Validity of Self-reported Cancers in a Prospective Cohort Study in Comparison with Data from State Cancer Registries

Manuela M. Bergmann,[1] Eugenia E. Calle,[2] Cynthia A. Mervis,[2] Heidi L. Miracle-McMahill,[2] Michael J. Thun,[2] and Clark W. Heath[2]

The accuracy of self-reported cancer diagnoses in a prospective study was compared with population-based cancer registry data in four states. The study cohort included 65,582 men and women aged 39–96 years who were participants in the Cancer Prevention Study II Nutrition Survey, begun by the American Cancer Society in 1992. Estimates of sensitivity (the proportion of study participants with a registry-documented cancer who self-reported the cancer) ranged from 0.79 for an exact match of cancer site and year of diagnosis (±1 year) to 0.93 for a match of any reported cancer. The sensitivity of exact matches varied considerably by cancer site and was highest for breast, prostate, and lung cancers (0.91, 0.90, and 0.90, respectively) and lowest for rectal cancer and melanoma (0.16 and 0.53, respectively). Sensitivity also varied slightly by the age, education, and smoking status of study participants. Estimates of sensitivity were virtually identical for each of the four states. The positive predictive value (the proportion of self-reported cancers that were confirmed by the registries) was 0.75 overall and also varied by cancer site. Unlike sensitivity, however, this proportion varied considerably by state. All self-reports of cancer that were not confirmed by the registries were further investigated by repeat questionnaires and acquisition of medical records. Low positive predictive values were due to underascertainment of true cancer cases by the registries, inaccurate reporting on the part of study participants, and problems with the algorithm used by the state to link the study population to the registry data. In conclusion, the ability of members of this cohort to report a past diagnosis of cancer accurately is quite high, especially for cancers of the breast, prostate, lung, and colon, or for the occurrence of any cancer. Am J Epidemiol 1998;147:556–62.

follow-up studies; neoplasms; recall; registries; sensitivity and specificity

Although self-reported cancer diagnoses are frequently used in case finding, to trigger collection of medical records, and to exclude persons with prevalent disease from prospective analyses, few studies have assessed the accuracy of a self-report of cancer compared with other, presumably more accurate, sources of information. Several early studies investigated the accuracy of self-reported cancers (combining all malignant neoplasms) by comparing the self-reports with medical records (1–3). While self-reports were confirmed by medical records 53–82 percent of the time, only 33–61 percent of all cancers reported in medical records were also self-reported (1–3). In more recent

studies that examined the accuracy of self-reports by specific cancer site (4, 5), medical records again confirmed the large majority of self-reported cancers (74–100 percent). However, more than one quarter of the cancers discovered through review of medical records were not reported by the respondent (5). Harlow and Linet (6), in reviewing previous validation studies, emphasized the need for additional studies that validate self-reported illnesses.

The American Cancer Society's Cancer Prevention Study II (CPS-II) is an ongoing prospective mortality study of approximately 1.2 million American adults, begun in 1982. In 1992 and 1993 the CPS-II Nutrition Survey was implemented to obtain updated and more detailed information on diet and other potential risk factors from a large subsample of CPS-II respondents and to establish ongoing cancer incidence follow-up for the CPS-II Nutrition Survey cohort. To investigate the validity of self-reported cancer and the completeness of cancer case ascertainment in this cohort, we compared the self-reported occurrence of cancer among the CPS-II Nutrition Survey participants with

cancer diagnoses documented by state cancer registries in four states of varying size and geographic region.

MATERIALS AND METHODS

Cancer Prevention Study II Nutrition Survey

CPS-II was begun by the American Cancer Society in 1982 when about 1.2 million American men and women in all 50 states, the District of Columbia, and Puerto Rico were enrolled by Society volunteers. In 1992 and 1993, approximately 185,000 of the original CPS-II participants completed a second, confidential questionnaire (the CPS-II Nutrition Survey) that included information on diet, lifestyle, and disease history. The Nutrition Survey was conducted to fulfill two objectives: 1) to obtain updated, more detailed information on diet and other potential risk factors for cancer and 2) to define a large subgroup of CPS-II participants in whom incident cancers as well as mortality could be identified through future follow-up. Thus, the Nutrition Survey subgroup of CPS-II participants was selected from residents of 21 states that had population-based cancer registries reporting ascertainment of approximately 90 percent of incident cancers. Before completing the survey, all participants were informed of our intention to identify cases of cancer through state cancer registries.

