Original Article

J Prev Med Public Health 2017;50:294-302 • https://doi.org/10.3961/jpmph.17.024

pISSN 1975-8375 eISSN 2233-4521

Journal of Preventive Medicine & Public Health

Level of Agreement and Factors Associated With Discrepancies Between Nationwide Medical History Questionnaires and Hospital Claims Data

Yeon-Yong Kim[1], Jong Heon Park[1], Hee-Jin Kang[1], Eun Joo Lee[1], Seongjun Ha[1], Soon-Ae Shin[2]

[1]Big Data Steering Department, National Health Insurance Service, Wonju; [2]Gwanak-Branch, National Health Insurance Service, Seoul, Korea

Objectives: The objectives of this study were to investigate the agreement between medical history questionnaire data and claims data and to identify the factors that were associated with discrepancies between these data types.

Methods: Data from self-reported questionnaires that assessed an individual's history of hypertension, diabetes mellitus, dyslipidemia, stroke, heart disease, and pulmonary tuberculosis were collected from a general health screening database for 2014. Data for these diseases were collected from a healthcare utilization claims database between 2009 and 2014. Overall agreement, sensitivity, specificity, and kappa values were calculated. Multiple logistic regression analysis was performed to identify factors associated with discrepancies and was adjusted for age, gender, insurance type, insurance contribution, residential area, and comorbidities.

Results: Agreement was highest between questionnaire data and claims data based on primary codes up to 1 year before the completion of self-reported questionnaires and was lowest for claims data based on primary and secondary codes up to 5 years before the completion of self-reported questionnaires. When comparing data based on primary codes up to 1 year before the completion of self-reported questionnaires, the overall agreement, sensitivity, specificity, and kappa values ranged from 93.2 to 98.8%, 26.2 to 84.3%, 95.7 to 99.6%, and 0.09 to 0.78, respectively. Agreement was excellent for hypertension and diabetes, fair to good for stroke and heart disease, and poor for pulmonary tuberculosis and dyslipidemia. Women, younger individuals, and employed individuals were most likely to under-report disease.

Conclusions: Detailed patient characteristics that had an impact on information bias were identified through the differing levels of agreement.

Key words: Information bias, Memory decay, Data accuracy, Self-report, Sensitivity and specificity, Kappa statistics

Received: February 11, 2017 Accepted: June 29, 2017

Corresponding author: Soon-Ae Shin, PhD
1485 Nambusunhwan-ro, Gwanak-gu, Seoul 08761, Korea
Tel: +82-2-860-5100, Fax: +82-2-3275-8421
E-mail: sashin513@gmail.com

INTRODUCTION

An individual's medical history provides key information for predicting the prognoses of patients with non-communicable diseases, particularly during initial visits in outpatient clinical settings. Most physicians consider the medical history to be more important than a physical examination or laboratory results [1]. The medical history is also pertinent to research that

deals with risk assessment for morbidity and mortality. Self-reported questionnaires and hospital medical records are the principal sources of collected information on medical history.

The accuracy of reporting an event of interest may be subject to information bias based on the degree of detail required for the respondent to recall, significance of the event to the respondent, social acceptance of the event, time since the event, and memory decay [2,3]. Information bias may reduce the accuracy of self-reported questionnaire data [4]. In several studies comparing data from self-reported questionnaires and patient medical records, the level of agreement was found to differ according to diagnosis, patient characteristics, and study design [5-9]. For example, the rates of agreement in several studies were typically high for patients with diabetes mellitus (DM) but were inconsistent for patients with hypertension, stroke, or myocardial infarction. Rates of agreement also tended to be lower in patients who were older, men, less educated, and had multiple comorbidities [5-8].

In secondary data sources that are used in health research, the medical history is generally obtained through self-reported questionnaires due to cost limitations and privacy concerns with obtaining data directly from hospital records. The accuracy of self-reported responses and the data discrepancies between subgroups of patients are important concerns that have not yet been fully investigated. Previous studies have been limited by small sample sizes (roughly 8000 participants or less) [5-9], and most studies have focused only on middle-aged or elderly patients [5-8]. In South Korea (hereafter Korea), claims data for the entire healthcare system (inpatient, outpatient, emergency, and pharmacy) are collected through a nationwide single insurer, the National Health Insurance Service (NHIS). Claims data from more than 99% of healthcare utilization services in Korea are transferred electronically to the NHIS and the Health Insurance Review and Assessment Service. The NHIS has also launched the National Health Screening Program, which enrolls more than 10 million individuals every year and includes detailed questionnaires on medical history. Screening programs that include a general health screening, cancer screening, transitional age screening (for individuals between 40 and 66 years old), and early childhood screening (for individuals younger than 7 years old) cover the Korean population across all age groups [10].

