A comparison of individual and population smoking data from a postal survey and general practice records

ANDREW WILSON

TERJINDER MANKU-SCOTT

DAVID SHEPHERD

BYRON JONES

SUMMARY

Background. *Data on smoking held by general practitioners (GPs) may contribute to clinical care and to an assessment of population health. However, these data are prone to several biases and their validity has not been tested.*

Aim. *To examine the accuracy of general practice data as an estimate for population prevalence of smoking and to estimate the accuracy of GP data on individuals' smoking habit compared with self-report.*

Method. *A postal questionnaire on smoking habit over the past six years was sent to a random sample of individuals aged 15 to 74 years and registered with five out of seven general practices in one part of Leicester. GP records of those sampled were examined for an entry of smoking status over this period.*

Results. *Response rate to the postal questionnaire was 1906 out of 2490 (76.5%). Reported smoking prevalence was 35.2%. Of those notes sampled, 1784 out of 2432 (73.4%) had an entry about smoking recorded between 1991 and 1996. Patients recorded as smokers were less likely to respond to the postal questionnaire than non-smokers. Using practice data to ascribe smoking status to non-responders produced an estimated prevalence of 38.6%. Using questionnaire data alone as the 'gold standard', the last practice record collected since 1991 overestimated current smoking prevalence by a factor of 1.22; using questionnaire data supplemented by practice data for non-responders as the 'gold standard' meant that the overestimate was by a factor of 1.11. Data from notes and the questionnaire were available for 1398 individuals and 2188 observations. Levels of agreement were high (k = 0.83).*

Conclusion. *GP-held data are valid for individuals but overestimate smoking prevalence at a population level.*

Keywords: primary care; survey; smoking.

Introduction

SMOKING is a major public health issue and an important risk factor for many conditions commonly presenting to general practitioners (GPs), so it is not surprising that data on cigarette consumption are collected in primary care, both as part of individual clinical care and in health promotion. Little is known about the accuracy of smoking data collected in general practice at either individual or population levels. In 1986, Mant and Phillips found that risk factors were recorded on an unrepresentative population and urged caution in using these data for population estimates.[1]

There are several potential sources of error in the general practice smoking record. First, no specific guidance was given to practices on how smoking habit should be defined.[2] Practices may differ in the duration of cessation required before ascribing 'ex-smoker' status. Such differences could affect prevalence estimates by up to 7%.[3]

Secondly, the recording of smoking data by GPs is subject to ascertainment bias; opportunistic collection of data leads to a bias towards frequent attenders. The fourth National Morbidity Study found that smokers aged 16 to 64 years were more likely to consult, although in the older age group smokers were less likely to consult.[4] Doctors may also be more likely to record smoking status in smokers than in non-smokers, the so-called 'worst first' bias.[1] Conversely, recording at well-person clinics — which smokers are less likely to attend[5] — will underestimate the true prevalence rate.

Thirdly, patients may misrepresent their smoking habit to doctors. In trials of cessation, up to 20% of smokers denied their habit.[6] Although it is likely that such a level of misrepresentation will be lower in the ordinary surgery setting, patients have several reasons to misrepresent their habit to the doctor. These include a desire to avoid criticism and to register as a non-smoker for life insurance proposals. Lastly data quality may decay with time.[7]

In this project we had the following aims:

- to examine the accuracy of general practice data as an estimate for population prevalence of smoking, and
- to estimate the accuracy of GP data on individuals' smoking habit compared with self-report.

Method

Five out of the seven practices in one part of Leicester defined by the boundaries of three wards agreed to take part. Of the 32450 residents in this locality, 22 480 (69.3%) were registered with a participating practice.

A postal questionnaire asking about smoking over the past six years was sent to a sample of those registered with participating practices. The medical records of those sampled were then examined for any entry of smoking during each of the years under scrutiny.

