Journal of the Korean Data & Information Science Society
2010, **21**(6), 1253–1261

한국데이터정보과학회지

데이터마이닝을 이용한 위암 예측모형 개발과 활용

박일수[1] · 한준태[2] · 강석복[3] · 지재훈[4]

[1,2]국민건강보험공단 건강보험정책연구원
[3]영남대학교 통계학과
[4]인제대학교 병원전략경영연구소

접수 2010년 10월 1일, 수정 2010년 11월 18일, 게재확정 2010년 11월 23일

요 약

본 연구는 국민건강보험공단의 건강검진데이터, 자격 및 보험료, 그리고 진료비 데이터를 활용하여 위암 발생 예측모형을 개발하고자 하였다. 모형개발에는 데이터마이닝 방법론에 의한 로지스틱 회귀모형을 활용하였으며, 모형개발은 남성, 여성 그리고 전체에 대해 각각 개발하여 각 모형에서 위암 발생 결정요인의 차이를 비교하였다. 그 결과 위암 발견 예측에 가장 큰 영향을 미치는 특성은 수검자의 연령이었고, 다음으로 음주, 가족병력 (암) 순으로 나타났다. 남자가 여자보다 위암 발견 가능성이 다소 높은 것으로 나타났으며, 남성의 경우는 연령, 여성의 경우는 음주유무가 위암 발생에 많은 영향을 미치는 것을 확인 할 수 있었다.

주요용어: 로지스틱 회귀분석, 예측모형, 위암, 위험요인.

1. 서론

세계보건기구 (WHO: World Health Organization)의 산하기구인 국제암연구소 (IARC: International Agency of Research on Cancer)의 자료에 의하면, 암은 심장혈관 질환에 이어 두 번째로 사망률이 높은 질환이이며, 비전염성 만성질환 중 예방율이 가장 높은 질환 중의 하나이다. 전 세계적으로 2005년에는 760만명이 암으로 사망하였고, 2015년에는 9백만명, 2030년에는 11.5백만명이 암으로 사망할 것으로 예상하고 있다. 그러나 암으로 인한 죽음의 40%는 흡연률의 감소, 식이습관 향상, 운동, 음주, 직장내 발암물질 제거, B형 간염 바이러스 그리고 인유두종 바이러스 (HPV: Human Papillomavirus) 면역체계 형성으로 예방가능하다고 밝히고 있으며, 이를 통제하지 못할 경우 꾸준히 증가할 것이라고 경고하고 있다 (WHO, 2007).

일반적으로 암 질환은 초기에 그 증상이 나타나지 않고 상당히 진행된 이후에 발견되는 경우가 많아 최근 암 예방에 대한 관심이 크게 증가하고 있다. 그러나 암 예방을 위해서 개인의 과거 건강상태 정보를 근거로 특정 질환이 발생할 가능성이 큰 고위험군 선정 시스템이나, 이를 이용한 고위험군의 특성별로 차별화된 사후관리체계는 아직 마련되지 않고 있다. 현재 전체 암 발생이 계속 증가하고 있는 가운데, 특정암 검진항목으로 들어간 위암, 대장암, 간암 그리고 유방암 중 위암을 제외한 다른 암 발생은 매년 급증하고 있다.

[1] (121-749) 서울 마포구 염리동 168-9, 국민건강보험공단 건강보험정책연구원, 부연구위원.
[2] (121-749) 서울 마포구 염리동 168-9, 국민건강보험공단 건강보험정책연구원, 부연구위원.
[3] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: sbkang@yu.ac.kr
[4] (633-165) 부산 부산진구 개금동 614-735, 인제대학교 병원전략경영연구소, 연구강사.

이러한 점에서 우리나라 암 건강검진사업의 궁극적 목표인 질병과 사망률 감소를 위해서는 건강한 생활습관 및 예방을 유도할 수 있는 관리 프로세스의 개발이 필요하다고 본다. 유럽 및 미국 등의 선진국에서는 대규모 코호트를 대상으로 고혈압, 당뇨, 고콜레스테롤혈증, 흡연 등의 주요한 위험요인을 밝혀내고 질병 예측모형 개발등을 통하여 적극적인 예방 노력을 펼쳤다 (D'Agostino 등, 2001). 또한 현재 세계적으로 가장 널리 사용되는 관상동맥심질환 발병위험도인 프래밍험 예측모형을 중국인에 적용 시켜본 결과 과추정되는 문제점도 제시하였다 (Liu 등, 2004).

데이터마이닝을 활용한 연구로는 건강증진사업에 데이터마이닝기법을 활용하는 방법을 제시한 연구 (강성홍과 최순호, 2001)와 이애경 등 (2006)은 대장암 발생 고위험군의 예측모형을 제안하였으며, 향후 예측모형의 활용 방안을 제안하였다. 그리고 용왕식 등 (2006)은 고혈압 발생위험 확률을 예측할 수 있는 모형을 개발하고, 개발된 모형을 이용한 고혈압 예방사업 활용방안을 제시하였으며, 박일수 등 (2008)은 국민건강보험공단의 검진데이터를 활용하여 고혈압 관리를 위한 맞춤형 고혈압 사후관리모형을 데이터마이닝 기법을 활용하여 개발하였다. 최근에는 건강검진 수검자를 10년간 추적하여 허혈성심질환에 대한 사망위험도를 분석하였다 (고민정과 한준태, 2010).

