經濟學硏究 제71집 제2호 5-55
DOI : 10.22841/kjes.2023.71.2.001

국민건강보험 빅데이터를 활용한 미래 질병부담 및 의료비 추정 연구*

홍 석 철** · 이 상 연*** · 김 세 현**** · 전 성 현*****

논문초록

고령화와 만성질환 유병률 증가로 국민 질병부담과 의료비 지출이 급증하여 사전적 건강관리에 대한 사회적·정책적 수요가 높아졌다. 그러나 적극적인 건강관리가 질병부담과 의료비 지출을 얼마나 경감시킬지에 대한 의료적·경제적 타당성 평가는 상대적으로 부족한 실정이다. 본 연구에서는 국민건강보험 맞춤형 DB와 빅데이터 분석기법을 활용하여 건강관리의 의료적·경제적 타당성을 평가하는 방법론과 응용 사례를 제시한다. 우선, 다양한 질병들이 삶의 질에 미치는 영향을 고려한 질병부담지수를 제안하고, 국민건강보험 일반건강검진 결과로 10년 누적 질병부담지수를 추정한다. 그리고 질병부담지수와 의료비 간 관계 분석을 기반으로 건강관리가 검진지표 및 질병부담지수의 개선을 통해 의료비 절감에 미치는 영향을 추정하고, 건강관리 프로그램의 예시를 통해 건강관리가 미래 질병부담과 의료비지출을 개선하는 정도를 평가한다. 마지막으로 보건의료정책 및 건강보험 재정의 지속가능성 관점에서 건강관리의 중요성을 강조한다.

핵심 주제어: 국민건강보험, 빅데이터, 질병부담, 의료비, 머신러닝

경제학문헌목록 주제분류: I10, I18

투고 일자: 2023. 4. 18. 심사 및 수정 일자: 2023. 5. 9. 게재 확정 일자: 2023. 5. 30.

* 본 연구는 국민건강보험공단과 서울대 건강금융연구센터의 연구지원으로 이루어진 연구이며, 서울대 생명윤리 규정을 준수하여 연구를 수행하였음(IRB No. E2105/001-008). 본 논문은 2023 경제학 공동학술대회의 경제학연구 특별세션(빅데이터, 비정형 자료, AI를 활용한 응용 경제 연구)에서 발표한 논문임.

** 제1저자, 서울대학교 경제학부 교수, e-mail: sokchul.hong@snu.ac.kr

*** 제2저자, 서울대학교 경제학부 석사과정, e-mail: stepano1992@snu.ac.kr

**** 제3저자, 서울대학교 경제학부 박사과정, e-mail: iocking2001@snu.ac.kr

***** 제4저자, 서울대학교 경제학부 박사과정, e-mail: jeon6884@snu.ac.kr

Ⅰ. 연구 배경 및 개요

만성질환 관리는 현재 우리 사회가 당면한 가장 중요한 보건의료 문제 중 하나이다. 통계청의 2021년 「사망원인통계」에 의하면 만성질환은 전체 사망원인 중 79.6%를 차지하며, 사망원인 상위 10개 중 8개가 암, 심장질환, 뇌혈관질환, 만성호흡기질환, 당뇨병, 고혈압성질환 등 만성질환들이다. 또한 조기사망이나 상병 및 장애로 인한 건강한 삶의 손실을 대표적인 질병부담 지표인 DALYs(disability-adjusted life years, 장애보정손실연수)로 평가할 때, 2019년 기준 국민들이 경험하는 전체 질병부담 중 만성질환에 의한 부담 비중은 81.5%에 달하는 것으로 추정된다. 주목할 사실은 한국인의 총 DALYs는 10년 전에 비해 14.5% 증가했고, 만성질환에 의한 질병부담 비중은 3% 증가했다는 점이다.[1] 이러한 질병부담 증가 추이는 질병 치료를 위한 의료이용과 의료비 지출 증가를 초래하고 있다. 2021년 우리나라의 GDP 대비 경상의료비 비중은 8.4% 수준으로 OECD 평균인 9.7% 보다는 다소 낮지만 증가 속도는 매우 높은 실정이다(보건복지부, 2020; OECD, 2022).

