Title
Validating California Teachers Study Self-Reports of Recent Hospitalization: Comparison with California Hospital Discharge Data

Journal
American Journal of Epidemiology, 158(10)

ISSN
0002-9262

Authors
Marshall, Sarah F
Deapen, Dennis
Allen, Mark
et al.

Publication Date
2003-11-15

DOI
10.1093/aje/kwg256

American Journal of Epidemiology

Vol. 158, No. 10
Printed in U.S.A.
DOI: 10.1093/aje/kwg256

Validating California Teachers Study Self-Reports of Recent Hospitalization: Comparison with California Hospital Discharge Data

Sarah F. Marshall[1], Dennis Deapen[1], Mark Allen[2], Hoda Anton-Culver[3], Leslie Bernstein[1], Pamela L. Horn-Ross[4], David Peel[3], Rich Pinder[1], Peggy Reynolds[5], Ronald K. Ross[1], Dee West[4], and Al Ziogas[3]

[1] Keck School of Medicine, University of Southern California, Los Angeles, CA.
[2] Public Health Institute, Sacramento, CA.
[3] School of Medicine, University of California, Irvine, CA.
[4] Northern California Cancer Center, Union City, CA.
[5] Environmental Health Investigations Branch, California Department of Health Services, Oakland, CA.

Received for publication October 24, 2002; accepted for publication June 3, 2003.

Determining an accurate method of obtaining complete morbidity data is a long-standing challenge for epidemiologists. The authors compared the accuracy and completeness of existing California hospital discharge data with self-reports of recent hospitalizations and surgeries from participants in the California Teachers Study. Self-reports were collected by questionnaire in 1997 from 91,433 female teachers and administrators residing in California. Of the 13,430 hospital discharge diagnoses identified for these women, cohort members reported 58%. Self-reporting was highest for neoplasms and musculoskeletal and connective tissue diseases and was most accurate for scheduled admissions, more recent admissions, longer lengths of stay, and less severe disorders. Hospitalizations for mental health and infectious disease were not as well reported. Among the 26,383 self-reports—including outpatient surgeries, which are not captured by the hospital discharge database—confirmation was lower, as expected, especially for disorders of the nervous system and sense organs and skin and subcutaneous tissue. Confirmation was highest for childbirth admissions. The hospital discharge database was more specific, but the self-reports were more comprehensive, since many conditions are now treated in outpatient settings. The combination of self-reports and secondary medical records provides more accurate and complete morbidity data than does use of either source alone.

data collection; epidemiologic methods; hospital records; prospective studies

Abbreviations: CI, confidence interval; ICD-9, *International Classification of Diseases*, Ninth Revision; OR, odds ratio; OSHPD, Office of Statewide Health Planning and Development.

Collection of morbidity data, even from the recent past, poses challenges for epidemiologists. Such data may be collected by asking research participants directly about their illnesses and surgeries. However, this information may also be present in varying degrees of completeness in existing medical databases. A comparison of data of both types will assist researchers in choosing the best method in specific circumstances. Some studies have found that the validity of self-reporting varies with diagnosis (1–10), but these studies have focused primarily on cancer (1–9, 11), circulatory system diseases (1, 5–8, 10, 12, 13), or one specific disease (14, 15). We evaluated the validity of self-reported hospitalization and surgery for the full range of diagnoses in a cohort of female California teachers by comparing their self-reports with statewide hospital discharge records. The accuracy of self-reporting depends on the population of interest (16, 17), so we investigated whether admission and patient characteristics accounted for any differences in reporting even in this highly educated cohort. We also examined the usefulness and completeness of the hospital discharge database.

Correspondence to Sarah F. Marshall, Department of Preventive Medicine, University of Southern California, 1420 San Pablo Street, PMB-B105, Los Angeles, CA 90033 (e-mail: smarshal@usc.edu).

MATERIALS AND METHODS

The California Teachers Study is a prospective study of 133,479 female teachers and administrators recruited in 1995 from the California State Teachers Retirement System, as described elsewhere (18). The cohort is primarily non-Hispanic White (87 percent), although there are over 3,500 African Americans, over 4,500 Asians/Pacific Islanders, and over 5,400 Hispanics in the group. The majority of the cohort members (59 percent) were aged ≥50 years at enrollment, and 15 percent were aged ≥70 years. Participants returned self-administered questionnaires by mail. In a questionnaire distributed in 1997–1998, participants were asked, "During the last 2 years, were you hospitalized for an illness or did you have a surgical procedure?" Response options were "no," "yes for heart disease," "yes for cancer (specify type)," and "yes for other reason (specify)." The specified hospitalizations or surgery reports were coded manually by one of us (S. F. M.) using the *International Classification of Diseases*, Ninth Revision (ICD-9), Clinical Modification (19) and ICD-9 coding software (20). Where possible, a code was assigned for the diagnosis rather than the hospital procedure or type of surgery. Values were checked for inadmissible codes and other data entry errors. Items with missing data, where neither "no" nor "yes" was checked and no conditions were specified, were classified under "no report."

