Journal of Data Science **8**(2010), 471-482

Interval Estimation for Ratios of Correlated Age-Adjusted Rates

Ram C. Tiwari[1], Yi Li[2] and Zhaohui Zou[3]
[1]*Food and Drug Administration,* [2]*Harvard School of Public Health and* [3]*Information Management Services*

Abstract: Providing reliable estimates of the ratios of cancer incidence and mortality rates across geographic regions has been important for the National cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) Program as it profiles cancer risk factors as well decides cancer control planning. A fundamental difficulty, however, arises when such ratios have to be computed to compare the rate of a subregion (e.g., California) with that of a parent region (e.g., the US). Such a comparison is often made for policy-making purposes. Based on F-approximations as well as normal approximations, this paper provides new confidence intervals (CIs) for such rate ratios. Intensive simulations, which capture the real issues with the observed mortality data, reveal that these two CIs perform well. In general, for rare cancer sites, the F-intervals are often more conservative, and for moderate and common cancers, all intervals perform similarly.

Key words: Cancer rate-ratio, F-approximation, normal-approximation, 2000 US standards.

1. Introduction and Preliminaries

Let Ω denote a region such as the entire US or a state in the US, and let X denote a subregion (a proper subset) of Ω. Denote the rest of the region by $X^c = \Omega/X$. Let R_X, R_{X^c} and R_Ω denote the age-adjusted rates for X, X^c and Ω, respectively, all of which are defined below.

$$R_X = \sum_{j=1}^{J} w_j \frac{d_{Xj}}{n_{Xj}},\ R_{X^c} = \sum_{j=1}^{J} w_j \frac{d_{X^cj}}{n_{X^cj}},\ R_\Omega = \sum_{j=1}^{J} w_j \frac{d_{\Omega j}}{n_{\Omega j}},$$

where w_j are known standards normalized to sum to 1 over the J age-groups; d_{Xj}, d_{X^cj}, $d_{\Omega j}$ and n_{Xj}, $n_{X^cj}, n_{\Omega j}$ are the number of cancer cases or deaths, and the number of person-years in X, X^c, Ω, respectively.

We define the underlying true rates as $\mu_X = E(R_X)$, $\mu_{X^c} = E(R_{X^c})$, $\mu_\Omega = E(R_\Omega)$, whose unbiased point estimates are $\widehat{\mu_X} = R_X$, $\widehat{\mu_{X^c}} = R_{X^c}$ and $\widehat{\mu_\Omega} = R_\Omega$,

respectively. Under the assumption that the counts are independent Poisson random variables, their variances can be approximated by

$$V_X \cong \sum_{j=1}^{J} w_j^2 \frac{d_{Xj}}{n_{Xj}^2},\ V_{X^c} \cong \sum_{j=1}^{J} w_j^2 \frac{d_{X^cj}}{n_{X^cj}^2},\ V_\Omega \cong \sum_{j=1}^{J} w_j^2 \frac{d_{\Omega j}}{n_{\Omega j}^2}.$$

Our interest is to construct an approximate $100(1-\alpha)\%$ confidence interval for the rate-ratio of X to Ω, namely that of the parameter $\theta = \frac{\mu_X}{\mu_\Omega}$, e.g the cancer rate-ratio comparing the rate of California with that of the US. Such comparisons are often necessary for policy-making purposes. However, statistical diffculties arise when the estimates of the rates (e.g. those of California and the US) are correlated.

In a simpler context, Fay (1999) and Tiwari *et al.* (2006) derived confidence intervals for $\phi = \frac{\mu_X}{\mu_{X^c}}$ using the F distribution as an approximation of the ratio of the two independent Gamma random variables. Specifically, denote by $\hat{\phi} = \frac{R_X}{R_{X^c}}$, the estimator of ϕ. The F-based confidence interval is given in Tiwari *et al.* (2006) as

$$\left(\frac{R_X}{R_{X^c}} F^{-1}_{((2R_X^2)/V_X,\,(2R_{X^c}^2/V_{X^c}))}\left(\frac{\alpha}{2}\right),\ \frac{R_X}{R_{X^c}} F^{-1}_{((2R_X^2)/V_X,(2R_{X^c}^2)/V_{X^c})}\left(1-\frac{\alpha}{2}\right) \right), \tag{1.1}$$

where $F^{-1}_{(a,b)}(p)$ is the p^{th} percentile of $F_{(a,b)}$. The derivation of (1.1) is to view the ratio $\hat{\phi}$ as an approximately F distributed random variable, given the independence of the numerator and the denominator in $\hat{\phi}$.

