

# **CASE COMPLETENESS AND DATA QUALITY ASSESSMENTS IN CENTRAL CANCER REGISTRIES AND THEIR RELEVANCE TO CANCER CONTROL**

**Steven D. Roffers, PA, CTR**

*"The great society is a place where (people) are more concerned with the quality of their goods than the quantity of their goods."<sup>4</sup>*

## **Introduction**

Methods to assess the quality of cancer registry data have received more attention, perhaps due to increased utilization of cancer registry data and the desire to know more about the quality of the data. Cancer registries can enable the study of the distribution and etiology of cancer, but the usefulness of cancer registries is governed by the quality as well as the quantity (case completeness) of the data they contain. As Brooke states, "Every year an enormous quantity of medical statistics is compiled and published, and very little is known about the quality of the data on which these statistics are based."<sup>2</sup>

## **Population-based Central Cancer Registries**

The population-based cancer registry "... represents an attempt to collect detailed information about all new cases of (cancer) in a population of known size and composition. Its essential feature is the effort to account for all (cancer) diagnoses, whether in hospital or not, and its specific use is to determine the risk of (cancer) in a population."<sup>3</sup>

The American Association of Central Cancer Registries (AACCR), established in 1988, is an organization of member registries that provides population-based cancer registration for approximately 80 percent of the United States' population and 100 percent of Canada. However, because of varying levels of registry quality, there has never been an attempt to aggregate all the data into a single cancer surveillance report or to view the linked registries as a national reporting system. The Centers for Disease Control and Prevention (CDC) has provided AACCR with funds and technical assistance for: 1) data evaluation and publication; 2) registry quality assessment; and 3) registry support and development. A fourth component of the CDC support, cancer control implementation planning, ensures that local population-based cancer data and central registry staff are included as an integral part of the planning and implementation of breast and cervical cancer control efforts.<sup>4</sup>

## **Using Cancer Registry Data to Evaluate Cancer Control**

Cancer registries are a vital part of the national effort to cut the United States' cancer mortality rates in half by the year 2000. Registries can provide data to focus programs and monitor progress toward meeting the Year 2000 cancer mortality reduction goal. This will require aggressive attention to the opportunities for cancer prevention, early detection, treatment, and applied cancer control research.<sup>5</sup>

The AACCR Cancer Surveillance and Control Program (CSCP) aims to identify cancer registry data items optimal for cancer control, to establish the minimum level of data quality necessary to carry out cancer control evaluations, and to determine whether the attainment of the minimum level of quality is practical for central cancer registries.

The CSCP has completed work to define cancer registry data measures for use in evaluating control programs for cancers of the breast, cervix, lung, colon/rectum, and prostate.<sup>6</sup> The CSCP has determined the levels of data accuracy and reporting completeness necessary, under various assumptions of bias, to use these data measures.<sup>7</sup> It is yet to be definitively determined whether these levels of completeness and accuracy are reasonable expectations for operating central cancer registries, but based on the results of a CSCP Case Completeness and Data Quality Audit of the Illinois State Cancer Registry, it appears these levels can be achieved.<sup>8</sup> A high level of cancer registry quality is essential for drawing correct conclusions about cancer incidence, survival, treatment efficacy, patterns of care, and also cancer control.

### **Case Completeness Assessments and Relevance to Cancer Control**

The completeness of cancer registry data is defined as the proportion of all cancer cases in the target population which appear in the cancer registry database. All cases of cancer in a defined population *theoretically* appear in a population-based cancer registry. It is important to routinely measure the completeness of population-based cancer registries because the result of systematic bias in case reporting is the calculation of misleading and erroneous rates of cancer in the defined population. In statewide population-based cancer registries, it is important to verify that all facilities that are required to report cancer cases do indeed report all of their cases. Within each facility, there are various routine casefinding procedures, such as review of the disease indices, pathology reports, radiation therapy logs and others (see *Casefinding Activities at the Hospitals* below). Rigorous case completeness audits can be difficult, time consuming and costly, especially in the larger or more populated states. Nevertheless, since valid inferences can only be made from studies based on valid and unbiased data, the difficulty and expense must be accepted as a necessary cost of registry operations.<sup>3</sup>

