using annualized regression would be relatively insensitive to our choice of influenza months and baseline months, by examining several different such choices (Web tables 2, 3, 4, 5, and 6). For all choices, each of the 15 estimates of deaths in a category attributed to a subtype was positive, and all produced estimated totals within 10 percent of the estimate in the main text.

Over the time span that we studied, the average surveyed prevalence of influenza increases through time, possibly because of better prescreening or improved test sensitivity. We constructed normalized surveillance time series, by fitting a linear trend through total measured influenza prevalence and dividing the surveillance observations for H3, H1, and B for each year by this trend line, thus removing overall prevalence trends. These normalized data showed a qualitatively similar pattern, with significantly positive regression coefficients for two of the three subtypes, and predicted a total number of influenza deaths around 90 percent of the number predicted by the raw prevalence estimates (Web table 7).

If, instead of using normalized excess deaths, we apply our regression model to Serfling estimates generated by the method of Simonsen et al. (14), the model attributes 33,100 (95 percent confidence interval: 20,500, 45,600) annual deaths to influenza infection.

**DISCUSSION**

The total number of deaths attributed to influenza by our surveillance model (41,400 per year from 1979 to 2001) is similar to that found in other studies (11, 13, 14). The fact that our estimate of the mortality burden of influenza, based on our conservative methodology, is similar to earlier estimates provides support for classical assumptions about the role of influenza in seasonal excess mortality. Our estimates are not consistent, however, with much lower estimates from authors who attribute most excess winter deaths to cold (3, 27) or with much higher numbers that can be inferred from cohort studies (2, 4–6, 14).

Although our results are consistent with other ecologic estimates, the conservative, annualized method used here increases the certainty that deaths attributed to influenza are in fact caused by influenza. Methods based on seasonal pattern begin from the assumption that influenza is the major source of excess winter death. Daily, weekly, and monthly regressions are powerful, but they are in danger of being confounded by other seasonal factors.

The annualized approach also measures effects at a different timescale than daily or weekly regressions. A person

### TABLE 3. Annual pneumonia and influenza deaths, with 95% confidence intervals, categorized by causes appearing anywhere in the death record, United States, 1979–2001*

<table>
<thead>
<tr>
<th>Cause</th>
<th>New York metropolitan area</th>
<th>Illinois and Indiana</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of deaths</td>
<td>95% confidence interval</td>
</tr>
<tr>
<td>H3N2</td>
<td>1,492</td>
<td>361, 2,624</td>
</tr>
<tr>
<td>H1N1</td>
<td>-88</td>
<td>-560, 384</td>
</tr>
<tr>
<td>B</td>
<td>774</td>
<td>-21, 1,571</td>
</tr>
<tr>
<td>Cold</td>
<td>1,640</td>
<td>-1,815, 5,097</td>
</tr>
<tr>
<td>Total</td>
<td>3,819</td>
<td>66, 7,572</td>
</tr>
</tbody>
</table>

* Threshold temperature is 10°C.
whose death in a given week is triggered by influenza or cold weather may be somebody who would otherwise have lived for many more years or somebody who was unlikely to survive the winter anyway. Deaths in the latter category associated with causes by time-series analysis are often referred to as the “harvesting effect,” meaning that a particular cause, such as cold or influenza, may be “harvesting” people with very short life expectancies, determining the exact time of death but shortening lives by only a small amount (28). Weekly analyses measure deaths in both categories, whereas an annualized analysis will omit most deaths in the latter category. Measuring at different timescales asks fundamentally different questions: Neither answer is more correct than the other. The annualized approach presented here is, however, a substantially more conservative way of asking whether influenza triggers a substantial proportion of seasonal excess deaths.

By making use of multiple-cause death information, we have more than doubled the number of pneumonia and influenza deaths available for analysis and made our analysis less sensitive to coding choices. We suggest that, where multiple-cause death data are available, all pneumonia and influenza deaths and all respiratory and circulatory deaths may be useful time series for investigating influenza-caused mortality.

On an annual timescale, we find preliminary evidence that cold weather does not predict winter deaths. A full-scale spatial analysis of influenza mortality, ideally incorporating spatial surveillance data, remains to be done. If this result holds up, however, it does not necessarily contradict the large body of evidence relating cold stress to mortality (27, 29, 30). It is likely that temperature determines the timing of deaths on a shorter timescale; that is, people whose deaths are triggered by a particular cold spell may be people who are unlikely to survive until June. It is also possible that cold weather has indirect effects on mortality, by accelerating the spread of influenza (1, 31).

We believe that we have made a simple, robust, and conservative model of influenza deaths. The price, however, is that our model is rather crude. The fact that we have only one data point per year limits the power and precision of our analysis. Our surveillance-based regression model does not allow for the fact that different strains of a given subtype may be more or less deadly from year to year. Additionally, virus surveillance by the Centers for Disease Control and Prevention is based largely on contributed samples and therefore does not reflect a systematic sampling scheme (32). On at least one occasion, the 1985–1986 season, the epidemic in the total US population was largely dominated by the B virus, while most cases in the elderly, and thus influenza-triggered deaths, were due to the more lethal influenza A(H3) virus (33).

In summary, a conservative, annualized analysis finds evidence that influenza virus infections are causing mortality on an annual timescale. Our estimates support earlier estimates of influenza burden based on direct excess deaths (13, 14) and weekly regressions (11), but they do not support higher estimates based on cohort studies (4–6). Our results suggest that the triggering of deaths by cold weather (3) may be a short-term phenomenon that has much less impact on an annual timescale. Further analyses based on disaggregated, systematic viral surveillance data have the potential to shed further light on these questions, as such data become publicly available.

ACKNOWLEDGMENTS

The authors thank Dr. Mark Miller for helpful conversations.
Conflict of interest: none declared.

REFERENCES
