건강보험심사평가원 환자표본자료(HIRA-NPS)의 소개

김록영 주임연구원
건강보험심사평가원 통계관리부

1. 서론

우리나라는 1960년대와 1970년대에는 표본자료 설계에 대한 구체적인 체계가 잡히지 않은 상태였고 1990년대에 통계청에서 보관하고 있는 조사구 자료가 일부 제공되면서 비교적 과학적인 표본설계를 하게 되었다. 2000년대에 비로소 보건분야에서 대규모 표본조사를 시작하여, 2000년도 초 불과 10개 정도의 보건조사 자료가 존재 했던 것에 비해 2010년을 기준으로 40개가 넘는 보건조사 자료와 패널조사 자료를 보유하게 되었다. 이러한 보건분야 자료의 증가는 현대 보건복지 분야에서 기초자료의 중요성이 부각되고 있음을 보여주는 예라고 할 수 있다. 우리나라는 현재 전 국민의 98%가 국민건강보험에 가입하고 있어 건강보험청구자료는 국가 보건의료를 대표하는 자료라고 할 수 있다. 그러나 지금까지 우리나라의 보건 분야의 자료는 주로 실사를 바탕으로 한 조사가 주를 이루고 있었으며 건강보험청구자료에 대한 표본자료의 제공 방안은 마련되지 않고 있다.

우리나라와 유사한 건강보험체계를 가지고 있는 대만은 1995년 3월 1일 단일 국가의료보험을 시작하였으며 2007년 기준으로 대만의 전체 인구 중 98.4%가 등록되어 있다. 대만 국민의료보험 데이터베이스(National Health Insurance Research Database)에는 의료상환을 위한 보험자 등록 자료와 건강보험청구자료' 포함되어 있다. 대만의 국민의료보험 데이터베이스는 국립건강보험국(Bureau of National Health Insurance)산하 국립보건연구소

(National Health Research Institutes)의 관리 하에 구축되어 연구용도로 제공된다. 대만의 국립보건연구소에서는 청구건 기준으로 실시되는 월단위 표본자료와 함께 환자기준으로 실시되는 1년 단위 표본자료부터, 5년 단위 패널자료 까지 다양한 방식으로 건강보험청구자료에 대한 표본자료를 제공하고 있다. 월단위 표본자료는 주로 계절에 민감하거나 유행성 질병에 시의 적절하게 대응하기 위한 자료로 활용되고 있으며 1년 단위 혹은 5년 단위의 패널 자료는 주로 연구용으로 활용되고 있다.

미국의 Agency for Healthcare Research and Quality(AHRQ)는 연방정부에 속한 연구기관으로 보건 분야에 관련된 연구를 지원한다. AHRQ는 37개 주의 정부 및 지역사회, 보건의료 산업체들로부터의 데이터를 수집하여 의료데이터 베이스를 구축하고 있다. AHRQ에서 관리하는 조사 프로그램 중 하나인 Healthcare Cost and Utilization Project(HCUP)는 미국에서 가장 큰 데이터베이스를 구축하고 있는데 HCUP의 제공 자료 중 가장 포괄적인 전국 입원환자 샘플 데이터(National Inpatient Sample)는 재활의료기관을 제외한 미국 병원협회(American Hospital Association)의 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 표본추출하여 추출된 의료기관의 전체 입원 자료(약 5백만~8백만 입원 건)를 포함하고 있다.

이에 반해 우리나라의 경우 건강보험심사평가원과 국민건강보험공단에서 자료처리실을 운영하고 있어 건강보험청구자료에 대하여 외부 연구자가 직접 자료를 가공하여 결과를 산출하도록 하고 있으나 접근성과 편의성 측면에서 한계가 존재한다. 또한 진료정보 자료는 연간 약 10억 건 이상으로 그 용량이 방대하여 사용자의 저장용량, 처리속도 등 수용능력의 한계로 인하여 시의 적절한 자료 확보를 불가능하게 한다. 따라서 다양한 수요층에 대한 접근성과 편의성, 즉시성의 확보를 위한 대안의 하나로 우리나라의 건강보험 청구자료에 대한 표본자료를 개발하였다.

국가별 비교자료에서 미국은 퇴원데이터의 성격상 비급여가 포함되어 있으며 환자가 특정의료기관에 입원하고 퇴원까지를 한단위로 하는 자료이므로 환자가 재입원하거나 다른 의료기관으로 이환 된 경우에는 환자구분이 되지 않는다.