When using cancer registry data for cancer control, case completeness is less important than the quality of the data for the cases that are included in the registry. This was demonstrated in the *Breast Cancer Mortality Model*, a CSCP modelling project undertaken to assess the completeness and accuracy required in a cancer registry database for measuring the impact of cancer control activities.<sup>6,7</sup> In the model, a variety of downstaging algorithms were employed with several resulting in a projection of at least a five percent reduction in mortality (a very reasonable assumption since community trials on screening mammography have accomplished 20 to 30 percent mortality reduction).<sup>7</sup> A cancer control evaluation strategy using *before* and *after* measurements would require comparing the stage distribution of cases collected in each of the two comparison periods. Sample size requirements to detect stage differences in the *before* and *after* groups were computed for one of the downstaging algorithms. The effect on sample size requirements of incomplete case ascertainment and by various rates of misclassification (data quality) was evaluated. While there is no effect if incompleteness is unbiased, with bias the sample size requirement is inflated. Sample size inflation resulting from different combinations of incompleteness and bias was computed. The effect of incompleteness with bias was not as dramatic as that seen for misclassification. Even with 10 percent incompleteness and a ratio of *in situ* cases to invasive cases three times that of the ascertained cases, inflation of the sample size was small (only 22 percent). Although incompleteness without bias has no effect on sample size requirements, it cannot simply be ignored, since it is never completely unbiased.

### **Data Quality Assessments and Relevance to Cancer Control**

Public and private health institutions recognize the need for the general public to be informed about cancer. Such information includes the probability of acquiring as well as surviving cancer. Population-based and hospital tumor registries must collect and disseminate this information.<sup>9</sup> More importantly, there must be an assessment of the quality of the data that are being collected and disseminated.

The intent of reabstracting and recoding studies is to standardize interpretation and abstracting of the medical record, to estimate rates of agreement, and to identify problems in data collection and interpretation. Thus, reabstracting is primarily an assessment and training tool.

The best method to assess the accuracy of cancer registry data is to compare them with the original medical records. The major advantage of reabstracting is that quality is evaluated for data already in the system, submitted under routine conditions. Reabstracted cancer cases become the *gold standard* with which to compare the previously abstracted cancer case, because more care and attention is directed to the reabstraction rather than to the routinely abstracted data. Reabstracting (and the subsequent recoding from the reabstracted data) attempts to measure the reproducibility of data collection and coding, but does not attempt to assess the accuracy of the underlying medical record.

When using cancer registry data for cancer control, a high level of data quality in the cancer registry is of primary concern. As described earlier, when using the cancer registry database to evaluate cancer control activities, data quality is more important than case completeness. High data quality (low rate of misclassification) is the primary concern for a cancer registry, with minimizing the potential bias from reporting incompleteness, a secondary concern.

### **Purpose of the AACCR Case Completeness and Data Quality Audits**

The primary purpose of the CSCP *Case Completeness and Data Quality Audits* is to assess the level of quality, and completeness of, cancer reporting in examples of statewide population-based cancer registries. The levels of completeness and accuracy are required to reasonably estimate important differences among potential target groups and changes in relevant sub-populations over appropriate lengths of time. State-funded population based central cancer registries<sup>a</sup> in states with CDC cooperative agreements for the conduct of cancer control programs in breast and cervix cancer are the first priority for these audits.<sup>4</sup> From 1992 to February 1994, these audits were performed in Illinois, Maine, Massachusetts, Colorado, and Nebraska. These five states agreed to be test sites to determine the concurrence between theoretically derived levels of data accuracy and completeness and actual field experience.