The aim of the present study was to measure the level of agreement between medical history questionnaires that were collected from a national general health screening database and disease status, which was collected from national health claims data that cover the entire population of Korea. We also investigated the factors that were associated with discrepancies between these data types in order to examine the role of information bias.

METHODS

Data Sources and Study Participants

The NHIS provides nationwide secondary data through the National Health Information Database, which includes an eligibility database, a general health screening database, and a healthcare utilization claims database [11]. The eligibility database contains data on income and other socio-demographic variables for the entire Korean population (roughly 50 million people). The general health screening database contains self-reported questionnaire data on individual and family medical history as well as data on lifestyle and behavior variables (smoking, alcohol consumption, and exercise). Anthropometric measurements (height, weight, waist circumference, and body mass index) and bioclinical laboratory results (e.g., systolic blood pressure and diastolic blood pressure, fasting blood glucose, hemoglobin, cholesterol, and liver enzyme levels) are also included. The entire Korean population 40 years of age or older as well as employed and self-employed individuals who are insured and younger than 40 years old are eligible for biennial screenings. Manual workers are eligible for annual screenings. In 2014, 74.8% of eligible individuals participated in a health screening [12].

A total of 13 281 550 individuals participated in health screenings in 2014. From this sample, those who were missing data, who did not respond to the medical history questionnaire, or who had temporarily lost eligibility during the study period (due to a long trip overseas or military enlistment, for example) were excluded from the analysis. The final analytic sample consisted of 12 668 931 participants.

Variables

Participant data from self-reported questionnaires in 2014 that assessed an individual's history of hypertension, DM, dyslipidemia, stroke, heart disease (myocardial infarction/ischemic heart disease), and pulmonary tuberculosis were collected from the general health screening database. The questions in the general health screening questionnaire asked whether the participants had ever been diagnosed with the aforemen-

tioned diseases. The history of these diseases was collected through primary and secondary diagnosis codes from a healthcare utilization claims database between 2009 and 2014 using the following Korean Standard Classification of Disease codes that originated from the International Classification of Disease, 10th revision: (1) hypertension: I10, I11, I12, I13, I15; (2) DM: E10, E11, E12, E13, E14; (3) stroke: I60, I61, I62, I63, I64; (4) heart disease: I20, I21, I22; (5) dyslipidemia: E78; and (6) pulmonary tuberculosis: A15, A16. The Charlson comorbidity index (CCI) for the status of multiple comorbidities was calculated from the primary and secondary diagnosis codes of all individuals in the healthcare utilization database in 2014.

Statistical Analysis

The overall level of agreement, sensitivity, and specificity were calculated. Sensitivity was defined as the percentage of participants who responded in the questionnaire that they had one of the listed diseases out of all patients in the claims data with that disease. Specificity was defined as the percentage of participants who responded in the questionnaire that they did not have one of the listed diseases out of all patients in the claims data who did not have that disease. Cohen kappa coefficients were also calculated. A kappa value of less than 0.40 was considered a poor level of agreement, a kappa value from 0.40 to 0.74 was considered a fair to good level of agreement, and a kappa value from 0.75 to 1.00 was considered an excellent level of agreement, as suggested by Fleiss [13].

In order to identify the impact of memory decay following diagnosis, claims data were analyzed by dividing the data into 3 overlapping time periods: 2013-2014 (from January 1, 2013 to 1 day before screening), 2011-2014 (from January 1, 2011 to 1 day before screening), and 2009-2014 (from January 1, 2009 to 1 day before screening). The 3 time periods thus included claims data up to 1, 3, and 5 years before the self-reported questionnaires in 2014 were completed. There was no information about the time of diagnosis in the medical history questionnaire, so the time periods were set according to the claims data. The analysis was performed based on 2 categories of diagnosis codes: 1) primary diagnosis codes only and 2) primary diagnosis codes as well as 4 secondary diagnosis codes (for a maximum of 5 diagnosis codes).

In order to investigate the factors associated with discrepancies between questionnaire data and claims data, multiple logistic regression analysis was performed using the primary diagnosis codes in claims data up to 1 year before the self-reported questionnaire. This was done to minimize bias related to memory decay, as the analysis was intended to be focused on the differences between self-reported data and medical claims data. The outcome variables that were calculated were sensitivity (having a disease according to the claims data) and specificity (not having a disease according to the claims data). The independent variables in the model included age, gender, insurance type, insurance contribution, residential area, and CCI score in 2014.