Tiwari *et al.* (2006) noted the need for comparing the rate of a subregion (e.g., California) with that of a larger region (e.g., the US), and gave a possible solution for computing the CIs for θ (the ratio of a subregion with its parent region) that accounts for the correlation between the age-adjusted rates. However, their method was not fully developed. In this paper, we will propose new F- and normal-based confidence intervals that can conveniently compute the CIS based on the ratio of two correlated estimates of rates. We demonstrate via simulations that the new intervals perform well and retain the nominal coverage probabilities.

The rest of the paper is organized as follows. In Section 2, we derive the new confidence intervals, and in Section 3 we evaluate their performance in terms of their empirical coverage probabilities. Section 4 gives a short discussion and Section 5 ends this paper with a conclusion.

2. Two New Confidence Intervals for θ

To derive the confidence interval for θ based on its point estimate $\hat{\theta}$, we first

assume that

$$\frac{n_{X1}}{n_{\Omega 1}} \doteq \frac{n_{X2}}{n_{\Omega 2}} \doteq \ldots \doteq \frac{n_{XJ}}{n_{\Omega J}} \doteq p_X$$

That is, the ratio of person years in X (e.g. California) to that of Ω (e.g. the US) is approximately the same across all age-groups. This so-called *proportional age-distribution assumption* is common in comparing the age-adjusted rates across different geographical areas and was found to be a good approximation for the US population; see, e.g., Pickle and White (1995).

Reasonable values for p_X is given by $\hat{p}_X = \frac{n_X}{n_\Omega}$, where $n_X = \sum_j n_{Xj}$ and $n_\Omega = \sum_j n_{\Omega j}$. Now, we can write

$$\begin{aligned} R_\Omega &= \sum_{j=1}^{J} w_j \frac{d_{\Omega j}}{n_{\Omega j}} \\ &= \sum_{j=1}^{J} w_j \frac{d_{Xj} + d_{X^c j}}{n_{\Omega j}} \\ &= \sum_{j=1}^{J} w_j \frac{n_{Xj}}{n_{\Omega j}} \left(\frac{d_{Xj}}{n_{Xj}}\right) + \sum_{j=1}^{J} w_j \frac{n_{X^c j}}{n_{\Omega j}} \left(\frac{d_{X^c j}}{n_{X^c j}}\right) \\ &\approx p_X R_X + p_{X^c} R_{X^c}, \end{aligned}$$

where $p_{X^c} = 1 - p_X$. Write the estimators for θ and ϕ by

$$\hat{\theta} = \frac{R_X}{R_\Omega} = \frac{R_X / R_{X^c}}{\hat{p}_X (R_X / R_{X^c}) + \hat{p}_{X^c}}$$

and

$$\hat{\phi} = \frac{R_X}{R_{X^c}} = \frac{\left(\frac{R_X}{R_\Omega}\right)(1 - \hat{p}_X)}{1 - \hat{p}_X \left(\frac{R_X}{R_\Omega}\right)}$$

respectively, where $\hat{p}_{X^c} = 1 - \hat{p}_X$. Also note that

$$\theta = \frac{\phi}{p_X \phi + p_{X^c}}.$$

Hence confidence intervals for ϕ would lead to those for θ (and vice versa), which will be derived below.

2.1 An F-approximation

Let $\left(\phi_{L(\alpha/2)}, \phi_{U(\alpha/2)}\right)$ be the $100(1-\alpha)\%$ F confidence interval for ϕ given in (1.1). Hence,

$$P(\phi_{L(\alpha/2)} < \phi < \phi_{U(\alpha/2)}) = 1 - \alpha. \tag{2.1}$$

Let $g(t) = 1 - p_{X^c}/(p_X t + p_{X^c})$ for short. Note that in this notation, $g(\phi)p_X = 1 - p_{X^c}/(p_X\phi + p_{X^c}) = \theta$. As $g(\phi)$ is increasing in ϕ, (2.1) is equivalent to

$$P\left(\frac{1}{p_X} g(\phi_{L(\alpha/2)}) < \theta < \frac{1}{p_X} g(\phi_{U(\alpha/2)})\right) = 1 - \alpha.$$

This finds the $100(1-\alpha)\%$ CI for ϕ.

