EXPLORING THE INTERNAL CONSISTENCY OF REGISTRY DATA ON STAGE OF DISEASE AT DIAGNOSIS: PART III: SUGGESTIONS FOR AN EDITS METAFILE

John P. Fulton, Pam Parrish, Steven D. Roffers, Catherine N. Correa, Nancy Jackson, Lilia O'Connor, Carin Perkins, Chris Newton, Gayle Clutter

This paper was prepared by the Data Quality Indicator Subcommittee of the NAACCR Data, Evaluation, and Publication Committee.

INTRODUCTION

In this report the Data Quality Indicator Subcommittee ("the Subcommittee") completes the study of inconsistencies in data on stage of disease at diagnosis begun in 1998,[1] and continued in 1999.[2] In Parts I and II, the Subcommittee focused on cancer of the female breast and cancer of the prostate. In Part III, the Subcommittee explores the internal consistency of registry data on stage of disease at diagnosis for all cancers combined, suggesting outputs for an EDITS metafile. EDITS is a software package designed by the Centers for Disease Control and Prevention (CDC) for performing internal data consistency edits.

Cancer registry data on stage of disease at diagnosis are particularly important for evaluating cancer control efforts designed to identify cancers at early stages of diagnosis, such as screening programs. "Summary Stage," an indicator of stage of disease at diagnosis pioneered by the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute, is used by a majority of member registries in the North American Association of Central Cancer Registries (NAACCR). Other indicators of stage, however, are also used among member registries, to comply with other national and international standards for the coding of cancer registry information, e.g., American Joint Commission on Cancer "Tumor-Nodes-Metastasis" (AJCC TNM), SEER "Extent of Disease" (EOD), and SEER "historic stage". These multiple indicators may be checked for internal consistency to assess the accuracy with which individual data items have been coded.

To date, information on stage of disease at diagnosis has not been subjected to widespread evaluation using internal consistency checks, electronic or otherwise. The development of simple edit checks with which such information may be evaluated is the subject of the present report. Such edit checks, when thoroughly developed and tested, may be used to identify and correct errors in the coding of summary stage and other indicators of stage of disease at diagnosis.

OBJECTIVES

Using 1996 incidence data for all cancers combined from four central cancer registries, the Subcommittee analyzed the internal consistency of data on stage of disease at diagnosis with several objectives:

1. to develop an analytical framework with which to identify consistencies and inconsistencies between staging schemes ***for all cancers combined***;
2. to analyze case reports with combinations of staging data not evaluated as "consistent;"
3. to measure the yield of errors from combinations of staging data not evaluated as "consistent;" and
4. to propose EDITS outputs for different combinations of staging data, based on the yield of errors.

METHODS

"EDITS"

EDITS is a software package designed by the Centers for Disease Control and Prevention (CDC) for analyzing the internal consistency of complex data sets, such as cancer registry data files. Checking the internal consistency of

values across variables is an essential approach to improving the quality of complex data. Impossible or unlikely combinations of values (e.g., codes indicating cancer of the prostate in a female or cancer of the cervix in a male) are identified, investigated, and corrected if found to be in error. Although such checks may be undertaken manually, computerized checks of internal consistency, such as those which may be used by EDITS, have been developed over time which increase the efficiency of the process. Once a year, NAACCR issues its own EDITS "metafile," containing consistency checks especially written for cancer incidence data in NAACCR's record exchange format, to be used by member registries before their annual submission of data for publication in *Cancer in North America* (*CINA*).

Variables Selected for Consistency Checks

Summary stage (SS) was selected as the standard for comparison, because of its pervasive use among NAACCR registries to categorize stage of disease at diagnosis. The other two indicators of stage were the SEER Program's EOD system and the AJCC TNM system. For examination of regional nodes, SS was compared with either SEER's number of regional lymph nodes positive (NRNP) or the node element of AJCC's TNM (TNM). For examination of site of distant metastases, SS was compared to SEER's sites of distant metastases (SDM) or the metastasis element of AJCC's TNM (TNM).

Abbreviations

Abbreviation	Staging System	Variable	*NAACCR #*
SS	Summary stage	Summary stage	760
NRNP	EOD	Number of regional nodes positive	820
SDM	EOD	Sites of distant metastases	1090
TNM	AJCC TNM	Node element	890 or 950
TNM	AJCC TNM	Metastasis	900 or 960

Sources of Data

Cancer registries in four states (Georgia - Comprehensive Cancer Registry, Arizona, California - Bay Area, and Illinois) contributed data for this study. Information on stage of disease at diagnosis was extracted on case reports for cancers diagnosed in 1996.