### **Stratification and Sampling Used in the AACCR Audits**

All facilities that are required by state law to report cancer cases were stratified according to the reported female breast, cervix, and prostate cancer caseloads for the year to be audited. Hospital caseloads were stratified by tertiles, each contributing one-third of the state's cancer cases.

In Nebraska, for example, there were 2,812 female breast, cervix, and prostate cancer cases (invasive and *in situ*) reported in 1991, from 84 reporting facilities. Facilities were stratified as follows:

- Low caseload: 1-91 breast (female), cervix, or prostate cancer cases  
(73 facilities, 921 cases)
- Medium caseload: 93-146 breast (female), cervix, or prostate cancer cases  
(8 facilities, 964 cases)
- High caseload: 209+ breast (female), cervix, or prostate cancer case (3 facilities, 927 cases)

---

<sup>a</sup> *State-funded population based cancer registries exist in nearly every one of the United States. The following states have NCI-funded SEER cancer registry programs and are therefore not eligible for the CSCP audits: Connecticut, Iowa, Utah, New Mexico, Hawaii, Washington (Seattle-Puget Sound), California (Greater Bay Area and Los Angeles), Michigan (Metropolitan Detroit), and Georgia (Metropolitan Atlanta).*

If a reliability interval of  $\pm 2.5$  percent is desired for a hypothesized completeness of 95 percent, then a sampling frame of 2,812 cases with the desired statistical reliability requires 274 cases (see Appendix for formulae).

$$D = \frac{(.025 * .025)}{(2 * 2)} = .00015625$$

$$n = \frac{(2,812 * .95 * .05)}{(2,811 * .00015625) + (.95 * .05)} = 274$$

To accomplish a sample of 274 cases given the above information, 92 cases were selected from each stratum, according to the following:

- For the **low caseload facilities**, we randomly selected a facility to be audited for 7 months, took seven-twelfths of its caseload and subtracted it from the 92 cases desired for the stratum. We continued until the sample size was met or exceeded.

<i>Low caseload: facilities</i>	<i>cases</i>
A	3
B	10
C	15
D	36
E	42
F	75

- For the **medium caseload facilities**, we randomly selected a facility to be audited for 4 months, took one-third of its caseload and subtracted it from the 92 cases desired for the stratum. We continued until the sample size was met or exceeded.

<i>Medium caseload: facilities</i>	<i>cases</i>
G	93
H	135
I	146

- For the **high caseload facilities**, we randomly selected a facility to be audited for 3 months, took one-quarter of its caseload and subtracted it from the 92 cases desired for the stratum. We continued until the sample size was met or exceeded.

<i>High caseload: facilities</i>	<i>cases</i>
J	453

Each stratum represents a third of the hospital-based cases and therefore missed cases may be summed to provide an estimate of the percent completeness of all hospital-based cases. Further, abstracting errors detected by the audit may also be projected to the entire annual caseload with a reliability of  $\pm 2.5$  percent.

## Performing the Actual On-Site Field Work

Using the Nebraska audit as an example, a two-person audit team visited the ten hospitals in the Nebraska random sample to perform the casefinding and reabstracting. Neither member of the audit team had abstracted or reported any of the cancer cases in the audit sample, and therefore were unbiased auditors.

The audit team reabstracted the medical records at the hospitals as described above. The audit team used portable computers with customized cancer registry software. The reabstracting portion of this audit was comprised of twelve data items, with emphasis on the value of the data item for cancer control evaluation or surveillance. For example, stage is more valuable and critical than is marital status at diagnosis, because cancer control efforts can result in shifts to earlier stages at diagnosis to decrease mortality. Marital status of the patient, on the other hand, is not an appropriate point for cancer control intervention to reduce mortality.