본 논문에서는 데이터마이닝을 이용하여 위암으로 진단 및 치료 받을 가능성이 큰 고위험 대상자 예측모형을 제안하고자 한다.

2. 고위험군 예측모형 개발

본 논문에서 개발하고자 하는 위암 고위험군 예측모형은 국민건강보험 가입자에서 암 검진일 이전에 암으로 진료 받은 경험이 있거나 검진결과 암 판정된 대상자를 제외하고, 검진 받은 일부터 2년 이내에 위암이 발견될 가능성이 큰 고위험군을 찾는 모형이다. 자료는 국민건강보험공단의 원천시스템 (operational data store) 및 데이터웨어하우스 (data warehouse)에서 2000년부터 2004년까지의 특정암 검진 및 문진자료, 1·2차 건강검진 및 문진자료, 현물급여 자료의 각 연도별 개인급여정보 (2005년 7월 지급기준), 상병정보 및 수검자의 자격정보 (수검월말 자격)를 이용하였다. 또한, 정확한 암발견 판정 기준을 위해, 통계청의 사망원인 자료를 연계하여 활용하였다. 단, 연구대상은 우리나라의 건강검진 대상자 선정기준에 근거하여 2000년부터 2004년 기간에 특정암 검진을 받은 국민건강보험가입자로 제한하였다. 분석 패키지는 SAS Enterprise Miner 4.1을 사용하였다.

2.1. 분석모형

국제암연구소와 미국 국립암협회지에서 발표한 암 발생의 위험요인과 국립암센터에서 권고하고 있는 위험요인들을 고려하여, 생활습관 (음주, 흡연, 운동, 식생활, 비만), 개인 과거병력, 가족력 그리고 건강검진 결과 등의 위험요인을 반영하였다. 또한 수검자의 인구사회학적 특성인 성별, 연령, 거주지 구분 (대도시, 중소도시, 소도시), 국민건강보험 가입자 자격 (직장가입자, 지역가입자)과 소득수준의 대리변수인 보험료를 설명변수로 포함하였다.

위암 발견 예측모형의 종속변수인 위암 발견 판정기준은 암검진 받기 전 암진료 경험이 없는 자가 2년 이내 '암검진 및 진료를 받은 자' 또는 '사망원인이 위암인 자' 또는 '암검진시 치료대상으로 판정된 자' 또는 '위암으로 입원한 경험이 1회 이상 있는 자 (단, 입원진료가 1회인 경우, 입원 후 외래진료 2회 이상 or 투약 2회 이상)'로 정의 하였다.

2.2. 예측모형 개발 프로세스

국민건강보험 가입자 중 암 건강검진을 받은 자를 중심으로 데이터마이닝을 이용하여 위암 발생 예측

모형을 개발하였고 분석 방법은 로지스틱 회귀분석을 적용하였다.

예측모형은 성별에 따른 암발생율과 치명율에 있어서 남녀 차이가 있음을 고려하여 (유근영과 신해림, 2003; 김정순, 2004), 전체 모형을 포함한 남녀 각각의 예측모형을 개발하였다.

분석데이터는 크게 분석용 (training data), 평가용 (validation data), 검정용 (test data)으로 구분하였고, 분석용과 평가용 데이터는 2000년, 2001년 암 건강검진 대상자를 기준으로 7 대 3의 비율로 분할하여 생성하였다. 검정용 자료는 2002년 암 검진 대상자를 기준으로 분석용 데이터와 동일한 기준으로 구축하였고, 이를 이용하여 개발된 예측모형의 일반화 검정을 통해 모형의 안정성을 평가하였다 (그림 2.1 참조).