The CPS-II Nutrition Survey collected information on lifetime history of cancer diagnoses and years of diagnoses. Specifically, participants were asked, "Has a physician ever diagnosed you with any type of cancer?" and "If yes, have you been diagnosed with any of these types of cancer: colon cancer, prostate cancer, breast cancer, skin cancer (non-melanoma-type), lung cancer, other cancer?" In answering "yes" to "other cancer," a participant could specify up to three types in open-ended fields. A field for year of diagnosis accompanied each cancer type. For the current validation study, we focused on the cancers (other than nonmelanoma skin cancer) reported in the Nutrition Survey that occurred during the years of registry operation.

Cancer registries

We compared cancer history reported at interview with reports from the cancer registries of four states. The self-reported completeness of these state cancer registries in 1990 ranged from 88 to 98 percent (7). All four registries are centralized, statewide cancer registries that became population based between 1981 and 1988. Thus, the time period available for validation of reported cancers varied from 5 to 12 years, from the beginning of registry operations through 1992 or 1993.

For this study, we did not choose registries belonging to the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute because the vast majority (89 percent) of Nutrition Survey participants are covered by non-SEER registries. Despite differences in their data collection systems and extent of data collection, all four of the registries follow general rules about reportable versus nonreportable diagnoses (8). Reportable cases typically include all new carcinomas, sarcomas, melanomas, leukemias, and lymphomas, both in situ and malignant, as identified in the latest edition of the *International Classification of Diseases for Oncology*.

Comparison of self-reported cancer with registry data

Identifying information for each of 65,714 CPS-II Nutrition Survey participants was submitted to the registries, including first and last name (for 100 percent of the subjects), middle initial (70 percent), day of birth (98 percent), month of birth (100 percent), year of birth (100 percent), Social Security Number (96 percent), and sex (100 percent). The linkage between CPS-II Nutrition Survey participants and cancer cases for each registry database was conducted by the individual registries. All linkages were completed between October 1994 and August 1995.
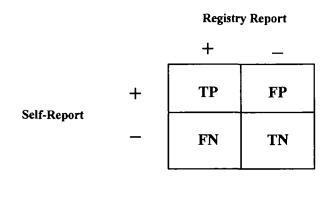
Three states used a probabilistic linkage procedure (9) and one a deterministic procedure. A deterministic approach links records from two files if they agree uniquely on all or on a predefined number of variables. In this approach, the user determines, based on subjective criteria, the level of agreement needed to define a match (10). In contrast, the more complex probabilistic approach assigns a numerical weight to every item on which matching is done, and a total weight is obtained indicating the probability that two records do or do not match. On a logarithmic scale, the weight statistic is an estimate of the odds that the two records refer to the same individual as opposed to different individuals (10).

Participants with more than one cancer reported by the registry (n = 117) were excluded from the main analysis and looked at separately. We also excluded participants for whom the registry reported a cancer with unknown primary site (n = 15). After these exclusions, the total population for the main analysis was 65,582 individuals (30,785 males and 34,797 females); at the time of the CPS-II Nutrition Survey, their mean age was 63 years, the youngest was 39, and the oldest was 96 years old.

Two types of conditions were reported occasionally but not uniformly by both the respondents and the registries. These were benign tumors or tumors of

uncertain behavior, and nonmelanoma cancers of the skin. In this analysis, reports of these conditions by either respondent or registry were considered equivalent to no report of cancer.

The definition of a true positive match (figure 1) between a self-report and a registry report considered two variables: the cancer site and the year of diagnosis. Initially, several levels of matches were defined across both cancer site and year of diagnosis. These levels represent the range of precision of all potential matches. For example, the most precise match defined in this study was a self-report and a registry report that matched exactly on cancer site and within 1 year on year of cancer diagnosis (an "exact" match). The least precise match was defined as any report of a cancer by the participant and any specified cancer and date of diagnosis by the registry.

