An Age-Stratified Poisson Model for Comparing Trends in Cancer Rates Across Overlapping Regions

Yi Li[*1], **Ram C. Tiwari** [2], and **Zhaohui Zou** [3]

[1] Harvard University, 44 Binney Street, Boston, MA 02115, USA
[2] National Cancer Institute, Bethesda, MD 20892-8317
[3] Information Management Services, Silver Spring, MD 20904

Received 15 November 2007, revised 28 February 2008, accepted 25 April 2008
Published online

Summary

The annual percent change (APC) has been used as a measure to describe the trend in the age-adjusted cancer incidence or mortality rate over relatively short time intervals. The yearly data on these age-adjusted rates are available from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute. The traditional methods to estimate the APC is to fit a linear regression of logarithm of age-adjusted rates on time using the least squares method or the weighted least squares method, and use the estimate of the slope parameter to define the APC as the percent change in the rates between two consecutive years. For comparing the APC for two regions, one uses a t-test which assumes that the two datasets on the logarithm of the age-adjusted rates are independent and normally distributed with a common variance. Two modifications of this test, when there is an overlap between the two regions or between the time intervals for the two datasets have been recently developed. The first modification relaxes the assumption of the independence of the two datasets but still assumes the common variance. The second modification relaxes the assumption of the common variance also, but assumes that the variances of the age-adjusted rates are obtained using Poisson distributions for the mortality or incidence counts. In this paper, a unified approach to the problem of estimating the APC is undertaken by modeling the counts to follow an age-stratified Poisson regression model, and by deriving a corrected Z-test for testing the equality of two APCs. A simulation study is carried out to assess the performance of the test and an application of the test to compare the trends, for a selected number of cancer sites, for two overlapping regions and with varied degree of overlapping time intervals is presented.

Key words: Age-adjusted incidence/mortality rates, age-stratified Poisson Regression, annual percent change (APC), surveillance, trends, hypothesis testing.

1 Introduction

The American Cancer Society (ACS) in its annual publication Cancer *ACS 2007* (http://www.cancer.org/) reports that in 2007 about 1.5 million new cancer cases are expected to be diagnosed, and approximately 560,000 Americans are expected to die of cancer. Cancer is the most common cause of death in US, exceeded only by heart disease, and accounts for 1 of every 4 deaths. The same report also reveals that, for a number of cancer sites (such as breast, stomach, colon and rectum, lung and bronchus and leukemia), the age-adjusted cancer mortality rates have been steadly decreasing in recent years. In addition, the National Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) periodically publishes similar reports on trends of cancer incidence at http://seer.cancer.gov/csr; see Ries et al. (2003) So much has been at stake in terms of human life and cost - for example, the government agencies such as the National Health of Institutes (NIH), and many private sectors spend billions of dollars every year on cancer research, health insurance and medical and other costs - that there is an urgent need for new methods that produce more accurate and reliable estimates of measures of cancer trends.

* Corresponding author: e-mail: yili@jimmy.harvard.edu, Phone: 617 632 5134, Fax: 617 632 2444

The annual percent change (APC) has been used as a measure of cancer trends over short time periods, and to compare the recent cancer trends by gender or by geographic regions, one compares their APC values using the two-sample pooled t-test (Kleinbaum et al., 1988) that assumes that the datasets on age-adjusted rates under the study are independent. 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. 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.

Recently, Li and Tiwari (2007) and Li et al. (2007) developed Z-tests which adjust for the dependence between the two APCs, and are more efficient than the naive test which assumes independence. However, these tests are based on the logarithmic transformation of the age-adjusted rates, and fits a simple linear regression model of the transformed data on time using either the ordinary least squares (OLS) or the other weighted least squares (WLS) procedures. The proposed test procedure is based on the natural assumption that the age-specific mortality or incidence counts are results of underlying Poisson processes (Brillinger, 1986), and hence are realizations of independent Poisson random variables. The age-specific instantaneous hazards are modeled by a log-link function, thus leading to an age-stratified Poisson model. The estimation of the parameters is then carried out using a likelihood-based approach.

The rest of the paper is organized as follows. In Section 2, we briefly review the existing tests, and derive the new test in Section 3. To compare the performance of the proposed test with respect to the above mentioned tests, a simulation study is carried out in Section 4. In this section, we also give application to breast cancer mortality data from California (CA) and the US extracted from the SEER*STAT software of the SEER Program. Section 5 ends this paper with a short discussion.

2 A Brief Review of Existing Tests

Consider two regions, and let d_{kji} denote the number of counts (deaths or new cancer cases) from the population at risk n_{kji} observed in Region k ($k = 1, 2$) in age-group j ($j = 1, \ldots, J$) and at times $T_1, \ldots, T_m$ for Region 1 and $T_{s+1}, \ldots, T_{s+n}$ for Region 2, where $T_1 \leq T_{s+1} < T_m \leq T_{s+n}$, with $0 \leq s < m$ leading to overlapping time intervals. Note that this formulation is general and allows one region to have fewer time points than the other. In the SEER program, it is common to choose n_{kji} (at year T_i) to be the mid-year population representing the total person-years in one year, with the assumption of "drop-outs" being uniform over the unit-intervals. The age-adjusted rates are defined as

$$r_{ki} = \sum_{j=1}^{J} w_j \frac{d_{kji}}{n_{kji}},$$

where $w_j > 0, j = 1, \ldots, J$, are the known standards for the age group j so that $\sum_{j=1}^{J} w_j = 1$. For the SEER analysis, there are $J = 19$ standard age-groups consisting of 0-1, 1-4, 5-9, $\ldots, 85+$, and w_j are chosen to be the year 2000 population standards (Fay et al. 2006).