Since $\hat{p}_X$ and $\hat{p}_{X^c}$ consistently estimate p_X and p_{X^c} respectively, we thus obtain the approximate $100(1-\alpha)\%$ confidence interval for θ as

$$\left(\frac{\phi_{L(\alpha/2)}}{\hat{p}_X\phi_{L(\alpha/2)} + \hat{p}_{X^c}}, \ \frac{\phi_{U\alpha/2)}}{\hat{p}_X\phi_{U(\alpha/2)} + \hat{p}_{X^c}}\right).$$

2.2 A normal approximation

From the perspective of a normal approximation, we can also derive a confidence interval for θ. First note that

$$\hat{\theta} = \frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}}$$

Using a Taylor series,

$$\begin{aligned}
&\frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}} \\
&= f\left(R_X, R_{X^c}\right) \\
&\approx f\left(\mu_X, \mu_{X^c}\right) + \left(R_X - \mu_X\right) f_X\left(\mu_X, \mu_{X^c}\right) + \left(R_{X^c} - \mu_{X^c}\right) f_{X^c}\left(\mu_X, \mu_{X^c}\right) \\
&\quad f\left(\mu_X, \mu_{X^c}\right) + \left(R_X - \mu_X\right)\left(\frac{\hat{p}_{X^c}\mu_{X^c}}{\left(\hat{p}_X\mu_X + \hat{p}_{X^c}\mu_{X^c}\right)^2}\right) \\
&\quad + \left(R_{X^c} - \mu_{X^c}\right)\left(\frac{\hat{p}_{X^c}\mu_X}{\left(\hat{p}_X\mu_X + \hat{p}_{X^c}\mu_{X^c}\right)^2}\right)
\end{aligned}$$

and

$$\begin{aligned}
\mu_\theta &= E\left(\frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}}\right) \\
&\approx \frac{\mu_X}{\hat{p}_X \mu_X + \hat{p}_{X^c} \mu_{X^c}} \\
V_\theta &= Var\left(\frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}}\right) \\
&\cong \left\{\frac{\hat{p}_{X^c}\mu_{X^c}}{\left(\hat{p}_X \mu_X + \hat{p}_{X^c}\mu_{X^c}\right)^2}\right\}^2 V_X + \left\{\frac{\hat{p}_{X^c}\mu_X}{\left(\hat{p}_X \mu_X + \hat{p}_{X^c}\mu_{X^c}\right)^2}\right\}^2 V_{X^c} \\
&= \left\{\frac{\hat{p}_{X^c}\mu_{X^c}\mu_X}{\left(\hat{p}_X \mu_X + \hat{p}_{X^c}\mu_{X^c}\right)^2}\right\}^2 \left(\frac{V_X}{\mu_X^2} + \frac{V_{X^c}}{\mu_{X^c}^2}\right).
\end{aligned}$$

Now, since (R_X, R_{X^c}) is asymptotically normal, and $\frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}} = f(R_X, R_{X^c})$ is a continuous function of (R_X, R_{X^c}), we have that $\frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}}$ is asymptotically normal with mean μ_θ and variance V_θ. Hence an approximate $100(1-\alpha)\%$ confidence interval for $\theta \cong \frac{\mu_X}{\mu_\Omega}$ is given by

$$\left(\frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}} - Z_{\alpha/2}\sqrt{\hat{V}_\theta}, \frac{R_X}{\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}} - Z_{\alpha/2}\sqrt{\hat{V}_\theta}\right),$$

where Z_α is the upper 100α percentile point of the standard normal distribution, and

$$\hat{V}_\theta = \left\{\frac{\hat{p}_{X^c} R_{X^c} R_X}{\left(\hat{p}_X R_X + \hat{p}_{X^c} R_{X^c}\right)^2}\right\}^2 \left(\frac{V_X}{R_X^2} + \frac{V_{X^c}}{R_{X^c}}\right) \doteq \left\{\frac{\hat{p}_{X^c} R_{X^c} R_X}{R_\Omega^2}\right\}^2 \left(\frac{V_X}{R_X^2} + \frac{V_{X^c}}{R_{X^c}^2}\right).$$

Thus, the normal confidence interval for θ is

$$\left(\frac{R_X}{R_\Omega} \pm Z_{\alpha/2}\frac{R_X}{R_\Omega^2}\sqrt{\hat{p}_{X^c} R_{X^c}^2 (\frac{V_X}{R_X^2} + \frac{V_{X^c}}{R_{X^c}^2})}\right). \tag{2.2}$$

3. Simulation Studies

We carried out simulations along the lines of Tiwari *et al.* (2006). We used the 2004 US cancer mortality data for tongue, esophagus, and lung cancer sites. These sites were selected to reflect the spectrum of cancer incidence; that is, from rare cancer (tongue), to moderate cancer (esophagus), to common cancer (lung).

