빅데이터 분석 방법(Shrinkage Methods)을 이용한 한국인의 골다공증에 연관된 질병 현황

연구책임자	연구실무자	분석지원
최제용 교수	신승연	신서희 주임연구원
경북대학교 의과대학	경북대학교 의과대학	건강보험심사평가원

※ '2017년 HIRA 빅데이터 분석 협업과제' 사례

1. 분석 배경

- 회귀분석(Regression Analysis)은 설명변수를 설정 · 활용하여 종속변수를 설명하는 분석 방법으로 보건의료 분야에서도 널리 쓰이는 분석 방법임
  - ▸ 하지만 설명변수가 많은 경우 불필요한 부분까지 설명하는 과적합 문제가 발생할 수 있음
  - ▸ 축소방법(Shrinkage Methods)을 이용한 회귀분석은 과적합 문제를 해결할 수 있는 장점이 있어 설명변수가 많은 경우 주로 활용함
  - ▸ 많은 질환들을 설명변수로 설정하여 연관 질환을 찾는 경우, 축소방법을 이용한 회귀분석을 실시한다면 유의미한 결과가 도출될 것으로 예상됨
- 골다공증은 큰 증상없이 진행되어 침묵의 병이라 불리는 만큼 진단이 어려워, 빠른 진단을 위해서는 연관 질환에 대한 분석이 필요함
  - ▸ 축소방법을 활용하여 골다공증에 연관된 질환들을 연구하여 질병 네트워크를 구축한다면, 골다공증의 빠른 진단과 예방에 도움을 줄 수 있을 것으로 기대
  - ▸ 본 연구의 결과를 PDN(Phenotype Disease Network)과 골대사학 기반 Gene Ontology Network 연구의 기초자료로 활용

2. 분석 방법과 내용

- 대상: 2013년~2015년 골다공증 진료환자

- 자료원: 2012년~2016년 건강보험 청구자료
- 분석방법
  - ▸ 기간 중 골다공증 환자 1,000명이 추출될 때까지 골다공증을 진단받지 않은 환자를 표본 추출한 후 클렌징 작업
  - ▸ 설명변수 개수가 많을 때 유용한 축소방법(Shrinkage Method)을 이용한 회귀분석을 통해 골다공증에 영향을 미치는 질병군 선택

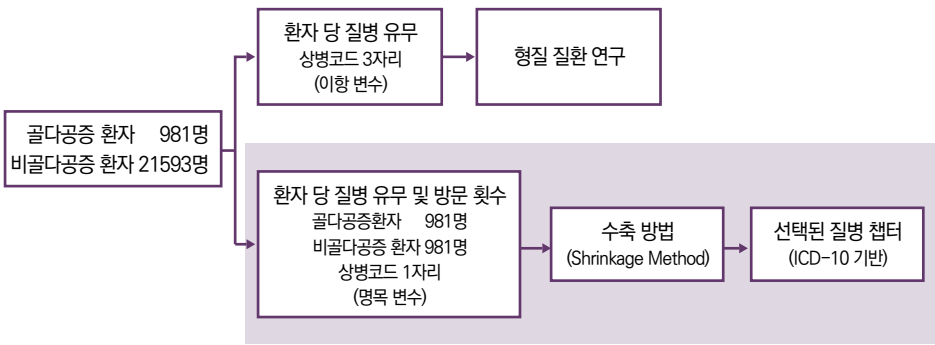
    *종속변수: 골다공증 발생 여부, 설명변수: 질병 유무 및 진료 횟수

    *질병의 개수가 많으므로, 과적합(Over-fitting)을 피하고 최적의 질병 개수 조합을 도출하고자 축소방법을 이용한 회귀분석 수행

    *축소모수 λ는 5 fold-Cross Validation에 의해 결정

[표 1] 축소방법(Shrinkage Method)를 이용한 회귀분석

회귀분석	축소방법(Shrinkage Method)를 이용한 회귀분석
● 설명변수 수가 많을 경우 과적합※ 문제 발생 ※ 과적합(Over-fitting): 자료의 불필요한 부분(noise)까지 설명하여 부정확한 결과를 도출	● 불필요한 변수들을 제거하여 과적합을 피할 수 있으며, 설명변수가 수백개일 때도 활용 가능하다는 장점이 있음
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ $\min imize \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=q}^{p}\beta_j x_{ij})^2$	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$ $\min imizse \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=q}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$
● RSS(Residual Sum of Square)을 최소화하는 β추정	● RSS에 Penalty 조건()을 추가하여 "RSS + Penalty"를 최소화하는 β추정 ● Penalty는 불필요한 변수들의 추정값을 0에 가깝게 만들어 제거하는 역할 ● 축소(Shrinkage) 정도는 λ에 의해 조절되며, λ가 커질수록 제거하는 변수 많아짐 * $\lambda \to 0$ (축소량 최소): 모든 변수를 추정 * $\lambda \to \infty$ (축소량 최대): 상수항(β_0)만 남음



