Published in final edited form as:
Biometrics. 2008 December ; 64(4): 1280–1286. doi:10.1111/j.1541-0420.2008.01002.x.

Comparing Trends in Cancer Rates Across Overlapping Regions

Yi Li and
Harvard School of Public Health and Dana-Farber Cancer Institute

Ram C. Tiwari
National Cancer Institute

Abstract

Monitoring and comparing trends in cancer rates across geographic regions or over different time periods has been one main task of the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) Program as it profiles health care quality as well as decides health care resource allocations within a spatial-temporal framework. A fundamental difficulty, however, arises when such comparisons have to be made for regions or time intervals that overlap, e.g. comparing the change in trends of mortality rates in a local area (e.g. the mortality rate of Breast Cancer in California) with a more global level (i.e. the national mortality rate of Breast Cancer). In view of sparsity of available methodologies, this paper develops a simple corrected Z-test that accounts for such overlapping. The performance of the proposed test over the two-sample “pooled” t-test that assumes independence across comparison groups is assessed via the Pitman asymptotic relative efficiency as well as Monte Carlo simulations and applications to the SEER cancer data. The proposed test will be important for the SEER*STAT software, maintained by the NCI, for the analysis of the SEER data.

Keywords

Age-adjusted cancer rates; Annual percent change (APC); Surveillance; Trends; Hypothesis testing; Pitman asymptotic relative efficiency (ARE)

1 Introduction

Cancer continues to be a major epidemic concern in the United States, contributing the second most deaths each year in the United States. For instance, cancer resulted in approximately 570,280 deaths in year 2005 (American Cancer Society, 2005), while the overall cost of cancer, including the costs of diagnosis, treatment, lost person-hours, and education and research, tallied as much as $189.8 billion for 2004 (Ghosh and Tiwari, 2007).

Many public and private agencies dealing with cancer and related problems depend on the rates of cancer deaths or new cases as an estimate of cancer burden for planning and resource allocation. Among these agencies, the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is the most authoritative and comprehensive source of information on cancer incidence and deaths in the United States, which currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately over a quarter of the entire US population. One main task of the SEER program is to routinely monitor and compare trends in cancer mortality and incidence rates across geographic regions or over different time periods. The data are analyzed by SEER*STAT software, which is maintained by the NCI, with the results periodically published at http://seer.cancer.gov/csr/. Indeed, this surveillance task has important social and economic ramifications, ranging from deciding which cancer programs get funded to deciding how the funds are allocated among various regions. Having reliable and accurate comparisons of trends of cancer rates is thus of tremendous importance.

However, a fundamental statistical difficulty arises when such comparisons, largely for policy making purposes, have to be made for regions or time intervals that overlap, e.g. comparing the most recent changes in trends of cancer rates in a local area (e.g. the mortality rate of breast cancer in California) with a more global level (i.e. the national mortality rate) over two overlapping time periods, because of availability of the data. For example, as detailed in the data analysis section, it is of substantial interest to compare the changes in California cancer mortality rates with the national cancer mortality rates in the last 15 years. However, for a 15-year block, the California cancer rates were available for 1990–2004, while the national data were available for 1988–2002.

As the current SEER*STAT software utilizes the two-sample pooled t-test (Kleinbaum et al., 1988) that assumes independence across comparison groups, it is not appropriate for the aforementioned settings. In this paper, we develop a simple corrected Z-test that accounts for the overlap and that will be available for the NCI SEER program.

The rest of this article is structured as follows. In Section 2, we introduce the cancer rate regression model that has been used in the SEER analysis, followed by the classical t-test, employed by the current SEER*STAT software for comparing the trends between two independent regressions in Section 3. In Section 4, we propose a corrected Z-test that properly accounts for correlation when the comparison has to be made across two overlapping regions or time intervals. The performance of the proposed test is assessed via applications to the SEER cancer data, with its validity confirmed by simulations in Section 5. We conclude with a short summary in Section 6. The technical detail is relegated to the WebAppendix (http://www.biometrics.tibs.org/).

2 Age-adjusted Cancer Rate Regression Model and Annual Percent Change

Let n_{ji} and d_{ji} be the mid-year population at risk and counts of deaths or incidents for age group j ($j = 1,\ldots,J$) at time t_i, $i = 1,\ldots,I$. The age-adjusted rate, at time t_i, is typically computed as

$$\tilde{r}_i=\sum_{j=1}^{J} w_j \frac{d_{ji}}{n_{ji}}, \tag{1}$$

where $w_j > 0$, $j = 1,\ldots,J$, are the known standards for the age group j so that $\sum_{j=1}^{J} w_j=1.$ In the SEER program, there are $J = 19$ standard age-groups consisting of 0–1, 1–4, 5–9,…,85+, and the specific weights w_j are given in Fay et al. (2006).

