Epidemiologic Perspectives & Innovations

BioMed Central

Methodology

Open Access

Applying the compound Poisson process model to the reporting of injury-related mortality rates

Scott R Kegler*

Address: Office of Statistics and Programming, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta GA, USA

Email: Scott R Kegler* - skegler@cdc.gov

* Corresponding author

Published: 16 February 2007

Epidemiologic Perspectives & Innovations 2007, **4**:1 doi:10.1186/1742-5573-4-1

Received: 29 March 2006

Accepted: 16 February 2007

This article is available from: http://www.epi-perspectives.com/content/4/1/1

Abstract

Injury-related mortality rate estimates are often analyzed under the assumption that case counts follow a Poisson distribution. Certain types of injury incidents occasionally involve multiple fatalities, however, resulting in dependencies between cases that are not reflected in the simple Poisson model and which can affect even basic statistical analyses. This paper explores the compound Poisson process model as an alternative, emphasizing adjustments to some commonly used interval estimators for population-based rates and rate ratios. The adjusted estimators involve relatively simple closed-form computations, which in the absence of multiple-case incidents reduce to familiar estimators based on the simpler Poisson model. Summary data from the National Violent Death Reporting System are referenced in several examples demonstrating application of the proposed methodology.

Introduction

Injury-related mortality rate estimates are often analyzed under the assumption that case counts follow a Poisson distribution. [1-4] Certain types of injury incidents occasionally involve multiple fatalities, however, resulting in dependencies between cases that are not reflected in the simple Poisson model and which can affect even elementary analyses of rates. This paper examines the application of the compound Poisson process model [5-7] to address this issue, emphasizing adjustments to some commonly used interval estimators for rates and rate ratios. Accompanying examples demonstrate the proposed adjustments and provide comparisons of results obtained under the Poisson and compound Poisson process models.

This paper was motivated by the need for basic statistical methods applicable to data from the National Violent Death Reporting System (NVDRS).[8] The NVDRS data provide a census of violent deaths occurring in the states covered by the reporting system. Data are collected on incidents and persons (victims/suspects), and records for all persons associated with each incident are linked. Two types of incidents (not mutually exclusive) are of particular note in the present context: (i) those involving multiple homicides and (ii) those involving homicide followed by suicide. The NVDRS data for the year 2004 (covering 13 states) indicate that over 4% of homicide-related incidents involved multiple homicides. Correspondingly, at the individual (person) level approximately 9% of homicides were associated with multiple-homicide incidents. These data also show that nearly 12% of homicide-suicide incidents involved multiple (usually two) homicides, while at the individual level approximately 23% of homicides associated with a homicide-suicide incident were part of a multiple homicide-suicide incident.

Analysis

An Analysis Framework Based on the Compound Poisson Process Model

Analyses of vital statistics data often rely on a conceptual framework in which case counts, even though based on a census, are considered inherently variable.[1,2,4,9,10] In selected National Center for Health Statistics reports, for example, mortality rate estimates are evaluated for statistical stability by assuming that census-level case counts (rate numerators) follow a Poisson distribution while at-risk population estimates (rate denominators) are assumed constant.[1,2]

The simple Poisson model includes the assumption that cases occur independently. Cases (fatalities) associated with multiple-case incidents are not independent, however, and for this reason such a model does not adequately characterize the types of incidents described above. The compound Poisson process model [5-7] provides a closer conceptual parallel, by incorporating a two-level counting process. Applying this model to the NVDRS data, incident counts represent the first level and are assumed to follow a simple Poisson distribution. The counts of cases associated with each incident represent the second level. These incident-specific case counts are assumed to follow a common (discrete) probability distribution of unspecified form. Incident-specific case counts are also assumed to be (i) independent across incidents and (ii) independent of the count of incidents. [5-7]

As an illustration of the basic aspects of the compound Poisson process model, suppose that occurrences of a specific type of incident are consistent with a Poisson process having person-year rate parameter λ. Letting person-years at risk be denoted by P, it follows that the incident count N has a Poisson distribution with (unknown) mean λP, denoted by $N \sim \text{Poisson}(\lambda P)$. At the next level, suppose that the incident-specific case counts $C_1, C_2, C_3, \ldots, C_N$ have a common underlying distribution with mean μ and variance σ^2 (both generally unknown) and satisfy the independence assumptions specified above. The total case count $C \equiv \sum_{k=1}^{N} C_k$ then conforms to a compound Poisson process model, with underlying mean $E[C] = \lambda P \times \mu$ and underlying variance $\text{Var}(C) = \lambda P \times (\mu^2 + \sigma^2)$. [5-7]

Effectively, the compound Poisson process model reduces to the simple Poisson model in analyses where multiple-case incidents do not occur.[7] In such situations $C_k = 1$ for every incident, so that the total case count C is equal to the incident count N, with the latter variable assumed to follow a simple Poisson distribution. In this way, the framework based on the compound Poisson process model encompasses the more customary analysis framework based on the simple Poisson model.

Rate Estimators and Variances

Using terms defined above, the typical estimate of the population-based case rate per 100,000 person-years is provided by $R \equiv C/P \times 100{,}000$. Under the compound Poisson process model $E[R] = E[C]/P \times 100{,}000 = \lambda \times \mu \times 100{,}000$. Since this latter quantity also corresponds to the underlying case rate per 100,000 person-years, it follows that R is an unbiased estimator.

The variance of the rate estimator is $\text{Var}(R) = \text{Var}(C)/P^2 \times 100{,}000^2$. An unbiased estimate of Var(C) under the compound Poisson process model is conveniently provided by $\hat{\text{Var}}(C) = \sum_{k=1}^{N} C_k^2$ (see Appendix A). Substituting $\sum_{k=1}^{N} C_k^2$ in place of Var(C) provides the unbiased variance estimate $\hat{\text{Var}}(R) = \sum_{k=1}^{N} C_k^2 / P^2 \times 100{,}000^2$.

