EVIDENCE BASED PUBLIC HEALTH POLICY AND PRACTICE

Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study

C Navarro, M D Chirlaque, M J Tormo, D Pérez-Flores, M Rodríguez-Barranco, A Sánchez-Villegas, A Agudo, G Pera, P Amiano, M Dorronsoro, N Larrañaga, J R Quirós, E Ardanaz, A Barricarte, C Martínez, M J Sánchez, A Berenguer, C A González

J Epidemiol Community Health 2006;**60**:593–599. doi: 10.1136/jech.2005.039131

See end of article for authors' affiliations

Correspondence to:
Dr C Navarro, Servicio de Epidemiología, Consejería de Sanidad, Ronda de Levante 11, E- 30008 Murcia, Spain; carmen.navarro@carm.es

Accepted for publication
16 January 2006

Introduction: This study aims to assess the validity of self reported diagnoses of cancer by persons recruited for the Spanish EPIC (European prospective investigation into cancer and nutrition) cohort study and to identify variables associated with correctly reporting a diagnosis of cancer.

Methods: 41 440 members of EPIC were asked at the time of recruitment whether they had been diagnosed with cancer and the year of diagnosis and site. The process of validating self reported diagnoses of cancer included comparison of the cohort database with the data from the population based cancer registries. Cancer diagnostic validity tests were calculated. The association between a correct report and certain sociodemographic, tumour related, or health related variables were analysed by logistic regression.

Results: The overall sensitivity of self reported diagnoses of cancer is low (57.5%; 95% CI: 51.9 to 63.0), the highest values being shown by persons with a higher level of education or with a family history of cancer and the lowest values by smokers. Breast and thyroid cancers are those with the highest diagnostic validity and uterus, bladder, and colon-rectum those with the lowest. In both sexes the variables showing a significant association with a correct report of cancer are: higher education level, number of previous pathologies, invasive tumour, and, in women, a history of gynaecological surgery.

Conclusions: The overall sensitivity of self reported diagnoses of cancer is comparatively low and it is not recommended in epidemiological studies for identifying tumours. However, self reported diagnoses might be highly valid for certain tumour sites, malignant behaviour, and average to high levels of education.

Validation of self reported diagnoses of cancer by means of a questionnaire has been undertaken extensively in epidemiological research, usually evaluating the utility of answers on family histories of cancer,¹ screening tests,² and personal histories of cancer.³ However, in cohort studies where the incidence of cancer is the end point, the aim of these studies is twofold: they are used in the recruitment stage to exclude persons who already have cancer (prevalent cases) and serve during follow up to identify incident tumours.^{4–5}

Although for many years studies have been conducted on the agreement between health data from questionnaires and those provided by medical records,^{6–10} not so many have been used to validate questions relating to cancer^{11–15} and even fewer to compare questionnaire data against data from the population based cancer registries that cover the areas of study.^{16–19} A wide difference has been seen in the ability to self report cancer across countries; thus, the highest valuable information has been reported in USA^{13–19} and the lowest in Japan.³ In Europe, sensitivity ranged between 67% and 82%.^{16–18}

The main factors associated with a correctly self reported diagnosis of cancer that are consistently given in the literature are age, sex, and level of education, although other associated variables have been found such as tobacco smoking, site, and behaviour of the tumour or time between diagnosis and interview. The differences seen among countries have been partly attributed to socio-cultural determinants and the variability in the practice by medical practitioners.^{3, 17–20}

Studies on the validity of self reported diagnoses of cancer are scarce in countries outside the USA and northern Europe,

and this is the first study conducted in a southern European country. The aims of the study are to estimate the validity of self reported diagnoses of cancer at enrolment (prevalent cases) by persons recruited for the cohort of the EPIC-Spain study and to identify the variables associated with correctly reported diagnoses.

METHODS

EPIC-Spain cohort

The Spanish EPIC (European prospective investigation into cancer and nutrition) cohort forms part of a larger cohort recruited in 10 European countries to study the relation between diet, cancer, and other chronic diseases.^{21–22} In Spain it is made up of 41 440 people recruited between 1992 and 1996. For this study on the validation of self reported diagnoses of cancer we have excluded a total of 202 cases of cancer, most of them (124 cases) for being non-melanocytic skin cancers, 74 cases because they self reported cancer at a time before the introduction of the corresponding cancer registries and four cases for being multiple primary tumours. Thus the final analysis includes 41 238 persons aged from 29 to 69 years.

Participation from five regions: three from the north (Asturias, Guipuzkoa, and Navarra) and two from the south (Granada and Murcia).²³ The participants are healthy volunteers, mainly blood donors, who received a letter of invitation and agreed to participate. At the time of recruitment each person gave information on their dietary habits together with their anthropometric measurements and a blood sample.²⁴ In addition they were given a questionnaire on lifestyles and other non-dietary factors, including tobacco smoking, physical activity, level of education, medical history

Table 1 2×2 table used in the calculation of validity indicators

		Cancer registry report	
		Yes	No
Cancer self reported in the questionnaire	Yes	True positives (TP)	False positives (FP)
	No	False negatives (FN)	True negatives (TN)

(heart attack, diabetes, cancer, hypertension, etc), and family history of cancer. The section on personal medical history included the question: "Has a doctor ever told you that you suffer or have suffered from cancer?" If so, they were to indicate the age they were when it began and specify, with an open response, the site.

