

데이터마이닝 기법을 이용한 건강보험공단의 수술 통계량 근사치 추정 -허니아 수술을 중심으로-

강성홍, 서숙경¹⁾, 양영자²⁾, 이애경³⁾, 배중면²⁾

인제대학교 보건행정대학, 서울아산병원의무기록실¹⁾, 제주대학교 의과대학 예방의학교실²⁾, 국민건강보험공단³⁾

Estimation of a Nationwide Statistics of Hernia Operation Applying Data Mining Technique to the National Health Insurance Database

Sunghong Kang, Seok Kyung Seo¹⁾, Yeong-Ja Yang²⁾, Aekyung Lee³⁾, Jong-Myon Bae²⁾

School of Health Administration, Inje University; Medical Record & Informatics Team, Asan Medical Center¹⁾;
Department of Preventive Medicine, Cheju National University College of Medicine²⁾; National Health Insurance Corporation³⁾

Objectives : The aim of this study is to develop a methodology for estimating a nationwide statistic for hernia operations with using the claim database of the Korea Health Insurance Cooperation (KHIC).

Methods : According to the insurance claim procedures, the claim database was divided into the electronic data interchange database (EDI_DB) and the sheet database (Paper_DB). Although the EDI_DB has operation and management codes showing the facts and kinds of operations, the Paper_DB doesn't. Using the hernia matched management code in the EDI_DB, the cases of hernia surgery were extracted. For drawing the potential cases from the Paper_DB, which doesn't have the code, the predictive model was developed using the data mining technique called SEMMA. The claim sheets of the cases that showed a predictive probability of an operation over the threshold, as was decided by the ROC curve, were identified in order to get the positive predictive value as an index of usefulness for the predictive model.

Results : Of the claim databases in 2004, 14,386 cases had hernia related management codes with using the EDI system. For fitting the models with applying the data mining technique, logistic regression was chosen rather than the neural network method or the decision tree method. From the Paper_DB, 1,019 cases were extracted as potential cases. Direct review of the sheets of the extracted cases showed that the positive predictive value was 95.3%.

Conclusions : The results suggested that applying the data mining technique to the claim database in the KHIC for estimating the nationwide surgical statistics would be useful from the aspect of execution and cost-effectiveness.

J Prev Med Public Health 2006;39(5):433-437

Key words : Estimation techniques, Logistic models, Predictive value of tests, Computer data processing, Korea Health Insurance Cooperation

서론

각종 수술에 관련한 통계는 국가의 기본적인 보건통계의 하나로써 보험업무 및 국가 보건 정책 수립의 기본 자료인 점에서, 국가차원에서 체계적으로 수술데이터를 수집하여 수술통계를 작성할 필요가 제기되고 있으나, 국가 차원의 수술통계 작성은 많은 시간과 비용이 소요되므로 단기간에 이루어지기는 매우 어려운 실정이다 [1].

우리나라는 1989년에 전국민 의료보험을 도입한 이후 세계 최대 규모의 보건의

료자료가 구축될 수 있게 되면서, 적은 시간과 비용으로 수술통계를 작성할 가능성이 제시되었다 [2]. 물론 건강보험 추진자료를 이용한 수술통계는 국가의 공식적인 수술통계로는 제한점이 있지만 건강보험업무나 국가보건정책 수립의 기초 자료로는 활용 가능성이 높음으로 건강보험공단의 수술데이터를 이용하는 수술통계를 작성할 필요성이 대두되고 있다 [3]. 그런데, 건강보험공단의 수술데이터는 의료보험 청구 서식에 따라 전산청구 자료와 서면청구 자료로 구성되어 있다 [2]. 우선 EDI 형태로 청구되는 전산청구 자료에서는 주

