Reliability of smoking history in medical charts in relation to interview data

DAPHNA POLLACK¹, JUDY JUBAS¹, ZUO-FENG ZHANG^{1,2}, STEVEN D. STELLMAN³ and SUSAN HARLAP^{1,2}

¹Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center; ²Department of Public Health, Cornell University Medical College; ³Division of Epidemiology, American Health Foundation, USA

Received October 3, 1995; Accepted October 23, 1995

Abstract. We previously reported on a series of studies exploring relationships between cigarette smoking and mutations of tumor suppressor genes; the studies relying on medical charts for data on smoking. To assess the accuracy of these data we compared them with data obtained using a conventional epidemiologic interview. As part of an ongoing case-control study of tobacco-related cancers, a trained interviewer questioned 144 patients about their smoking habits using a structured interview instrument; the medical charts were then reviewed. Using the Spearman test to compare the data obtained from the two sources, the correlation coefficient was 0.85 for cigarettes per day; 0.88 for years of smoking and 0.95 for pack-years, suggesting that the data in the charts were reliable. The kappa coefficient was 0.96 for current smoking status, indicating a very strong agreement between the two sources of data. The sensitivity and specificity for chart review were 96.3% and 100% respectively, compared with the 'gold standard' of a structured interview. These results show that when patients cannot be interviewed, for example when performing retrospective studies using existing collections of archived tumors or other biologic materials, then data on smoking abstracted from medical charts may be a reliable substitute.

Introduction

Collections of tumors and other archived biologic materials exist in many cancer centers. They are now being used extensively by oncologists, clinicians and pathologists to investigate how biomarkers such as tumor suppressor genes and oncogenes correlate with progression and survival of patients with cancer. It is especially important to investigators who intend to use medical records to obtain information on smoking exposure of patients for whom they have archived biological specimens. More recently, we (1-5) and others (6,7) have shown that such studies can provide

important insights into mechanisms of carcinogenesis, based on data of previous exposure history abstracted from medical charts and biomarkers from archived biological specimens. Our approach has been to investigate how abnormalities of tumor suppressor genes relate to previous exposure to epidemiologic risk factors. We have extracted data on smoking and other risk factors recorded in the medical record and studied tobacco use, alcohol, family history and body size in relation to overexpression of p53 nuclear protein and other biomarkers in cancer of the bladder (1,2), prostate (3), colon (4) and stomach (5).

Epidemiologic studies of this type can be done only if stored data on exposure to the risk factors, such as smoking, and archived tumor specimens exist. Molecular epidemiologic studies can maximize the use of recently available molecular biomarkers and supplement this information with data obtained from retrospective chart review. It allows epidemiologists to link patients' previous exposures to their molecular genetic abnormalities. In addition, information on risk factors abstracted from medical charts may also be useful for oncologists to study the relationship between various risk factors and cancer progression and survival.

Very few studies have assessed the reliability and validity of data on smoking abstracted from medical charts. Although many studies have been done to compare the quality of information obtained from questionnaires with that from chart review, most have focussed on chronic illness or medical care, use of medication or female reproductive history (8). The majority of these studies considered the medical record as the standard measurement and thus evaluated the accuracy of data obtained from questionnaires against that obtained from the chart. The purpose of our study is to assess the reliability of data on smoking extracted from medical charts compared to information collected by an epidemiologic interview.

Materials and methods

Patient population. This study is based on patients seen at Memorial Sloan-Kettering Cancer Center from July 1993 through March 1994, who participated in an ongoing case-control study of tobacco-related cancers. Included were patients aged 20-80 admitted for cancer diagnoses as well as controls attending screening clinics.

Correspondence to: Dr Zuo-Feng Zhang, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, Box 44, New York, N.Y. 10021, USA

Key words: biomarkers, chart review, interview; molecular epidemiology, neoplasia, reliability, smoking

Data collection. All subjects were interviewed by a trained interviewer using a structured questionnaire to obtain detailed history of smoking. The interviewer then abstracted comparable information on smoking from the medical records of the same admission. Data in the charts in our institution had been systematically recorded in a standardized format by physicians or nurses upon admission or during the patients stay in the hospital. Information was available on whether the patient had ever smoked, and if so, the number of cigarettes per day and number of years smoked.