이에 비해 우리나라와 대만의 청구자료는 비 급여는 포함되지 않으며 환자단위의 자료이므로 재입원하거나 이환하더라도 환자구분이 가능하다.

표 1. 국가별 표본자료 비교

국가별비교	한국(HNPS)	미국(NIS)	대만(NHIRD)
추출 단위	환자	병원	환자
제공 단위	환자단위	기관별 퇴원 환자 단위 (퇴원자료)	환자단위
층화 변수	성, 연령구간	병원 특징 지리학적 위치	단순무작위추출
제공 대상	모든 연구자	모든 연구자	국가 연구기관 및 연구자 (일반인은 학습용 자료 이용)
표본규모	입원 환자 13% (약 70만 명) 외래 환자 1%(약 40만 명)	약 700만 건 정도의 기관 단위(입원자료)	건강보험 등록자의 100만 명

2. 환자표본자료 개발 배경 및 목적

건강보험 청구자료는 제한적 실험환경이 아닌 실제 보건의료 환경을 반영하는 데이터 이다. 따라서 비교적 일반화가 용이하며 이미 구축된 자료를 활용함으로써 연구에 들이는 시간과 비용 등을 단축시킬 수 있다. 세계적으로 드물게 우리나라는 전 국민을 대표할 수 있는 국가가 운영하는 단일 건강보험자료를 보유하고 있으며 이러한 자원은 보건의료분야의 국가정책 수립 및 국민의 건강증진에 관련된 연구에 기초자료로 활용될 수 있다. 현재 다양한 분야에서 건강보험 청구자료를 활용한 연구 수요가 급증하고 있는 추세이다. 건강보험심사평가원에서는 건강보험 청구자료에 대한 다양한 분야의 수요에 부응하기 위한 방안의 일환으로 대표성 있는 우리나라의 건강보험자료의 표본자료를 개발하게 되었다. 표본자료의 개발 및 제공의 목적은 건강보험 자료에 대한 이해 및 활용도를 높여 보건의료 연구를 활성화시키기 위함이다.

본 연구에서는 건강보험 청구자료에 대한 접근성을 확대하기 위한 방안으로 효율적인 표본설계 방법론에 대하여 연구하였다. 표본자료의 공개를 통해 국가 사회적으로 편익을 제고 시킬 수 있으며 지속적인 환류과정을 통해 다양화되고 개선된 자료가 개발 될 것으로 기대한다.

3. 표본 대상 및 추출 방법

가. 환자표본자료의 개요

건강보험심사평가원에서 개발한 환자표본자료의 개요는 다음과 같다.

○ 2009년 1년간 의료이용을 한 모든 환자 대상(약 4600만 명)으로 성별, 연령구간(5세 단위)에 따른 환자단위 층화계통 추출
- 표본추출환자 1년간의 모든 진료내역과 처방내역을 포함
- 표본 기준 변수는 1년간 환자 당 총 진료금액 (최대분산을 가지는 변수)
- 입원환자 추출 비율 13%(모집단 1년간 입원 환자 1인당 평균 진료금액 표준편차의 0.5%에 해당하는 금액을 표본의 허용 오차 범위로 정함)
- 외래환자 추출 비율 1%(입원과 동일한 오차 범위에서 필요표본 인원 계산하여 0.05%로 산출되었으나 1%로 확대 제공)

○ 해당년도 1월부터 익년 6월까지의 심결자료를 기준으로 해당년도 1년분의 진료내역을 구축

○ 용 량 : 33G(압축 후 3.44G), 40,989,560줄(건), DVD 1장

나. 모집단(건강보험청구자료)의 특징

환자표본자료의 대상 모집단이 되는 건강보험 청구자료란 의료기관에서 환자의 진료비용 중 국민건강보험이 부담하는 부분에 대해 지급의뢰를 하기위해 건강보험심사평가원에 청구하는 자료이다. 국민건강보험공단은 국민을 대상으로 국민 건강보험 재정에 대한 징수, 관리 및 지급 업무를 담당하며 건강보험심사평가원은 의료기관에서 청구하는 청구자료에 대한 진료비심사 결과를 건강보험관리공단에 통보하여 지급을 요청한다. 우리나라의 1년간 건강보험 청구 환자 수는 2009년 기준 45,969,893명으로 연앙인구 수 49,773,145명의 92.4%에 달하는 수치이다.

