hpm 보건행정학회지 2013;23(2):152–161 | ISSN 1225-4266
Health Policy and Management Vol.23 No.2, 152-161
http://dx.doi.org/10.4332/KJHPA.2013.23.2.152

ORIGINAL ARTICLE

입원 환자 표본 개발에 관한 연구: 국민건강보험 청구자료를 중심으로

김록영[1] · 사공진[2] · 김 윤[1] · 김세라[1] · 김수경[1] · 최병호[3] · 정형선[4] · 이태림[5]

[1]건강보험심사평가원, [2]한양대학교 경제학부, [3]한국보건사회연구원, [4]연세대학교 보건행정학과, [5]한국방송통신대학교 정보통계학과

Developing the Inpatient Sample for the National Health Insurance Claims Data

Logyoung Kim[1], Jin Sakong[2], Yoon Kim[1], Sera Kim[1], Sookyeong Kim[1], Byongho Tchoe[3], Hyoungsun Jeong[4], Taerim Lee[5]

[1]Health Insurance Review and Assessment Service; [2]Department of Economics, Hanyang University; [3]Korea Institute for Health and Social Affairs; [4]Department of Health Administration, Yonsei University; [5]Department of Information Statistics, Korea National Open University, Seoul, Korea

Korea has a single National Health Insurance program and all citizens are covered under this program, accounting 97% of the population, approximately 50 million people. Claims submitted by Health care providers are reviewed by Health Insurance Review and Assessment (HIRA) for the reimbursement. HIRA database contains not only individual beneficiary's information, but also healthcare service information such as diagnosis, procedures, prescriptions and tests for them. HRA database has gained attention as importance source for research due to its rich healthcare information and the demand of HIRA database has increased. Due to its tremendous size, however, researchers have had problems in accessing the database to conduct research. To meet this demand, we conducted a study to develop the inpatient sample data from HIRA database for research. This study has two purposes: 1) to determine a needed sample size; 2) to test reliability and validity of the sample data. We determined an adequate sample size to ensure representativeness and generality with additional consideration for convenience of calculation. The minimum sample size was 729,904 for the generality, and 488,861 for representativeness. After considering the convenience of calculation, our final sample size was 13% of the population, which was about 7.7 million beneficiaries. Age (5 years interval) and gender were used as stratification variables for sampling. In order to examine whether this sample data appropriately reflect population, we tested the reliability and validity of the sample data. From the sample data, we computed average expenditure of total claims per inpatient for 2011, frequency of top 30 disease, estimation of the number of stroke patients from the sample data, and then compared them to those from the population. Results confirmed reliability and validity of the sample data .

Keywords: Administrative data; National Health Insurance; Claims data; Inpatient; Sampling; Generality; Reliability; Validity

서 론

우리나라는 1960년대와 1970년대에는 표본자료 설계에 대한 구체적인 체계가 잡히지 않은 상태였고 1990년대에 통계청에서 보관하고 있는 조사구 자료가 일부 제공되면서 비교적 과학적인 표본설계를 하게 되었다. 보건의료 분야에서 2000년대에 실시한 대규모 표본조사자료는 약 10개 정도였고, 2010년에 이르러 약 40개가 넘는 조사자료와 패널조사자료를 구축하게 되었다.

최근까지의 보건의료 분야에서 조사자료의 다양화는 기초자료의 중요성이 부각되고 있음을 보여주는 예라고 할 수 있으나 보건의료 관련 기초자료는 주로 실사를 바탕으로 한 조사자료가 주를 이루고 있었으며 건강보험 청구자료에 대한 직접 제공방안은 마련되지 않고 있었다. 우리나라는 현재 전 국민의 98%가 국민건강보험에 가입되어 있어 건강보험 청구자료는 우리나라 보건의료를 대표

Correspondence to: Logyoung Kim
Health Insurance Review and Assessment Service, 267 Hyoryeong-ro, Seocho-gu, Seoul 137-706, Korea
Tel: +82-2-2182-2515, Fax: +82-2-6710-7610, E-mail: kimlog2@hiramail.net
*본 논문은 2010년 진행된 건강보험심사평가원의 '진료정보의 표본자료 제공 방안에 관한 연구'의 사례를 활용하여 작성되었음.
Received: February 27, 2013 / Accepted after revision: April 20, 2013

하는 자료라고 할 수 있다. 우리나라는 1989년에 전 국민 의료보험을 시작하여 2000년 국민의료보험공단이 직장의료보험조합과 통합되면서 조합마다 다르게 운영되던 건강보험이 국민건강보험으로 통합하여 출범하게 되었고, 동시에 전산기술의 발달로 건강보험 급여비의 전산 청구율은 2001년 90%에서 꾸준히 증가하여 2005년부터 99% 이상으로 장기간의 데이터가 누적되고 있다.

우리나라의 국민건강보험 청구자료는 제한적 실험 환경이 아닌 국민건강보험체계하의 실제 보건의료 환경을 반영하는 데이터이므로 비교적 일반화가 용이하며 이미 구축된 자료를 활용함으로써 연구에 소요되는 시간과 비용 등을 단축시킬 수 있다. 국민건강보험 청구자료는 보건의료 분야의 국가정책수립 및 국민의 건강증진에 관련한 연구에 기초자료로 활용되어질 수 있기 때문에 다양한 분야에서 건강보험 청구자료에 대한 수요가 급증하고 있다.

우리나라와 유사한 건강보험체계를 가지고 있는 대만은 1995년 3월 1일 단일 국가의료보험을 시작하였으며 2007년 기준으로 대만의 전체 인구 중 98.4%가 등록되어 있다. 이렇게 국가의료보험으로부터 수집된 자료는 대만 '전민건강보험 데이터베이스(National Health Insurance Research Database, NHIRD)'에 축적되어지며, 대만의 건강보험국(Bureau of National Health Insurance) 산하 국립보건연구소(National Health Research Institutes)의 관리하에 구축되어 연구용도로 활용되어진다. 대만의 NHIRD는 의료급여비용 상환을 위한 '보험자 등록자료'와 '건강보험 청구자료'를 포함하고 있으며, 청구 건 기준으로 추출되는 월 단위 표본자료와 함께 환자기준으로 추출되는 1년 단위 표본자료부터, 5년 단위 패널자료까지 다양한 방식으로 건강보험 청구자료에 대한 표본자료를 구축하고 있다. 월 단위 표본자료는 주로 계절에 민감하거나 유행성 질병에 시의적절하게 대응하기 위한 단기 분석자료로 활용되고 있으며 1년 단위 혹은 5년 단위의 패널자료는 주로 연구용으로 활용되고 있다[1].