The persons who agreed to participate in the study gave their informed consent. They were informed that all the information given was confidential and that the databases were registered at the Spanish Data Protection Authority as stipulated by law.

Cancer registries

The five regions in which the EPIC-Spain study is conducted have population based cancer registries,²⁵ form part of the European Network of Cancer Registries,²⁶ publish their data regularly in the series *Cancer Incidence in Five Continents* by the International Agency for Research on Cancer,²⁷ and began operating between 1970 (Navarra) and 1986 (Basque country). Registrable cases are all new malignant tumours, including in situ carcinomas, in any anatomical site, except non-melanoma skin tumours in Asturias and the Basque country. The *International Classification of Diseases for Oncology*, 2nd edition,²⁸ was used to codify the anatomical site and morphology of tumours according to the information obtained from the questionnaire.

Record linkage of self reported cases with registered cases

The history of cancer reported in the interviews was compared with the data from the cancer registries in the five regions, with the validation period ranging from 6 to 26 years before the date of interview (median: 10 years) depending on the date of introduction of the registries.

The record linkage method is a widely contrasted procedure^{29–31} used by disease registries to validate information, detect cases and in cohort studies establish the presence or absence of disease and the person/years at risk.³² With this aim a software application was developed to perform a probabilistic linkage between the data from the EPIC study and those from the population cancer registries of the five participating communities. The first stage of the record linkage process consists in the specific normalisation of each of the variables that is going to form part of the probabilistic linkage. In the second stage a duplicate control generates a table containing the persons where some or all of the cross-variables match. The programme then performs the probabilistic linkage (EPIC-Link), assigning different values to the persons in the EPIC study depending on the similarity of

the cross-variables to those contained in the cancer registries. According to data available from other studies,^{33–34} the simulations carried out and the calculations of the odds of a random match between two person, scores are assigned to the most important identification variables. After multiple tests to determine the cut off point in the score beyond which the record linkage results should be reviewed, one was established that would maintain a high positive predictive value (PPV) without losing a high sensitivity. The cases finally detected as possible matches were rigorously inspected with a view to their confirmation. Other variables that were available in both sources were also used.

Clinical records of reported cases not linked with the cancer database were reviewed in each region before accepting a case as a false positive. If it really was a true positive case missed by the cancer registry it was added to the cancer registry and accounted as a true positive self reported case.

Study variables

The sociodemographic variables included in the analysis are sex, age (29–44, 45–54, and 55–69 years), and maximum level of education attained (incomplete primary school, primary, and secondary or further education). The associations with disease related variables were analysed, such as age at the time of diagnosis, year of diagnosis (≤ 1989 , ≥ 1990), number of years between questionnaire and diagnosis (defined as the interval between the date of the cancer diagnosis, and the date of the baseline questionnaire and grouped into four categories: 0–2, 3–4, 5–7, and >7 years), level of tumour infiltration (infiltrating or in situ), and the basis of diagnosis (microscopical or clinical confirmation). Also included were variables related to health status and lifestyles, such as tobacco smoking, family history of cancer, gynaecological surgeries in women, number of previous pathologies (myocardial infarction, angina, ictus, other circulatory problems, hypertension, hyperlipaemia, diabetes, kidney stones or gallstones, polyps, peptic ulcer, asthma, and urinary infection), being a blood donor, and body mass index (BMI).

Analysis of data

Sensitivity, specificity, PPV, and negative predictive value (NPV) with their corresponding 95% confidence intervals were calculated using the reports from the cancer registry as a standard reference (table 1). Sensitivity was calculated as the proportion of persons with a cancer report in a registry who also self reported the cancer; specificity as the proportion of persons not found in the registry who did not self report

Table 2 2×2 table used in the calculation of validity indicators by site. For example, breast cancer

		Cancer registry report: indicates breast cancer	
		Yes	No
Breast cancer self reported in the questionnaire	Yes	Breast cancer self reported and registered (TP)	Other site confirmed or cancer not confirmed (FP)
	No	Other site reported or cancer not reported (FN)	No cancer self reported and no case registered (TN)

Table 3 Sociodemographic characteristics and variables related to health status according to whether persons of the EPIC-Spain cohort self report cancer