The sample was drawn from individuals aged 15 to 74 years registered with the five participating practices. Sampling was weighted by practice size, and stratified by sex and ten-year age bands. The weighting given to practice size was chosen to enable a reasonably precise estimate of smoking prevalence in all practices, requiring at least 300 responders from each practice. The initial sample was 2500.

The questionnaire enquired about current and previous smoking habits using identical questions to those included on the bio-

A Wilson, MD, FRCGP, senior lecturer in general practice; T Manku-Scott, BA, MSc, research associate; and D Shepherd, MBChB, MRCGP, clinical research fellow, Department of General Practice and Primary Health Care, University of Leicester. B Jones, PhD, professor of medical statistics, Department of Medical Statistics, De Montfort University, Leicester.
Submitted: 23 April 1999; final acceptance: 18 February 2000.

chemically validated questionnaire reported by the Scottish Heart Health Study.[8] Reminder letters and further questionnaires were sent out twice at two-weekly intervals as appropriate. The study was conducted in 1996 and 1997.

General practice records of those included in the sample were examined for presence or absence of a record about smoking in each of the years from 1991 to 1996. If there was more than one record for a given year then the record nearest the mid-year was taken. Recording methods varied in each practice, ranging from one practice where all data were available on the computer to one where only manual records were used. A protocol for each practice was developed, and good inter-rater reliability (k>0.6) achieved before substantive data collection was started. A post hoc inter-rater reliability estimate was calculated for each practice based on examination of 50 records by the two data extractors (TMS and SB). Kappa values were calculated on extraction of data about smoking habit. Levels of agreement were very good, with coefficients ranging from 0.74 to 0.94.

Unpaired data were compared using chi-square, Mann–Whitney, and Student's *t*-test for categorical, ordinal, and interval data respectively. Paired observations were compared using the McNemar test and agreement assessed using Cohen's kappa.[9] Regression techniques were used where more than one variable was associated with an outcome. Descriptive statistics include an adjustment for weighting of the sample by practice size. The study was approved by Leicestershire Health's ethics committee.

Results

After substitution for questionnaires returned by the post office, 2490 names were selected (1233 females and 1232 males, 25 were missing). Numbers from each practice ranged from 337 in the smallest practice (sampling fraction = 0.19) to 703 in the largest (sampling fraction = 0.07).

Overall response was 1906 out of 2490 (76.5%), (practice range = 74.4% to 81.5%, $\chi^2 = 7.71$, d.f. = 4, $P = 0.10$). Response rates were higher for females than males (82.2% versus 71.2%, $\chi^2 = 42.1$, d.f. = 1, $P<0.0001$).

Smoking prevalence from postal questionnaire

Responses to questions about current smoking (defined as one or more per day) and smoking in the five previous years are shown in Table 1. Results show prevalences of smoking higher than UK averages[10] but declining over the six years under scrutiny (χ^2 for trend = 2.46, $P = 0.014$). There was no significant difference in current smoking rates reported by men and women (both 35.2%).

Prevalences of current smoking in the five practices were 26.7%, 28.1%, 32.6%, 38.2%, and 41.6% ($\chi^2 = 28.0$, d.f. = 4, $P<0.0001$). This difference remained significant when adjusted for age and postcode sector.

Smoking prevalence from GP records

Records were available for scrutiny in 2432 (96.7%) of the sample. Questionnaire data on current smoking were available for all those with missing records and showed that their smoking behaviour did not differ from those whose records were examined.

Any entry recording smoking behaviour was extracted from the notes for the years 1991 to 1996. Prevalences calculated from data collected in each of these years and the last entry during the period is shown in Table 2. Over the period 1991 to 1995 (the last full year included), there was a small but statistically significant increase in data collection year on year (χ^2 for trend = 8.65, $P = 0.003$).

A record about smoking between 1991 and 1996 was present in 1784 of the 2432 records examined, i.e. an ascertainment rate of 73.4% (practice range = 67.1% to 81.6%). Women were more likely than men to have smoking status recorded and those with a record of smoking status were older than those with no record, although this only applied to men. Detailed results are shown in Table 3. There was no significant difference between men's and women's smoking rates recorded in the notes (44.4% and 41.8% respectively) when adjusted for sampling fractions.