미국의 Agency for Healthcare Research and Quality (AHRQ)는 연방정부에 속한 연구기관으로 보건의료서비스 분야에 관련된 연구를 수행 · 지원한다. AHRQ는 37개 주정부 및 지역사회, 보건의료 산업체들로부터 데이터를 수집하여 의료 데이터베이스를 구축하고 있다. AHRQ의 조사 프로그램 중 하나인 Healthcare Cost and Utilization Project (HCUP)는 미국에서 가장 큰 보건의료 데이터베이스를 구축하고 있고, HCUP의 제공자료 중 가장 포괄적인 전국 입원 환자 표본자료(National Inpatient Sample, NIS)는 재활의료기관을 제외한 미국병원협회(American Hospital Association)에 속해있는 모든 의료 커뮤니티를 포함한다. NIS는 커뮤니티에 가입된 37개 주의 약 3,900개의 의료기관으로부터 수집된 데이터를 기반으로 하고 있으며, 가입 의료기관 중 매년 약 20% (800-1,100개 기관)를 표본 추출하여 추출된 의료기관의 전체 입원 자료(약 5백만-8백만 입원 건)를 포함하고 있다[2,3].

우리나라의 경우 건강보험심사평가원(심평원)과 국민건강보험공단에서 자료처리실을 운영하고 있어 건강보험 청구자료에 대하여 외부 연구자가 직접 자료를 가공하여 결과를 산출하도록 하고 있으나 직접 내방하여 이용해야 하는 번거로움으로 접근성과 편의성 측면에서 한계가 존재하며, 연간 약 10억 건 이상의 방대한 용량의 자료는 사용자의 저장용량, 처리속도 등 수용능력의 한계로 인하여 시의적절한 자료 확보를 불가능하게 한다. 따라서 다양한 수요층에 대한 접근성과 편의성, 즉시성의 확보를 위한 대안의 하나로 우리나라의 건강보험 청구자료에 대한 표본자료를 개발하게 되었다.

Table 1에서 미국의 표본자료는 메디케어(Medicare), 메디케이드(Medicaid), 민간보험, 무보험 환자의 퇴원자료로 구축되어 비급여 부분까지 포함되어 있으며, 환자가 특정 의료기관에 입원하고 퇴원하기까지를 한 단위로 하는 자료이므로 환자가 재입원하거나 다른 의료기관으로 이환된 경우에는 동일인 구분이 되지 않는다. 우리나라와 대만의 표본자료는 1년간 환자의 의료이용내역을 포함하는 자료이므로 급여가 보장되는 범위에서는 재입원하거나 이환하더라도 동일인 구분이 가능하다. 우리나라의 국민건강보험 입원 환자 표본자료 개발의 목적은 건강보험 청구자료에 대한 접근성 및 활용도를 높여 보건의료 관련 연구를 활성화시키기 위함이다.

모집단 개요

1. 건강보험 청구자료의 개요 및 수집방식

환자 표본자료의 대상 모집단이 되는 건강보험 청구자료란 의료기관에서 환자의 진료비용 중 국민건강보험이 부담하는 부분에 대해 지급의뢰를 하기 위해 심평원에 청구하는 자료이다. 국민건강보험공단은 국민을 대상으로 국민건강보험 재정에 대한 징수, 관리 및 지급업무를 담당하며, 심평원은 의료기관에서 청구하는 청구자

Table 1. Comparison of the nations sample dataset

Country	Republic of Korea (HIRA)	USA (AHRQ)	Twain (NHRI)
Unit of sampling	Patient	Hospital	Patient
Unit of provision	Patient level	Episode of care by healthcare providers (discharge information)	Patient level
Variables for stratified sampling	Demographic characteristics	Hospital characteristics	Simple random sampling
		Geographic location	

HIRA, Health Insurance Review and Assessment Service; AHRQ, Agency for Healthcare Research and Quality; NHRI, National Health Research Institutes.

료에 대한 진료비 심사결과를 국민건강보험공단에 전달하여 지급을 요청한다.

우리나라의 1년간 건강보험 청구 환자 수는 2011년 기준 45,804,866명으로 주민등록인구 50,734,284명의 90.3%에 달하는 수치이며, 건강보험 청구에 대한 심사건수와 청구 진료비 총액은 꾸준히 증가하여 2011년 기준으로 심사건수는 약 13억 건, 총 진료비는 약 51조 5천억 원에 달하고 있다(Figure 1). 건강보험에 등록된 요양기관은 1980년대 7,289개소에서 2011년에 와서는 82,948개소로 증가하였다(Figure 2) [4].

병원에서는 환자를 진료하고 급여상환을 위해 심평원으로 요양급여비용을 청구하게 된다. 요양기관으로부터 진료비 청구명세서가 접수되면, 전산점검 및 심사단계를 거쳐 데이터가 누적된다. 심사처리절차를 그림으로 표시하면 Figure 3과 같다.

2. 모집단 현황

Figure 3에서와 같이 축적된 데이터는 data warehouse 시스템으로 저장되어, 급여적정성평가나 통계분석, 연구자료로 활용된다. 본 연구에서 선정된 대상 모집단은 2011년 전체 입원 환자[1] 5,889,784명이다(Table 2).

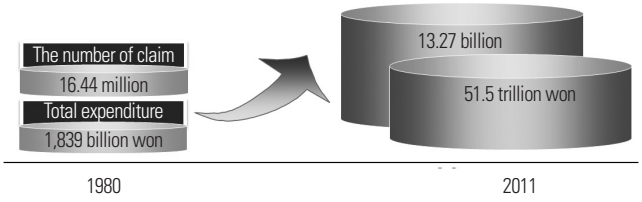

Figure 1. Total expenditure and the number of claim data review on National Health Insurance in 2011. Reprinted from Health Insurance Review and Assessment Service. Guidelines for review and assessment for healthcare services of 2011. Seoul: Health Insurance Review and Assessment Service; 2011 [4].