Let $y_{ki} = \log(r_{ki})$, be the logarithmic transformations of the age-adjusted rates. Consider the linear regression models

$$y_{ki} = \beta_{0k} + \beta_{1k} t_{ki} + e_{ki}, i = 1, \ldots, I_k, \tag{1}$$

for $k = 1, 2$, flagging Regions 1 and 2, respectively. Here e_{ki} are random errors with mean 0, and t_{ki} corresponds to the calendar times of data collection in region k with $I_1 = m$ and $I_2 = n$. More specifically, $(t_{11}, \ldots, t_{1I_1}) = (T_1, \ldots, T_m)$, while $(t_{21}, \ldots, t_{2I_2}) = (T_{s+1}, \ldots, T_{s+n})$. For the two regions, the

annual percent change (APC) are defined as $APC_k = 100(e^{\beta_{1k}} - 1) \doteq 100\beta_{1k}$, for a small β_{1k}, e.g. in the order of 10^{-2} (Kim et al., 2000; Fay et al., 2006; Tiwari et al., 2006).

Then under the assumptions that e_{ki} are independent and have common variance σ^2, the two-sample pooled t-test (Kleinbaum et al., 1988) for testing the null hypothesis $H_0 : APC_1 = APC_2$ versus the alternative $H_a : APC_1 \neq APC_2$ is given by

$$T_t = \frac{\hat{\beta}_{11} - \hat{\beta}_{12}}{\sqrt{\hat{\sigma}^2 \left((\sum_{i=1}^{I_1} (t_{1i} - \bar{t}_1)^2)^{-1} + (\sum_{i=1}^{I_2} (t_{2i} - \bar{t}_2)^2)^{-1} \right)}} \sim t_{(I_1+I_2-4)}, \tag{2}$$

where $\bar{t}_k = \sum_{i=1}^{I_k} t_{ki}/I_k$ for $k = 1, 2$, and $\hat{\sigma}^2$ is the "pooled" unbiased estimate of σ^2 given by

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

where $\hat{y}_{ki} = \hat{\beta}_{0k} + \hat{\beta}_{1k} t_{ki}$ are the predictions for $k = 1, 2$. Here, $\hat{\beta}_{0k}$ and $\hat{\beta}_{1k}$ are obtained from the least squares estimation. That is,

$$\hat{\beta}_{1k} = \frac{\sum_{i=1}^{I_k} (t_{ki} - \bar{t}_k)(y_{ki} - \bar{y}_k)}{\sum_{i=1}^{i_k} (t_{ki} - \bar{t}_k)^2}, \ \hat{\beta}_{0k} = \bar{y}_k - \hat{\beta}_{1k} \bar{t}_k$$

where $\bar{y}_k = \sum_{i=1}^{I_k} y_{ki}/I_k$.

The above test is not appropriate, however, when there is an overlap between the two regions or the two time periods. For this case, Li and Tiwari (2007) proposed the following corrected Z-test

$$Z_{CT} = \frac{\hat{\beta}_{11} - \hat{\beta}_{12}}{\left\{ \hat{\sigma}^2 \left(\sigma_1^{-2} + \sigma_2^{-2} - 2\sigma_{12}\sigma_1^{-2}\sigma_2^{-2} \frac{(n^{(O)})^2}{n_1 n_2} \right) \right\}^{1/2}}, \tag{3}$$

where $\sigma_k^2 = \sum_{i=1} (t_{ki} - \bar{t}_k)^2$, $\sigma_{12} = \sum_{s+1}^{m} (T_i - \bar{t}_1)(T_i - \bar{t}_2)$, $n_k = \sum_{i=s+1}^{m} \sum_{j=1}^{J} n_{kji}$ for $k = 1, 2$, $n^{(O)} = \sum_{i=s+1}^{m} \sum_{j=1}^{J} n_{ji}^{(O)}$ and and $n_{ji}^{(O)}$ are the numbers of at-risk population in the overlapping region. Note that there is no suffix k in $n_{ji}^{(O)}$. The sign of σ_{12} determines, whether the covariance between $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$ is positive or negative, and when there is no overlap in time intervals, $\sigma_{12} = 0$. Under the log-normal model, the corrected Z_{CT} test was shown to follow a standard normal distribution under the null hypothesis, and to be more efficient than the pooled t-test; see Li and Tiwari (2007).

However, one assumption in Li and Tiwari (2007) is the equal variance in both regression models, which may not be realistic, especially for rare cancers. A further refinement has been made to derive the variance of y_{ki} by using the Poisson assumptions on the first two moments of the counts d_{kji}, i.e. $E(d_{kji}) = var(d_{kji})$. Under these assumptions, the consistent estimate of the error variance of e_{ki} is given by $v_{ki}^2 = \frac{1}{r_{ki}^2} \sum_{j=1}^{J} w_j^2 \frac{d_{kji}}{n_{kji}^2}$, leading to the following weighted least squares test proposed by Li et al. (2007), referred to as Z_{WLS}:

$$Z_{WLS} = \frac{\tilde{\beta}_{11} - \tilde{\beta}_{12}}{\left\{ \tilde{\sigma}_1^{-2} + \tilde{\sigma}_2^{-2} - 2\tilde{\sigma}_{12}\tilde{\sigma}_1^{-2}\tilde{\sigma}_2^{-2} \frac{(n^{(O)})^2}{n_1 n_2} \right\}^{1/2}}. \tag{4}$$

with $\tilde{\sigma}_1^2 = \sum_{i=1}^{m} (T_i - \tilde{t}_1)^2 / v_{1i}^2$ and $\tilde{\sigma}_2^2 = \sum_{i=s+1}^{s+n} (T_i - \tilde{t}_2)^2 / v_{2i}^2$,

$$\tilde{\sigma}_{12} = \sum_{i=s+1}^{m} (T_i - \tilde{t}_1)(T_i - \tilde{t}_2) \frac{v_{12i}^{(o)}}{v_{1i}^2 v_{2i}^2},$$

where

$$v_{12i}^{(o)} = \frac{1}{r_{1i}r_{2i}} \sum_{j=1}^{J} w_j^2 \frac{d_{ji}^{(o)}}{(n_{ji}^{(o)})^2},$$

and $d_{ji}^{(o)}$ are the counts in the overlapping region and during the overlapping period. Here,

$$\tilde{t}_1 = \frac{\sum_{i=1}^{m} T_i/v_{1i}^2}{\sum_{i=1}^{m} 1/v_{1i}^2}, \qquad \tilde{t}_2 = \frac{\sum_{i=s+1}^{s+n} T_i/v_{2i}^2}{\sum_{i=s+1}^{s+n} 1/v_{2i}^2},$$

and $\tilde{\beta}_{11}, \tilde{\beta}_{12}$ are weighted least square estimates of β_{11}, β_{12}.

