EXPLORING THE INTERNAL CONSISTENCY OF REGISTRY DATA ON STAGE OF DISEASE AT DIAGNOSIS

Richard Porter
Catherine N. Correa
John P. Fulton
Holly L. Howe
Chris Newton
Judy Nowak
Steven D. Roffers

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

INTRODUCTION

An essential approach to improving the quality of a cancer registry data set is to check each case for the internal consistency of its data. Impossible or unlikely combinations of data (e.g., cancer of the prostate in a female; 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 have been developed over time which increase the efficiency of the process. The North American Association of Central Cancer Registries (NAACCR) has developed a set of computerized internal consistency checks for evaluating the accuracy of data submitted for use in its annual publication, *Cancer in North America (CINA).*

Cancer registry information on stage of disease at diagnosis is 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 NAACCR's member registries in the United States. 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. 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 evaluation in NAACCR's set of computerized internal consistency checks. The development of computerized edit checks with which information on stage of disease at diagnosis 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 data on female breast cancer from four state cancer registries, the Data Quality Indicator Subcommittee of NAACCR Data Evaluation and Publication Committee studied the internal consistency of data on stage of disease at diagnosis with several objectives: (1) to develop a theoretical framework with which to identify inconsistencies between staging schemes for breast cancer; (2) to test the framework by reviewing case reports that had combinations of staging data evaluated as "inconsistent;" and (3) to explore the efficiency with which inconsistencies may be identified from case reports that had combinations of staging data evaluated as "sometimes inconsistent," but not always. The Subcommittee undertook the study with the goal of eventually developing a new reliability check to assess data submitted to NAACCR for publication in *CINA.*

METHODS

Sources of Data

Cancer registries in four states (Arizona, Illinois, Louisiana, and Rhode Island) contributed data for this study. Information on stage of disease at diagnosis was extracted from 39,675 reports of female breast cancer diagnosed from 1990 through 1995. Arizona and Louisiana contributed 1990-1994 data, Rhode Island contributed 1990-1995 data, and Illinois contributed 1995 data. (The last began collecting stage information using the extent of disease scheme in 1995.)

Data Analysis

Summary stage ("SS," NAACCR variable #760) was selected as the focus of analysis, because of its pervasive use among NAACCR registries to categorize stage of disease at diagnosis. SS was compared with other data on positive regional lymph nodes (either number of regional nodes positive, "NRNP," NAACCR variable #820; or the node element of the American Joint Committee on Cancer's [AJCC's] tumor-node-metastasis [TNM] system, "TNM," NAACCR variable #890 or #950), and with data on distant metastases (either site of distant metastasis, "SDM," NAACCR variable #1090; or the metastasis element of the AJCC's TNM system, "TNM;" NAACCR variable #900 or #960). Certain inconsistencies between SS and other data elements were investigated by referring to case narratives and other available sources of information. Results were used to assess three criteria for the adoption of a reliability check for SS: validity, effectiveness, and efficiency.

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 identified 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.

As the theoretical framework, combinations of SS and either NRNP *or* TNM were evaluated using coding manuals and the expertise of committee members in tumor classification and were designated as consistent, infrequently inconsistent, sometimes inconsistent, or inconsistent (Table 1). Combinations of SS and SDM were evaluated and designated in the same manner (Table 2), as were combinations of SS and TNM (Table 3).

SS data were cross-tabulated with recoded data on nodes and metastases (Tables 4–6) and their internal consistencies were examined. Case narratives (the text fields in cancer abstracts) and other available sources of information were consulted to assess the validity of all data designated as "inconsistent" (Table 7). True inconsistencies were corrected by revising SS, NRNP, TNM, SDM, or TNM. Some inconsistencies could not be assessed because additional information was unavailable at the central registry, e.g., text fields were incomplete or empty.

A limited, exploratory analysis was undertaken to evaluate cells of indeterminate consistency and to assess the potential yield of "inconsistent" information from cells designated as "sometimes inconsistent" or "infrequently inconsistent." Case narratives and other available sources of information were consulted to assess all cases in Table 4 ("SS versus NRNP *or* TNM") which had data combinations considered to be "sometimes inconsistent" (Table 8). These were the only cells in which information designated as "sometimes inconsistent" or "infrequently inconsistent" were assessed. A more comprehensive assessment was beyond the scope of the present study, but may be undertaken in the future.