The study protocol was approved by the institutional review board of the NHIS (Sa-2016-HR-02-015). Despite the approval to use the data for research (NHIS-2017-1-019), the results of our study do not represent the official opinion of the NHIS.

RESULTS

General Characteristics of Study Participants

General characteristics of the study participants are presented in Table 1. The largest age group was ≤39 years old (26.2%), followed by 40-49 years old (25.8%), 50-59 years old (24.5%), 60-69 years old (14.4%), and ≥70 years old (9.2%). There were slightly more men (53.4%) than women (46.6%) in the study. The majority of participants were employed and insured (60.7%), followed by dependents of employed individuals (20.2%), self-employed individuals (10.9%), family members of self-employed individuals (7.0%), household Medical Aid beneficiaries (1.0%), and family members of Medical Aid beneficiaries (0.2%). Insurance contributions served as a proxy indicator for income level. The majority of participants were in the third quintile (22.7%), followed by the fourth quintile (22.1%), the second quintile (20.9%), the fifth quintile (19.8%), and the first quintile (14.6%). Roughly 80% of participants had a CCI score of 0 or 1. Most participants lived in a metropolitan city region (47.6%), followed by metropolitan areas (44.3%) and rural areas (8.1%). Hypertension was the most common of the 6 diseases according to both the questionnaire data (18.0%) and claims data (17.1%, primary diagnosis codes from 2013-2014). Stroke was the least common condition in the questionnaire data (0.8%), and pulmonary tuberculosis was the least common condition in the claims data (0.2%, primary diagnosis codes from 2013-2014). In both the questionnaire data and the claims data, DM was the second most common condition (6.6 and 7.0%, respectively), and dyslipidemia was the third most common condition (4.2 and 5.5%, respectively).

Table 1. General characteristics of study participants

Variables	n	%
Total (n)	12 668 931	
Gender		
Men	6 769 570	53.4
Women	5 899 361	46.6
Age (y)		
≤39	3 316 318	26.2
40-49	3 265 034	25.8
50-59	3 103 475	24.5
60-69	1 824 886	14.4
≥70	1 159 218	9.2
Insurance type		
Self-employed individuals	1 385 168	10.9
Family of self-employed individuals	889 863	7.0
Employed individuals	7 690 634	60.7
Dependents of employed individuals	2 552 395	20.2
Medical Aid beneficiaries	125 008	1.0
Family of Medical Aid beneficiaries	25 863	0.2
Income level, quintile (based on insurance contribution)		
1st	1 844 637	14.6
2nd	2 646 843	20.9
3rd	2 871 931	22.7
4th	2 795 547	22.1
5th	2 509 973	19.8
Charlson comorbidity index		
0-1	9 896 606	78.1
≥2	2 772 325	21.9
Region		
Metropolitan city	5 617 821	44.3
City	6 031 389	47.6
Rural	1 019 721	8.1
Diagnosis in questionnaire		
Hypertension	2 278 189	18.0
Diabetes mellitus	840 782	6.6
Stroke	103 282	0.8
Pulmonary tuberculosis	155 587	1.2
Dyslipidemia	530 526	4.2
Heart disease	261 683	2.1
Diagnosis in claims data (2013-2014, primary diagnosis)		
Hypertension	2 167 431	17.1
Diabetes mellitus	890 825	7.0
Stroke	157 809	1.3
Pulmonary tuberculosis	25 958	0.2
Dyslipidemia	693 190	5.5
Heart disease	253 144	2.0

Overall Agreement, Sensitivity, Specificity, and Kappa Statistics

The overall agreement, sensitivity, specificity, and kappa values are presented in Table 2. The claims data based on primary diagnosis codes up to 1 year before the self-reported questionnaire showed higher levels of agreement with questionnaire data than did claims data based on primary and secondary diagnosis codes up to 5 years before self-reported data. Specificity was highest when comparing the questionnaire data with claims data based on primary and secondary diagnosis codes up to 5 years prior to the self-reported data. When comparing data based on primary diagnosis codes up to 1 year before the self-reported data, the overall agreement, sensitivity, specificity, and kappa values for the 6 diseases ranged from 93.2 to 98.8%, 26.2 to 84.3%, 95.7 to 99.6%, and 0.09 to 0.78, respectively. When comparing data based on primary and secondary diagnosis codes up to 5 years before the self-reported data, the overall agreement, sensitivity, specificity, and kappa values for the 6 diseases ranged from 67.4 to 98.0%, 10.7 to 66.3%, 99.0 to 99.8%, and 0.13 to 0.73, respectively.