Confidence Intervals for Rates

Confidence intervals for rates and rate ratios frequently involve an initial logarithmic transformation to the point estimate. Applying this transformation to the rate estimator R, the "delta" method [11-14] suggests the following general form (not model-dependent) for an approximate 95% confidence interval for the underlying rate [15]:

$$\left(\exp\left\{ \ln(R) - 1.96 \times \sqrt{\frac{\hat{\text{Var}}(R)}{R^2}} \right\}, \ \exp\left\{ \ln(R) + 1.96 \times \sqrt{\frac{\hat{\text{Var}}(R)}{R^2}} \right\} \right). \quad (1)$$

When R is based on a total case count C that is assumed to follow a simple Poisson distribution, interval (1) simplifies to the more recognizable form [10,16,17]:

$$\left(\exp\left\{ \ln(R) - 1.96 \times \sqrt{\frac{1}{C}} \right\}, \ \exp\left\{ \ln(R) + 1.96 \times \sqrt{\frac{1}{C}} \right\} \right). \quad (1a)$$

Alternatively, when the total case count C is assumed to conform to a compound Poisson process model, substitution of the earlier expression for $\hat{\text{Var}}(R)$ into (1) provides the following adjustment to interval (1a) to account for multiple-case incidents:

$$\left(\exp\left\{ \ln(R) - 1.96 \times \sqrt{\frac{\sum_{k=1}^{N} C_k^2}{C^2}} \right\}, \ \exp\left\{ \ln(R) + 1.96 \times \sqrt{\frac{\sum_{k=1}^{N} C_k^2}{C^2}} \right\} \right) \quad (1b)$$

where $C_1, C_2, C_3, \ldots, C_N$ are the incident-specific case counts defined previously. By inspection of the square root terms in (1a) and (1b), it can be seen that the estimated variance of ln(R) is increased by the factor

$\sum_{k=1}^{N} C_k^2 / C$ under the compound Poisson process model. When no multiple-case incidents appear in the data it follows that $\sum_{k=1}^{N} C_k^2 = \sum_{k=1}^{N} C_k \equiv C$ (because C_k = 1 for each incident covered), whereupon interval (1b) reduces to interval (1a) and the distinction between models becomes academic.

Example 1

This example briefly demonstrates the calculations required for intervals (1a) and (1b), referencing NVDRS summary data for the year 2004. These data (for 13 states) show that there were 25 homicide-suicide incidents involving homicide victims under 21 years of age. The left half of Table 1 provides a summary of the incident-specific homicide counts associated with these incidents. For example, the first summary line represents 19 separate incidents, each involving one homicide victim under 21 years old. The right half of the table shows the calculation of the sums appearing in intervals (1a) and (1b).

The totals at the bottom of the right half of Table 1 are used to calculate the unadjusted and adjusted confidence interval estimates for the population-based rate. There were approximately 19.8 million persons under 21 years of age in the states covered by NVDRS for 2004.[18] Since the reporting period covered one year this translates to approximately 19.8 million person-years at risk for this age group, so the estimated rate per 100,000 person-years is:

$R = (31/19.8M) \times 100{,}000 \approx 0.157.$

Recalling the earlier shorthand notation $C \equiv \sum_{k=1}^{N} C_k$, the unadjusted 95% confidence interval for the rate is obtained by substituting $R \approx 0.157$ and C = 31 into (1a):

$$\left(\exp\left\{ \ln(0.157) - 1.96 \times \sqrt{\frac{1}{31}} \right\}, \ \exp\left\{ \ln(0.157) + 1.96 \times \sqrt{\frac{1}{31}} \right\} \right)$$

resulting in the interval estimate (0.110, 0.223).

The 95% confidence interval adjusted for multiple-case incidents is obtained by substituting the values $R \approx 0.157$, C = 31, and $\sum_{k=1}^{N} C_k^2 = 43$ into (1b):

$$\left(\exp\left\{ \ln(0.157) - 1.96 \times \sqrt{\frac{43}{31^2}} \right\}, \ \exp\left\{ \ln(0.157) + 1.96 \times \sqrt{\frac{43}{31^2}} \right\} \right).$$

The resulting interval estimate (0.104, 0.238) is approximately 19% wider than the unadjusted interval estimate.

Example 2

The coverage properties of the unadjusted and adjusted confidence intervals (1a) and (1b) were compared using stochastic simulation. Assuming that the Poisson distribution is an appropriate model for incident counts, the relevant simulation inputs are the underlying mean incident count λP and the underlying distribution of cases within incidents. The ranges of the simulation input values were selected in part to cover the observed values from Example 1.

The underlying mean incident count λP can be varied either through λ (the incident occurrence rate per person-year) or through P (person-years at risk) and for purposes of simulation the choice of which to vary is arbitrary. Therefore, P was held constant at 20 million person-years while λ was varied across the values 0.050×10^{-5}, 0.125×10^{-5}, 0.250×10^{-5}, and 0.500×10^{-5}. These values correspond to underlying mean incident counts of λP = 10, 25, 50, 100 for the hypothetical period of observation.

Each value of the incident occurrence rate λ was considered in combination with five different within-incident case count distributions. The within-incident case count distributions are denoted by quadruples (p_1, p_2, p_3, p_4) indicating the probabilities that an incident involves 1, 2, 3, or 4 cases of interest, respectively. The first within-incident distribution (0.76, 0.24, 0.00, 0.00) matches the observed distribution from Example 1 in which each incident involved one or two cases. The second within-incident distribution (0.95, 0.05, 0.00, 0.00) reflects a comparatively small fraction of incidents involving multiple cases. The remaining three distributions reflect a successively increasing fraction of incidents involving multiple cases as well as larger numbers of cases. Consistent with earlier notation, μ and σ^2 denote the mean and variance for each within-incident case count distribution.

