
Health Services & Outcomes Research Methodology 2: 67–76, 2001

Using Hospital Tumor Registries to Identify Research Subjects

MICHAEL SCHOENBAUM*, MEREDITH L. KILGORE, BEVERLY A. WEIDMER, SANDRA H. BERRY, JOSE J. ESCARCE, DANA P. GOLDMAN
RAND, 1700 Main Street, Santa Monica, CA 90407

JOY H. LEWIS
University of California, 405 Hilgard Ave., Los Angeles, CA 90095

NIKHIL WAGLE
Harvard University, 25 Shattuck Street, Boston, MA 02115

Received June 20, 2000; revised February 20, 2001; accepted February 20, 2001

Abstract. Hospital tumor registries may permit efficient identification of relatively large numbers of patients for clinical and health services research. This study investigates the feasibility and cost of using hospital tumor registry data for identifying cancer patients with particular clinical characteristics, corresponding to the protocol entry criteria of four randomly sampled Phase III clinical treatment trials for cancer. We screened tumor registry data on 717 patients diagnosed with cancer in 1997 and 1998 who had been identified and abstracted by the registry of a large academic medical center by April, 1999; and we abstracted the medical records of the 122 patients who passed the registry screen. For each clinical profile, the registry screen eliminated a substantial fraction of patients who did not meet the relevant criteria. Of the patients identified from the tumor registry as potential matches, 41% matched the relevant clinical profile based on detailed medical records review. The cost and effort of the registry-based method were substantially lower than would have been necessary if we had reviewed medical records without the registry screen, suggesting that tumor registry data can be a relatively efficient tool for identifying research subjects.

Keywords: cancer, tumor registry, research methods, research costs

1. Introduction

Clinical researchers often face the task of identifying patients with specific characteristics in order to study treatment costs, patient satisfaction with care or quality of care, or to find patients for clinical trials. Small numbers of patients at individual institutions can be identified in a variety of ways, including conversations with providers or by reviewing medical records. However, such methods are likely to be impractical for identifying relatively larger numbers of patients with specific characteristics and, in the case of *ad hoc* methods such as identification by providers, may produce biased samples of patients.

*Corresponding author: Michael Schoenbaum, RAND, 1200 South Hayes St., Arlington, VA 22202.
Fax: 703-413-8111, E-mail: mikels@rand.org

Searching automated data systems may be a relatively efficient alternative (Cooper et al., 1999; Doebbeling et al., 1999). Although some hospitals have implemented medical information systems that could support such a search, such systems are not yet universal, may not be accessible to researchers, and are not generally standardized across institutions, a particular limitation in the case of multi-institution studies. In the particular case of cancer patients, however, most hospitals report clinical information to a local tumor registry on a timely basis. Furthermore, many tumor registries use a common standard: the Registry Operations and Data Standard established by the Commission on Cancer of the American College of Surgeons (American College of Surgeons, 1998).

Previous studies have assessed the quality of tumor registry data, including the completeness of case-finding and vital status (Harvel et al., 1996; Hawkins and Swerdlow, 1992; Lapham and Waugh, 1992; Mukherjee et al., 1991; Zippin et al., 1995) and the sensitivity and specificity of information on the first course of treatment (Bickell and Chassin, 2000). However, many of these studies use regional or national rather than hospital tumor registry data. This study tests the feasibility and costs of using tumor registry data at a large academic medical center to help identify newly diagnosed cancer patients eligible for Phase III treatment trials in a way that minimizes costs and staff burden. We then compare this registry-based strategy with direct medical records abstraction. We know of no previous work that has systematically evaluated the value of using tumor registry data for case identification.

2. Methods

2.1 Tumor Registry

This research was conducted at a large academic medical center with a tumor registry accredited by the American College of Surgeons (ACOS). The ACOS standard requires a hospital tumor registry to identify and abstract all newly diagnosed ("analytic") cancer patients presenting for consultation or treatment at the hospital; the tumor registry is required to complete the abstraction of medical records for patients presenting in a calendar year by August of the following year (American College of Surgeons, 1998). In addition, the tumor registry is required to assess the vital status of each newly diagnosed patient for which they have abstracted information, at 18 month intervals until death. The ACOS standard does not require abstraction of patients presenting at the hospital with progressed or relapsed cancer ("non-analytic" patients), and we do not consider such patients here.

