Statistical Research and Applications Branch, NCI, Technical Report # 2009-01

Effect of Reporting Year on Delay Modeling

Zou J, Huang L, Midthune D, Horner MJ, Krapcho M, Feuer EJ.

Reporting year

Cancer cases are reported to the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute within 22 months of initial diagnosis. This means that cases diagnosed in 2006 are reported in the November data submission of 2008. Case reports for a given diagnosis year are updated with information received in subsequent data submissions. Updates include adding new, previously unreported cases as well as deleting or correcting information on existing cases which may include corrections to race, cancer site, sex, and age of diagnosis. Thus, by comparing cases from one data submission to the next, adds and drops of cases can be computed within specific age-race-sex stratifications. The corresponding cancer incidence rate for a particular year of diagnosis is, therefore, continually updated as additional case information is received (referred to as "adds") or deleted (referred to as "drops") in each stratification in each data submission. Usually there are more adds than drops, which leads to an increase in the number of reported cancer cases over reporting years. The time lag that occurs in the reporting of newly diagnosed cases or the reporting of case updates is referred to as reporting delay. The model used to adjust for reporting delay is referred to as the delay adjustment model.

The purpose of this technical report is to explain how reporting year affects the delay adjustment model and how we identify specific instances of the problem and incorporate it into the delay model.

Effect of reporting year on delay modeling

For specific reporting years, a sharp increase or decrease in the adds or drops of case counts may occur across all diagnosis years. These secular trends are known as reporting year effects. Past instances of reporting year effects are the sharp changes in case counts that have occurred due to registry operations such as systematic review of unknown race category and registry-wide changes in case identification numbers. For example, in 2005 a correction designed to provide consistency in statewide case identification numbers resulted in an erroneous dip in the delay adjustment factors for male lung cancer. Because the original delay adjustment model (Midthune *et al*, 2005) was not structured to take into account sudden changes in counts in a particular reporting year, these sudden changes may cause bias in the estimates of the delay adjustment factors.

Solution

To reduce the impact of a secular trend for a given reporting year on the delay adjustment factors, we identify reporting years with sudden changes in adds and drops and add indicator variables to the delay models to represent the fixed effect of these reporting years.

Identifying reporting year effects

In the following table, we use 2008 as the most recent reporting year to present the structure of the data with varying years of diagnosis and reporting years in SEER 9 registries. The first reporting year is 1983 and corresponds to diagnosis year 1981. The 1981 diagnosis year has been reported 26 times in 26 distinct submissions. From the table, we see that the minimum reporting delay is two years and occurs at the first submission. The first submission for a given diagnosis year contains only adds, while all subsequent submissions contain both adds and drops.

Table 1. Number of times data have been submitted across reporting year and year of diagnosis (dx) for SEER 9 registries.

Reporting Year	Dx Year 1981	Dx Year 1982	Dx Year 1983	Dx Year 1984	Dx Year 1985	Dx Year 1986	Dx Year 1987	…	Dx Year 2006
1983	1								
1984	2	1							
1985	3	2	1						
1986	4	3	2	1					
1987	5	4	3	2	1				
1988	6	5	4	3	2	1			
1989	7	6	5	4	3	2	1		
…	…	…	…	…	…	…	…	…	…
2008	26	25	24	23	22	21	20		1

When there is a reporting-year effect, the number of add or drops in that year will be increased or reduced across all diagnosis years in one row in Table 1.

Note that we do not allow reporting-year effects in the two most recent reporting years. This is because the data from the last two reporting years are not included in the model selection procedure.

Reporting Year Effects in SEER 9: 1983-2008 data submissions

We first consider possible reporting-year effects for the add counts. We compare add counts across reporting years in a systematic manner to determine which reporting years have an unusually large increase or decrease in the add counts. To make a valid comparison across reporting years we need to compare equivalent submissions. For example, we could compare all of the third submissions in Table 2 across reporting years (i.e. across 1985 through 2006) and look for outliers. However, because add and drop counts tend to have a lot of variability, to reduce the variability, we compare equivalent groups of submissions, rather than single submissions, across reporting years (e.g. the sum of submissions 2 and 3 from reporting year 1985 through 2006). Below we describe precisely how equivalent groups of submissions are compared to search for outliers. First submissions are not included for consideration because by definition a reporting year effect represents a change from a prior submission (either adds or drops).