The simulation was programmed using the Statistical Analysis System (SAS).[19] One-hundred thousand simulation replicates were generated for each combination of simulation inputs. Each replicate involved simulation of an incident count according to a Poisson distribution having the indicated mean λP, followed by simulation of incident-specific case counts according to the indicated discrete distribution. The 95% confidence intervals (1a) and (1b) were calculated for each simulation replicate, and for each interval estimate it was noted whether the underlying case rate per 100,000 person-years ($\lambda \times \mu \times 100{,}000$) fell between the interval endpoints. The relative frequencies with which intervals (1a) and (1b) covered

Table 1: NVDRS Summary Data and Calculations for a Rate Confidence Interval.

Incident Summary Data		Calculation of Sums	
Incident Count	Homicides <21 Years of Age	ΣC_k	ΣC_k^2
19	1	19 × 1 = 19	19 × 1^2 = 19
6	2	6 × 2 = 12	6 × 2^2 = 24
25		31	43

the true case rate for the various combinations of simulation inputs are displayed in Tables 2 and 3, respectively.

The results in Table 2 indicate that as incidents involving multiple cases and larger numbers of cases become more frequent, coverage for the unadjusted interval (1a) drops well below the nominal 95% level. In the first line of the table (paralleling the NVDRS data from Example 1) coverage falls to about 90%. In the second line, where multiple-case incidents are less common, the difference between effective and nominal coverage is minimal. In the fifth line of the table, representing the most extreme departure from one case per incident, coverage is reduced to just over 85%. In contrast, the results in Table 3 show that coverage for adjusted interval (1b) conforms closely to the nominal 95% level under all of the simulation parameter combinations considered.

Further simulation results (not shown) indicate that the performance of interval (1b) is sensitive to extra-Poisson variability in the incident counts. Although more general models compensating for such extra-Poisson variability might be considered, the primary interest here is the effect of multiple-case incidents. Moreover, multiple-case incidents are unmistakable when represented in the data, whereas extra-Poisson variability in the incident counts may be more difficult to detect. The remaining presentation therefore continues to rely on the compound Poisson process model, which adequately addresses the issue of multiple-case incidents and results in tractable computational expressions.

Confidence Intervals for Rate Ratios

The analysis is somewhat more complicated when considering rate ratios as opposed to individual rates. Letting R_{S1} and R_{S2} denote rate estimators for two demographic subgroups (for example, persons under 21 years of age and persons 21 years of age or older) the estimated rate ratio is defined in the usual way as $RR \equiv R_{S1}/R_{S2}$. Applying a logarithmic transformation to this ratio, the delta method [11-14] suggests the following general form for an approximate 95% confidence interval for the underlying rate ratio:

$$\left(\exp\left\{ \ln(RR) - 1.96 \times \sqrt{\frac{\hat{V}ar(R_{S1})}{R_{S1}^2} + \frac{\hat{V}ar(R_{S2})}{R_{S2}^2} - 2 \times \frac{\hat{C}ov(R_{S1}, R_{S2})}{R_{S1} \times R_{S2}}} \right\}, \right.$$
$$\left. \exp\left\{ \ln(RR) + 1.96 \times \sqrt{\frac{\hat{V}ar(R_{S1})}{R_{S1}^2} + \frac{\hat{V}ar(R_{S2})}{R_{S2}^2} - 2 \times \frac{\hat{C}ov(R_{S1}, R_{S2})}{R_{S1} \times R_{S2}}} \right\} \right). \quad (2)$$

Let C_{S1} and C_{S2} denote the case counts used to calculate R_{S1} and R_{S2}, respectively. These counts (and hence R_{S1} and R_{S2}) are often considered independent, and the covariance term in (2) thus omitted. Assuming that these counts follow a simple Poisson distribution, interval (2) reduces to the more customary interval [10,16,20] for the underlying rate ratio:

$$\left(\exp\left\{ \ln(RR) - 1.96 \times \sqrt{\frac{1}{C_{S1}} + \frac{1}{C_{S2}}} \right\}, \exp\left\{ \ln(RR) + 1.96 \times \sqrt{\frac{1}{C_{S1}} + \frac{1}{C_{S2}}} \right\} \right). \quad (2a)$$

The adjustment to interval (2a) to account for multiple-case incidents must address dependencies not only within the two subgroups, but also dependencies extending across the subgroups. The latter type of dependency occurs when some multiple-case incidents include cases in both subgroups, and in such instances the covariance term in interval (2) cannot simply be omitted. Let $C_{S1 \cdot 1}$, $C_{S1 \cdot 2}$, $C_{S1 \cdot 3}$, ..., $C_{S1 \cdot N}$ denote the incident-specific case counts (some possibly zero) for subgroup 1 (so $C_{S1} \equiv \sum_{k=1}^{N} C_{S1 \cdot k}$). Similarly, let $C_{S2 \cdot 1}$, $C_{S2 \cdot 2}$, $C_{S2 \cdot 3}$, ..., $C_{S2 \cdot N}$ denote the incident-specific case counts for subgroup 2 (so $C_{S2} \equiv \sum_{k=1}^{N} C_{S2 \cdot k}$). With P_{S1} and P_{S2} denoting person-years at risk for the respective subgroups, an unbiased estimate of $Cov(R_{S1}, R_{S2})$ under the compound Poisson process model is given by $\sum_{k=1}^{N} (C_{S1 \cdot k} \times C_{S2 \cdot k}) / (P_{S1} \times P_{S2}) \times 100{,}000^2$ (see Appendix B). Substituting the appropriate variance and covariance estimates into (2) and simplifying provides the adjustment to interval (2a) to reflect both within-group and cross-group dependencies:

Table 2: Estimated Coverage for Unadjusted (Poisson) 95% Confidence Interval (1a).