요 수술에 해당되는 수술코드가 있어 이를 활용하여 단순히 전산처리만 하면 되기 때문에 주요 수술 통계 산출이 용이하다. 반면에 서면 형태로 청구된 자료는 수술을 나타내는 코드가 전산입력이 되어 있지 않아, 통계 산출을 위해서는 추가적으로 대규모의 조사비용과 시간이 투입되어야 한다. 다시 말해서, 서면청구자료 전체를 수작업 방식으로 조회를 하여서 이중 구축대상이 되는 주요 수술청구만을 추출하여 전산 입력을 새로 해야만 한다. 이렇듯, 서면청구 자료에 대한 추가적인 요구 작업 때문에 건강보험관리공단의 수진청구 자료를 활용한 수술통계를 산출하려면 서면청구 자료에 대한 방대한 조사

접수: 2006년 5월 12일, 채택: 2006년 7월 6일

본 연구는 2004년도 국민건강보험공단 연구용역사업 지원으로 이루어진 것임.

책임저자: 배중면 (제주특별자치도 제주시 아라1동 1번지, 전화: 064-755-5567, 팩스: 064-725-2593, E-mail: jmbae@cheju.ac.kr)

비용과 조사시간을 단축할 수 있는 방안을 강구하는 것이 관건이다 [4].

최근 개발되어 다방면에서 활용되고 있는 데이터 마이닝은 방대한 자료에서 의미 있는 정보를 창출하는 기법으로, 분류, 추정, 예측, 유사 집산화, 군집화, 서술의 6개 영역에서 활용이 되고 있으며 [5], 건강보험공단자료에 적용하여 예측하는 시도가 있어 왔다 [6-9]. 이런 배경 하에 데이터 마이닝 기법을 건강보험공단의 전산청구 자료에 적용하여 수술통계 작성에 필요한 자료를 선정할 수 있는 모델을 만든 후 이 모델을 서면청구 자료에 적용하여 수술통계작성에 필요한 자료의 확률을 예측한 것이 유용하다면, 공단자료를 활용한 수술통계를 산출하는 하나의 방법으로 제시될 수 있을 것이다 [4]. 이런 배경에서 건강보험공단 자료 중 허니아 수술을 중심으로 서면청구 자료에 대한 조사모형을 데이터마이닝 기법을 통해 개발하여 그 정확성과 유용성을 알아보는 것이 본 연구의 목적이다.

연구방법

1. 수술통계 산출을 위해 활용 가능한 자료

국가 차원의 수술 통계량 중 허니아 수술을 본 연구 대상으로 삼았다. 그 이유는 OECD (Organization for Economic Cooperation and Development)가 요구하는 수술통계 중의 하나이면서, 해당 수술 코드의 호환성에 문제가 없고, 응답이 아닌 선택 수술이 이루어지기 때문이다. 또한 실제 수술이 이루어진 경우는 외래가 아닌 입원으로 의료보험 청구가 되는 점을 반영하고, 실제 수술여부를 직접 확인하기에 용이한 종합병원에서의 진료로 한정하기 위하여, 진료개시일이 2004년 1월1일부터 12월 31일까지 종합병원에서 이루어진 총 1,112,592건의 입원 현물급여자료를 확보하였다. 이들 현물급여자료에서, 본 연구를 위하여 총 12개 - 요양기관, 진료과, 주민번호, 주상병, 부상병, 진료개시일자, 투약일수, 입내원일수, CT진료비, 심사결정 총진료비, 심사결정공단분담금, 청구매체 구분-의 변수를 추출하였다. 그러나 현물급여자료에는 급여비 지불에 대한 요약 정

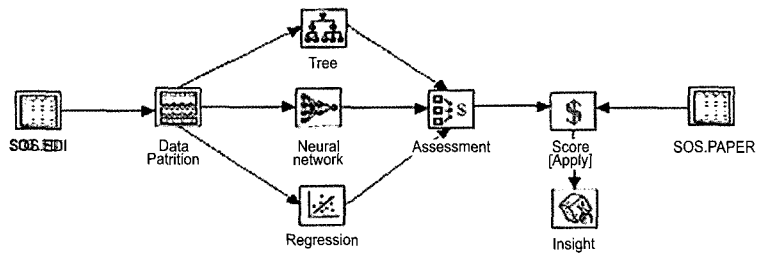


Figure 1. Flowchart of data mining.

