Time-Series Analyses of Count Data to Estimate the Burden of Seasonal Infectious Diseases

William W. Thompson,[a] *Benjamin L. Ridenhour,*[b] *John P. Barile,*[c] *and David K. Shay*[d]

In this issue of EPIDEMIOLOGY, Goldstein and colleagues[1] present "a new statistical method for estimating influenza-related mortality" using recent US mortality and influenza data sources. They propose that their method is an advance over previous methods[2–13] for estimating influenza disease burden. Strengths of their approach include: (1) the accommodation of nonlinear mortality effects of specific influenza types and subtypes over time; (2) the use of a logical method to model exposure to influenza; and (3) an assessment of autocorrelation in the time series. There are also some limitations, which include (1) questionable statistical assumptions in the use of linear regression models for weekly data with low frequency counts; (2) neglect of recent advances in time-series modeling; (3) a complex and nonparsimonious model; and (4) lack of comparisons of influenza-associated mortality estimates with those from the published literature using similar outcomes.

A BRIEF HISTORY

Indirect statistical methods have long been used to estimate the number of deaths associated with influenza, often making use of vital statistics. Estimates of influenza-associated deaths were first made in 1847 in England by William Farr[14] and were subsequently made for the 1892 Massachusetts pandemic and 1918 pandemic[15]—decades before actual human influenza viruses were isolated.[16] For many years, investigators relied on indirect methods because they lacked inexpensive, sensitive, and widely available diagnostic tests for influenza infection. More recently, reverse transcription polymerase chain reaction (RT-PCR) assays provide a practical gold standard for detecting infections with influenza and other respiratory virus infections.[17,18] Recent studies have estimated influenza-associated hospitalizations on a population-level through prospective enrollment of hospitalized subjects followed by RT-PCR testing of respiratory specimens.[19,20] Still, the availability of molecular diagnostics has not made time-series methods obsolete. Influenza-associated deaths will always be difficult to count. Older adults, who are at greatest risk of influenza-associated mortality, may not die from the initial assault of the virus but from later bacterial superinfection, or from exacerbation of chronic respiratory or cardiac conditions. Such deaths can occur a week or two after the acute infection, by which time the virus is no longer replicating.[21,22]

GENERAL MODEL STRATEGIES

There are three broad classes of models to estimate deaths associated with infectious agents. The etiologic fraction method takes the proportion of specific infection among proximal outcomes and applies this to an outcome for which it is more difficult to obtain etiologic data, such as mortality.[23] (For example, the proportion of all pneumonia deaths

From the [a]Division of Behavioral Surveillance, Office of Surveillance, Epidemiology and Laboratory Services, Centers for Disease Control and Prevention, Atlanta, GA; [b]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN; University of Hawaii at Mānoa, Mānoa, HI; and [d]Influenza Division, Centers for Disease Control and Prevention.

Editors' note: *Related articles appear on pages 829 and 843.*

Correspondence: William W. Thompson, CDC, 1600 Clifton Rd, MS E-61, Atlanta, GA 30333. E-mail: wct2@cdc.gov.

ISSN:1044-3983/12/2306-0839
DOI:10.1097/EDE.0b013e31826cc1df

due to pneumococcus can be estimated using the proportion of all pneumonia hospitalizations due to pneumococcus.)[24] The obvious question is whether the agent of interest infects similar proportions of those hospitalized and of those dying. For influenza, this remains unknown.

A second class of models is the probabilistic multiplier approach,[25] which estimates the mortality rate as the product of the symptomatic attack rate and the risk of death given infection (case-fatality rate).[26] Data of this type are seldom available, limiting this model's main use to influenza pandemics or other epidemics that attract the needed resources to estimate the needed parameters.

The third class of models, and the one most frequently used in influenza studies, is time-series regression. The model proposed by Goldstein and colleagues[1] is an example. Time-series models have been used for at least five decades to estimate influenza-related deaths in the United States and other temperate countries with relatively consistent, well-defined periods of influenza circulation during winter months.[11,27–29] We would suggest that the time-series approach of Goldstein and colleagues does not take into account some of the advances in time-series modeling as applied to infectious diseases.