The data were used to generate Poisson counts d_{Xj}, where X represents each of the 51 regions (50 states and Washington D.C.) and j indexes the 19 age-groups. The true means of the Poisson distributions are taken to be the observed values of d_{Xj}. We generated 10,000 Poisson counts, and the computed age-adjusted rates, using the 2000 US standards, so that $\sum_{j=1}^{19} w_j = 1$. Approximate 95% confidence intervals were obtained for the ratios of the age-adjusted rates for each of the 51 regions as compared to the overall US rate using the modified versions of the two CIs, as discussed in Tiwari *et al.* (2006).

- **F-interval:** $\left(\frac{\phi_{L(\alpha/2)}}{\hat{p}_X\phi_{L(\alpha/2)}+\hat{p}_{X^c}}, \frac{\phi_{U(\alpha/2)}}{\hat{p}_X\phi_{U(\alpha/2)}+\hat{p}_{X^c}}\right)$
  with

$$\begin{aligned}\phi_{L(\alpha/2)} &= \frac{R_X}{\tilde{R}_{X^c}} F^{-1}_{(2R_X^2/V_X, 2\tilde{R}_{X^c}^2/\tilde{V}_{X^c})}(\alpha/2)\,;\\ \phi_{U(\alpha/2)} &= \frac{\tilde{R}_X}{R_{X^c}} F^{-1}_{(2\tilde{R}_X^2/\tilde{V}_X, 2R_{X^c}^2/V_{X^c})}(\alpha/2)\,;\end{aligned}$$

- **Normal Interval:** $\left(\frac{\tilde{R}_X}{\tilde{R}_\Omega} \pm Z_{\alpha/2}\frac{\tilde{R}_X\tilde{R}_{X^c}}{\tilde{R}_\Omega^2}\sqrt{\hat{p}_{X^c}\left(\frac{\tilde{V}_X}{\tilde{R}_X^2}+\frac{\tilde{V}_{X^c}}{\tilde{R}_{X^c}^2}\right)}\right)$;
  where

$$\tilde{R}_X = \sum_{j=1}^{J} w_j \frac{d_{Xj}+\frac{1}{J}}{n_{Xj}},\ \tilde{R}_{X^c} = \sum_{j=1}^{n} w_j \frac{d_{X^cj}+\frac{1}{J}}{n_{X^cj}}$$

$$\tilde{V}_X = \sum_{j=1}^{J} w_j^2 \frac{d_{Xj}+\frac{1}{J}}{n_{Xj}^2},\ \tilde{V}_{X^c} = \sum_{j=1}^{n} w_j^2 \frac{d_{X^cj}+\frac{1}{J}}{n_{X^cj}^2};\ \tilde{\rho}_{X,\Omega} = \sum_{j=1}^{J} w_j^2 \frac{d_{Xj}+\frac{1}{J}}{n_{Xj}n_{\Omega j}}.$$

For the normal-approximation based intervals, if the lower limit is negative, we replace it by 0. Also note that the correction fraction added to the counts d_{Xj} and d_{X^cj} does not make any significant difference numerically. It merely avoids the zero rates.

Each of the Tables 1-3 gives the ratio of age-adjusted rates, $\frac{\tilde{R}_X}{\tilde{R}_\Omega}$, the estimate, $\hat{p}_X$, of the ratio of population for region X to that of the US, the empirical coverage probabilities, and the width of the 95% intervals for the two intervals, namely, the F- and normal-approximation based CIs. Because of the space, we only report the selected states in these tables and the full tables are available from the authors. The results show that the two CIs perform reasonably close. For the tongue cancer, all intervals have higher coverage probabilities and larger

Table 1: Performance of the Derived Confidence Intervals based on the 2004 Tongue Cancer Mortality Data. The empirical coverage probabilities and the widths of the intervals were based on 10,000 simulations.