To describe the trend in mortality or incidence, we often use a logarithm transformation of $\tilde{r}_i$ and fit a linear regression on the calendar time. However, for rare cancers, $\tilde{r}_i$ defined in (1) can be zero, making its logarithm transformation overflow in computation. To avoid this situation, we introduce a correction factor, which amounts to distributing a count of 1 uniformly to all J categories, and hence adding $1/J$ to d_{ji}, yielding a zero-corrected rate (Tiwari et al., 2006)

$$r_i=\sum_{j=1}^{J} w_j \frac{d_{ji}+1/J}{n_{ji}}. \tag{2}$$

Numerically, the difference between (1) and (2) is negligible; however, the logarithm of the latter is always defined. A simple linear regression has been established by a number of authors (Kim et al. 2000;Tiwari et al., 2005;Fay et al., 2006) to link the logarithm transformation of mortality or incidence rate r_i, say, $y_i = \log(r_i)$, to the calendar time t_i, via

$$y_i=\beta_0+\beta_1 t_i+e_i, \tag{3}$$

where the e_i are i.i.d. normal with mean 0 and variance σ^2, which measures the fluctuation of rates over years.

Model (3) is commonly referred to as the (transformed) Cancer Rate Regression Model in the SEER analysis (see e.g. Kim et al. 2000;Tiwari et al., 2005;Fay et al., 2006), which can be conveniently fitted for observed data (t_i, y_i), i = 1,…,I, using the least squares or the maximum likelihood estimation methodologies. The resulting estimates of $\beta = (\beta_0, \beta_1)$ are denoted by $\widehat{\beta}=(\widehat{\beta}_0,\widehat{\beta}_1)$.

Regression coefficient β_1 in (3) has been of main interest, as it transcribes the trends of mortality or incidence. Indeed, the annual percent change (APC), defined as $APC = 100(e^{\beta 1} - 1)$, has been used by the NCI (see e.g. Fay et al., 2006) for describing the trends of cancer incidence and mortality. Its estimate, $\widehat{APC}=100(e^{\widehat{\beta}_1}-1)$, along with its variance, obtained via the delta method (Ries et al., 2003;Fay et al., 2006), $\widehat{V}=10^4 e^{2\widehat{\beta}_1}\widehat{\sigma}^2_{\widehat{\beta}_1}$, constitutes the basis of drawing inference on the trend, e.g. constructing confidence intervals or testing hypothesis. Here, $\widehat{\sigma}^2_{\widehat{\beta}_1}=\widehat{\sigma}^2/\sum_{i=1}^{I}(t_i-\overline{t})^2$ and the unbiased estimator $\widehat{\sigma}^2=\sum_{i=1}^{I}(y_i-\widehat{y}_i)^2/(I-2)$, where $\overline{t}=\sum_{i=1}^{I}t_i/I$ and $\widehat{y}_i$ is a prediction of y_i based on (3), namely, $\widehat{y}_i=\widehat{\beta}_0+\widehat{\beta}_1 t_i$.

For the purpose of health-care evaluations, it is of substantial interest to compare the APC of one region (e.g., county or state level) to that of another region, or to a more global level (e.g. state or national level). One may also be interested in comparing the APCs over two overlapping intervals. In the following, we derive the tests for comparing APCs of two overlapping regions within two overlapping time intervals, which includes the aforementioned local-vs-global comparison as a special case.

3 Test for Equality of APCs for Two Independent Regressions

To start, we briefly review the test for comparing APCs for two independent comparison groups, e.g. for two non-overlapping regions or time intervals. That is, we consider two independent linear regressions

$$y_{ki}=\beta_{k0}+\beta_{k1}t_{ki}+e_{ki}, i=1,\ldots,I_k, \tag{4}$$

for k = 1, 2, flagging groups 1 and 2, respectively.