The best-known of the time-series models is the original Serfling regression model.[30] This model includes cyclical regression terms (sine and cosine terms) to account for the obvious seasonality of mortality in the United States. Both the initial model and subsequent iterations predict that deaths exceeding a seasonally adjusted baseline are all influenza-associated, given that most if not all of the excess deaths occurred during the same time period that influenza viruses were circulating. Serfling proposed to do this by setting death counts in particular weeks determined to be significant outliers to an estimated baseline constructed after omitting those data points from the original time series. Variations in the original Serfling regression model have been driven largely by the type and quality of data available. Time-series methods have been used to estimate influenza burden for a wide range of health outcomes, including outpatient visits,[31,32] hospitalizations,[4,7,31,32] and mortality.[2,3,8,30,33,34]

One of the most important advances in modeling influenza-associated deaths was made by Clifford and colleagues.[35,36] They proposed an alternative to the Serfling model; in the Clifford model, laboratory-diagnosed virus infections and influenza diagnosis rates were used in a stepwise regression analysis that also included terms for secular trends, seasonal patterns, air temperature, and number of years since a major change was detected in the influenza virus hemagglutinin antigen. The Clifford model was novel in that it did not impute the mortality baseline. Rather, it estimated excess deaths by subtracting the predicted baseline values (with the influenza terms set to zero) from the estimated values during the same time period when influenza viruses were circulating. The Clifford model was subsequently adopted by Alling and colleagues[37,38] and used to estimate excess influenza deaths in the United States.

During the late 1990s, weekly US data became available for the numbers of respiratory specimens tested, and the numbers of samples that tested positive for influenza and respiratory syncytial virus (RSV). These data provided a new opportunity to model influenza-associated deaths for both influenza and RSV simultaneously, using methods that directly account for the temporal identification of virus activity. Several research groups began to explore such models. For example, quasi-Poisson regression models were applied to US viral surveillance data for influenza A(H1N1), B, A(H3N2), and RSV viruses to estimate burden of pneumonia and influenza deaths, respiratory and circulatory deaths, and all-cause mortality.[8] This general methodology was subsequently applied to hospitalizations and other health outcomes[7,39] and compared with other burden of disease models and estimates.[9] Time-series methods have also provided an opportunity to add other important covariates and confounders into the models, including time-varying exposure models for air pollution effects on mortality.[40,41,42] These models are increasingly used in regions with more complex seasonal influenza patterns.[27,43] Based on this biostatistics and influenza literature, it appears that Goldstein and colleagues[1] might have missed an opportunity to provide stepping stones to the next iteration of influenza burden of disease methods. It is difficult to evaluate how likely violations of statistical assumptions might have affected their results. This assessment is complicated by the use of different outcome measures in Goldstein's models relative to the previous literature—thus precluding direct historical comparisons of results from other studies.

MODELING LOW-FREQUENCY-COUNT DATA

Our second concern with the approach proposed by Goldstein et al[1] is that it does not capitalize on the advances in time-series methods for low-frequency-count data. Their approach instead uses linear regression models for outcomes likely to have very low frequencies. An important advance in generalized linear models in the early 1970s[44] allows for more complicated forms of regression modeling that relax the assumption of linearity between predictor and response variables. Generalized linear modeling permits the expected value of the response variable to be a smoothed (eg, nonlinear) monotonic function of the linear predictors. Generalized linear modeling also relaxes the assumption of normal distribution for the response variable, allowing any distribution from the exponential family. For count data, means and variances can be modeled by discrete probability distributions, the most common being the Poisson, binomial, negative binomial, and geometric distributions. Poisson regression is a generalized linear model with Poisson-distributed response and a natural logarithm for the link function. This has become a common choice for epidemiologic models because it is well suited to low-frequency-count data like mortality.[40,41,45,46] Although the choice of the link function used in generalized linear modeling is somewhat arbitrary, the function should nonetheless support the same range of values as the probability distribution (eg,

a Poisson distribution is supported on [0, ∞], so a good link function would have a domain of [0, ∞], such as the natural logarithm function).

Poisson distributions are themselves limited by having a mean equal to the variance. This assumption is frequently violated when working with low-frequency-count data, and thus quasi-Poisson or negative binomial regression techniques are widely used in time-series studies using death certificate data.[43,47–49] In these models, a second parameter is used to account for either overdispersion (variance > mean) or underdispersion (variance < mean). Other models—including zero-inflated,[50] hurdle,[51] finite mixture,[52] generalized estimating equations,[45] mixed/hierarchical generalized linear models,[53] and generalized additive models[54]—can also provide flexible variance structures when working with mortality data. When modeling mortality data from national vital records, quasi-Poisson or negative binomial estimation techniques have statistical advantages (eg, fewer assumptions required) as well as practical advantages (interpretability) over standard linear regression. The limitations of traditional linear regression in estimating count-derived data have been noted in many fields.[55–60] A serious limitation in the approach of Goldstein et al[1] is their assumption that mortality count data can be adequately modeled with a normal probability distribution and corresponding error terms.