Under the null hypothesis, because of the normal approximation, Z_{WLS} approximately follows a standard normal distribution. The Z_{WLS} has been shown to be more conservative than Z_{CT} in retaining the size of the test, but is more powerful for the common cancer sites; see Li et al. (2007). However, there are several disadvantages of the existing methods. First, one key step of Li et al. (2007) is the normal approximation of the age-adjusted rates. Secondly, both Li et al. (2007) and Li and Tiwari (2007) need adjustments for zero counts.

3 Age-stratified Poisson Regression Model

As the existing approaches to dealing with age-adjusted cancer rates were all based on the normal approximation, we take a more natural route in the sequel by considering the Poisson nature of the underlying count data and propose an age-stratified Poisson regression to describe the change trend of incident (or death) counts on time. Based on this model, a proper test that accounts for overlapping is proposed.

Specifically, since d_{kji}, the number counts (deaths or new cancer cases) observed in Region k ($k = 1, 2$) in age-group j, is a count, we assume that

$$d_{kji} \overset{ind}{\sim} Pois(n_{kji}\lambda_{kji}),$$

with

$$\log \lambda_{kji} = \beta_{0kj} + \beta_{1k} t_{ki}, \tag{5}$$

which is referred to as the *Age-stratified Poisson Regression Model* as the age-specific intercept β_{0kj} is assumed for age-group j. The common slope β_{1k} is of particular importance as it transcribes the trends of mortality or incidence and, in particular, determines the APC value.

Again let APC_1 and APC_2 be the corresponding APC values for these two Poisson regressions. A natural test for the null hypothesis $H_0 : APC_1 = APC_2$ versus the alternative hypothesis $H_1 : APC_1 \neq APC_2$ would be

$$Z_{POIS} = \frac{\hat{\beta}_{11} - \hat{\beta}_{12}}{\sqrt{Var\{\hat{\beta}_{11} - \hat{\beta}_{12}\}}},$$

where $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$ are the maximum likelihood estimates of β_{11} and β_{12} derived in the Appendix.

Because of the possible overlapping of Regions 1 and 2, $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$ may be correlated. Thus the key to the derivation of the test lies in a correct evaluation of $Cov(\hat{\beta}_{11}, \hat{\beta}_{12})$.

3.1 Derivation of the Test

To proceed, we let $\beta_k = (\beta_{1k}, \beta_{0k1}, \ldots, \beta_{0kJ})'$, $k = 1, 2$, whose estimates $\hat{\beta}_k$ can be obtained by solving the score equations [based on (5)]

$$U_k(\beta_k) = 0,$$

for $k = 1, 2$, where $U_k = (U_{1k,1}, U_{0k,1}, \ldots, U_{0k,J})'$ and

$$\begin{aligned} U_{1k,1} &= \sum_{i=1}^{I_k}\sum_{j=1}^{J} n_{kji} t_{ki} \exp(\beta_{0kj} + \beta_{1k} t_{ki}) - \sum_{i=1}^{I_k}\sum_{j=1}^{J} d_{kji} t_{ki}, \\ U_{0k,j} &= \sum_{i=1}^{I_k} n_{kji} \exp(\beta_{0kj} + \beta_{1k} t_{ki}) - \sum_{i=1}^{I_k} d_{kji}, \end{aligned}$$

for $j = 1, \ldots, J$.

As $U_k(\hat{\beta}_k) = 0$, expanding it around the true value β_k, and ignoring the higher order terms yields

$$0 \equiv \frac{1}{\sqrt{I_k}} U_k(\hat{\beta}_k) = \frac{1}{\sqrt{I_k}} U_k(\beta_k) + \frac{1}{I_k} U_k^{(1)}(\beta_k) \left\{ \sqrt{I_k} \left(\hat{\beta}_k - \beta_k \right) \right\} + o_p(1),$$

where $U_k^{(1)} = \partial U_k(\beta_k)/\partial \beta_k$ are $(J+1)\times(J+1)$ matrices with its 1^{st} row as $(U_{1k,11}^{(1)}, U_{1k,01}^{(1)}, \ldots, U_{1k,0J}^{(1)})$ and the $(j+1)^{th}$ row $(j = 1, \ldots, J)$ as $(U_{0k,j1}^{(1)}, U_{0k,0j1}^{(1)}, \ldots, U_{0k,0jJ}^{(1)})$, for $k = 1, 2$. Here

$$\begin{aligned} U_{1k,11}^{(1)} &= \frac{\partial U_{1k,1}}{\partial \beta_{1k}} = \sum_{i=1}^{I_k}\sum_{j=1}^{J} n_{kji} t_{ki}^2 \exp(\beta_{0kj} + \beta_{1k} t_{ki}), \\ U_{1k,0j}^{(1)} &= \frac{\partial U_{1k,1}}{\partial \beta_{0kj}} = \sum_{i=1}^{I_k} n_{kji} t_{ki} \exp(\beta_{0kj} + \beta_{1k} t_{ki}), \\ U_{0k,j1}^{(1)} &= \frac{\partial U_{0k,j}}{\partial \beta_{1k}} = \sum_{i=1}^{I_k} n_{kji} t_{ki} \exp(\beta_{0kj} + \beta_{1k} t_{ki}), \\ U_{0k,0jj'}^{(1)} &= \frac{\partial U_{0k,j}}{\partial \beta_{0kj'}} = \delta_{jj'} \sum_{i=1}^{I_k} n_{kji} \exp(\beta_{0kj} + \beta_{1k} t_{ki}), \end{aligned}$$

and $\delta_{jj'} = 1$ if $j = j'$ and 0 otherwise.