State	Rate Ratio with US	Overlap ratio (p_X)	Empirical Coverage Prob.		Average Width	
			F-based	Normal-based	F-based	Normal-based
California	1.094	0.122	0.956	0.963	0.270	0.286
Colorado	0.984	0.016	0.957	0.952	0.821	0.802
Connecticut	0.892	0.012	0.964	0.955	0.797	0.771
Delaware	0.900	0.003	0.979	0.961	1.790	1.706
Washington DC	2.073	0.002	0.968	0.957	3.399	3.257
Florida	1.281	0.059	0.956	0.960	0.388	0.391
Georgia	0.921	0.030	0.959	0.955	0.555	0.548
Hawaii	0.463	0.004	0.973	0.971	1.054	1.001
Idaho	1.441	0.005	0.964	0.955	1.751	1.686
Illinois	1.019	0.043	0.955	0.955	0.450	0.450
Indiana	0.886	0.021	0.958	0.953	0.601	0.588
Iowa	0.883	0.010	0.965	0.958	0.838	0.807
Kansas	0.970	0.009	0.961	0.951	0.963	0.928
Kentucky	0.872	0.014	0.962	0.957	0.738	0.715
Louisiana	1.016	0.015	0.962	0.956	0.785	0.765
Maine	1.306	0.004	0.973	0.967	1.566	1.496
Maryland	0.817	0.019	0.958	0.951	0.624	0.609
Massachusetts	0.932	0.022	0.961	0.957	0.601	0.589
Michigan	0.953	0.034	0.958	0.956	0.481	0.476
Minnesota	0.906	0.017	0.963	0.956	0.691	0.674
Mississippi	0.377	0.010	0.971	0.956	0.618	0.593
Missouri	0.945	0.020	0.959	0.955	0.641	0.626
Montana	0.737	0.003	0.980	0.977	1.514	1.441
Nebraska	0.786	0.006	0.963	0.955	1.124	1.079
Nevada	1.214	0.008	0.963	0.955	1.230	1.188
New Hampshire	1.087	0.004	0.969	0.962	1.560	1.496
New Jersey	1.191	0.030	0.955	0.953	0.573	0.567
New Mexico	0.936	0.006	0.965	0.956	1.202	1.157
New York	1.129	0.066	0.953	0.956	0.363	0.369
North Carolina	0.928	0.029	0.959	0.956	0.533	0.526
Pennsylvania	0.841	0.042	0.954	0.954	0.384	0.383
Texas	1.044	0.077	0.954	0.955	0.356	0.365

widths than those for esophagus and lung cancers. Indeed, for the latter two cancers, the empirical coverage probabilities get much closer to 95%. For large states such as California, Florida, New York, Pennsylvania, and Texas, all intervals perform similarly in terms of interval widths and the coverage probabilities.

4. Discussion of the Results

The SEER Program of NCI has implemented the F-intervals of Fay (1999) and the modified F-interval of Tiwari *et al.* (2006) in the SEER*STAT software to compare the age-adjusted rates for two nonoverlapping regions. However, as pointed out in Tiwari *et al.* (2006), there is an emerging need of obtaining confidence interval formulae for comparing the age-adjusted rates of a subregion

Table 2: Performance of the Derived Confidence Intervals based on the 2004 Esophagus Cancer Mortality Data. The empirical coverage probabilities and the widths of the intervals were based on 10,000 simulations.

State	Rate Ratio with US	Overlap ratio (p_X)	Empirical Coverage Prob.		Average Width	
			F-based	Normal-based	F-based	Normal-based
California	0.842	0.122	0.952	0.961	0.092	0.097
Colorado	0.980	0.016	0.953	0.951	0.299	0.296
Connecticut	0.968	0.012	0.957	0.953	0.306	0.302
Delaware	1.591	0.003	0.956	0.951	0.826	0.805
Washington DC	1.286	0.002	0.961	0.954	0.939	0.908
Florida	0.951	0.059	0.957	0.964	0.122	0.125
Georgia	1.044	0.030	0.951	0.951	0.221	0.222
Hawaii	0.563	0.004	0.963	0.958	0.395	0.382
Idaho	0.864	0.005	0.956	0.949	0.493	0.480
Illinois	1.037	0.043	0.956	0.958	0.170	0.172
Indiana	1.080	0.021	0.951	0.950	0.247	0.246
Iowa	0.976	0.010	0.953	0.949	0.324	0.319
Kansas	0.913	0.009	0.954	0.951	0.344	0.338
Kentucky	1.009	0.014	0.953	0.951	0.291	0.288
Louisiana	1.034	0.015	0.950	0.948	0.292	0.290
Maine	1.336	0.004	0.960	0.956	0.565	0.551
Maryland	0.951	0.019	0.955	0.954	0.251	0.249
Massachusetts	1.167	0.022	0.952	0.951	0.246	0.245
Michigan	1.099	0.034	0.956	0.957	0.193	0.194
Minnesota	1.016	0.017	0.953	0.951	0.268	0.266
Mississippi	0.946	0.010	0.953	0.950	0.345	0.339
Missouri	1.044	0.020	0.952	0.951	0.246	0.245
Montana	1.062	0.003	0.956	0.947	0.632	0.613
Nebraska	0.989	0.006	0.958	0.953	0.447	0.438
Nevada	0.941	0.008	0.957	0.953	0.404	0.397
New Hampshire	1.193	0.004	0.957	0.952	0.580	0.566
New Jersey	0.941	0.030	0.955	0.955	0.190	0.191
New Mexico	0.973	0.006	0.957	0.954	0.437	0.428
New York	1.019	0.066	0.950	0.956	0.130	0.133
North Carolina	1.024	0.029	0.952	0.953	0.206	0.206
Pennsylvania	1.160	0.042	0.953	0.957	0.166	0.167
Texas	0.875	0.077	0.954	0.957	0.125	0.129

