

COMMENTARY

Potential for Selection Bias with Tumor Tissue Retrieval in Molecular Epidemiology Studies

JANE A. HOPPIN, ScD, PAIGE E. TOLBERT, PhD, JACK A. TAYLOR, MD, PhD,
JANE C. SCHROEDER, DVM, PhD, AND ELIZABETH A. HOLLY, PhD

Molecular epidemiological studies of cancer generally require tumor tissue to evaluate somatic genetic alterations. Frequently this requires retrieval of fixed tissue blocks from hospital pathology archives. The availability of this material may be associated with disease severity, diagnostic practices, hospitals, or risk factors for disease. Tumor material is not available when the diagnosis is made clinically without histological confirmation. These characteristics create difficulties in defining the study base population. Incomplete access to tumor tissue has implications for description of the natural history of disease, estimates of the prevalence of mutation in the population, and evaluation of environmental exposures and critical target gene mutations. Differential diagnostic practices by age groups or across hospitals may create a biased population with respect to potential risk factors. However, this will not bias case-case comparisons unless the mutation of interest is associated both with the exposure of interest and the presence of a tumor block. When subjects with less severe disease are more likely to have biopsies, information regarding the natural history of the disease will be obscured. Investigation of the interaction of environmental agents and critical target gene mutations may be limited if, for example, an environmental agent is associated with a more aggressive form of the disease. Using an ongoing pancreatic cancer case-control study as an example, we discuss the potential for bias associated with differential availability of tumor blocks including consideration of tumor, patient, and hospital characteristics. Due to incomplete retrieval of tissue, the determinants of selection should be described in all studies using tumor tissue, and the implications for generalizability, power, and interpretation of findings in population-based studies should be considered. *Ann Epidemiol* 2002;12:1–6. © 2001 Elsevier Science Inc. All rights reserved.

KEY WORDS: Molecular Epidemiology, Selection Bias, Tumor Tissue.

BACKGROUND

In molecular epidemiological studies that rely on tumor tissue to assess genetic changes, the availability of tumor tissue may be a function of the disease, hospital diagnostic practices, pathology laboratory preservation and storage protocols, and the study population. Thus, the availability of tumor tissue, primarily formalin-fixed paraffin-embedded tissue blocks (referred to throughout as tissue blocks), may

influence the results of studies designed to describe the prevalence of genetic alterations in tumors, the natural history of disease in a population, and the interactions between exposure and mutations, since the study base may be difficult to characterize. Here, we describe factors associated with incomplete retrieval of tumor blocks, explore the implications for epidemiological studies, and suggest strategies to assess potential differences in sample retrieval. We limit our discussion to issues of specimen retrieval but these issues are also relevant to tumor tissue quality and the ability to obtain adequate DNA for analysis.

Molecular epidemiological studies of critical target gene mutation usually require tumor tissue to assess genetic changes (1, 2). Tumor tissue samples may be available as frozen tissue, fixed tissue, fine needle aspirates, or pathology slides; many molecular epidemiology analyses rely on tissue blocks from biopsies or resections to assess genetic characteristics, as fixed tumor tissue is frequently retained by pa-

From the National Institute of Environmental Health Sciences, Epidemiology Branch, Research Triangle Park, NC (J.A.H., J.A.T., J.C.S.); Rollins School of Public Health, Emory University, Atlanta, GA (J.A.H., P.E.T.); and Department of Epidemiology and Biostatistics, University of California, San Francisco, CA (E.A.H.).

Address correspondences to: Jane A. Hoppin, Sc.D., NIEHS, Epidemiology Branch, MD A3-05, PO Box 12233, Research Triangle Park, NC 27709-2233.

Received January 26, 2001; revised May 4, 2001; accepted May 9, 2001.