Denote by $A_k = plim_{I_k \to \infty} - U_k^{(1)}(\beta_k)/I_k$ for $k = 1, 2$, where $plim$ denotes the limit in probability. Then for large $I_1 (\equiv m)$ and $I_2 (\equiv n)$, standard probabilistic arguments yield

$$\left\{ \sqrt{I_1}(\hat{\beta}_1 - \beta_1), \sqrt{I_2}(\hat{\beta}_2 - \beta_2) \right\} \overset{d}{\sim} \left\{ A_1^{-1} \frac{1}{\sqrt{I_1}} U_1(\beta_1), A_2^{-1} \frac{1}{\sqrt{I_2}} U_2(\beta_2) \right\}.$$

Here, $\overset{d}{\sim}$ denotes approximate equivalence in joint distribution functions. Hence,

$$\begin{aligned} Var(\sqrt{m}(\hat{\beta}_1 - \beta_1), \sqrt{m}(\hat{\beta}_1 - \beta_1)) &\doteq A_1^{-1} \frac{1}{m} Cov(U_1(\beta_1), U_1(\beta_1)) A_1^{-T}, \\ Var(\sqrt{n}(\hat{\beta}_2 - \beta_2), \sqrt{n}(\hat{\beta}_2 - \beta_2)') &\doteq A_2^{-1} \frac{1}{n} Cov(U_2(\beta_2), U_2(\beta_2)) A_2^{-T}, \\ Cov(\sqrt{m}(\hat{\beta}_1 - \beta_1), \sqrt{n}(\hat{\beta}_2 - \beta_2)') &\doteq A_1^{-1} \frac{1}{\sqrt{mn}} Cov(U_1(\beta_1), U_2(\beta_2)) A_2^{-T}. \end{aligned}$$

Let $V = \frac{1}{m}Cov(U_1(\beta_1), U_1(\beta_1))$, $W = \frac{1}{n}Cov(U_2(\beta_2), U_2(\beta_2))$, $\Sigma = \frac{1}{\sqrt{mn}}Cov(U_1(\beta_1), U_2(\beta_2))$, and $\hat{V}, \hat{W}, \hat{\Sigma}$ be the corresponding estimates, whose derivations are given in the Appendix.

Then,

$$\begin{aligned}\hat{C}ov(\hat{\beta}_1, \hat{\beta}_1) &= m\{U_1^{(1)}(\hat{\beta}_1)\}^{-1}\hat{V}\{U_1^{(1)}(\hat{\beta}_1)\}^{-T};\\ \hat{C}ov(\hat{\beta}_2, \hat{\beta}_2) &= n\{U_2^{(1)}(\hat{\beta}_2)\}^{-1}\hat{W}\{U_2^{(1)}(\hat{\beta}_2)\}^{-T};\\ \hat{C}ov(\hat{\beta}_1, \hat{\beta}_2) &= \sqrt{mn}\{U_1^{(1)}(\hat{\beta}_1)\}^{-1}\hat{\Sigma}\{U_2^{(1)}(\hat{\beta}_2)\}^{-T}.\end{aligned}$$

These are three $(J+1)\times(J+1)$ matrices, the $(1,1)$ entries of which are $\hat{\sigma}_1^2 = \hat{V}ar(\hat{\beta}_{11})$, $\hat{\sigma}_2^2 = \hat{V}ar(\hat{\beta}_{12})$ and $\hat{\sigma}_{12} = \hat{C}ov(\hat{\beta}_{11}, \hat{\beta}_{12})$, respectively. From this we compute $\hat{V}ar(\hat{\beta}_{11}-\hat{\beta}_{12}) = \hat{V}ar(\hat{\beta}_{11})+\hat{V}ar(\hat{\beta}_{12})-2\hat{C}ov(\hat{\beta}_{11}, \hat{\beta}_{12})$. Hence, the Z-test for comparing APC values is given by

$$Z_{POIS} = \frac{\hat{\beta}_{11} - \hat{\beta}_{12}}{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\sigma}_{12})^{1/2}}, \tag{6}$$

which follows the standard normal distribution under $H_0 : \beta_{11} = \beta_{12}$. The computation of $\hat{\beta}_{11}, \hat{\beta}_{12}$ is given in the Appendix.

3.2 ARE Comparison with the WLS test

It is of substantial interest to evaluate the gains in efficiency of the proposed test compared with the WLS test. First note (5) implies that $E(r_{ki}) \equiv E(\sum_j w_j \frac{d_{kji}}{n_{kji}}) = (\sum_j w_j e^{\beta_{0kj}})e^{\beta_{1k}T_i}$. This in turn implies that

$$E(\log r_{ki}) \doteq \log E(r_{ki}) = \log(\sum_j w_j e^{\beta_{0kj}}) + \beta_{1k}t_{ki}, \tag{7}$$

when $Var(r_{ki})$ is small, which is often the case for the cancer incidence and mortality data (Kim et al., 2000).

A comparison between (1) and (7) reveals that models (1) and (5) approximately specify the same first moment of the age-adjusted cancer rates, making it possible to compare the efficiency of the tests based on these two models via the measure of the Pitman Asymptotic Relative Efficiency. Specifically, standard asymptotic analysis will yield

$$\hat{\beta}_{11} - \hat{\beta}_{12} \sim N(\beta_{11} - \beta_{12}, \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\sigma}_{12}),$$

while, for the WLS estimates,

$$\tilde{\beta}_{11} - \tilde{\beta}_{12} \sim N(\beta_{11} - \beta_{12}, \tilde{\sigma}_1^{-2} + \tilde{\sigma}_2^{-2} - 2\tilde{\sigma}_{12}\tilde{\sigma}_1^{-2}\tilde{\sigma}_2^{-2}\frac{(n^{(O)})^2}{n_1 n_2}).$$

Hence, the Pitman Asymptotic Relative Efficiency (ARE) comparing tests (6) and (4), which is the ratio of the noncentralities of the above two normal distributions, is given by

$$\text{ARE} = \frac{\tilde{\sigma}_1^{-2} + \tilde{\sigma}_2^{-2} - 2\tilde{\sigma}_{12}\tilde{\sigma}_1^{-2}\tilde{\sigma}_2^{-2}\frac{(n^{(O)})^2}{n_1 n_2}}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\sigma}_{12}}. \tag{8}$$

The evaluation of (8) typically involves numerical computations.