Statistical methods. We compared the data from the two sources, interview and medical records, in several ways. Continuous variables were compared using the Spearman correlation coefficient. Correlation was assessed for the following variables; cigarettes smoked per day, number of years smoked, and number of pack years. In order to assess if the correlation between sources was affected by other factors, the analysis was then stratified by sex, type of diagnosis, and age group. Additionally a concordance matrix was constructed to compare the values of daily cigarette use from both sources. Agreement between the chart and the interview for the dichotomous variable current smoking was evaluated using the kappa statistic (9). Agreement was then classified as either poor to fair (<0.40), moderate (0.40-0.59), substantial (0.60-0.79), strong (0.80-0.89), and very strong (≥ 0.90) (10). To determine the accuracy of chart review versus patient interview, sensitivity and specificity were measured.

Results

Characteristics of study population. One hundred and forty-four inpatient records were reviewed. The mean age of the subjects was 68, and 69% were male (Table I). Approximately 70% of the patients were smokers and 76% had smoking related diagnoses. Ninety-nine percent of the patients were white.

Never/ever smoking status was available for all patients. Information on number of years smoked was available for 141 (97.9%) patients, and 140 (97.2%) charts had data on number of cigarettes per day. Eighty percent of the patients claimed to be smokers when interviewed and 77 percent classified themselves as smokers in the medical record. The mean number of cigarettes smoked per day was 29.8 according to the chart and 31.9 based on the interview.

Table I. Characteristics of study population.

Characteristic	N	(%)
Age		
<50	19	13.2
50-59	31	21.5
60-69	63	43.8
≥ 70	31	21.5
Sex		
Male	99	68.8
Female	45	31.2
Cancer site		
Lung	83	57.6
Mesothelioma	1	0.69
Bladder	26	18.1
Prostate	16	11.1
Leukemia/lymphoma	18	12.5

Correlation. Among smokers correlation between the chart data and the interview data was significant for all three measures of smoking; cigarettes per day, years smoked, and pack-years, with correlation coefficients of 0.70, 0.78 and 0.81 respectively (Table II). These results did not differ significantly upon further stratification by age, sex, and smoking related diagnosis. The group with the lowest correlation was patients aged fifty to fifty-nine (Table III).

Kappa coefficient and concordance. The kappa coefficient was 0.96 for never/ever smoking status indicating almost perfect agreement between the two sources of information. Table IV shows the results of cross-classification of the number of cigarettes, grouped in increments of 10 cigarettes per day. The total number of exact concordances, represented by the diagonal from the top left corner of the matrix to the bottom right corner, is 104 (74.3%). When the data were cross-classified with a larger cut-point at 20 cigarettes (not shown), corresponding to one pack per day, the concordance increased to 116 (82.8%). The overall sensitivity was 96.3% and the specificity was 100% compared with the epidemiologic interview.

Table II. Comparison of smoking history information from medical records and interview data.

Variable	Chart			Interview		Spearman correlation coefficient
	N	Median	Range	Median	Range	
Cigarettes/day	107	20	3-80	30	3-80	0.70
Years smoked	108	30	1-65	37	1-63	0.78
Pack years	107	45	1-168	50	1-148	0.81

Missing values were excluded from the analysis.

Table III. Correlation between interview and record smoking history by sex, diagnosis, and age group.

Strata	No. of pack years Chart			No. of pack years Interview		Spearman correlation coefficient
	N	Median	Range	Median	Range	
Sex						
Male	75	50	1-168	53	1-148	0.89
Female	32	40	1.5-108	43	1.9-118	0.71
Diagnosis						
Smoking related	95	50	1.5-150	53	1.9-148	0.78
Non smoking related	12	29	1-168	32	1-144	0.96
Age						
<50	16	28.5	1-90	25	1-72	0.73
50-59	22	30	10.5-110	48	4-112	0.67
60-69	45	50	15-168	60	10-148	0.87
≥70	24	50	20-130	52	7-126	0.89

Missing values were excluded from the analysis.

Table IV. Cross-classification of daily cigarette smoking evaluated by epidemiologic interview and by medical record review.

Information from chart review	Number of cigarettes per day Information from questionnaire						Total
	0	1-10	11-20	21-30	31-40	41+	
0	29	0	3	0	1	0	33
1-10	0	4	0	0	1	0	5
11-20	0	3	35	7	4	2	51
21-30	0	0	1	10	1	0	12
31-40	0	0	3	3	17	4	27
41+	0	1	0	0	2	9	12
Total	29	8	42	20	26	15	140

Assessment of information bias. In order to assess the possibility of information bias on the part of the data reviewer, comparable data on smoking was obtained electronically from the hospital tumor registry database. Information in the cancer registry is abstracted from the medical record by trained data registrars. The registry records smoking history and the number of pack years. Smoking status is classified as current, past or never use. Comparing never/ever smoking status from the cancer registry with that from the interview we obtained a kappa coefficient of 0.82. Similarly, comparison of smoking history from the chart with that from the registry yielded a kappa statistic of 0.80. The overall concordance for the number of pack years as recorded in the chart versus that recorded in the tumor registry was 88%, which is slightly higher than the concordance from cigarettes per day obtained by interview versus that

abstracted from medical charts (82.8%). The results suggested that the information bias was minimal.