Let APC_1 and APC_2 be the corresponding APC values for these two regressions. Often, we wish to test the null hypothesis $H_0 : APC_1 = APC_2$ versus the alternative hypothesis $H_1 : APC_1 \neq APC_2$, which is equivalent to testing $H'_0 : \beta_{11} = \beta_{21}$ versus $H'_1 : \beta_{11} \neq \beta_{21}$. Under the assumption that error variances for the two groups are equal, a test for the latter is given by Kleinbaum et al. (1988):

$$t=\frac{\widehat{\beta}_{11}-\widehat{\beta}_{21}}{\sqrt{S_p^2\left[\{\sum_{i=1}^{I_1}(t_{1i}-\overline{t}_1)^2\}^{-1}+\{\sum_{i=1}^{I_2}(t_{2i}-\overline{t}_2)^2\}^{-1}\right]}} \sim t(I_1+I_2-4), \tag{5}$$

where

$$\overline{t}_k=\sum_{i=1}^{I_k}t_{ki}/I_k$$

for k = 1, 2, and S_p^2 is the “pooled” unbiased estimate of σ^2 given by

$$S_p^2=\frac{\sum_{i=1}^{I_1}(y_{1i}-\widehat{y}_{1i})^2+\sum_{i=1}^{I_2}(y_{2i}-\widehat{y}_{2i})^2}{I_1+I_2-4},$$

where $\widehat{y}_{ki}=\widehat{\beta}_{k0}+\widehat{\beta}_{k1}t_{1i}$ are the predictions for k = 1, 2. Test (5) is currently employed by the NCI SEER*STAT software (http://seer.cancer.gov/seerstat). We remark that in these tests an implicit assumption is that the total population in each time period is stable so that the variance of y_{ki} is time-independent. This is indeed the case for the SEER incidence/mortality data. Hence, we will make the same assumption throughout.

4 A Corrected Z-test for Two Dependent Regressions

Much difficulty arises as (5) is no longer valid if the independence assumption is violated. Suppose we are interested in comparing the APCs of two overlapping regions, say, Region 1 and Region 2, with data collected over two time intervals $[t_1, t_m]$ and $[t_{s+1}, t_{s+I}]$, which possibly overlap. That is, $t_1 \le t_{s+1} < t_m \le t_{s+I}$. We modify (4) to accommodate this situation

$$y_{1i}=\beta_{10}+\beta_{11}t_i+e_{1i},i=1,\ldots,m, \tag{6}$$

for Region 1, and

$$y_{2i}=\beta_{20}+\beta_{21}t_i+e_{2i},i=s+1,\ldots,s+1, \tag{7}$$

for Region 2. Let $\widehat{\beta}_{11}$ and $\widehat{\beta}_{21}$ be the estimates of the slope parameters of the regression lines for these two regions respectively. In particular,

$$\widehat{\beta}_{11}=\frac{\sum_{i=1}^{m}(t_i-\bar{t}_1)(y_{1i}-\bar{y}_1)}{\sum_{i=1}^{m}(t_i-\bar{t}_1)^2},\widehat{\beta}_{21}=\frac{\sum_{i=s+1}^{s+I}(t_i-\bar{t}_2)(y_{2i}-\bar{y}_2)}{\sum_{i=s+1}^{s+I}(t_i-\bar{t}_2)^2}.$$

where $\bar{y}_1=\sum_{i=1}^{m}y_{1i}/m,\ \bar{t}_1=\sum_{i=1}^{m}t_i/m,\ \bar{y}_2=\sum_{i=s+1}^{s+I}y_{2i}/I$ and $\bar{t}_2=\sum_{i=s+1}^{s+I}t_i/I.$

When Regions 1 and 2 are overlapping, the two regressions may not be independent and, hence, (5) will not be valid as it fails to account for the correlation between $\widehat{\beta}_{11}$ and $\widehat{\beta}_{21}$. Indeed, under the assumption that, errors e_{1i} and e_{2i} are i.i.d. normal with mean 0 and equal variance σ^2 for the two regions,

$$\widehat{\beta}_{11}-\widehat{\beta}_{21}\sim N\left\{\beta_{11}-\beta_{21},\sigma^2(\sigma_1^{-2}+\sigma_2^{-2})-2Cov(\widehat{\beta}_{11},\widehat{\beta}_{22})\right\},$$

where $\sigma_1^2=\sum_{i=1}^{m}(t_i-\bar{t}_1)^2,\ \sigma_2^2=\sum_{i=s+1}^{s+I}(t_i-\bar{t}_2)^{2}\cdot$ It turns out that the derivation of $Cov(\widehat{\beta}_{11},\widehat{\beta}_{21})$, when the two time intervals $[t_1, t_m]$ and $[t_{s+1}, t_{s+I}]$ under consideration are overlapping, is nontrivial as it requires a careful consideration of the overlapping of two regions. The detailed derivation is given in the WebAppendix, which shows that