4 SEER mortality data analysis and Simulations

Li et al. (2007) demonstrated that Z_{WLS} performs better than Z_{CT}, via the calculation of ARE, as Z_{CT} relies on the common variance assumption, which may not be realistic. Hence, we focus this paper on comparing Z_{POIS} with Z_{WLS}. To comprehensively evaluate these tests, we consider several scenarios of overlap in two regions and in two different time intervals. Specifically, we assume that Region 1 consists of Georgia (GA), South Carolina (SC), and North Carolina (NC), and that Region 2 consists of NC, Virginia (VA) and Maryland (MD); with NC as the overlapping state between the two regions. The three different time intervals, with varying degree of overlap in the intervals, are taken to be : (a) [1980,1989] for Region 1, and [1990,1999] for Region 2 so that there is no overlap between the two time intervals and, hence, $\sigma_{12} = 0$; (b) [1980,1989] for Region 1, and [1983,1992] for Region 2 so that there a considerable overlap of six years between the two intervals and $\sigma_{12} = 12.25$; (c) [1980,1989] for Region 1, and [1987, 1996] for Region 2 so that there is a little overlap of three years' between the two intervals and $\sigma_{12} = -34.75$.

The counts, d_{kji} were generated based on model (5) with t_{ki} taking values in the intervals corresponding to the two regions stated above. More specifically, the t_{1i} take values of $\{0, 1, \ldots, 9\}$, while the t_{2i} take values of $\{10, \ldots, 19\}$, $\{3, \ldots, 12\}$, and $\{7, \ldots, 16\}$, respectively for cases (a)-(c).

In order to fully specify λ_{kji} in (5), we assume that $\beta_{0kj} = \log(d_{kj1}/n_{kj1}) - \beta_{1k}t_{k1}$, where d_{kj1} and n_{kj1} are respectively the observed number of deaths and the number of at-risk population at the beginning of the intervals considered, and take $\beta_{1k} = \log(APC_k/100 + 1)$, based on the specified values of APC_k. The age-specific counts for the overlapping state, NC, are generated from Poisson distributions with means, $n_{ji}^{(o)} \times \frac{1}{2}(\lambda_{1ji} + \lambda_{2ji})$, where $n_{ji}^{(o)}$ denotes the at-risk population in the overlapping region. When $\lambda_{1ji} = \lambda_{2ji}$, this reduces to the situation specified by the null hypothesis. 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. The values of APC were assumed to range from -0.3% to 3.0%. For each parameter configuration, a total of 1000 simulated data were obtained. The results for the three time-overlapping cases are summarized in Tables 1-3.

We remark that that, even though both Z_{WLS} and Z_{POIS} are derived under different model assumptions, they are both valid tests for testing the equality of two APCs and hence the ARE defined in (8) is valid. The tables show that the ARE of Z_{POIS} with respect to Z_{WLS} is greater than 1 for all the three cases, meaning that Z_{POIS} would be more powerful than Z_{WLS} when the alternative hypothesis is true. The tables also show that in most situations Z_{POIS} outperforms the Z_{WLS} in retaining the Type I error probabilities and, hence, yields a more valid test. Also the powers of both WLS and Poisson-based tests are sensitive to the delta values (the differences of APC values). The larger the delta values are, the more powerful the tests are. The larger delta values also lead to slightly larger AREs, though the differences are not so obvious.

It is of substantial interest to compare the changes in cancer mortality rates in California with the national levels starting late 1980's as a California law (Health and Safety Code, Section 103885) was passed then, which 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 weighted linear models as well as the age-stratified Poisson model, and applied both tests to compare the age-adjusted mortality rates of female breast cancer in CA for the 16-year period from 1989-2004 to that of US for the 16-year period from 1987-2002, for which the national mortality data were available. The observed values of the log-transformed annual age-adjusted rates and fitted regression lines from the Z_{POIS} test procedure are plotted in Figure 1. The parameter estimates and the values of the test statistics are summarized in Table 4. The results indicated the mortality rates of Breast cancer for California and the US have decreased. Both tests reject the null hypothesis of equality of the two APCs, indicating that the annual percent change (APC) of California, is significantly different from the national level. However, the p-value for Z_{POIS} is much smaller than that of Z_{WLS}, rendering more evidence again the null hypothesis.

5 Discussion

In this paper, we have considered an important problem where comparisons have to be made for regions or time intervals that overlap. As opposed to the existing work, e.g Li and Tiwari (2007) and Li et al. (2007), this project advances this area in two distinct ways. First, the developed test does not rely on the normal approximation of the cancer rate, but directly model the counts to follow a Poisson regression model. The parameters are then estimated using their maximum likelihood estimates, and the Z-test is derived for testing the equality of two APCs. Secondly, the developed Poisson regression model can easily accommodate 0 count data (for the rare cancers), as opposed to the log normal model (Li and Tiwari, 2007; Li et al., 2007) which needs to involve extra zero-corrected adjustment. 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 existing tests. A Bayesian Poisson regression would be a useful approach. However, choice of priors is always difficult and computation may not be so straightforward compared to this current work, wherein analytical solutions have been derived. Hence, we envision that the proposed method would be preferable because of simple interpretation of the model parameters, natural choice of the model and computational readiness.

In our technical development, we have modelled the logarithm transformation of the age-adjusted rates as a linear regression on time in (5) and have indeed explicitly assumed parallelism across age groups. That is, the growth curves of the cancer rates for various age groups share the same slope, which carries the information for the APC. Indeed, linearity parallelism for the cancer rates could be a debatable issue in cancer surveillance, which is likely to be violated for some cancers. One alternative, along the line of generalized mixed models, is to assume a random slope (as opposed to a constant slope) across age groups. This ongoing work will be reported in a subsequent communication.