보험 청구는 진료행위, 약품, 치료재료의 세 가지 항목으로 구성되어 있다. 명세서는 입원의 경우 입원-퇴원 단위 또는 월 단위 분리청구(입원이 지속되는 경우 명세서가 분리되어 청구됨), 외래의 경우 월간 진료내역 통합 청구(단, 의원급은 방문일자별로 구분청구), 약국의 경우 일자별로 구분되어(2005년 이후) 청구된다.

다. 건강보험 청구자료의 표본추출 단위 설정

건강보험자료를 활용한 표본추출 방법의 이해를 돕기 위해 몇 가지 그림을 소개한다. 앞서 기술했던 바와 같이 의료기관에서는 월별 혹은 일자별로 건강보험 심사평가원에 환자 당 한건의 건강보험 요양급여를 청구 한다.

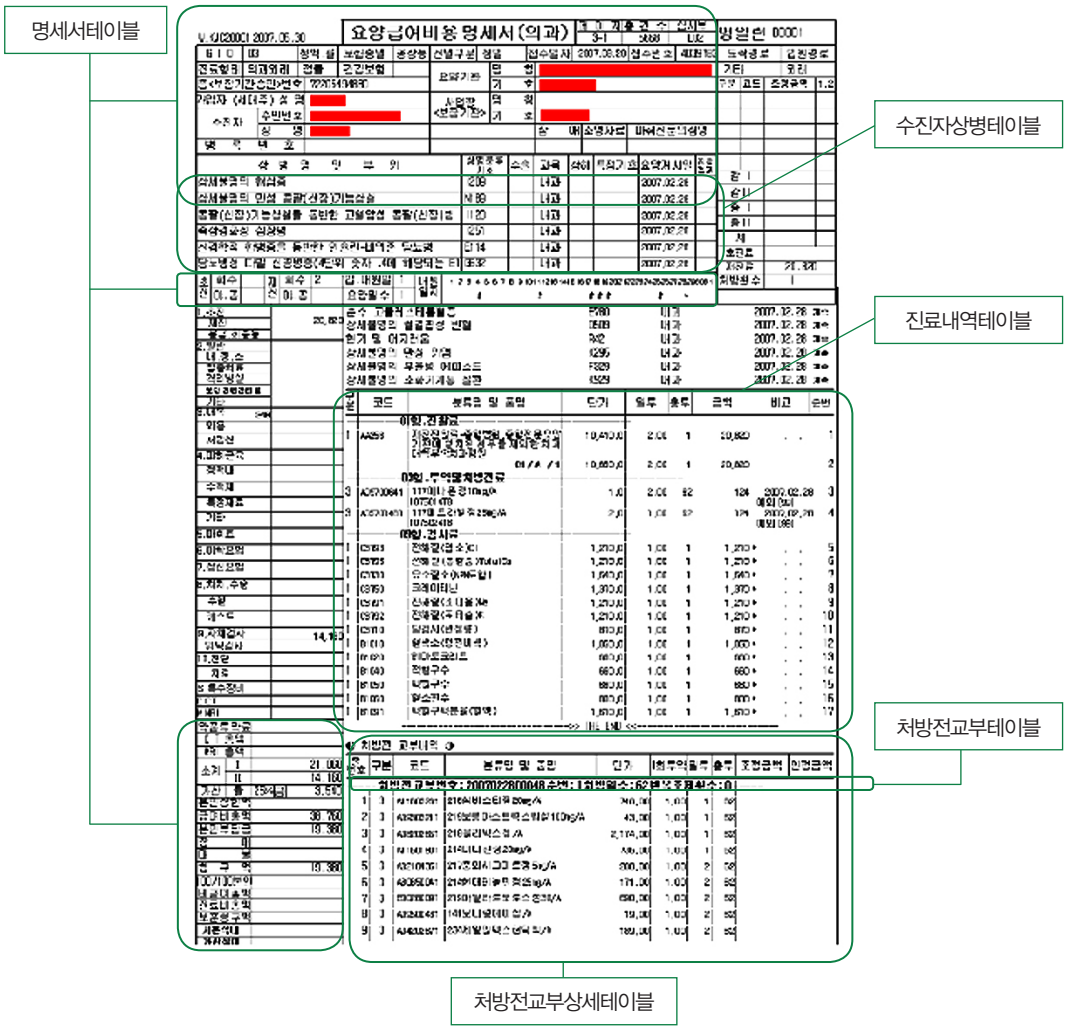

그림 1. 의과 청구명세서 형식

표본자료는 모든 에피소드 및 질병정의에 대한 기준을 만족시키기 위해서 1년간 건강보험으로 청구된 환자를 한 단위로 하여 추출 되었다. 환자를 한 단위로 하여 추출하게 되면 각각의 질병 특성 및 정의에 따라 질병의 에피소드를 구축하여 연구에 활용할 수 있다.