모집단은 입원 환자 1년간의 모든 진료, 처치, 처방내역을 포함하고 있으며, 대상이 되는 자료는 크게 요양급여비용 청구명세서자료와 요양기관 현황자료로 구성된다. 요양급여비용 청구명세서자료란 의료기관 및 약국 등에서 환자에게 진료 또는 조제한 후 요양급여비용 청구방법에 따라 작성한 진료내역이 기재된 자료이다. 이 명세서는 electronic data interchange, 전산매체(디스켓, CD). 또는 서면으로 청구 가능하며, 건강보험 환자, 의료급여 환자, 보훈국비 환자의 진료비 청구내용을 모두 포함하고 있다(Table 3) [5].

요양기관 현황자료는 심평원에 청구한 요양급여비용을 심사·평

Table 2. The number of inpatients by age group in 2011

Age groups	Gender		Sum
	Man	Woman	
1-4	163,266	136,499	299,765
5-9	90,589	72,740	163,329
10-14	79,925	50,538	130,463
15-19	111,485	72,858	184,343
20-24	92,869	106,024	198,893
25-29	121,746	251,813	373,559
30-34	144,228	351,767	495,995
35-39	162,611	224,121	386,732
40-44	189,190	205,487	394,677
45-49	213,686	237,513	451,199
50-54	253,496	282,936	536,432
55-59	216,078	223,277	439,355
60-64	193,540	194,635	388,175
65-69	181,924	205,897	387,821
70-74	170,885	228,825	399,710
≥ 75	226,360	432,976	659,336
Total	2,611,878	3,277,906	5,889,784

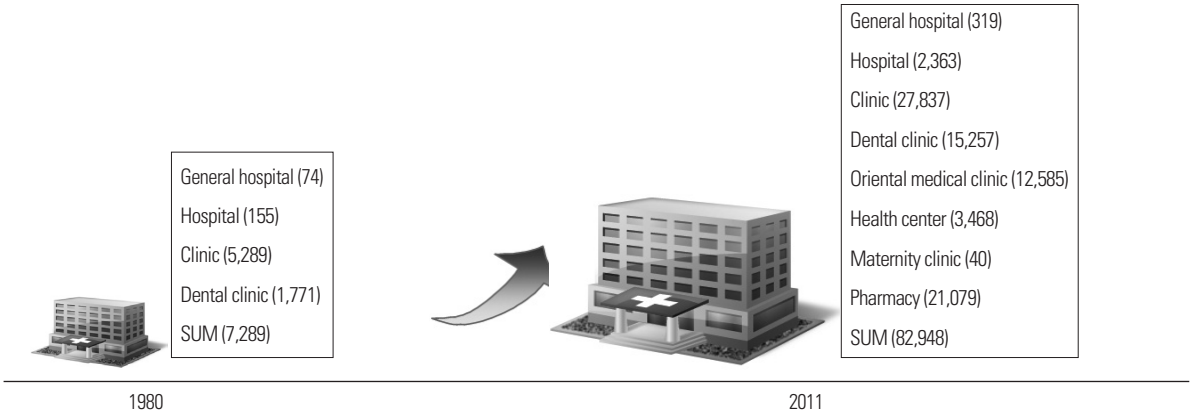

Figure 2. Registration status of medical care institution in National Health Insurance in 2011. Reprinted from Health Insurance Review and Assessment Service. Guidelines for review and assessment for healthcare services of 2011. Seoul: Health Insurance Review and Assessment Service; 2011 [4].

1) 2011년 1년간 청구된 환자 중에 단 1건이라도 입원 진료 경험이 있는 환자를 입원 환자로 분류함.

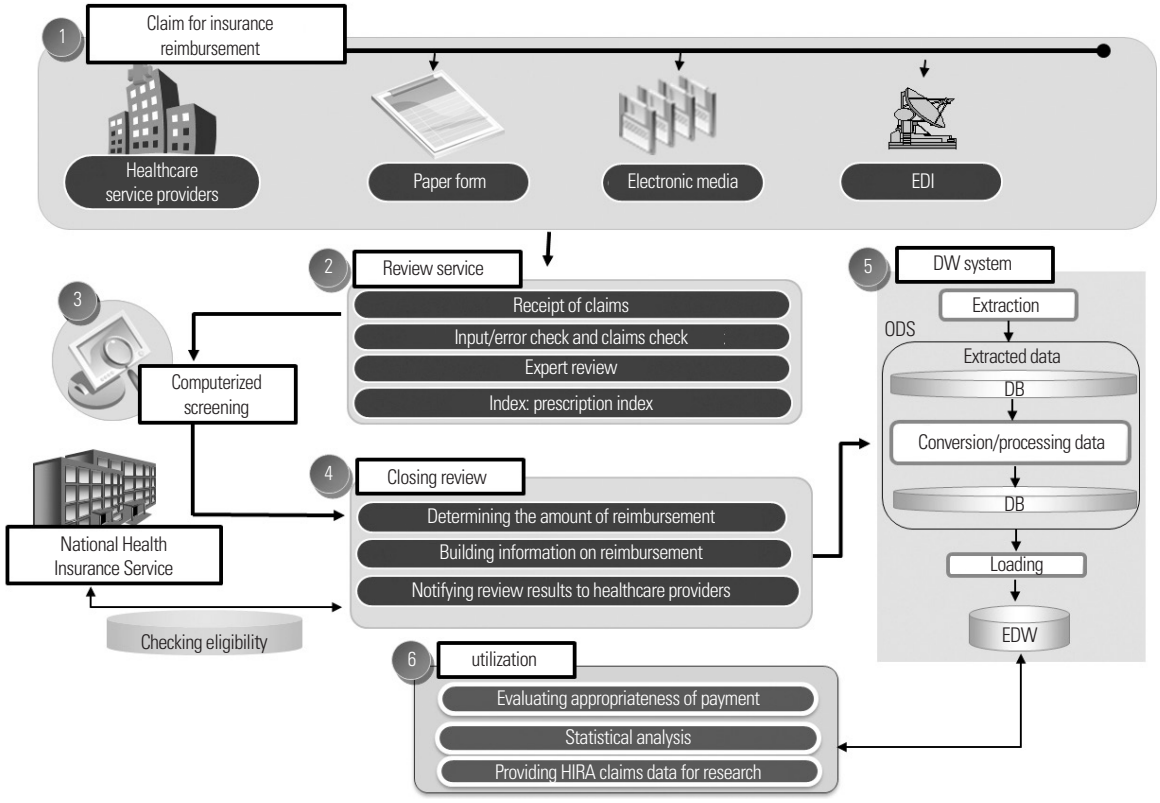

Figure 3. Flow of claim data processing. EDI, electronic data interchange; DW, data warehouse; EDW, enterprise data warehouse; HIRA, Health Insurance Review and Assessment Service. Reprinted from Health Insurance Review and Assessment Service. Guidelines for review and assessment for healthcare services of 2011. Seoul: Health Insurance Review and Assessment Service; 2011 [4].