Variable	Self reported cancer Yes		Self reported cancer No	
	n = 260	%	n = 40978	%
Age (years)				
29–44	52	20.0	13289	32.4
45–54	121	46.5	16086	39.3
55–69	87	33.5	11603	28.3
Sex				
Male	55	21.2	15510	37.8
Female	205	78.8	25468	62.2
Level of education				
Incomplete primary school	88	33.8	14119	34.7
Primary	122	46.9	19225	47.2
Secondary/further education	50	19.2	7356	18.1
Tobacco				
Smoker	51	19.7	11026	26.9
Former smoker	53	20.5	7233	17.7
Non-smoker	155	59.8	22695	55.4
Family history of cancer				
Yes	39	15.0	4200	10.2
No	221	85.0	36778	89.8
Gynaecological surgery in women				
Yes	156	76.1	5659	22.2
No	49	26.4	19809	77.8
Blood donors				
Yes	168	64.6	30942	75.5
No	92	35.4	10036	24.5

cancer. The PPV was calculated as the proportion of persons with a self reported cancer during the years of registry operation who had a matching cancer report in the registry; and NPV as the percentage of persons who denied having cancer and whose name did not appear in the cancer registry. The analysis was done for all the persons overall and, stratified, by sociodemographic variables, lifestyles, and health status. Certain variables had missing values (tobacco smoking 0.1% of the cohort and level of education 0.7%). Excluded were non-melanocytic skin tumours, for not being

recorded in all the cancer registries, and multiple primary tumours.

Also analysed were the validity indicators according to tumour site, including as false negatives any persons identified in the cancer registries not reporting cancer or reporting it in a different site to that analysed, and as false positives any non-confirmed cancers or cancers confirmed in another site (table 2).

A multivariate logistic regression analysis was used to identify variables associated with a correctly self reported

Table 4 Indicators of diagnostic validity and 95% confidence intervals (CI) of self reported prevalent cancer. EPIC-Spain cohort

	Number	True positives	False negatives	False positives	Sensitivity	(CI)	Specificity	(CI)	PPV	(CI)	NPV	(CI)
Overall	41238	184	136	76	57.5	(51.9 to 63.0)	99.8	(99.8 to 99.9)	70.8	(64.8 to 76.2)	99.7	(99.6 to 99.7)
Sex												
Male	15565	45	48	10	48.4	(37.9 to 59.0)	99.9	(99.9 to 100.0)	81.8	(69.1 to 90.9)	99.7	(99.6 to 99.8)
Female	25673	139	88	66	61.2	(54.6 to 67.6)	99.7	(99.7 to 99.8)	67.8	(60.9 to 74.2)	99.7	(99.6 to 99.7)
Age (years)												
29–44	13341	32	24	20	57.1	(43.2 to 70.2)	99.9	(99.8 to 99.9)	61.5	(47.1 to 74.7)	99.8	(99.7 to 99.9)
45–54	16207	91	48	30	65.5	(60.0 to 73.3)	99.8	(99.7 to 99.9)	75.2	(66.6 to 82.6)	99.7	(99.6 to 99.8)
55–69	11690	61	64	26	48.8	(39.8 to 57.9)	99.8	(99.7 to 99.9)	70.1	(59.4 to 79.5)	99.5	(99.3 to 99.6)
Level of education												
Incomplete primary school	14207	67	70	21	48.9	(40.3 to 57.6)	99.9	(99.8 to 99.9)	76.1	(65.9 to 84.6)	99.5	(99.4 to 99.6)
Primary	19347	80	51	42	61.1	(52.2 to 69.4)	99.8	(99.7 to 99.8)	65.6	(56.4 to 73.9)	99.7	(99.7 to 99.8)
Secondary/further education	7406	37	13	13	74.0	(59.7 to 84.4)	99.8	(99.7 to 99.9)	74.0	(59.7 to 85.4)	99.8	(99.7 to 99.9)
Tobacco												
Smoker	10094	28	37	19	43.1	(30.9 to 56.0)	99.8	(99.7 to 99.9)	59.6	(44.3 to 73.6)	99.6	(99.5 to 99.7)
Former smoker	8269	47	28	10	62.7	(50.7 to 73.3)	99.9	(99.8 to 99.9)	82.5	(70.1 to 91.3)	99.7	(99.5 to 99.8)
Non-smoker	22850	108	71	47	60.3	(52.8 to 67.6)	99.8	(99.7 to 99.9)	69.7	(61.8 to 76.8)	99.7	(99.6 to 99.8)
Family history of cancer												
Yes	4239	28	13	11	68.3	(51.9 to 81.9)	99.7	(99.5 to 99.9)	71.8	(55.1 to 85.0)	99.7	(99.5 to 99.8)
No	36999	156	123	65	55.9	(49.9 to 61.8)	99.8	(99.8 to 99.9)	70.6	(64.1 to 76.5)	99.7	(99.6 to 99.7)
Gynaecological surgery in women												
Yes	5815	114	58	42	66.3	(58.7 to 73.3)	99.3	(99.0 to 99.5)	73.1	(65.4 to 79.9)	99.0	(98.7 to 99.2)
No	19858	25	30	24	66.3	(65.1 to 67.5)	99.3	(99.1 to 99.5)	73.1	(0.72 to 73.7)	99.0	(98.7 to 99.3)
Blood donors												
Yes	31110	109	89	59	65.1	(64.6 to 65.6)	99.8	(99.8 to 99.9)	81.5	(81.1 to 81.7)	99.5	(99.4 to 99.6)
No	10128	75	47	17	61.5	(52.2 to 70.1)	99.8	(99.7 to 99.9)	81.5	(72.1 to 88.6)	99.5	(99.4 to 99.7)

PPV, positive predictive value; NPV, negative predictive value.