라. 표본 추출 방법

표본추출은 층을 나눌 수 있는 충분한 정보가 존재한다면 단순무작위 방식보다는 층화추출 방식을 우선 고려한다. 병원청구자료의 경우 환자의 연령 및 성별 구분이 층을 나눌 수 있는 충분한 정보가 되므로 층화추출방식을 사용하였다. 층화 표본추출은 단순 무작위 표본추출에 비해 적은 표본수로도 전체 모집단의 특성을 잘 대표할 수 있는 표본추출방법이므로 표본자료의 효율성을 높이기 위해 층화 추출방식을 사용하였다. 층화변수는 성별, 연령 5세 단위 16개 구간으로 총 32개의 층을 가지며 층화추출방식에서 가장 일반적으로 사용되는 층화계통비례확률 추출 방식을 사용하였다.

4. 결과 및 타당도 검증

가. 표본 추출 결과

2009년 1년간 건강보험심사평가원으로 요양급여가 청구된 전체 환자 수는 약 4,600만 명으로 입원경험이 있는 환자(전체 입원환자)는 약 500만 명, 입원경험 없이 외래로만 내원한 환자(전체 외래환자)는 약 4,000만 명으로 나타났다. 전체 입원환자의 13%, 전체 외래환자의 1%를 표본추출하여, 환자표본자료에 포함된 표본 환자 수는 약 110만 명이다.

표 2. 표본자료의 개요

전체 입원 환자	5,472,670명	표본 입원 환자	711,457명
전체 외래 환자	40,497,223명	표본 외래 환자	404,583명
모집단 환자수	45,969,893명	표본자료 환자수	1,116,040명

표 3. 표본자료의 용량

구 분	용 량
명세서일반내역	27,320,505줄(3.82G)
진료내역	313,011,694줄(21.13G)
상병내역	68,807,094줄(1.56G)

처방전상세내역	65,477,122줄(3.93G)
요양기관현황	80,418기관 (6.84MB)
전체	34Gbyte (DVD_CD 한 장)

나. 타당도 검증

1) 통계적 타당도 평가 결과

표본 검정 방법은 분산분석(일원분산분석)을 사용하였으며 95% 신뢰수준에서 두 가설 모두 채택하여 모집단과 표본간의 차이가 없는 것으로 나타났다. 추출된 표본의 효율성을 평가하기 위해서 모집단과 표본집단의 분산의 상대비율을 사용하였다. 환자단위의 모집단에 대한 분산의 상대 비율은 98.02%이다.

건강보험 청구자료를 환자단위로 하였을 경우 최대분산을 가지는 연속변수는 1년간의 환자 1인당 진료비이다. 모집단과 환자표본자료의 환자당 진료비를 월별로 비교하였고 그 결과 (그림 2)의 표본자료와 모집단의 월별 환자당 평균 진료비 추이가 거의 일치하였으며 분산의 상대 비율은 0.98로 높은 일치율을 보였다.

입원환자 다빈도 질환에서 주상병 기준으로 모집단 대비 표본자료의 발병건수를 비교하였고 상대비율은 입원환자 추출 비율인 13%에 근사한 값을 가지는 것으로 나타났다.

