JP Journal of Biostatistics
Volume 6, Number 1, 2011, Pages 1-18

Published by Pushpa Publishing House, Allahabad, INDIA

STATISTICAL METHODS TO CHECK AGREEMENT BETWEEN TWO CODING SYSTEMS IN THE ABSENCE OF DOUBLE-CODED DATA

Qingzhao Yu[1,*], **Xiaocheng Wu**[2], **Patricia Andrews**[2], **Donald Mercante**[1], **Praveen Ranganath**[2], **Umed Ajani**[3], **Baozhen Qiao**[4] and **Brad Wohler**[5]

[1]Biostatistics Program
School of Public Health
LSU Health Sciences Center, U. S. A.
e-mail: qyu@lsuhsc.edu

[2]Epidemiology Program
School of Public Health
LSU Health Sciences Center, U. S. A.

[3]Cancer Prevention and Control
Chronic Disease Prevention and Health Promotion, CDC, U. S. A.

[4]New York State Cancer Registry, U. S. A.

[5]University of Miami Miller School of Medicine
Florida Cancer Data System and Sylvester Comprehensive Cancer Center, U. S. A.

Abstract

In the past 25 years, the cancer surveillance community has revised several times the coding systems used to record cancer stage at diagnosis.

2010 Mathematics Subject Classification: 62P10.

Keywords and phrases: agreement, cancer stage distribution, EM algorithm, likelihood ratio test, NAACCR, SEER.

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

[*] Corresponding author

Received December 3, 2010

Most recently, the Collaborative Staging (CS) System replaced the directly coded Surveillance, Epidemiology and End Results (SEER) Summary Staging 2000 (SS2000) system for cancer cases diagnosed in 2004 and after in the United States. SS2000 can be derived from CS, producing the so-called Derived SS2000. But CS and SS2000 have not been used to code the same cancer cases simultaneously. Checking the agreement of these two systems has proved challenging. When comparing the stage of cancer cases diagnosed before and after January 1, 2004, observed differences in stage distributions could be attributed to true changes in stage trends, to decreases in the percentage of cases with unknown stage due to improved staging schema in the CS, to differences in the staging instructions between the two staging systems, or to other factors. This paper proposes a method to check the agreement of the two staging systems by comparing the cancer stage distributions, when cases in 2001-2003 are staged using SS2000 and cases in 2004 by CS. We build models to separate the impact of non-coding factors, such as natural trends, from those related to changes in coding instructions. If the non-coding factors satisfactorily explain the differences in stage distributions from pre-CS to CS diagnosis years, then we may conclude that the two systems have no significant coding differences. Otherwise, we provide directions for determining the nature of the discrepancies. Cancer data from the North American Association of Central Cancer Registries are presented to illustrate the method used in checking the agreement between the CS and SS2000 coding systems.

1. Introduction

Cancer stage describes the extent of cancer and how far it has spread from its origin at diagnosis. Cancer cases have been historically staged by two major staging systems, each with its unique purpose. The American Joint Committee on Cancer (AJCC) stage (Greene et al. [5]) is used mainly by clinicians who need clinically relevant data to make decisions on treatment and to evaluate prognosis. The directly coded Surveillance, Epidemiology and End Results (SEER) Summary Stage (Young et al. [10]) and its predecessors are mainly used by epidemiologists who require less complex data to examine stage distributions in different populations, to monitor stage trends, and to evaluate the effectiveness of early detection intervention programs. To meet the needs of clinicians and epidemiologists, tumor abstractors coded cancer cases with both stage systems until 2004.

In order to have a single staging system to meet the needs of both sectors, the

Collaborative Staging (CS) System was implemented, effective for cancer cases diagnosed in 2004 and after in the United States. Both the SEER Summary Stage and AJCC stage can be automatically derived using CS data items. The transition to CS has raised a question among epidemiological researchers studying staging trends as to the comparability of stage coded directly using the SEER Summary Stage 2000 (SS2000) and stage coded using CS and derived SEER Summary Stage 2000 (CS DRSS2000). The challenge with assessing comparability is that cases diagnosed in post-CS years (such as 2004) and pre-CS years (such as 2001-2003) were staged using one staging system or the other, but not both. Hence, traditional statistical methods for assessing agreement, such as the Kappa measure of agreement (Sim and Wright [8]), cannot be used.

If we assume that the cancer stage distributions for each cancer site remained constant over time, the presence of significantly different distributions of coded stage, say, from 2003 (using SS2000) to 2004 (using CS DRSS2000) would suggest the possible disagreement of the two coding systems. Under this assumption, we can use the likelihood ratio test to examine the consistency of stage distribution, which reflects the agreement of the two coding systems. However, this assumption may not be valid. Variations in cancer stage distributions can happen due to yearly trends or to factors such as changes in insurance policies or improvements in early detection. To evaluate discrepancies in stage at the time of a major coding revision, we propose new statistical approaches for checking the agreement between two staging systems when no cancer cases were staged by both systems. We assume that, after adjusting for known factors that potentially affect cancer stage distributions, cancer stage distributions will be consistent over time.

In this paper, we use a dataset from the North American Association of Central Cancer Registries (NAACCR) described in the next section. Section 3 presents the models adjusting the data for possible factors, such as linear trend, that might influence cancer stage distribution as well as the agreement tests on the model adjusted data. The corresponding computational methods are also described. Section 4 presents the results of comparing the two cancer staging coding systems using Cancer in North America (CINA) Deluxe data. Finally, limitations of the method and future research are discussed in Section 5.