Social Security number, date of birth, and race/ethnicity (White, African American, Hispanic, Asian/Pacific Islander, or other/mixed/not specified) were self-reported by each participant on the baseline questionnaire mailed in 1995. Social Security numbers in the California Teachers Study are 99 percent complete and were validated using a checking algorithm that excludes numbers outside of the possible range. Dates of birth were recorded for all study participants and were validated by checking with the California Department of Motor Vehicles and the California State Teachers Retirement System. Age in years was calculated as time from the birth date to the date of questionnaire completion (when date of completion was unknown, the date of questionnaire return was used). Socioeconomic status was obtained by linking the residential street addresses of cohort members to US Census neighborhood ("block")-level data (21). A summary index was generated from median family income, proportion of adults with a college degree or higher, and proportion of adults employed in managerial or professional occupations. Women who had not continuously lived in California for 2 years before administration of the 1997 questionnaire were excluded from the analyses. California residence was self-reported on both questionnaires and was supplemented by information obtained from the US Postal Service National Change of Address database and the California Department of Motor Vehicles.

Patient discharge data for 1995–1999 were obtained from the California Office of Statewide Health Planning and Development (OSHPD). The OSHPD file provides a record of all inpatient stays in acute-care hospitals licensed by the state of California (22). For comparability with the self-reports, patient discharges were only included if the patient had been admitted during the 2 years prior to the date of questionnaire completion. If a participant had multiple discharges for the same principal diagnosis during this time period, only the first discharge was included.

OSHPD discharge records include data on principal diagnosis (ICD-9 code), principal procedure, principal external cause of injury (E-code), up to 24 other diagnoses, severity of patient illness (minor, moderate, major, or extreme loss of function), admission date, type of admission (unscheduled or scheduled), and length of stay in days. The California Health and Safety Code requires that all hospitals provide discharge data, with the exception of 11 state mental health and developmental hospitals (22). Data from the hospitals are checked by the OSHPD and are required to meet data error tolerance levels of 0.1 percent for sex and date of birth (23). Of 3,775,711 discharges occurring in 1999, 777,718 (20.6 percent) were missing a Social Security number (73 percent of these were for patients aged 10 years or younger) and 120 (0.003 percent) were missing a date of birth. California Teachers Study records were linked with the OSHPD database using AutoMatch probabilistic record linkage software (MatchWare Technologies, Inc., Kennebunk, Maine). Participants were linked by Social Security number, date of birth, and sex; diagnostic information was not considered. Discrepancies between possible matches were resolved visually.

Statistical analysis

In the first set of analyses, the principal diagnoses from the OSHPD were used as the "gold standard" against which self-reports were compared. "Exact" matches were defined as those in which the first three digits of the ICD-9 codes from the two sources were the same for the same patient. "Nonspecific" matches were defined as those in which the self-report was too general for precise coding but was judged a match in visual review, such as when the respondent specified a surgical procedure rather than a diagnosis and this matched the OSHPD principal procedure or when the respondent reported a matching hospitalization rather than a diagnosis. For example, a self-report of "heart disease" (ICD-9 code 429.9) was considered a nonspecific match with several types of heart disease (ICD-9 codes 390–398, 402, and 404–428). The rate of confirmation of the OSHPD data, defined as the number of self-reports matching the OSHPD file divided by the total number of OSHPD discharges among eligible California Teachers Study members, was calculated for each ICD-9 diagnosis group and for each level of agreement (exact and nonspecific matches).

Characteristics associated with the hospitalization (scheduled or unscheduled admission (<24 hours or ≥24 hours), time since admission in quartiles, length of stay (≤1 day, 2–3 days, or >4 days), and illness severity (minor, moderate, major, or extreme loss of function)) and patient characteristics (age (<50 years, 50–69 years, or ≥70 years), race/ethnicity (White or non-White), and socioeconomic status (low, medium, or high)) were considered as potential predictors of a "nonspecific" match (yes/no). Hospitalization covariates were already in categories set by the OSHPD, and the data on patient characteristics were grouped into tertiles for analysis. Logistic regression methods were used to calculate

TABLE 1. Distribution of diagnoses made in the 2 years prior to questionnaire completion for 91,433 women in the California Teachers Study, by report source, 1995–1997

	Self-report	OSHPD* discharge record
Total no. of women	91,433	91,433
Women with no report	69,882	80,754
Women with one or more reports	21,551	10,679
Total no. of diagnoses reported	26,383	13,430

* OSHPD, Office of Statewide Health Planning and Development.

unadjusted odds ratios for each patient and admission characteristic and 95 percent confidence intervals (24). Adjusted odd ratios were calculated by entering the variables into a multivariate logistic regression model.