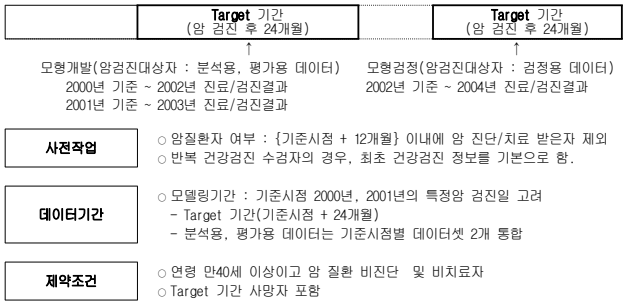


그림 2.1 고위험군 예측모형 개발 프로세스

예측모형 개발 프로세스의 첫 번째 단계로 위암 발생 고위험군 예측모형 개발에서 위암 발생의 유형을 정하고 분석대상, 분석기간, 분석주제의 정의, 평가 그리고 예측기간을 정의하였다. 두 번째 단계는 예측모형의 정확성을 높이기 위해 데이터 탐색 및 데이터 정제 작업을 수행하고, 분석용 데이터를 구축하였다. 세 번째 단계에서는 구축된 분석용 데이터를 활용하여 모델링 작업을 수행하였다. 로지스틱 회귀분석을 통해 모델링을 수행하였다. 분석용 데이터와 평가용 데이터를 통해 모형을 생성하고 생성된 모형을 검정용 데이터에 적용했을 경우 모형 구축에서 발생할 수 있는 과적합 문제를 해결하였다. 또한 다양한 알고리즘을 통해 만들어진 모형들 중에서 가장 좋은 모형을 평가하고 선정하기 위해 향상도 도표와 ROC (Receiver Operating Characteristic)분석과 정확도를 계산하였다.

2.3. 위암 발견 주요요인의 일반적 현황

2004년 위암검진 (연령 만40세 이상)을 받은 건강보험가입자들의 위암판정결과에 대해 국민건강보험 자격, 보험료, 진료량, 1차, 2차 검진·문진 및 암검진·문진에 나타난 인구사회학적 특성, 건강행위 특성, 건강위험요인 등을 토대로 현황분석을 실시하였다.

2004년 기준 국민건강보험 가입자의 인구사회학적 특성을 중심으로 위암판정결과를 살펴보면, 총 1,150,829명의 위암수검자들 중 정상으로 판명된 자가 541,240명 (47.03%), 재검대상인 자가 38,734명 (3.37%), 위암치료대상인 자가 1,366명 (0.12%) 그리고 기타질환자가 569,849명 (49.49%)으로 나타났다. 연령대별로는 위암검진결과 정상인 사람들 중 40대가 211,466명 (39.07%)로 가장 많았고, 연령이 증가할수록 감소하였다. 위암 재검대상 및 기타질환으로 판명된 경우 또한, 연령이 증가할수록 해당판정비율이 감소하는 경향을 보였다. 위암치료대상의 경우엔 다른 판정결과와는 달리 연령이 증가할수록 위암치료대상자가 증가하였다. 특히 위암치료대상의 61.79%가 60대 이상이었다. 성별로 위암판정결과를 살펴보면, 여자 (60.05%)가 남자 (39.95%)보다 정상인 경우가 많았으나, 재검대상 및 위암치료대상

인 경우에는 모두 남자의 비율이 여자의 비율보다 높은 것으로 나타났다. 거주지역별로 위암판정결과를 살펴보면, 위암치료대상자 중 18.89%가 농어촌에 살고 있었으며, 정상인 사람들의 46.38%가 대도시에 거주하고 있는 것으로 나타났다. 직역별로는 위암치료대상자의 13.25%가 지역가입자로 다른 판정결과들의 지역가입자 비율보다 상대적으로 높게 나타났으며, 재검대상인 경우는 다른 판정결과에 비해 공교가입자가 차지하는 비율이 54.17%로 상대적으로 가장 높았다. 보험료 등급이 높을수록 위암수검비율이 낮아졌으며, 의료이용량 부분에서는 위암치료대상자들 중 입원한 비율이 19.40%로 정상 (2.57%), 재검대상 (2.92%), 기타질환 (2.73%)보다 상대적으로 높게 나타났다. 가족력에서는 위암치료대상자인 경우 2.63%가 간장질환에 대한 가족병력이 있었으며, 고혈압은 7.39%, 뇌졸중은 3.61%, 심장병은 0.66%, 당뇨병은 5.25%, 암은 13.88%의 사람들이 해당 가족병력이 있으며, 음주를 거의 하지 않을 경우 정상판정자들의 비율은 62.73%로 다른 판정결과들의 해당판정비율보다 상대적으로 높게 나타났으며, 위암치료대상자들이 다른 판정결과에 속하는 그룹들보다 상대적으로 운동을 하지 않는 비율이 높은 것으로 나타났다 (표 2.1).

표 2.1 국민건강보험 가입자의 일반적 특성 (2004년: 위암수검자)