The first step of this process is to consider only those reporting years with at least 3 or more diagnosis years available. Reporting years 1985 and forward meet this criteria (see table 1). For each reporting year beginning with reporting year 1985, we calculate with the second and third newest submission counts for a given diagnosis year. In Table 1 for example, for reporting year 1985, we sum the counts for diagnosis year 1981 and 1982; for reporting year 1986, we sum the counts for diagnosis year 1982 and 1983, etc. In the second step, the mean and the standard deviation of these summed counts across reporting years are calculated; then each reporting year is scored by its difference from the mean. This set of scores are shown in blue in Table 2 and are denoted Score(reporting year,2). If the difference from the mean is greater than 1.96 (97.5 percentile point of the normal distribution) standard deviations, the reporting year receives a score 1; if the difference is greater than 2.58 (99.5 percentile point of the normal distribution) standard deviations, the reporting year receives a score 3. Otherwise, the reporting year receives a score 0.

The process above is reiterated for each reporting year that has at least k diagnosis years (k=3 to 19). The add counts of the k most recent diagnosis years, *excluding* the newest diagnosis year, are summed and scores are assigned to the corresponding reporting year. To establish a stable mean and standard deviation to search for outliers, we have used a minimum of five reporting years for any comparison. Thus the maximum value for k is 19. The score corresponding to reporting year j and the k most recent diagnosis years is called score(j, k), j=1985 to 2006, k = 2 to K_j, where K_j is the maximum number of diagnosis years used for reporting year j. The iterative process of comparing equivalent groups of submissions and scoring is shown in table 2 below:

	Diagnosis Year										
Reporting year		1981	1982	1983	1984	1985	1986	1987	…	2003	2004
	1985	Score(1985,2)		(exclude)							
	1986		Score(1986,2)		(exclude)						
		Score(1986,3)			(exclude)						
	1987			Score(1987,2)		(exclude)					
			Score(1987,3)			(exclude)					
		Score (1987,4)				(exclude)					
	1988				Score(1988,2)		(exclude)				
				Score(1988,3)			(exclude)				
			Score(1988,4)				(exclude)				
		Score (1988,5)					(exclude)				
	…	…	…	…	…	…	…	…	…	…	
	2006	Score (2006,19)									(exclude)

To be concise, the summary of scores may be written as score(1985-2006,2), score(1986-2006,3), …, score(2002-2006,19). We next determine if a particular reporting year has enough scores that are outliers to qualify as a reporting year effect.

A particular reporting year is identified as having a reporting year effect if it has at least 4 scores equal to 1 or a total score equal to or greater than 5. The average score for reporting year j is

$$(\sum_{k=2}^{K_j} \text{score}(j,k)\ ) / (K_{j}\text{-}1).$$

The same procedure is performed to identify reporting-year effects for the drop counts. If more than three reporting-year effects (add plus drop) are identified, we keep only the three that have the highest average scores. If a particular reporting year has both add and drop reporting-year effects, then the greater average score is used in the comparison.

In the future when we receive new submissions beyond 2008, the maximum above mentioned *k* will increase by one each year to still keep comparing a minimum of five reporting years. For example, for the 2009 submission, the scores to be calculated are: score(1985-2007,2), score(1986-2007,3), …, score(2002-2007,19) , score(2003-2007,20).

Reporting Year Effects in SEER 13-9: 1994-2008 data submissions

For SEER 13 – 9, the first available year of diagnosis is 1992 and the first reporting year is 1994. Similar to the steps for SEER 9, we calculate the scores for both add and drop counts for SEER 13-9. These scores are summarized as score(1996-2006,2), score(1997-2006,3), …, score(2002-

2006,8). To be considered as a reporting year effect, the scores corresponding to the reporting year must have at least three scores equal to 1 or a single score equal to 3. For each qualified reporting year, a final score is calculated. If there are more than two qualified reporting years, only the two with the highest final scores are kept.

Similar to SEER9, in the future when the most recent reporting year goes beyond 2008, more scores will be calculated. For example, for the 2009 submission, the scores to be calculated are: score(1996-2007,2), score(1997-2007,3), …, score(2002-2007,8), score(2003-2007,9).

The reporting year effects for the two data sets SEER9 and SEER13-9 are available in the summary table of delay model covariates at: http://www.srab.cancer.gov/delay/covariates.html

References

Midthune DN, Fay MP, Clegg LX, Feuer EJ. Modeling reporting delays and reporting corrections in cancer registry data. *J Am Stat Assoc* 2005;100(469):61-70.