Data Preparation

NRNP data and TNM data were recoded into four categories: 1) no regional nodes identified; 2) some regional nodes identified; 3) regional nodes not examined or cannot be assessed medically; 4) no information on regional nodes. SDM data were recoded into three categories: 1) no distant metastases identified; 2) specific sites of distant metastases identified; 3) other sites of distant metastases or no information on distant metastases. TNM data were recoded into four categories: 1) no metastases identified; 2) some metastases identified; 3) metastases cannot be assessed medically; 4) no information on metastases.

Analytical Framework and Table Construction

Combinations of SS and either NRNP or TNM were evaluated using coding manuals and the expertise of committee members in tumor classification, and one of three EDITS outputs was originally proposed for each cell in Table 1. Combinations of SS and SDM were evaluated and designated in the same manner for each cell in Table 2, as were combinations of SS and TNM for each cell in Table 3.

Criteria for Proposing EDITS Outputs

Edits Output	*Meaning*	*Percentage Yield of Errors*
Pass	Consistent	0 – 14
Warning	Sometimes inconsistent	15 – 84
Error	Inconsistent	85 – 100

Data Analysis

SS data were cross-tabulated with recoded data on nodes and metastases. Then, the analytical framework was tested by reviewing case reports with combinations of staging data evaluated as "error," or "warning." All case reports with data combinations originally evaluated as "error" were reviewed, and a sample of case reports with data combinations evaluated as "warning" were reviewed. Each participating registry was asked to contribute as many cases as it had identified, up to a maximum of 10 cases, to the sample in each cell evaluated as "warning."

Case narratives (the text fields in cancer abstracts) and other available sources of information were consulted to assess whether or not each case had truly inconsistent staging information. Those that did were recorded as such and corrected. Those that did not were recorded as such. The percent of errors found in each "error" and "warning" cell were recorded in Tables 1, 2, and 3. The sources of the inconsistencies and the reasons for them were not recorded as part of this study.

Modifications to the Original Proposed Cell Designations for EDITS Outputs

Modifications were made to the original proposed cell designations for EDITS outputs to bring them into line with the criteria described in *Analytical Framework*, above. After careful discussion, four exceptions were made in applying the criteria. Final proposed EDITS outputs are listed in Tables 1, 2, and 3, with the original proposed EDITS outputs.

Summary of Table Construction

For ease of comparison, Tables 1, 2, and 3 contain four sets of information:

- Original proposed EDITS output for each cell
- Aggregate frequency of data combinations for the central registries contributing to the table
- Percent of errors (error yield) found in each cell originally designated as "error" or "warning," using data from 2-4 central registries contributing to the table
- Final proposed EDITS output for each cell

Evaluation of Data Comparisons as the basis of an Edit Check

As in Parts I and II of this three-part series, the Subcommittee evaluated its approach to the design of proposed EDITS outputs, using three criteria formulated as questions:

(1) *Validity:* Is the approach valid? (Does it make sense? Is the logic correct?)
(2) *Effectiveness:* Is the approach effective? (Does it identify inconsistent data?)
(3) *Efficiency:* Is the approach efficient? (Can the relative yield of true errors be predicted from different combinations of staging data?)

RESULTS

Internal Consistency

The frequency of data combinations from the participating registries are listed in Tables 1, 2, 3, to give the reader a sense of the distribution of cases across cells for (114,239 cases in 1996). They are summarized here by table and by *final* proposed EDITS output:

Summary of frequencies from four participating registries, 1996, Table 1:
Combinations of summary stage and NRNP or TNM codes

• Pass:	113,941	99.74%
• Warning:	129	0.11%
• Error:	169	0.15%

Summary of frequencies from two participating registries, 1996, Table 2:
Combinations of summary stage and SDM codes

• Pass:	66,957	99.75%
• Warning:	48	0.07%
• Error:	121	0.18%

Summary of frequencies from two participating registries, 1996, Table 3:
Combinations of summary stage and TNM codes

• Pass:	63,493	99.60%
• Warning:	105	0.16%
• Error:	156	0.24%

Validation of Inconsistencies

Of six cells in Tables 1 and 3 originally proposed as "warning" cells:

- Four yielded less than 15 percent errors. For these four cells "warning" was changed to "pass" (Tables 1 and 3).
- Two yielded 100 percent errors (Table 1). Proposed EDITS outputs for these cells were not changed, in exception to the criteria listed in Table 1, because in some circumstances, some cases in these cells may not be in error.