특성		정상		재검대상		암치료대상		기타질환		전체	
		인원	(%)	인원	(%)	인원	(%)	인원	(%)	인원	(%)
연령	40대	211,466	(39.07)	14,356	(37.06)	174	(12.74)	208,810	(36.67)	434,806	(37.78)
	50대	165,943	(30.66)	12,450	(32.14)	348	(25.48)	186,361	(32.72)	365,102	(31.73)
	60대이상	163,831	(30.27)	11,928	(30.79)	844	(61.79)	174,318	(30.61)	350,921	(30.49)
성별	남	216,221	(39.95)	20,854	(53.84)	974	(71.30)	282,940	(49.68)	520,989	(45.27)
	여	325,019	(60.05)	17,880	(46.16)	392	(28.70)	286,549	(50.32)	629,840	(54.73)
거주지역	대도시	251,024	(46.38)	21,107	(54.49)	552	(40.41)	260,789	(45.79)	533,472	(46.36)
	중소도시	208,366	(38.50)	14,829	(38.28)	556	(40.70)	234,300	(41.14)	458,051	(39.80)
	농어촌	81,850	(15.12)	2,798	(7.22)	258	(18.89)	74,400	(13.06)	159,306	(13.84)
직역	지역	63,094	(11.66)	4,664	(12.04)	181	(13.25)	72,053	(12.65)	139,992	(12.16)
	직장	206,798	(38.21)	13,088	(33.79)	472	(34.55)	205,880	(36.15)	426,238	(37.04)
	공교	271,348	(50.13)	20,982	(54.17)	713	(52.20)	291,556	(51.20)	584,599	(50.80)
입원유무	무	527,337	(97.43)	37,602	(97.08)	1,101	(80.60)	553,372	(97.17)	1,119,412	(97.27)
	유	13,903	(2.57)	1,132	(2.92)	265	(19.40)	16,117	(2.83)	31,417	(2.73)
간장질환	무	454,877	(89.33)	31,573	(86.67)	1,092	(89.66)	477,246	(89.46)	964,788	(89.30)
	유	13,973	(2.74)	1,684	(4.62)	32	(2.63)	14,753	(2.77)	30,442	(2.82)
	무응답	40,357	(7.93)	3,172	(8.71)	94	(7.72)	41,488	(7.78)	85,111	(7.88)
고혈압	무	427,347	(83.92)	29,485	(80.94)	1,037	(85.14)	448,904	(84.15)	906,773	(83.93)
	유	43,694	(8.58)	4,001	(10.98)	90	(7.39)	45,707	(8.57)	93,492	(8.65)
	무응답	38,166	(7.50)	2,943	(8.08)	91	(7.47)	38,876	(7.29)	80,076	(7.41)
뇌졸중	무	443,716	(87.14)	30,606	(84.02)	1,085	(89.08)	465,482	(87.25)	940,889	(87.09)
	유	26,141	(5.13)	2,766	(7.59)	44	(3.61)	27,925	(5.23)	56,876	(5.26)
	무응답	39,350	(7.73)	3,057	(8.39)	89	(7.31)	40,080	(7.51)	82,576	(7.64)
심장병	무	456,763	(89.70)	31,696	(87.01)	1,116	(91.63)	479,658	(89.91)	969,233	(89.72)
	유	12,232	(2.40)	1,568	(4.30)	8	(0.66)	12,466	(2.34)	26,274	(2.43)
	무응답	40,212	(7.90)	3,165	(8.69)	94	(7.72)	41,363	(7.75)	84,834	(7.85)
당뇨병	무	437,642	(85.95)	30,210	(82.93)	1,064	(87.36)	460,084	(86.24)	929,000	(85.99)
	유	32,726	(6.43)	3,214	(8.82)	64	(5.25)	33,826	(6.34)	69,830	(6.46)
	무응답	38,839	(7.63)	3,005	(8.25)	90	(7.39)	39,577	(7.42)	81,511	(7.54)
암	무	402,503	(79.05)	27,465	(75.39)	967	(79.39)	419,015	(78.54)	849,950	(78.67)
	유	70,435	(13.83)	6,172	(16.94)	169	(13.88)	78,195	(14.66)	154,971	(14.34)
	무응답	36,269	(7.12)	2,792	(7.66)	82	(6.73)	36,277	(6.80)	75,420	(6.98)
음주빈도	거의안마심	319,408	(62.73)	20,028	(54.98)	674	(55.34)	312,133	(58.51)	652,243	(60.37)
	월2 3회	68,580	(13.47)	5,361	(14.72)	124	(10.18)	74,800	(14.02)	148,865	(13.78)
	주 1 2회	65,227	(12.81)	5,868	(16.11)	174	(14.29)	78,006	(14.62)	149,275	(13.82)
	주 3 4회	28,669	(5.63)	2,536	(6.96)	121	(9.93)	37,363	(7.00)	68,689	(6.36)
	거의 매일	19,754	(3.88)	1,724	(4.73)	113	(9.28)	24,825	(4.65)	46,416	(4.30)
	무응답	7,569	(1.49)	912	(2.50)	12	(0.99)	6,360	(1.19)	14,853	(1.37)

표 2.1 (계속)국민건강보험 가입자의 일반적 특성 (2004년: 위암수검자) (단위: 명, %)