보이므로 수술코드 등의 정보를 가지고 있지 않는 한계가 있다 [2]. 이에, 수술코드를 가지고 있는 전산청구자료를, 앞서 확보한 현물급여자료와 결합 (merge)한 결과 79%인 878,818건에 대하여 수술 및 처치코드를 붙일 수 있었다. 이렇게 결합되는 자료는 공단에 전산청구방식으로 급여를 요청한 것이기에 '매체DB' (Figure 1의 SOS.EDI)라 명명하였고, 결합되는 과정에서 수술 및 처치코드를 가지게 되었다. 반면, 나머지 21%인 233,774건은 결합이 안되는 자료로서, 이들은 공단에 서면청구방식으로 급여를 요청한 것이기에 수술 및 처치코드가 없으며 '서면DB' (Figure 1의 SOS.PAPER)라 구분하였다.

2. 변수 가공 및 변환

전산청구자료가 가지고 있는 항목 중 수술 및 처치코드가 Table 1과 같은 경우에 허니아 수술을 시행한 경우로 정의하고 수술 유무 변수를 생성하였다. 이때, 단 일회 입원한 가운데 동일한 수술군으로 여러 번의 수술 청구가 있을 경우는 하나의 수술로 간주하였다. 진료개시일과 생년월일을 이용하여 연령을 산출한 뒤, 0-4세, 5-9세와 같이 5세 간격으로 연령군을 만들었다. 주민번호의 7번째 숫자를 활용하여 성별 변

수를 만들었고, 심사결정 총진료비를 입내원일수로 나누어 일일 진료비를 새로 만들었다. 허니아 수술을 받은 대상이 가진 주/부 진단코드의 빈도를 이용하여 주/부 진단분율이라는 새로운 변수를 생성하였다. 진료과, 투약일수, 입내원일수, CT 진료비는 자료를 그대로 활용하였다.

3. 데이터 마이닝 분석

변수 가공 및 변환을 거쳐 구축된 매체 DB에 데이터마이닝 기법을 이용하여 수술 유무를 예측하고자 하는 모델을 만든 다음, 해당 모델에 근거하여 서면DB에 대해 해당 수술에 대한 수술 확률값을 예측하였다. 이를 위하여 SAS사의 데이터마이닝 소프트웨어인 SAS E-miner를 사용하였다 [10]. 데이터마이닝을 이용하는 모델링에 사용된 방법론은 SAS사의 SEMMA라는 방법론으로 sampling, exploration, modification, modeling, assessment 단계에 따라 모델링을 하는 방법론이다 [11].

Figure 1은 본 연구의 마이닝 분석과정을 도식화한 것이다. 우선 '매체DB'에서 허니아 수술군의 특성을 살펴본 뒤 (exploration), 포이송 회귀 분석 결과를 활용하여 변수들을 재조정 하였다 (modification). 다음으로 설명변수를 input 변수로 지정하고, 수술

Table 1. Operating and management codes of hernia operation

Codes	
Q2721	Operation of umbilical hernia (with resection of Intestine)
Q2722	Operation of umbilical hernia (others)
Q2731	Operation of incisional hernia (with resection of intestine)
Q2732	Operation of incisional hernia (others)
Q2740	Manual reduction of incarcerated hernia
Q2753	Operation of inguinal hernia (with resection of intestine, high ligation)
Q2754	Operation of inguinal hernia (with resection of intestine, high ligation and posterior repair)
Q2755	Operation of inguinal hernia (others, high ligation)
Q2756	Operation of inguinal hernia (others, high ligation and posterior repair)
Q2757	Operation of femoral hernia
QA753	Operation of recurrent inguinal hernia (with resection of intestine, high ligation)
QA754	Operation of recurrent inguinal hernia (with resection of intestine, high ligation and posterior repair)
QA755	Operation of recurrent inguinal hernia (others, high ligation)
QA756	Operation of recurrent inguinal hernia (others, high ligation and posterior repair)

Table 2. Characteristics of subjects taking hernia operation in electronic data interchange database*