Evaluation of the Theoretical Framework as the Basis of an Edit Check

Staff in each of the four state registries evaluated the theoretical framework as the basis of an edit check for use in the annual NAACCR *Call for Data*, using three criteria formulated as questions:

(1) *Validity:* Is the theoretical framework valid? (Does it make sense? Is the logic correct?)
(2) *Effectiveness:* Is the theoretical framework effective? (Does it identify inconsistent data?)
(3) *Efficiency:* Is the theoretical framework efficient? (What was the yield of inconsistencies in the primary analysis? What was the yield of inconsistencies in the exploratory analysis of cells designated as "sometimes inconsistent" after consulting case narratives and other sources of information?)

RESULTS

Internal Consistency

Data on both SS and either NRNP *or* TNM were available for 39,675 cases of female breast cancer. Of these, 39,168 (99 percent) had data combinations designated as "consistent" and "infrequently inconsistent," 341 (1 percent) had data combinations designated as "sometimes inconsistent," and 166 (0.4 percent) had data combinations designated as "inconsistent."

Data on both SS and SDM were available for 30,819 cases of female breast cancer. Of these, 30,244 (98 percent) had data combinations designated as "consistent" and "infrequently inconsistent," 521 (2 percent) had data combinations designated as "sometimes inconsistent," and 54 (0.2 percent) had data combinations designated as "inconsistent."

Data on both SS and TNM were available for 13,773 cases of female breast cancer. Of these, 12,476 (90 percent) had data combinations designated as "consistent" and "infrequently inconsistent," 1267 (10 percent) had data combinations designated as "sometimes inconsistent," and 30 (0.2 percent) had data combinations designated as "inconsistent."

Validation of Inconsistencies

Of those data combinations designated as "inconsistent" in Tables 4, 5, and 6, 231 of 250 had sufficient supplementary information (from case narratives and other available sources of information) to evaluate the potential inconsistency thoroughly. Most (219 of 231, or 95 percent) proved to be truly inconsistent after consulting case narratives and other available sources of information (such as information from multiple case reports for the same tumor). A very few (12 of 231, or 5 percent) proved to be consistent (Table 7).

Of those data combinations designated as "sometimes inconsistent" for nodal involvement, as shown in Table 4, 168 of 341 had sufficient supplementary information (from case narratives and other available sources of information) to evaluate the potential inconsistency thoroughly. Forty percent (67 of 168) proved to be inconsistent after consulting case narratives and other available sources of information (such as information from multiple case reports for the same tumor). A majority (101 of 168, or 60 percent) proved to be consistent (Table 8). The yield of inconsistent cases from individual cells varied from 33 percent to 100 percent.

DISCUSSION

With the goal of developing a computerized check of the internal consistency of information on stage of disease at diagnosis, the Data Quality Indicator Subcommittee of NAACCR developed a theoretical framework with which to identify inconsistencies between staging schemes, and used information on stage of disease at

diagnosis from cases of breast cancer recently reported to four central cancer registries to test the framework for validity, effectiveness, and efficiency.

Is the theoretical framework valid?

The logic of the comparison appears to be correct. When cases with "inconsistent" data were examined to make corrections, the vast majority were found to have obvious coding errors for summary stage or the alternative source of information on staging. Similarly, when cases with "sometimes inconsistent" data 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. It should be applicable to other cancer sites with little or no modification.

Is the theoretical framework effective?

The approach is effective in finding coding errors with available information. The approach is accessible without complicated technology. It requires one-way cross-tabulation and the ability to identify cases within specific cells.

Is the theoretical framework efficient?

The approach identifies a very high proportion of errors (95 percent) among data in cells designated as "inconsistent." The yield of errors from cells designated as "sometimes inconsistent" is 40 percent overall, varying from 33 percent to 100 percent in the four cells examined. This yield may vary in other cells similarly designated (e.g., those in Tables 5 and 6 which were not examined) and should be smaller in cells designated as "infrequently inconsistent." Also, the yield may vary considerably from the results reported here when applied to data pertaining to other cancer sites.

Future Directions

Although the approach promises to be efficient, it should be applied to data from other registries for a broader evaluation of yield. It may be useful to pilot the approach as a computerized edit check in one of NAACCR's annual *Calls for Data*, and to evaluate the results closely at that time. Further analysis, such as the use of additional variables, may suggest modifications to increase the yield of inconsistent data from particular cells. Weighing the costs and benefits of the latter would also be facilitated by expanding the scope of the analysis to include data from other registries. Finally, it would be useful to conduct similar studies for other cancer types, e.g., colon or prostate.