2.2 Clinical Criteria

As part of a larger study of the incremental difference in treatment costs between patients participating in NCI-sponsored clinical trials and patients receiving care as usual, the Cost

of Cancer Treatment Study (CCTS; Goldman et al., 2000, 2001), we randomly sampled four ongoing Phase III treatment trials from the National Cancer Institute's Physician Data Query (PDQ) Clinical Trials Database. The sample included two trials for colon cancer, one trial for lung cancer and one for breast cancer. Table 1 presents summary information on each sampled trial.

2.3 Participants

In the spring of 1999, the hospital tumor registrar provided complete registry data on all newly diagnosed patients who presented with cancer of the lung, colon, or breast (i.e., corresponding to our four sampled trials) during 1997 and part of 1998. The total sample size was 717 patients: colon, 84; lung, 179; breast, 454.

2.4 Analysis Method

For each sampled trial, we identified which protocol entry criteria could be assessed in the tumor registry data. The screening was based only on data fields required by ACOS, which should be available in any ACOS-accredited tumor registry. Specifically, we screened patients according to age and sex; cancer site; cancer stage at initial evaluation; histology; and the number of lymph nodes that were sampled and that were positive. Prior treatment was listed in the cancer registry only if it had occurred at that hospital. Specific screening criteria corresponded to the respective protocol entry criteria of the four sampled trials. For example, for trials that precluded prior chemotherapy, we excluded patients listed in the registry as having received this. However, for trials that required prior treatment, we opted not to exclude patients when the registry did not have a record of prior treatment, because patients could have received it elsewhere.

Based on this screen, we selected all patients who appeared to meet the eligibility criteria of any of our sampled trials based on tumor registry data. A trained research coordinator in the hospital's oncology research program (who was also a certified Tumor Registrar) then abstracted the medical record for each patient, using an abstraction form listing the detailed protocol entry criteria for the trial to which the patient was being matched. For each criterion, the abstractor identified whether the patient met the criterion and, if not, in what way the criterion was not met.

Based on these abstraction forms, a physician supervisor on the research team assessed whether each patient met the full list of protocol entry criteria of the trial to which he/she was being matched. We classified areas of mismatch into six domains: tumor stage, lab values, comorbidity, concurrent therapy, prior therapy, and performance status. We note that the medical records for a small number of patients did not contain sufficient information to assess whether the patients met the protocol entry criteria of the trial to which they were being matched; these patients were excluded from subsequent analyses.

Table 1. Clinical characteristics to be matched using tumor registry data

Trial Cancer Site	NCCTG-934653 Colon	E-EB-193 Breast	E-1594 Lung	SWOG-9415 Colon
Sex and age	Not specified	Female; 65+ or postmenopausal	Not specified	Not specified
Stage	I, II or III; no acutely obstructed or perforated colon cancer requiring urgent surgery	T1-3, N1-2, M0; no clinical or pathological T4 disease; no clinical N2 disease	IV or IIIb with pleural effusion	C1/C2 (T1-4a, N1-3, M0); poor prognosis stage B2 (T4a, N0, M0) with evidence of nearly total bowel obstruction or perforation
Histology	Adenocarcinoma	Adenocarcinoma; no adenoid cystic, squamous, or sarcomatous histology	Non-small cell, brochogenic; squamous cell, adenocarcinoma, large cell anaplstic, bronchoalveolar, or non-small cell not otherwise specified	Adenocarcinoma
Distant metastases	No advanced local disease that makes laparoscopic resection impossible	None	Not specified	None
Lymph nodes sampled/positive	Not specified	At least 6 sampled (minimum waived for patients 70+ with at least 1 positive node)	Not specified	Not specified
Prior treatment	Not specified	No prior chemotherapy or hormonal therapy for breast cancer except up to 1 month of tamoxifen	No prior chemotherapy or biological response modifier therapy; recovered from any prior radiotherapy	No prior chemotherapy or radiotherapy; complete en bloc resection required

3. Results

3.1 Registry-based Screen

Table 2 presents the results of the registry-based screen. Of the 717 newly diagnosed patients, 125 passed the registry screen for one or more of our four clinical profiles. Five patients passed the screen for both sampled colon cancer trials, for a total of 130 potential patient-trial matches. Of these, medical records could not be located in eight cases. The total number of medical records reviewed was thus 122.

3.2 Medical Record Review

Table 3 lists the results of the medical record review. Of the 122 records, six patients matched some of the criteria but lacked sufficient information to assess eligibility definitively. Of the remaining 116 patients, 47 matched all the relevant criteria. Across the four trials, the positive predictive value of the registry-based screen was thus 41%. The three trials with the largest number of potential matches all had match rates close to the mean; the rate was twice as high for colon cancer trial SWOG-9415, but the number of patients assessed for that clinical profile was small.