Variables		% (N=14,366)
Sex	Male	82.08
	Female	17.92
Age	0-4	27.29
	5-9	17.28
	10-14	4.16
	15-19	1.36
	20-24	2.06
	25-29	1.98
	30-34	2.35
	35-39	2.60
	40-44	3.05
Clinics	45-49	4.02
	50-54	4.26
	55-59	5.09
	60-64	6.37
	65-	18.24
	General Surgery	87.62
	Urology	10.81
	Others	1.57
	Days for medication	Q1
Q2		8
Q3		10
Days for admission	Q1	3
	Q2	4
	Q3	6
Cost of CT	Q1	0
	Q2	0
	Q3	0
Daily fee	Q1	145,898
	Q2	180,084
	Q3	220,248
% of major diagnosis*	Q1	88.10
	Q2	88.10
	Q3	88.10
% of minor diagnosis*	Q1	2.71
	Q2	58.54
	Q3	58.54

* Q1: less than 25 percentile, Q2: 25-50 percentile, Q3: 50 percentile and over

* percentage of major/minor disease code among cases taken the hernia operation

여부를 결과변수(target)로 정한 뒤, 분석용(training) 데이터와 검증용(test) 데이터로 50:50 비율로 분할(partition)하였다(sampling). 이후 데이터 마이닝의 대표적인 모형화 알고리즘인 Decision tree, Neural Network, Logistic regression 을 적용하여 상호 비교를 수행하였다(modeling). 모델의 타당성과 안전성을 평가하기 위하여 이익도표(Lift chart)와 ROC 곡선을 구하였다(assessment). 이렇게 선정된 예측모형을 '서면DB'에 적용하여 수술 예측값을 구하였고, ROC 곡선에서 얻어낸 기준점을 중심으로 그 이상의 수술 예측값을 보인 서면청구 건을 추출하였다. 추출된 이들 청구물의 명세서를 보관창고에서 확인 대조하여 실제 수술유무를 확인한 결과를 근거로 양성예측도를 산출하여 수술통계를 위한 서면청구 자료의 마이닝 기법 적용

가능성을 평가해 보았다. 결과 해석에서 1종 통계오류 한계를 5%로 간주하였다.

결 과

1. 매체DB 자료의 허니아 수술 시행 건의 특성

매체DB 자료에서 Table 1의 수술 및 처치 코드에 따라 종합병원에서 허니아 수술을 받은 경우는 총 14,366건이었다(Table 2). 이들의 의료수진과 관련한 특성을 살펴보면 남자가 많았고, 영유아에서 많이 시행되었으며, 주로 외과와 비뇨기과에서 수술이 이루어졌다. 투약일수, 재원일수, 일일진료비의 중앙값은 각각 8일, 4일, 18만원이었다. 수술질환의 특성상 CT 시행은 되지 않았다.

2. 다변량 분석 모델링용 범주화 및 포아송 회귀분석 결과

매체DB 자료에서 표 2의 대상 건수에게 수술 여부라는 새로운 파생변수에 1의 값을 부여하고, 나머지 건은 0의 값을 부여하였다. 한편, 원 자료 그대로 데이터 마이닝에 적용할 수 있겠으나, 허니아 수술의 진료 특성이 반영될 수 있도록 Table 2의 변수 분포를 이용하여 다음과 같이 변수들을 범주화 하였다. 예를 들어, 진료과가 일반외과와 비뇨기과에서 주로 이루어진 특

성을 반영하여 해당 과에서 시술한 경우와 아닌 경우로 묶고; 투약일수, 재원일수, 일일진료비는 25-75% 분포에 들어온 경우와 아닌 경우로 나누며; CT 진료비는 0인 경우를 중심으로 범주화 하였다. 또한 주진단분율이 88.1% 이상인 경우와 부진단분율이 2.71% 이상인 경우를 중심으로 변수를 각각 재정의하였다.

이상의 범주화된 변수들을 수술여부를 종속변수로 한 포아송 회귀분석모형에 적용한 결과, 투약일수와 일일 진료비를 제외한 성별, 연령군, 진료과, 투약일수, 재원일수, CT 비용, 주진단분율, 부진단분율이 1종 오류 5% 수준에서 통계적으로 유의하게 나왔다(Table 3).