For the purposes of these audits, the twelve most important and critical data items for cancer control applications were: *site* (first three digits of the four digit International Classification of Diseases for Oncology (ICD-O) topography code); *stage at diagnosis* (SEER Summary Stage: in situ, localized, regional (either direct extension, metastases to regional lymph nodes, both direct extension and metastases to regional lymph nodes, or regional, not otherwise specified (NOS)), or distant; *diagnosis year*; *diagnosis date* (month and day, or month alone if the registry does not collect day); *race*; *date of birth*; *state of residence at diagnosis*; *subsite of primary* (fourth digit of the four digit ICD-O topography code); *histology* (first four digits of the six digit ICD-O morphology code); *grade* (cell differentiation — the sixth digit of the six digit ICD-O morphology code); *sequence number*; and *laterality* (applicable only to breast cancer as the cervix uteri and the prostate are not paired organs).

In addition to the reabstracting, the audit team performed casefinding using the following sources, if available in the hospital:

- Medical record disease indices;
- All pathology reports (including bone marrows, autopsies, and other specialized pathology reports);
- Surgical log books and *same day surgery* log or appointment books;
- SNOMED listing of final histologic diagnoses (if available);
- Cytology reports (if filed separately from the pathology reports);
- Outpatient hematology/oncology clinic log books and other special clinics that may diagnosis and/or treat a breast, cervix or prostate cancer patient (i.e., Breast Health Clinics, etc);
- Radiation therapy clinics [patient visit log book, radiotherapy treatment summary sheets, clinic *shadow* charts, radioactive implant (Iridium-192, Iodine-125 and Cesium for example) log books, etc.);
- Nuclear medicine clinic (if the implant log books are kept here rather than in the Radiation Therapy Department);
- Any other sources in the hospital where a breast, cervix, or prostate cancer case may have been documented as being diagnosed or treated at that facility.

When a case was found in one or more of the aforementioned sources and did not match with the reported cases, a *dummy case accession* was created and stored on the portable computers for reconciliation later by the audit team and the central registry staff.

## Analysis of the Reabstracted Data and Casefinding

After the actual field work was completed, the reabstracted data were analyzed. Any discrepancies between the reabstracted data and the data originally submitted were sent to the central registry staff for reconciliation. In addition, the unmatched cases were compiled and sent to the central registry staff for reconciliation.

It is through the reconciliation process that identification of the patterns of inaccuracy and missed cases occurs. Identifying the problems, errors, and inadequacies is the first step in addressing how best to rectify these deficiencies. It also directs the training which is needed in order to reduce and avoid these reported problems in the future.

Detailed, intensive, statistically valid, and expensive case completeness and data quality audits must be accepted as a necessary cost of registry operations.

## Results

The final products of these case completeness and data quality audits are:

- 1) an estimate of the percent completeness of all female breast, cervix, and prostate cancer cases from reporting facilities in the state being audited for the audit year; and
- 2) an estimate of the accuracy of those data items used in the measures recommended for cancer control by the AACCR.<sup>6</sup>

Final results and reports have been completed for case completeness and data quality audits performed in Illinois and Maine. Copies of these reports can be obtained by contacting central registry staff in these states. Results and reports from the audits performed in Massachusetts, Colorado, and Nebraska will be completed and available by August 1994.

## Future Activities and Direction

Surveillance and evaluation are crucial for effective cancer control efforts. Cancer registries need to be incorporated as a tool for cancer control and utilized accordingly. There is an exciting future for cancer registries with the enactment of Public Law 102-515 (the *National Cancer Registries Act* — allocating \$16.8 million for the establishment and support of cancer registries in every state in the United States) administered by the CDC. Linking the activities of the Breast and Cervical Cancer Control Program of the CDC with the National Cancer Registry Act activities is not only appropriate, but also logical.

These case completeness and data quality audits can serve as a baseline measurement prior to implementation of cancer control interventions. Post-intervention audits would serve to demonstrate the effectiveness of the intervention. Future audits in these same states, after the implementation of control programs for cancers of the breast, cervix and prostate in these states, can demonstrate the effectiveness of the control programs by assessing pre-intervention versus post-intervention stage distribution for each state.