Table 5 Indicators of diagnostic validity and 95% confidence intervals (CI) of self reported prevalent cancer by tumour site. EPIC-Spain cohort (n = 41238)

	n (%) ^a	True positives	False negatives	False positives	Sensitivity (CI)	Specificity (CI)	VPP (CI)	VPN (CI)
Breast	84 (26.3)	71	13 (9)	15 (13)	84.5 (75.0 to 91.5)	100.0 (99.9 to 100.0)	82.6 (72.9 to 89.9)	100.0 (100.0 to 100.0)
Oral cavity	17 (5.3)	4	13 (9)	6 (3)	23.5 (6.8 to 49.9)	100.0 (100.0 to 100.0)	40.0 (12.2 to 73.8)	100.0 (100.0 to 100.0)
Stomach	12 (3.8)	6	6 (3)	0	50.0 (21.1 to 78.9)	100.0 (100.0 to 100.0)	100.0 (54.1 to 100.0)	100.0 (100.0 to 100.0)
Colorectal	23 (7.2)	4	19 (18)	1 (1)	17.4 (5.0 to 38.8)	100.0 (100.0 to 100.0)	80.0 (28.4 to 99.5)	100.0 (99.9 to 100.0)
Uteri SAI	74 (23.1)	23	51 (46)	32 (27)	31.1 (20.8 to 42.9)	99.9 (99.9 to 100.0)	41.8 (28.7 to 55.9)	99.9 (99.8 to 99.9)
Cervix uteri	53 (16.6)	7	46 (33)	7 (5)	13.2 (5.5 to 25.3)	100.0 (100.0 to 100.0)	50.0 (23.0 to 77.0)	99.9 (99.9 to 99.9)
Corpus uteri	20 (6.3)	3	17 (13)	5 (5)	15.0 (3.2 to 37.9)	100.0 (100.0 to 100.0)	37.5 (8.5 to 75.5)	100.0 (99.9 to 100.0)
Bladder	23 (7.2)	5	18 (17)	0	21.7 (7.5 to 43.7)	100.0 (100.0 to 100.0)	100.0 (47.8 to 100.0)	100.0 (99.9 to 100.0)
Thyroid	21 (6.6)	13	8 (8)	2 (2)	61.9 (38.4 to 81.9)	100.0 (100.0 to 100.0)	86.7 (59.5 to 98.3)	100.0 (100.0 to 100.0)
Lung	8 (2.5)	4	4 (3)	3 (2)	50.0 (15.7 to 84.3)	100.0 (100.0 to 100.0)	57.1 (18.4 to 90.1)	100.0 (100.0 to 100.0)

^aNumber of confirmed cases and percentage of the total. False negatives = another site reported + cancer not reported (in parentheses the number not reporting cancer). False positives = another site confirmed + cancer not confirmed (in parentheses the number with non-confirmed cancer).

Table 6 Variables associated with a correctly reported cancer in the members of the EPIC-Spain cohort (cases of cancer included)

n = 320	Number	ORa* (95% CI)
Sex		
Male	93	1
Female	227	2.04 (0.99 to 4.19)
Age (years)		
29–44	56	1
45–54	139	1.71 (0.66 to 4.47)
55–69	125	1.62 (0.43 to 6.13)
Level of education		
Incomplete primary school	137	1
Primary	131	2.03 (1.13 to 3.65)
Secondary/further education	50	3.82 (1.54 to 9.48)
Tobacco		
Non-smoker	179	1
Former smoker	75	0.89 (0.39 to 2.02)
Smoker	65	0.49 (0.23 to 1.05)
Family history of cancer		
No	41	1
Yes	279	1.23 (0.55 to 2.77)
Gynaecological surgery in women		
No	55	1
Yes	172	2.64 (1.22 to 5.68)
Age at diagnosis		
<45	109	1
45–54	131	0.51 (0.21 to 1.26)
>54	80	0.39 (0.11 to 1.46)
Years between questionnaire and diagnosis		
0–2	132	1
3–4	79	0.86 (0.44 to 1.68)
5–7	62	1.16 (0.25 to 5.45)
>7	47	0.70 (0.12 to 4.12)
Year of diagnosis		
≤ 1989	104	1
≥ 1990	216	2.06 (0.46 to 9.26)
Number of pathologies		
0	105	1
1	111	0.50 (0.27 to 0.95)
2	65	1.02 (0.48 to 2.18)
≥ 3	39	0.44 (0.18 to 1.08)
Tumour behaviour		
In situ	36	1
Infiltrating	280	3.59 (1.45 to 8.90)
Blood donor		
No	122	1
Yes	198	0.64 (0.37 to 1.18)
Microscopical confirmation		
No	7	1
Yes	313	0.21 (0.20 to 2.27)
Body mass index		
Underweight and normal (BMI <25)	58	1
Overweight and obesity (BMI ≥25)	260	1.01 (0.48 to 2.15)