Table 3 also summarizes the domains of information where the patients did not match. In the majority of such cases, patients did not match on disease characteristics that could not be assessed using tumor registry data, for instance because they presented with complications that were precluded by the respective protocol entry criteria or had been diagnosed using methods other than what was required. A substantial number of patients also did not match the comorbidity exclusion criteria.

Table 2. Screen of tumor registry data to identify potential matches

Trial Cancer Site	NCCTG-934653 Colon*	E-EB-193 Breast	E-1594 Lung	SWOG-9415 Colon*	Total
No. with type of cancer in registry	84	454	179	84	717
Ineligible domains** – No. (%)					
Sex or age	n/a	190 (42%)	n/a	n/a	
Stage	36 (43%)	383 (84%)	114 (64%)	74 (55%)	
Histology	4 (5%)	3 (1%)	34 (19%)	4 (10%)	
Prior treatment	n/a	n/a	6 (3%)	0	
Distant Metastasis	n/a	n/a	n/a	33 (39%)	
Lymph nodes	n/a	131 (29%)	n/a	n/a	
No. who passed screen	44	29	50	7	136
% who passed screen	52%	6%	28%	8%	18%
No. charts reviewed	41	27	48	6	122

Note: * Five patients screened eligible for both colon cancer trials.
** Not mutually exclusive; n/a means domain not applicable for particular trial or could not be assessed.

Table 3. Positive predictive value of tumor registry screen

Trial Cancer Site	NCCTG-934653 Colon	E-EB-193 Breast	E-1594 Lung	SWOG-9415 Colon	Total
No. based on registry screen	44	29	50	7	130
No. whose medical record could be reviewed[a]	41	27	48	6	122
No. whose eligiblity could be assessed[b]	40	26	44	6	116
No. who match clinical criteria (%)[c]	14 (35%)	9 (35%)	19 (43%)	5 (83%)	47 (41%)
No. who don't match (%)	26 (65%)	17 (65%)	25 (57%)	1 (17%)	69 (59%)
Ineligible Domains[d] – No. (%)					
Disease characteristics	19 (73%)	9 (53%)	15 (60%)	0	43 (62%)
Lab values	1 (4%)	1 (6%)	3 (12%)	1 (100%)	6 (9%)
Comorbidity	12 (46%)	8 (47%)	2 (8%)	0	22 (32%)
Concurrent therapy	0	10 (59%)	0	0	10 (14%)
Prior therapy	1 (4%)	0	1 (4%)	0	2 (3%)
Performance status	0	0	7 (28%)	0	7 (10%)

Note: [a] Inability to review due to inability to access medical record when it was requested by local staff.
[b] Inability to assess was due to incomplete information in medical record.
[c] Among patients whose eligibility could be assessed.
[d] Ineligible domains are not mutually exclusive.

3.3 Level of Effort

The left side of Table 4 lists the level of effort necessary to screen the 717 patients and identify the 47 matches. We report effort separately for different types of employees due to differences in wages. The registry-based strategy required approximately 87 minutes of researcher time per matched patient; 26 minutes of registrar time, including the hospital registrar as well as the registrar consultant hired by RAND; 42 minutes of clerical time to recall records; and 83 minutes of nurse time. For comparison, the right side of Table 4 lists estimates of the effort that would have been required had we elected not to use the registry-based screen and had instead reviewed all 717 medical records. Given the 717 newly diagnosed patients presenting with the relevant cancer types during our study window, we estimate that this approach would have required 98 minutes of researcher time, 229 minutes of clerical time, 251 minutes of nurse time, and no registrar time per final matched patient.

If we assume hourly wages of $75 for researchers, $25 for registrars, $10 for clerical staff, and $30 for nurses, the registry-based strategy would cost $168 per confirmed match – compared with $286 per match for the records-based approach. In practice, of course, the costs of different strategies will vary with wages. This analysis excludes costs that would be common to both strategies, such as developing abstraction forms and obtaining human subjects approval for the research.

