

# Web-based Supplementary Materials for “Comparing trends in cancer rates across overlapping regions” by Y. Li and R. Tiwari

## Derivation of Equation (8)

To proceed, we assume that  $t_1 \leq t_{s+1} < t_m \leq t_{s+I}$  and note that

$$\begin{aligned} Cov(\hat{\beta}_{11}, \hat{\beta}_{21}) &= \frac{1}{\sigma_1^2 \sigma_2^2} Cov \left\{ \sum_{i=1}^m (t_i - \bar{t}_1) y_{1i}, \sum_{s+1}^{s+I} (t_i - \bar{t}_2) y_{2i} \right\} \\ &= \frac{1}{\sigma_1^2 \sigma_2^2} \sum_{s+1}^m (t_i - \bar{t}_1)(t_i - \bar{t}_2) Cov(y_{1i}, y_{2i}). \end{aligned} \quad (14)$$

Recall that we use superscript ‘O’ to denote the intersection of Regions 1 and 2 and ‘NO’ the non-overlapping subset. We further introduce the following notation. Let  $n_{kji}$ ,  $n_{kji}^{(O)}$  and  $n_{kji}^{(NO)}$  be the numbers of underlying population at risk for age group  $j$  at time  $t_i$  in Region  $k$  ( $k = 1, 2$ ), in the overlapping subregion and in the non-overlapping subregions, respectively. Similarly, define  $d_{kji}$ ,  $d_{kji}^{(O)}$  and  $d_{kji}^{(NO)}$  the corresponding numbers of events (e.g. deaths or cancer cases). Denote by  $n_{ki} = \sum_{j=1}^J n_{kji}$ ,  $n_{ki}^{(O)} = \sum_{j=1}^J n_{kji}^{(O)}$ ,  $n_{ki}^{(NO)} = \sum_{j=1}^J n_{kji}^{(NO)}$ . Also define  $d_{ki}$ ,  $d_{ki}^{(O)}$  and  $d_{ki}^{(NO)}$  in the similar fashion. In fact,  $d_{kji}^{(O)}$  and  $n_{kj}^{(O)}$  are independent of index  $k$  (for region) as they correspond to the same common subregion for  $k = 1, 2$ .

Let  $y_i^{(O)} = \log(r_i^{(O)}) = \log \left( \sum_{j=1}^J w_j \frac{d_{ji}^{(O)} + 1/J}{n_{ji}^{(O)}} \right)$  be the logarithm of the (zero corrected) age-adjusted rate  $r_i^{(O)}$  at time  $t_i$  for the overlapping region, and let  $y_{1i}^{(NO)}$  and  $y_{2i}^{(NO)}$  be defined similarly based on  $r_{1i}^{(NO)}$  and  $r_{2i}^{(NO)}$ , respectively, for the non-overlapping regions/intervals for the two groups.

Dropping the subscript  $i$  (for time), we assume the age groups have the same distribution across the overlapping and non-overlapping regions, that is,

$$\frac{n_{k1}^{(O)}}{n_{k1}} = \frac{n_{k2}^{(O)}}{n_{k2}} = \dots = \frac{n_{kJ}^{(O)}}{n_{kJ}} = p_k^{(O)}, \text{ and } \frac{n_{k1}^{(NO)}}{n_{k1}} = \frac{n_{k2}^{(NO)}}{n_{k2}} = \dots = \frac{n_{kJ}^{(NO)}}{n_{kJ}} = p_k^{(NO)}, \quad (15)$$

for  $k=1,2$ . This assumption is common in comparing the age-adjusted rates across different geographical areas (see, e.g., Pickle and White, 1995), under which, we have

$$r_k = \sum_{j=1}^J w_j \frac{d_{kj}}{n_{kj}} = \sum_{j=1}^J w_j \frac{d_{kj}^{(O)} + d_{kj}^{(NO)}}{n_{kj}}$$

$$\begin{aligned}
&= \sum_{j=1}^J w_j \frac{n_{kj}^{(O)} d_{kj}^{(O)} + 1/J}{n_{kj}^{(O)}} + \sum_{j=1}^J w_j \frac{n_{kj}^{(NO)} d_{kj}^{(NO)} + 1/J}{n_{kj}^{(NO)}} + c_k \\
&= p_k^{(O)} r_k^{(O)} + p_k^{(NO)} r_k^{(NO)} + c_k,
\end{aligned}$$

where  $c_k = -\frac{1}{J} \sum_{j=1}^J \frac{w_j}{n_{kj}}$ , a negligible constant. Again, since  $r_1^{(O)} = r_2^{(O)}$ , let  $r^{(O)}$  denote this common value, and let  $y^{(O)} = \log(r^{(O)})$ . Now, since  $\text{Cov}(r_1^{(NO)}, r_2^{(NO)}) = 0$  and  $\text{Cov}(r^{(O)}, r_k^{(NO)}) = 0$ ,  $k = 1, 2$ , using the delta method, we have,

$$\begin{aligned}
\text{Cov}(y_1, y_2) &= \text{Cov}(\log(r_1), \log(r_2)) \\
&\approx \frac{1}{E(r_1)E(r_2)} \text{Cov}(r_1, r_2) \\
&= \frac{1}{E(r_1)E(r_2)} p_1^{(O)} p_2^{(O)} \text{Var}(r^{(O)}) \\
&= \frac{1}{E(r_1)E(r_2)} p_1^{(O)} p_2^{(O)} \text{Var}(e^{y^{(O)}}).
\end{aligned}$$

Let  $y^{(O)}$  satisfy the regression model (3), and let  $\mu^{(O)} = E(y^{(O)})$ . Since  $y^{(O)} \sim N(\mu^{(O)}, \sigma^2)$ , using the properties of that log normal distribution, we have that

$$\begin{aligned}
E(r^{(O)}) &= E(e^{y^{(O)}}) = e^{\mu^{(O)}} e^{\sigma^2/2}, \\
\text{Var}(e^{y^{(O)}}) &= e^{2\mu^{(O)}} e^{\sigma^2} (e^{\sigma^2} - 1).
\end{aligned}$$

Furthermore, the null hypothesis implies that  $E(y_1) = E(y_2) = E(y^{(O)})$ . Hence, adding back the time index  $i$ , we will have

$$\begin{aligned}
\text{Cov}(y_{1i}, y_{2i}) &= (e^{\sigma^2} - 1) p_{1i}^{(O)} p_{2i}^{(O)} \\
&\approx \sigma^2 p_{1i}^{(O)} p_{2i}^{(O)},
\end{aligned}$$

when  $\sigma^2$  is small. For the US population,  $p_{1i}^{(O)}$  and  $p_{2i}^{(O)}$  were found to be constant over years (as confirmed by the SEER population data base). We then write  $p_{ki}^{(O)} \equiv p_k^{(O)}$  for  $i = s+1, \dots, m$ , an estimate of which is given by  $\hat{p}_k^{(O)} = \frac{n^{(O)}}{n_k}$ , where  $n_k = \sum_{i=s+1}^m \sum_{j=1}^J n_{kji}$  and  $n^{(O)} = \sum_{i=s+1}^m \sum_{j=1}^J n_{ji}^{(O)}$ . Hence,  $\text{Cov}(y_{1i}, y_{2i}) \approx \sigma^2 \frac{(n^{(O)})^2}{n_1 n_2}$  for  $i = s+1, \dots, m$ . Inserting it back to (14) yields (8).

## **Tables for Data Analysis**