Of 22 cells in Tables 1-3 originally proposed as "error" cells:

- Thirteen yielded 86-100 percent errors. Proposed EDITS outputs for these cells were not changed (Tables 1 and 3).
- Nine yielded less than 85 percent errors. Of these:
  - Six yielded 16-40 percent errors. Proposed EDITS outputs for these cells were changed to "warning" (Tables 2 and 3).
  - One yielded four percent errors. The proposed EDITS output for this cell was changed to "pass" (Table 2).

- Two yielded 27 percent and 18 percent errors, respectively. Proposed EDITS outputs for these cells were changed to "pass" (Tables 2 and 3), in exception to the criteria listed above, because of the large number of cases which would have to be searched to find the errors in these cells.

DISCUSSION

Is the analytical framework valid?

The logic of the comparisons appears to be largely correct. When cases in "error" cells were examined, the vast majority were found to have obvious coding errors. Similarly, when cases in "warning" cells were examined, many were found to have obvious coding errors. Also, the logic of the comparison is simple. It is not specific to particular cancer sites.

Is the analytical framework effective?

The approach is effective in finding coding errors with available information. The approach can be implemented simply. It requires one-way crosstabulation and the ability to identify cases within specific cells.

Is the analytical framework efficient?

The approach identifies very high proportions of error (85-100 percent) in "error" cells, and moderate proportions of error (15-84 percent) in "warning" cells.

Future Directions

The Subcommittee has shared its work with other NAACCR subcomittees: the Data, Evaluation, and Publication Committee's Staging Subcommittee and the Uniform Standard Committee's EDITS Subcommittee. The Cancer Staging Subcommittee is currently in the process of polling all cancer registries in North America regarding the cancer staging practices that are in place in each registry. The EDITS Subcommittee has designed a set of computerized edit checks based on staging data whose outputs closely conform to the final proposed outputs in Tables 2-4. There will be some exceptions in translation, which have to do with the realities of the NAACCR data exchange format, but most of the logic as proposed herein will be preserved. It is hoped that the edit checks thus designed may be incorporated into the next version of NAACCR's EDITS metafile.

In the meantime, it might be fruitful for registries to explore the internal consistency of their staging data, using the logic and procedures described in the present report. Staging data are of such essential value to the evaluation of screening interventions that they will quickly become the subject of national calls for data, with all the quality assurance checks inherent in such a process.

Conclusions

The results presented in previous reports for cancer of the female breast[1] and cancer of the prostate,[2] and the results presented herein for all cancers combined may be used to develop a computerized edit check to be incorporated in NAACCR's annual call for data. The proposed approach to the identification of inconsistent staging data has high face validity, and appears to be effective and efficient in finding true coding errors. Its use may help clarify the relationships among sources of staging data for registrars. Using data from four participating central registries, previously undetected errors were found in SS. In anticipation of increased national demand for accurate staging data, registries may wish to explore the internal consistency of their data using the logic and procedures illustrated in this report.

REFERENCES

1. Porter R, Correa CN, Fulton JP, Howe HL, Newton C, Nowak J, Roffers SD. Exploring the internal consistency of registry data on stage of disease at diagnosis. VI.1-12. In Chen VW, Wu XC, Andrews PA (eds). *Cancer in North America 1990-1994. Volume One: Incidence.* Sacramento, CA: North American Association of Central Cancer Registries, April, 1998.

2. Parrish P, Fulton JP, Correa CN, Roffers SD, Jackson N, Newton C, O'Connor L, Perkins C, Porter R, Tucker T, Weir H. Exploring the internal consistency of registry data on stage of disease at diagnosis: Part II: Cancer of the Prostate. VI.1-14. In Chen VW, Wu XC, Andrews PA (eds). 1999 *Cancer in North America 1991-1995. Volume One: Incidence.* Sacramento, CA: North American Association of Central Cancer Registries, April, 1999.

Table 1. Original / final proposed EDITS outputs for combinations of summary stage and NRNP or TNM codes, with:

- **Frequency of combinations from four central registries (GA - Comprehensive Cancer Registry, AZ, CA - Bay Area, and IL), 1996 (114,239 total cases)**
- **Percentage of errors found in E (error) and W (warning) cells in four central registries (GA - Comprehensive Cancer Registry, AZ, CA - Bay Area, and IL), 1996**
- **Shading to indicate revised proposed EDITS outputs, based on empirical evaluation**