Acknowledgements The authors thank the editor, an AE and two referees for their insightful suggestion, which led to a better version of this manuscript.

References

American Cancer Society (2007) *Cancer Facts & Figures.* Atlanta, Georgia.

Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics (with discussion). *Biometrics* **42,** 693-734.

Fay, M., Tiwari, R., Feuer, E. and Zou, Z. (2006). Estimating Average Annual Percent Change for Disease Rates without Assuming Constant Change. *Biometrics* **62,** 847–854.

Kim, H., Fay, M., Feuer, E., Midthune, D. (2000) Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* **19,** 335-351.

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

Li, Y. and Tiwari, R. (2007). Comparing Trends in Age-Adjusted Cancer Rates Across Overlapping Regions or Time Intervals for the NCI SEER Program. *Technical Report*, http://www.bepress.com/harvardbiostat/paper71.

Li, Y., Tiwari, R., Walters, K. and Zou, Z. (2007) A Weighted-Least-Squares Estimation Approach to Comparing Trends in Age-Adjusted Cancer Rates Across Overlapping Regions. *Technical Report*, http://biowww.dfci.harvard.edu/∼yili/apc2.pdf

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

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

Appendix A: Derivation of $\hat{V}, \hat{W}, \hat{\Sigma}$

Write

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{pmatrix}_{(J+1)\times(J+1)}; W = \begin{pmatrix} W_{11} & W_{12} \\ W_{12}' & W_{22} \end{pmatrix}_{(J+1)\times(J+1)}; \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}_{(J+1)\times(J+1)}$$

where

$$V_{11} = \frac{1}{m} Var(U_{11,1}) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{J} T_i^2 Var(d_{1ji}) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{J} T_i^2 E(d_{1ji}).$$

Hence a consistent estimate is $\hat{V}_{11} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{J} T_i^2 d_{1ji}$.

Similarly,

$$\hat{V}_{12} = (\hat{V}_{12,1}, \ldots, \hat{V}_{12,J})$$

where

$$\hat{V}_{12,j} = \hat{Cov}(U_{11,1}, U_{01,j}) = \frac{1}{m} \sum_{i=1}^{m} T_i d_{1ji}.$$

Also, $V_{22} = ((V_{22,jj'}))$ with $\hat{V}_{22,jj'} = \hat{Cov}(U_{01,j}, U_{01,j'}) = 0$, $j \neq j'$; $\frac{1}{m} \sum_{i=1}^{m} d_{1ji}$, $j = j'$.

Next compute the estimate of W:

$$W_{11} = \frac{1}{n} Var(U_{12,1}) = \frac{1}{n} \sum_{i=s+1}^{s+n} \sum_{j=1}^{J} T_i^2 Var(d_{2ji}) = \frac{1}{n} \sum_{i=s+1}^{s+n} \sum_{j=1}^{J} T_i^2 E(d_{2ji})$$

so that $\hat{W}_{11} = \frac{1}{n} \sum_{i=s+1}^{s+n} \sum_{j=1}^{J} T_i^2 d_{2ji}$ Similarly, $\hat{W}_{12} = (\hat{W}_{12,1}, \ldots, \hat{W}_{12,J})$ where $\hat{W}_{12,j} = \hat{Cov}(U_{12,1}, U_{02,j}) = \frac{1}{n} \sum_{i=s+1}^{n} T_i d_{2ji}$. Also, $\hat{W}_{22} = ((\hat{W}_{22,jj'}))$ with $\hat{W}_{22,jj'} = 0$, $j \neq j'$; $= \frac{1}{n} \sum_{i=s+1}^{s+n} d_{2ji}$, $j = j'$ Finally, the estimate of Σ is computed as follows:

$$\begin{aligned} \Sigma_{11} &= \frac{1}{\sqrt{mn}} Cov(U_{11,1}, U_{12,1}) \\ &= \frac{1}{\sqrt{mn}} Cov\left(\sum_{i=1}^{m} \sum_{j=1}^{J} T_i d_{1ji}, \sum_{i=s+1}^{s+n} \sum_{j=1}^{J} T_i d_{2ij} \right) \\ &= \frac{1}{\sqrt{mn}} Cov\left(\sum_{i=1}^{m} \sum_{j=1}^{J} T_i \left(d_{1ji}^{(NO)} + d_{ji}^{(O)} \right), \sum_{i=s+1}^{s+n} \sum_{j=1}^{J} T_i \left(d_{2ji}^{(NO)} + d_{ji}^{(O)} \right) \right) \end{aligned}$$

where $d_{kji} = d_{jki}^{(NO)} + d_{ji}^{(O)}$ and the superscripts "NO" and "O" denote the non-overlapping and overlapping regions, respectively. Thus, $\hat{\Sigma}_{11} = \frac{1}{\sqrt{mn}} \sum_{i=s+1}^{m} \sum_{j=1}^{J} T_i^2 d_{ji}^{(O)}$. Let $\hat{\Sigma}_{12} = \left(\hat{\Sigma}_{12,1}, \ldots, \hat{\Sigma}_{12,J} \right)$ where $\hat{\Sigma}_{12,j} = \frac{1}{\sqrt{mn}} \sum_{i=s+1}^{m} T_i d_{ji}^{(O)}$. Also, $\hat{\Sigma}_{22} = ((\hat{\Sigma}_{22,jj'}))$ with $\hat{\Sigma}_{22,j,j'} = 0$, $j \neq j'$; $\frac{1}{\sqrt{mn}} \sum_{i=s+1}^{m} d_{ji}^{(O)}$, $j = j'$.