MODEL COMPLEXITY

Our third concern is with regard to parsimony. Goldstein et al[1] have chosen to fit a relatively complex model that has many underlying statistical and biological assumptions. The authors apply a very specific model (eg, modeling changes in the effects of influenza A(H3N2) across a very short time period. This is likely to be too short a period for modeling such differences and is unlikely to be generalizable to other modeling efforts. Goldstein and colleagues also choose to include an autocorrelation term between subsequent observations. This is a common technique in many time-series applications (e.g. market analysis), but based on our own experience, the addition of autocorrelated terms reduces estimates substantially because the model absorbs some of the variance likely associated with influenza. Goldstein and colleagues suggest that, even though their model is different (and more complex), it results in estimates similar to those based on less complex models. If parsimony is a virtue, one might then prefer instead to select the less complex model with fewer assumptions, at least until arguments of biological plausibility supporting such complexity could be developed from other data.

MODEL VALIDITY

Our last concern is with the suggestion that the proposed methodology is an advance over previous approaches. It is a truism that "all models are wrong; some are useful."[61] Most burden of disease estimates have been consistent regardless of statistical method, showing substantial season-to-season variations in influenza-associated deaths.[9] Given that evidence of recent influenza infection is undetectable in most persons dying during peak influenza circulation, there is no gold standard for determining a "best model." It is therefore important for researchers proposing new models to make direct comparisons of their results with those in the literature.

In sum, we believe most epidemiologists would agree on certain guidelines in model development: (1) use a model type suitable for the distribution and error structure of the data; and (2) invest resources in validation and cross-validation techniques that can help verify whether indirect modeling is both internally and externally consistent. The full contribution of the model proposed by Goldstein and colleagues is difficult to assess before such information is in hand.

ABOUT THE AUTHORS

WILLIAM THOMPSON is the Associate Director of Science for the Division of Behavioral Surveillance in the Office of Surveillance, Epidemiology and Laboratory Services, CDC. His interests include influenza modeling and behavioral risk factor research. BENJAMIN RIDENHOUR is an Assistant Professor in the Department of Biological Sciences at Notre Dame. His interests include biostatistical and biomathematical modeling of disease systems. JOHN BARILE is an Assistant Professor at the University of Hawaii at Mānoa. His interests include the study of environmental contributors of health. DAVID SHAY is a medical epidemiologist with CDC's Influenza Division. His interests include burden of disease models for viral respiratory pathogens and vaccine effectiveness studies.

REFERENCES

1. Goldstein E, Viboud C, Charu V, Lipsitch M. Improving the estimation of influenza-related mortality over a seasonal baseline. *Epidemiology*. 2012;23:829–838.
2. Simonsen L, Clarke MJ, Stroup DF, Williamson GD, Arden NH, Cox NJ. A method for timely assessment of influenza-associated mortality in the United States. *Epidemiology*. 1997;8:390–395.
3. Simonsen L, Clarke MJ, Williamson GD, Stroup DF, Arden NH, Schonberger LB. The impact of influenza epidemics on mortality: introducing a severity index. *Am J Public Health*. 1997;87:1944–1950.
4. Simonsen L, Fukuda K, Schonberger LB, Cox NJ. The impact of influenza epidemics on hospitalizations. *J Infect Dis*. 2000;181:831–837.
5. Reichert TA, Simonsen L, Sharma A, Pardo SA, Fedson DS, Miller MA. Influenza and the winter increase in mortality in the United States, 1959-1999. *Am J Epidemiol*. 2004;160:492–502.
6. Viboud C, Miller M, Olson D, Osterholm M, Simonsen L. Preliminary Estimates of Mortality and Years of Life Lost Associated with the 2009 A/H1N1 Pandemic in the US and Comparison with Past Influenza Seasons. *PLoS Curr*. 2010;2:RRN1153.
7. Thompson WW, Shay DK, Weintraub E, et al. Influenza-associated hospitalizations in the United States. *JAMA*. 2004;292:1333–1340.
8. Thompson WW, Shay DK, Weintraub E, et al. Mortality associated with influenza and respiratory syncytial virus in the United States. *JAMA*. 2003;289:179–186.
9. Thompson WW, Weintraub E, Dhankhar P, et al. Estimates of US influenza-associated deaths made using four different methods. *Influenza Other Respi Viruses*. 2009;3:37–49.
10. Schanzer DL, Tam TW, Langley JM, Winchester BT. Influenza-attributable deaths, Canada 1990-1999. *Epidemiol Infect*. 2007;135:1109–1116.
11. Yang L, Ma S, Chen PY, et al. Influenza associated mortality in the subtropics and tropics: results from three Asian cities. *Vaccine*. 2011;29:8909–8914.