2. CINA Deluxe Dataset

The 2001-2004 cancer incidence data were originally submitted to NAACCR by

40 population-based cancer registries in December 2007. These registries participate in the National Cancer Institute’s SEER Program and/or the Centers for Disease Control and Prevention’s (CDC) National Program of Cancer Registries (NPCR). For more information about the CINA datasets, the readers are referred to the website: http://www.naaccr.org/.

We used SEER*Stat, software provided by SEER (check the website http://seer.cancer.gov/statistics/ for details), to generate cancer counts by year of diagnosis, stage, registry, and cancer site. We focused on four main stage categories: localized, regional, distant and unknown. We used the data for the diagnostic years 2001-2004, the most up-to-date years at the time of analysis. We did not use data before 2001 because they had been recorded with yet another stage coding system, Summary Stage 1977.

Figure 1 gives examples of stage distributions from 2001 to 2004 for selected cancer sites: other non-epithelial skin (ONES), corpus and uterus, soft tissue including heart (STIH), colon and rectum, as well as larynx. To examine the agreement between SS2000 and CS DRSS2000, we focus on the variations of cancer stage distributions (proportion of each stage) between 2003 and 2004. We pose the following two questions: (1) If a difference in stage distribution exists between the two years, can we attribute it to non-coding factors such as natural trends in stage distributions over years or to the disagreement of the two staging systems, or both? (2) If the discrepancy can be attributed to coding differences, is it due to the fact that CS can stage some cases that would be classified as unknown by SS2000, or is it due to a more essential disagreement between the two stage systems?

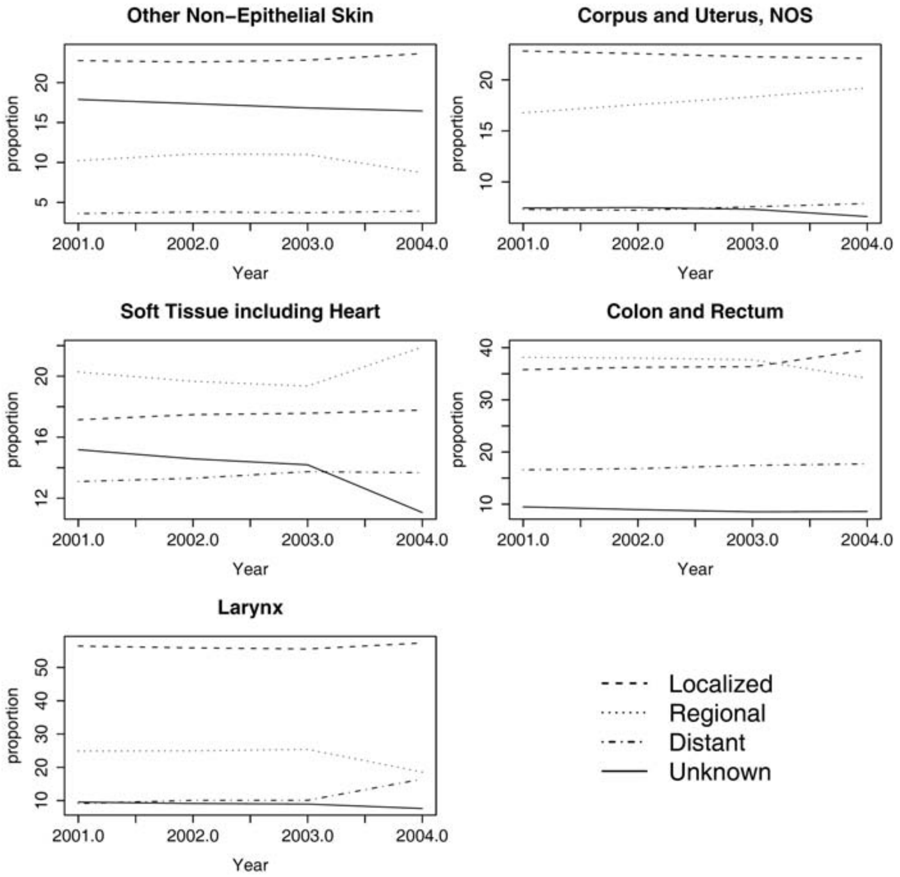


Figure 1. Stage distributions of some example sample sites. The proportions of localized ONES, corpus and uterus and STIH cancers are divided by 4 to reduce the vertical scale and accommodate the plots from all four stages.

From Figure 1, we observe that trends of stage distributions vary by cancer site. For example, stage distributions for ONES cancer vary little across years, whereas the corpus and uterus cancer stage distributions follow a linear trend – the proportions of localized and unknown stages were decreasing while those for regional and distant stages were increasing over time. Likewise, for STIH, stage distributions also follow a linear trend until 2003, when there were a noticeable drop in the proportion of unknown stage and a corresponding increase in the proportion of distant stage from 2003 to 2004. The latter might be explained by the fact that the CS system includes more detailed coding instructions, thus enabling the staging of some cancer cases that were classified as unknown by the SS2000 system. For colon and rectum cancer, there seems to be a linear trend in the stage distributions until 2003, but from 2003 to 2004 a sudden increase in the proportion of localized cancer and a corresponding decrease in regional cancer emerge. This change may also be related to the implementation of the CS system. If this difference in distribution is significant, then we should consider the possibility that some cancer cases that were staged as regional by SS2000 would be staged as localized by CS DRSS2000. This suggests that epidemiologists need to compare the coding instructions of the two staging systems in detail with special attention to the instructions for localized and regional stages. Finally, for laryngeal cancer, the changes in stage distributions are very irregular, necessitating detailed comparison of the staging instructions for the two systems.