$$Cov(\widehat{\beta}_{11},\widehat{\beta}_{21})\doteq\frac{\sigma^2\sigma_{12}}{\sigma_1^2\sigma_2^2}\frac{\{n^{(O)}\}^2}{n_1n_2}, \tag{8}$$

where $n_k=\sum_{i=s+1}^{m}\sum_{j=1}^{J}n_{kji}$ for $k=1,2,\ n^{(O)}=\sum_{i=s+1}^{m}\sum_{j=1}^{J}n_{ji}^{(O)}\cdot$ Here, we have used superscript ‘O’ to denote the intersection of Regions 1 and 2, and denoted by n_{kji} and $n_{ji}^{(O)}$ the numbers of underlying population at risk for age group j at time t_i in Region k(k = 1, 2), and in the overlapping subregion, respectively.

The cross term in (8)

$$\sigma_{12}=\sum_{i=s+1}^{m}(t_i-\bar{t}_1)(t_i-\bar{t}_2),$$

merits attention as it determines the sign of (8) and is completely decided by how $[t_1, t_m]$ overlaps with $[t_{s+1}, t_{s+I}]$. For example, when $[t_1, t_m]$ coincides with $[t_{s+1}, t_{s+I}]$ (i.e. $s = 0$, $m = I$), then $\bar{t}_1=\bar{t}_2$, and hence $\sigma_{12}=\sum_{i=1}^{m}(t_i-\bar{t}_1)^2>0.$ On the other hand, when $[t_1, t_m]$ only partially overlaps with $[t_{s+1}, t_{s+I}]$, σ_{12} can be negative, causing a negative covariance in (8). For example, when s is close to m such that $t_{s+1}>\bar{t}_1$ and $t_m<\bar{t}_2$, then $t_i-\bar{t}_1>0$ and $t_i-\bar{t}_2<0$ for any $i \in [s + 1, m]$, leading to $\sigma_{12} < 0$.

Note that when the overlapping region is an empty set, $n^{(O)} = 0$, and $Cov(\widehat{\beta}_1,\widehat{\beta}_2)=0$. When $s + 1 > m$ (i.e. the time intervals are non-overlapping), $\sigma_{12} = 0$ and, hence, $Cov(\widehat{\beta}_{11},\widehat{\beta}_{21})=0$ as well. On the other hand, if, for example, Region 1 is completely contained in Region 2, then $n_1 = n^{(O)}$, and

$$Cov(\widehat{\beta}_{11},\widehat{\beta}_{21})=\frac{\sigma^2\sigma_{12}}{\sigma_1^2\sigma_2^2}\frac{n_1}{n_2}.$$

So, in summary, if the two regions are non-overlapping (or time intervals are non-overlapping)

$$\widehat{\beta}_{11}-\widehat{\beta}_{21}\sim N\{\beta_{11}-\beta_{21},\sigma^2(\sigma_1^{-2}+\sigma_2^{-2})\} \tag{9}$$

and if Region 1 is completely contained in Region 2,

$$\widehat{\beta}_{11}-\widehat{\beta}_{21}\sim N\left\{\beta_{11}-\beta_{21},\sigma^2\left(\sigma_1^{-2}+\sigma_2^{-2}-2\sigma_{12}\sigma_1^{-2}\sigma_2^{-2}\frac{n_1}{n_2}\right)\right\},$$

where n_1/n_2 is typically termed as the *overlapping ratio*. In general, for two regions that overlap partially,

$$\widehat{\beta}_{11}-\widehat{\beta}_{21}\sim N\left\{\beta_{11}-\beta_{21},\sigma^2\left(\sigma_1^{-2}+\sigma_2^{-2}-2\sigma_{12}\sigma_1^{-2}\sigma_2^{-2}\frac{(n^{(O)})}{n_1n_2}\right)\right\}. \tag{10}$$

Eq. (10) reveals that its asymptotic efficacy (AE), defined by its noncentrality, is

$$\frac{(\beta_{11}-\beta_{21})^2}{\sigma^2\left\{\sigma_1^{-2}+\sigma_2^{-2}-2\sigma_{12}\sigma_1^{-2}\sigma_2^{-2}\ \frac{(n^{(O)})^2}{n_1n_2}\right)\}}, \tag{11}$$

compared to the AE of the naive test that ignores overlapping [cf. (9) or (5)]

$$\frac{(\beta_{11}-\beta_{21})^2}{\sigma^2\left(\sigma_1^{-2}+\sigma_2^{-2}\right)}. \tag{12}$$