12. 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;138:836–842.
13. van Asten L, van den Wijngaard C, van Pelt W, et al. Mortality attributable to 9 common infections: significant effect of influenza a, respiratory syncytial virus, influenza B, norovirus, and parainfluenza in elderly persons. *J Infect Dis*. 2012;206:628–639.
14. Farr W. *Annual Report of the Registrar General*. In: HMSO, ed. London, 1847.
15. Dauer CC, Serfling RE. Mortality from influenza: 1957–1958 and 1959–1960. *Am Rev Respir Dis*. 1961;83:15–28.
16. Smith W, Andrewes C, Laidlaw P. A virus obtained from influenza patients. *Lancet*. 1933;2:66–68.
17. Templeton KE, Scheltinga SA, Beersma MF, Kroes AC, Claas EC. Rapid and sensitive method using multiplex real-time PCR for diagnosis of infections by influenza a and influenza B viruses, respiratory syncytial virus, and parainfluenza viruses 1, 2, 3, and 4. *J Clin Microbiol*. 2004;42:1564–1569.
18. Weinberg GA, Erdman DD, Edwards KM, et al; New Vaccine Surveillance Network Study Group. Superiority of reverse-transcription polymerase chain reaction to conventional viral culture in the diagnosis of acute respiratory tract infections in children. *J Infect Dis*. 2004;189:706–710.
19. Jules A, Grijalva CG, Zhu Y, et al. Estimating age-specific influenza-related hospitalization rates during the pandemic (H1N1) 2009 in Davidson Co, TN. *Influenza Other Respi Viruses*. 2012;6:e63–e71.
20. Poehling KA, Edwards KM, Weinberg GA, et al; New Vaccine Surveillance Network. The underrecognized burden of influenza in young children. *N Engl J Med*. 2006;355:31–40.
21. Taubenberger JK, Morens DM. The pathology of influenza virus infections. *Annu Rev Pathol*. 2008;3:499–522.
22. Collins SD, Lehmann J. Trends and epidemics of influenza and pneumonia: 1918-1951. *Public Health Rep*. 1951;66:1487–1516.
23. Tate JE, Burton AH, Boschi-Pinto C, Steele AD, Duque J, Parashar UD; WHO-coordinated Global Rotavirus Surveillance Network. 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. *Lancet Infect Dis*. 2012;12:136–141.
24. O'Brien KL, Wolfson LJ, Watt JP, et al; Hib and Pneumococcal Global Burden of Disease Study Team. Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. *Lancet*. 2009;374:893–902.
25. Simonsen L, Viboud C. The art of modeling the mortality impact of winter-seasonal pathogens. *J Infect Dis*. 2012;206:625–627.
26. Dawood FS, Iuliano AD, Reed C, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *Lancet Infect Dis*. 2012.
27. Chow A, Ma S, Ling AE, Chew SK. Influenza-associated deaths in tropical Singapore. *Emerging Infect Dis*. 2006;12:114–121.
28. Kyncl J, Prochazka B, Goddard NL, et al. A study of excess mortality during influenza epidemics in the Czech Republic, 1982-2000. *Eur J Epidemiol*. 2005;20:365–371.
29. Thompson MG, Shay DK, Zhou H, et al. Estimates of deaths associated with seasonal influenza—United States, 1976–2007. *MMWR Morb Mortal Wkly Rep*. 2010;59:1057–1062.
30. Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep*. 1963;78:494–506.
31. Izurieta HS, Thompson WW, Kramarz P, et al. Influenza and the rates of hospitalization for respiratory disease among infants and young children. *N Engl J Med*. 2000;342:232–239.
32. O'Brien MA, Uyeki TM, Shay DK, et al. Incidence of outpatient visits and hospitalizations related to influenza in infants and young children. *Pediatrics*. 2004;113(3 Pt 1):585–593.
33. Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962-1979. I. Time series forecasts of expected pneumonia and influenza deaths. *Am J Epidemiol*. 1981;113:215–226.
34. Choi K, Thacker SB. An evaluation of influenza mortality surveillance, 1962-1979. II. Percentage of pneumonia and influenza deaths as an indicator of influenza activity. *Am J Epidemiol*. 1981;113:227–235.