The statistics questions to investigate are: first, whether the differences from 2003 to 2004 in stage distributions for a specific cancer site are significant and, secondly, can the statistically significant changes be explained by linear trends in cancer stage distributions and/or other known factors?

3. Statistical Models to Differentiate the Cancer Stage Distribution Variation from 2003 to 2004

First, we compare stage distributions of 2003 cases coded directly by using SS2000 with the stage distributions of the 2004 cases coded by using CS DRSS2000 to identify cancer sites with statistically significant differences in stage distributions from 2003 to 2004. The null hypothesis is that the stage distributions do not change from 2003 to 2004. We used the likelihood ratio test at a significance level of 0.05. For cancer sites that showed significant change in stage distributions from 2003 to 2004, we tested hypotheses related to the following two research questions: is the change in stage distributions attributable only to the 2001 to 2004 linear trends? Or is the change in the distribution attributable to a combination of the 2001-2004 linear trends and the presence of fewer unknown stage cases based on the CS coding system? In both tests, we assume that the overall linear trends do not change significantly over the years 2001-2004. If the answers are no to both research questions, then other factors may have an impact on the changes in stage distributions. If we know the factors, then we can add them to the model and test their significance. Otherwise, we manually review the coding instructions for SS2000 and CS DRSS2000 to check whether the two systems are equivalent. In the following, we describe the models for the two tests and the computational methods. If other factors should be considered, then similar methods could be used.

3.1. A model to adjust linear trend

We use the corpus and uterus cancer rates (the upper right plot of Figure 1) as an example to test whether the changes of stage distribution are attributable to the linear trends from 2001 to 2004. Table 1 displays the stage proportions (i.e., stage distribution) of corpus and uterus cancer cases by year. The number of cases by stage in each year is presented in parentheses.

Table 1. Proportions of corpus and uterus cancer cases

	Stages			
	Localized	Regional	Distant	Unknown
2001	68.50(20461)	16.77(5010)	7.29(2179)	7.44(2222)
2002	67.72(20342)	17.58(5281)	7.22(2169)	7.48(2247)
2003	66.80(20386)	18.33(5594)	7.22(2308)	7.31(2232)
2004	66.33(20831)	19.20(6030)	7.48(2473)	6.60(2073)

Note: Number of cases by stage within year are given in the parentheses.

Figure 1 and Table 1 show that, roughly, the percentages of localized (*L*) and unknown (*U*) stage cases were decreasing, while those of the regional (*R*) stage cases were increasing from 2001 to 2004. The percentage of distant (*D*) cases did not change too much, with the proportion in the year 2004 only a little higher than that in 2001. We assume that the variations in stage distributions are due to consistent linear trends in stage distributions and that as the proportions of localized and unknown stages decrease from 2001 to 2004, part of the decreases comes from the increases in the proportions of regional and/or distant stage cases.

If the complete data underlying Table 1 were observable, then the changes in stage distribution over time would be determined. The form of the complete (but unobserved) data for stage distributions is shown in Table 2:

Table 2. Form of unobserved complete corpus and uterus cancer data

Stages	Regional (↗)			Distant (↗)			Localized (↘)		Unknown (↘)	
Year	base	↗ from *L*	↗ from *U*	base	↗ from *L*	↗ from *U*	base	↘	base	↘
2001	y_{111}	0	0	y_{121}	0	0	y_{131}	y_{132}	y_{141}	y_{143}
2002	y_{211}	y_{212}	y_{213}	y_{221}	y_{222}	y_{223}	y_{231}	y_{232}	y_{241}	y_{243}
2003	y_{213}	y_{312}	y_{313}	y_{321}	y_{322}	y_{323}	y_{331}	y_{332}	y_{341}	y_{343}
2004	y_{411}	y_{412}	y_{413}	y_{421}	y_{422}	y_{423}	y_{431}	0	y_{441}	0

Note: ↘ means decreasing and ↗ means increasing.

We now present a model to analyze the probabilistic relationship between the underlying complete data and the observed data. To make the notation easy to be extended to other cancer sites, the model presents the situation where the proportions of the first two stages (here, stages 1 and 2 are regional and distant) increase while those of the last two stages (here, localized and unknown) decrease from 2001 to 2004. The following notation is employed:

π_i = The base proportion of stage i, $i \in \{1, 2, 3, 4\}$,

τ_j = The annual increasing proportion of stage j from stage 3, $j \in \{1, 2\}$,

ρ_k = The annual increasing proportion of stage k from stage 4, $k \in \{1, 2\}$. (1)

Under the assumptions described above, the probabilities underlying the unobserved complete data are shown in Table 3:

Table 3. Probabilities underlying the unobserved complete corpus and uterus cancer data

Stages	Stage 1 (↗)			Stage 2 (↗)			Stage 3 (↘)		Stage 4 (↘)	
Year	base	↗ from 3	↗ from 4	base	↗ from 3	↗ from 4	base	↘	base	↘
2001	π_1	0	0	π_2	0	0	π_3	$3(\tau_1+\tau_2)$	π_4	$3(\rho_1+\rho_2)$
2002	π_1	τ_1	ρ_1	π_2	τ_1	ρ_2	π_3	$2(\tau_1+\tau_2)$	π_4	$2(\rho_1+\rho_2)$
2003	π_1	$2\tau_1$	$2\rho_1$	π_2	$2\tau_2$	$2\rho_2$	π_3	$\tau_1+\tau_2$	π_4	$\rho_1+\rho_2$
2004	π_1	$3\tau_1$	$3\rho_1$	π_2	$3\tau_2$	$3\rho_2$	π_3	0	π_4	0

Note: ↘ means decreasing and ↗ means increasing.