In the second set of statistical analyses, the self-reports were treated as the gold standard and OSHPD patient discharges were compared with them, using the "nonspecific" matching criteria as before. The rate of confirmation of the self-reports by the OSHPD data (the number of OSHPD discharge records matching the self-reports divided by the total number of self-reports) was calculated for each diagnosis group.

RESULTS

Of the 99,519 women who returned a 1997–1998 questionnaire, 8,080 were excluded because they had not lived continuously in California during the 2 years prior to completion of the questionnaire. Also excluded were six women who completed their second questionnaire after 1999, since their full hospital records may not have been included in the OSHPD data. Therefore, 91,433 members of the original cohort remained in the analysis; 23.6 percent reported having had one or more hospitalizations or surgeries in the prior 2 years (table 1), and 11.7 percent had an OSHPD discharge record in the same period. Some women had multiple hospitalizations or surgeries in this time frame. This provided us with a total of 26,383 self-reported diagnoses and 13,430 OSHPD diagnoses for analysis. Eight hundred and ten women reported neither "yes" nor "no"; these missing values were treated as "no report" in the analysis.

OSHPD records

Of the 10,679 participants with hospitalizations recorded in the OSHPD database, 79.7 percent reported a hospitalization on their questionnaire. Of the 13,430 unique diagnoses in OSHPD, 7,798 (58.1 percent) matched with a self-report when nonspecific matches in diagnoses were accepted. Of

TABLE 2. Percentage of reports from the California Office of Statewide Health Planning and Development that were confirmed by California Teachers Study self-reports, by diagnosis group, 1995–1997*

Diagnosis group (ICD-9† codes)	No. of OSHPD† reports	Percentage self-reported (exact three-digit ICD-9 code matches only)		Percentage self-reported (includes nonspecific matches)	
		%	95% CI†	%	95% CI
Blood and blood-forming organs (280–289)	55	32.7	20.7, 46.7	43.6	30.3, 57.7
Circulatory system (390–459)	1,670	16.4	14.6, 18.2	66.5	64.2, 68.8
Congenital anomalies (740–759)	27	29.6	13.8, 50.2	77.8	57.7, 91.4
Digestive system (520–579)	1,218	26.0	23.6, 28.6	62.6	59.8, 65.3
Endocrine, nutritional, metabolic, and immune system (240–279)	242	30.6	24.8, 36.8	43.4	37.1, 49.9
Genitourinary system (580–629)	1,153	18.3	16.1, 20.7	70.0	67.3, 72.6
Infectious and parasitic diseases (001–139)	145	20.7	14.4, 28.2	34.5	26.8, 42.8
Injury and poisoning (800–999)	1,012	30.2	27.4, 33.2	54.9	51.8, 58.0
Mental disorders (290–319)	204	5.9	3.1, 10.0	27.5	21.5, 34.1
Musculoskeletal system and connective tissue (710–739)	1,346	16.1	14.1, 18.1	83.7	81.6, 85.6
Neoplasms (140–239)	1,692	42.2	39.8, 44.6	85.3	83.6, 87.0
Pregnancy and childbirth (630–676)	2,952	5.0	4.2, 5.8	38.9	37.1, 40.7
Nervous system and sense organs (320–389)	138	36.2	28.2, 44.8	52.2	43.5, 60.7
Respiratory system (460–519)	502	34.1	29.9, 38.4	55.8	51.3, 60.2
Skin and subcutaneous tissue (680–709)	111	34.2	25.5, 43.8	50.5	40.8, 60.1
Supplementary factors influencing contact with health services (V codes)	466	4.5	2.8, 6.8	20.0	16.4, 23.9
Symptoms and ill-defined conditions (780–799)	497	13.5	10.6, 16.8	17.5	14.3, 21.1

* The ICD-9 group "Conditions Originating in the Perinatal Period" (codes 760–779) is not shown; there were no OSHPD discharges for the California Teachers Study cohort in this category.

† ICD-9, *International Classification of Diseases*, Ninth Revision; OSHPD, Office of Statewide Health Planning and Development; CI, confidence interval.