that is a part of a larger region. This paper fills that gap by providing the needed confidence interval formulae, namely, the F-based and the normal-based intervals. It is noticeable that these two intervals depend on the ratio of the population sizes for the subregion and the region, and on their age-adjusted rates. The results in Tables 1-3 show the effect of the size of the overlap in the two populations on the confidence intervals. To avoid the situation where the observed age-adjusted rates are zero, we adopted a corrected version of age-adjusted rates by adding a small constant as in Tiwari *et al.* (2006).

We note the framework of the normal-approximation method allows the computation for the following two common scenarios in cancer surveillance. The first concerns with a partial overlapping situation. Consider X and Ω two regions

Table 3: Performance of the Derived Confidence Intervals based on the 2004 Lung Cancer Mortality Data. The empirical coverage probabilities and the widths of the intervals were based on 10,000 simulations.

State	Rate Ratio with US	Overlap ratio (p_X)	Empirical Coverage Prob.		Average Width	
			F-based	Normal-based	F-based	Normal-based
California	0.797	0.122	0.932	0.959	0.026	0.027
Colorado	0.740	0.016	0.952	0.953	0.075	0.076
Connecticut	0.936	0.012	0.955	0.956	0.085	0.085
Delaware	1.234	0.003	0.948	0.946	0.203	0.201
Washington DC	0.959	0.002	0.953	0.949	0.228	0.225
Florida	1.012	0.059	0.951	0.961	0.036	0.037
Georgia	1.101	0.030	0.948	0.950	0.065	0.066
Hawaii	0.703	0.004	0.956	0.955	0.123	0.122
Idaho	0.772	0.005	0.953	0.951	0.133	0.132
Illinois	1.029	0.043	0.946	0.952	0.048	0.049
Indiana	1.145	0.021	0.950	0.951	0.072	0.073
Iowa	1.009	0.010	0.952	0.952	0.094	0.093
Kansas	1.017	0.009	0.952	0.952	0.102	0.102
Kentucky	1.448	0.014	0.953	0.954	0.099	0.099
Louisiana	1.255	0.015	0.954	0.954	0.091	0.091
Maine	1.145	0.004	0.952	0.951	0.147	0.146
Maryland	0.998	0.019	0.948	0.949	0.074	0.074
Massachusetts	0.980	0.022	0.953	0.955	0.064	0.065
Michigan	1.068	0.034	0.955	0.958	0.054	0.055
Minnesota	0.881	0.017	0.948	0.949	0.071	0.072
Mississippi	1.282	0.010	0.952	0.952	0.114	0.113
Missouri	1.224	0.020	0.950	0.952	0.075	0.076
Montana	0.921	0.003	0.953	0.950	0.164	0.162
Nebraska	0.888	0.006	0.952	0.951	0.119	0.118
Nevada	1.075	0.008	0.950	0.950	0.122	0.122
New Hampshire	1.041	0.004	0.957	0.955	0.153	0.152
New Jersey	0.913	0.030	0.952	0.954	0.053	0.054
New Mexico	0.673	0.006	0.953	0.951	0.103	0.102
New York	0.862	0.066	0.952	0.961	0.034	0.035
North Carolina	1.111	0.029	0.953	0.955	0.061	0.061
Pennsylvania	0.998	0.042	0.949	0.954	0.044	0.044
Texas	0.977	0.077	0.948	0.952	0.037	0.039

with a partial overlap, for example $X = \{\text{Geogia, North Carolina, South Carolina}\}$ and $\Omega = \{\text{North Carolina, Virginia}\}$, with $X\Omega = \{\text{North Carolina}\}$. Also, a Taylor series expansion, similar to that in Section 2.2, for the rate-ratio

$$\frac{R_X}{R_\Omega} \doteq \frac{p_{X\Omega}R_{X\Omega} + p_{X\Omega^c}R_{X\Omega^c}}{q_{X\Omega}R_{X\Omega} + q_{X^c\Omega}R_{X^c\Omega^c}}$$