Selected Abbreviations and Acronyms

NCCC = Northern California Cancer Center
 sd = standard deviation
 SEER = Surveillance, Epidemiology and End Results

thology laboratories and is relatively easy to retrieve and ship for subsequent population-based molecular epidemiological investigations (1). However, tumor tissue often is not available for all subjects in a study. The inability to obtain tumor tissue falls into three broad categories: 1) no tissue sample was collected (e.g., the patient was diagnosed clinically without histological confirmation); 2) tissue sample was collected but is unavailable (e.g., the tissue was lost or destroyed or the hospital will not release the sample); and 3) the sample preparation is inappropriate for the molecular analysis (e.g., the sample is too small, degraded to assay, lacks normal DNA, or sample preservation method is incompatible with the molecular analytical protocol). Slatery and coworkers (3) have shown that tumor tissue from more advanced colon tumors was less likely to yield useful DNA.

Diagnostic methods directly influence the availability of tissue and, to some extent, the subpopulation with tumor tissue. Diagnostic practices and use of non-invasive methods can vary by country, region, hospital, and by calendar time period. Limited and differential availability of tumor tissue has implications for epidemiological studies, including generalizability, evaluation of critical target gene mutations and environmental exposures, and power. While the magnitude and direction of potential biases are study-specific, researchers should consider these issues when evaluating studies utilizing tumor blocks or other biological materials.

IMPLICATIONS

Generalizability

Tumor mutation analyses are used to describe the prevalence of genetic changes in tumors and the natural history of disease, i.e., molecular genetic changes as the tumor progresses over time. Differences in diagnostic practices within a study region or between countries may result in different estimates of the prevalence of a specific mutation. For example, in some countries such as Japan, almost all lung cancer cases are resected regardless of stage, so available tumor tissue represents the entire progression of disease. However, in other regions, biopsies may be restricted to those with early stage disease, and therefore, may have fewer disease-related changes. Thus, description of the genetic characteristics of invasive disease in this population would be limited to early disease. Another issue is the appropriateness of hospital pathology laboratory populations to characterize the disease experience of the general population.

Teaching hospitals and tertiary treatment hospitals often have different referral patterns, including difficult and unusual cases, than those that exist in the general population. For example, skin cancer resection and treatment in a dermatologist's office may differ from practices at a university hospital. While pathology laboratory-based studies are useful, especially for rare cancers, they may not represent the characteristics of disease in the general population, and thus, may over- or underestimate the prevalence of a genetic alteration.

While each individual study may be internally valid, drawing conclusions based on multiple studies may introduce bias. For example, comparisons of the prevalence of genetic alterations between two populations with different diagnostic practices could be biased if the mutation occurs in more advanced disease and advanced cases are less likely to have biopsies in one of the populations. Stratification by stage may minimize this bias. However, it cannot be completely eliminated since stage, as defined for clinical or prognostic purposes, may not accurately reflect tumor progression and evolution. Analysis by grade is difficult because subjects without tumor tissue will lack information regarding grade. In order to interpret these studies in a larger context, a complete description of the sample collection methodology and the procedures giving rise to the available sample are necessary.

Evaluation of Critical Target Gene Mutations and Environmental Exposures

Differential diagnostic or retrieval patterns of tumor tissue within the study population may result in selection bias for evaluation of the interaction of environmental agents and critical target gene mutations. Diagnostic practices may create a population sample with a different exposure prevalence than the general population, leading to bias in case-control comparisons. Tests regarding critical target gene mutations in case-case comparisons are unlikely to be biased unless the factors that determine the availability of the tumor block are associated both with the exposure and the mutation. As with other examples of selection bias in epidemiology, the comparisons within the group with tumor blocks will only be biased if the sampling fractions differ based on the joint distribution of exposure of interest and the outcome (4). Another potential concern may be introduced when retrieving historical tumor blocks. Some hospitals may keep their blocks longer or may selectively retain unique or intriguing cases, thereby creating a differential sample of tissue, that includes cases with unusual exposure, histology or mutations.

Power

Restrictions in tissue availability regardless of whether they arise as a result of diagnostic or surgical practices or speci-

men retrieval will reduce the sample size. Other types of material, such as cytology brushings, fine-needle aspirates, or sputum samples, may be available for investigation of genetic changes within tumors. However, use of these may have increased measurement error because they may not be as definitive as using tumor blocks to characterize mutations. While limited power is not a bias issue, it is a key study design concern as tissue availability will influence the feasibility to conduct molecular epidemiological studies.