TABLE 3. Reports from the California Office of Statewide Health Planning and Development with a low proportion of confirmation by California Teachers Study self-reports (for selected diagnoses), 1995–1997

Diagnosis group and diagnosis (ICD-9* code)	No. of OSHPD* reports	Percentage self-reported (includes nonspecific matches)	
		%	95% CI*
Infectious and parasitic diseases			
Septicemia (038)	74	17.6	9.7, 28.2
Mental disorders			
Depressive disorder (311)	8	25.0	3.2, 65.1
Alcohol dependence syndrome (303)	13	7.7	0.3, 36.0
Adjustment reaction (309)	9	0	0, 28.3
Pregnancy and childbirth			
Other amniotic cavity problems (658)	115	35.7	26.9, 45.1
Umbilical cord complications (663)	158	27.2	20.4, 34.9
Normal delivery (650)	353	23.5	19.2, 28.3
Perineal trauma with delivery (664)	627	21.5	18.4, 25.0
Supplementary factors influencing contact with health services			
Rehabilitation procedure (V57)	202	1.0	0.1, 3.5
Convalescence (V66)	19	0	0, 14.6
Symptoms and ill-defined conditions			
Other abdomen/pelvis symptoms (789)	40	15.0	5.7, 29.8
Respiratory system/other chest symptoms (786)	288	13.2	9.5, 17.7

* ICD-9, *International Classification of Diseases*, Ninth Revision; OSHPD, Office of Statewide Health Planning and Development; CI, confidence interval.

these, 2,673 (19.9 percent) matched exactly with the self-reported three-digit ICD-9 diagnosis. The proportion self-reported by the teachers varied substantially by principal diagnosis (table 2). Self-reporting, at the nonspecific matching level, was highest for neoplasms (85.3 percent), which included two subgroups: malignant neoplasms (83.6 percent, not shown) and benign, in-situ, unknown, and unspecified neoplasms (87.4 percent, not shown). Musculoskeletal and connective tissue diseases (83.7 percent), congenital anomalies (77.8 percent), and genitourinary disorders (70.0 percent) also had high proportions of matches at the nonspecific level. The diagnosis group with the highest proportion of exact ICD-9 code matches was neoplasms (42.2 percent). The proportion of exact matches for malignant neoplasms only was 66.0 percent (not shown). Other categories with high proportions of exact matches were disorders of the nervous system and sense organs (36.2 percent), disorders of the respiratory system (34.1 percent), and disorders of the skin and subcutaneous tissue (34.2 percent).

Diagnoses found in the hospital records with the lowest self-reporting by the teachers, including the nonspecific matches, were symptoms and ill-defined conditions (17.5 percent), supplementary factors influencing contact with health services (20.0 percent), mental disorders (27.5 percent), infectious and parasitic diseases (34.5 percent), and pregnancy and childbirth (38.9 percent). With the exception of the infectious disease group, these groups also had the lowest percentages of exact matches. Within these groups, there were certain diagnoses with particularly low levels of self-reporting, as shown in table 3. For example, convalescence, a general medical or follow-up examination, and a diagnosis of adjustment reaction, alcoholic psychosis, schizophrenia, or drug dependence were never self-reported.

Characteristics of OSHPD discharges that were self-reported

The odds of accurately matching the diagnosis group of the self-reported condition and the diagnosis group recorded in the OSHPD database were calculated according to patient and admission characteristics. A scheduled admission, a length of stay longer than 1 day, and a shorter time between admission and report were more likely to be accurately reported (table 4). Patient illnesses causing minor loss of function were more likely to be reported than those with moderate, major, or extreme loss. Participants with self-reports were significantly older than those without self-reports, although the relation was not linear. After pregnancy and childbirth admissions were excluded, the relation reversed and became linear, so that reporting became negatively and significantly associated with age: For women aged 50–69 years relative to those aged <50 years, the odds ratio was 0.78 (95 percent confidence interval (CI): 0.69, 0.89); for women aged ≥70 years, the odds ratio was 0.53 (95 percent CI: 0.47, 0.60) ($p < 0.0001$ for trend; data not shown). Accurate self-reporting was higher among Whites than among non-Whites. Logistic regression analysis

TABLE 4. Odds ratios for confirmation of reports from the California Office of Statewide Health Planning and Development by California Teachers Study self-reports, according to patient characteristics (*n* = 13,414), 1995–1997