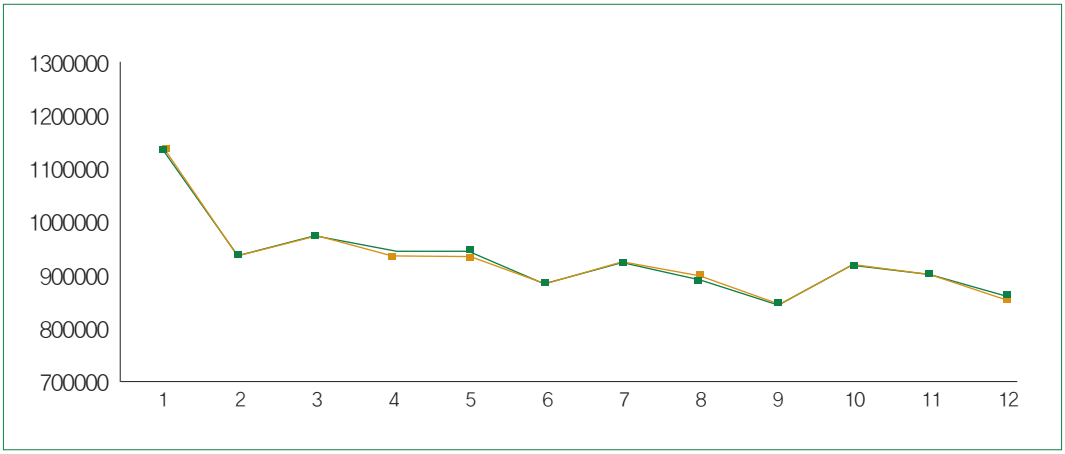

그림 2. 월별 1인당 평균 청구비용의 모집단과 표본집단 비교

표 4. 입원환자 30개 다빈도 질환에서 표본자료와 모집단의 상대비율

다빈도상병순위	상병코드	모집단 빈도	표본 빈도	백분율
1	O800	188,840	24,484	12.97
2	J189	139,603	18,025	12.91
3	A09	127,301	16,373	12.86
4	I841	126,508	16,510	13.05
5	H259	79,074	10,243	12.95
6	O820	78,332	10,298	13.15
7	K359	74,155	9,581	12.92
8	M511	69,381	9,052	13.05
9	I639	59,495	7,847	13.19
10	I10	53,087	6,860	12.92
11	S335	52,243	6,681	12.79
12	H251	45,861	5,903	12.87
13	O821	43,856	5,739	13.09
14	N10	42,737	5,519	12.91
15	H258	42,503	5,645	13.28
16	J459	41,512	5,431	13.08
17	N393	39,918	5,257	13.17
18	J209	38,260	4,958	12.96
19	J180	36,630	4,675	12.76
20	J039	33,588	4,257	12.67
21	D259	32,834	4,311	13.13
22	M4806	32,383	4,122	12.73
23	I209	30,714	4,018	13.08
24	M512	29,983	3,866	12.89
25	I200	29,760	3,858	12.96
26	C169	29,430	3,819	12.98
27	J157	28,704	3,714	12.94
28	S3200	27,184	3,479	12.80
29	M170	26,832	3,497	13.03
30	E119	26,500	3,469	13.09

95%신뢰수준 모집단 추정 오차구간을 산출한 결과, 만성신부전증(N18) 전체환자는 1,159,366,039명의 추정 값을 가지며 실제 모집단의 만성신부전증 환자는 110,354명으로 전체 환자와 각 연령 구간 모두 신뢰 구간 내에 참값이 존재함을 확인할 수 있다.

표 5. 표본자료를 이용한 만성신부전증(N18)환자의 모집단 추정

(단위: 명)

연령구간	95%신뢰하한	95%신뢰상한	모집단 환자 빈도
1~4세	11	81	48
5~9세	16	92	88
10~14세	78	199	192
15~19세	127	749	430
20~24세	123	1,053	625
25~29세	1,034	2,574	1,635
30~34세	1,820	3,715	2,802
35~39세	3,657	6,275	4,918
40~44세	5,454	8,600	6,696
45~49세	8,888	12,763	9,897
50~54세	9,986	13,928	12,298
55~59세	10,797	14,849	11,796
60~64세	11,266	15,349	12,808
65~69세	13,960	18,386	14,940
70~74세	12,290	16,323	13,858
75세이상	15,854	20,330	17,323
Total	109,263	121,252	110,354

2) 관련 학회[1]에 의한 자료의 타당성 검토

건강보험심사평가원은 환자표본자료의 타당도 평가를 위해 보건의료관련 5개 학회와 MOU를 맺었으며 MOU학회 회원을 대상으로 한 표본자료 활용 연구의 주요 결과를 보면 '우리나라 당뇨병 유병률 추정 및 DPP-4 억제제 사용 양상 평가[2]'에서 표본자료를 이용한 당뇨병 유병률 추정결과가 모집단 분석결과와 일치하였고, 혈당강하제 각 약효군별 처방률 추정결과가 모집단 분석결과와 일치하는 결과를 보였다. 또한 외래환자에서 각 약효군별 처방률은 모두 추정치의 95%신뢰구간 내에 참값 존재하였다.

1) 대한예방의학회, 보건경제정책학회, 보건정보통계학회, 보건행정학회, 한국역학회.
2) 박병주, 성종미. 우리나라 당뇨병 유병률 추정 및 DPP-4 억제제 사용 양상 평가.