When comparing different diseases based on primary diagnosis codes up to 1 year before self-reported data, the overall agreement and specificity were highest for stroke (98.8 and 99.6%, respectively), and the sensitivity and kappa value were highest for hypertension (84.3 and 0.78, respectively). The level of agreement was excellent for hypertension (0.78) and DM (0.76), fair to good for stroke (0.41) and heart disease (0.48), and poor for pulmonary tuberculosis (0.09) and dyslipidemia (0.26).

Logistic Regression Analysis for Discrepancies Between Data Types

The results of the multiple logistic regression analysis that identified discrepancies between the questionnaire data and the claims data regarding participant characteristics are presented in Tables 3 and 4. Men patients had lower odds of a negative response on the questionnaire, except for hypertension and dyslipidemia (adjusted odds ratios [aOR], 0.64 to 0.82). Similarly, patients aged 70 years or older had lower odds of a negative response on the questionnaire than those aged 39 years or younger, except for pulmonary tuberculosis (aOR, 0.16 to 0.99). Patients who were employed had higher odds of a negative response on the questionnaire than those who were unemployed (aOR, 1.07 to 1.53). Lastly, patients who lived in a metropolitan city had lower odds of a negative response on the questionnaire than those who lived in rural ar-

Table 2. Overall agreement, sensitivity, specificity, and Cohen kappa statistics according to the diagnosis code and time period of claims data