[그림 1] 연구를 위한 작업흐름도(음영처리 부분이 주된 내용임)

3. 분석 결과

- 골다공증 환자 981명 중 남성은 69명, 여성은 912명으로 여성의 비율이 크게 높았으며, 평균 연령이 60세 이상으로 고령층에서 많이 발생함

[표 2] 골다공증과 비골다공증 환자수와 평균 연령

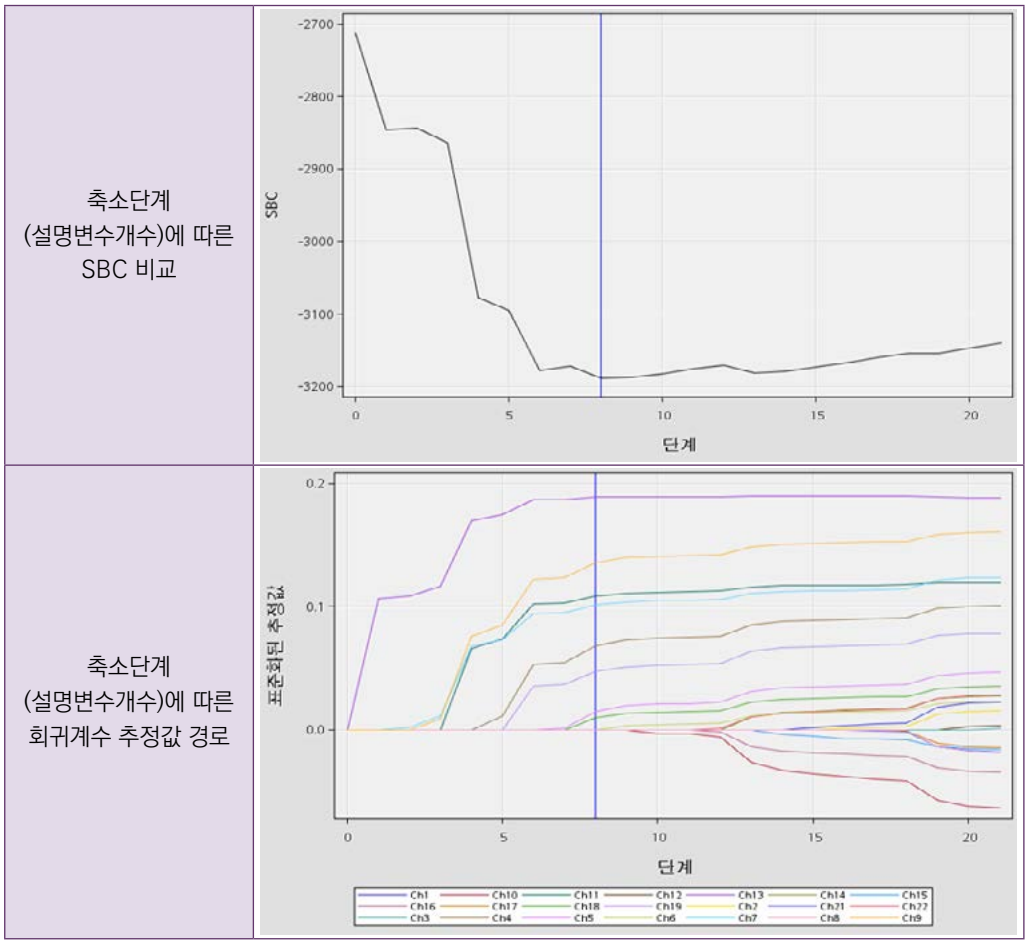
	성별	환자수	연령(평균±분산)
골다공증	남	69	68.75±10.42
	여	912	64.71±10.63
비골다공증	남	11,668	35.03±20.24
	여	9,925	33.20±20.17

- ICD-10 기준으로 질환을 22개의 범주로 구분하였으며[표 3], 이들 질환 유무를 설명변수로 골다공증 발생 여부를 종속변수로 하여 축소방법을 이용한 회귀분석을 실시함