EXAMPLE: PANCREATIC CANCER

Retrieval of tumor tissue for some cancer types can be more difficult than for others. Pancreatic cancer represents an extreme example of issues associated with tumor block retrieval due to the high prevalence of clinical diagnosis and the high case fatality rate. Exploration of factors related to tissue specimen retrieval in this group are relevant to studies that use tumor tissue. For this example, data from an ongoing population-based case-control study of pancreatic cancer at the University of California in San Francisco were used to describe the selection process for obtaining tumor blocks using data from both the study questionnaire and the Northern California Cancer Center (NCCC). The NCCC collected data on all incident pancreatic cancer cases as part of the Surveillance, Epidemiology and End Results (SEER) program. This included information on demographics (age, sex, race, county of residence), method of diagnosis, tumor characteristics (size, stage, grade), and vital status. The study used rapid case ascertainment using data obtained from pathology labs to identify all newly diagnosed cases in the San Francisco Bay Area. Rapid case ascertainment involves weekly review of pathology laboratory records for all participating hospitals in the SEER region to identify pancreatic cancer cases prior to receipt of the SEER abstracts.

Figure 1 illustrates the selection process for cases in the study and the subset for whom tumor tissue was potentially available during a 20-month period of the study. To represent the case population, data for all 1130 incident cases diagnosed between May 1995 and December 1997 were obtained from the NCCC, regardless of whether the subject was included in the case-control study. Even with rapid case ascertainment over half of the pancreatic cancer cases died prior to study recruitment, due to the high case-fatality rate. Following losses due to non-response and lack of tumor tissue, tumor tissue potentially was available for approximately 21% of all identified study cases. Tumor tissue was retrieved for 80% of subjects with tissue, resulting in 46 tumor blocks representing 16% of all cases interviewed and 4% of all pancreatic cancer cases diagnosed in this time period in this region. Even with rapid case-ascertainment, excellent response rates, and good tissue retrieval rates, tumor

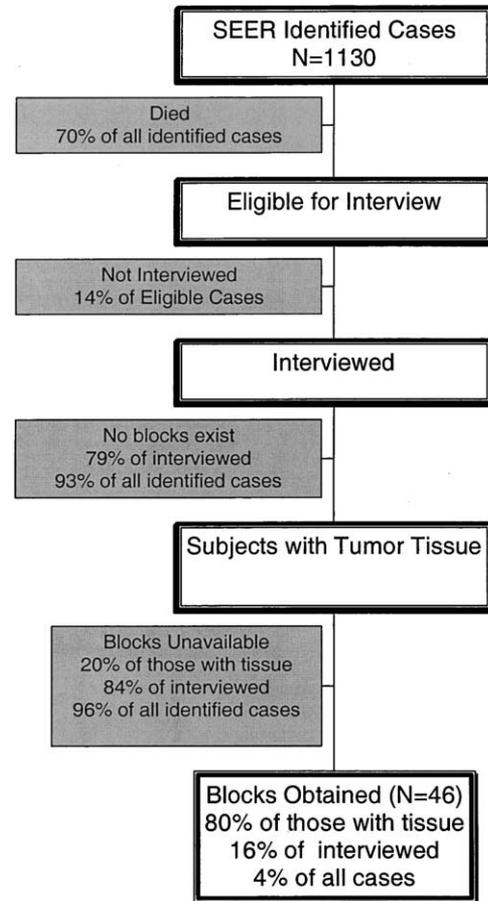


FIGURE 1. Selection process for pancreatic cancer cases and tumor tissue, San Francisco Bay Area, May 1995 to December 1997.

specimens were available for a very small fraction of the base population.