Year/disease diagnosis code	Overall agreement		Sensitivity		Specificity		Cohen kappa statistic	
	Primary	5 Codes[1]	Primary	5 Codes[1]	Primary	5 Codes[1]	Primary	5 Codes[1]
Hypertension								
2009-2014	93.03 (93.02, 93.05)	90.60 (90.59, 90.62)	75.38 (75.33, 75.43)	66.26 (66.21, 66.31)	97.93 (97.92, 97.94)	99.33 (99.32, 99.33)	0.782 (0.781, 0.782)	0.731 (0.730, 0.731)
2011-2014	93.63 (93.61, 93.64)	92.30 (92.29, 92.32)	79.22 (79.17, 79.27)	70.98 (70.93, 71.03)	97.20 (97.19, 97.21)	99.22 (99.22, 99.24)	0.793 (0.792, 0.793)	0.772 (0.771, 0.772)
2013-2014	93.73 (93.72, 93.75)	94.12 (94.11, 94.13)	84.25 (84.20, 84.30)	77.15 (77.10, 77.20)	95.70 (95.68, 95.71)	98.98 (98.98, 98.99)	0.784 (0.783, 0.784)	0.818 (0.817, 0.818)
Diabetes mellitus								
2009-2014	95.49 (95.48, 95.50)	87.80 (87.79, 87.83)	60.71 (60.63, 60.80)	34.96 (34.90, 35.02)	99.33 (99.32, 99.33)	99.78 (99.78, 99.79)	0.704 (0.704, 0.705)	0.462 (0.461, 0.463)
2011-2014	96.24 (96.23, 96.25)	90.38 (90.36, 90.39)	66.29 (66.20, 66.38)	40.56 (40.49, 40.63)	99.14 (99.13, 99.14)	99.75 (99.75, 99.75)	0.737 (0.736, 0.737)	0.528 (0.527, 0.528)
2013-2014	96.89 (96.88, 96.90)	93.87 (93.86, 93.88)	75.09 (75.00, 75.18)	52.08 (52.00, 52.16)	98.54 (98.53, 98.55)	99.65 (99.65, 99.66)	0.756 (0.755, 0.757)	0.643 (0.642, 0.644)
Stroke								
2009-2014	97.96 (97.96, 97.97)	96.78 (96.78, 96.79)	24.25 (24.10, 24.40)	17.22 (17.12, 17.33)	99.75 (99.75, 99.76)	99.81 (99.81, 99.81)	0.353 (0.351, 0.355)	0.272 (0.271, 0.274)
2011-2014	98.37 (98.37, 98.38)	97.50 (97.48, 97.50)	28.09 (27.90, 28.27)	20.45 (20.32, 20.58)	99.70 (99.70, 99.70)	99.76 (99.76, 99.77)	0.383 (0.381, 0.385)	0.310 (0.308, 0.311)
2013-2014	98.80 (98.80, 98.81)	98.27 (98.26, 98.27)	34.75 (34.51, 34.98)	26.35 (26.18, 26.53)	99.61 (99.61, 99.62)	99.69 (99.69, 99.69)	0.414 (0.412, 0.417)	0.364 (0.362, 0.366)
Pulmonary tuberculosis								
2009-2014	98.42 (98.42, 98.43)	97.98 (97.97, 97.99)	24.33 (24.05, 24.62)	16.66 (16.47, 16.85)	98.93 (98.92, 98.93)	98.96 (98.95, 98.96)	0.166 (0.163, 0.168)	0.153 (0.152, 0.155)
2011-2014	98.58 (98.58, 98.60)	98.31 (98.30, 98.31)	27.76 (27.38, 28.15)	18.48 (18.24, 18.73)	98.88 (98.88, 98.89)	98.90 (98.90, 98.91)	0.134 (0.132, 0.136)	0.131 (0.130, 0.133)
2013-2014	98.71 (98.70, 98.71)	98.57 (98.57, 98.58)	33.56 (32.99, 34.14)	22.93 (22.54, 23.32)	98.84 (98.83, 98.84)	98.85 (98.84, 98.86)	0.093 (0.091, 0.095)	0.999 (0.097, 0.101)
Dyslipidemia								
2009-2014	88.19 (88.17, 88.21)	67.44 (67.42, 67.47)	18.88 (18.82, 18.94)	10.72 (10.69, 10.74)	97.86 (97.85, 97.87)	99.50 (99.50, 99.51)	0.234 (0.233, 0.234)	0.127 (0.126, 0.127)
2011-2014	90.20 (90.18, 90.22)	71.97 (71.94, 71.99)	20.84 (20.76, 20.91)	12.07 (12.04, 12.11)	97.58 (97.57, 97.59)	99.43 (99.42, 99.43)	0.246 (0.245, 0.247)	0.150 (0.150, 0.151)
2013-2014	93.20 (93.20, 93.22)	79.61 (79.59, 79.63)	26.20 (26.09, 26.30)	15.40 (15.35, 15.44)	97.09 (97.08, 97.10)	99.24 (99.23, 99.24)	0.262 (0.261, 0.263)	0.205 (0.204, 0.205)
Heart disease								
2009-2014	96.65 (96.63, 96.65)	93.22 (93.21, 93.24)	33.45 (33.32, 33.58)	20.13 (20.05, 20.21)	99.21 (99.20, 99.21)	99.48 (99.48, 99.49)	0.422 (0.421, 0.423)	0.296 (0.295, 0.297)
2011-2014	97.29 (97.28, 97.30)	94.85 (94.84, 94.86)	39.35 (39.20, 39.51)	24.61 (24.52, 24.71)	99.10 (99.09, 99.10)	99.39 (99.39, 99.40)	0.454 (0.452, 0.456)	0.347 (0.346, 0.349)
2013-2014	97.93 (97.92, 97.94)	96.54 (96.53, 96.55)	49.83 (49.63, 50.02)	32.98 (32.86, 33.11)	98.91 (98.90, 98.91)	99.26 (99.25, 99.26)	0.479 (0.478, 0.481)	0.423 (0.421, 0.424)

Values are presented as odds ratio (95% confidence interval).
[1]Primary diagnosis codes in addition to 4 secondary diagnosis codes.

Table 3. Adjusted logistic regression results for the tendency to provide negative responses on the questionnaire and the presence of a positive history in the claims data[1]

	Hypertension	Diabetes mellitus	Stroke	Heart disease	Dyslipidemia	Pulmonary tuberculosis
Total (n)	2 167 431	890 825	157 809	253 144	693 190	25 958
Gender (reference: women)						
Men	1.17*	0.82*	0.68*	0.64*	1.44*	0.82*
Age (reference: 60-69, y)						
≤39	3.21*	1.91*	1.18*	5.64*	2.16*	0.36*
40-49	1.58*	1.11*	0.83*	2.00*	1.55*	0.51*
50-59	1.27*	1.07*	0.89*	1.32*	1.26*	0.77*
≥70	1.01*	1.09*	1.17*	0.88*	1.08*	1.39*
Employed (reference: no)						
Yes	1.14*	1.23*	1.53*	1.22*	1.07*	1.11*
Charlson comorbidity index (reference: 0-1)						
≥2	1.04*	0.45*	0.81*	0.83*	1.02*	1.07*
Insurance contribution (reference: 4th, quintile)						
1st	1.06*	0.98*	0.95*	1.01	1.09*	1.11*
2nd	1.01*	0.99	1.03	1.07*	1.10*	1.19*
3rd	1.02*	1.01	1.02	1.02	1.07*	1.06
5th	1.03*	1.03*	1.03	1.02	0.87*	0.94
Region (reference: metropolitan city)						
City	1.01*	1.03*	1.09*	1.02*	1.01	1.08*
Rural	1.04*	1.06*	1.17*	1.07*	1.14*	1.49*

[1]We used claims data from 2013-2014 with primary diagnosis codes.
*$p<0.05$.