			Incident Occurrence Rate per 100,000 Person-years ($\lambda \times 10^5$)			
			0.050	0.125	0.250	0.500
Within-Incident Case Count Distribution			Relative Frequency of Coverage			
(p_1, p_2, p_3, p_4)	μ	σ^2				
(0.76, 0.24, 0.00, 0.00)	1.24	0.1824	0.912	0.908	0.906	0.899
(0.95, 0.05, 0.00, 0.00)	1.05	0.0475	0.944	0.941	0.937	0.944
(0.85, 0.10, 0.05, 0.00)	1.20	0.2600	0.914	0.896	0.908	0.900
(0.80, 0.15, 0.03, 0.02)	1.27	0.3771	0.891	0.884	0.892	0.891
(0.70, 0.20, 0.07, 0.03)	1.43	0.5651	0.867	0.864	0.864	0.854

$$\left(\exp\left\{ \ln(RR) - 1.96 \times \sqrt{\frac{\sum_{k=1}^{N} C_{S1\cdot k}^2}{C_{S1}^2} + \frac{\sum_{k=1}^{N} C_{S2\cdot k}^2}{C_{S2}^2} - 2 \times \frac{\sum_{k=1}^{N} (C_{S1\cdot k} \times C_{S2\cdot k})}{C_{S1} \times C_{S2}}} \right\}, \right.$$
$$\left. \exp\left\{ \ln(RR) + 1.96 \times \sqrt{\frac{\sum_{k=1}^{N} C_{S1\cdot k}^2}{C_{S1}^2} + \frac{\sum_{k=1}^{N} C_{S2\cdot k}^2}{C_{S2}^2} - 2 \times \frac{\sum_{k=1}^{N} (C_{S1\cdot k} \times C_{S2\cdot k})}{C_{S1} \times C_{S2}}} \right\} \right) \quad (2b)$$

As with the adjusted interval for a rate, when no multiple-case incidents are represented in the data it follows that $\sum_{k=1}^{N} C_{S1\cdot k}^2 = C_{S1}$, $\sum_{k=1}^{N} C_{S2\cdot k}^2 = C_{S2}$, all cross-products $C_{S1\cdot k} \times C_{S2\cdot k}$ are zero, and the adjusted interval (2b) reduces to the unadjusted interval (2a).

Example 3

The calculations required for intervals (2a) and (2b) are illustrated using NVDRS summary data for the year 2004. These data indicate a total of 144 incidents (13 states) involving homicide followed by suicide. Of these, 127 incidents involved a single homicide, 15 involved a double homicide, one involved a triple homicide, and one involved a quadruple homicide, for a total of 164 homicides.[21] Expanding on Example 1, Table 4 provides a summary of the incident-specific homicide counts associated with all 144 homicide-suicide incidents, with homicides classified into the two demographic subgroups described earlier (<21 years of age, 21+ years of age).

The left half of Table 4 contains summary data for the homicide-suicide incidents. The right half of the table shows the calculation of the sums appearing in intervals (2a) and (2b). For example, the third summary line represents four separate incidents, each with two homicides in the first age group and none in the second age group. Since $C_{S1\cdot k} = 2$ and $C_{S2\cdot k} = 0$ for each of the four incidents represented by this line, it adds $4 \times 2 = 8$ to the sum $\sum_{k=1}^{N} C_{S1\cdot k}$ and $4 \times 2^2 = 16$ to the sum $\sum_{k=1}^{N} C_{S1\cdot k}^2$.

The totals at the bottom of the right half of Table 4 are used to calculate the unadjusted and adjusted confidence interval estimates for the rate ratio. In the states covered by NVDRS for 2004, the size of the population for the first age group was approximately 19.8 million, and approximately 48.9 million for the second age group.[18] Since the reporting period covered one year, these population figures approximate person-years at risk in the respective age groups, so the estimated rate ratio is:

Table 3: Estimated Coverage for Adjusted (Compound Poisson) 95% Confidence Interval (1b).

			Incident Occurrence Rate per 100,000 Person-years ($\lambda \times 10^5$)			
			0.050	0.125	0.250	0.500
Within-Incident Case Count Distribution			Relative Frequency of Coverage			
(p_1, p_2, p_3, p_4)	μ	σ^2				
(0.76, 0.24, 0.00, 0.00)	1.24	0.1824	0.948	0.953	0.950	0.950
(0.95, 0.05, 0.00, 0.00)	1.05	0.0475	0.958	0.950	0.952	0.951
(0.85, 0.10, 0.05, 0.00)	1.20	0.2600	0.955	0.947	0.949	0.949
(0.80, 0.15, 0.03, 0.02)	1.27	0.3771	0.945	0.948	0.949	0.949
(0.70, 0.20, 0.07, 0.03)	1.43	0.5651	0.942	0.948	0.948	0.948

Table 4: NVDRS Summary Data and Calculations for a Rate Ratio Confidence Interval.

Incident Summary Data				Calculation of Sums				
Incident Count	Homicides in Incident	Age <21	Age 21+	$\Sigma C_{S1\cdot k}$	$\Sigma C_{S1\cdot k}^2$	$\Sigma C_{S2\cdot k}$	$\Sigma C_{S2\cdot k}^2$	$\Sigma C_{S1\cdot k} \times C_{S2\cdot k}$
14	1	1	0	14	14			
113	1	0	1			113	113	
4	2	2	0	8	16			
5	2	1	1	5	5	5	5	5
6	2	0	2			12	24	
1	3	2	1	2	4	1	1	2
1	4	2	2	2	4	2	4	4
144				31	43	133	147	11

$RR \equiv R_{S1}/R_{S2} = (31/19.8M)/(133/48.9M) = (31/133) \times (48.9/19.8) \approx 0.576$.