4. Discussion

For many types of studies, researchers want to identify patients with particular clinical characteristics. Methods for doing so must be logistically feasible and affordable and must

Table 4. Level of effort for alternative strategies to identify patients "hours"[a]

Activity	Registry-based Strategy				Systematic Record Review[b]			
	Fixed Effort	Effort per Patient	Number of Patients	Total Effort	Fixed Effort	Effort per Patient	Number of Patients[c]	Total Effort
Researcher								
Develop data request to tumor registry	8	0	n/a	8	0	0	n/a	0
Develop & conduct registry-based screen	32	0	n/a	32	n/a	n/a	n/a	0
Supervise initial medical records screen	n/a	n/a	n/a	n/a	4	0.1	521	52
Supervise medical records abstraction	4	0.2	122	28	1	0.2	122	24
Researcher hours				*68*				*77*
Registrar/Registry consultant								
Create registry data extract	12	0	n/a	12	n/a	n/a	n/a	0
Advise development of registry-based screen	8	0	n/a	8	n/a	n/a	n/a	0
Registrar hours				*20*				*0*
Clerical								
Recall medical records	n/a	0.25	130	33	n/a	0.25	717	179
Clerical hours				*33*				*179*
Nurse								
Conduct initial medical records screen	n/a	n/a	n/a	0	4	0.25	521	134
Abstract medical records for screened patients	4	0.5	122	61	1	0.50	122	62
Nurse hours				*65*				*196*

Note: [a] Both strategies assume the same outcome of 47 confirmed matches, as in Tables 2 and 3.
[b] Initial medical records screen, followed by detailed abstraction for patients who pass screen.
[c] Assumes same rate of unavailable medical records as reported in Table 3.

meet the scientific requirements of the particular study for which patients are sought. In many cases, automated data offer several distinct advantages for meeting these goals: they are generally designed to cover defined patient populations (so that, even if coverage is not complete, it may be relatively easy to identify what classes of patients are missing); and the marginal cost of screening additional patients is low. Hospital tumor registry data are particularly appealing for identifying cancer patients because the data are reported in timely fashion; tumor registries are designed to abstract salient clinical characteristics; most hospitals report data to a local tumor registry; and, at least for hospitals with ACOS-accredited registries, the data components covered by the registry are standard.

We can evaluate the empirical utility of a registry-based screen in two ways. The first is the extent to which the registry-based screen focused the search, an indication of the value added of such a screen versus directly reviewing medical records. Our results in Table 2 indicate that, for each set of clinical criteria studied here, the tumor registry screen eliminated a substantial number of newly diagnosed patients who did not meet the relevant criteria.

A second way to assess the utility of the registry-based approach is the positive predictive value of the registry screen, which gives the fraction of patients passing the registry screen who were found to meet the criteria to which they were being matched. The mean fraction of patients identified via the registry who matched based on the medical record review was 39% (41% excluding patients whose match could not be assessed due to incomplete medical records). For one set of clinical criteria, a high fraction of screened patients failed to match due to details of their first course of treatment, which tumor registries do abstract; however, evidence suggests that tumor registries are not very sensitive in capturing these data (Bickell and Chassin, 2000). Most commonly, however, screened patients failed to match on criteria that most tumor registries do not abstract, such as disease characteristics other than stage, comorbidities, and lab values.

Finally, using tumor registry data to target medical records abstraction appears to be highly cost-effective, relative to abstracting medical records without such a screening mechanism. The results in Table 4 indicate that the registry-based approach requires many times fewer person-hours, with correspondingly lower costs. Furthermore, our effort estimates assume a very favorable environment for screening medical records, including that the hospital's record system permits researchers to recall charts by cancer site and year of diagnosis. If these do not apply, the approach based on medical records would require even more effort. This difference between the two strategies will be smaller if fewer patients can be ruled out by the registry screen, as when the clinical profiles are broader or based more on characteristics that can not be assessed in tumor registries.

Our results suggest that hospital tumor registries can be a relatively efficient method for identifying newly diagnosed cancer patients meeting certain clinical criteria – but this finding is of little practical value unless researchers are permitted to use registry data. The laws and regulations for using tumor registries (and medical records) for research vary by state, and a formal survey of these procedures is beyond the scope of this study. Our experience, however, suggests that access to tumor registry data for research purposes is no more restricted than access to data from medical records. The current study was approved by the Institutional Review Boards (IRBs) of RAND and the participating medical center;

neither IRB required us to obtain explicit patient approval to use tumor registry or medical records data. For CCTS, a national study in which we are enrolling 1500 cancer patients who meet specific clinical criteria, we have asked staff at approximately 150 participating institutions to use tumor registry and medical record data to identify potential study subjects (Goldman et al., 2000, 2001). In all cases, local staff contact the potential subjects, describing the study and requesting patient consent for RAND to contact them; RAND only received identifying information on consenting patients. Only two institutions declined to allow us to use tumor registry data to identify potential CCTS subjects.