Table 4. Adjusted logistic regression results for tendency of positive responses in questionnaire and negative history in claims data[1]

	Hypertension	Diabetes mellitus	Stroke	Heart disease	Dyslipidemia	Pulmonary tuberculosis
Total (n)	10 501 500	11 778 106	12 511 122	12 415 787	11 975 741	12 642 973
Gender (reference: women)						
Men	1.66*	1.68*	1.40*	1.16*	0.83*	1.75*
Age (reference: ≤39, y)						
40-49	1.69*	3.19*	2.24*	1.64*	2.35*	1.87*
50-59	3.59*	6.53*	5.18*	3.89*	4.65*	2.11*
60-69	8.31*	11.38*	9.88*	8.62*	8.37*	2.28*
≥70	14.50*	14.18*	18.43*	16.33*	7.71*	2.09*
Employed (reference: yes)						
No	1.10*	1.09*	2.03*	1.37*	1.05*	0.94*
Charlson comorbidity index (reference: 0-1)						
≥2	4.49*	5.71*	2.82*	2.50*	2.42*	0.93*
Insurance contribution (reference: 3rd, quintile)						
1st	1.08*	1.10*	1.33*	1.09*	1.03*	1.01
2nd	1.00	1.03*	1.05*	0.99	0.97*	0.98*
4th	1.01*	1.00	1.00	1.03*	1.10*	1.09*
5th	1.06*	0.98*	0.96*	1.05*	1.39*	1.26*
Region (reference: rural)						
Metropolitan city	1.19*	1.24*	0.94*	0.95*	1.47*	1.71*
City	1.19*	1.20*	1.01	0.97*	1.42*	1.50*

[1]We used claims data from 2013-2014 with primary diagnosis codes.
*$p<0.05$.

eas (aOR, 0.67 to 0.96). Among patients not identified by the claims data, the following groups showed greater odds of a positive response on the questionnaire: men (the aOR for men compared to women ranged from 1.16 to 1.75, except for dyslipidemia), older individuals (the aOR for individuals aged 70 years or older compared to individuals aged 39 or younger ranged from 2.09 to 18.43), participants who were unemployed (the aOR for employed individuals compared to unemployed individuals ranged from 0.49 to 0.92, except for dyslipidemia and pulmonary tuberculosis), and those who lived in a non-rural area (the aOR for those living in a metropolitan city compared to a rural area ranged from 1.19 to 1.71, except for stroke and heart disease).

DISCUSSION

In this nationwide study including self-reported data and claims data for more than 12 million individuals, the level of agreement between the questionnaire data and medical claims data varied by disease and participant characteristics. Based on kappa statistics, the level of agreement was excellent for hypertension and DM, fair to good for stroke and heart disease, and poor for tuberculosis and dyslipidemia. Women, younger participants, and those who were employed were most likely to under-report their disease status, whereas men, older participants, and those who were unemployed tended to over-report their disease status.

The discrepancy between the self-reported disease data and claims data also differed by diagnosis time and the priority of the disease codes in the claims data. Agreement was greatest between the questionnaire data and the claims data that contained more recent diagnoses and were based on primary diagnosis codes. Six major factors have been identified as possible influences on memory decay, including characteristics related to the event of interest (recency, attributes, and complexity) and characteristics related to the context of the event of interest (salience, patient experience, and mood) [14]. The effects of recency were confirmed by the analysis of different diagnosis periods, while the importance of salience was supported by results for the different categories of diagnosis codes. Our results regarding the time lag between diagnosis and self-reported disease were consistent with the results of

previous studies that investigated information bias in relation to the timing of memory decay [15].