*ORa, odds ratio adjusted for the rest of the variables (sex, age, level of education, tobacco smoking, family history of cancer, surgical operations, age at diagnosis, years between questionnaire and diagnosis, year of diagnosis, number of pathologies, tumour behaviour, blood donor, microscopical confirmation, BMI).

history of cancer using the crude and adjusted odds ratio including the 95% confidence interval.

RESULTS

The final analysis included 41 238 persons. Of the 260 persons who reported a cancer, 184 were confirmed as malignant tumours by the population cancer registries, 177 by record linkage, and seven after reviewing the clinical record of those not linked. A hundred and thirty six persons not self reporting a diagnosis of cancer were identified with cancer by the registries, accounting for a 43% rate of false negatives in this population.

Table 3 shows the main characteristics of the persons who did or did not self report a diagnosis of cancer. Of note is the older age, higher percentage of women, family history of cancer, and, in women only, gynaecological operations among the persons who correctly self reported cancer, whereas in the same group tobacco smoking or being a blood donor was less common.

The diagnoses of tumour confirmed by the cancer registry with regard to the total number of cases reported (table 4) give an overall sensitivity of 57.5% (95% CI: 51.9 to 63.0) and a specificity of 99.8% (95% CI: 99.8 to 99.9). The total positive predictive value was 70.8% (95% CI: 64.8 to 76.2) and negative predictive value 99.7% (95% CI: 99.6 to 99.7). The highest sensitivity values are shown by persons with a higher level of education (74.0; 95% CI: 59.7 to 84.4) or with a family history of cancer and the lowest values by smokers (43.1; 95% CI: 30.9 to 56.0). Men show one of the lowest sensitivity values (48.4%; 95% CI: 37.9 to 59.0) but a high positive predictive value of 81.8% (95% CI: 64.8 to 76.2).

Table 5 shows the validation analysis taking into account the topography of the tumour. Breast cancer is seen to present the highest sensitivity (84.5; 95% CI 75.0 to 91.5), followed by thyroid cancer (61.9; 95% CI 38.4 to 81.9); cancer of the cervix uteri and corpus uterus, bladder cancer and colorectal cancer have the lowest sensitivity, with very low values of between 13% and 17%. When the cervix uteri, corpus uteri, and uterus with no other specification are grouped in a single site, sensitivity improves slightly.

Identification of variables associated with a correctly reported history of cancer was done including the 320 cases of cancer identified in the cancer registries (true positives and false negatives). The univariate analysis shows a higher frequency in self reporting correctly and being a woman, having a higher level of education, having had the diagnosis more recently, presenting with an infiltrating cancer and, in the case of women, having a history of gynaecological surgery. Many variables lose their significance in the adjusted analysis, but higher level of education (with an adjusted OR of 6.41; 95% CI 1.98 to 20.82), infiltrating nature of the tumour (OR_{adjusted}: 3.15; 95% CI: 1.17 to 8.52) and, in the

case of women, history of gynaecological surgery (OR_{adjusted}: 2.68; 95% CI 1.10 to 6.52) are maintained. Being a woman is very close to statistical significance (table 6).

DISCUSSION

The likelihood of a correctly self reported cancer in this study is not very high (58%) and in the range of other studies.¹⁶⁻¹⁸ However, big differences have been seen among countries: in USA, Bergmann *et al* report in the cancer prevention study the highest overall sensitivities¹⁹ (79% for an exact agreement of site and year of diagnosis (± 1 year) and 93% for any diagnosis of cancer) and in the California teachers study have been reported values between 70% and 96% depending on site.²⁰ Only in Connecticut were lower values found, but this study has a community based design and the former were done on selected groups in cancer prevention studies.¹⁷ In Japan³ (36% for any type of cancer) and France⁵ (21% in persons aged over 75 years) they are much lower. In Europe, our study—the first in a Mediterranean country—shows lower sensitivity than the two previously published in north European countries.¹⁶⁻¹⁸ On the other hand, all the studies published show very high specificities, where the presence of cancer can be ruled out with relative certainty if the person denies having received a medical diagnosis of the disease. In other words the persons in this study tend to underreport rather than overreport a history of cancer, which is why estimated specificity and NPV agree with those of other studies that report values of over 90%.¹⁶⁻¹⁸

The high sensitivity encountered in the Spanish EPIC cohort for breast and thyroid cancer (85% and 62% respectively) coincides with the sites that present a high sensitivity in the California teachers study²⁰ (96% and 93% respectively), although the values in the Spanish EPIC cohort are considerably lower; they also coincide in that cervical and endometrial cancer are those with the lowest sensitivity. The JPHC study cohort³ finds a high sensitivity for breast (81%) and low sensitivity for uterus (42%) and colon-rectum (14%); the latter site is similar in the EPIC study (17%), the same as in the cancer prevention study 2 nutrition survey,¹⁹ with a 91% sensitivity for breast and 16% for colon-rectum.