The non-links were categorized as one of the following: a registry report of a cancer with no report of the cancer by the respondent (i.e., false negative), a report of cancer by the respondent during the years of registry operation with no report by the registry (i.e., false positive), and no report of cancer by both the respondent and the registry (i.e., true negative) (figure 1). All false positive reports were further investigated to determine why the cancer had not been reported by the registry. This investigation began in July 1995 and was largely completed by June 1996. Subjects were recontacted by letter and asked to give more detailed information about their reported cancer including the following: type of cancer, date of diagnosis, place of diagnosis, treating physician, and whether the same cancer had been diagnosed at an earlier date. In addition, subjects were asked to authorize the release of their medical records.

If the reason for the apparent false positive was clear (the respondent denied the original report of cancer or clarified that the cancer was a recurrence or metastasis, had occurred outside the years of the registry's operation, or had been diagnosed out of state in a state with which the registry did not have a data exchange agreement), we did not obtain medical records, and we judged the registry report to be correct (i.e., the individual did not have an eligible cancer). If the explanation of the false positive was not obvious, we requested medical records. In several instances where we had only the date of death and no information on the place of diagnosis, we requested death certificates. We judged the self-report to be correct (i.e., the individual did have cancer) if we were able to confirm the cancer diagnosis through medical records or a death certificate (both for in-state diagnoses and for diagnoses occurring in a data exchange state).

In addition to obtaining medical records, we resubmitted all individuals with a false positive self-report to the registries for manual review. In each state, this review resulted in positive links that were missed by the initial computerized linkage procedure. We considered these individuals to be false positives since they did not link with registry data during the normal linkage process; they are noted separately in the presentation of false positive data.

We further investigated the false negative self-reports (a registry report of a cancer and no report of the cancer by the respondent) by examining responses to questions about noncancerous conditions in the Nutrition Survey (e.g., colon polyps, prostate problems). We hypothesized that a proportion of the false negatives might be due to a lack of awareness that a given condition was cancer.

Data analysis

Sensitivity and the positive predictive value were computed using the registry reports as the "gold standard" (figure 1). Sensitivity was calculated as the proportion of persons with a cancer report in a registry who also self-reported the cancer. The positive predictive value 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; thus, it is the predictive value of a positive report.

Initially, sensitivity was calculated for different levels of agreement on cancer site and year of diagnosis. This allowed us to assess the influence of different definitions of a true positive match on overall sensitivity. For all subsequent analyses of sensitivity and all analyses of positive predictive value, the most stringent definition of a true positive was used: exact

Registry Report

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Positive Predictive Value} = \frac{TP}{TP + FP}$$

FIGURE 1. Notation for statistical quantities described in text, Cancer Prevention Study Nutrition Survey, United States, 1992–1993. TP, true positive; FP, false positive; FN, false negative; TN, true negative.
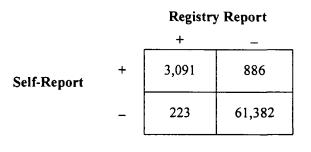
agreement on cancer site and agreement within 1 year on year of diagnosis.

RESULTS

After linkage with the cancer registries, the total study population of 65,582 individuals resulted in 3,091 true positives (all definitions), 223 false negatives, 886 false positives, and 61,382 true negatives (figure 2). The great majority (84 percent) of the total true positives were exact matches (figure 2); an additional 4.5 percent of the total true positives matched on organ system and year of diagnosis. Estimates of sensitivity based on the different match definitions ranged from 0.79 for an exact match to 0.93 for a match of any reported cancer.

The sensitivity of exact matches varied considerably by cancer site and was highest for breast, prostate, and lung cancers (0.91, 0.90, and 0.90, respectively) (table 1; figure 3). A low sensitivity was observed for rectal cancer and melanoma of the skin (0.16 and 0.53, respectively). The low sensitivity for rectal cancer was due, in large part, to participants' misreporting rectal cancer as colon cancer. When self-reported rectal cancers were allowed to match with cancers of the same organ system, the sensitivity increased to 0.77. The positive predictive value was highest for reports of breast, prostate, and uterine cancers (0.85, 0.80, and

Any Type of Match

Registry Report

		+	−
Self-Report	+	3,091	886
	−	223	61,382

Exact Match

Registry Report

		+	-
Self-Report	+	2,610	886
	−	704	61,382

FIGURE 2. Numbers of true positives, false positives, false negatives, and true negatives for all cancers combined, any type of match and exact match, Cancer Prevention Study Nutrition Survey, United States, 1992–1993.