Hence, the Pitman Asymptotic Relative Efficiency (ARE), which is the ratio of (11) and (12), and measures the gain of efficiency by accounting for overlapping, is

$$\text{ARE}=\frac{\sigma_1^{-2}+\sigma_2^{-2}}{\sigma_1^{-2}+\sigma_2^{-2}-2\sigma_{12}\sigma_1^{-2}\sigma_2^{-2}\frac{(n^{(O)})^2}{n_1n_2}}.$$

Several points are worth mentioning. First, when $n(^{O}) = 0$ (corresponding to disjoint regions) or $t_{s+1} > t_m$ (corresponding to disjoint time intervals), the Pitman ARE is 1, justifying the use of the classical test (as used in the current SEER*STAT software). Secondly (and interestingly), depending on the sign of σ_{12}, i.e the mixing of the time intervals, the ARE can be greater or less than 1. Specifically, when $\sigma_{12} > 0$, then $ARE > 1$, indicating the naive test will be too conservative; otherwise, $ARE < 1$, hinting that the naive test will be too aggressive and will not maintain the nominal type I error, all of which calls for a new test that accounts for overlapping. Finally, as a simple example, when $s = 0$, $m = I$ (i.e. two time intervals are identical), then $\sigma_{12}=\sigma_1^2=\sigma_2^2, \text{ and,hence, } ARE=\left\{1-\frac{(n^{(O)})^2}{n_1 n_2}\right\}^{-1}$, indicating that the naive test will always be too conservative and the efficiency loss will become more severe as the overlapping population $n^{(O)}$ becomes larger.

In practice, as σ^2 is unknown, we have to replace it with a consistent estimate, leading to the following Z-test,

$$Z=\frac{\widehat{\beta}_{11}-\widehat{\beta}_{21}}{\left\{\widehat{\sigma}^2\left(\sigma_1^{-2}+\sigma_2^{-2}-2\sigma_{12}\sigma_1^{-2}\sigma_2^{-2}\frac{(n^{(O)})^2}{n_1 n_2}\right)\right\}^{1/2}}. \qquad (13)$$

Under the null hypothesis, Z in (13) approximately follows a normal distribution, where an unbiased estimate for σ^2 is given by

$$\widehat{\sigma}^2=\frac{\sum_{i=1}^{m}(y_{1i}-\widehat{y}_{1i})^2+\sum_{i=s+1}^{s+I}(y_{2i}-\widehat{y}_{2i})^2}{m+I-4}.$$

5 Analysis of SEER Mortality Data and Simulation Studies

It is of substantial interest to compare the changes in cancer mortality rates in California with the national levels as a California law (Health and Safety Code, Section 103885) was passed in late 1980's that mandated the reporting of malignancies diagnosed throughout the state. For this purpose, we applied the proposed methodology to compare the annual percent change (APC) in the age-adjusted mortality rates for the United States (US) for the period from 1988–2002 to that of California (CA)for the period from 1990 to 2004. We fitted the simple linear models (4) to the logarithms of the age-adjusted mortality rates for both male and female for a number of cancer sites from the *Cancer Facts & Figures* (American Cancer Society, 2007). The mortality data for the United States are compiled by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (www.cdc.gov/nchs) and are available from the National Cancer Institute’s Surveillance, Epidemiology, and End Results SEER) Program (http://www.seer.cancer.gov). The ratio of the total population for all age-groups combined for CA to that for the US for the overlapping years (i.e. n_1/n_2) was around 11% for females, and 11.5% for males. Because of the space contraint, the results are summarized in Tables A.1 and A.2 in the WebAppendix. The tables give the estimates of the slope parameters for CA and US and their standard errors, along with the p-values for the comparisons based on the naive t- and the corrected Z- tests. The estimate of common residual variance σ^2 is also provided. We also calculated the residual variances for all the cancer sites for CA and the US separately (not reported in the tables), and found that they were close, confirming our common variance assumption.

The table shows that the corrected Z-test seems to more aggressively detect the difference between the two APCs than the t-test, yielding smaller p-values for all the cancer sites. For example, the corrected Z-test detected a significant difference in the APC between CA and the US on the site of Stomach in men (meaning CA has a more rapid decrease of Stomach cancer mortality rate compared to the US), while the naive t-test failed to detect such a difference at 5% type I error rate level.