can be used to obtain an approximate normal CI. Here, $p_{X\Omega} = n_{X\Omega}/n_X, p_{X\Omega^c} = n_{X\Omega^c}/n_X, q_{X\Omega} = n_{X\Omega}/n_\Omega, q_{X^c\Omega} = n_{X^c\Omega}/n_\Omega$. The second scenario concerns with comparing two age-adjusted incidence rate ratios, namely (i) $R^{(L)}_{XW}/R^{(T)}_{XW}$ with $R^{(L)}_{XB}/R^{(T)}_{XB}$, and (ii) $R^{(L)}_{XW}/R^{(T)}_{XW}$ with $R^{(L)}_{\Omega W}/R^{(T)}_{\Omega W}$, where $R^{(L)}_{XW} = \sum_j w_j d^{(L)}_{XWj}/n_{XWj}$ is the age-adjusted incidence rate for white women with localized breast cancer

in subregion $X \subset \Omega$, and $R^{(T)}_{XW} = \sum_j w_j d^{(T)}_{XWj}/n_{XWj}$ is the age-adjusted rates for white women with breast cancer with localized, regional, and distant stages combined in X, while $R^{(L)}_{XB} = \sum_j w_j d^{(L)}_{XBj}/n_{XBj}$ and $R^{(T)}_{XB} = \sum_j w_j d^{(T)}_{XBj}/n_{XBj}$ are the corresponding age-adjusted incidence rates for black women in subregion X respectively. The terms $R^{(L)}_{\Omega W} = \sum_j w_j d^{(L)}_{\Omega Wj}/n_{\Omega Wj}$ and $R^{(T)}_{\Omega W} = \sum_j w_j d^{(T)}_{\Omega Wj}/n_{\Omega Wj}$ are defined for the age-adjusted incidence rates for white women in Region Ω similarly. Our focus centers on constructing the confidence intervals based on the following differences of ratios, namely,

$$DR_{X,W-B} = \frac{R^{(L)}_{XW}}{R^{(T)}_{XW}} - \frac{R^{(L)}_{XB}}{R^{(T)}_{XB}}, \quad \text{and} \quad DR_{X-\Omega,W} = \frac{R^{(L)}_{XW}}{R^{(T)}_{XW}} - \frac{R^{(L)}_{\Omega W}}{R^{(T)}_{\Omega W}}.$$

Indeed, the approximate $(1-\alpha)\times 100\%$ confidence intervals based on $DR_{X,W-B}$ and $DR_{X-\Omega,W}$ are $DR_{X,W-B} \mp Z_{\alpha/2} \times (\widehat{V}(DR_{X,W-B}))^{1/2}$ and $DR_{X-\Omega,W} \mp Z_{\alpha/2} \times (\widehat{V}(DR_{X-\Omega,W}))^{1/2}$, respectively, where $\widehat{V}(DR_{X,W-B})$ and $\widehat{V}(DR_{X-\Omega,W})$ are the variance estimates of $DR_{X-\Omega,W}$ and $DR_{X-\Omega,W}$ given in the Appendix. As an application, we computed the 95% confidence interval for $DR_{X,W-B}$, the difference of the age-adjusted rate ratio for white women with localized breast cancer to all stages combined with that for the women from the SEER-9 database (year 2003). The observed difference was 10.87% with the 95% CI being (8.32%, 13.41%), demonstrating racial disparities as the rate-ratio for the whites were significantly different from the black women. On the other hand, an application of $DR_{X-\Omega,W}$ was made to compare the rate-ratio for white women in Iowa with that for all white women in the SEER-9 database. The calculated difference was 1.01% with the 95% CI being (-1.20%, 3.22%), revealing no significant differences.

5. Conclusion

We derived confidence intervals based on F and normal approximations for the ratio of age-adjusted rates for a subregion to a parent region containing the subregion. Through simulations, we showed that all the proposed intervals performed well, in terms of retaining the nominal coverage. Our work fills the gap of non-availability of methods that compare the ratio of age-adjusted rates for two overlapping regions. We will suggest that these formulae be implemented in the NCI SEER*STAT software for public use.

References

Fay, M.P. (1999). Approximate Confidence Intervals for Rate Ratios from Directly Standardized Rates with Sparse Data. *Communications in Statistics, Theory and Methods* **28,** 2141-2160.

Tiwari, R.C., Clegg, L. and Zou, Z. (2006) Efficient interval estimation for age-adjusted cancer rates. *Statistical Methods in Medical Research* **15,** 547-569.

Pickle, L.W. and White, A.A. (1995). Effects of the choice of age-adjustment method on maps of death rates. *Statistics in Medicine* **14,** 615-627.