It has been suggested that cancers with very clear cut diagnostic criteria, such as breast and thyroid cancer, are more likely to be reported than cancers that have more ambiguous diagnostic procedures.²⁰ Chambers *et al* suggested that reporting might be lower for cancers that have a large proportion of less severe histological types, such as cervical cancer.³⁵

In our study, older age at diagnosis and higher number of years between diagnosis and date of interview were associated with self reporting incorrectly although these associations were not statistically significant, possibly because of the relatively low number of cases. Desai *et al*¹⁷ also find that old age and longer time between diagnosis and interview are associated with self reporting incorrectly. Likewise, in the study by Parikh-Patel *et al*²⁰ old age reduces the validity of the diagnosis. In one of the cohorts of the EPIC-Sweden study Manjer *et al*¹⁸ find that the overreporting or underreporting of cancer occurs more often among elderly persons. In our study women report better than men, although this association is not statistically significant. Sex also influences the report of tumours in other studies as occurs in the Swedish EPIC cohort,¹⁸ but in this case a higher frequency of overreporting is seen among women.

We found that invasive tumours are self reported better than those in situ. Parikh-Patel *et al*²⁰ report an increase in sensitivity when tumours are invasive (98.1%) or not in situ (87.8%). In women, a history of gynaecological surgery is usually associated with tumour related operations, which is

What is already known on this subject

- A few studies have been carried out on the validity of self reported diagnoses of cancer that use population based cancer registries as a reference method.
- Most of the studies show acceptable or high sensitivity depending on the tumour site and very high specificity.
- The main factors associated with a correctly self reported diagnosis of cancer are age, sex, and level of education
- Self reported diagnosis of breast cancer has a very high validity

What this study adds

- First study conducted in a Mediterranean country.
- The sensitivity of self reported diagnoses of cancer in this cohort is comparatively low and it is not recommended in epidemiological studies for identifying tumours.
- For certain tumour sites, malignant behaviour, and medium to high levels of education a self reported diagnosis may have a high validity

why the result here gives greater validity to the information analysed.

The most important factor associated with a correct self report is level of education. This result is in agreement with other studies.^{3 5 19} It seems that the most plausible reasons for the wide differences seen among countries are sociocultural factors and variability in medical practice. This is in concordance with the finding that patients are more often informed about a cancer diagnosis in the USA and Scandinavia, than in southern Europe and Asia.¹⁸ Recent studies³⁶ in oncological patients that have inquired into a real knowledge of their process have shown that only a third of the patients know for certain, whereas another third do not know but suspect, and the other third are completely unaware. A study carried out in Spain showed that perceived intelligence and emotional control in the patients were the best predictors of the decision by doctors to give information. Age and socioeconomic status were also significantly associated with the doctors' information giving practices.³⁷

One limitation of the study is the external validity of the results, as although sensitivity and specificity are tests of internal validity, the predictive values are indeed influenced by the selection of the population. Although in this cohort study the participating persons are the general population, they were not selected as representative of this population. However, as they do belong to different Spanish regions, include a large number of people of different levels of education, reside in rural and urban areas, and present a smoking habit distribution similar to the general population, the cohort does not show any notable differences when compared with Spain's adult population.³⁸

The population cancer registries included in this study show good indicators of quality²⁷ with a high coverage. There are no self reported cases in which it was not possible to ascertain if they were tumours or not, as whenever there was doubt the patient's medical record was consulted or more information on the tumour was requested.

Incompleteness of the cancer registries could bias the results in two ways. First of all, increasing false positives, particularly for cases with date of incidence in the beginning of launching each registry. We think that this bias is unlikely because there are no self reported cases in which it was not possible to ascertain if they were tumours or not. For all self reported cases not linked with the cancer registry database the patient's medical record was consulted or more information on the tumour was requested. Incidence date for the cases lost to registration was distributed along the study period, between 1987 and 1995. Incompleteness also could underestimate sensitivity. Seven of 184 true positive cases were lost to registration by cancer registries. If the proportion of cases lost to registration among false negative cases were the same, the completeness of registration for all registries together would be 96%. Thus, the true sensitivity would be slightly lower. Other possible bias could be originated because we have not validated self reported tumours

diagnosed before cancer registries were started. As for 85% of self reported cases the years between questionnaire and diagnosis was less than eight (table 6) and the overlap between the cancer database and the reported incidence date is 16 years this bias is much unlikely to have occurred.