In the observed data, some of the cells from the complete data are collapsed (see, e.g., Chen and Feinberg [2, 3]). Hence, we observe sums of several cells rather than all 40 possible cells represented in the complete-data tables. Table 4 presents the notation for the observed data table and indicates which cell counts from the complete data table are summed together to create the observed data table. The probabilities underlying the observed data are similarly just the sums of the probabilities underlying the unobserved complete data and are shown in Table 5.

Table 4. Form of observed data

	Stages			
	Stage 1	Stage 2	Stage 3	Stage 4
2001	$n_{11} = y_{111}$	$n_{12} = y_{121}$	$n_{13} = y_{131} + y_{132}$	$n_{14} = y_{141} + y_{142}$
2002	$n_{21} = y_{211} + y_{212} + y_{213}$	$n_{22} = y_{221} + y_{222} + y_{223}$	$n_{23} = y_{231} + y_{232}$	$n_{24} = y_{241} + y_{243}$
2003	$n_{31} = y_{311} + y_{312} + y_{313}$	$n_{32} = y_{321} + y_{322} + y_{323}$	$n_{33} = y_{331} + y_{332}$	$n_{34} = y_{341} + y_{343}$
2004	$n_{41} = y_{411} + y_{412} + y_{413}$	$n_{42} = y_{421} + y_{422} + y_{423}$	$n_{43} = y_{431}$	$n_{44} = y_{441}$

Table 5. Probabilities underlying the observed data

	Stages			
	Stage 1	Stage 2	Stage 3	Stage 4
2001	π_1	π_2	$\pi_3 + 3(\tau_1 + \tau_2)$	$\pi_4 + 3(\rho_1 + \rho_2)$
2002	$\pi_1 + \tau_1 + \rho_1$	$\pi_2 + \tau_2 + \rho_2$	$\pi_3 + 2(\tau_1 + \tau_2)$	$\pi_4 + 2(\rho_1 + \rho_2)$
2003	$\pi_1 + 2(\tau_1 + \rho_1)$	$\pi_2 + 2(\tau_2 + \rho_2)$	$\pi_3 + \tau_1 + \tau_2$	$\pi_4 + \rho_1 + \rho_2$
2004	$\pi_1 + 3(\tau_1 + \rho_1)$	$\pi_2 + 3(\tau_2 + \rho_2)$	π_3	π_4

We may estimate the π, τ and ρ parameters under this model using the EM-algorithm (see, for example, Dempster et al. [4]). The M-step involves maximizing the complete data likelihood function obtained using the cell probabilities shown in Table 3 and the complete data from Table 2, subject to the constraint that $\sum_{i=1}^{4}\pi_i + 3\sum_{j=1}^{2}\tau_j + 3\sum_{k=1}^{2}\rho_k = 1.$ The likelihood function, written so that the functions of the three types of parameters are obvious, is proportional to the following ("+" in a subscript indicates summation over the corresponding index):

$$L \propto [\pi_1^{y_{+11}} \times \pi_2^{y_{+21}} \times \pi_3^{y_{+31}} \times \pi_4^{y_{+41}}] \times [\tau_1^{y_{+12}} \times \tau_2^{y_{+22}} \times (\tau_1 + \tau_2)^{y_{+32}}]$$

$$\times [\rho_1^{y_{+13}} \times \rho_2^{y_{+23}} \times (\rho_1 + \rho_2)^{y_{+43}}].$$

With the restriction on the parameters and using a Lagrange multiplier, we wish to find estimators for π, τ and ρ that maximize the function $P = \ln L + \lambda\left(\sum_{i=1}^{4}\pi_i + 3\sum_{j=1}^{2}\tau_j + 3\sum_{k=1}^{2}\rho_k - 1\right).$ Taking the first partial derivatives of P and setting these derivatives equal to zero, we find the closed-form estimators for the parameters as follows:

$$\hat{\pi}_i = \frac{y_{+i1}}{n}, \qquad i = 1, 2, 3, 4;$$

$$\hat{\tau}_j = \frac{y_{+j2} \times y_{++2}}{3n(y_{+12} + y_{+22})}, \qquad j = 1, 2;$$

$$\hat{\rho}_k = \frac{y_{k3} \times y_{++3}}{3n(y_{+13} + y_{+23})}, \qquad k = 1, 2;$$

where n is the total number of cases in Table 2.

The E-step of the EM-algorithm consists of obtaining the expected cell counts for the complete data matrix (Table 5), given the observed data and the current estimates of the π, τ and ρ parameters. These expectations are particularly simple in the case of discrete data (see, for example, Little and Rubin [6]) and amount to proportionally allocating the n_{ij} of the observed data as shown in Table 4 to the y_{ijk} cells of Table 2 according to the current parameter estimates. For example,

$$\hat{y}_{412} = n_{41} \times \frac{3\hat{\tau}_1}{\hat{\pi}_1 + 3\hat{\tau}_1 + 3\hat{\rho}_1}.$$

Other expected cell counts may be found similarly and, hence, are not shown here.