Again recalling the shorthand notation $C_{S1} \equiv \sum_{k=1}^{N} C_{S1\cdot k}$ and $C_{S2} \equiv \sum_{k=1}^{N} C_{S2\cdot k}$, the unadjusted 95% confidence interval for the rate ratio is obtained by substituting $RR \approx 0.576$ and the values $C_{S1} = 31$ and $C_{S2} = 133$ into (2a):

$$\left(\exp\left\{ \ln(0.576) - 1.96 \times \sqrt{\frac{1}{31} + \frac{1}{133}} \right\}, \ \exp\left\{ \ln(0.576) + 1.96 \times \sqrt{\frac{1}{31} + \frac{1}{133}} \right\} \right)$$

resulting in the interval estimate (0.390, 0.852).

To calculate the adjusted 95% confidence interval, the values $RR \approx 0.576$, $C_{S1} = 31$, $\sum_{k=1}^{N} C_{S1\cdot k}^2 = 43$, $C_{S2} = 133$, $\sum_{k=1}^{N} C_{S2\cdot k}^2 = 147$, and $\sum_{k=1}^{N} \left(C_{S1\cdot k} \times C_{S2\cdot k} \right) = 11$ are substituted into expression (2b):

$$\left(\exp\left\{ \ln(0.576) - 1.96 \times \sqrt{\frac{43}{31^2} + \frac{147}{133^2} - 2 \times \frac{11}{31 \times 133}} \right\}, \right.$$
$$\left. \exp\left\{ \ln(0.576) + 1.96 \times \sqrt{\frac{43}{31^2} + \frac{147}{133^2} - 2 \times \frac{11}{31 \times 133}} \right\} \right)$$

resulting in the interval estimate (0.375, 0.884).

The adjusted interval estimate is approximately 10% wider than the unadjusted estimate. In this example, the influence of the increased variance estimates for the rate ratio components (the numerator and denominator rate estimates) is partially offset by the covariance term.

When cases associated with multiple-case incidents are concentrated mostly within subgroups the covariance term in adjusted interval (2b) will be relatively small and this interval will generally be wider than unadjusted interval (2a). For example, if the subgroups represent separate geographic regions, multiple-case incidents will rarely involve both subgroups and the covariance term will be negligible. Conversely, in situations where multiple-case incidents frequently involve both subgroups, the covariance term in (2b) offsets the influence of cases concentrated within subgroups (as in the previous computational example). In some instances the covariance term can dominate to the extent that the adjusted interval is narrower than the unadjusted interval.

The performance of intervals (2a) and (2b) under the types of conditions just described was evaluated using stochastic simulation, assuming various combinations of the incident occurrence rate, the underlying rate ratio, and the within-incident case count distribution (p_1, p_2, p_3, p_4). The initial set of simulations randomly assigned all cases associated with any given multiple-case incident to one of the two subgroups (according to probabilities consistent with the assumed rate ratio) thereby reducing the covariance term in interval (2b) to zero. Provided that the underlying mean incident count for each subgroup was not less than 10, the estimated coverage for adjusted interval (2b) was within 1% of the nominal level (95%) for all simulation inputs considered. By contrast, the estimated coverage for unadjusted interval (2a) dropped to about 85% when assuming the most extreme within-incident case count distribution (0.70, 0.20, 0.07, 0.03) from Example 2.

When modified to include multiple-case incidents simultaneously involving both subgroups, the simulations again indicated an effective coverage for adjusted interval (2b) close to the nominal level for the inputs considered. However, the cross-group dependencies introduced by this simple change also resulted in an adjusted interval with average width about the same as that of the unad-

justed interval. Consequently, the estimated coverage of unadjusted interval (2a) was also close to the nominal level.

Confidence Intervals for Age-Standardized Rates and Rate Ratios

The methods considered thus far pertain to *crude* rates (and ratios of crude rates) estimated using a single case count (numerator) and a single value for person-years at risk (denominator). The treatment of *age-standardized* rates and ratios of age-standardized rates follows from straightforward extensions of results already presented.

To illustrate the proposed extension for age-standardized rates, assume that there are M age groups into which the data are partitioned and denote the corresponding age-group rate estimators by R_{G1}, R_{G2}, R_{G3}, ..., R_{GM}. Let ω_{G1}, ω_{G2}, ω_{G3}, ..., ω_{GM} denote corresponding age-group population fractions (assumed fixed) in the referent (standard) population, such that $\sum_{\ell=1}^{M} \omega_{G\ell} = 1$. Applying the direct method of standardization [14] the age-standardized rate estimator is given by:

$$R_a \equiv \sum_{\ell=1}^{M} \omega_{G\ell} \times R_{G\ell}.$$

The usual formula for the variance of a weighted sum provides the following expression for the estimated variance of R_a:

$$\hat{\mathrm{Var}}(R_a) = \sum_{\ell=1}^{M} \omega_{G\ell}^2 \times \hat{\mathrm{Var}}(R_{G\ell}) + 2 \times \sum_{\ell=1}^{M-1} \sum_{m=\ell+1}^{M} \omega_{G\ell} \times \omega_{Gm} \times \hat{\mathrm{Cov}}(R_{G\ell}, R_{Gm}). \quad (3)$$

Here, dependencies between cases within any given age group will affect the variance of the age-group rate estimator, while dependencies between cases in different age groups will result in nonzero covariances between the age-group rate estimators. The analog to interval (1) for age-standardized rates is given by:

$$\left(\exp\left\{ \ln(R_a) - 1.96 \times \sqrt{\frac{\hat{\mathrm{Var}}(R_a)}{R_a^2}} \right\}, \ \exp\left\{ \ln(R_a) + 1.96 \times \sqrt{\frac{\hat{\mathrm{Var}}(R_a)}{R_a^2}} \right\} \right). \quad (4)$$

Appendix equations (A.1) and (B.1) provide variance and covariance estimation formulas applicable to the age-group rate estimators (with rate scale per 100,000 person-years) assuming a compound Poisson process model. These can be substituted into (3) to obtain a computational formula for $\hat{\mathrm{Var}}(R_a)$, which when used in (4) provides an interval adjusting for multiple-case incidents.