Another problem is the possibility that some persons not self reporting a malignant tumour were recorded as cases in the cancer registries and not detected by record linkage. But the EPIC databases are annually cross checked by record linkage with the data from the corresponding cancer registries and national population death database, and the rate of follow up of this cohort is very high, with less than 2% lost to follow up.³⁸

It can therefore be concluded that the validity of self reported diagnoses of cancer in this large Spanish cohort is comparatively low, which is the reason why the overall information provided by the participants cannot be used individually for tumour detection. Population based cancer registries seem to be essential for ascertaining cancer cases in cohort studies carried out in Spain. However, for certain tumour sites, malignant behaviour, and medium to high levels of education a self reported diagnosis may have a high validity.

Authors' affiliations

C Navarro, M D Chirlaque, M J Tormo, M Rodríguez-Barranco, Department of Epidemiology, Murcia Health Council, Murcia, Spain

D Pérez-Flores, Department of Socio-Medical Sciences, Faculty of Medicine, Murcia University, Spain

A Sánchez-Villegas, Department of Clinical Sciences, University of Las Palmas de Gran Canaria, Spain

A Agudo, G Pera, A Berenguer, C A González, Catalan Institute of Oncology, Barcelona, Spain

P Amiano, M Dorronsoro, N Larrañaga, Department of Public Health of Guipuzkoa, San Sebastián, Spain

J R Quirós, Consejería de Salud y Servicios Sanitarios del Principado de Asturias, Oviedo, Spain

E Ardanaz, A Barricarte, Public Health Institute of Navarra, Pamplona, Spain

C Martínez, M J Sánchez, Andalusian School of Public Health, Granada, Spain

Funding: the project was financed by the European Union Europe Against Cancer Programme, the Spanish Health Research Fund (project no 99/0024) and the Spanish Autonomous Communities and participating Institutions. Some centres receive aid from the ISCIII Network RCESP (C03/09).

Competing interests: none.

REFERENCES

- 1 **Kadan-Lottick NS,** Friedman DL, Mertens AC, *et al.* Self-reported family history of cancer: the utility of probing questions. *Epidemiology* 2003;**14**:737-40.
- 2 **Madlensky L,** McLaughlin J, Goel V. A comparison of self-reported colorectal cancer screening with medical records. *Cancer Epidemiol Biomarkers Prev* 2003;**12**:656-9.
- 3 **Yoshinaga A,** Sasaki S, Tsugane S. Sensitivity of self-reports of cancer in a population-based prospective study: JPHC study cohort I. *J Clin Epidemiol* 2001;**54**:741-6.
- 4 **Kato I,** Toniolo P, Koenig KL, *et al.* Comparison of active and cancer registry-based follow-up for breast cancer in a prospective cohort study. *Am J Epidemiol* 1999;**149**:372-8.
- 5 **Berthier F,** Grosclaude P, Bocquet H, *et al.* Prevalence of cancer in the elderly: discrepancies between self-reported and registry data. *Br J Cancer* 1997;**75**:445-7.
- 6 **Harlow SD,** Linet MS. Agreement between questionnaire data and medical records. The evidence for accuracy of recall. *Am J Epidemiol* 1989;**129**:233-48.
- 7 **de Sanjose LS,** Anto Boque JM, Caballero JA. A comparison of the information obtained in a health survey by interview with primary care registries. *Gac Sanit* 1991;**5**:260-4.
- 8 **Haapanen N,** Mäkelä P, Pasanen M, *et al.* Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. *Am J Epidemiol* 1997;**145**:762-9.
- 9 **Burns RB,** Moskowitz MA, Ash A, *et al.* Self-report versus medical record functional status. *Med Care* 1992;**30**:85-95.

