Eur J Cancer, Vol. 28, No. 1, pp. 116–117, 1992.
Printed in Great Britain

0964–1947/92 $5.00 + 0.00

Comparison of Questionnaire-derived and Tumour Registry-derived Smoking Histories

Jack L. Mayer, Paolo Boffetta and Maxine M. Kuroda

Information on cigarette smoking, an exposure of great epidemiological interest, is occasionally obtained from tumour registry data. We compared smoking histories from a hospital tumour registry with those from a questionnaire administered to 94 lung cancer patients. Reliability of tumour registry data was good for classifying individuals as ever-smokers and non-smokers (sensitivity 0.96, specificity 0.86). There was higher discrepancy in classifying smokers as current or former smokers. Current smokers had lower reliability for amount of smoking than former smokers, whereas both groups had high reliability for duration of smoking. These results suggest that tumour registry derived data on smoking must be used with caution.
Eur J Cancer, Vol. 28, No. 1, pp. 116–117, 1992.

INTRODUCTION

SMOKING HISTORIES by personal interview have been shown to be reliable. For example, when smoking histories from interviews were compared with longitudinal records available 20 years prior to their study, Krall *et al.* [1] found agreement was 87% for smoking status and 71% for amount smoked. However, epidemiological studies of cancer may rely upon tumour registry data derived from medical records for measures of exposure [2], and to our knowledge, there has been one report comparing smoking data contained in a tumour registry with smoking histories from a telephone-administered questionnaire [3].

PATIENTS AND METHODS

The Columbia Presbyterian Medical Center (CPMC) Tumour Registry was established in 1980 and contains more than 22 000 records. Information from patients' charts is abstracted by individuals specially trained to review clinical oncology records. Data are then entered by a data processor from the abstraction forms. 10–20% of all entries are randomly checked by the registry coordinator. All entries are checked by a program designed to flag possible keying errors and to assess consistency of values.

Smoking histories were extracted from a validated environmental exposure questionnaire that had been developed by the Division of Environmental Sciences at Columbia University. This questionnaire was administered to patients enrolled in a lung cancer case-control study at CPMC between 1983 and 1988 [4]. Smoking status was defined as never a smoker, current smoker (including those who had stopped within the last 2 months), and former smoker (stopped 2 or more months ago). Lifetime smoking history comprised information on amount of smoking (calculated as packs per day [ppd]) and duration. Three measures of tobacco exposure: smoking status (never, current, former), amount (ppd) and total years smoked were obtained from the CPMC registry database for this study. The definition of former smoker in the tumor registry is consistent with the definition derived from the questionnaire.

Registry and questionnaire entries for age and ethnicity were abstracted as a check for data that are expected to have high agreement. The interval between questionnaire and registry data collection was less than 2 weeks for most subjects. Reliability of smoking status was measured by the kappa statistic as well as by sensitivity and specificity indices [5]. Reliabilities for amount of smoking and duration were measured by Pearson product moment correlation coefficients using subjects classified as smokers in both data sets.

RESULTS

94 cases were reported in the tumour registry. In only three records (3.2%) was there a discrepancy in age of more than one year (2, 6 and 8 years, respectively). Five other records showed discrepancies of one year that could be explained by rounding procedures. The overall Pearson correlation coefficient for age was 0.995 ($P < 0.001$). There were six discrepancies (6.5%) in ethnicity (3 white vs. black and 3 white vs. hispanic).

Table 1 compares smoking status between the two sources of data. No cases had missing information on smoking status from the questionnaire. The kappa statistic was 0.623, which is considered to represent good agreement beyond chance [5]. It should be noted that 5/7 patients with missing data from the tumour registry were non-smokers.

Questionnaire-derived smoking histories are first-hand accounts from the patient. Using this source as the standard,

Table 1. Smoking status according to questionnaire and tumour registry datasets

	Questionnaire			
	Non-smoker	Current smoker	Former smoker	Total
Tumour registry				
Non-smoker	6	1	2	9
Current smoker	1	31	7	39
Former smoker	0	8	31	39
Missing data	5	1	1	7
Total	12	41	41	94

Correspondence to J.L. Mayer.
J.L. Mayer and M.M. Kuroda are at the Division of Environmental Sciences, Columbia University School of Public Health, 60 Haven Avenue, Room B1–109, New York City, New York 10032, U.S.A.; P. Boffetta is at the International Agency for Research on Cancer, Lyon, France.
Received 18 July 1991; accepted 27 Sept. 1991.

Table 2. Correlation of amount and duration of smoking among current and former smokers

Smoking status		Amount			Duration		
Questionnaire	Tumour registry	r	n	P	r	n	P
Current	Current	0.184	28	0.34	0.854	22	< 0.001
Current	Former	0.569	7	0.18	0.870	7	0.01
Former	Current	0.394	5	0.51	0.958	5	0.01
Former	Former	0.621	25	0.001	0.749	18	<0.001
All smokers		0.426	65	< 0.001	0.824	52	< 0.001

sensitivity and specificity can be calculated for the second-hand accounts of smoking history contained in tumour registry data, which is abstracted from the medical record. There was a higher discrepancy in classifying a smoker as a current or former smoker (sensitivity 0.795, specificity 0.816) than in classifying any individual as a smoker or non-smoker (sensitivity 0.963, specificity 0.857).