Appendix B: Computation of the MLEs

Note that the MLEs of β_{0jk} and β_{1k} satisfy:

$$\begin{aligned}
e^{\beta_{0jk}} &= \frac{\sum_i d_{kji}}{\sum_i e^{\beta_{1k}T_i} n_{kji}};\\
\tilde{U}(\beta_{1k}) &= \sum_{j,i} n_{kji} T_i \left(\frac{\sum_i d_{jki}}{\sum_i e^{\beta_{1k}T_i} n_{kji}} \right) e^{\beta_{1k}T_i} - \sum_{j,i} d_{kji} T_i \\
&= \sum_j d_{kj.} \left(\frac{\sum_i n_{kji} T_i e^{\beta_{1k}T_i}}{\sum_i e^{\beta_{1k}T_i} n_{kji}} \right) - \sum_i d_{k.i} T_i \\
&= \sum_j d_{kj.} [A_{kj}(1) A_{kj}^{-1}(0)] - \sum_i d_{k.i} T_i = 0;\ k = 1, 2,
\end{aligned}$$

where $d_{kj.} = \sum_i d_{kji}$, $d_{k.i} = \sum_j d_{kji}$, and $A_{kj}(a) = \sum_i n_{kji} T_i^a e^{\beta_{1k}T_i}$ for $a = 0, 1, 2$.

Since $\tilde{U}(\beta_{1k})$ is a monotonic function of β_{1k}, there is a unique solution to this equation. Let $\hat{\beta}_{1k}$ be the solution. This is the MLE of β_{1k}. We can use a Newton-Raphson method to obtain $\hat{\beta}_{1k}$ as follows. Let $\hat{\beta}_{1k}(l)$ be the estimate at the l^{th} iteration, then the estimate at the $(l+1)^{th}$ iteration is given by $\hat{\beta}_{1k}(l+1) = \hat{\beta}_{1k}(l) - [\tilde{U}^{(1)}(\hat{\beta}_{1k}(l))]^{-1} \tilde{U}(\hat{\beta}_{1k}(l))$, where $\tilde{U}^{(1)}(\beta)$ is the first (partial) derivative of $\tilde{U}(b)$ with respect to b evaluated at $b = \beta$. We stop iterating wh en $|\hat{\beta}_{1k}(l+1) - \hat{\beta}_{1k}(l)| < \varepsilon$ for some pre-specified value of ε. Note that

$$\begin{aligned}
\tilde{U}^{(1)}(\beta_{1k}) &= \sum_{j,i} n_{jki} T_i^2 e^{\beta_{1k}T_i} \left(\frac{\sum_i d_{kji}}{\sum_i e^{\beta_{1k}T_i} n_{kji}} \right) - \sum_{j,i} n_{jki} t_i e^{\beta_{1k}t_i} \frac{\sum_i n_{kji} t_i e^{\beta_{1k}t_i}}{\left(\sum_i e^{\beta_{1k}t_i} n_{kji} \right)^2} \\
&= \sum_j \left(\frac{\left(\sum_i n_{kji} t_i^2 e^{\beta_{1k}t_i} \right) \left(\sum_i d_{kji} \right)}{\sum_i e^{\beta_{1k}t_i} n_{kji}} \right) - \sum_j \frac{\left(\sum_i n_{kji} t_i e^{\beta_{1k}t_i} \right)^2}{\left(\sum_i e^{\beta_{1k}t_i} n_{kji} \right)^2} \\
&= \sum_j d_{kj.} \left(\frac{\sum_i n_{kji} t_i^2 e^{\beta_{1k}t_i}}{\sum_i e^{\beta_{1k}t_i} n_{kji}} \right) - \sum_j \frac{\left(\sum_i n_{kji} t_i e^{\beta_{1k}t_i} \right)^2}{\left(\sum_i e^{\beta_{1k}t_i} n_{kji} \right)^2} \\
&= \sum_j d_{kj.} [A_{kj}(2)(A_{kj}(0))^{-1}] - \sum_j [(A_{kj}(1))^2 (A_{kj}(0))^{-1}],\ k = 1, 2.
\end{aligned}$$

Substituting the MLE $\hat{\beta}_{1k}$ in place for β_{1k} in $e^{\beta_{0kj}} = \frac{\sum_i d_{kji}}{\sum_i e^{\beta_{1k}t_i} n_{kji}}$ gives the MLE $\hat{\beta}_{0kj} = \log\left(\sum_i d_{kji} \right) - \log\left(\sum_i e^{\hat{\beta}_{1k}t_i} n_{kji} \right)$.

Table 1 Comparison of the power functions under various hypotheses between the Poisson-based test Z_{POIS} and the weighted-least-squares based test Z_{WLS} for two overlapping regions over disjoint time intervals, Region 1 (1980-1989) vs Region 2 (1990-1999). APC1 and APC2 are the annual percent changes in Regions 1 and 2, respectively.

Cancer Sites	APC_1	APC_2	ARE	Z_{WLS}	Z_{POIS}
All Malignant	0.100	0.100	1.1490	0.059	0.050
	-0.300	-0.300	1.1491	0.049	0.045
	0.500	0.500	1.1492	0.060	0.049
	1.000	1.000	1.1493	0.048	0.053
	3.000	3.000	1.1504	0.047	0.057
	0.100	0.500	1.1507	0.877	0.907
	-0.300	0.300	1.1509	0.998	1.000
	1.000	2.000	1.1515	1.000	1.000
	1.000	3.000	1.1518	1.000	1.000
Prostate	0.100	0.100	1.1610	0.043	0.045
	-0.300	-0.300	1.1611	0.041	0.047
	0.500	0.500	1.1612	0.051	0.041
	1.000	1.000	1.1613	0.046	0.051
	3.000	3.000	1.1614	0.045	0.051
	0.100	0.500	1.1617	0.182	0.212
	-0.300	0.300	1.1619	0.357	0.40
	1.000	2.000	1.1622	0.773	0.830
	1.000	3.000	1.1634	1.000	1.000

Table 2 Comparison of the power functions between the Poisson based Z_{POIS} and the weighted-least-squares based Z_{WLS} for two overlapping regions over roughly the same time intervals, Region 1 (1980-1989) vs Region 2 (1983-1992). APC1 and APC2 are the annual percent changes in Regions 1 and 2, respectively.