만성질환에 따른 질병부담과 의료비 지출 증가의 주된 원인은 인구고령화이다. 나이가 들수록 다양한 만성질환에 걸릴 위험이 높아지며 그에 따라 고비용의 의료이용 역시 증가하기 때문이다. 통계청의 「2022년 고령자 통계」에 의하면 우리나라 65세 이상 고령인구 비중은 17.5%이며, 2025년에는 20%를 넘기며 초고령사회로 진입할 전망이다. 건강보험심사평가원의 「급여의약품·치료재료청구현황」에 의하면 2021년에 65세 이상 고령인구의 급여항목 진료비 청구액은 약 39조 6491억 원, 약품 청구금액은 9조 6955억 원으로 건강보험 총 진료비 대비 각각 약 42.4%, 45.5%를 차지하고 있다. 이러한 지출 중 상당한 비중이 만성질환 진료에서 기인하는 것으로 판단되는데, 2021년 12대 만성질환 진료 환자는 약 2000만 명으로 전년 대비 6.1% 증가하였으며 진료비 역시 전년 대비 8.1% 증가한 39조 2109억 원으로 추정된다.[2]

나이가 들수록 만성질환 발병 확률이 높아지는 주된 이유 중 하나는 만성질환을 유발하는 나쁜 생활습관과 환경 요인에 노출되는 기간이 증가하기 때문이다. 그런

1) DALYs 값과 만성질환 질병부담 비중은 IHME GHDx DB의 자료를 활용함. (https://vizhub.healthdata.org/gbd-results/)

2) 출처: 건강보험공단 「2021년 건강보험 통계연보」.

만큼 고령화 사회에서 만성질환 관리를 위해서는 흡연, 음주, 신체활동, 식습관 등 생활습관의 관리가 중요하다. 하지만 최근 우리나라의 교정 가능한 생활습관 개선율은 상당히 정체되어있다. 질병관리청의 「2021 만성질환 현황과 이슈」에 의하면, 고위험 음주율은 2007년 12.5%에서 2019년 12.6%, 걷기실천율은 2007년 45.7%에서 2019년 43.5%로 개선이 거의 되고 있지 않거나 오히려 악화 추이에 있다. 비만 유병률 역시 2007년 31.7%에서 2019년 33.8%로 증가 추이에 있다. 이러한 추이는 지난 10년간 모든 연령대에서 당뇨병과 고혈압 진단 유병자 비중이 증가한 사실을 잘 뒷받침한다(홍석철, 2021). 또한 생활습관에 기인한 만성질환 발병률의 증가 추이가 모든 연령대에서 관측되고 있다는 사실은 향후 인구고령화가 심화될수록 만성질환에 따른 질병부담과 의료비 부담이 빠르게 증가할 수 있음을 강하게 시사한다(Choi, Y., 2020). 특히 저성장 기조 하에서 소득 분배의 악화와 고령화가 동시에 진행되는 현재의 상황은 건강보험 재정 건전성과 지속 가능성에 있어 위험요인으로 작용할 수 있다(김준경 · 김준일, 2021).

위와 같은 문제를 해결하기 위한 정책 대응은 크게 두 가지로 나뉜다. 하나는 증가하는 만성질환의 진단과 치료를 위한 건강보험의 보장성을 높이는 방안이고, 다른 하나는 만성질환 발병률을 낮추기 위해 생활습관 개선 등 사전적인 건강관리 지원을 늘리는 방안이다. 물론 두 대응 방안은 완전히 대체할 수 있는 것은 아니며, 상호 보완적이다. 하지만 비용-효과성 측면에서 보면 사후적 치료보다는 사전적 건강관리와 예방이 더 효율적이라 기대된다.