We also compared annual percent change (APC) in the recent 15 years' age-adjusted mortality rates for California (1990 to 2004) to the national mortality rates during eighties and early nineties (1980–1994). Indeed, it was a common practice for policy-makers to evaluate the progress made at a state level by comparing with the historical national trends (see e.g. http://statecancerprofiles.cancer.gov/historicaltrend). Statistically, this comparison is also of interest. In particular, as $\sigma_{12} < 0$ in this case, the theoretical results in Section 4 hinted that the naive t-test would be too aggressive, and, hence, might 'exaggerate' the progress made in California. The ratio of the total population for all age-groups combined for CA to that for the US for the overlapping years (1990–1994) (i.e. n_1/n_2) was around 11.1% for females, and 11.4% for males. The results are summarized in Tables B.1 and B.2 in the WebAppendix. These tables show that the naive t-test was a bit more aggressive than the corrected Z-test, yielding slightly smaller p-values for all the cancer sites. This confirmed our theoretical results.

To further confirm our analysis results, simulation studies were performed to compare the characteristics of the naive t-test, based on (5), with the corrected Z-test (13) that properly accounts for overlapping. We conducted the following simulations to compare the APCs for two regions. We mimicked the comparision between, say, the Southern Region (Region 1) consisting of Georgia (GA), South Carolina (SC) and North Carolina (NC), and the Eastern Region (Region 2) consisting of NC, Virginia (VA) and Maryland (MD), with NC the overlapping state. The different time periods, with varying degree of overlap in the time intervals, are taken to be : **(Scenario 1)** years [1986,…,2001] for Region 1, and years [1989, …,2004] for Region 2 so that there a considerable overlap of 12 years between the two intervals and $\sigma_{12} = 152.75$; **(Scenario 2)** years [1978,…,1993] for Region 1, and years [1989,…,2004] for Region 2 so that there is a little overlap of three years between the two intervals and $\sigma_{12} = -141.25 < 0$. For generating the counts, d_{kji}, we assume that $d_{kji} \sim^{ind} Poisson(n_{kji}\lambda_{kji})$, where $\log(\lambda_{kji}) = \beta_{kj,0} + \beta_{k1}t_i$, with t_i taking values in the intervals corresponding to the two regions stated above. Define the age-adjusted rate at time t_i in Region k as $r_{ki}=\sum_{j=1}^{J} w_j d_{kji}/n_{kji}$. Then the specification for λ_{kji} leads to

$$\begin{aligned} E(r_{ki}) &= \exp(\beta_{k1}t_i)\sum_{j=1}^{J} w_j \exp(\beta_{kj,0}) \\ &= \exp(\beta_{k1}t_i)\beta_{k,0}. \end{aligned}$$

Hence, the delta method would yield that $E(y_{ki}) \equiv E(\log r_{ki}) = \log(B_{k,0}) + \beta_{k1}t_i$, where $APC_k = 100(e^{\beta k1} - 1)$.

Now to specify the regression for λ_{kji}, we take $\beta_{k1} = \log(100^{-1}APC_k + 1)$, based on the specified values of APC_k ranging from −0.3% to 1.0%, and compute $\beta_{kj,0}=\log(\frac{d_{kj,0}}{n_{kj,0}})$ where $d_{kj,0}$ and $n_{kj,0}$ are, respectively, the observed number of deaths and the number of at risk population at the "baseline" year, the beginning of the time interval considered for Region k. The age-specific counts for the overlapping state, NC at the overlapping time t_i are generated from Poisson distributions with means $n_{ji}^{(0)}(\lambda_{1ji}+\lambda_{2ji})/2$, where $n_{ji}^{(o)}$ denotes the number of at-risk population in the overlapping region (e.g. NC) in year t_i. In our simulation, the number of at-risk population and the observed number of deaths were obtained from the SEER database for all malignant male cancers and prostate cancer within the time intervals specified in Scenarios 1 and 2.

Table 1 and Table 2 display the powers of the corrected Z-test and the naive Kleinbaum's t-test (5) as a function of various APCs for the two regions, based on the two time-overlapping scenarios listed above. For each parameter configuration, a total of 10000 monte carlo samples were generated and the empirical powers were calculated. The results in Table 1 and Table 2 clearly showed that corrected test maintained the nominal type I error under both time-overlapping scenarios and had good power, which approached 1 quickly as the difference between the two APCs increased. On the contrary, the naive test did not maintain the nominal type I error. It was too conservative in Scenario 1 (where $\sigma_{12} > 0$ as in Table 1), with the type I error being around 0.031 under the null hypothesis, almost half less than the nominal level, and its power was obviously less than the corrected test, while in Scenario 2 (where $\sigma_{12} < 0$ as in Table 2), its type I error rate was around 0.075, almost 50% more than the nominal level. Hence, our simulation results verified the theoretical results.