3. 매체DB 자료의 데이터마이닝을 이용한 수술여부 예측모형 개발 결과

Table 3에서 유의성을 보인 설명변수를 input 변수로 지정하고, 수술 여부를 결과변수(target)로 정한 뒤, 매체DB 자료를 분석용(training) 데이터와 검증용(test) 데이터로 50:50 비율로 재분할 하였다. 이후 데이터 마이닝의 대표적인 모형화 알고리즘인 Decision tree, Neural Network, Logistic regression 을 적용한 뒤, 모델의 타당성과 안전성을 평가하기 위하여 이익도표(Lift chart)와 ROC 곡선을 구한 결과 3가지 모형 모두 별 차이가 없는 것으로 나왔다(Figure

Table 3. Results of univariate analysis by Poisson regression model in electronic data interchange database*

Variables	Baseline	Comparison	Estimate	p-value
Sex	Female	Male	0.2604	<0.0001
Age	0-4	5-9	0.7877	<0.0001
		10-14	-1.7483	<0.0001
		15-19	-2.4629	<0.0001
		20-24	-2.4165	<0.0001
		25-29	-2.5929	<0.0001
		30-34	-2.4415	<0.0001
		35-39	-2.1188	<0.0001
		40-44	-1.9269	<0.0001
		45-49	-1.6682	<0.0001
		50-54	-1.2321	<0.0001
		55-59	-1.0245	<0.0001
Clinics	Others	Surgery/Urology	-3.7026	<0.0001
		Q1-Q3	-0.1025	0.0832
		Q1-Q3	0.5902	<0.0001
Days for medication	Others	Q1-Q3	-0.1025	0.0832
Days for admission	Others	Q1-Q3	0.5902	<0.0001
Cost of CT	≠0	0	-0.2066	0.0341
Daily fee	Others	Q1-Q3	-0.0000	1.0000
% of major diagnosis*	Others	Q1-	-9.1879	<0.0001
% of minor diagnosis*	Others	Q1-	-1.3212	<0.0001

* Q1: less than 25 percentile, Q2: 25-50 percentile, Q3: 50 percentile and over

* percentage of major/minor disease code among cases taken the hernia operation

Table 4. Results of logistic regression analysis by data mining procedure

Variables	Estimate	p-value	Odds ratio
Sex	0.2746	<0.0001	1.732
Age	-0.0603	<0.0001	0.942
Clinics	-1.6003	<0.0001	0.041
Days for admission	0.3882	<0.0001	2.174
% of major diagnosis*	-4.5677	<0.0001	0.001
% of minor diagnosis*	-0.5873	<0.0001	0.309

*percentage of major/minor disease code among cases taken the hernia operation

2,3). 3가지 알고리즘이 비슷한 것으로 평가됨에 따라 결과변수에 영향을 끼치는 설명변수의 중요도를 가장 쉽게 알 수 있는 Logistic regression을 수술여부의 예측모형의 최종적인 모델로 선정하였다. 선정된 Logistic regression의 threshold의 값이 10 일때 민감도와 특이도는 각각 98.84%와 99.84%로 평가되었다. Logistic regression 모델을 이용하여 Link function을 Logit으로 정하고, profit/loss 기준에 따라 backward 선택법을 적용한 결과, 매체DB 자료에 있어 모형상수와 상대위험도 (OR)는 Table 4와 같았다.

이렇게 개발된 예측모형을 이용하여 서면DB 자료에 적용한 결과, ROC 곡선을 통해 기준점 (threshold) '10' 이 얻어졌다 (Figure 2). 이를 근거로, 수술 예측값이 0.90 이상인 건수를 확보한 결과 전체의 0.17%인 1,019건으로 나왔다. 이들의 서면청구 명세서를 일일이 찾아 실제 허니아 수술을 받았는가를 직접 조사 확인한 결과, 이 중 961건이 실제 허니아 수술환자로 판명되었기에 데이터마이닝을 이용한 양성예측도는 95.3%이었다. 또한 수술 예측값이 0.90 미만인 경우 중 다른 수술이 의심이 되어 명세서를 확인한 39,910건에 있어 허니아 수술 사례로 밝혀진 경우가 159건으로 음성예측도는 99.6%이었다.