Discussion

These results show that the smoking history recorded in medical charts is reliable when compared to data obtained from epidemiologic interviews. This conclusion validates our previous studies of biomarkers of tumor suppressor genes that have relied on the medical charts for information on smoking. No differences were noted for accuracy of recall among those patients with smoking related diagnoses versus those with other diagnoses. However, smoking related sites made up 76% of the study population.

We obtained a kappa statistic of 0.96 for current smoking status which demonstrates the reproducibility of this

information from the alternate data source. We observed a strong correlation for level of smoking. The concordance matrix showed that chart data are somewhat limited in detecting exact values when they fall below a cut-off point of one pack per day as compared to data obtained in the interview. Smokers versus non-smokers were detected with a sensitivity of 96.3% and a specificity of 100%.

Similarly in a study of risk factors for osteoporosis, Beard *et al* found a very high agreement ($K=0.96$) for never/ever smoking in a random sample of Minnesota women (11). Brownson *et al* studied risk factors for cancer, comparing data from interview with data in a central cancer registry (12). Their results show 85% agreement between registry and interview information for packs consumed per day with a high correlation ($r=0.93$); and a sensitivity and specificity of 80% and 91% respectively.

We found that medical staff can accurately record smoking information in the medical record, although information from medical chart may lack the detailed smoking history such as age at initiation of smoking, changes in smoking patterns over time, and types of tobacco used. Smoking data obtained from medical records can provide summary statistics pertaining to levels of previous tobacco exposure. These data are reliable and suitable for preliminary studies and for initial descriptive studies whose aim is to generate hypotheses. In summary, information on smoking history obtained from medical records can serve as a useful resource for epidemiologists, oncologists, and other medical researchers working with biomarkers in cancer, providing that the data in the chart are systematically collected in a standardized format.

Acknowledgements

This publication was made possible in part by grant number R01 ES-06718 (Dr Zuo-Feng Zhang) from the National Institute of Environmental Health Science, and grant number P01 CA-32617 (Dr Steven Stellman) from the National Cancer Institute, National Institute of Health, Department of Health and Human Services. We thank Ms Claudette Green for her assistance in typing and editing the manuscript.

References

1. Zhang ZF, Sarkis AS, Cordon-Cardo C, *et al*: Tobacco smoking, occupation, and p53 nuclear overexpression in early stage bladder cancer. *Cancer Epidemiol Biomarkers Prev* 3: 19-24, 1994.
2. Begg CB and Zhang ZF: Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prev* 3: 173-175, 1994.
3. Zhang ZF, Aprikian A, Sarkis AS, *et al*: Factors associated with p53 accumulation in prostatic adenocarcinoma. *Int J Oncol* 4: 897-901, 1994.
4. Zhang ZF, Zeng ZS, Sarkis AS, *et al*: Family history of cancer, body weight and p53 nuclear overexpression in Duke's C colorectal cancer. *Br J Cancer* 71: 888-893, 1995.
5. Zhang ZF, Karpeh MS, Lauwers GY, *et al*: A case-series study of p53 nuclear overexpression in early stage intestinal-type stomach cancer. *Cancer Detection Prev* 19: 156-164, 1995.
6. Spruck CH III, Rideout WM III, Olumi AF, *et al*: Distinct pattern of p53 mutations in bladder cancer: Relationship to tobacco usage. *Cancer Res* 53: 1162-1166, 1993.
7. Ozturk M, *et al*: P53 mutation in hepatocellular carcinoma after aflatoxin exposure. *Lancet* 338: 1356-1359, 1991.
8. Harlow SD and Linet MS: Agreement between questionnaire data and medical records. *Am J Epidemiol* 129: 233-248, 1989.
9. Cohen J: A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20: 37-47, 1960.
10. Landis JR and Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174, 1977.
11. Beard CM, Melton LJ, Cedel SL, *et al*: Ascertainment of risk factors for osteoporosis: comparison of interview data with medical record review. *J Bone Miner Res* 5: 691-699, 1990.
12. Brownson RC, Davis JR, Chang JC, *et al*: A study of the accuracy of cancer risk factor information reported to a central registry compared with that obtained by interview. *Am J Epidemiol* 129: 616-624, 1989.