Comparison of the two data sources

Comparison of prevalence estimates from notes and questionnaire. Our results show that using the last GP record of smoking over six years produces a higher prevalence (42.8%) than was found in the postal survey (35.2% in 1996). If practice data were used to estimate current smoking prevalence they would overestimate it by a factor of 42.8/35.2, i.e. by 1.22. This factor was calculated for each practice and ranged from 1.12 to 1.34. It was not related to ascertainment rates.

The overestimation factor was also calculated for each year's data in predicting current smoking. This ranged from 1.34 (in 1991) to 1.47 (in 1996), with no consistent association with age of data.

Influence of smoking habit on response rate. Non-responders and responders to the questionnaire were compared for smoking data held by the practice over the past six years. Non-responders were more likely to have no record of smoking status (33.9% versus 24.4%) and those with a record were more likely to be smokers (54.1% versus 38.8%). This suggests that the questionnaire data underestimated smoking prevalence. A further estimate of smoking prevalence was made by substituting information from the medical record (if present) when there was no response to the postal questionnaire. Of the 2493 cases, questionnaire data were available for 1906 and GP record data for a further 386, i.e. a total of 2292 (92.1%). Prevalence estimates were: 35.2% from questionnaire alone (n = 1906); and 38.6% from questionnaire supplemented by notes for non-responders (n = 2292).

We conclude that the true prevalence of smoking is between 35.2% and 38.6% (assuming smoking prevalence in those for whom we have no GP or questionnaire data is within this range). Relying solely on data from medical records would overestimate prevalence by a factor of between 1.22 (42.8/35.2) and 1.11 (42.8/38.6).

Analysis of paired data

Where smoking data for an individual over the same period were available in the GP record and from questionnaire response, these were compared for agreement as shown in Table 4.

Levels of agreement were high with an overall kappa coefficient of 0.83, and over 0.8 for all but one year. In Table 5 the last entry in the medical record is compared with current reported smoking. Levels of agreement were less than for individual years producing an overestimation of smoking prevalence of 1.09.

Discussion

A clear assumption in addressing both aims of this study was that individuals would respond truthfully to the postal questionnaire. We feel this is justified by previous work that concluded that self-report by questionnaire is accurate and more reproducible than biochemical markers.[11] The questionnaire we used has been tested in a similar context (i.e. a questionnaire from an academic institution with assurances about confidentiality) and found to correlate well with biochemical markers.[8] The authors' best estimate of deception using this questionnaire was 2.2% of self-

Table 1. Responses from postal questionnaire on previous and current cigarette smoking (n = 1906).

	Did/do you smoke cigarettes regularly?						
	Yes		No		Uncertain		
Year	n	%	n	%	n	%	Missing
1991	710	37.9	1150	61.4	13	0.7	33
1992	703	37.6	1155	61.7	14	0.7	48
1993	687	36.8	1170	62.6	12	0.6	49
1994	687	36.8	1175	62.9	5	0.3	39
1995	675	36.3	1179	63.5	4	0.2	48
Currently (1996)	655	34.4	1251	65.6	0	0	0
Adjusted for sampling fraction[a]	-	35.2	-	64.8	-	-	-

[a]Adjustment to account for larger sampling fraction in smaller practices.

Table 2. Smoking status recorded in GP notes, 1991–1996 (n = 2432).

	Recorded as:		
Year	Smoker (valid %)	Non-smoker (valid %)	No record of smoking (%)
1991	188 (46.2)	219 (53.8)	2025 (83.3)
1992	211 (49.8)	213 (50.2)	2008 (82.6)
1993	307 (41.5)	433 (58.5)	1692 (69.6)
1994	189 (43.2)	249 (56.8)	1994 (82.0)
1995	233 (47.3)	260 (52.7)	1939 (79.7)
1996 (incomplete year)	180 (50.6)	176 (49.4)	2076 (85.4)
Last entry 1991–1996	752 (42.2)	1032 (57.8)	648 (26.6)
Adjusted for sampling fraction[a]	42.8%	57.2%	-

[a]Adjustment to account for larger sampling fraction in smaller practices.