6 Discussion

In this paper, we have considered an important problem where comparisons have to be made for regions or time intervals that overlap. We have shown that the existing methodology, which does not properly account for such overlapping, will be be inappropriate as it will not maintain the type I error. We have proposed a simple test that solves this fundamental difficulty and correctly accounts for overlapping. Simulations have indicated good performance of the proposed methodology. We have applied the developed methodology to the analysis of the major cancer sites from the SEER Program and have found that the corrected Z-test renders more power than the naive t-test. Hence, the proposed Z-test will be an important addition to the SEER*STAT software, which only handles independent comparisons at this time.

We have focused on the local linearity for the cancer rates by considering time periods of short or moderate length. Indeed, linearity assumption for the cancer rates is a debatable issue in cancer surveillance, which is likely to be violated over a longer period (e.g. ≥ 30 years). A detailed discussion on this issue has been made in Fay et al. (2006), which proposed a joinpoint linear regression for long-term cancer rate analysis. In a similar context, we plan to pursue APC comparisons for longer periods by considering joinpoint linear regressions, and will report the results in a subsequent communication.

7 Supplementary Materials

Web Appendix for the derivation of Equation (8) and Tables A.1, A.2, B.1 and B.2 referenced in Section 5 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank the Editor, the AE and an anonymous referee for insightful suggestions that improved the orginal manuscript. The authors thank Steve Scoppa and Joe Zou of Information Management Services (IMS), Inc., the company that provides biomedical computing support to the NCI, for their valuable contributions. The authors also thank Dr Rocky Feuer for carefully reading an early draft of this work and for many insightful suggestions.

Reference

American Cancer Society. Cancer Facts & Figures. Atlanta: Georgia; 2007.

Fay M, Tiwari R, Feuer E, Zou Z. Estimating average annual percent change for disease rates without assuming constant change. Biometrics 2006;62:847–854. [PubMed: 16984328]

Ghosh K, Tiwari R. Prediction of US cancer mortality counts using semi-parametric Bayesian techniques. Journal of the American Statistical Association 2007;102:7–15.

Kim H, Fay M, Feuer E, Midthune D. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine 2000;19:335–351. [PubMed: 10649300]

Kleinbaum, D.; Kupper; Muller, P. Applied Regression Analysis and Other Multivariable Methods. 2nd edition. Boston: PWS-Kent; 1988.

Pickle LW, White AA. Effects of the choice of age-adjustment method on maps of death rates. Statistics in Medicine 1995;14:615–627. [PubMed: 7792452]

Ries, L.; Eisner, M.; Kosary, C.; Hankey, B.; Miller, B.; Clegg, L.; Mariotto, A.; Feuer, E.; Edwards, BK., editors. SEER Cancer Statistics Review, 1975–2002. Bethesda, MD: National Cancer Institute; 2003. http://seer.cancer.gov/csr/1975-2002/

Tiwari R, Cronin K, Davis W, Feuer E, Yu B, Chib S. Bayesian model selection for join point regression with application to age-adjusted cancer rates. Journal of the Royal Statistical Society: Series C (Applied Statistics) 2005;54:919–939.

Tiwari R, Clegg L, Zou Z. Effcient interval estimation for age-adjusted cancer rates. Statistical Methods in Medical Research 2006;15:547–569. [PubMed: 17260923]

Table 1 (for time-overlapping scenario 1)

Comparison of the corrected Z-test and the naive t-test for two overlapping regions over roughly the same time interval (Region 1: 1986–2001 vs. Region 2: 1989–2004), in which case $\sigma_{12} = 152.75 > 0$. APC1 and APC2 are the annual percent changes in Regions 1 and 2, respectively.