The E- and M-steps of the EM-algorithm are repeated until parameter estimates have converged to the desired degree of accuracy, in our case, when all estimated probabilities had a sum of absolute differences of no more than 10^{-8} between two iterations. Convergence occurred in 921 iterations for the corpus and uterus cancer data.

Our model has four π parameters, two τ and two ρ parameters with a single constraint. Thus, there are 7 free parameters to be estimated. The observed data of Table 4 have 16 cells with one constraint for each row: that the row sum is the total sample size of each year. Hence, we have 12 free cells and therefore 5 degrees of freedom for our model (see, for example, Bishop et al. [1]).

3.2. A model adjusting for linear trend and changes in unknown stages

In addition to linear trends over time, changes in cancer stage distributions may be because the new coding system provides more detailed coding instructions, thus enabling abstractors to stage some cancer cases that were categorized as unknown by the old system. If two staging systems do not agree, we want to identify how much of the disagreement is due to this improvement in the new staging system and, after adjusting for this improvement and linear trend, to determine whether the two systems now agree. The model we propose here is an extension of the model from Subsection 3.1. We use the site soft tissue including heart (STIH) cancer as an illustration. The distribution of the cancer is shown in the middle of the first column in Figure 1.

Table 6 lists the stage distributions of STIH cancer over the years 2001 to 2004. We found that the proportions of localized and distant stages were increasing. Overall, the proportion of regional stage decreased, with a sudden increase observed from 2003 to 2004. This may be caused by some cases that would be classified as unknown by SS2000 being staged as distant by the CS DRSS2000.

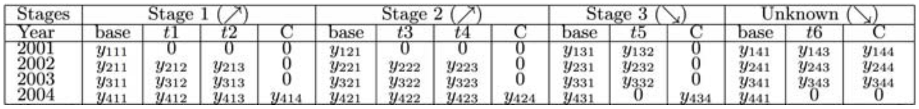
Table 6. Stage proportion of soft tissue including heart cancer cases by year

	Stages			
	Localized	Regional	Distant	Unknown
2001	51.44(3456)	20.27(1362)	13.10(880)	15.18(1020)
2002	52.44(3574)	19.66(1340)	13.31(907)	14.59(994)
2003	52.71(3873)	19.35(1422)	13.75(1010)	14.19(1043)
2004	53.34(4087)	21.91(1679)	13.68(1048)	11.09(848)

Note: The number of cases by stage and year are given in parentheses.

In this model, we assume that the proportions of localized and distant stages increased linearly over time. These increases came partly from the decreases occurring in the regional and unknown stages. From 2003 to 2004, part of the increases in all known stages reflects the greater precision of CS, allowing some previously unknown stage cases to be categorized as “known” using CS DRSS2000. If the complete data underlying Table 6 were observable, then the changes in stage distribution over time could be determined. The form of the complete (but unobserved) data for stage distributions is shown in Table 7. To make the tables easily adoptable for other cancer sites, we assume that the first two stages increased and the last two stages decreased from 2001 to 2004. Stage 4 is the Unknown stage (we observed that for almost all cancer sites, the proportion of unknown was decreasing over the years). For the STIH cancer, stages 1 and 2 are localized and regional, while stage 3 is distant.

Table 7. Form of unobserved complete soft tissue including heart cancer data

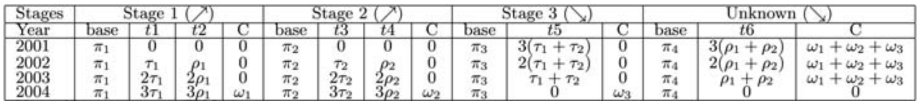


Stages	Stage 1 (↗)				Stage 2 (↗)				Stage 3 (↘)			Unknown (↘)		
Year	base	$t1$	$t2$	C	base	$t3$	$t4$	C	base	$t5$	C	base	$t6$	C
2001	y_{111}	0	0	0	y_{121}	0	0	0	y_{131}	y_{132}	0	y_{141}	y_{143}	y_{144}
2002	y_{211}	y_{212}	y_{213}	0	y_{221}	y_{222}	y_{223}	0	y_{231}	y_{232}	0	y_{241}	y_{243}	y_{244}
2003	y_{311}	y_{312}	y_{313}	0	y_{321}	y_{322}	y_{323}	0	y_{331}	y_{332}	0	y_{341}	y_{343}	y_{344}
2004	y_{411}	y_{412}	y_{413}	y_{414}	y_{421}	y_{422}	y_{423}	y_{424}	y_{431}	0	y_{434}	y_{441}	0	0

Note: ↘ means decreasing and ↗ means increasing. *t*1 and *t*2 mean increases in stage 1 from stage 3 and unknown, respectively. *t*3 and *t*4 mean increases in stage 2 from stage 3 and unknown, respectively. *t*5 and *t*6 are the linear trends in stage 3 and unknown, respectively. C means changes from unknown stages due to the use of new coding system.

In addition to the parameters defined by Equations in (1), parameter ω is estimated to account for the changes in unknown stages. Let ω_l be the proportion of change from the unknown stage using directly coded SS2000 to stage l by CS DRSS2000, where $l \in \{1, 2, 3\}$. The underlying probabilities of Table 7 are shown in Table 8.