Characteristic	Unadjusted OR*,†	95% CI*	Adjusted OR†,‡	95% CI
Type of hospital admission				
Unscheduled (<24 hours) (*n* = 7,009)	1.0		1.0	
Scheduled (≥24 hours) (*n* = 6,405)	2.58	2.40, 2.77	2.33	2.17, 2.51
Severity of illness				
Minor loss of function (*n* = 6,576)	1.0		1.0	
Moderate loss of function (*n* = 5,193)	0.79	0.73, 0.85	0.72	0.66, 0.78
Major loss of function (*n* = 1,402)	0.67	0.60, 0.75	0.53	0.47, 0.61
Extreme loss of function (*n* = 243)	0.61	0.47, 0.79	0.44	0.34, 0.58
Length of hospital stay (days)				
≤1 (*n* = 3,839)	1.0		1.0	
2 or 3 (*n* = 5,424)	2.23	2.05, 2.43	2.22	2.03, 2.42
≥4 (*n* = 4,151)	2.24	2.05, 2.45	2.53	2.28, 2.81
No. of days between hospital admission and self-report				
≥541 (*n* = 3,327)	1.0		1.0	
351–540 (*n* = 3,381)	1.25	1.13, 1.37	1.24	1.12, 1.37
176–350 (*n* = 3,397)	1.40	1.27, 1.54	1.45	1.31, 1.61
0–175 (*n* = 3,309)	1.50	1.36, 1.66	1.53	1.38, 1.70
Age (years)				
<50 (*n* = 4,579)	1.0		1.0	
50–69 (*n* = 4,275)	1.95	1.79, 2.12	1.64	1.50, 1.80
≥70 (*n* = 4,560)	1.32	1.22, 1.44	1.23	1.13, 1.35
Socioeconomic status				
Lowest third (*n* = 4,572)	1.0		1.0	
Middle third (*n* = 4,773)	0.94	0.87, 1.02	0.93	0.85, 1.01
Highest third (*n* = 4,069)	1.07	0.98, 1.17	0.98	0.89, 1.07
Race/ethnicity				
Non-White (*n* = 1,569)	1.0		1.0	
White (*n* = 11,845)	1.23	1.11, 1.37	1.14	1.02, 1.27

* OR, odds ratio; CI, confidence interval.

† Odds of accurately matching the diagnosis group recorded in the Office of Statewide Health Planning and Development database and the self-reported condition. Ratios greater than 1 mean that the characteristic was associated with more accurate self-reporting in comparison with the baseline group.

‡ Adjusted for the other variables in the table.

revealed that African Americans (*n* = 257; odds ratio (OR) = 0.73, 95 percent CI: 0.57, 0.94) and Asians/Pacific Islanders (*n* = 384; OR = 0.77, 95 percent CI: 0.63, 0.95) had lower rates of reporting than Whites (*n* = 11,845), whereas Hispanics (*n* = 476; OR = 0.83, 95 percent CI: 0.69, 1.00) and persons of other/mixed/unspecified race/ethnicity (*n* = 452; OR = 0.87, 95 percent CI: 0.72, 1.05) had slightly lower rates. Socioeconomic status did not predict accurate self-reporting. In a final multivariate model including all of the variables—that is, length of stay, time since admission, scheduled admission versus unscheduled admission, degree of loss of function, age, socioeconomic status, and race/ethnicity—the effect of each covariate was attenuated slightly but not altered.

Self-reported diagnoses

Overall, 58 percent of the 26,383 self-reports corresponded with an OSHPD diagnosis. The number confirmed by the OSHPD varied (table 5), with the highest level of confirmation being seen for pregnancy and childbirth (75.5 percent), followed by mental disorders (45.3 percent), circulatory system disorders (43.1 percent), and treatment for external causes of injury (e.g., car accidents) (42.7 percent).

TABLE 5. Percentage of California Teachers Study self-reports that were confirmed by reports from the California Office of Statewide Health Planning and Development, by diagnosis group, 1995–1997*

Diagnosis group (ICD-9† codes)	No. of self-reports	Percentage confirmed by the OSHPD†	
		%	95% CI†
Blood and blood-forming organs (280–289)	69	30.4	19.9, 42.7
Circulatory system (390–459)	2,125	43.1	40.9, 45.2
Congenital anomalies (740–759)	39	23.1	11.1, 39.3
Digestive system (520–579)	2,545	28.3	26.5, 30.0
Endocrine, nutritional, metabolic, and immune system (240–279)	341	31.7	26.8, 36.9
Genitourinary system (580–629)	2,603	15.5	14.1, 17.0
Infectious and parasitic diseases (001–139)	242	15.7	11.4, 20.9
Injury and poisoning (800–999)	1,674	25.8	23.7, 27.9
Mental disorders (290–319)	95	45.3	35.0, 55.8
Musculoskeletal system and connective tissue (710–739)	3,454	33.4	31.9, 35.0
Neoplasms (140–239)	4,517	22.3	21.1, 23.5
Nervous system and sense organs (320–389)	2,002	3.4	2.6, 4.2
Pregnancy and childbirth (630–676)	1,393	75.5	73.1, 77.7
Respiratory system (460–519)	989	25.4	22.7, 28.2
Skin and subcutaneous tissue (680–709)	358	14.0	10.5, 18.0
Supplementary factors influencing contact with health services (V codes)	600	17.5	14.5, 20.8
Symptoms and ill-defined conditions (780–799)	490	22.7	19.0, 26.6
Diagnosis unspecified (includes treatment and external-causes-of-injury/poisoning reports)	2,847	42.7	40.9, 44.6

* The ICD-9 group "Conditions Originating in the Perinatal Period" (760–779) is not shown; there were no California Teachers Study self-reports in this category.