특성		정상		재검대상		암치료대상		기타질환		전체	
		인원	(%)	인원	(%)	인원	(%)	인원	(%)	인원	(%)
운동실태	안한다	281,241	(55.23)	17,475	(47.97)	712	(58.46)	290,845	(54.52)	590,273	(54.64)
	1 2회	109,825	(21.57)	9,891	(27.15)	248	(20.36)	120,773	(22.64)	240,737	(22.28)
	3 4회	51,386	(10.09)	4,314	(11.84)	90	(7.39)	57,732	(10.82)	113,522	(10.51)
	5 6회	14,237	(2.80)	1,142	(3.13)	23	(1.89)	15,756	(2.95)	31,158	(2.88)
	거의 매일	39,568	(7.77)	2,554	(7.01)	129	(10.59)	39,970	(7.49)	82,221	(7.61)
	무응답	455,402	(89.43)	32,733	(89.85)	1,066	(87.52)	477,761	(89.55)	966,962	(89.51)
전체		541,240	(100.00)	38,734	(100.00)	1,366	(100.00)	569,489	(100.00)	1,150,829	(100.00)

주) 2004년 위암중복수검자 제외

3. 연구결과

전체, 남자, 여자의 추정된 위암 발생 로지스틱 회귀모형은 아래와 같다.

$$\begin{aligned}\widehat{y}_{\text{전체}} =& -5.682 + 0.0088X_{\text{남자}} - 0.0209X_{\text{50대}} + 1.0158X_{\text{60대이상}} \\ &+ 0.1716X_{\text{중소도시}} + 0.0096X_{\text{대도시}} - 0.1178X_{\text{지역가입자}} - 0.0783X_{\text{5만원이하}} \\ &+ 0.0134X_{\text{5만원 - 7만5천원}} - 0.0416X_{\text{7만5천원 - 10만원}} + 0.1304X_{\text{운동안함}} \\ &+ 0.8338X_{\text{소주한병이상}} + 0.0825X_{\text{육식}} + 0.1854X_{\text{암}} \end{aligned} \tag{3.1}$$

$$\begin{aligned}\widehat{y}_{\text{남자}} =& -5.3899 - 0.0196X_{\text{50대}} + 1.1375X_{\text{60대이상}} \\ &+ 0.0856X_{\text{중소도시}} + 0.0098X_{\text{대도시}} - 0.0535X_{\text{지역가입자}} - 0.0313X_{\text{5만원이하}} \\ &- 0.0088X_{\text{5만원 - 7만5천원}} - 0.1415X_{\text{7만5천원 - 10만원}} + 0.2623X_{\text{운동안함}} \\ &+ 0.7753X_{\text{소주한병이상}} + 0.2793X_{\text{암}} + 0.2472X_{\text{위수술}} \end{aligned} \tag{3.2}$$

$$\begin{aligned}\widehat{y}_{\text{여자}} =& -5.0036 - 0.0872X_{\text{50대}} + 0.7703X_{\text{60대이상}} \\ &+ 0.0233X_{\text{중소도시}} + 0.0555X_{\text{대도시}} + 0.0525X_{\text{지역가입자}} - 0.6681X_{\text{5만원이하}} \\ &+ 0.0962X_{\text{5만원 - 7만5천원}} + 0.4073X_{\text{7만5천원 - 10만원}} + 0.0479X_{\text{운동안함}} \\ &+ 1.3565X_{\text{소주한병이상}} + 0.0747X_{\text{비만}} + 0.2447X_{\text{암}} + 0.0894X_{\text{위십이지궤양}} \end{aligned} \tag{3.3}$$

위암 발생 로지스틱 회귀예측모형에 대한 평가는 평가용과 검정용 데이터에서 ROC 곡선과 향상도에 근거하였다. 그림3.1에서와 같이, ROC 곡선의 밑면적을 나타내는 C-통계량이 검정용 데이터에서의 개발된 로지스틱 예측모형의 C-통계량은 전체 모형에서 0.735, 남자와 여자 예측모형에서는 각각 0.730과 0.673의 값을 보여 임의의 모형 (random model)이 가지는 C-통계량 0.5보다 크게 나타나 상대적으로 모형이 우수함을 나타내 주었다 (그림 3.1).

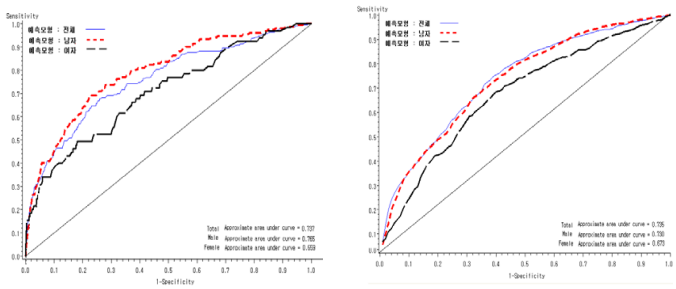
다음 누적 향상도 도표는 추정된 사후확률의 분위수에 따른 반응률 (%response)을 도표화 한 것으로, 위암 발견 로지스틱 회귀 예측모형은 평가용과 검정용 데이터 모두에서 상위 분위수에 대응되는 리프트가 더 큰 값을 보여, 개발된 예측모형의 안정성과 효과성을 확인하였다. 제시된 표3.1의 검정용 데이터에서 상위 10% 수준의 리프트값을 보면 전체 예측모형에서는 4.79, 남자의 경우는 4.91 그리고 여자의 경우는 5.31 값을 나타내, 위암 발견 고위험군 예측모형을 활용하여 상위 10%에서 관리할 경우 전체 위암 수검자 대비 각각 4.79배, 4.91배 그리고 5.31배의 효율을 남녀별로 각각 기대할 수 있을 것이다.