Table 3. Ascertainment of smoking status by age and sex. Adjusted percentages are to account for the larger sampling fraction in smaller practices.

	Record		No record		
Characteristic		Adjusted %		Adjusted %	P-value
Female n (%)	986 (81.8)	80.9	220 (18.2)	-	<0.001[a]
Male n (%)	789 (65.0)	63.5	424 (35.0)	-	
Mean age (years)					
All	43.1	42.9	38.2	38.2	<0.001[b]
Female	42.1	41.9	41.0	41.0	0.44[b]
Male	44.4	44.3	36.8	36.9	<0.001[b]

[a]Chi-square analysis; bunpaired t-test.

declared non-smokers.[12] The accuracy of GP data depends not just on truthfulness but also the other sources of error discussed earlier. Extraction of routinely collected data from the GP record is prone to error because of difficulties in identification and interpretation. While higher levels of inter-rater agreement would have made our estimates more secure, it is unlikely that these could be achieved in analysis of manual records.

Although the response rate to our questionnaire was highly satisfactory, unsuprisingly there was selection bias. As in previous studies, smokers were less likely to respond,[13] as demonstrated by documentation on smoking in the GP record. Using both sources of data we feel secure that the true prevalence of smoking was between 35.2% (from the questionnaire alone) and 38.6% (with GP data added for non-responders, accepting that GP data may overestimate smoking, as discussed below).

If these assumptions are accepted then there is clear evidence from all practices that GP data will overestimate smoking because of ascertainment bias. Although logically this bias will become less when ascertainment is more complete, there is no evidence of this in the range of ascertainments achieved by practices in this study. Using data from only the previous year increased ascertainment bias compared with using the last GP record in the previous five years, probably because the status of smokers is recorded more frequently than non-smokers. Thus, using data for a five-year period produces a better estimate of smoking prevalence, despite the declining prevalence over this period shown in Table 1.

We found significantly different smoking prevalences between populations in the same locality, not explained by age or postcode sector. Consequently, data from one practice may well not represent the population smoking prevalence and so health authorities and researchers should be reluctant to rely on 'spotter' or research network practices, such as the UK General Practice Research Database[14] for this purpose.

How far our results reflect the situation outside our study has to be speculative. It is likely that the distribution of individuals between practices is similar to that in many urban and suburban

Table 4. Agreement between questionnaire (Q) and notes (N) about whether an individual was positive (+) or negative (-) for smoking for years 1991 to 1996 (total of 2188 observations).

Year	+Q +N	-Q -N	+Q -N	-Q +N	Kappa
1991	108	169	16	18	0.77
1992	141	163	19	12	0.81
1993	194	322	25	25	0.81
1994	118	186	13	15	0.82
1995	150	195	7	12	0.89
1996	119	137	9	15	0.83
Total	830	1172	89	97	0.83

Table 5. Contingency table for last smoking record in notes and current smoking status from questionnaire (1398 pairs).

Notes	Questionnaire		Totals
	Smoker	Non-smoker	
Smoker	452	91	543 (38.8%)
Non-smoker	48	807	
Totals	500 (35.8%)		Kappa = 0.79

populations, and so our finding that smoking prevalence varies between practices in such settings may be generalisable. We cannot know whether such differences reflect the selection of practices by individuals or individuals by practices or are the result of practice interventions. In rural locations, where practices may be in a more 'monopolistic' situation, they are more likely to reflect true population prevalence. Similarly, our finding that there is no association between ascertainment rate and validity of smoking prevalence estimates may not hold where ascertainment rates are higher or lower than those found in our practices.