Table 3. Major variables of claim data

Categories of bariables		Variable
General items	Identification codes	Patient' name, unidentifiable patient' ID, insurance no., business no., types of healthcare providers, etc.
	Other variables	Diagnoses, surgical operation, medical department, start date of care, length of stay at a hospital, no. of outpatient visits, no. of prescriptions, days covered by prescription, first visit, no. of re-visits, disposition of the patient at discharge, total amount to be reimbursed, patient's out-of-pocket cost, insurer's payment, etc.
Detailed information on care		Items (procedures, medical products including drugs, ingredients and materials for treatment, etc.)
		Prescription related information (information on drugs prescribed such as dosage, strength, days of supplies, etc.)

Reprinted form Health Insurance Review and Assessment Service. Manual: management of healthcare providers. Seoul: Health Insurance Review and Assessment Service; 2011 [5].

Table 4. Major variables of medical care institution data

Categories of bariables	Variable
General information	Founder of the personal information, address, type of hospital, current management information and treatment parts
Hospital beds information	Hospital room, special-purpose room (intensive care unit, operating room, emergency room, hemodialysis room, isolation ward, etc. 12 classification), day care room, etc.
Medical personnel information	Doctor, oriental doctor, chemist, nurse, clinical pathologist, shadow gazer, nutritionist, etc.
Medical equipment information	Computed tomography, magnetic resonance imaging, positron emission tomography, etc. (radiation diagnosis and radiation treatment, medical check-up, physical therapy, surgery and treatment, oriental related equipment etc.

Reprinted form Health Insurance Review and Assessment Service. Manual: management of healthcare providers. Seoul: Health Insurance Review and Assessment Service; 2011 [5].

가하는데 필요한 기초자료로 활용하기 위하여 요양기관으로부터 법정서식인 '요양기관현황통보서' 및 '요양기관변경사항통보서'를 통해 최소 월 1회 전산으로 제출받는다. 제출받은 요양기관 현황자료에는 요양기관의 일반 현황, 병상 현황, 의료인력 현황, 의료장비 현황이 포함되어 있다. 요양기관 현황자료는 청구명세서자료와 함께 심평원 데이터베이스에 저장 관리된다(Table 4).

진료에피소드 개념 및 청구자료의 특성

1. 진료에피소드의 개념과 필요성

질병에피소드는 환자의 상병이 발병하고 완치되기까지를 하나의 기간으로 묶는 단위이다. 그러나 건강보험으로 청구되는 자료로

는 질병의 발병시기와 완치시기를 정확히 알아내는 것이 거의 불가능하다. 청구자료를 활용하여 특정 상병으로 처음 외래진료를 한 시기와 입원기간, 퇴원일, 최종 외래진료일은 추적이 가능하여 이러한 에피소드 구분방식을 진료에피소드(episode of care)라고 한다. 진료에피소드는 환자가 특정 상병으로 병원을 처음 방문하고 그 이후 마지막 외래진료시점까지의 기간을 한 단위로 묶은 것이다. 일반적으로 특정 질환으로 인하여 입원을 한 후에 퇴원하더라도 바로 해당 질환으로 인한 의료서비스가 종료되는 것이 아니라 외래진료까지 추적 관찰할 필요가 있다. 이 단계에서 동일한 에피소드로 간주할 일정 기간 즉, 무 진료기간이 적용된다. 동일 질병군으로 며칠 이내에 또 다시 의료이용이 있었을 때에 이를 동일한 에피소드로 간주하는 기준이 무 진료기간이다. 만성질환과 같이 한번 이환되면 사실상 완치가 불가능한 질환들의 무 진료기간은 무한대라고 할 수 있는 반면 짧은 이환기간과 잦은 재발을 특징으로 하는 감기와 같은 질환에서는 무 진료기간을 어느 수준으로 정하느냐에 따라 발병건수가 달라질 수 있다. 현재 청구방식에서는 요양기관 종별로 청구하는 진료기간이 달라 하나의 에피소드에 대해 여러 청구건이 발생할 수 있다. 또한 전원을 한 경우도 하나의 에피소드에 대해 청구자료가 분리될 수 있다. 따라서 청구자료를 표본 추출할 경우 진료에피소드의 개념이 필요한 이유는 청구자료의 특성상 동일인의 동일 상병에 대하여 여러 건의 청구자료가 발생할 수 있기 때문에, 이러한 경우 여러 건의 청구자료를 하나의 상병 단위로 묶어주기 위하여 에피소드의 개념이 필요할 것이다[6].

진료에피소드의 무 진료기간은 연구자마다, 상병 특성마다 각기 다르기 때문에 범용의 표본자료에서는 무 진료기간에 대한 어떠한 기준도 적용시킬 수 없다. 따라서 모든 에피소드의 기준을 만족시키기 위한 방법이 환자 단위 표본 추출이라 할 수 있다. 환자 단위로 추출하면 이를 다시 에피소드 묶음[2]으로 나눠 준다. 예를 들어 한 환자가 1년 동안 세 가지 상병으로 병원을 방문했을 경우 세 개의 에피소드 묶음이 생기며, 각각의 에피소드 묶음은 상병의 성격과 연구자의 판단에 따라 무 진료기간을 조정하여 에피소드의 기준을 정의할 수 있다(Table 5).

2. 청구자료의 특성

건강보험자료를 활용한 표본 추출방법의 이해를 돕기 위해 다음의 몇 가지 그림을 소개하기로 한다. 앞서 기술했던 바와 같이 의료기관에서는 월별 혹은 일자별로 심평원에 환자당 한 건의 건강보험 요양급여를 청구한다(Figure 4).

명세서의 진료기간은 환자의 입원기간에 상관없이 1개월을 기준으로 청구하여 장기간 입원 환자는 여러 건의 청구명세서가 발생할 수 있다. 외래의 경우 월간 진료내역을 통합 청구하고(단, 의원급은 방문일자별로 구분 청구), 약국의 경우 일자별로 구분(2005년 이후) 청구한다.