	NRNP or TNM											
	No +Reg Nodes ID'd NRNP = 00; TNM N = 0			Some +Reg Nodes ID'd NRNP = 01-97; TNM N = 1, 2, or 3			Reg Nodes Not Examined or Can't be Assessed Medically NRNP = 98; TNM N = X			No Info on Reg Nodes NRNP = 99 or blank; TNM N is N/A		
Summary Stage	*	ALL #	ALL %	*	ALL #	ALL %	*	ALL #	ALL %	*	ALL #	ALL %
(0) In Situ	P / P	1098	-	E / E	1	100	P / P	5621	-	P / P	209	-
(1) Localized	P / P	19227	-	E / E	38	100	P / P	22917	-	P / P	3398	-
(2) Reg, DE **	P / P	5072	-	E / E	43	100	P / P	4046	-	P / P	683	-
(3) Reg, +N ***	E / E	19	95	P / P	6129	-	W / P	1153	14	W / W	85	100
(4) Reg, DE & +N	E / E	28	100	P / P	4354	-	W / P	1276	7	W / W	84	100
(5) Reg, NOS	P / P	119	-	P / P	240	-	P / P	634	-	P / P	928	-
(7) Dist Mets	P / P	978	-	P / P	2677	-	P / P	11960	-	P / P	6670	-
(9) Unknown	P / P	421	-	P / P	247	-	P / P	7438	-	P / P	6446	-

*P = Pass; W = Warning; E = Error**DE = Direct Extension***+N = Positive Nodes

Table 2. Original / final proposed EDITS outputs for combinations of summary stage and SDM codes, with:

- **Frequency of combinations from GA - Comprehensive Cancer Registry and IL, 1996 (67,126 total cases)**
- **Percentage of errors found in E (error) and W (warning) cells in two central registries (GA - Comprehensive Cancer Registry and IL), 1996**
- **Shading to indicate revised proposed EDITS outputs, based on empirical evaluation**

	SDM								
	No Dist Mets ID'd SDM=0			Specific Sites of Dist Mets ID'd SDM = 1 thru 8			Other Sites of Dist Mets ID'd or No Info on Dist Mets SDM = 9; Blanks are ignored		
Summary Stage	*	GA,IL #	GA,IL %	*	GA,IL #	GA,IL %	*	GA,IL #	GA,IL %
(0) In Situ	P / P	3899	-	E / E	0	None	P / P	5	-
(1) Localized	P / P	26204	-	E / E	6	100	P / P	235	-
(2) Reg, DE *	P / P	5645	-	E / W	56	40	P / P	75	-
(3) Reg, +N **	P / P	4638	-	E / E	6	100	P / P	39	-
(4) Reg, DE & +N	P / P	3445	-	E / W	50	16	P / P	32	-
(5) Reg, NOS	P / P	1423	-	E / W	15	20	P / P	29	-
(7) Dist Mets	E / P	859	27	P / P	7623	-	E / P	4296	4
(9) Unknown	P / P	4419	-	E / E	36	97	P / P	4091	-

* P = Pass; W = Warning; E = Error ** DE = Direct Extension *** +N = Positive Nodes

Table 3. Original / final proposed EDITS outputs for combinations of summary stage and TNM codes, with:

- **Frequency of combinations from GA - Comprehensive Cancer Registry and IL, 1996 (63,754 total cases)**
- **Percentage of errors found in E (error) and W (warning) cells in two central registries (GA - Comprehensive Cancer Registry and IL), 1996**
- **Shading to indicate revised proposed EDITS outputs, based on empirical evaluation**

	TNM											
	No Mets ID'd TNM M = 0			Some Mets ID'd TNM M = 1			Mets Cannot be Assessed Medically TNM M = X			No Info on Mets TNM M = N/A		
Summary Stage	*	GA,IL #	GA,IL %	*	GA,IL #	GA,IL %	*	GA,IL #	GA,IL %	*	GA,IL #	GA,IL %
(0) In Situ	P / P	2845	-	E / E	0	None	P / P	769	-	P / P	448	-
(1) Localized	P / P	16764	-	E / E	15	100	P / P	3650	-	P / P	2649	-
(2) Reg, DE **	P / P	4327	-	E / W	56	21	P / P	727	-	P / P	652	-
(3) Reg, +N ***	P / P	3926	-	E / E	18	100	P / P	354	-	P / P	381	-
(4) Reg, DE & +N	P / P	2875	-	E / W	79	37	P / P	397	-	P / P	358	-
(5) Reg, NOS	P / P	446	-	E / W	21	38	P / P	218	-	P / P	752	-
(7) Dist Mets	E / P	699	18	P / P	5037	-	W / P	1687	0	W / P	5117	11
(9) Unknown	P / P	694	-	E / E	72	86	P / P	3029	-	P / P	4692	-

* P = Pass; W = Warning; E = Error ** DE = Direct Extension *** +N = Positive Nodes