site	APC1	APC2	naive T-test	Corrected Z-test	estimated residual variance
All Malignant Cancers	−0.5	−0.5	0.030	0.050	5.71E-05
	−0.5	−0.3	0.940	0.960	5.68E-05
	−0.5	−0.1	0.99	1.00	5.63E-05
	−0.3	−0.3	0.032	0.047	5.63E-05
	−0.3	−0.1	0.930	0.960	5.59E-05
	−0.3	0.1	0.99	1.00	5.56E-05
	−0.1	−0.1	0.032	0.048	5.55 E-05
	−0.1	0.1	0.94	0.96	5.52E-05
	−0.1	0.3	0.99	1.00	5.49E-05
	0.1	0.1	0.030	0.050	5.47E-05
	0.1	0.3	0.94	0.96	5.44E-05
	0.1	0.5	0.99	1.00	5.44E-05
	0.3	0.3	0.032	0.050	5.40E-05
	0.3	0.5	0.94	0.96	5.36E-05
	0.3	0.7	0.99	1.00	5.33E-05
	0.5	0.5	0.031	0.050	5.32E-05
	0.5	0.7	0.94	0.97	5.26E-05
	0.5	1.0	0.99	1.00	5.26E-05
Prostate Cancer	−0.5	−0.5	0.031	0.050	0.000480
	−0.5	−0.3	0.188	0.242	0.000477
	−0.5	−0.1	0.655	0.716	0.000475
	−0.5	0.1	0.94	0.97	0.000472
	−0.3	−0.3	0.031	0.050	0.000474
	−0.3	−0.1	0.191	0.245	0.000471
	−0.3	0.1	0.660	0.721	0.000468
	−0.3	0.3	0.95	0.97	0.000466
	−0.1	−0.1	0.031	0.049	0.000467
	−0.1	0.1	0.193	0.247	0.000465
	−0.1	0.3	0.665	0.724	0.000462
	−0.1	0.5	0.95	0.97	0.000459
	0.1	0.1	0.031	0.049	0.000461
	0.1	0.3	0.196	0.250	0.000458
	0.1	0.5	0.670	0.727	0.000456
	0.1	0.7	0.953	0.970	0.000453
	0.3	0.3	0.031	0.049	0.000455
	0.3	0.5	0.198	0.250	0.000452
	0.3	0.7	0.673	0.733	0.000450
	0.3	1.0	0.990	0.994	0.000446
	0.5	0.5	0.032	0.050	0.000449
	0.5	0.7	0.200	0.253	0.000446
	0.5	1.0	0.860	0.894	0.000443

Table 2 (for time-overlapping scenario 2)

Comparison of the corrected Z-test and the naive t-test for two overlapping regions over partially overlapping intervals (Region 1: 1978–1993 vs. region 2: 1989–2004), in which case $\sigma_{12} = -141.25 < 0$. APC1 and APC2 are the annual percent changes in Regions 1 and 2, respectively.

site	APC1	APC2	naive T-test	Corrected Z-test	estimated residual variance
All Malignant Cancers	−0.5	−0.5	0.075	0.052	6.81E-05
	−0.5	−0.3	0.840	0.830	6.84E-05
	−0.5	−0.1	1.00	0.99	6.82E-05
	−0.3	−0.3	0.075	0.058	6.77E-05
	−0.3	−0.1	0.843	0.831	6.78E-05
	−0.3	0.1	1.00	1.00	6.74E-05
	−0.1	−0.1	0.074	0.052	6.67E-05
	−0.1	0.1	0.851	0.831	6.65E-05
	−0.1	0.3	1.00	1.00	6.62E-05
	0.1	0.1	0.073	0.054	6.58E-05
	0.1	0.3	0.854	0.831	6.54E-05
	0.1	0.5	1.00	0.99	6.53E-05
	0.3	0.3	0.071	0.054	6.62E-05
	0.3	0.5	0.859	0.846	6.51E-05
	0.3	0.7	1.00	1.00	6.44E-05
	0.5	0.5	0.075	0.054	6.42E-05
	0.5	0.7	0.862	0.849	6.40E-05
	0.5	1.0	1.00	1.00	6.35E-05
Prostate Cancer	−0.5	−0.5	0.076	0.052	0.00061
	−0.5	−0.3	0.198	0.189	0.00061
	−0.5	−0.1	0.540	0.517	0.00061
	−0.5	0.1	0.850	0.837	0.00061
	−0.3	−0.3	0.075	0.050	0.00060
	−0.3	−0.1	0.199	0.183	0.00059
	−0.3	0.1	0.544	0.523	0.00058
	−0.3	0.3	0.854	0.840	0.00057
	−0.1	−0.1	0.075	0.058	0.00057
	−0.1	0.1	0.201	0.185	0.00057
	−0.1	0.3	0.545	0.526	0.00057
	−0.1	0.5	0.95	0.97	0.00057
	0.1	0.1	0.074	0.053	0.00057
	0.1	0.3	0.203	0.186	0.00057
	0.1	0.5	0.550	0.530	0.00057
	0.1	0.7	0.859	0.845	0.00057
	0.3	0.3	0.075	0.051	0.00057
	0.3	0.5	0.205	0.188	0.00057
	0.3	0.7	0.555	0.533	0.00057
	0.3	1.0	0.941	0.933	0.00057
	0.5	0.5	0.075	0.052	0.00056
	0.5	0.7	0.205	0.190	0.00056
	0.5	1.0	0.736	0.717	0.00056