Table 8. Probabilities underlying the unobserved complete data for soft tissue including heart cancer



Stages	Stage 1 (↗)				Stage 2 (↗)				Stage 3 (↘)			Unknown (↘)		
Year	base	$t1$	$t2$	C	base	$t3$	$t4$	C	base	$t5$	C	base	$t6$	C
2001	π_1	0	0	0	π_2	0	0	0	π_3	$3(\tau_1+\tau_2)$	0	π_4	$3(\rho_1+\rho_2)$	$\omega_1+\omega_2+\omega_3$
2002	π_1	τ_1	ρ_1	0	π_2	τ_2	ρ_2	0	π_3	$2(\tau_1+\tau_2)$	0	π_4	$2(\rho_1+\rho_2)$	$\omega_1+\omega_2+\omega_3$
2003	π_1	$2\tau_1$	$2\rho_1$	0	π_2	$2\tau_2$	$2\rho_2$	0	π_3	$\tau_1+\tau_2$	0	π_4	$\rho_1+\rho_2$	$\omega_1+\omega_2+\omega_3$
2004	π_1	$3\tau_1$	$3\rho_1$	ω_1	π_2	$3\tau_2$	$3\rho_2$	ω_2	π_3	0	ω_3	π_4	0	0

Note. Notation as in Table 7.

As in Subsection 3.1, some cells in the observed data are collapsed from the complete data. We can easily determine how the complete data are related to the observed data, which are similar to those in Table 4. Again, we can estimate the π, τ, ρ and ω parameters under this model using the EM-algorithm. For the M-step, the likelihood function is proportional to the following:

$$L \propto [\pi_1^{y_{+11}} \times \pi_2^{y_{+21}} \times \pi_3^{y_{+31}} \times \pi_4^{y_{+41}}] \times [\tau_1^{y_{+12}} \times \tau_2^{y_{+22}} \times (\tau_1 + \tau_2)^{y_{+32}}]$$

$$\times [\rho_1^{y_{+13}} \times \rho_2^{y_{+23}} \times (\rho_1 + \rho_2)^{y_{+43}}]$$

$$\times [\omega_1^{y_{414}} \times \omega_2^{y_{424}} \times \omega_3^{y_{434}} \times (\omega_1 + \omega_2 + \omega_3)^{y_{+44}}].$$

Obtaining the MLEs for the parameters involves maximizing the likelihood function subject to the constraint that $\sum_{i=1}^{4} \pi_i + 3\sum_{j=1}^{2} \tau_j + 3\sum_{k=1}^{2} \rho_k + \sum_{l=1}^{3} \omega_l = 1.$ Again, adding a Lagrange multiplier and taking derivatives, we find the closed-form estimators for the parameters as follows:

$$\hat{\pi}_i = \frac{y_{+i1}}{n}, \qquad i = 1, 2, 3, 4;$$

$$\hat{\tau}_j = \frac{y_{+j2} \times y_{++2}}{3n(y_{+12} + y_{+22})}, \qquad j = 1, 2;$$

$$\hat{\rho}_k = \frac{y_{+k3} \times y_{++3}}{3n(y_{+13} + y_{+23})}, \qquad k = 1, 2;$$

$$\hat{\omega}_l = \frac{y_{4l4} \times y_{++4}}{ny_{4+4}}, \qquad l = 1, 2, 3;$$

where *n* is the total number of cases in Table 7.

The E-step, as previously described in Subsection 3.1, involves obtaining the expected cell counts for the complete data (Table 7). For the EM algorithm, we use the same convergence rule as in Subsection 3.1. For the STIH cancer data, convergence occurred in 2418 iterations (this took less than 10 seconds running in *R*, version 2.8.0). This model has 10 free parameters and 12 free cells, resulting in 2 degrees of freedom.

3.3. Variance estimation

We use the jackknife method to estimate the variances of the estimators. More specifically, we consider each tumor registry as an independent sampling unit (SU). This results in 40 sampling units for our application. We explain the method through an example. Assuming that we would like to estimate the variance of the estimator for parameter τ_1, the best estimate of τ_1, $\hat{\tau}_1$, is obtained using all the 40 SUs. Then by deleting one SU, we use the jackknife data set of 39 resampled SUs to get another estimate, say, $\hat{\tau}_{11}$. In the next step, a new resampling is performed with a different SU being deleted, and a new estimate $\hat{\tau}_{12}$ is obtained from the second jackknife data set. The process is repeated for each sample unit, resulting in a set of estimates, $\hat{\tau}_{1i}$, $i = 1, \ldots, 40$. The variance for $\hat{\tau}_1$ is then estimated by

$$\hat{\sigma}^2_{\hat{\tau}_1} = \frac{39\sum_{i=1}^{40}(\hat{\tau}_{1i} - \hat{\tau}_1)^2}{40}.$$

Variance calculations are similar for the other parameter estimates.

4. Results and Conclusions

To evaluate the agreement of the SS2000 and CS DRSS2000 for each cancer site, we follow the following steps. First, we check the hypothesis that the stage distributions keep constant over the years (test 1 in Table 9). If the hypothesis is accepted, then we conclude that SS2000 and CS DRSS2000 agree. Otherwise, we go to step 2, where we assume that the distributions of stages followed a linear trend from 2001 to 2004. We adjust for the linear trend using the model described in Subsection 3.1 and check whether the stage distributions differ significantly from the assumed linear trend (test 2 in Table 9). If not, then we conclude that the SS2000 and CS DRSS2000 are consistent. Otherwise, we go to step 3, where we check whether the difference in stage distributions can be explained by linear trend and/or the ability of CS DRSS2000 to reduce the number of unstageable cases (test 3 in Table 9). We use model described in Subsection 3.2. This test provides a guide for further comparisons of the staging instructions of the two staging systems. If linear trends and/or decreases in percentage of unknown stage disease cannot explain the differences in stage distribution between 2003 and 2004, then we have to consider more complicated trends and/or other possible influencing factors.