실제로 최근 만성질환 위험군에 대한 건강관리 프로그램 활성화에 대한 적극적인 논의와 정책 개발이 추진되고 있다. 서울특별시에서 실시 중인 대사증후군관리사업이 대표적인데, 상담 및 건강 SMS 제공을 통해 대사증후군 위험이 있는 개인의 생활습관 요인을 개선함으로써 조기에 만성질환 발병 위험을 관리하는 방식으로 진행되고 있다. 또한 국민건강보험공단은 2022년 7월부터 당뇨병, 고혈압 등 만성질환 위험군을 대상으로 건강생활을 실천하면 실천 과정과 개선 정도에 따라 지원금을 제공하는 「건강생활실천지원금제」 시범사업을 시행하고 있다. 이러한 프로그램 도입의 목적은 사전적 건강관리를 유도하여 중증 · 고액의 질병 발생을 예방하고 질병으로 인한 불필요한 의료비 지출을 감소시키는 것으로 목적으로 하고 있다.[3]

3) 출처: 보건복지부 보도자료(2021.7.28.), “스스로 건강관리, 이제 국가가 지원합니다.”

사전적 건강관리 사업들이 정책 목적을 성공적으로 달성하기 위해서는 한정된 자원으로 더욱 효과적인 프로그램을 기획하고 운영할 수 있도록 건강관리 사업의 성과를 정량적으로 예측하고 평가하는 것이 중요하다. 무엇보다 건강관리 프로그램의 대상이 되는 만성질환 위험군에 속하는 개인들이 프로그램 참여 이후 건강 개선과 의료비 절감의 효과가 얼마나 되는지를 정량적으로 예측·평가할 수 있는 체계 구축이 필수적이다. 본 연구는 이러한 평가 체계 구축을 목표로 한다.

건강관리 사업의 성과는 가장 직접적으로는 혈압, 혈당, 콜레스테롤, 체질량지수 등 임상지표의 변화 그리고 금연, 규칙적 운동, 건강한 식습관 등 건강행태의 변화에서 나타나게 된다. 그리고 임상지표와 건강행태의 개선은 미래의 건강 지표를 개선하게 되며, 궁극적으로 의료비 절감 등으로 이어질 것이다. 따라서 건강관리 사업 성과 평가 체계 구축의 핵심은 미래 건강 지표를 설정하고, 임상지표와 건강행태 등 건강위험요인의 개선이 미래 건강 지표에 미치는 영향을 추정하는 것이다.

기존 연구에서는 특정 질환의 발병률이나 사망률을 미래 건강 지표로 설정하고, 현재의 위험요인 조건 하에서 미래 발병률과 사망률을 예측하였다. 미국에서는 질병부담이 높은 심뇌혈관 질환 발병률을 예측하는 Framingham Risk Score와 ACC/AHA 심뇌혈관질환 예측 모형 등이 개발되었으며 영국의 QRISK도 특정 만성질환 발병률을 미래 건강 지표로 설정한 대표적인 모형이다(Kannel et al., 1976; Wolf et al., 1991; Fleisher et al., 2014; Hippisley-Cox, 2017). 이들 연구는 흡연이나 고혈압 유병 여부, 혈중 콜레스테롤, 당뇨병 유병 여부 등에 따라 미래 심뇌혈관질환 발병률을 추정하여 건강 위험도를 제공하고 있다. 우리나라에서는 국민건강보험공단의 건강iN(건강인)의 심뇌혈관질환 위험평가 서비스가 대표적인데, 건강검진 결과의 임상지표와 문진정보를 활용하여 미래 뇌졸중과 심장질환 추정 발병률을 제공하고 있다.[4] 또한 건강보험공단이 제공하는 건강나이는 현재 건강검진 결과가 미래 사망확률에 미치는 영향을 추정한 결과들을 기본적으로 활용한다.

그러나 기존 연구들은 주로 특정 질환의 발병 위험 평가에 초점을 두고 있다는 측면에서 미래의 종합적인 건강 위험을 예측하고 평가하는데 한계를 갖는다. 사전적 건강관리를 통해 임상지표와 건강행태가 개선되면 심뇌혈관질환과 같은 특정 질환뿐만 아니라 다양한 경로를 통해 수많은 경증·중증질환의 위험도 함께 낮출 수 있기 때문이다. 대안적인 지표로 활용되어온 사망위험은 좀 더 종합적인 건강 위험

4) 이밖에 단일 질환 발병률, 사망률 등을 추정한 연구로 홍석철(2017; 2018; 2019) 등이 있다.