표 3.2에서와 같이 위암발견 요인별 분석을 살펴보면 다음과 같다. 그 결과 전체모형에서 수검자의 성별은 여자보다 남자의 경우가 위암 발견 가능성이 상대적으로 높은 것으로 나타났다. 연령대로는 40대보다 60대 이상의 경우가 상대적으로 위암발견 가능성이 7.17배 높았고, 그 중 남자의 경우는

표 3.1 Score대별 위암 발견 고위험군 예측모형의 적중률

표본	등급	전체				남자				여자			
		건수	CRes (%)	Res (%)	Lift	건수	CRes (%)	Res (%)	Lift	건수	CRes (%)	Res (%)	Lift
분석용	10%	12,214	48.01	1.08	4.79	7,176	48.71	1.32	4.91	4,988	53.14	1.00	5.31
	20%	24,368	64.62	0.73	3.23	14,361	63.07	0.85	3.17	9,978	61.70	0.58	3.08
	30%	36,626	74.72	0.56	2.48	21,747	77.94	0.69	2.59	14,881	63.83	0.40	2.13
	50%	61,075	87.00	0.39	1.73	35,890	87.17	0.47	1.75	24,930	77.67	0.29	1.56
	100%	122,043	100.0	0.23	1.00	72,274	100.0	0.26	1.00	49,769	100.0	0.14	1.00
평가용	10%	5,201	51.20	1.23	5.15	3,072	52.94	1.46	5.33	2,151	39.28	0.51	3.89
	20%	10,448	61.61	0.74	3.08	6,185	61.18	0.84	3.06	4,274	2.32	0.30	2.32
	30%	15,745	71.22	0.57	2.37	9,316	75.29	0.69	2.05	6,400	1.54	0.20	1.55
	50%	26,144	82.43	0.39	1.65	15,587	89.41	0.49	1.78	10,662	1.35	0.18	1.36
	100%	52,307	100.0	0.23	1.00	30,978	100.0	0.27	1.00	21,329	1.00	0.13	1.00
검정용	10%	58,336	35.55	1.12	3.58	26,340	35.50	1.72	3.55	32,365	24.42	0.42	2.44
	20%	117,328	49.48	0.77	2.48	52,371	48.75	1.19	2.45	64,363	42.42	0.37	2.13
	30%	176,532	62.93	0.65	2.09	76,577	60.42	1.01	2.08	100,017	58.47	0.33	1.89
	50%	294,785	82.47	0.51	1.64	130,501	81.35	0.80	1.64	161,481	74.69	0.26	1.50
	100%	586,815	100.0	0.31	1.00	263,264	100.0	0.48	1.00	323,551	100.0	0.17	1.00

주: 1) Res (%) = (해당 Group에서 위암 발견자의 수 / 해당 Group에서 전체 수검자의 수)×100
2) CRes (%) = (해당 Group에서 위암 발견자의 수/ 전체 위암 발견자의 수)×100
3) Base Res (%)는 전체 위암발견자의 구성비
4) Lift = 해당 Group에서 위암 발견율/ 전체 위암 발견율
5) Res (%): Response (%), CRes (%): Captured Response (%)



평가용 데이터 검정용 데이터

그림 3.1 위암 발견 로지스틱 회귀모형의 ROC 곡선

9.54배, 여자의 경우는 4.27배 높은 것으로 나타났다. 다만, 기타 거주지역, 가입자의 직역 그리고 보험료 수준은 위암 발견에 통계학적으로 영향력이 유의하지 않은 것으로 나타났다.

개인의 건강행위 위험요인으로 고려된 운동습관은 운동을 전혀 하지 않는 경우가 그렇지 않은 경우보다 상대적으로 위암발견 위험도가 1.29배 높았고, 남자는 1.69배 높은 것으로 나타났다. 또한 음주습관의 경우는 소주를 1병 이상, 주 3회 이상 마시는 경우는 그렇지 않은 경우보다 상대적으로 위암발견 위험도가 5.3배 높았으며, 남자의 경우는 4.71배 그리고 여자의 경우는 무려 15.7배 높게 나타났다. 한편, 가족 중 암병력이 있을 경우 그렇지 않은 경우보다 상대적으로 위암발견 위험도가 1.45배 높았는데, 그 중 남자의 경우는 1.74배 높은 것으로 나타났다.

그러나, 개인의 과거 병력 중에서 위·십이지장궤양이나 위 수술을 받은 경험이 있는 경우는 위암 발견에 있어 양의 관계를 갖고 있으나, 통계학적으로 유의하지 않은 것으로 나타났다.