To assess potential differences between individuals with and without tumor tissue, we used data from NCCC for all San Francisco Bay Area cases and data from the ongoing study for study participants. SEER data represent a standardized source to explore issues associated with tissue block availability; however, it may not be as detailed as data collected in epidemiology studies. "Potential tumor tissue", or biopsy, was defined based on the report in the SEER abstract of any cancer-directed surgery. This excluded fine needle aspirates, but may included cancer-related surgeries that did not involve the pancreas and therefore, may overestimated the availability of blocks. Table 1 presents the tumor characteristics of the cases by biopsy status. Tumors were smaller among subjects who had had biopsies. In the SEER data, tumor size on non-biopsied subjects was determined from data obtained via x-ray or imaging scans and scopes where the lesion was observed. Tumor tissue appeared to be more commonly available among

TABLE 1. Tumor characteristics of subjects with and without potential tumor tissue. Pancreatic cancer cases diagnosed in San Francisco Bay SEER Area, May 1995 to December 1997. Source: Northern California Cancer Center

Tumor characteristic	Potential biopsy tissue ^a				P-value ^b
	Yes N = 142		No N = 988		
	Mean	sd	Mean	sd	
Tumor size ^c (mm)	179	343	570	474	<0.0001
	n	%	n	%	
Histologic grade ^d					
G1 (lowest grade)	16	11	49	5	0.001
G2	59	42	107	11	
G3	41	29	186	19	
G4 (highest grade)	0	0	17	2	
Unknown	26	18	629	64	
Summary stage					0.001
In situ	2	1	0	0	
Localized	22	15	59	6	
Regional, extension only	49	35	222	23	
Regional, nodes only	7	5	12	1	
Regional, extension and nodes	45	32	58	6	
Remote	13	9	498	51	
Unstaged	4	3	131	13	

^aPotential biopsy tissue defined as any pancreatic cancer directed surgery.

^bP-value for t-test for continuous data, chi-squared test for categorical data.

^cTumor size for non-biopsied tumors is based on description of primary tumor from x-ray or imaging scans and scopes where the lesion was observed.

^dHistologic grade for non-biopsied tumors is based on cytology reports from needle or incisional biopsy.

those with low-grade tumors. However, many pancreatic cancer patients without tumor tissue blocks were missing grade information. These data suggest that while descriptions of genetic alterations in high-grade disease and in large tumors will be limited, analysis of lower-grade tumors is feasible.

Subjects who had biopsies lived longer than those who did not ($p = 0.0001$) (Figure 2), suggesting that these individuals were more likely to be alive and able to participate in the case-control study. As seen in Table 2, individuals with tumor blocks were more likely to be younger and white race. The sample size for non-white cases was small, particularly among those who had biopsies. Since availability of blocks was associated with survival, studies of prognostic factors may be limited to those factors associated with early disease. There was some suggestion of differences in surgery rates among the six counties with rates ranging from 8 to 17%, although, differences across counties were not statistically significant ($p = 0.14$). If the difference in surgery rates by county was correlated with an exposure of interest (such as drinking water source), then statistical analysis of environmental exposures and critical target gene mutations may be affected.

To explore whether potential exposure related factors differed between those with and without tumor blocks, we utilized the detailed study data for the 294 cases identified during this time period. We observed no difference in the distribution of potential risk factors (smoking, diabetes) between individuals with and without tumor tissue. However, women were more likely to have tumor blocks obtained than men. If women were less likely to have an exposure of interest than men, such as an occupational exposure, this would limit the power to evaluate that hypothesis but would not introduce bias unless women with tissue had a different probability of both the mutation and the exposure than women without tissue. For this illustration, we did not explore differences in occupation among subjects with and without tumor blocks given the small sample size.

STRATEGIES TO ASSESS SAMPLE LOSS

Although problems with sample selection are often easy to identify and consider conceptually, they may be intractable and difficult to quantify, due to non-identifiability of the selection probabilities (5, 6). Selection bias cannot be evaluated empirically since all analyses are limited to the population for which there are data, in this case, tumor tissue. Even with this limitation, the following strategies can be employed to explore the extent of selection bias in the sample, through sensitivity analyses and statistical methods to account for missing data.