고 찰

국가 차원의 수술통계가 국가의 기본통계로서 보험업무 및 국가 보건 정책 수립의 기본자료로 필요하나 수집비용이나 시간상의 문제로 단시간에 산출하기는 어려운 실정이다 [12]. 이러한 어려운 여건 하에서 세계 최대 규모의 보건의료자료를 전산화한 건강보험공단의 청구데이터를 이용하여 건강보험의 수술통계를 작성할

필요성이 대두되었다 [13]. 물론 건강보험 수진자료를 이용한 수술통계는 국가의 공식적인 수술통계로는 제한점이 있지만 건강보험업무나 국가보건정책 수립에는 효율적으로 활용될 수 있으므로 건강보험공단의 청구자료를 이용한 수술통계를 작성할 필요가 제시되었다 [14].

한편, 데이터마이닝은 산업 전반에서 널리 활용되고 있으며, 보건의료분야에서는 의료의 질 향상, 병원경영, 의학연구 등에 널리 활용되기 시작하고 있다 [15-17]. 그러나 조사방법 분야에는 활용된 사례가 극히 미비한데 본 연구에서 데이터마이닝 기법을 이용한 조사방법을 새로이 개발했다는 점에 의미를 둘 수 있다. 즉, 건강보험공단의 청구자료를 이용하여 수술통계를 작성시 매체청구 자료는 자료수집이 용이하나 전체 청구의 10% 정도인 서면청구에 대한 수술데이터의 수집은 많은 시간과 비용이 소요되므로 데이터마이닝이란 새로운 조사방법을 적용하여 서면청구 자료

중 조사하고자 하는 수술을 한 확률이 높은 청구건수 만을 추출하는 조사방법을 개발한 것이다 [4]. 이런 배경 하에 본 연구가 가지는 의미는, 종합병원의 서면청구 자료 233,774건이지만 이중 0.17%에 해당하는 1,019건만을 추출하여 조사함에 따라 많은 시간과 비용을 절감할 수 있었으며, 향후 건강보험공단에서 비용-효과적인 수술통계 조사방법으로 활용할 가능성을 제시한 것이다.

이번에 개발한 조사방법은 종합병원의 허니아 수술에 국한하였기 때문에, 다른 의뢰기관이나 다른 수술에 대한 통계 산출을 위해서는 활용 가능한 변수의 선정과 범주화를 각각 달리 해야만 할 것이다 [4]. 또한, 년도를 달리한 허니아 수술 자료도 지속적으로 변수의 특성을 비교 분석하여 수술 예측을 높이도록 변수들을 정밀화 시켜야 할 것이다. 그러나 서면청구 자료가 있는 한, 명세서 확인이란 시간과 인력 투입을 요하는 추가 작업이 필요하

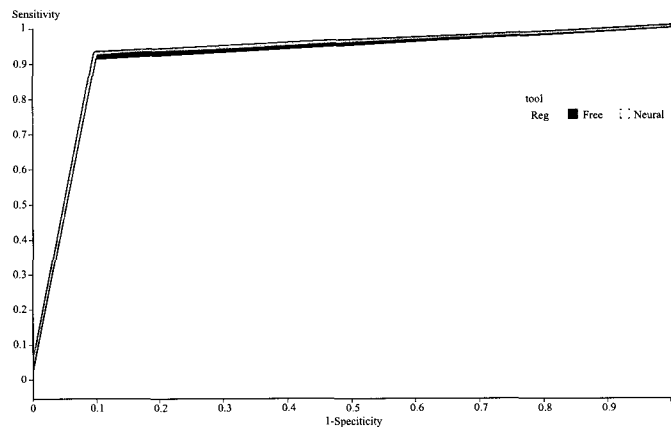


Figure 2. ROC curves of logistic regression (Reg), decision tree (Tree), and neural network (Neural).