When considering a ratio of age-standardized rate estimates for two subgroups, there are three potential types of dependency associated with multiple-case incidents: (i) between cases within the same subgroup and age group, (ii) between cases in different age groups within the same subgroup, and (iii) between cases in different subgroups. All three effects can be simultaneously illustrated by considering the ratio of age-standardized rates for males and females. A multiple-case incident may variously involve several males (or females) in the same age group; males (or females) in different age groups (dependency between case counts contributing to the same age-standardized rate estimate); or both males and females (dependency between case counts contributing to both numerator and denominator rate estimates).

Letting $R_{a\cdot S1}$ and $R_{a\cdot S2}$ denote the respective age-standardized rate estimators for two subgroups, the age-standardized rate ratio is estimated by $RR_a \equiv R_{a\cdot S1}/R_{a\cdot S2}$. The analog to interval (2) for age-standardized rate ratios is:

$$\left(\exp\left\{ \ln(RR_a) - 1.96 \times \sqrt{\frac{\hat{\mathrm{Var}}(R_{a\cdot S1})}{R_{a\cdot S1}^2} + \frac{\hat{\mathrm{Var}}(R_{a\cdot S2})}{R_{a\cdot S2}^2} - 2 \times \frac{\hat{\mathrm{Cov}}(R_{a\cdot S1}, R_{a\cdot S2})}{R_{a\cdot S1} \times R_{a\cdot S2}}} \right\}, \right.$$
$$\left. \exp\left\{ \ln(RR_a) + 1.96 \times \sqrt{\frac{\hat{\mathrm{Var}}(R_{a\cdot S1})}{R_{a\cdot S1}^2} + \frac{\hat{\mathrm{Var}}(R_{a\cdot S2})}{R_{a\cdot S2}^2} - 2 \times \frac{\hat{\mathrm{Cov}}(R_{a\cdot S1}, R_{a\cdot S2})}{R_{a\cdot S1} \times R_{a\cdot S2}}} \right\} \right). \quad (5)$$

Computational formulas for $\hat{\mathrm{Var}}(R_{a\cdot S1})$ and $\hat{\mathrm{Var}}(R_{a\cdot S2})$ under the compound Poisson process model can be obtained as described above for interval (4). Appendix equation (C.1) provides a computational formula for $\hat{\mathrm{Cov}}(R_{a\cdot S1}, R_{a\cdot S2})$ (with rate scale per 100,000 person-years) assuming the compound Poisson process model. These formulas can be substituted into (5) to obtain an interval adjusting for the effects of multiple-case incidents.

When there are no multiple-case incidents represented in the data, the covariance estimates in (3) vanish as does the term $\hat{\mathrm{Cov}}(R_{a\cdot S1}, R_{a\cdot S2})$ appearing in (5). Under such circumstances (4) and (5) reduce to intervals appropriate when case counts are assumed to follow a simple Poisson distribution and subgroups are assumed independent.[22]

Stochastic simulation was used to evaluate the coverage properties of adjusted intervals (4) and (5) when case counts conform to a compound Poisson process model. When the age distributions of the study groups of interest do not depart too greatly from that of the referent population, the simulation results suggest that coverage levels are comparable to those reported earlier for the adjusted confidence intervals for crude rates and rate ratios. In particular, estimated coverage was close to the nominal level provided that underlying mean subgroup incident counts were not less than 10.

Assessment of Bias

It was noted at the outset that crude rate estimators are unbiased; by extension age-standardized rate estimators are also unbiased. Consequently, the coverage properties of adjusted intervals for crude and age-standardized rates depend on the appropriateness of the compound Poisson process model as well as the accuracy of the normal approximation implied when applying the delta method.

While the crude and age-standardized rate estimators are unbiased, the rate ratio estimators are only asymptotically unbiased. A supplementary assessment of finite-sample bias in the simulations described following Example 3 suggests that it is relatively small compared to the standard error of the rate ratio estimator, for the simulation inputs considered. In none of the simulations was the bias strong enough to cause the effective coverage of the adjusted (compound Poisson) interval to differ substantially from the nominal level.

Conclusion

By referring to the compound Poisson process model in place of the simple Poisson model for case counts, confidence intervals for injury-related mortality rates and rate ratios can be adjusted to account for statistical dependencies associated with multiple-case incidents. The adjustments rely on closed-form computations and offer meaningful improvements in the accuracy of statistical statements. The adjusted interval estimators described in this paper have been programmed as general routines using SAS.[19]

When the data show any pattern of multiple-case incidents, the adjusted intervals for rates will be wider than their unadjusted counterparts. This does not hold for the adjusted intervals for rate ratios, however; different patterns in the data can variously widen or narrow these intervals relative to their unadjusted counterparts.

It is evident that in situations where multiple-case incidents are very infrequent and involve small numbers of cases when they do occur, there will be little difference between the statistical results obtained using the methods based on the compound Poisson process model and those based on the simple Poisson model. In the context of the NVDRS data, for example, when suicides are considered separately there is almost no distinction between suicide-related incident counts and suicide case counts (because multiple-suicide incidents are extremely infrequent). Conversely, there may be situations covered by other reporting systems where multiple-case incidents are more prominent and/or involve larger numbers of cases than in the examples considered in this paper. Simulations show that in such instances, the gaps between nominal and effective coverage probabilities for the unadjusted interval estimators become quite substantial.