[표 3] ICD-10 기준 질환 분류

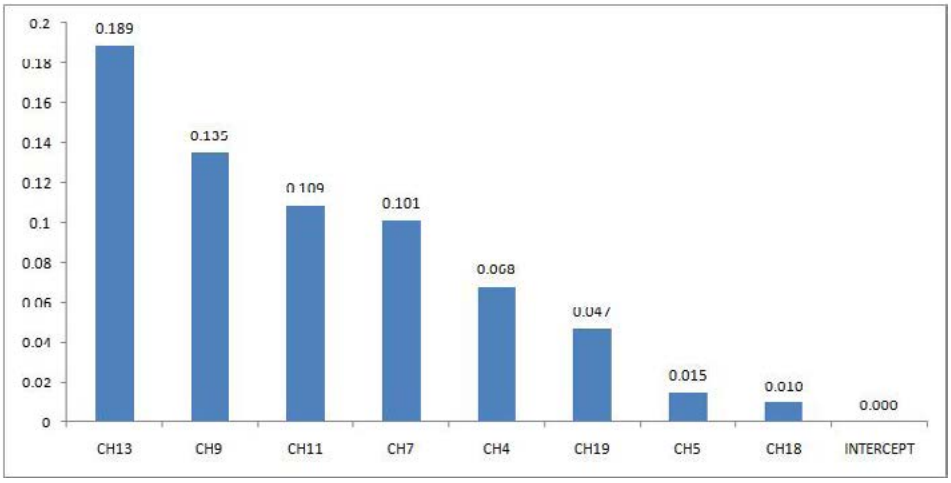
구분	ICD-10코드	질환명
Ch1	A00-B99	특정 전염병 및 기생충병
Ch2	C00-D48	종양
Ch3	D50-D89	혈액과 조혈 기관의 질병 및 면역 체계 관련 특정 장애
Ch4	E00-E90	내분비, 영양 및 대사 질병
Ch5	F00-F99	정신 및 행동 장애
Ch6	G00-G99	신경 계통의 질병
Ch7	H00-H59	눈과 그에 딸린 기관의 질병
Ch8	H60-H95	귀와 유양돌기
Ch9	I00-I99	순환 계통의 질병
Ch10	J00-J99	호흡기 계통의 질병
Ch11	K00-K93	소화 계통의 질병
Ch12	L00-L99	피부와 피하 조직의 질병
Ch13	M00-M99	근육과 연결 조직의 질병
Ch14	N00-N99	생식, 배설 계통의 질병
Ch15	O00-O99	임신, 출산, 산후 조리
Ch16	P00-P96	출산 전후 기간에 일어나는 어떤 상태
Ch17	Q00-Q99	선천성 기형, 변형, 염색체의 이상
Ch18	R00-R99	다른 곳에서 분류되지 않은 증상, 증세나 임상 또는 연구에서 발견한 비정상
Ch19	S00-T98	상처, 중독과 외부 원인에 의한 것들
Ch20	V01-Y98	질병이나 사망의 외부적 원인
Ch21	Z00-Z99	건강 상태에 영향을 미치는 원인들과, 보건 서비스와의 관계
Ch22	U00-U99	특별 목적을 위한 코드

- 축소방법을 이용한 회귀분석에서 최적화된 설명변수 개수는 SBC(Schwarz-Bayesian Criterion)를 기준으로 결정하였으며, 설명변수 개수에 따른 SBC 값과 회귀계수 추정값은 [그림 2]와 같음
  - ▸ 설명변수가 8개일 때 SBC가 가장 작게 나타났으며(SBC가 작을수록 좋음), 선택된 8개 설명변수는 회귀계수 추정값 크기 순으로 Ch13, Ch9, Ch11, Ch7, Ch4, Ch19, Ch5, Ch18임



[그림 2] 축소단계(설명변수개수)에 따른 SBC, 회귀계수 추정값 경로

- 선택된 8개 챕터의 계수추정값을 보면 '(Ch13)근육과 연결 조직의 질병'이 가장 높은 영향력을 보였으며, '(Ch18)다른 곳에서 분류되지 않은 증상, 증세나 임상 또는 연구에서 발견한 비정상'이 포함된 것이 특징임



[그림 3] 선택된 질환(챕터)의 계수추정값

4. 결론

- 축소방법(Shrinkage Methods)을 이용한 회귀분석의 수행 결과, '근육과 연결 조직의 질병'의 영향력이 가장 큰 것으로 나타남
  - ▸ 골다공증에 미치는 영향이 가장 큰 질환은 '근육과 연결 조직의 질병'이었으며, '순환 계통의 질병', '소화계통의 질병' 등 순으로 나타남
  - ▸ 다소 당연한 결과이지만, 보건의료 분야에서 동반 질환에 대한 새로운 분석 방법의 적용 가능성을 확인한 것으로 의미가 있음
- 위 결과는 미국의 골다공증 동반질환 연구결과와 양상은 비슷하나 세부적인 질환은 다르기 때문에, 한국인 특이적인 질환이 존재하는지 더욱 깊은 연구가 필요함
  - ▸ 이를 위해 PDN(Phenotype Disease Network)과 골대사학 기반 Gene ontology Network를 맵핑한 추가 연구를 추진하고 있음
  - ▸ 분석 환경의 한계로 인해 22개 카테고리로 분석을 실시하였지만 좀 더 세분화된 질병코드를 사용한다면 정확한 결과를 얻을 수 있을 것임