Cancer Sites	APC_1	APC_2	ARE	Z_{WLS}	Z_{POIS}
All Malignant	0.100	0.100	1.1710	0.055	0.057
	-0.300	-0.300	1.1710	0.056	0.057
	0.500	0.500	1.1711	0.048	0.050
	1.000	1.000	1.1712	0.047	0.049
	3.000	3.000	1.1723	0.051	0.056
	0.100	0.500	1.1726	0.872	0.908
	-0.300	0.300	1.1729	0.994	0.997
	1.000	2.000	1.1733	1.000	1.000
	1.000	3.000	1.1737	1.000	1.000
Prostate	0.100	0.100	1.1820	0.043	0.047
	-0.300	-0.300	1.1821	0.043	0.043
	0.500	0.500	1.1822	0.046	0.041
	1.000	1.000	1.1823	0.035	0.053
	3.000	3.000	1.1823	0.034	0.055
	0.100	0.500	1.1825	0.170	0.197
	-0.300	0.300	1.1827	0.341	0.393
	1.000	2.000	1.1832	0.742	0.802
	1.000	3.000	1.1835	1.000	1.000

Table 3 Comparison of the power functions between the Poisson based Z_{POIS} and the weighted-least-squares based Z_{WLS} for two overlapping regions over slightly overlapping time intervals, Region 1 (1980-1989) vs Region 2 (1987-1996). APC1 and APC2 are the annual percent changes in Regions 1 and 2, respectively.

Cancer Sites	APC_1	APC_2	ARE	Z_{WLS}	Z_{POIS}
All Malignant	0.100	0.100	1.1350	0.044	0.044
	-0.300	-0.300	1.1350	0.044	0.046
	0.500	0.500	1.1351	0.046	0.051
	1.000	1.000	1.1351	0.045	0.051
	3.000	3.000	1.1352	0.043	0.047
	0.100	0.500	1.1354	0.834	0.891
	-0.300	0.300	1.1355	0.994	0.994
	1.000	2.000	1.1362	1.000	1.000
	1.000	3.000	1.1367	1.000	1.000
Prostate	0.100	0.100	1.1410	0.045	0.051
	-0.300	-0.300	1.1410	0.043	0.049
	0.500	0.500	1.1412	0.060	0.051
	1.000	1.000	1.1423	0.061	0.055
	3.000	3.000	1.1426	0.052	0.052
	0.100	0.500	1.1428	0.170	0.184
	-0.300	0.300	1.1431	0.319	0.361
	1.000	2.000	1.1436	0.706	0.759
	1.000	3.000	1.1439	0.998	1.000

Table 4 Comparing APC of Breast Cancer Mortality Between CA and the US

	CA	US
APC_{WLS}	-2.33	-1.94
SE_{WLS}	0.084	0.027
Z_{WLS}	-4.95	(p=0.000000757)
APC_{POIS}	-2.29	-1.84
SE_{POIS}	0.083	0.026
Z_{POIS}	-5.62	(p= 0.000000019593)

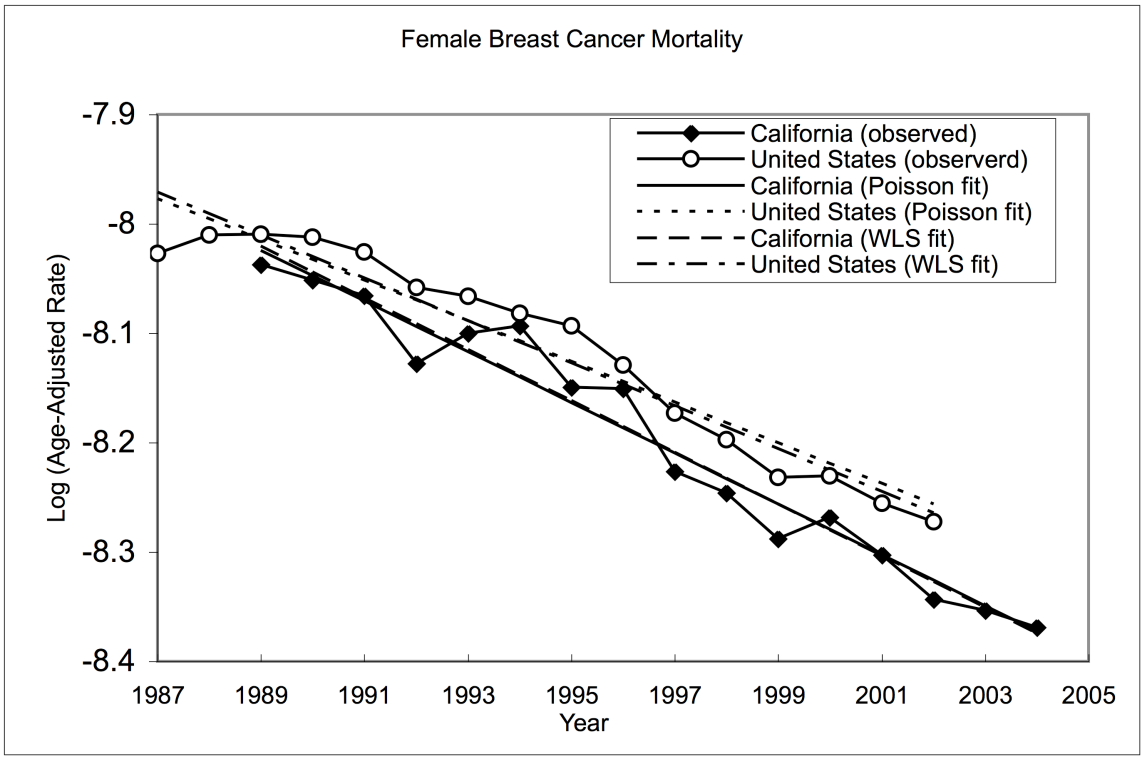

Fig. 1 Observed and fitted log-transformed age-adjusted breast cancer mortality rates in CA [1989-2004] and US [1987-2002]