Figure 5의 한 단위(원)를 청구 건 한 건으로 가정할 때 원의 크기는 중증도, 요양일수, 비용의 크기로 정의할 수 있으며, 작은 원은 외래진료 건이고 큰 원은 입원진료 건으로 볼 수 있다. 큰 원 안에 작

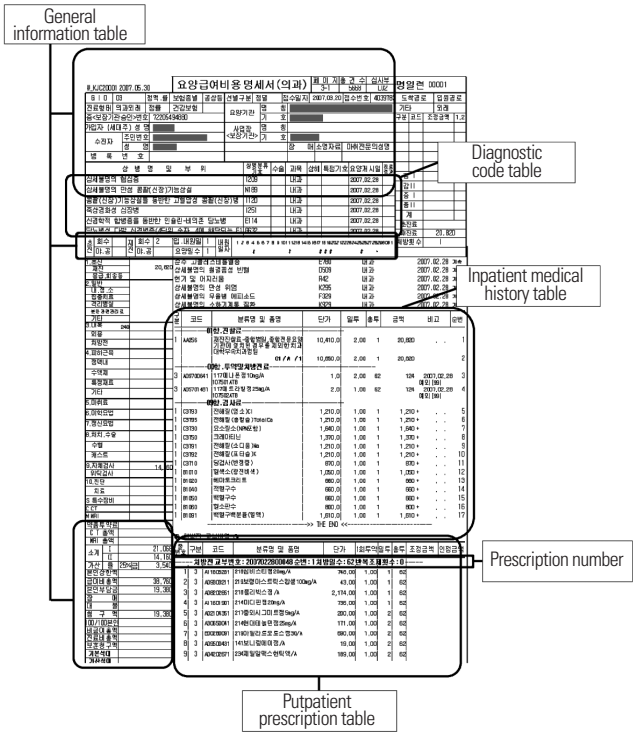

Figure 4. Paper form of medical claim data.

Table 5. Episode types

Episode type	Definition	Window period
Episode of healthcare providers	Collection of inpatient care that a patient used during the stay at a same healthcare institution	Collection of re-admission within one day
Episode of patient	Collection of care including inpatient services that a patient used to treat a particular condition with all different healthcare providers	Collection of re-admission within two days
Episode of care	Collection of inpatient and outpatient care that a patient uses to treat a particular condition within a specified window period	Depending on characteristics of health condition (disease)

Reprinted from Kim JY et al. Development of risk adjustment and prediction methods for care episodes using National Health Insurance database. Seoul: Health Insurance Review and Assessment Service; 2007 [6].

2) 에피소드 묶음이란 표본 추출의 해당 기간 내에 동일인의 동일 상병에 대해 발생하는 입원, 외래에 대한 모든 청구 건의 묶음을 말함.

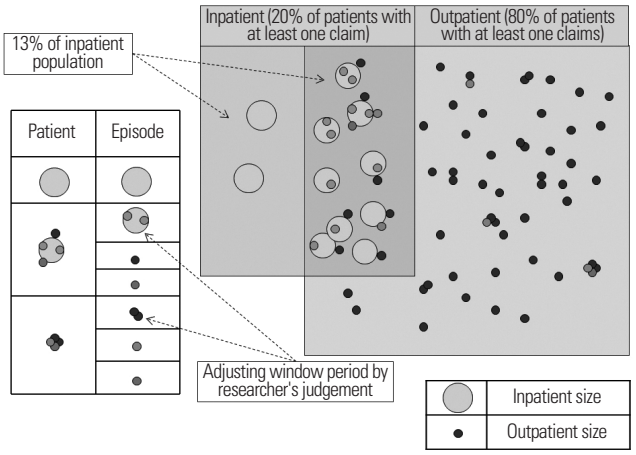

Figure 5. Distribution of patient types in claims data: inpatient and Outpatient.

은 원은 한 명의 환자에 대한 입원진료와 동일한 상병의 외래진료이며, 큰 원 밖에 작은 원은 입원진료와 다른 상병의 외래진료이다.

질병의 기간적인 측면까지 고려하면, 한 환자가 종합병원에 3개월 동안 입원하는 경우 병원급 이상은 월별로 분리 청구하므로 총 세 장의 청구명세서가 발생하게 된다. 또한 환자가 다른 병원으로 이환했을 경우에도 새로운 청구명세서가 발생하게 된다. 예를 들어 어떤 환자가 만성질환으로 계속해서 병원을 이동할 경우 그 환자는 만성질환 유병 환자로 간주하여야 한다. 또 감기와 같은 경증질환자가 수개월 간격으로 다시 감기로 병원에 내원한다 할지라도 그 환자는 동일 감기의 유병 환자로 보기 어렵다. 이와 같이 질병의 성격에 따라 그 시작과 끝을 정의하는 에피소드의 기준이 모두 다르다. 심평원 입원 환자 표본자료는 모든 에피소드 및 질병정의에 대한 기준을 만족시키기 위해서 1년간 건강보험으로 청구된 입원 환자를 기준으로 하여 표본 추출되었다. 따라서 환자를 한 단위[3]로 하여 추출하게 되면 질병 특성 및 조작적 정의에 따라 질병의 에피소드를 구축하여 연구에 활용할 수 있게 될 것이다.

표본 설계

1. 표본 추출방법

표본 추출방법은 층을 나눌 수 있는 충분한 정보가 존재한다면 단순무작위방식보다는 층화추출방식을 우선 고려하여야 한다. 병원 청구자료의 경우 환자의 연령 및 성별 구분은 층을 나눌 수 있는 충분한 정보가 되므로 층화추출방식을 사용하였다. 층화표본 추출은 단순무작위 표본 추출에 비해 적은 표본 수로도 전체 모집단의 특성을 잘 대표할 수 있는 표본 추출방법이므로 표본자료의 효율성을 높이기 위해 이를 사용하였다. 통계분석을 위해 SAS Enterprise Guide ver. 4.3 (SAS Institute Inc., Cary, NC, USA)과 SAS ver. 9.1 (SAS Institute Inc.)을 사용하였다.