CONCLUSIONS

The results presented herein may be used to develop a computerized edit check in NAACCR's annual *Call for Data* after further study and development. The method for identifying inconsistent staging data in case reports of female breast cancer has high face validity. Its use may help clarify the relationships among sources of staging data for registrars. Using data from the four participating central registries, previously undetected errors were found in SS. Although the efficiency of the queries indicated by certain cells was not evaluated fully in this pilot study, it seems reasonable to study them further, possibly in one round of the annual *Call for Data*, carefully evaluating them on the basis of those results. It would also be useful to conduct similar studies for other cancer types, e.g., colon or prostate.

Table 1. Evaluation of Summary Stage / NRNP or Summary Stage / TNM Combinations

Summary Stage	NRNP or TNM			
	No Positive Regional Nodes Identified	**Some Positive Regional Nodes Identified**	**Regional Nodes Not Examined *or* Cannot be Assessed Medically**	**No Information on Regional Nodes**
(0) In Situ	Consistent	Inconsistent	Consistent	Infrequently Inconsistent
(1) Localized	Consistent	Inconsistent	Consistent	Infrequently Inconsistent
(2) Regional by Direct Extension	Consistent	Inconsistent	Consistent	Consistent
(3) Regional by Positive Nodes	Inconsistent	Consistent	Sometimes Inconsistent	Sometimes Inconsistent
(4) Regional by Direct Extension & Nodes	Inconsistent	Consistent	Sometimes Inconsistent	Sometimes Inconsistent
(5) Regional, NOS	Infrequently Inconsistent	Infrequently Inconsistent	Infrequently Inconsistent	Infrequently Inconsistent
(7) Distant Metastases	Consistent	Consistent	Consistent	Infrequently Inconsistent
(9) Unknown	Consistent	Consistent	Consistent	Consistent

Table 2. Evaluation of Summary Stage / SDM Combinations

Summary Stage	SDM		
	No Distant Metastases Identified	**Specific Sites of Distant Metastases Identified**	**Other Sites of Distant Metastases Identified *or* No Information on Distant Metastases**
(0) In Situ	Consistent	Inconsistent	Sometimes Inconsistent
(1) Localized	Consistent	Inconsistent	Sometimes Inconsistent
(2) Regional by Direct Extension	Consistent	Inconsistent	Sometimes Inconsistent
(3) Regional by Positive Nodes	Consistent	Inconsistent	Sometimes Inconsistent
(4) Regional by Direct Extension & Nodes	Consistent	Inconsistent	Sometimes Inconsistent
(5) Regional, NOS	Consistent	Inconsistent	Sometimes Inconsistent
(7) Distant Metastases	Inconsistent	Consistent	Consistent
(9) Unknown	Consistent	Inconsistent	Sometimes Inconsistent

Table 3. Evaluation of Summary Stage / TNM Combinations

Summary Stage	TNM			
	No Metastases Identified	**Some Metastases Identified**	**Metastases Cannot be Assessed Medically**	**No Information on Metastases**
(0) In Situ	Consistent	Inconsistent	Sometimes Inconsistent	Sometimes Inconsistent
(1) Localized	Consistent	Inconsistent	Sometimes Inconsistent	Sometimes Inconsistent
(2) Regional by Direct Extension	Consistent	Inconsistent	Sometimes Inconsistent	Sometimes Inconsistent
(3) Regional by Positive Nodes	Consistent	Inconsistent	Sometimes Inconsistent	Sometimes Inconsistent
(4) Regional by Direct Extension & Nodes	Consistent	Inconsistent	Sometimes Inconsistent	Sometimes Inconsistent
(5) Regional, NOS	Consistent	Inconsistent	Sometimes Inconsistent	Sometimes Inconsistent
(7) Distant Metastases	Inconsistent	Consistent	Sometimes Inconsistent	Sometimes Inconsistent
(9) Unknown	Consistent	Inconsistent	Consistent	Consistent

Table 4. Summary Stage versus NRNP *or* TNM Information

Summary Stage	NRNP *or* TNM				
	No Positive Regional Nodes Identified	**Some Positive Regional Nodes Identified**	**Regional Nodes Not Examined *or* Cannot be Assessed Medically**	**No Information on Regional Nodes**	**Total**
(0) In Situ	2237	**5**	1988	219	4449
(1) Localized	17462	**102**	2676	658	20898
(2) Regional by Direct Extension	551	**24**	278	51	904
(3) Regional by Positive Nodes	**26**	7751	75	147	7999
(4) Regional by Direct Extension & Nodes	**9**	1271	94	25	1399
(5) Regional, NOS	11	103	175	8	297
(7) Distant Metastases	166	617	949	134	1866
(9) Unknown	450	80	889	444	1863
Total	20912	9953	7124	1686	39675

Note: Bolded cells are those which have been designated as “inconsistent.”