Appendices

A. The Estimated Variance of a Total Case Count

Let $N \sim \text{Poisson}(\lambda P)$ denote the count of incidents and let $C_1, C_2, C_3, \ldots, C_N$ denote the incident-specific case counts. That $\sum_{k=1}^{N} C_k^2$ is an unbiased estimator for the variance of the total case count $C \equiv \sum_{k=1}^{N} C_k$ under a compound Poisson process model can be demonstrated using a basic conditioning argument. Employing the assumptions specified in the text (particularly the independence assumptions) it follows that:

$$\begin{aligned} E[\textstyle\sum_{k=1}^{N} C_k^2] &= E_N[E[\textstyle\sum_{k=1}^{N} C_k^2 \mid N]] \\ &= E_N[\textstyle\sum_{k=1}^{N} E[C_k^2 \mid N]] \\ &= E_N[\textstyle\sum_{k=1}^{N} E[C_k^2]] \\ &= E_N[N \times (\mu^2 + \sigma^2)] \\ &= \lambda P \times (\mu^2 + \sigma^2). \end{aligned}$$

Because the last term in the sequence of equalities corresponds to the underlying variance of the total case count C under the compound Poisson model, it follows that $\hat{\text{Var}}(C) = \sum_{k=1}^{N} C_k^2$ is an unbiased estimator for Var(C). Since the estimated case rate (per 100,000 person-years) is given by $R \equiv C/P \times 100{,}000$, an unbiased estimate of Var(R) is:

$$\hat{\text{Var}}(R) = \hat{\text{Var}}(C)/P^2 \times 100{,}000^2 = \textstyle\sum_{k=1}^{N} C_k^2 / P^2 \times 100{,}000^2. \qquad (A.1)$$

B. The Covariance of Rate Estimators

Let C_{S1} denote the total case count for subgroup 1 and let $C_{S1 \cdot 1}, C_{S1 \cdot 2}, C_{S1 \cdot 3}, \ldots, C_{S1 \cdot N}$ denote the incident-specific case counts (some possibly zero) for subgroup 1 (so $C_{S1} \equiv \sum_{k=1}^{N} C_{S1 \cdot k}$). Similarly, for subgroup 2 let C_{S2} denote the total case count and let $C_{S2 \cdot 1}, C_{S2 \cdot 2}, C_{S2 \cdot 3}, \ldots, C_{S2 \cdot N}$ denote the incident-specific case counts (so $C_{S2} \equiv \sum_{k=1}^{N} C_{S2 \cdot k}$). The usual formula for the variance of a sum:

$$\text{Var}(C_{S1} + C_{S2}) = \text{Var}(C_{S1}) + \text{Var}(C_{S2}) + 2 \times \text{Cov}(C_{S1}, C_{S2})$$

can be rearranged to get:

$$\mathrm{Cov}(C_{S1}, C_{S2}) = (\mathrm{Var}(C_{S1} + C_{S2}) - \mathrm{Var}(C_{S1}) - \mathrm{Var}(C_{S2}))/2.$$

The results of Appendix A provide unbiased estimators for all of the terms appearing on the right-hand side of the last equality. Substituting these unbiased estimators provides an unbiased estimate of the covariance:

$$\hat{\mathrm{Cov}}(C_{S1}, C_{S2}) = (\sum_{k=1}^{N}(C_{S1\cdot k} + C_{S2\cdot k})^2 - \sum_{k=1}^{N} C_{S1\cdot k}^2 - \sum_{k=1}^{N} C_{S2\cdot k}^2)/2$$
$$= \sum_{k=1}^{N}(C_{S1\cdot k} \times C_{S2\cdot k}).$$

Let P_{S1} and P_{S2} denote the respective person-years at risk in subgroups 1 and 2. The corresponding rate estimators (per 100,000 person-years) are $R_{S1} \equiv C_{S1}/P_{S1} \times 100{,}000$ and $R_{S2} \equiv C_{S2}/P_{S2} \times 100{,}000$. It follows immediately that an unbiased estimate of $\mathrm{Cov}(R_{S1}, R_{S2})$ under a compound Poisson process model is given by:

$$\hat{\mathrm{Cov}}(R_{S1}, R_{S2}) = \sum_{k=1}^{N}(C_{S1\cdot k} \times C_{S2\cdot k})/(P_{S1} \times P_{S2}) \times 100{,}000^2. \quad (B.1)$$

C. The Covariance of Age-Standardized Rate Estimators

Consider age-standardized rate estimators $R_{a\cdot S1}$ and $R_{a\cdot S2}$ for two subgroups based on a partition of the data into M age groups. Let $R_{S1\cdot G1}$, $R_{S1\cdot G2}$, $R_{S1\cdot G3}$, ... $R_{S1\cdot GM}$ denote the age-group rate estimators (per 100,000 person-years) for subgroup 1 and similarly let $R_{S2\cdot G1}$, $R_{S2\cdot G2}$, $R_{S2\cdot G3}$, ... $R_{S2\cdot GM}$ denote the age-group rate estimators for subgroup 2. Referring to the (fixed) age-group population fractions ω_{G1}, ω_{G2}, ω_{G3}, ..., ω_{GM} defined in the text, the covariance of the age-standardized rate estimators is given by:

$$\mathrm{Cov}(R_{a\cdot S1}, R_{a\cdot S2}) = \mathrm{Cov}(\sum_{\ell=1}^{M} \omega_{G\ell} \times R_{S1\cdot G\ell}, \sum_{m=1}^{M} \omega_{Gm} \times R_{S2\cdot Gm})$$
$$= \sum_{\ell=1}^{M}\sum_{m=1}^{M} \omega_{G\ell} \times \omega_{Gm} \times \mathrm{Cov}(R_{S1\cdot G\ell}, R_{S2\cdot Gm}).$$