Since the tumor characteristics of the subjects without tissue cannot be determined, efforts should be made to identify the determinants of biopsies to assess whether disease characteristics or subject-related factors, such as age, medical conditions, or exposures, may be associated with the presence or absence of a tumor block. In addition to characterizing the demographic factors that relate to the patients who provided tumor blocks, exposure prevalence should be compared. To begin, the population giving rise to the samples should be described thoroughly (Figure 1 in our example). Frequently in pathology laboratory-based samples, the selection process is not described fully or description is limited by the lack of complete information on the cases. Both the demographic characteristics of the population giving rise to the study population (Tables 2 and 3) and the tumor characteristics of the case population (Table 1) should be described. The characteristics of the subset with tumor blocks should be described, using any available data. As seen in Table 1, comparisons by tumor grade are limited since 64% of subjects without biopsy tissue had no information regarding tumor grade.

Statistical methods are available to explore the potential for selection bias through sensitivity analyses and missing data methods. New methods have been developed to assess non-random missingness and the impact of selection bias in

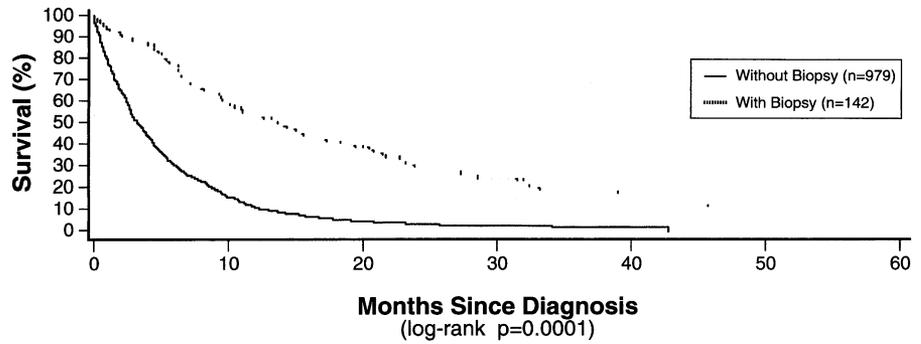


FIGURE 2. Survival curves for incident pancreatic cancer cases by biopsy status, San Francisco Bay Area, May 1995 to December 1997.

observational studies (7). These methods use non-parametric identified models indexed by the selection bias function that quantifies the magnitude of selection bias due to unobservable data (7). Adjusting for selection bias requires knowledge of the selection probabilities for each cell in the table and weighting accordingly (5, 6). Expectation-maximization (E-M) models have been developed that can be used to obtain asymptotically unbiased estimates of case-control associations when exposure data are available for all cases. However, these methods will introduce bias when the availability of biopsy material is related to mutation status (8, 9). Case-case comparisons will provide unbiased estimates of exposure-outcome associations when block availability is related either to exposure status or mutation

status, but will be biased when tissue availability is related to both (8). Even with improved analytical methods to evaluate or reduce the impact of selection bias on study findings, selection bias is still best considered and addressed in the study design phase rather than the analytical phase.

Aggressive study recruitment methods and improved analytical techniques are necessary to analyze all available samples and tumor tissue. To minimize the impact of limited tissue availability, efforts must be made to maximize retrieval efforts and to ensure cooperation of all hospitals and pathology laboratories in the study region. Pathology labo-

TABLE 2. Patient characteristics of subjects with and without potential tumor tissue. Pancreatic cancer cases diagnosed in San Francisco Bay SEER Area, May 1995 to December 1997. Source: Northern California Cancer Center

Subject characteristic	Potential Biopsy Tissue ^a				p-value ^b
	Yes n = 142		No n = 988		
	Mean n	sd %	Mean n	sd %	
Age at diagnosis	64.4	10.8	68.7	10.7	0.0001
Race					0.02
White	120	84%	720	74%	
Black	10	7%	127	13%	
Other	12	8%	129	13%	
Sex					0.72
Female	74	48%	499	49%	
Male	68	52%	489	51%	
County					0.14
1	32	23%	253	26%	
2	20	14%	159	16%	
3	8	6%	56	6%	
4	13	9%	147	15%	
5	21	15%	133	14%	
6	48	34%	240	24%	

^aPotential biopsy tissue defined as any pancreatic cancer directed surgery.