위암 건강검진 정보를 기초로 최종 결정된 위암발견 요인은 표3.2에 제시된 바와 같이 연령, 음주, 가

표 3.2 위암 발견 로지스틱 회귀모형 결과

특성요인			모형: 전체			모형: 남자			모형: 여자		
			추정계수		OR	추정계수		OR	추정계수		OR
절편			-5.6820	***		-5.3899	***		-5.0036	***	
인구사회학적특성	성별 (Ref: 여자)	남자	0.0088	**	1.01						
	연령 (Ref: 40대)	60대 이상	1.0158	***	7.17	1.1375	***	9.54	0.7703	***	4.27
		50대	-0.0209		2.65	-0.0196		2.99	-0.0872		1.81
	거주지역 (Ref: 농어촌)	대도시	0.0096		2.21	0.0098	*	1.11	0.0555		1.14
		중소도시	0.1716		1.42	0.0856		1.14	0.0233		1.11
국민건강보험	직역구분 (Ref: 직장)	지역 가입자	-0.1178		0.79	-0.0535	*	0.89	0.0525		1.11
가입자특성	보험료 (Ref: 10만원 이상)	5만원 이하	-0.0783		0.83	0.0313		0.92	-0.6681	**	0.43
		5만원 7만5천원	0.0134		0.91	-0.0088		0.88	0.0962		0.93
		7만5천원 10만원	-0.0416		0.86	-0.1415		0.77	0.4073	*	1.27
생활습관 특성	운동 (Ref: 운동함)	전혀 안함	0.1304	*	1.29	0.2623	***	1.69	0.0479		1.10
	음주 (Ref: 기타)	소주 1병 이상 (주 3회 이상)	0.8338	***	5.30	0.7753	***	4.71	1.3565	***	15.07
	식생활 (Ref: 채식&육식)	주로 육식	0.0825		1.18						
	비만도 (Ref: 25미만)	비만 (25이상)							0.0747		1.16
가족병력	암 (Ref: 없음)	있다	0.1854	**	1.45	0.2793	***	1.74	0.2448		1.63
개인 과거병력	위수술 (Ref: 없음)	있다				0.2472		1.64			
	위·십이지궤양 (Ref: 없음)	있다							0.0894		1.14

주) *: 유의확률 p <0.1, **: p <0.05, ***: p <0.01, Ref: 기준변수 (Reference)

족병력 중 암, 운동, 직역구분, 거주지역, 보험료, 식생활, 성별을 최종 위암 발견 가능 특성으로 선별하여 로지스틱 회귀모형에 반영하였다. 단, 모든 유효 변수는 목표 (target) 변수와의 발견율을 고려하여 범주화 시켜 더미 (dummy) 변수로 모형에 적합하였다.

이러한 변수 중 전체 위암 발견 고위험군 예측에 영향을 미치는 특성을 로지스틱 회귀모형결과의 Wald χ^2통계량을 통해 살펴본 결과, 위암 발견 예측에 가장 큰 영향을 미치는 특성은 수검자의 연령이었고, 다음으로 음주, 가족병력 (암) 순으로 나타났다. 그러나 여성의 경우는 위암발견에 음주유무가 가장 많은 영향을 미치는 것으로 나타났다 (표 3.3).

표 3.3 위암 발견의 결정요인

유효변수	자유도	전체		남		여		p 값		
		순위	Wald χ^2-통계량	순위	Wald χ^2-통계량	순위	Wald χ^2-통계량	전체	남	여
연령	2	1	130.01	1	80.68	2	16.71	<.0001	<.0001	<.0001
음주	1	2	102.09	2	60.60	1	53.67	<.0001	<.0001	<.0001
가족병력 (암)	1	3	5.42	4	6.84	4	2.40	0.0199	0.0089	0.1210
운동	1	4	3.38	3	7.33	7	0.10	0.0660	0.0068	0.7493
직역구분	1	5	2.43	7	0.28	8	0.10	0.1185	0.5990	0.7494
거주지역	2	6	2.14	8	0.28	9	0.05	0.3414	0.8712	0.9739
보험료	3	7	1.10	5	0.95	3	6.40	0.7757	0.8138	0.0936
식생활	1	8	0.48	-	-		-	0.4872	-	-
성별	1	9	0.01	-	-		-	0.9228	-	-
과거병력 (위수술)	1	-	-	6	0.69		-	-	0.4053	-
비만도 (BMI)	1	-	-	-	-	5	0.26	-	-	0.6113
과거병력 (위·십이지장궤양)	1	-	-	-	-	6	0.23	-	0.4053	0.6285

4. 결론

위암의 위험요인들이 상대적으로 위암 발견에 미치고 있는 영향력을 수검자의 성별, 연령, 거주지역, 국민건강보험 가입자의 직역 그리고 보험료를 고려한 데이터마이닝 방법론에 의한 로지스틱 회귀모형을 활용하여 위험요인들의 상대적 위험도를 살펴보았다.