There are a number of potential caveats to our research. For instance, for each of the four clinical profiles studied here, the registry-based screen identified a high fraction of newly diagnosed patients who were unlikely to match the respective profiles. However, the positive predictive value of the registry-based screen varied by clinical profile. Results using tumor registries at other hospitals may also vary, depending on factors such as the nature and quality of the abstracted data.

In addition, we have argued that one of the potential benefits of using a registry-based approach for identifying patients is the representativeness of the resulting sample. However, we did not attempt to verify the registry's coverage, and it is possible that some relevant patients might not be identified and abstracted by the registry, or might be abstracted with greater delay than that specified by ACOS. Further, although it seems unlikely, it is possible that we excluded patients in the registry screen who actually would have met the relevant clinical criteria. Such "false negatives" could arise if, for instance, the medical records were updated after registry abstraction took place, or if the medical record were abstracted with error (for reference, we did not identify any patients passing the registry screen whose medical record contradicted the registry data). It is also possible that some information in the medical record is incorrect, so that a patient may pass the registry screen and the medical record review but not in fact meet the relevant clinical criteria.

Finally, the clinical profiles to which we sought to match patients in this study could be considered narrowly defined, corresponding to the scientific requirements of the randomized controlled trials for which they were specified. However, they could be considered broad for other purposes, in particular for study designs that seek to match individual cases and controls. While the tumor registry may still be a valuable starting point in such cases – indeed, especially valuable in such cases, given their screening demands – the binding constraint of such a design may be the lack of availability of matching patients.

Our results from one hospital suggest that tumor registry data may be a very valuable tool for health services, epidemiological, and clinical research. However, the generalizability of our findings should be the subject of future research.

Acknowledgments

Principal funding for this study was provided by the National Cancer Institute, with additional funding from the Office of the Director, National Institutes of Health; and by the National Science Foundation as part of its support for the White House's Office of Science

and Technology Policy. We are grateful to Amalia Rincon, Nancee Relles and Eunice Little for their invaluable help in collecting the data used here; Kathryn Davis for coordinating data collection; Jane Weeks, Richard Kaplan and Robert Figlin for their help in reviewing protocol entry criteria; Louise Schuman for guidance in using tumor registry data; and an anonymous reviewer for helpful comments. The opinions expressed here are those of the authors and do not necessarily reflect those of the funders.

REFERENCES

American College of Surgeons, *Standards of the commission on cancer, Volume II: Registry operations and data standards*, American College of Surgeons, Chicago, 1998.

N. A. Bickell and M. R. Chassin. "Determining the quality of breast cancer care: Do tumor registries measure up?" *Ann Intern Med*, 132, pp. 705–710, 2000.

G. S Cooper, Z. Yuan, K. C. Stange, et al. "Agreement of Medicare claims and tumor registry data for assessment of cancer-related treatment," *Medical Care*, 38(4), pp. 411–421, 1999.

B. N. Doebbeling, D. K. Wyant, K. D. McCoy, et al. "Linked insurance-tumor registry database for health services research," *Medical Care*, 37(1), pp. 1105–1115, 1999.

D. P. Goldman, J. Adams, S. Berry, K. Davis, J. Escarce, M. Kilgore, Joy. Lewis, M. Schoenbaum, M. Schonlau, N. Wagle and B. Weidmer. *The cost of cancer treatment study design and methods*, RAND MR-1169, Santa Monica, RAND, 2000.

D. P. Goldman, M. Schoenbaum, A. L. Potosky, et al. "Measuring the incremental cost of clinical cancer research: The cost of cancer treatment study," *Journal of Clinical Oncology*, 19(1), pp. 105–110, 2001.

S. Harvel, S. Tretli and F. Langmark. "Quality of prostate cancer data in the cancer registry of Norway," *European Journal of Cancer*, 32A, pp. 104–110, 1996.

M. M. Hawkins and A. J. Swerdlow. "Completeness of cancer and death follow-up obtained through the National Health Service Central Register for England and Wales," *British Journal of Cancer*, 66, pp. 408–413, 1992.

R. Lapham and N. R. Waugh. "An audit of the quality of cancer registration data," *British Journal of Cancer*, 66, pp. 552–554, 1992.

A. K. Mukherjee, I. Leck, F. A. Langley and C. Ashcroft. "The completeness and accuracy of health authority and cancer registry records according to a study of ovarian neoplasms," *Public Health*, 105, pp. 69–78, 1991.

C. Zippin, D. Lum and B. R. Hankey. "Completeness of hospital cancer case reporting from the SEER program of the National Cancer Institute," *Cancer*, 76, pp. 2353–2350, 1995.