TABLE 1.   Sensitivity and positive predictive value of self-reported cancer by selected sites, Cancer Prevention Study II Nutrition Survey, 1992–1993

	No. of true positive reports*	Sensitivity	Positive predictive value
All sites	2,610	0.79	0.75
Colon	229	0.85	0.54
Rectum	12	0.16	0.71
Lung	151	0.90	0.72
Melanoma	86	0.53	0.34
Breast	779	0.91	0.85
Uterus	121	0.71	0.79
Prostate	793	0.90	0.80
Bladder	102	0.67	0.72
Leukemia	27	0.61	0.41
Lymphoma (non-Hodgkin's)	70	0.64	0.69

* Numbers do not add up to all sites because all sites are not presented separately.

0.79, respectively) and lowest for melanoma and leukemia (0.34 and 0.41, respectively).

When the registry reported colon cancer and the self-report indicated no cancer at all ($n = 17$), 82 percent of respondents answered "yes" to having colon polyps in another section of the Nutrition Survey. Similarly, with false negative self-reports of prostate cancer ($n = 28$), 89 percent of respondents reported having "prostate problems." Among the false negative self-reports of melanoma ($n = 46$), 59 percent of respondents reported having nonmelanoma skin cancer or skin cancer not otherwise specified.

Relative to cancer site, there were less variability in sensitivity and very little variability in positive predictive value for total cancer across levels of age, sex, smoking status, and education (table 2). Sensitivity increased slightly with increasing education and was somewhat higher among nonsmokers than among current smokers. Sensitivity decreased slightly with increasing age of the respondent. While sensitivity did not vary among the four states, the positive predictive value varied considerably by state, from 0.63 to 0.89, due to a much higher proportion of false positive reports in states B and C than in states A and D.

We were able to obtain additional information from the registry, the respondent, or medical records for 567 of the 886 false positive reports (64 percent). Overall, 61 percent of the self-reports of cancer that were not initially reported by the registry were ultimately confirmed to be true cancers (table 3). This proportion varied considerably by state (from 28 percent for state D to 78 percent for state C) and was due primarily to missed cancer cases that were diagnosed in-state during the years of registry operation (table 3). Over half of these missed cases comprised prostate cancers (26 percent), melanomas (20 percent), and breast cancers (16 percent). The remaining missed cancers were leu-
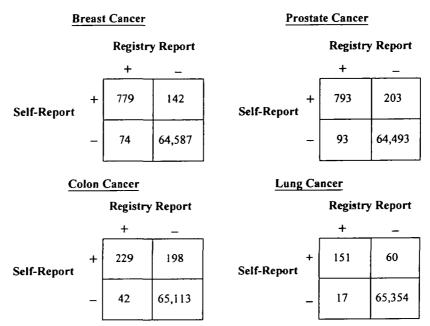
Breast Cancer

Prostate Cancer

FIGURE 3.   Numbers of true positives, false positives, false negatives, and true negatives for breast, prostate, colon, and lung cancers, exact match, Cancer Prevention Study Nutrition Survey, United States, 1992–1993.

kemia (6 percent), colon (5 percent), lymphoma (4 percent), bladder (4 percent), lung (2 percent), and other cancers (16 percent). While linkage failure explained fewer than 10 percent of all false positive reports in three of the four states, for state C the number of cancer cases missed due to linkage failure was substantial (table 3).

Of the 117 individuals with more than one cancer reported by the registry (who were excluded from the main analysis), 93 percent had two cancer diagnoses, and 7 percent had three or more cancer diagnoses. The sensitivity of self-reports decreased when the registry reported two or more cancer diagnoses. While 91 percent of the 117 respondents reported at least one cancer accurately (year and site), only 35 percent reported all cancers accurately. No false negatives occurred among participants with multiple registry reports.