† ICD-9, *International Classification of Diseases*, Ninth Revision; OSHPD, Office of Statewide Health Planning and Development; CI, confidence interval.

Confirmation by the OSHPD was particularly low for disorders of the nervous system and sense organs, such as cataracts (3.4 percent); disorders of the skin and subcutaneous tissue (14.0 percent); disorders of the genitourinary system (15.5 percent), particularly breast lumps and cervical dysplasia; infectious and parasitic diseases (15.7 percent); and supplementary factors influencing contact with health services (17.5 percent). Specific diagnoses within these groups with very low confirmation are shown in table 6. There are no OSHPD records for certain diagnoses—for example, benign mammary dysplasia, glaucoma, corneal opacity, sebaceous gland disease, and prosthesis fitting or adjustment.

DISCUSSION

Overall, 80 percent of the participants with hospitalizations in the OSHPD database accurately reported having undergone hospitalization during the 2-year period of inquiry, and 58 percent of the self-reported conditions matched the OSHPD diagnoses. These results are not easily compared with those from other studies, since to our knowledge this is the first study to present data on variability in reporting of hospitalization or surgery for the full range of ICD-9 diagnosis groups. Having a large study population available and having access to an established hospital discharge database allowed us to complete such analyses.

Two previous studies examined the reporting of a range of chronic diseases and found that 32 percent of insurance reports (6) and 53 percent of physician records (7) were matched by in-person interviews for all diagnoses. Norrish et al. (25) found that 52 percent of hospital admissions over a 4-year period were recalled by telephone; however, they focused on the self-reporting of hospital admissions rather than the actual disease diagnoses. Research examining the self-reporting of diseases usually employs a checklist of specific conditions (1, 3, 5, 9, 10, 12, 13), whereas our query was an open-ended question with prompts for cancer and heart disease. The prompted conditions were well reported. Self-reporting of malignant neoplasms (83.6 percent) was higher than the 61–79 percent range reported in other cancer validation studies (2, 9). Our observed proportion of self-reported circulatory system disorders (66.5 percent) was slightly lower than the 73–80 percent agreement reported by Haapanen et al. (10). Unprompted conditions were not always as well reported, though the sensitivity of reporting of congenital disorders (77.8 percent) and musculoskeletal disorders (83.7 percent) was fairly high. Even so, a broad, open question such as ours is unlikely to elicit complete recall of all of the conditions of interest.

TABLE 6. Percentage of California Teachers Study self-reports with a low proportion of confirmation by reports from the California Office of Statewide Health Planning and Development (for selected diagnoses), 1995–1997

Diagnosis group and diagnosis (ICD-9* code)	No. of self-reports	Percentage confirmed by the OSHPD*	
		%	95% CI*
Genitourinary system			
Disorders of menstruation (626)	102	7.8	3.4, 14.9
Disorders of the uterus (621)	597	3.5	2.2, 5.3
Other disorders of the breast (611)	382	1.6	0.6, 3.4
Noninflammatory disorders of the cervix (622)	116	0.9	0, 4.7
Benign mammary dysplasia (610)	141	0	0, 2.1
Infectious and parasitic diseases			
Intestinal infection (008)	22	13.6	2.9, 34.9
Bacterial infection in other conditions (041)	44	6.8	1.4, 18.7
Nervous system and sense organs			
Mononeuritis in lower limb (355)	91	1.1	0, 6.0
Mononeuritis in upper limb (354)	202	1.0	0.1, 3.5
Cataract (366)	957	0.1	0, 0.6
Skin and subcutaneous tissue			
Sebaceous gland disease (706)	136	0	0, 2.2
Other skin disorders (709)	36	0	0, 8.0
Diseases of nail (703)	32	0	0, 8.9
Supplementary factors influencing contact with health services			
Elective surgery (V50)	327	1.5	0.5, 3.5
Fitting/adjustment of prosthesis (V52)	35	0	0, 8.2

* ICD-9, *International Classification of Diseases*, Ninth Revision; OSHPD, Office of Statewide Health Planning and Development; CI, confidence interval.

The teachers were less likely to identify hospitalizations for mental health problems, infectious diseases, or pregnancy. Zhu et al. (11) suggested that poor reporting is associated with less explicit diagnostic criteria, which may be the case with psychoses, depressive disorders, and chest and abdominal symptoms. The respondents may be confusing these diagnoses with others. For example, several OSHPD records of affective psychoses were self-reported as depression, stress, or an adverse reaction to a medicine, and respiratory and chest symptoms were frequently reported as heart disease or as an adverse reaction to medicine. Mental health problems, such as alcohol or drug dependence and psychoses, may be intentionally underreported because of a perception of societal unacceptability. Additionally, self-reporting of conditions like rehabilitation, convalescence, and pregnancy and childbirth may be lower because respondents do not associate these conditions with illness as such and may need specific prompting to report them.