2. 표본 층화 설정

층화변수는 인구학적 구조인 성별, 연령(16개 구간)변수로 32개 구간을 설정하였다. 건강보험 청구자료에서 신생아는 주민등록번호의 부재 등 부정확성의 이유로 0세 구간은 제외하였다. 층화변수로 종별, 상병별, 병원지역 등의 변수를 사용할 경우 층 내에는 동질이고 층간에는 이질적인 층화변수의 특징을 반영할 수 없다. 자료의 특성상 시차를 두고 동일 환자가 다른 상병에 걸리거나 다른 지역의 병원으로 이동하는 경우 여러 번의 추출 확률을 갖게 되며 층화변수의 기본 가정인 층간 이질을 만족할 수 없다. 이러한 층화방식은 네트워크 추출에서 사용되는 방식으로 네트워크방식은 동일하지 않은 추출 확률 때문에 표본 평균이 모평균의 불편추정량이 될 수 없다. 따라서 요양기관의 종별 구분에 따라 일자별 혹은 월별 분리 청구되는 병원 청구자료의 특성과 질병마다 다른 에피소드 기준을 갖는 의학자료의 시계열 측면의 특성을 감안하여 인구학적 층화 후 환자 단위 표본 추출하는 방식이 가장 바람직하다고 판단하여 시행하였다.

3. 필요 표본 수 산출

환자 표본자료의 개발목적 중 고려해야 할 부분은 범용성과 대표성이다. 따라서 보다 많은 연구에 활용될 수 있는 표본자료의 생산을 위해서는 필요 표본 수 산출에서 다양한 부분이 고려되어야 한다. 환자 표본자료에서 범용성과 대표성 확보를 위한 필요 표본 수는 두 가지 목표를 만족시켜야 한다. 첫 번째로 범용성 확보를 위해서는 목표 인구집단의 발생 확률을 만족시키는 표본 수여야 하며, 두 번째로 모집단의 최대 분산을 갖는 연속변수[4]에 대하여 대표성을 갖는 표본 수여야 한다.

그 결과 범용성 확보를 위한 필요 표본 수는 729,904명으로 산출되었으며, 연속변수의 대표성 확보를 위한 필요 표본 수는 488,861명으로 산출되었다. 최종 필요 표본 수는 범용성과 대표성을 충족시키는 수준인 729,904명 이상이 되어야 하며, 실제 표본 환자 수는 계산상의 편의를 위해 모집단의 13%(표본 가중치 7.692)인 765,603명으로 결정하였다.

1) 범용성 확보를 위한 필요 표본 수

n의 요구된 표본 크기는 다음과 같다.

$$n = \frac{[1.96^2(r)(1-r)]}{(0.1r)^2h} \quad (1)$$

3) 환자 구분 코드를 기준으로 한 환자의 모든 청구 건 묶음.
4) 표본자료를 구성하는 모든 연속변수 중에 가장 큰 분산을 가진 변수(1년간 환자당 진료비 총액).

식(1)에서는 95% 신뢰도를 획득하기 위한 요인이며 r은 예측 또는 예상된 값으로 비율의 형태로 표현하였다. 따라서 0.1r은 95% 신뢰도에서 허용되는 오차 한계로, r의 10%로 정의된다(r의 상대적인 오차 한계). 는 의 예측 값에 대한 전체 인구에서 대상 집단이 차지하는 비율이다. 입원 환자 표본자료에서 r의 관측확률은 5%이며, h는 1%이다. 이 결과로 계산된 필요 표본 환자 수는 729,904명이었으며, 이는 전체 입원 환자를 대표하기 위해 필요한 최소 표본크기이다. 여기에서 관측확률 r과 인구집단(연령, 성별 계층 등)의 상대비율 h는 특정하지 않으며 상호 교호작용한다. 따라서 전체 인구집단을 대상으로 할 경우 r의 관측확률은 약 0.0524% 이상이 되고, 이는 250개 다빈도 질환 중에서 최저 관측 비율인 0.0591%보다 작은 수치이다.

2) 연속변수의 대표성 확보를 위한 필요 표본 수

모집단에서 최대분산을 가지는 연속변수는 환자당 입원청구총액이며 필요 표본 수를 구하기 위한 추정오차의 한계는 모집단의 환자당 입원 진료 평균액의 0.5% (19,956원)로 하였다. 따라서 95% 신뢰구간에서 평균에 대한 오차범위는 3,991,289원 ± 0.25% (9,978원)이 된다. 크기가 인 추정오차의 한계를 갖는 필요 표본 수를 계산하기 위해서 다음의 식을 사용한다. 추정오차 한계는 $B = 1.96\sqrt{V(\bar{y})} = 1.96\sqrt{\frac{\sigma^2}{n}}$ 이 되며, 추정오차 한계 식을 다시 n에 대하여 정리하고, 필요 표본 수를 계산하는 식은 다음과 같다.

$$n = \frac{(1.96)^2\sigma^2}{(B)^2} \quad (2)$$

최대분산을 갖는 연속변수를 사용하여 표본집단과 모집단의 상대 효율을 검정할 수 있으며, 식(2)에서 평균의 0.5% (19,956원)를 오차범위로 하는 최소 필요 표본 수는 488,861명 이상이다.

4. 모비율 추정 및 추정오차의 한계

전체 집단에서 연구에 목표로 하는 특정 질환의 환자가 있을 경우 모비율 추정을 통해 환자 수의 추정오차의 한계를 구할 수 있다. 모비율 추정량은 다음과 같이 계산한다.

$$\hat{p} = \sum_{i=1}^{n} \frac{y_i}{n} \quad (3)$$

식(3)에서 $\hat{p}$은 표본비율이고 $\sum_{i=1}^{n} y_i$ 는 크기 n인 표본에서 구하고자 하는 비율을 결정하는 특정한 속성을 갖는 원소들의 총합이다. 만약 y_i 이면 i번째 환자가 특정 속성을 갖고 있지 않는 경우이고, $y_i = 1$이면 번째 환자가 특정 속성을 갖고 있는 경우이다. 이 경우에 $E(\hat{p}) = p$이므로 표본비율은 모비율에 대한 불편추정량이다.