Analyses of amount of smoking and duration were conducted on the 77 individuals who were classified as current or former smokers in both datasets. As shown in Table 2, correlation coefficients were 0.426 for amount of smoking ($P < 0.001$) and 0.824 for duration ($P < 0.001$). There was no difference in ppd for 22/65 subjects (35%), and 17/65 subjects (26%) differed by no more than 0.5 ppd, a difference which may be explainable by rounding procedures. The overall mean difference was 0.23 ppd, with tumour registry records producing the higher mean. This difference was not significant by paired t-test ($P > 0.05$). There was no difference in duration of smoking for 13/52 subjects (25%), and the difference was within 5 years for 22/52 subjects (42%). Values for duration from the questionnaire tended to be slightly higher than values from the registry. The mean difference of 2.13 years was not significant by paired t-test ($P > 0.05$).

As shown in Table 2, the correlation for amount of smoking was highest among subjects classified as former smokers in both datasets ($r = 0.621$). For current smokers in both sources, the correlation for amount of smoking was notably low ($r = 0.184$). Among individuals classified as current smokers in one source and as former smokers in the other, the correlations for amount of smoking were intermediate ($r = 0.569$ and $r = 0.394$). In contrast to the correlations for amount, the correlations for duration of smoking appeared to be consistently high and statistically significant in all the combinations of smoking status (r ranged from 0.749 to 0.958).

DISCUSSION

In our investigation of the concordance between smoking histories obtained from an interviewer-administered questionnaire and the CPMC Tumour Registry, the kappa statistic for smoking status was 0.623. Although this is considered a high value for random data [5], such agreement may be too low when assessing exposures in epidemiological studies. The effect of non-differential misclassification in biasing odds ratios toward the null in case-control studies is well known and can lead to dramatically different interpretations of results [6].

Our analysis demonstrates good correlation for duration of smoking but poorer correlation for amount smoked in current smokers. On the other hand, for former smokers, both duration of smoking and amount smoked were reliably represented by hospital tumour registry data. Former smokers may have firmly fixed their accounts of consumption whereas current smokers, especially individuals recently diagnosed with lung cancer, may alter reporting because of concern about the possible contribution of smoking to the causation of their disease ('wish bias' [7]) or changes in smoking habits subsequent to early symptoms of disease.

A previous study reported a comparison of smoking status (smoker vs non-smoker) and average number of ppd (<1, 1–2, >2) among 439 cancer patients reported to the population based Missouri Cancer Registry, who answered by telephone to a brief, standardised questionnaire [3]. 83 of these patients had tobacco-related cancers. Information on smoking in the cancer registry is collected from hospital medical records by trained registrars. The registry identified smokers from non-smokers with a sensitivity of 0.80 and a specificity of 0.91; for tobacco related cancers, these values were 0.97 and 0.67 (based on 3 cases), respectively. The correlation on ppd was very high ($r = 0.93$). Our study closely replicated these results with respect to binary classification of smoking status (sensitivity 0.96, specificity 0.86). Our lower correlation coefficient on amount of smoking might be due to a more precise measure used in our study.

Although our analysis used only the CPMC Tumour Registry, it is conceivable that vulnerability to misclassification is not restricted to this hospital's tumour registry. Hence, while smoking histories derived from tumour registries might be useful when classifying individuals as smokers or non-smokers, evidence from this study suggests that classification of individuals as current or former smokers and data on amount smoked, are likely to introduce an important misclassification.

1. Krall EA, Valadian I, Dwyer JT, Gardner J. Accuracy of recalled smoking data. *Am J Public Health* 1989, **79**, 200–202.
2. Muir CS, Demaret E, Boyle P. The cancer registry in cancer control: An overview. In: Parkin DM, Wagner G, Muir CS, eds. *The Role of the Registry in Cancer Control*. International Agency for Research on Cancer, Lyon, 1985, 13–26.
3. Brownson RC, Davis JR, Chang JC, DiLorenzo TM, Keefe TJ, Bagby JR. A study of the accuracy of cancer risk factor information reported to a central registry compared with that obtained by interview. *Am J Epidemiol* 1989, **129**, 616–624.
4. Perera FP, Mayer JL, Jaretzki A. *et al.* Comparison of DNA adducts and sister chromatic exchange in lung cancer cases and controls. *Cancer Res* 1989, **49**, 4446–4451.
5. Fleiss JL. *Statistical Methods for Rates and Proportions*. New York, John Wiley, 1981.
6. Kelsey JL, Thompson WD, Evans AS. *Methods in Observational Epidemiology*. New York, Oxford University Press, 1986.
7. Wynder EL. Guidelines for the epidemiology of weak associations. *Prev Med* 1987, **16**, 211–212.

Acknowledgements—The authors thank Dr Frederica Perera for her guidance and assistance with this study. We are grateful to Gabriella Simon-Cereijido and Agnès Hanss-Cousseau for help with the preparation of the manuscript and Dennis Timony of the CPMC tumour registry for his assistance with data retrieval.