DISCUSSION

Using the registry reports of cancer as the standard, this study indicates that the sensitivity for self-reports of all cancer sites varies from 0.79 to 0.93 depending on the definition of a true positive match. Sensitivity is very high when any mention of cancer by the respondent is sufficiently informative. This might be the case if all mentions of cancer are followed by review of medical records to obtain complete information or if prevalent cancer cases are being excluded from a prospective data analysis. Where complete accuracy of both the site and year of diagnosis is essential, the

TABLE 2.   Sensitivity and positive predictive value of self-reported cancer by age, sex, smoking, and education, Cancer Prevention Study II Nutrition Survey, 1992–1993

	No. of true positive reports	Sensitivity	Positive predictive value
Total	2,610	0.79	0.75
Age (years)			
<60	421	0.81	0.73
60–<70	1,405	0.80	0.75
≥70	784	0.76	0.75
Sex			
Males	1,344	0.78	0.74
Females	1,266	0.80	0.76
Smoking			
Never smoker	971	0.79	0.73
Former smoker	1,206	0.81	0.76
Current smoker	257	0.75	0.75
Unknown	176	0.69	0.72
Education			
<High school	160	0.75	0.75
High school	558	0.78	0.72
Some college	839	0.77	0.75
College graduate	1,036	0.81	0.76
Unknown	17	0.74	0.77
State			
State A	971	0.78	0.84
State B	611	0.80	0.67
State C	600	0.78	0.63
State D	428	0.80	0.89

TABLE 3. Resolution of false positive reports using medical records and additional follow-up, Cancer Prevention Study II Nutrition Survey, 1992–1993

| | Total | | State | | | | | | | |
| | | | A | | B | | C | | D | |
	No.	%	No.	%	No.	%	No.	%	No.	%
Self-report correct										
Total	343	60.5	44	38.6	100	59.2	187	77.6	12	27.9
Diagnosed in state	167	29.5	33	28.9	67	39.6	58	24.1	9	20.9
Diagnosed out of state*	79	13.9	0	0.0	24	14.2	54	22.4	1	2.3
Linkage failure	97	17.1	11	9.6	9	5.3	75	31.1	2	4.7
Registry correct										
Total	224	39.5	70	61.4	69	40.8	54	22.4	31	72.1
Noncancer†	84	14.8	33	28.9	19	11.2	20	8.3	12	27.9
Noneligible cancer‡	90	15.9	36	31.6	19	11.2	18	7.5	17	39.5
Diagnosed out of state§	50	8.8	1	0.9	31	18.3	16	6.6	2	4.7

* Diagnosed in a state with which the registry had a data exchange agreement.
† Respondent modified original report of cancer.
‡ Respondent reported a cancer recurrence, metastasis, or a cancer diagnosed outside the years of registry operation.
§ Diagnosed in a state with which the registry had no data exchange agreement.

overall sensitivity for self-reported cancers is lower (i.e., 0.79) but still reasonably high. The overall sensitivity found in this study is higher than the sensitivity found in a recent study where cancer cases documented in medical records were reported by 71 percent of the study participants (5). In earlier studies, the sensitivity was much lower, ranging from 0.33 when health insurance records were compared with household interviews (2) to 0.61 when Kaiser Foundation physician records were compared with personal interviews (3). The very low sensitivity found in early studies may indicate that cancer patients in the 1960s and 1970s were less well informed about their disease status than they are today.

Sensitivity and the positive predictive value varied considerably by cancer site. High sensitivity was observed for self-reports of lung, breast, and prostate cancer. Sensitivity for both breast and prostate cancer was considerably higher than had been observed in a previous study (5). The high predictive value of breast and prostate cancer self-reports is consistent with previous studies using medical records for confirmation (4, 5). The somewhat lower predictive value observed in this study for lung cancer is probably due in part to self-reports of lung cancers that are metastatic rather than primary cancer diagnoses. Similar results were seen in the Nurses' Health Study where the predictive value for lung cancer self-reports was 67 percent (4). A very low sensitivity of 0.16 was observed for self-reports of rectal cancer; study participants' inability to accurately report a diagnosis of rectal cancer has been previously observed (5). In the great majority of cases

where the site reported by the registry was "rectum," the self-reported site was "colon." The sensitivity for colon, bladder, and uterine cancers in this study is comparable with that of others (5) and varies from 0.67 to 0.85.

Lower sensitivity was found for self-reports by individuals aged 70 years and older, current smokers, and individuals who did not graduate from high school. While no other study has investigated the accuracy of cancer self-reports by sociodemographic characteristics, Nevitt et al. (11) found that self-reports of fractures in elderly women were more accurate among individuals with a college education.