여기서 $\hat{q} = 1-\hat{p}$이고 비율 $\hat{p}$의 분산추정량은 $V(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}$ 이며, 95% 신뢰구간에서 추정오차의 한계는 $1.96\sqrt{\hat{V}(\hat{p})}$ 이다. 표본비율 $\hat{p}$은 모비율 p에 대한 불편 추정량이므로 해당 질환의 추정모비율 $(\hat{p}) \approx \frac{\text{표본에서 해당 질환의 대상이 되는 환자 수}}{\text{표본전체환자 수}}$ 를 이용하여 식(5)와 같이 $\hat{p}$의 95% 신뢰구간을 계산할 수 있다[8].

$$\hat{p} \pm 1.96\sqrt{\hat{V}(\hat{p})} \quad (5)$$

표본 추출결과 및 타당도 검증

1. 표본 추출결과

2011년 1년간 심평원으로 요양급여가 청구된 전체 환자 수는 약 4,600만 명으로 이 중 입원경험이 있는 환자(전체 입원 환자)는 약 589만 명이며, 전체 입원 환자에서 13%를 표본 추출한 결과 입원

Table 7. The number of inpatients by age group on sample

Age groups	Gender		Sum
	Man	Woman	
1-4	21,225	17,745	38,970
5-9	11,777	9,456	21,233
10-14	10,391	6,570	16,961
15-19	14,495	9,472	23,967
20-24	12,074	13,784	25,858
25-29	15,829	32,736	48,565
30-34	18,754	45,727	64,481
35-39	21,143	29,136	50,279
40-44	24,600	26,714	51,314
45-49	27,784	30,877	58,661
50-54	32,958	36,783	69,741
55-59	28,093	29,027	57,120
60-64	25,162	25,303	50,465
65-69	23,651	26,767	50,418
70-74	22,216	29,748	51,964
≥ 75	29,372	56,234	85,606
Total	339,524	426,079	765,603

Table 6. Overview of inpatient sample

	Description
Population	Patients who used inpatient services for 2011 (approximately 5.89 million)
Sample	Stratified proportional sampling approximately 0.77 million (13% of population)
Variables used for stratification	Gender, age group (the sample containing all medical and prescription claims for one year)

Table 8. Descriptive statistics of population and sample

	Unit of inpatient	
	Population	Sample
Number	58,889,784	765,603
Variance	4.97296E+13	5.06803E+13
Standard deviation	7,051,922	7,119,009
Mean (won)	3,991,289	3,984,645

Table 9. Hypothesis test

99% Confidence interval (the null hypothesis was accepted)		
Population mean test	H0=no difference between means	Pr> =0.9993

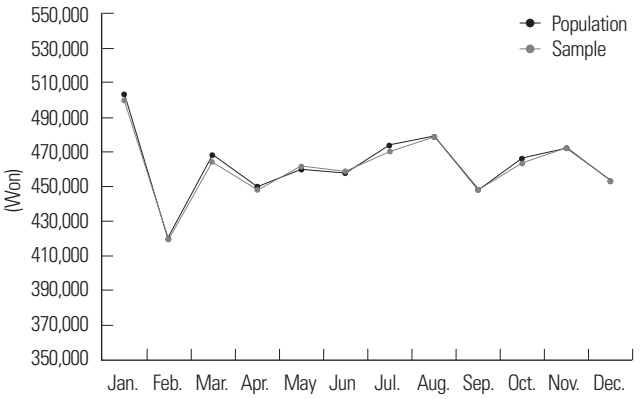

Figure 6. Monthly average expenditure of a patient: population vs. sample.

환자 표본자료에 포함된 표본 환자 수는 약 77만 명이다. 심평원에서 개발한 입원 환자 표본자료의 개요는 다음의 Table 6과 같다. 2011년 표본 층별 입원 환자 현황은 다음의 Table 7과 같다. Table 8의 표본과 모집단의 기초통계량에서 표본의 모집단에 대한 상대분산의 비율은 98.12%이다.

2. 타당도 검증

환자당 진료비 총액[5]을 가설검정변수로 채택하여 모집단과 표본자료의 분산과 평균에 대하여 가설검정을 하였다. 가설검정방법은 모집단 표준편차를 알고 있으므로 Z-검정을 사용하였으며 99% 신뢰구간에서 가설을 채택하여 모집단과 표본집단이 차이가 없는 것으로 나타났다(Table 9).

통계학에서는 추출된 표본의 효율성을 평가하기 위해서 모집단과 표본집단의 분산의 상대비율을 이용한다. 건강보험 청구자료를 환자 단위로 하였을 경우 최대분산을 가지는 연속변수는 1년간의 환자 1인당 진료비이다. 모집단과 환자 표본자료의 환자당 진료비를 월별로 비교하였고 그 결과 Figure 6의 표본자료와 모집단의 월별 환자당 평균 진료비 추이가 거의 일치하였으며 상대분산 0.98의 일치율을 보였다.

다음 Table 10의 입원 환자 상위 30개 다빈도 질환에서 주상병(ICD-10: International Classification of Diseases-10) 기준으로 모집단 대비 표본자료의 발병건수를 비교하였고 상대비율은 입원 환자 추출비율인 13%에 근사한 값을 가지는 것으로 나타났다. Table 11에서 95% 신뢰수준 모집단 추정오차구간 산출결과 뇌졸중 전체 환자는 195,905±2,414의 추정 값을 가지며 실제 모집단의 뇌졸중 환자는 196,677명으로 전체 환자와 각 연령구간 모두 신뢰구간 내에 참값이 존재함을 확인할 수 있었다.

Table 10. Top 30 frequent diseases of population and sample

Frequency ranking	diagnostic code	Population frequency	Sample frequency	Relative proportions
1	J189	211,942	27,630	13.04
2	O800	195,192	25,364	12.99
3	M511	129,333	16,616	12.85
4	A099	86,299	11,224	13.01
5	S3350	86,069	11,160	12.97
6	A090	82,669	10,721	12.97
7	O820	73,260	9,386	12.81
8	I639	69,109	9,021	13.05
9	M4806	65,510	8,485	12.95
10	K358	65,277	8,347	12.79
11	I8418	64,544	8,406	13.02
12	J209	56,027	7,258	12.95
13	M170	54,781	7,239	13.21
14	I109	52,309	6,849	13.09
15	J180	51,685	6,677	12.92
16	C73	49,045	6,365	12.98
17	I8411	47,905	6,266	13.08
18	M512	47,755	6,104	12.78
19	O821	47,576	6,231	13.10
20	F102	47,446	6,063	12.78
21	N10	47,193	6,147	13.03
22	J157	46,202	5,935	12.85
23	F009	41,176	5,358	13.01
24	H2590	40,775	5,282	12.95
25	H2591	40,457	5,247	12.97
26	E119	39,445	5,099	12.93
27	J0390	37,576	4,877	12.98
28	I209	36,907	4,820	13.06
29	I839	35,637	4,572	12.83
30	D259	35,148	4,571	13.01

심평원은 환자 표본자료의 타당도 평가를 위해 보건의료 관련 5개 학회[6]와 memorandom of understanding (MOU)를 맺었으며 MOU 체결 학회 회원을 대상으로 표본자료 활용연구들을 수행한 바 있다. 수행된 연구들 중에서 주요 연구결과를 보면 '우리나라 당뇨병 유병률 추정 및 dipeptidyl-peptidase 4 억제제 사용 양상 평가' 연구에서 당뇨병 유병률 및 혈당강하제의 각 약효군별 처방률이 모집단과 표본자료가 일치하는 결과를 보였으며, '시력손실과 실명으로 인한 사회적 질병 부담비용 추계'에서도 백내장, 녹내장, 황반변성의 발생빈도 및 연령별 추이가 표본자료와 모집단이 일치하는 결과를 보였다.