The CSCP audits can contribute to the goals of the national program of cancer registries.<sup>11</sup> These are 1) timely feedback for evaluating progress toward achieving Healthy People 2000 cancer control objectives, among others;<sup>12</sup> 2) data to identify cancer incidence variation for specific race and ethnic groups and among counties, states, and regions; 3) guidance for health resource allocation; 4) data to evaluate state cancer control activities; and 5) information to improve planning for future health care needs.

## References

1. United States Statesman Speech, delivered on May 22, 1964 by Lyndon B. Johnson.
2. Brooke, EM. *The Current and Future Use of Registries in Health Information Systems*. Geneva: World Health Organization, 1974.

3. Goldberg J, Gelfand HM, and Levy PS. Registry Evaluation Methods: A Review and Case Study. *Epidemiological Reviews* 2:210-220, 1980.
4. *Implementation of the Breast and Cervical Cancer Mortality Prevention Act: A Progress Report*; U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, 1991.
5. Greenwald, P, Sondik, EJ, and Young, JL. Emerging Roles for Cancer Registries in Cancer Control. *The Yale Journal of Biology and Medicine* 59:561-566, 1986.
6. Roffers SD and Austin DF. Analysis of AACCR Data Items and Interim Outcome Measures. *AACCR-CDC Technical Report #1 (and Addendum)* April 1992.
7. Austin DF, Fu CX, Roffers SD, Shields JM, and Aubert RE. Breast Cancer Mortality Model. *AACCR-CDC Technical Report #2*, February 1993.
8. Roffers SD, Snodgrass J, Howe HL, and Austin DF. American Association of Central Cancer Registries (AACCR) - Illinois State Cancer Registry Audit of Female Breast Cancer Cases Diagnosed in 1990 in Illinois: An Assessment of Completeness of Reporting and Quality of the Reported Data. *AACCR-CDC Analytical Report*, October 1993.
9. Bender AP and Olsen GW. A survey of the American College of Surgeons Hospital based Tumor Registries. *Journal of the American Medical Record Association* 55:20-23, 1984.
10. Hilsenbeck SG, Glaefke GS, Feigel P, Lane WW, Golenzer H, Ames C, and Dickson C. *Quality Control for Cancer Registries*. U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, May 1985.
11. P.L. 102-515, Cancer Registries Amendment Act, October 24, 1992.
12. Healthy People 2000: National Health Promotion and Disease Prevention Objectives. USDHHS, PHS No. 91-50212, 1991.

## APPENDIX

### Formulae Used to Estimate Completeness

If a case is found in the state's file from any source, and was diagnosed in the calendar year of study, it was classified as a "match." If a case was abstracted at the hospital but was not on the state file, then it was counted as "non-match." If a reportable neoplasm was found in one of the casefinding sources but was not on the state file, then it was counted as a "non-match". Resolved cases, those diagnosed prior to the year of study or with no cancer, are deleted from the counts of non-matches.

The equations for the stratified estimate are:

$N$  = number of expected cases for the diagnosis year for the state

$N_i$  = number of expected cases in stratum  $i$

$n_i$  = sample size from stratum  $i$

$p_i$  = number of matched cases in stratum  $i$   
(number of matched + unmatched cases) in stratum  $i$

$q_i = 1 - p_i$

$$p_{st} = \frac{1}{N} \sum_{i=1}^3 N_i p_i$$

**Variance:**

$$V(p_{st}) = \frac{1}{N^2} \sum_{i=1}^3 (N_i^2) \left( \frac{N_i - n_i}{N_i} \right) \left( \frac{p_i * q_i}{n_i - 1} \right)$$

**Standard deviation:**

$$S(p_{st}) = \sqrt{V(p_{st})}$$

**The 95% confidence interval for  $p_{st}$  is:**

$$CI. = p_{st} \pm (1.96 * p_{st})$$