Table 5. Summary Stage versus SDM Information

Summary Stage	SDM			
	No Distant Metastases Identified	**Specific Sites of Distant Metastases Identified**	**Other Sites of Distant Metastases Identified *or* No Information on Distant Metastases**	**Total**
(0) In Situ	3505	**0**	2	3507
(1) Localized	16144	**3**	23	16170
(2) Regional by Direct Extension ***or*** (5) Regional, NOS	972	**2**	5	979
(3) Regional by Positive Nodes ***or*** (4) Regional by Direct Extension & Nodes	7297	**11**	10	7318
(7) Distant Metastases	**29**	1398	83	1510
(9) Unknown	845	**9**	481	1335
Total	28792	1423	604	30819

Note: Bolded cells are those which have been designated as "inconsistent."

Table 6. Summary Stage versus TNM Information

Summary Stage	TNM				
	No Metastases Identified	Some Metastases Identified	Metastases Cannot be Assessed Medically	No Information on Metastases	Total
(0) In Situ	1537	**0**	126	61	1724
(1) Localized	6720	**5**	282	443	7450
(2) Regional by Direct Extension	246	**1**	20	19	286
(3) Regional by Positive Nodes	2439	**3**	53	137	2632
(4) Regional by Direct Extension & Nodes	358	**1**	23	21	403
(5) Regional, NOS	222	**2**	14	3	241
(7) Distant Metastases	**17**	581	31	34	663
(9) Unknown	63	**1**	165	145	374
Total	11602	594	714	863	13773

Note: Bolded cells are those which have been designated as “inconsistent.”

Table 7. Validation of cells designated as "inconsistent"

Summary Stage	Other Source of Stage Information	Cases with Stage Information Designated as "Inconsistent," and Sufficient Additional Information to Validate	Number and % Found to be Inconsistent	
(0) In Situ	NRNP or TNM = Some regional nodes identified	5	5	100%
	SDM = Specific sites of distant metastases identified	0		NA
	TNM = Some metastases identified	0		NA
(1) Localized	NRNP or TNM = Some regional nodes identified	99	98	99%
	SDM = Specific sites of distant metastases identified	3	3	100%
	TNM = Some metastases identified	3	3	100%
(2) Regional by Direct Extension *or* (5) Regional, NOS	NRNP or TNM = Some regional nodes identified	23	23	100%
	SDM = Specific sites of distant metastases identified	2	1	50%
	TNM = Some metastases identified	1	1	100%
(3) Regional by Positive Nodes *or* (4) Regional by Direct Extension & Nodes	NRNP or TNM = No regional nodes identified	35	26	74%
	SDM = Specific sites of distant metastases identified	11	11	100%
	TNM = Some metastases identified	1	1	100%
(7) Distant Metastases	SDM = No distant metastases identified	29	28	97%
	TNM = No distant metastases identified	10	10	100%
(9) Unknown	SDM = Specific sites of distant metastases identified	9	9	100%
Total		231	219	95%

Note: Numbers may differ from Tables 4-6 because unresolvable cases (included in Tables 4-6) were omitted from Table 7.

Table 8. Evaluation of cells designated as "sometimes inconsistent"

Summary Stage	Other Source of Stage Information	Cases with Stage Information Designated as "Sometimes Inconsistent," and Sufficient Additional Information to Validate	Number and % Found to be Inconsistent ("Yield")	
(3) Regional by Positive Nodes	NRNP or TNM = Not examined or cannot be assessed medically	64	24	38%
	NRNP or TNM = No information on regional nodes	11	11	100%
(4) Regional by Direct Extension & Nodes	NRNP or TNM = Not examined or cannot be assessed medically	91	30	33%
	NRNP or TNM = No information on regional nodes	2	2	100%
Total		168	67	40%

Note: Numbers may differ from Table 4 because unresolvable cases (included in Table 4) were omitted from Table 8.