'시력손실과 실명으로 인한 사회적 질병 부담비용 추계[3]' 연구에서는 표본자료와 모집단 모두에서 여자가 남자에 비해 모든 주요안과질환(백내장, 녹내장, 황반변성, 당뇨망막변중, 망막정맥폐쇄)에서 의료이용 환자 비율이 높은 것으로 나타났으며 백내장, 녹내장, 환반변성의 연령별 추이는 모집단과 표본자료가 비슷한 양상을 가지는 것으로 나타났다.

5. 결론 및 제언

본 연구의 결과로 모집단인 2009년 건강보험 청구자료에 대한 대표성과 개인정보보호에서 일정 수준을 만족하는 표본자료를 산출할 수 있게 되었다. 향후 표본자료는 지속적인 개발 및 보완을 통해 자료의 제공 영역을 확대해 나갈 계획에 있다.

환자 표본자료 활용 시 주의사항으로 급여가 인정된 의료이용 내역만 포함되어 있기 때문에 비급여 내역 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 입원환자, 다빈도 경증질환자 보다는 위중한 환자에서 높게 나타나며 의원급보다는 종합병원급 요양기관에서 더 높은 경향이 있다. 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.

환자표본자료의 제한점은 모든 표본 자료가 갖게 되는 공통의 한계점으로서 표본자료내의 관측치는 확률에 의해 추출되는 자료이기 때문에 적정수준이상의 표본수를 확보해야 대표성, 유의성을 보장받을 수 있다. 예를 들어 본 환자표본자료에서 특정 연령대의 희귀질환 발생빈도의 경우 표본추출 빈도가 너무 적어 대표성과 설명력이 떨어질 수 있다. 따라서 표본자료의 설명력은 다빈도 상병 일수록 커지며, 상병의 빈도가 떨어지면 감소하게 된다.

이를 보완하는 방안으로 500건 이하 발생빈도의 상병은 전수 제공하는 방안이 있으며 자료의 제공영역을 1년 단위가 아닌 일정기간 동안의 패널자료를 구축해 추출 빈도율을 높이는 방안도 고려해 볼 수 있다. 또한 표본자료 변수의 제공 범위를 확대하고 치과, 산부인과 등의 특정 진료과나 65세 이상의 고령 환자로 표본자료를 세분화시켜 개발하는 방안도 고려해 볼 수 있다.

건강보험심사평가원 환자 표본자료를 통해 다양한 보건의료 분야에서 연구에 활용되어 공공의 편익을 제고할 수 있을 것으로 기대된다.

3) 최상은. 시력손실과 실명으로 인한 사회적 질병 부담비용 추계.

참고문헌

1. 김재용, 임지혜, 김화영. 진료에피소드 중증도 보정 및 예측용 지표 개발. 서울: 건강보험심사평가원; 2007.
2. 남궁평. 표본조사설계와 분석. 2007.
3. 건강보험심사평가원 조사연구실. 표본조사 방법론 개발에 관한 연구. 서울: 건강보험심사평가원; 2003.
4. 신의철 등. 진료정보의 적정활용 방안에 대한 연구. 서울: 대한예방의학회·건강보험심사평가원; 2008.
5. 한국보건정보통계학회. 보건사회 조사 연구에서의 통계와 정보. 2009 한국보건정보통계학회 추계 학술대회.
6. AHRQ. Design of the Nationwide Inpatient Sample(NIS). Agency for Healthcare Research and Quality; 2005.
7. AHRQ. Intriduction to the HCUP Nationwide Inpatient Sample(NIS). Agency for Healthcare Research and Quality; 2007.
8. Bethel. J. W. An Optiomum Allocation Algorithm for Multivariate Surveys, Proceedings of th Social Statistics Section. American Statistical Association; 1985.
9. Cheng. S. H. Introduction to National Health Insurance database. Chin. J. Public Health 1999; 18: 235-236.
10. Chen. C. J. Lin. L. H. Use of National Health Insurance database in academic research: experiences from analysis of major disease certification profile. Chin. J. Public Health 1997; 16: 513-521.
11. Chatterjee. S. Multivariate Stratified Survey. Journal of American Statistical Assocation 1968.
12. Hidiroglou. M. A. Srinath. K. P. Problems associate with designing sub- annual business surveys. Journal of Business and Economic Statistics 1993.
13. Inho P. Hyunshik L. THE DESIGN EFFECT: DO WE KNOW ALL ABOUT IT?. Proceeding of the Annual Meeting of the American Statistical Association 2001.
14. National Health Research Institute. Researches of National Health Insurance claim database. http://www.nhri.org.tw