^bp-value for t-test for continuous data, chi-squared test for categorical data.

TABLE 3. Patient characteristics of study subjects with and without biopsy tissue. Participants in an ongoing pancreatic cancer case-control study of cases diagnosed in San Francisco Bay Area, May 1995 to December 1997

Characteristics	All study cases n = 294		Cases with biopsy tissue n = 46		p-value ^a
	Mean	sd	Mean	sd	
Age at Diagnosis	64.9	10.9	64.4	10.9	.70
	n	%	n	%	
Race					.50
White	239	81%	40	85%	
Black	29	10%	4	9%	
Other	26	9%	3	6%	
Sex					.01
Female	133	45%	29	63%	
Male	161	55%	17	37%	
Smoking > 100 cigarettes in a lifetime					.31
Yes	198	67%	28	61%	
Diagnosed with diabetes more than one year earlier	45	15%	5	11%	.50

^ap-values are based on results of analyses to compare characteristics of patients with tumor blocks (n = 46) with those w/out tumor blocks (n = 248). T-test analysis was used to compare difference in mean age, Fisher's exact test was used for 2 × 2 contingency table analyses (sex; smoking; diabetes) and Wilcoxon rank sum test was used for analyses of 2 × 3 tables (race).

ratories should be encouraged to retain specimens as long as possible, especially for rare diseases that may require many years to obtain a sufficient sample size. Since diagnostic methods are moving toward non-invasive technologies, development of molecular analyses that utilize other tissues, such as serum or pathology slides, to assess genetic alterations will be helpful. With the expansion of molecular biological techniques, the demand for tumor tissue will increase. As a result, epidemiological studies will need to maximize the efficiency for using available tissue and conserve this resource for future investigations.

Tumor tissue and other biological samples are critical to understand the disease process of cancers (1). While availability of these tissues may be limited, tumor tissue is important to understand the natural history of disease, the prevalence of genetic changes in tumors, and the environmental and medical factors associated with these alterations. When using tumor tissue, epidemiologists and molecular biologists need to provide a thorough description of the population giving rise to the sample and the determinants of tissue availability. As new statistical methods to address problems with missing samples are developed, and new molecular biological techniques to work with small and degraded samples are created, the impact of selection bias will be reduced but not eliminated.

This work was supported in part by grant number EDT-101 from the American Cancer Society and grant number R01-CA59706 from the National Cancer Institute, National Institutes of Health. We thank Paige

Bracci and Rebecca Zhang for their statistical and database support and Northern California Cancer Center for use of the SEER data.

REFERENCES

1. Taylor JA. Oncogenes and their applications in epidemiologic studies. *Am J Epidemiol.* 1989;130:6–13.
2. Toniolo P, Boffetta P, Shuker D, Rothman B, Pearce N. Methodological issues in the use of tumour markers in cancer epidemiology. *Applications of Biomarkers in Cancer Epidemiology.* IARC Scientific Publications. Lyon: IARC; 1977:201–213.
3. Slattery ML, Edwards SL, Palmer L, Curtin K, Morse J, Anderson K et al. Use of archival tissue in epidemiologic studies: Collection procedures and assessment of potential sources of bias. *Mutat Res.* 2000;432:7–14.
4. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. *Am J Epidemiol.* 1992;135:1019–1028.
5. Kleinbaum DG, Morgenstern H, Kupper LL. Selection bias in epidemiologic studies. *Am J Epidemiol.* 1981;113:452–463.
6. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol.* 1996;25:1107–1116.
7. Robins J, Rotnitzky A, Scharfstein D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran M, Berry D, eds. *Statistical Models in Epidemiology: The Environment and Clinical Trials.* New York: Springer-Verlag; 1999:1–92.
8. Schroeder JC, Weinberg CR. Use of Missing-data methods to correct bias and improve precision in case-control studies where cases are subtyped but subtype information is incomplete. *Am J Epidemiol.* In press.
9. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM Algorithm. *J R Statistic Soc B.* 1977;39:1–38.