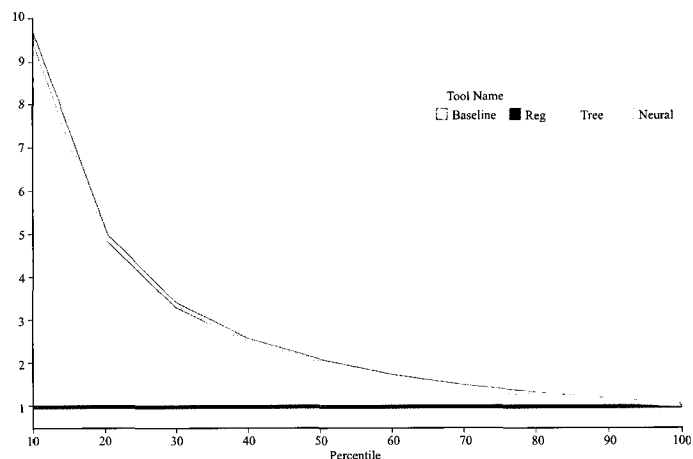


Figure 3. Lift charts of logistic regression (Reg), decision tree (Tree), and neural network (Neural).

다는 점에서 서면 청구 접수 때 수술코드를 추가로 입력하는 것이 향후 안정적이고 지속적인 통계 산출을 위해서 더 효율적이다 [2].

이상과 같이 다각도로 통계 방법론적인 노력을 한다고 해도 국민건강보험공단의 수진자료가 가지고 있는 몇 가지 특성들을 염두에 두어야만 공단의 수술 통계 연보가 타당성과 신뢰성을 확보할 수 있다. 첫째, 의료비 청구의 특성상 특정 해의 연말에 진료개시가 된 건강보험 수진 건은 통상 그 다음해의 상반기에 청구되므로, 이들 자료를 모두 수합하여 정리하려면 그 다음해의 연말에야 가능하다는 것이다. 경우에 따라서는 다음해의 상반기 이후에도 청구할 가능성이 있어 연보 작성 상 이들 누락을 극복할 방안을 강구해야 한다. 둘째, 청구 자료는 수진에 대한 의료비 청구 목적으로 작성된 것으로, 통계 산출을 위해서는 청구 자료의 타당성에 대한 의문이 제기될 수 있다. 건강보험 청구 자료는 진료비 청구를 관리할 목적으로 구축된 분류 시스템으로, 수술통계를 산출하기 위해서 필요한 ICD-9-CM 분류시스템과는 사용목적과 용도가 달라 이종간의 호환 코드표를 확보해야만 가능하기 때문이다. 즉, 보험 청구에 사용된 내용과 실제 의무기록에 기재된 수술 내용에는 상당한 차이가 있어 [2], 연보 작성에 포함되는 수술 종류는 제한될 수밖에 없는 것이다 [4].

요약 및 결론

본 연구에서는 허니아 수술을 대상으로 하여 데이터마이닝을 이용하는 조사방법을 개발하고 이의 효과성을 평가하였다. 모델개발단계는 매체청구자료를 이용하여 허니아 수술여부의 예측모델을 개발하였다. 예측모델로 검토된 알고리즘은 의사결정나무, 신경망분석, 로지스틱 회귀 분석이었다. 3가지 알고리즘의 우수성을 ROC 곡선과 Lift chart 등을 이용하여 비교한 결과 알고리즘의 성능은 비슷하게 나오며 따라 결과 해석의 용이성을 고려하

여 Logistic regression을 최종 모델로 선정하였다. 선정된 모델의 민감도와 특이도는 각각 98.84%와 99.84%로 평가되어서 통계량 생성에 활용할 가치가 있다고 보겠다.