An unbiased estimate of the last expression on the right-hand side follows from the results of Appendix B. Specifically, for age group ℓ in subgroup 1, let $C_{S1\cdot G\ell\cdot 1}$, $C_{S1\cdot G\ell\cdot 2}$, $C_{S1\cdot G\ell\cdot 3}$, ..., $C_{S1\cdot G\ell\cdot N}$ denote the incident-specific case counts (some possibly zero) and let $P_{S1\cdot G\ell}$ denote the person-years at risk. Similarly, for age group m in subgroup 2, let $C_{S2\cdot Gm\cdot 1}$, $C_{S2\cdot Gm\cdot 2}$, $C_{S2\cdot Gm\cdot 3}$, ... $C_{S2\cdot Gm\cdot N}$ denote the incident-specific case counts and let $P_{S2\cdot Gm}$ denote the person-years at risk. From (B.1) it follows that an unbiased estimate for $\mathrm{Cov}(R_{a\cdot S1}, R_{a\cdot S2})$ is:

$$\hat{\mathrm{Cov}}(R_{a\cdot S1}, R_{a\cdot S2}) =$$
$$\sum_{\ell=1}^{M}\sum_{m=1}^{M} \omega_{G\ell} \times \omega_{Gm} \times \sum_{k=1}^{N}(C_{S1\cdot G\ell\cdot k} \times C_{S2\cdot Gm\cdot k})/(P_{S1\cdot G\ell} \times P_{S2\cdot Gm}) \times 100{,}000^2. \quad (C.1)$$

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

The findings and conclusions in this article are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. Anderson RN, Minino AM, Fingerhut LA, Warner M, Heinen MA: *Deaths: Injuries, 2001. National Vital Statistics Reports Volume 52*. Issue 21 Hyattsville MD: National Center for Health Statistics; 2004.
2. Kochanek KD, Murphy SL, Anderson RN, Scott C: *Deaths: Final Data for 2002. National Vital Statistics Reports Volume 53*. Issue 5 Hyattsville MD: National Center for Health Statistics; 2004.
3. Vyrostek SB, Annest JL, Ryan GW: **Surveillance for fatal and nonfatal injuries – United States, 2001.** *MMWR Surveill Summ* 2004, **SS-7:**1-57. Available at: http://www.cdc.gov.
4. Berry JG, Harrison JE: *A Guide to Statistical Methods for Injury Surveillance. Injury Technical Paper Series; 5* Adelaide: Australian Institute of Health and Welfare; 2005. Available at: http://www.nisu.flinders.edu.au.
5. Feller W: *An Introduction to Probability Theory and Its Applications Volume I*. 3rd edition. New York: Wiley; 1968.
6. Taylor HM, Karlin S: *An Introduction to Stochastic Modeling* Orlando: Academic Press; 1984.
7. Ross SM: *Introduction to Probability Models* 4th edition. San Diego: Academic Press; 1989.
8. Serpi TL, Wiersma B, Hackman H, Ortega L, Jacquemin BJ, Weintraub KS, Kohn M, Millet L, Carter LP, Weis MA, Head KE, Powell V, Mueller M, Paulozzi LJ, White D, Ryan G: **Homicide and suicide rates – National Violent Death Reporting System, six states, 2003.** *MMWR Morb Mortal Wkly Rep* 2005, **54:**377-380 [http://www.cdc.gov]. Available at: http://www.cdc.gov.
9. Brillinger DR: **The natural variability of vital rates and associated statistics.** *Biometrics* 1986, **42:**693-734.
10. Greenland S, Rothman KJ: **Introduction to categorical statistics.** In *Modern Epidemiology* 2nd edition. Edited by: Rothman KJ, Greenland S. Philadelphia: Lippincott-Raven Publishers; 1998:231-252.
11. Serfling RJ: *Approximation Theorems of Mathematical Statistics* New York: Wiley; 1980.
12. van der Vaart AW: *Asymptotic Statistics* New York: Cambridge University Press; 1998.
13. Agresti A: *Categorical Data Analysis* 2nd edition. Hoboken NJ: Wiley; 2002.
14. Fleiss JL, Levin B, Paik MC: *Statistical Methods for Rates and Proportions* 3rd edition. Hoboken NJ: Wiley; 2003.
15. Carriere KC, Roos LL: **Comparing standardized rates of events.** *Am J Epidemiol* 1994, **140(5):**472-482.
16. Clayton D, Hills M: *Statistical Models in Epidemiology* New York: Oxford University Press; 1993.
17. Breslin C, Koehoorn M, Smith P, Manno M: **Age related differences in work injuries and permanent impairment: a comparison of workers' compensation claims among adolescents, young adults, and adults.** *Occup Environ Med* 2003, **60:**e10.
18. U.S. Census Bureau: *SC-EST2004-AGESEX-RES: Estimates of the Resident Population by Single-Year of Age and Sex for the United States and States: July 1, 2004* . Available at: http://www.census.gov. Downloaded July 26, 2005.
19. SAS Institute Inc: *SAS® Language Reference Dictionary, Version 8* Cary NC: SAS Institute Inc; 1999.
20. Rosner B: *Fundamentals of Biostatistics* 4th edition. Belmont CA: Wadsworth; 1995.
21. Bossarte RM, Simon TR, Barker L: **Characteristics of homicide followed by suicide incidents in multiple states, 2003 – 2004.** *Inj Prev* 2006, **12(Suppl 2):**33-38.
22. Greenland S, Rothman KJ: **Introduction to stratified analysis.** In *Modern Epidemiology* 2nd edition. Edited by: Rothman KJ, Greenland S. Philadelphia: Lippincott-Raven Publishers; 1998:253-279.