5) 최대분산을 가지는 연속변수.
6) MOU 체결 학회: 대한예방의학회, 한국보건경제 · 정책학회, 한국보건정보통계학회, 한국보건행정학회, 한국역학회.

Table 11. Population estimates of patients with stroke (diagnostic code in = I60, I61, I62, I63, I64, I67, I68, I69) using the sample data

Gender	Age groups	2011			
		Population estimates (sample × weight)	Lower 95% confidence limit	Upper 95% confidence limit	Population
Man	1-9	169	97	241	168
	10-19	308	210	405	333
	20-29	631	491	770	657
	30-39	2,892	2,594	3,191	2,678
	40-49	10,469	9,902	11,036	9,842
	50-59	20,184	19,398	20,971	20,312
	60-69	23,461	22,613	24,309	23,688
	≥ 70	36,330	35,276	37,384	36,885
Woman	1-9	231	147	315	167
	10-19	215	134	297	311
	20-29	669	526	813	574
	30-39	1,631	1,407	1,855	1,633
	40-49	6,200	5,763	6,636	6,187
	50-59	13,315	12,676	13,955	13,540
	60-69	18,677	17,920	19,434	18,923
	≥ 70	60,522	59,165	61,880	60,779
Total		195,905	193,491	198,319	196,677

결론 및 제언

본 논문은 모집단인 2011년 건강보험 청구자료에 대한 대표성과 범용성 부분에서 일정 수준을 충족하는 표본자료를 추출하는 방법과 과정에 대해 분석, 설명하고 있다. 향후 표본자료는 지속적인 개발 및 보완을 통해 자료의 제공 영역을 확대해 나갈 계획으로 있다.

입원 환자 표본자료 활용 시 주의사항으로 급여가 인정된 의료이용내역만 포함되어 있기 때문에 비급여내역 또는 처방전 없이 구입할 수 있는 아스피린 등의 일반의약품에 대한 정보는 표본자료에서 확인할 수 없다. 또한 진단명의 정확성에 대한 연구자의 고려가 필요하다. 진단명의 정확성은 외래보다는 입원 환자, 다빈도 경증질 환자보다는 위중한 환자에서 높게 나타나며 의원급보다는 종합병원급 요양기관에서 더 높은 경향이 있다[7]. 진단명 및 시술에서 의사의 개인차, 관습적 요인을 완전히 배제하기 어렵기 때문에 자료의 특성, 환자의 의료이용행태와 질병의 고유특성, 의사의 진료과정과 임상환경, 병원의 전산망과 청구과정, 건강보험급여제도 등을 충분히 파악해야 올바른 해석이 가능하다. 이와 더불어 진단명의 타당도 등을 주기적으로 평가하여 청구자료의 지속적인 품질관리를 통해 신뢰성을 높이는 것도 중요한 과제라 하겠다.

환자 표본자료의 제한점은 모든 표본자료 공통의 한계점으로서 표본자료 내의 관측치는 확률에 의해 추출되는 자료이기 때문에 적정 수준 이상의 표본 수를 확보해야 대표성, 유의성을 보장받을 수 있다는 것이다. 예를 들어 본 환자 표본자료에서 특정 연령대의 희귀질환 발생빈도의 경우 표본 추출 빈도가 너무 적어 대표성과 설명력이 떨어질 수 있다. 따라서 표본자료의 설명력은 다빈도 상병일수록 커지며, 상병의 발생빈도가 떨어지면 감소하게 된다.

이를 보완하는 방안으로 희귀 상병은 전수 제공하는 방안이 있으며 자료의 제공 영역을 1년 단위가 아닌 일정기간 동안의 패널자료를 구축해 표본 추출 빈도를 높이는 방안도 고려해 볼 수 있겠다.

표본자료의 다양화와 확대방안으로 우선 표본자료의 제공변수 영역을 확대하고, 전체 환자, 65세 이상 고령 환자, 영유아 환자, 진료과 혹은 질환별 코호트 표본 등으로 표본자료를 세분화시켜 개발하는 방안도 고려해 볼 수 있다. 건강보험 입원 환자 표본자료의 지속적인 환류과정을 통해 표본자료의 영역이 확대되면, 다양한 분야에서 표본자료의 활용을 통해 국가 · 사회적 편익을 제고할 수 있을 것으로 기대된다.

REFERENCES

1. National Health Research Institutes. Researches of National Health Insurance claim database [Internet]. Miaoli County: National Health Research Institutes [cited 2013 May 20]. Available from: http://www.nhri.org.tw.
2. Agency for Healthcare Research and Quality. Design of the nationwide inpatient sample (NIS). Rockville (MD): Agency for Healthcare Research and Quality; 2005.
3. Agency for Healthcare Research and Quality. Introduction to the HCUP nationwide inpatient sample (NIS). Rockville (MD): Agency for Healthcare Research and Quality; 2007.
4. Health Insurance Review and Assessment Service. Guidelines for review and assessment for healthcare services of 2011. Seoul: Health Insurance Review and Assessment Service; 2011.
5. Health Insurance Review and Assessment Service. Manual: management of healthcare providers. Seoul: Health Insurance Review and Assessment

Service; 2011.
6. Kim JY, Im JH, Kim HY. Development of risk adjustment and prediction methods for care episodes using National Health Insurance database. Seoul: Health Insurance Review and Assessment Service; 2007
7. Park BJ, Seong JH, Park GD Seo SW Kim SH, Studying on improving diagnosis codes in National Health Insurance claims data. Seoul: Health Insurance Review and Assessment Service; 2003.
8. Namkung P. Design and analysis of sample survey. 2nd ed. Seoul: Tamjin; 2007.