Appendix: Technical Detail

Using the delta method and applying the results derived in Section 2.1, we obtain the estimates of the variance expressions for $DR_{X,W-B}$ and $DR_{X-\Omega,W}$ as follows:

$$\begin{aligned}
&\widehat{V}(DR_{X,W-B}) = \\
&\frac{1}{(R_{XW}^{(T)})^4}\left\{(R_{XW}^{(T)})^2\widehat{V}(R_{XW}^{(L)}) + (R_{XW}^{(L)})^2\widehat{V}(R_{XW}^{(T)}) - 2R_{XW}^{(T)}R_{XW}^{(L)}\widehat{Cov}(R_{XW}^{(L)}, R_{XW}^{(T)})\right\} \\
&\quad + \frac{1}{(R_{XB}^{(T)})^4}\left\{(R_{XB}^{(T)})^2\widehat{V}(R_{XB}^{(L)}) + (R_{XB}^{(L)})^2\widehat{V}(R_{XB}^{(T)}) - 2R_{XB}^{(T)}R_{XB}^{(L)}\widehat{Cov}(R_{XB}^{(L)}, R_{XB}^{(T)})\right\}
\end{aligned}$$

with $\widehat{Cov}(R_{XW}^{(L)}, R_{XW}^{(T)}) = \sum_j w_j^2 d_{XWj}^{(L)}/n_{XWj}^2$ and $\widehat{Cov}(R_{XB}^{(L)}, R_{XB}^{(T)}) = \sum_j w_j^2 d_{XBj}^{(L)}/n_{XBj}^2$, and

$$\begin{aligned}
&\widehat{V}(DR_{X-\Omega,W}) = \\
&\frac{1}{(R_{XW}^{(T)})^4}\left\{(R_{XW}^{(T)})^2\widehat{V}(R_{XW}^{(L)}) + (R_{XW}^{(L)})^2\widehat{V}(R_{XW}^{(T)}) - 2R_{XW}^{(T)}R_{XW}^{(L)}\widehat{Cov}(R_{XW}^{(L)}, R_{XW}^{(T)})\right\} \\
&+\frac{1}{(R_{\Omega W}^{(T)})^4}\left\{(R_{\Omega W}^{(T)})^2\widehat{V}(R_{\Omega W}^{(L)}) + (R_{\Omega W}^{(L)})^2\widehat{V}(R_{\Omega W}^{(T)}) - 2R_{\Omega W}^{(T)}R_{\Omega W}^{(L)}\widehat{Cov}(R_{\Omega W}^{(L)}, R_{\Omega W}^{(T)})\right\} \\
&\quad - \frac{2}{(R_{XW}^{(T)}R_{\Omega W}^{(T)})^2}\left\{R_{XW}^{(T)}R_{\Omega W}^{(T)}\widehat{Cov}(R_{XW}^{(L)}, R_{\Omega W}^{(L)}) + R_{XW}^{(L)}R_{\Omega W}^{(L)}\widehat{Cov}(R_{XW}^{(T)}, R_{\Omega W}^{(T)})\right. \\
&\quad \left. -R_{XW}^{(L)}R_{\Omega W}^{(T)}\widehat{Cov}(R_{XW}^{(T)}, R_{\Omega W}^{(L)}) - R_{XW}^{(T)}R_{\Omega W}^{(L)}\widehat{Cov}(R_{XW}^{(L)}, R_{\Omega W}^{(T)})\right\}
\end{aligned}$$

with $\widehat{Cov}(R_{\Omega W}^{(L)}, R_{\Omega W}^{(T)}) = \sum_j w_j^2 d_{\Omega Wj}^{(L)}/n_{\Omega Wj}^2$ and $\widehat{Cov}(R_{XW}^{(L)}, R_{\Omega W}^{(L)}) = \widehat{Cov}(R_{XW}^{(T)}, R_{\Omega W}^{(T)}) = \widehat{Cov}(R_{XW}^{(T)}, R_{\Omega W}^{(L)}) = \widehat{Cov}(R_{XW}^{(L)}, R_{\Omega W}^{(T)}) = \sum_j w_j^2 d_{XWj}^{(L)}/(n_{XWj} \times n_{\Omega Wj})$, and $\widehat{Cov}(R_{XW}^{(L)}, R_{XW}^{(T)})$ defined earlier.

Received September 24, 2008; accepted October 29, 2008.

Ram C. Tiwari
Food and Drug Administration
Center for Drug Evaluation & Research, FDA
10903 New Hampshire Ave.
WO Bldg. 21, Rm. 3524
Silver Spring, MD, 20993-0002, USA
ram.tiwari@fda.hhs.gov

Yi Li
Harvard School of Public Health
44 Binney Street, Boston MA 02115
yili@jimmy.harvard.edu

Zhaohui Zou
Information Management Services
12501 Prosperity Dr, Suite 200
Silver Spring, MD 20904, USA
ZouJ@imsweb.com