Table 9 shows the test results for the cancer sites presented in Figure 1. We found that for other non-epithelial skin (ONES) cancer, the distributions of stage do not change much over the years. For corpus and uterus cancer, linear trends can explain the changes in stage distributions very well. Linear trends and changes in unknown stage can explain the changes in distribution of soft tissue including heart (STIH) cancer stages very well. Neither model we proposed can explain larynx or colon and rectum cancer well, but we found linear trends in both cancers and the significant changes in proportion of unknowns in larynx cancer, by the significant decreases in corresponding likelihood.

Table 9. Test results for the sample cancer sites presented in Figure 1

Cancer sites	Test 1			Test 2			Test 3		
	lkhd	df	*p*-value	lkhd	df	*p*-value	lkhd	df	*p*-value
ONES	19.38	12	0.08	–	–	–	–	–	–
CAU	99.71	12	< 0.001	8.5	5	0.13	–	–	–
STIH	81.07	12	< 0.001	27.26	5	< 0.001	0.59	2	0.74
CAR	794.96	12	< 0.001	212	5	< 0.001	211	2	< 0.001
LAR	436.55	12	< 0.001	151.69	5	< 0.001	104.72	2	< 0.001

Note: ONES stands for the site other non-epithelial skin, CAU for corpus and uterus, STIH for soft tissue including heart, CAR for colon and rectum and LAR for larynx.

The following two subsections show the detailed results for analyzing corpus and uterus cancer and soft tissue including heart cancer. For the comparison results of many other cancer sites, the readers are referred to Wu et al. [9].

4.1. Analysis of the corpus and uterus cancer data

We fit the model described in Subsection 3.1 for corpus and uterus cancer. The original data are shown in Table 1, and the estimators of parameters obtained from the modeling procedure are shown in Table 10. The values in parentheses are the estimated standard deviations for the corresponding estimators. All the estimates are significantly different from 0. It is estimated that for corpus and uterus cancer, the proportion of regional stage increased 0.59% and the proportion of distant stage increased 0.15% per year because of the decrease in localized stage. The proportion of regional stage increased 0.22% and the proportion of distant stage increased 0.06% per year because of the decrease in proportion of unknown stage.

Table 10. Estimates for corpus and uterus cancer

	Stages			
	Localized	Regional	Distant	Unknown
2001	20443.88	5007.66	2143.91	2276.56
2002	20337.07	5277.28	2218.26	2206.39
2003	20438.07	5607.28	2317.15	2157.50
2004	20800.90	6022.88	2449.70	2133.53

EM parameter estimates. Values in parenthesis are the estimated standard deviations of the corresponding estimates:

$$\hat{\pi}_1 = 0.17(1.13 \times 10^{-5}), \quad \hat{\pi}_2 = 0.07(6.27 \times 10^{-6}), \quad \hat{\pi}_3 = 0.66(7.36 \times 10^{-5}),$$

$$\hat{\pi}_4 = 0.07(1.03 \times 10^{-2}), \quad \hat{\tau}_1 = 0.0059(1.79 \times 10^{-6}), \quad \hat{\tau}_2 = 0.0015(2.65 \times 10^{-7}),$$

$$\hat{\rho}_1 = 0.0022(1.70 \times 10^{-6}), \quad \hat{\rho}_2 = 0.0006(2.00 \times 10^{-7}).$$

The likelihood ratio test to check whether the linear trends could satisfactorily explain the data showed the likelihood ratio statistic to be 8.5 with 5 degrees of freedom (p-value $= 0.13$). Comparing this with the simple model of consistent stage distribution (likelihood ratio statistics = 99.71, $d.f. = 12$ and p-value < 0.001) gave a difference in likelihood ratio statistics of 91.21 with 7 degrees of freedom. The p-value is smaller than 0.0001. Therefore, we conclude that there are significant linear trends in the corpus and uterus cancer stage distributions over the years 2001 to 2004 and that the linear trends alone can explain the changes in stage distributions. We conclude that SS2000 and CS DRSS2000 agree in staging corpus and uterus cancer.

4.2. Analysis of soft tissue including heart (STIH) cancer

The model described in Subsection 3.2 is used on cancer of the soft tissue including heart as an illustration. Table 11 shows the estimated results, compared with the original data in Table 6. We see that all estimators for parameters are significantly different from 0 except for ω_1. Also, ω_2 is barely 0. This is in correspondence with Figure 1. We conclude that 2.79% of the unknown cases using SS2000 rules were staged as regional by CS DRSS2000. Therefore, we suggest that epidemiologists evaluate this difference by comparing the staging instructions.