The main disadvantage of collecting self-reported data on hospitalizations through an open-ended question is the variability in the specificity of the replies. Only 20 percent of the self-responses were specific to the degree that their three-digit ICD-9 code matched exactly with the hospital diagnosis. However, open-ended self-reports can provide researchers with a general indication of the diagnosis. For instance, the self-report alone accurately predicted the main diagnosis group of the OSHPD record in 58 percent of cases. Therefore, self-reports may be useful in identifying populations for which medical records could be obtained for more specific diagnoses.

The accuracy of self-reporting depends on the population of interest. The California Teachers Study cohort comprises women with a college education; female respondents and those with higher levels of education have been found to have greater reporting accuracy (1, 11, 16). The cohort is predominantly White, but it is unclear from previous studies whether reporting is associated with race/ethnicity (1, 2, 17). We found that Whites had more accurate reporting, but this result was not independent of age. Reporting did not vary by socioeconomic status, which is in agreement with the findings of Reijneveld and Stronks (26). Self-reporting of OSHPD diagnoses was associated with admission characteristics. For instance, a scheduled admission, a short recall period, and a longer length of stay were associated with increased recall. Given this, it was surprising that illnesses causing more severe loss of function were not better reported, since we would expect such diagnoses to result in longer hospital stays for the patient. The severity indicator may refer to cost of provision rather than disease severity; the most frequent "major" or "severe" diagnosis in our

sample was osteoarthritis, and the most common "severe" procedure was rehabilitation. Length of stay and the degree of loss of function remained independent in logistic regression analysis.

Completeness of the hospital database

There were only half as many diagnoses in OSHPD records as were self-reported. Consequently, the overall percentage of self-reports confirmed by OSHPD was low. The majority of these unconfirmed self-reports are likely to have been outpatient procedures not found in the OSHPD database; our question asked about hospitalizations or surgeries. A number of surgical procedures are now routinely performed in doctors' offices and other ambulatory or outpatient surgical centers; in 1996, US women had almost as many surgical procedures in these settings (11.6 million) as in hospital inpatient settings (11.9 million) (27). The types of conditions that were unconfirmed support this theory, being more frequently those requiring elective surgery or those treated in ambulatory settings, such as disorders of the eyes, skin, or genitourinary system (27). Our questionnaire was specifically designed to capture these outpatient events as well, but unfortunately combining them with inpatient illnesses in one question hampered our ability to compare them with OSHPD data.

Such a general open-ended question may also encourage overreporting of conditions that do not result in either surgery or hospitalization. Another form of overreporting is the inaccurate reporting of chronic conditions or other diseases for which participants have not actually undergone surgery or hospitalization in the past 2 years. This exact time frame may be problematic for respondents and may result in both under- and overreporting. We estimated the prevalence of overreporting due to inaccurate recall of the time of an event by rerunning the analyses using a wider time frame for the OSHPD discharge records. Allowing matches with hospital admissions from 1995–1999 rather than just the 2 years prior to questionnaire completion increased the overall confirmation by OSHPD data from 29 percent to 35 percent (by 1,505 matches).

Another type of reporting error probably includes participants' recalling a diagnosis or condition from their visit that was not actually the "principal" reason for the visit. Analysis accepting "nonspecific" matches with any top five "other" OSHPD diagnoses, rather than just the one identified as "principal," increased overall matching from 29 percent to 32 percent. In addition, the extra reports may be valid reports of hospitalizations that took place out of state or in one of the 11 California hospitals that are not included in the OSHPD database. The proportion of out-of-state hospitalizations is unknown, but it is unlikely to be high in California, where only a small proportion of the population resides near state borders.

Further disparities may be due to errors in the OSHPD data. Green and Wintfeld (28) found a 9 percent difference in principal diagnosis between original hospital records and the OSHPD database, and there may be ICD-9 coding differences between OSHPD data and the questionnaire data due to differences in coders and coding procedures. There may be error in the linkage of participants with the OSHPD file, since linkage was based on date of birth and Social Security number but not on name. There may be errors in Social Security numbers and dates of birth in either source, although the use of probabilistic record linkage methods, which can produce high matching success even if individual variables do not match exactly, reduces the likelihood that such errors would result in a failed linkage.

One scenario that might produce inaccurate linkage is one in which a cohort member provides her husband's Social Security number instead of her own, although cross-checking with date of birth and sex should have prevented this. There may have been errors in the ICD-9 coding of the teachers' written responses, although indecipherable comments were coded as unknown and therefore would have been included under nonspecific matches. Furthermore, cases in which the teacher's report was in the same diagnosis group as the OSHPD diagnosis were visually reviewed for a match. Coding of self-reports into the wrong diagnosis group would have resulted in a failed match.