35. Clifford RE, Smith JW, Tillett HE, Wherry PJ. Excess mortality associated with influenza in England and Wales. *Int J Epidemiol*. 1977;6:115–128.
36. Tillett HE, Smith JW, Clifford RE. Excess morbidity and mortality associated with influenza in England and Wales. *Lancet*. 1980;1:793–795.
37. Alling DW, Blackwelder WC, Stuart-Harris CH. A study of excess mortality during influenza epidemics in the United States, 1968-1976. *Am J Epidemiol*. 1981;113:30–43.
38. Blackwelder WC, Alling DW, Stuart-Harris CH. Association of excess mortality from chronic nonspecific lung disease with epidemics of influenza. Comparison of experience in the United States and in England and Wales, 1968 to 1976. *Am Rev Respir Dis*. 1982;125:511–516.
39. Zhou H, Thompson WW, Viboud CG, et al. Hospitalizations associated with influenza and respiratory syncytial virus in the United States, 1993-2008. *Clin Infect Dis*. 2012;54:1427–1436.
40. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42:121–130.
41. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44:1049–1060.
42. Dominici F, McDermott A, Zeger SL, Samet JM. On the use of generalized additive models in time-series studies of air pollution and health. *Am J Epidemiol*. 2002;156:193–203.
43. Feng L, Shay DK, Jiang Y, et al. Influenza-associated mortality in temperate and subtropical Chinese cities, 2003–2008. *Bull World Health Organ*. 2012;90:279–288B.
44. Nelder J, Wedderburn. Generalized linear models. *J R Stat Soc Ser A*. 1972;135:370.
45. Liang K, Zeger S. Longitudinal data-analysis using generalized linear-models. *Biometrika*. 1986;73:13–22.
46. Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. *Stat Med*. 1992;11:1825–1839.
47. Peng RD, Bobb JF, Tebaldi C, McDaniel L, Bell ML, Dominici F. Toward a quantitative estimate of future heat wave mortality under global climate change. *Environ Health Perspect*. 2011;119:701–706.
48. Katsouyanni K, Samet JM, Anderson HR, et al; HEI Health Review Committee. Air pollution and health: a European and North American approach (APHENA). *Res Rep Health Eff Inst*. 2009;142:5–90.
49. Gasparrini A, Armstrong B. The impact of heat waves on mortality. *Epidemiology*. 2011;22:68–73.
50. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992;34:1–14.
51. Mullahy J. Specification and testing of some modified count data models. *J Econometrics*. 1986;33:341–365.
52. McLachlan GJ, Peel D. *Finite Mixture Models*. New York: Wiley; 2000.
53. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. *J Am Stat Assoc*. 1993;88:9–25.
54. Hastie T, Tibshirani R. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*. 1990;46:1005–1016.
55. Afifi AA, Kotlerman JB, Ettner SL, Cowan M. Methods for improving regression analysis for skewed continuous or counted responses. *Annu Rev Public Health*. 2007;28:95–111.
56. Coxe S, West SG, Aiken LS. The analysis of count data: a gentle introduction to poisson regression and its alternatives. *J Pers Assess*. 2009;91:121–136.
57. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychol Bull*. 1995;118:392–404.
58. O'Hara RB, Kotze DJ. Do not log transform count data. *Methods in Ecology and Evolution*. 2010;1:118–122.
59. Winkelmann R, Zimmermann KF. Recent developments in count data modelling: Theory and application. *Journal of Economic Surveys*. 1995;9:1–24.
60. Yang Z, Hardin JW, Addy CL, Vuong QH. Testing approaches for overdispersion in poisson regression versus the generalized poisson model. *Biom J*. 2007;49:565–584.
61. Box GEP, Draper NR. *Empirical Model Building and Response Surfaces*. New York: Wiley; 1987.