Hypertension and DM exhibited high levels of agreement between self-reported data and medical claims data. These findings were consistent with results from previous studies of DM [5-9]. Stroke and heart disease showed fair to good levels of agreement in our study, a finding that is also consistent with the results of previous studies on hypertension [8,9]. The lowest levels of agreement in our study were observed for pulmonary tuberculosis and dyslipidemia. The poor levels of agreement for these conditions may have stemmed from several factors. Patients with infectious diseases, such as tuberculosis and human immunodeficiency virus, are frequently exposed to societal stigma [16]. This may partly explain the low sensitivity observed for pulmonary tuberculosis, as patients may have been reluctant to report their disease status. Additionally, dyslipidemia may also be referred to as hyperlipidemia, hypertriglyceridemia, and hypercholesterolemia. The use of these various terms may affect an individual's recognition of dyslipidemia as well as adherence to medication [17], which in turn could plausibly reduce the accuracy of self-reported disease status.

Employed participants were more likely to under-report their own disease status than unemployed participants. Since questionnaire data were obtained through the national health screening program in Korea, responses may be have been influenced by employment status. Although confidentiality concerns prevent data obtained from an individual during a general health screening from being reported to a company's health manager, the health manager may be notified of rates of disease at an aggregated company level. Thus, participants may still choose to hide their disease status in order to avoid further evaluation.

The prevalence and severity of chronic disease were notably lower among younger individuals than older individuals. This could potentially be attributed to younger individuals who take their diagnoses of a chronic disease and their disease status less seriously than older individuals, which may lead to a decrease in disease awareness and decreased adherence to medication [18]. Conversely, a tendency to over-report was observed among the elderly participants. However, the limit of 5 diagnosis codes in the claims data could have possibly excluded even more secondary diagnosis codes for hospitalized elderly patients than younger patients.

Women were more likely to under-report disease status than men in our investigation. This trend has been observed in previous studies [6-8]. However, studies investigating pain sensitivity and activity limitations found that women were more sensitive to symptoms than men, and it has been hypothesized that this is due in part to cultural-specific gender roles [19,20].

Various methods, including self-reported questionnaires and laboratory testing, are used to calculate the prevalence of disease at the population level. As a result, prevalence rates calculated from the same population over the same time period may differ depending on the data sources used. For example, the prevalence of DM in a study in the US was found to be 4.1% in men and 5.6% in women based on survey questionnaire data from the Behavioral Risk Factor Surveillance System in 1990 [21]. Conversely, the prevalence of DM in another study from the US was estimated as 8.4% in men and 7.7% in women based on laboratory tests and survey questionnaire data from the National Health and Nutrition Examination Survey (NHANES) for the time period of 1988-1994 [22]. We found a higher prevalence for DM in women than in men according to survey questionnaire data, which may partly stem from under-reporting by men and over-reporting by women. In the NHANES that was conducted in the US, hypertension was defined as a systolic blood pressure $\geq$140 mmHg, a diastolic blood pressure $\geq$90 mmHg, or current use of prescription medication to lower blood pressure. In one study using this definition, the prevalence of hypertension did not differ between men and women, but rates of treatment and awareness, which were based solely on questionnaire data, were higher among women than men [23]. In light of different reporting patterns across subgroups, these findings highlight the importance of equity issues that are related to gender, age, and employment status. Our results suggest that prevalence rates that are calculated from questionnaire data may underestimate the true disease prevalence among women, younger individuals, and those who are employed. This may lead to bias away from the null in epidemiologic studies [4], which could reduce internal validity by exhibiting larger differences than those that truly exist. Furthermore, under-reporting of disease may lead policy makers to neglect opportunities for intervention based on the reporting of inaccurate prevalence rates.

In light of the reporting patterns among patients, more precise diagnoses and treatments are required to achieve good prognoses for those who under-report disease. Estimating accurate, national-level descriptive statistics is essential for es-

tablishing effective and equitable health policies. Our study identified differing levels of agreement between self-reported data and claims data according to disease type, diagnosis period, and patient characteristics using a large, nationwide, population-based data set. Our study was also able to identify more patient characteristics that were associated with information bias than previous studies. However, there were several limitations to our research. First, the accuracy of diagnoses may be influenced by the fee-for-service reimbursement system, which can result in upcoding in claims data [24]. This means that claims data may not be reliable and that there was a possibility that both the questionnaire data and the claims data were inaccurate. However, we applied various analytic methods to the claims data in order to overcome this limitation. Second, we did not consider the impact of medication. While medication was not a primary concern in this study, there was a potential interaction between reporting a diagnosed disease and history of medication treatment.

CONFLICT OF INTEREST

The authors have no conflicts of interest associated with the material presented in this paper.