Analysis of paired data (Table 4) shows that the GP record of smoking status is very accurate for individuals and that the few discrepancies appear to be random rather than systematic. Levels of agreement between self-report and GP records were similar to those achieved in the inter-rater reliability test for data extraction and so the apparent discrepancies could be due to inaccuracies in interpretation of the medical record. This should be reassuring to researchers considering using this source of data in epidemiological studies. However, there was evidence that data on individuals decay over time and that using data up to five years old will overestimate prevalence.

The main source of data on smoking prevalence has been the General Household Surveys.[15] Although these provide useful data at national level they do not allow linkage with clinical data and are not precise enough for individual health authorities. GP data have the potential to make a major contribution in tracking progress towards the targets set in *Our Healthier Nation*.[16] We hope that this paper has helped to describe their potential and limitations.

References

1. Mant D, Phillips A. Can the prevelance of disease risk factors be assessed from general practice records? *BMJ* 1986; **292:** 102-104.
2. Haste FM. Value of data provided for health promotion programmes. [Letter.] *BMJ* 1994; **309:** 959.
3. Roberts H, Dengler R, Zamorski A. *Trent Health Lifestyle Survey 1994*. Nottingham: Department of Public Health Medicine and Epidemiology, University of Nottingham, 1995.
4. McCormick A, Fleming D, Charlton J. *Morbidity statistics from general practice. Fourth National Study 1991-1992*. [MB5.] London: HMSO, 1995.
5. Thorogood M, Coulter A, Jones L, *et al*. Factors affecting response to an invitation to attend for a health check. *J Epidemiol Comm Health* 1993; **47:** 224-228.
6. Jamrozik K, Vessey M, Fowler G, *et al*. Controlled trial of three different anti-smoking interventions in general practice. *BMJ* 1984; **288:** 1499-1503.
7. Ozasa K, Watanabe Y, Higashi A, *et al*. Reproducibility of a self-administered questionnaire for dietary habits, smoking, and drinking. [In Japanese.]. *Nippon Eiseigaku Zasshi* 1994; **48(6):** 1048-1057.
8. Tunstall-Pedoe H, Smith WC, Tavendale R. Smoking characteristics and inhalation biochemistry in the Scottish population. *J Clin Epidemiol* 1991; **44(12):** 1405-1410.
9. Landis JR, Kock GG. The measurement of observer agreement for catergorical data. *Biometrics* 1997; **33:** 159-174.
10. Dawe F, Goddard E. *Smoking related behaviour and attitudes. A report on research using the ONS Omnibus Survey produced on behalf of the Department of Health*. London: The Stationery Office, 1997.
11. Petitti DB, Friedman GD, Kahn W. Accuracy of information on smoking habits provided on self-administered research questionnaires. *Am J Pub Health* 1981; **71:** 308-311.
12. Woodward M, Tunstall-Pedoe H. An iterative technique for identifying smoking deceivers with application to the Scottish Heart Study. *Preventive Medicine* 1992; **21(1):** 88-97.
13. Sheikh K, Mattingley S. Investigating non-response bias in mail surveys. *J Epidemiol Comm Health* 1981; **35:** 283-286.
14. Walley T, Mantagin A. The UK General Practice Research Database. *Lancet* 1997; **350:** 1097-1099.
15. *OPCS General Household Survey, 1992*. [Series GHS No23.] London: HMSO, 1994.
16. Department of Health. *Saving lives: our healthier nation.* [CM 4385.] London: HMSO, 1999.

Acknowledgements

We are grateful to participating practices for allowing access to their data and to Leicestershire Health Authority for their help in developing the sampling frame. The study was funded by NHS Trent Research Scheme. Thanks also to Susan Brunskill for help in data collection and to Tim Coleman for his comments on an earlier draft.

Address for correspondence

Dr A Wilson, Department of General Practice and Primary Health Care, University of Leicester, Gwendolen Road, Leicester LE5 4PW. E-mail: aw7@le.ac.uk