그 결과, 남자가 여자보다 위암 발견 가능성이 상대적으로 1.01배로 다소 높은 것으로 나타났으며, 위암 발견 예측에 가장 큰 영향을 미치는 특성은 남성의 경우는 수검자의 연령, 여성의 경우는 음주유무가 가장 많은 영향을 미치는 것을 확인 할 수 있었다. 따라서 제안된 위암 예측모형의 결과 성별에 따라 위암 발생 결정요인이 차이가 있으므로, 대상 집단을 세분화하고 대상자의 선별적 관리가 필요가 필요하다.

본 연구에서 국민건강보험공단의 건강검진을 받은 수검자를 중심으로 예측모형을 개발했고, 현재 국민건강보험 가입자의 연령, 성별, 거주 지역별, 직역별에 따라 수검율이 큰 차이를 보이고 있어 제안된 예측 모형이 모든 국민들에게 적용되는 일반화된 예측모형으로 사용함에 있어서는 한계점을 가지고 있다. 그러나 본 연구는 이러한 한계에도 불구하고 우리나라 전국민 자료가 구축된 국민건강보험공단의 방대한 데이터베이스와 최신 정보기술인 데이터마이닝을 활용함으로써 위암뿐만 아니라 향후 우리나라 국민을 위한 맞춤형 건강정보 제공, 관리대상자의 효율적 선정 및 관리서비스 제공 등의 적극적인 건강관리사업으로 발전시킬 수 있는 정보기술의 인프라 구축이라는 측면에서 중요한 의의를 갖는다고 할 수 있겠다.

참고문헌

강성홍, 최순호 (2001). 데이터마이닝을 이용한 보건소의 건강증진사업의 효율화 방안. <대한의료정보학회지>, **7**, 37-48.

고민정, 한준태 (2010). 주요 위험요인별 허혈성심질환 사망위험도 분석. <한국데이터정보과학회지>, **21**, 201-209.

김정순 (2004). <역학원론>, 신광출판사, 서울.

박일수, 용왕식, 김유미, 강성홍, 한준태 (2008). 데이터마이닝 기법을 활용한 맞춤형 고혈압 사후관리 모형 개발. <응용통계연구>, **21**, 639-647.

용왕식, 박일수, 강성홍, 김원중, 김공현, 김광기, 박노래 (2006). 고혈압 발생 예측 모형 개발. <보건교육·건강증진학회지>, **22**, 13-28.

유근영, 신해림 (2003). 암의 위험요인과 예방. <한국역학회지>, **25**, 1 -15.

이애경, 이상이, 박일수, 김수영, 윤태호, 정백근 (2006). 대장암 발생 고위험군의 예측모형 개발과 활용. <예방의학회지>, **39**, 438-446.

D'Agostino, Sr R. B., Grundy, S., Sullivan, L. M. and Wilson, P. (2001). Validation of the Framingham coronary heart disease prediction scores. *Journal of the American Medical Association*, **286**, 180-187.

Liu, J., Hong, Y., D'Agostino, Sr R. B., Wu, Z., Wang, W., Sun, J., Wilson, P. W. F., Kannel, W. B. and Zhao D. (2004). Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese multi-provincial cohort study. *Journal of the American Medical Association*. **291**, 2591-2599.

The World Health Organization's Fight Against Cancer (2007), WHO.

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1253–1261

한국데이터정보과학회지

Developing the predictive model for stomach cancer using data mining

Il Su Park[1] · Jun Tae Han[2] · Suk Bok Kang[3] · Jae Hoon Ji[4]

[12]Health Insurance Policy Research Institute, National Health Insurance Corporation
[3]Department of Statistics, Yeungnam University
[4]Center for Health Care Strategic Management, Inje University

Received 1 October 2010, revised 18 November 2010, accepted 23 November 2010

Abstract

We develope the predictive model for the incidence of the stomach cancer by utilizing the health screening data of the National Health Insurance in Korea. We also explore the characteristics for the stomach cancer. We perform the logistic regression analysis using the data mining methodology and use SAS Enterprise Miner 4.1. This study shows that there exists a higher rate of the stomach cancer for males than females. Our study confirms that the major influencing factors for the incidence of the stomach cancer are age, drinking and a family history of cancer, lack of exercise. For man, the age is the most important determinant of the stomach cancer incidence, whereas the drinking is the most important determinant of the stomach cancer incidence for women.

Keywords: Logistic regression, predictive model, risk factor, stomach cancer.

[1] Senior researcher, National Health Insurance Policy Research Institute, National Health Insurance Corporation, Seoul 121-749, Korea.
[2] Senior researcher, National Health Insurance Policy Research Institute, National Health Insurance Corporation, Seoul 121-749, Korea.
[3] Corresponding author: Professor, Department of Statistics, Yeungnam University, Gyeongsan 712-749, Korea. E-mail: sbkang@yu.ac.kr
[4] Research professor, Center for Health Care Strategic Management, Inje University, Busan 633-165, Korea.