With the exception of a few ICD-9 categories such as mental illness or infectious disease, the cohort participants accurately self-reported most of the diagnoses found in the hospital discharge database. The self-reports were also more comprehensive than the hospital records, since they included outpatient surgeries. The self-reports may be subject to overreporting of both events that took place outside of the specified time period and illnesses that did not in fact require surgery or hospitalization. The hospital database avoids the problems of respondent overreporting and was more useful for obtaining exact diagnoses, although prompts can increase the specificity of reports for certain conditions. In addition, hospital and other medical records are available even when the questionnaire is not completed, though such a developed hospital discharge record system may not exist in all places. We conclude that the combination of self-reports and secondary medical records provides more accurate and complete morbidity data than does either source alone.

ACKNOWLEDGMENTS

The California Teachers Study is supported by the California Breast Cancer Act of 1993, the California Department of Health Services, and the National Cancer Institute (grant CA77398).

The authors thank the California State Teachers Retirement System for their help in establishing the cohort.

REFERENCES

1. 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.
2. 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.

3. 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.
4. Hewitt M, Breen N, Devesa S. Cancer prevalence and survivorship issues: analyses of the 1992 National Health Interview Survey. J Natl Cancer Inst 1999;91:1480–6.
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. National Center for Health Statistics. Health interview responses compared with medical records. (Vital and health statistics, series 2, no. 7). Rockville, MD: National Center for Health Statistics, 1965.
7. Madow WG. Net differences in interview data on chronic conditions and information derived from medical records. (Vital and health statistics, series 2, no. 57). Rockville, MD: National Center for Health Statistics, 1973.
8. 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.
9. 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.
10. Haapanen N, Miilunpalo S, 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.
11. 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.
12. Walker MK, Whincup PH, Shaper AG, et al. Validation of patient recall of doctor-diagnosed heart attack and stroke: a postal questionnaire and record review comparison. Am J Epidemiol 1998;148:355–61.
13. Mackenbach JP, Looman CW, Van Der Meer JB. Differences in the misreporting of chronic conditions, by level of education: the effect on inequalities in prevalence rates. Am J Public Health 1996;86:706–11.
14. Barr RG, Herbstman J, Speizer FE, et al. Validation of self-reported chronic obstructive pulmonary disease in a cohort study of nurses. Am J Epidemiol 2002;155:965–71.
15. Karlson EW, Lee I, Cook NR, et al. Comparison of self-reported diagnosis of connective tissue disease with medical records in female health professionals: The Women's Health Cohort Study. Am J Epidemiol 1999;150:652–60.
16. 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.
17. Reijneveld SA. The cross-cultural validity of self-reported use of health care: a comparison of survey and registration data. J Clin Epidemiol 2000;53:267–72.
18. Bernstein L, Allen M, Anton-Culver H, et al. High breast cancer incidence rates among California teachers: results from the California Teachers Study (United States). Cancer Causes Control 2002;13:625–35.
19. Practice Management Information Corporation. International classification of diseases, ninth revision, clinical modification. 6th ed. Los Angeles, CA: Practice Management Information Corporation, 2000.
20. Medicode, Inc. ICD-9 Code It Fast Software for Windows. Salt Lake City, UT: Medicode, Inc, 1999.
21. Bureau of the Census, US Department of Commerce. Census of population and housing, 1990: modified age/race, sex and Hispanic origin (MARS) state and county file (data file). Washington, DC: Bureau of the Census, 1992.
22. California Office of Statewide Health Planning and Development. Patient discharge data file documentation: full calendar year 1998. Sacramento, CA: Office of Statewide Health Planning and Development, 2000.
23. California Office of Statewide Health Planning and Development, Healthcare Information Division, Patient Discharge Section. California patient discharge data reporting manual, third edition, for discharge data for the years 1999 and 2000. Sacramento, CA: Office of Statewide Health Planning and Development, 2000.
24. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York, NY: John Wiley and Sons, Inc, 2000.
25. Norrish A, North D, Kirkman P, et al. Validity of self-reported hospital admission in a prospective study. Am J Epidemiol 1994;140:938–42.
26. Reijneveld SA, Stronks K. The validity of self-reported use of health care across socioeconomic strata: a comparison of survey and registration data. Int J Epidemiol 2001;30:1407–14.
27. Owings MF, Kozak LJ. Ambulatory and inpatient procedures in the United States, 1996. (Vital and health statistics, series 13, no. 139). Hyattsville, MD: National Center for Health Statistics, 1998.
28. Green J, Wintfeld N. How accurate are hospital discharge data for evaluating effectiveness of care? Med Care 1993;31:719–31.