Table 11. Estimates for soft tissue including heart cancer

	Stages			
	Localized	Regional	Distant	Unknown
2001	3466.14	1354.68	882.46	1014.72
2002	3556.51	1348.64	909.85	999.99
2003	3878.14	1426.53	996.81	1046.53
2004	4089.20	1672.31	1055.88	844.62

EM parameter estimates. Values in parenthesis are the estimated standard deviations of the corresponding estimates:

$$\hat{\pi}_1 = 0.52(4.63 \times 10^{-5}), \quad \hat{\pi}_2 = 0.13(4.11 \times 10^{-5}), \quad \hat{\pi}_3 = 0.19(1.59 \times 10^{-4}),$$

$$\hat{\pi}_4 = 0.11(2.15 \times 10^{-4}), \quad \hat{\tau}_1 = 0.0028(8.55 \times 10^{-6}), \quad \hat{\tau}_2 = 0.0010(7.47 \times 10^{-7}),$$

$$\hat{\rho}_1 = 0.0032(8.25 \times 10^{-6}), \quad \hat{\rho}_2 = 0.0011(2.35 \times 10^{-6}), \quad \hat{\omega}_1 = 2.92 \times 10^{-11}(3.60 \times 10^{-6}),$$

$$\hat{\omega}_2 = 7.02 \times 10^{-9}(2.54 \times 10^{-22}), \quad \hat{\omega}_3 = 0.0279(8.11 \times 10^{-5}).$$

The likelihood ratio statistics for checking whether the model described in Subsection 3.2 could satisfactorily explain the data is 0.59 with 2 degrees of freedom, and the *p*-value is 0.74. Therefore, we conclude that after accounting for the linear trends in stage distributions and the decrease in percentage of unknown stage, the two coding systems are consistent in staging STIH cancer. The decrease in unknown stage proportions results from the expanded coding instructions in Collaborative Staging, so some cases that would have been coded as unknown by

SS2000 now could be staged as regional stage. A detailed comparison of the staging instructions of SS2000 and CS DRSS2000 reveals that for STIH, the CS manual includes additional sites/subsites to define extension and lymph node involvement, which results in fewer cases classified as unknown (see Wu et al. [9]).

5. Discussion and Future Research

In this paper, we proposed methods to check the agreement between two coding systems when no case is coded by both systems. We take into account other factors such as the linear trends that might influence the stage distributions. If the known factors cannot account for all observed discrepancies in stage distribution, then further investigation of the agreement of the two coding systems is necessary. We applied the method to check SS2000 and CS DRSS2000 systems for coding cancer stage at diagnosis. Corpus and uterus cancer and soft tissue including heart cancer data from CINA Deluxe dataset were used as examples to test the two staging systems.

The data sets are updated every year, and using updated data (2001-2005, for example) might result in different conclusions for the same cancer sites. Therefore, a future research topic would be to check the robustness of the method. We hope that minor changes in data would not result in significantly different final results. Second, because we use only 4 years' data, we could test only a few factors together. For example, the tables in this paper have only 12 free cells, so we can have at most 12 free parameters. As more data are collected, we might be able to test more factors such as more complicated trends, the change of insurance coverage or the invention of new cancer screening methods. Another direction of research would be to use the previous years' data as prior information and implement the Bayesian method for analysis. For example, we did not use data before 2001 because those cancer cases were staged with still another coding system - SS1977. Although the stage system changes, most likely the natural trends in cancer stage distribution would not change. Therefore, we can distill the information from previous data as prior knowledge and incorporate the information in our models (Yu et al. [11]). This could result in more precise estimates.

References

[1] Y. M. M. Bishop, S. E. Feinberg and P. W. Holland, Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, MA, 1975.

[2] T. Chen and S. E. Feinberg, Two-dimensional contingency tables with both completely and partially cross-classified data, Biometrics 30 (1974), 629-642.

[3] T. Chen and S. E. Feinberg, The analysis of contingency tables with incompletely classified data, Biometrics 32 (1976), 133-144.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. Roy. Statist. Soc. Ser. B 39 (1977), 1-38.

[5] F. L. Greene, D. L. Page, I. D. Fleming, A. G. Fritz, C. M. Balch, D. G. Naller and M. Morrow, eds., AJCC Cancer Staging Manual, 6th ed., Springer-Verlag, 2002.

[6] R. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 2002.

[7] L. A. G. Ries, J. L. Young, G. E. Keel, M. P. Eisner, Y. D. Lin and M.-J. Horner, Editors, SEER Survival Monograph: Cancer Survival Among Adults: U.S. SEER Program, 1988-2001, Patient and Tumor Characteristics. National Cancer Institute, SEER Program, NIH Pub. No. 07-6215, Bethesda, MD, 2007.

[8] J. Sim and C. C. Wright, The Kappa statistic in reliability studies: use, interpretation, and sample size requirements, Physical Therapy 85 (2005), 257-268.

[9] X. C. Wu, Q. Yu, P. A. Andrews, P. Ranganath, B. Qiao, U. Ajani, B. Wohler and Z. Zhang, Comparisons of directly coded SEER summary stage 2000 and collaborative staging derived SEER summary stage 2000, J. Registry Manag. 37(4) (2010), 137-140.

[10] J. L. Young, S. D. Roffers, L. A. G. Ries, A. G. Fritz and A. A. Hurlbut, Editors, SEER Summary Staging Manual-2000: Codes and Coding Instructions, National Cancer Institute, NIHPub. No. 01-4969, Bethesda, MD, 2001.

[11] Q. Yu, E. A. Stasny and B. Li, Bayesian models to adjust for response bias in survey data: an example in estimating rape and domestic violence rates from the NCVS, Ann. Appl. Statist. 2 (2008), 665-686.