모델 개발시에 모델의 내적 타당도를 고려하여 전체 데이터를 training 50%, test 50%로 구분하여 모델 개발과 평가에 사용하였다. 그러나 이와 같은 방법만으로는 과적합이 없다는 것을 입증하기에는 부족한 점이 있음에 따라 본 연구에서는 외적 타당성을 알아보기로 하였다. 즉, 서면청구물중 허니아 수술건수로 예측된 1,019건에 대해 실제 서면청구물을 찾아서 수술 여부를 확인하였다. 확인한 결과 이중 961건이 실제 허니아 수술환자로 판명되어서 양성예측도 95.3%, 음성예측도 99.6%로 나왔다. 이에 따라, 본 연구에서 개발한 조사방법론이 정확하고 신뢰할 수 있는 것으로 판명되었으며, 향후 건강보험관리공단의 수술통계 연보 작성에 활용할 수 있을 것이다.

참고문헌

1. Jang YS, Ko YH, Doh SR, Lee LH, Seo SW. Study on the Production of 2004 Health Data in OECD. Ministry of Health welfare Korea Institute of Health and Social Affairs, 2004
2. Sung JY, Kang SH, Kim ON, Research for Surgical Operations Management and Constructing DB for Major Surgical Operations in NHIC. National Health Insurance Cooperation, 2003
3. Song TM, Lee YH, Lee GH, Lim GY, Cha YM, Jung SY, Lee CK. A Study for Establish and Application of Database in NHIC of the based Knowledge. Korea Institute of Health and Social Affairs, 2002
4. Bae JM, Kang SH, Kim ON, Park BJ. Statistics for Major Surgical Operations in the National Health Insurance, 2004 by Constructing DB for Major Surgical Operations. National Health Insurance cooperation, 2005 (Korean)
5. Cho GL, Jo DH, Lee SL. Theory and Exercise for Data Mining. Chung Goo. 2001
6. Kim YM, Park IL, Kang SH. Development of HRA of hypertension using data of NHIC data. *J Health Inform Institute*, 2004; 10: 47-56

- (Korean)
7. Lee Ak, Jung HJ, Park IS. Development of a Health Management Model for Tailored Health Information Service using the Data Mining Technique. National Health Insurance Cooperation, Health Insurance Research Center, 2004
8. Hong DH, Lee JG, Jo My, Park GD, Lee MS, Lee SI, Kim CY, Kim YI. Efficient DRG fraud candidate detection method using data mining techniques. *Korean J Prev Med* 2003; 26(2): 147-152 (Korean)
9. Lee SM, Kang JO, Se YM. Predictive model of cancer patient's medical care expense using data mining. *J Korea Soc Manage Inform Syst* 2003; 1: 664-671
10. Greg R, Ellen J. Data Mining for Health Care Quality Improvement SAS institute, Gary, NC, 1998
11. Bae HS, Cho DH, Seok KH, Kim BS, Choi KL, Lee JY, No SY, Lee SC, Sohn YH. Data Mining Using SAS Enterprise Miner. Kyow Oosa, 2004.
12. Jang YS, Ge HB, Doh SR, Go KH, Seo JS, Seo SW, Bu YK. A Study on Health Promotion Statistics in 2000 OECD. Ministry of Health welfare, 2000
13. Korea Health Industry Development. A Study on Statistics Index Development using Medical Insurance Data. Ministry of Health welfare, 1999
14. Jang YS, Doh SR, Ko KH, Seo JS, Seo SW, Bu YK, Lee LH. A Study of Health Data Production : OECD Comparison and Improvement Strategies. Ministry of Health Welfare Korea Institute of Health and Social Affairs, 2003
15. Judith H. Hibbard, Ellen P. Supporting Informed customer health care decision: Data presentation approaches the facilitate the sue of information in choice. *Ann Rev Public Health* 2003; 24: 413-433
16. Park YS, Kang SH, GU BB, Kim BC, Kim YN, Kim WJ, Ru Mi, Song SK, Song HD, Ha JY. Hospital Information Management. Goryeo Medical Science, 2005
17. Kang HC, Han ST, Che JH, Lee SK, Kim YS, Yem YH, Kim MK. Data Mining Using SAS Enterprise Miner 4.0 -Methodology and Application - Third Edition. Freedom Academy, 2001