ORCID

Yeon-Yong Kim *https://orcid.org/0000-0003-2179-8931*
Jong Heon Park *https://orcid.org/0000-0002-4749-5878*
Hee-Jin Kang *https://orcid.org/0000-0003-2788-6262*
Eun Joo Lee *https://orcid.org/0000-0002-4294-9471*
Seongjun Ha *https://orcid.org/0000-0001-9664-0827*
Soon-Ae Shin *https://orcid.org/0000-0001-8858-0801*

REFERENCES

1. Rich EC, Crowson TW, Harris IB. The diagnostic value of the medical history. Perceptions of internal medicine physicians. Arch Intern Med 1987;147(11):1957-1960.
2. Coughlin SS. Recall bias in epidemiologic studies. J Clin Epidemiol 1990;43(1):87-91.
3. Murray JK, Singer ER, Morgan KL, Proudman CJ, French NP. Memory decay and performance-related information bias in the reporting of scores by event riders. Prev Vet Med 2004;63(3-4):173-182.
4. Raphael K. Recall bias: a proposal for assessment and control. Int J Epidemiol 1987;16(2):167-170.
5. Simpson CF, Boyd CM, Carlson MC, Griswold ME, Guralnik JM, Fried LP. Agreement between self-report of disease diagnoses and medical record validation in disabled older women: factors that modify agreement. J Am Geriatr Soc 2004;52(1):123-127.
6. Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, Rodeheffer RJ. Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. J Clin Epidemiol 2004;57(10):1096-1103.
7. Haapanen N, Miilunpalo S, Pasanen M, Oja P, Vuori I. Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. Am J Epidemiol 1997;145(8):762-769.
8. Chun H, Kim IH, Min KD. Accuracy of self-reported hypertension, diabetes, and hypercholesterolemia: analysis of a representative sample of Korean older adults. Osong Public Health Res Perspect 2016;7(2):108-115.
9. Bergmann MM, Jacobs EJ, Hoffmann K, Boeing H. Agreement of self-reported medical history: comparison of an in-person interview with a self-administered questionnaire. Eur J Epidemiol 2004;19(5):411-416.
10. National Health Insurance Service. Health insurance guide [cited 2016 Dec 26]. Available from: http://www.nhis.or.kr/static/html/wbd/g/a/wbdga0606.html.
11. Cheol Seong S, Kim YY, Khang YH, Park JH, Kang HJ, Lee H, et al. Data resource profile: the national health information database of the National Health Insurance Service in South Korea. Int J Epidemiol 2016;pii:dyw253.
12. National Health Insurance Service. 2014 National health screening statistical yearbook. Seoul: National Health Insurance Service; 2015, p. 5 (Korean).
13. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981, p. 218.
14. Stull DE, Leidy NK, Parasuraman B, Chassany O. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Curr Med Res Opin 2009;25(4):929-942.
15. Muntner P, Coresh J, Klag MJ, Whelton PK, Perneger TV. History of myocardial infarction and stroke among incident end-stage renal disease cases and population-based controls: an analysis of shared risk factors. Am J Kidney Dis 2002;40(2):323-330.
16. Weiss MG, Ramakrishna J, Somma D. Health-related stigma: rethinking concepts and interventions. Psychol Health Med

2006;11(3):277-287.
17. Brewer NT, Chapman GB, Brownlee S, Leventhal EA. Cholesterol control, medication adherence and illness cognition. Br J Health Psychol 2002;7(Part 4):433-447.
18. Morris AB, Li J, Kroenke K, Bruner-England TE, Young JM, Murray MD. Factors associated with drug adherence and blood pressure control in patients with hypertension. Pharmacotherapy 2006;26(4):483-492.
19. Khadr Z, Yount K. Differences in self-reported physical limitation among older women and men in Ismailia, Egypt. J Gerontol B Psychol Sci Soc Sci 2012;67(5):605-617.
20. Merrill SS, Seeman TE, Kasl SV, Berkman LF. Gender differences in the comparison of self-reported disability and performance measures. J Gerontol A Biol Sci Med Sci 1997;52(1):M19-M26.
21. Mokdad AH, Ford ES, Bowman BA, Nelson DE, Engelgau MM, Vinicor F, et al. Diabetes trends in the U.S.: 1990-1998. Diabetes Care 2000;23(9):1278-1283.
22. Harris MI, Flegal KM, Cowie CC, Eberhardt MS, Goldstein DE, Little RR, et al. Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. The Third National Health and Nutrition Examination Survey, 1988-1994. Diabetes Care 1998;21(4):518-524.
23. Egan BM, Zhao Y, Axon RN. US trends in prevalence, awareness, treatment, and control of hypertension, 1988-2008. JAMA 2010;303(20):2043-2050.
24. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res 2005;40(5 Pt 2):1620-1639.