- 10 **Fowles JB**, Fowler EJ, Craft C. Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. *J Ambul Care Manage* 1998;**21**:24–34.
- 11 **Aitken J**, Bain C, Ward M, et al. How accurate is self-reported family history of colorectal cancer? *Am J Epidemiol* 1995;**141**:863–71.
- 12 **Bergmann MM**, Byers T, Freedman DS, et al. Validity of self-reported diagnoses leading to hospitalization: a comparison of self-reports with hospital records in a prospective study of American adults. *Am J Epidemiol* 1998;**147**:969–77.
- 13 **Colditz GA**, Martin P, Stampfer MJ, et al. Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *Am J Epidemiol* 1986;**123**:894–900.
- 14 **Douglas FS**, O'Dair LC, Robinson M, et al. The accuracy of diagnoses as reported in families with cancer: a retrospective study. *J Med Genet* 1999;**36**:309–12.
- 15 **Zhu K**, McKnight B, Stergachis A, et al. Comparison of self-report data and medical records data: results from a case-control study on prostate cancer. *Int J Epidemiol* 1999;**28**:409–17.
- 16 **Schrijvers CT**, Stronks K, van de Mheen DH, et al. Validation of cancer prevalence data from a postal survey by comparison with cancer registry records. *Am J Epidemiol* 1994;**139**:408–14.
- 17 **Desai MM**, Bruce ML, Desai RA, et al. Validity of self-reported cancer history: a comparison of health interview data and cancer registry records. *Am J Epidemiol* 2001;**153**:299–306.
- 18 **Manjer J**, Merlo J, Berglund G. Validity of self-reported information on cancer: determinants of under- and over-reporting. *Eur J Epidemiol* 2004;**19**:239–47.
- 19 **Bergmann MM**, Calle EE, Mervis CA, et al. Validity of self-reported cancers in a prospective cohort study in comparison with data from state cancer registries. *Am J Epidemiol* 1998;**147**:556–62.
- 20 **Parikh-Patel A**, Allen M, Wright WE, et al. Validation of self-reported cancers in the California teachers study. *Am J Epidemiol* 2003;**157**:539–45.
- 21 **Riboli E**, Kaaks R. The EPIC Project: rationale and study design. European prospective investigation into cancer and nutrition. *Int J Epidemiol* 1997;**26**(suppl 1):S6–14.
- 22 **Riboli E**. The European prospective investigation into cancer and nutrition (EPIC): plans and progress. *J Nutr* 2001;**131**:170–5S.
- 23 **Gonzalez CA**, Navarro C, Martinez C, et al. El Estudio Prospectivo Europeo sobre Cáncer y Nutrición (EPIC). *Rev Esp Salud Publica* 2004;**78**:167–76.
- 24 **Navarro C**, Tormo MJ, Chirlaque MD. Evaluación del estado nutricional de la cohorte EPIC-Murcia. Murcia: Consejería de Sanidad y Consumo, Dirección General de Salud. *Pública Serie Informes*, 1999;**27**:175–96 (<http://www.murciasalud.es/recursos/ficheros/1483-EvaluacNutricionalEpicMurcia-1.pdf>).
- 25 **Miñarro R**, Black RJ, Martinez C, et al. *Cancer incidence and mortality in Spain. Patterns and trends*. Lyon: IARC, 2000:15.
- 26 **Tyczynski JE**, Demaret E, Parkin DM. *Standards and guidelines for cancer registration in Europe*. Lyon: IARC, 2003:89.
- 27 **Parkin DM**, Whelan SL, Ferlay J, et al. *Cancer incidence in five continents*. Vol 8. Lyon: IARC, 2002:424–39.
- 28 **Percy C**, Van Holten V, Muir C. *International classification of diseases for oncology*. 2nd ed. Ginebra: World Health Organisation, 1990.
- 29 **Cook LJ**, Olson LM, Dean JM. Probabilistic record linkage: relationships between file sizes, identifiers and match weights. *Methods Inf Med* 2001;**40**:196–203.
- 30 **Fair M**, Cyr M, Allen AC, et al. An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. The Fetal and Infant Health Study Group. *Chronic Dis Can* 2000;**21**:8–13.
- 31 **Storm HH**. Completeness of cancer registration in Denmark 1943–1966 and efficacy of record linkage procedures. *Int J Epidemiol* 1988;**17**:44–9.
- 32 **Dal Maso L**, Zanetti R, Orengo MA, et al. Methodological issues and first results of a record linkage between AIDS and cancer registries in Italy. *Epidemiol Prev* 2000;**24**:109–16.
- 33 **Felt MJ**. The development of matching criteria for epidemiological studies using record linkage techniques. *Int J Epidemiol* 1984;**13**:351–5.
- 34 **van den Brandt PA**, Schouten LJ, Goldbohm RA, et al. Development of a record linkage protocol for use in the Dutch Cancer Registry for Epidemiological Research. *Int J Epidemiol* 1990;**19**:553–8.
- 35 **Chambers LW**, Spitzer WO, Hill GB, et al. Underreporting of cancer in medical surveys: a source of systematic error in cancer research. *Am J Epidemiol* 1976;**104**:141–5.
- 36 **Centeno C**, Núñez JM. Studies performed on the communication of the diagnostic of cancer in Spain. *Med Clin* 1998;**110**:744–50.
- 37 **Rodríguez Marín J**, López-Roig S, Pastor MA. Doctors' decision-making on giving information to cancer patients. *Psychol Health* 1996;**11**:839–44.
- 38 **Agudo A**, Pera G, Rodríguez M, et al. Changes in smoking habits in adults: results from a prospective study in Spain. *Ann Epidemiol* 2004;**14**:235–43.

bmjupdates+

bmjupdates+ is a unique and free alerting service, designed to keep you up to date with the medical literature that is truly important to your practice. bmjupdates+ will alert you to important new research and will provide you with the best new evidence concerning important advances in health care, tailored to your medical interests and time demands.

Where does the information come from?

bmjupdates+ applies an expert critical appraisal filter to over 100 top medical journals. A panel of over 2000 physicians find the few 'must read' studies for each area of clinical interest.

Sign up to receive your tailored email alerts, searching access and more...

www.bmjupdates.com