Unlike sensitivity, the proportion of self-reported cancers that was confirmed by the registries varied considerably by state. Such differences could imply either differences in the quality of self-reports by state or differences in the quality of the registry data or linkage procedures by state. Further follow-up of these unconfirmed self-reports revealed that 61 percent of them overall were true cancers that had, indeed, been missed by the registries. Thus, the terminology initially used to describe these reports as "false positives" is, in this case, somewhat of a misnomer. The exact reasons for the missed cases are not known; most were diagnosed in-state, and the majority were common cancers with good survival (prostate, melanoma, breast).

Registries may miss a significant proportion of melanoma skin cancer cases because they are being diagnosed and treated in an outpatient setting; however, outpatient treatment is less likely for prostate cancer

and quite unlikely for breast cancer. The follow-up of false positive reports also revealed that registries may differ in their ability to accurately link an external population to their case files. In this study, the one state that used a deterministic linkage algorithm (state C) had a much higher level of linkage failure than did the three states that used more complex probabilistic linkage methods. Close to 40 percent of all false positive reports could not be subsequently confirmed as eligible cancers. Most were due to respondents' misreporting of noncancerous conditions as cancer or to reporting recurrent or metastatic disease.

This study is the first to use state-based registry data to validate self-reported cancer diagnoses. Our study population is considerably larger than in any previous validation study of self-reported cancer, and consequently we were able to investigate the accuracy of reports of specific cancer sites as well as all cancers combined. In addition, unlike some studies that were able to validate only positive reports of disease, we were able to look at the accuracy of both positive and negative reports of cancer diagnoses and to investigate to some extent the completeness of cancer reporting by state cancer registries. One limitation of our study is that the CPS-II Nutrition Survey participants represent a relatively educated and health-conscious group of older adults whose ability to accurately report their cancer diagnoses may differ from that of other groups.

In conclusion, the ability of members of this cohort to report a past diagnosis of cancer accurately is quite high, especially for cancers of the breast, prostate, lung, and colon, or for the occurrence of any cancer. For other specific sites (e.g., rectal cancer and melanoma), self-reports are less likely to be accurate and usually must be supplemented with other sources of information. In addition, this study has shown that linkage of a large cohort with multiple state cancer registries is feasible and will ascertain the large majority of cancer cases. However, some true cases will be missed, and the extent of this underascertainment will likely depend on the individual registries, both on the quality of their data and on their ability to link successfully with external populations.

REFERENCES

1. Krueger DE. Measurement of prevalence of chronic disease by household interviews and clinical evaluations. Am J Public Health 1957;47:953–60.
2. National Center for Health Statistics. Health interview responses compared with medical records. Rockville, MD: National Center for Health Statistics, 1965. (Vital and health statistics, Series 2: no. 7) (USPHS publication no. 1000).
3. Madow WG. Net differences in interview data on chronic conditions and information derived from medical records. Rockville, MD: National Center for Health Statistics, 1973. (Vital and health statistics, Series 2: no. 57) (DHEW publication no. (HSM) 73–1331).
4. 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.
5. Paganini-Hill A, Chao A. Accuracy of recall of hip fracture, heart attack, and cancer: a comparison of postal survey data and medical records. Am J Epidemiol 1993;138:101–6.
6. Harlow SD, Linet MS. Agreement between questionnaire data and medical records. Am J Epidemiol 1989;129:233–48.
7. Howe HL, ed. Cancer incidence in North America, 1988–1991. Sacramento, CA: North American Association of Central Cancer Registries, April 1995.
8. Seiffert JE, Hoyler SS, McKeen K, et al. Casefinding, abstracting, and death clearance. In: Menck H, Smart C, eds. Central cancer registries: design, management, and use. Chur, Switzerland: Harwood Academic Publishers, 1994:35–63.
9. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14:491–8.
10. Roos LL, Wajda A. Record linkage strategies: estimating information and evaluating approaches. Methods Inf Med 1991;30:117–23.
11. Nevitt MC, Cummings SR, Browner WS, et al. The accuracy of self-report of fractures in elderly women: evidence from a prospective study. Am J Epidemiol 1992;135:490–9.