을 측정한다는 장점을 갖는다. 하지만 (특히 젊은 층의 경우) 사망 빈도가 낮아 건강관리의 성과를 효과적으로 예측하기 어렵다는 단점을 가진다.

이런 문제들을 고려하여 본 연구에서는 '질병부담지수'라는 개념의 좀 더 포괄적이고 종합적인 개인 건강 지표를 제안한다.[5] 질병부담지수는 특정 질환의 위험만을 평가하는 것이 아니라 270여 개에 달하는 다양한 질환과 사망 위험을 포괄하는 지표이며, 질환별로 부여된 장애 가중치를 활용하여 각 질환의 질병부담 차이를 반영한다. 따라서 현재의 건강관리사업의 건강 개선 성과는 미래의 질병부담지수의 개선 정도를 정량적으로 평가한다.

다음으로 건강관리의 직접적인 성과지표라 할 수 있는 임상지표와 건강행태 등 위험요인과 미래 질병부담지수 간의 연관성 분석은 국민건강보험공단의 맞춤형 빅데이터를 활용하였다. 이때 임상지표와 건강행태 변수는 건강보험공단 일반검진기록의 변수를, 그리고 질병부담지수의 추정은 진료기록을 활용하였다. 기존 연구들은 건강보험공단의 100만 명 표본코호트 빅데이터를 활용해왔으나, 일반건강검진 수검자를 추출하는 과정에서 분석 표본이 현저히 줄어드는 문제를 가진다. 이런 단점을 보완하기 위해 본 연구에서는 인구 연령 분포를 고려하여 2010년 기준 일반검진 수검자 150만 명 표본을 맞춤형으로 구축하여 분석에 활용하였다.

검진기록에서 관측되는 위험요인 변수와 미래 질병부담지수 간의 연관성은 다양한 빅데이터 분석모형을 활용하여 비교·분석하였으며, 모형의 성능·예측력·활용성을 고려하여 Lasso 모형을 최종적으로 선택하고 다양한 실증 분석을 수행하였다. 다만 분석의 목적은 위험요인의 수준을 기반으로 미래 질병부담지수를 예측하는 것이므로, 인과성보다는 예측력을 극대화하는데 초점을 맞추고 있음에 유의할 필요가 있다.

한편 사전적 건강관리의 또 다른 주요 목적은 미래의 의료비 절감이다. 이때 의료비 절감은 현재의 건강관리를 통해 미래 질병부담이 줄어들게 되면 의료이용이 감소하여 발생하게 된다. 따라서 의료비 절감 정도를 예측·평가하는 것은 질병부담지수가 의료비에 미치는 인과적 영향 분석에 기반할 필요가 있다. 본 연구에서는 패널 분석 모형을 적용하여 개인 단위에서 질병부담지수가 개선될 때 의료비 절감 정도를 정량적으로 분석한다.

---

5) 포괄적 건강 상태를 측정하고 지표화하는 기존 연구는 Murray(1994), Sassi(2006), Vos(2020), WHO(2020) 등을 참고하였다.

마지막으로 위에서 수행한 분석 결과를 기반으로 가상의 건강관리 사업 성과 평가 사례를 제시하여 모형의 활용 방안에 대해 논의한다. 허리둘레, 혈압, 혈당, 콜레스테롤, 중성지방 등이 위험수준에 속하는 대사증후군의 가상 표본을 가정하고 체중감량 프로그램이 적극적으로 제공되어 체중을 10% 감량할 때 향후 체중감량이 없을 때와 비교해서 향후 10년 동안 질병부담지수와 의료비가 얼마나 개선될지를 예측하여 건강관리 사업의 의료적·경제적 타당성을 논하고자 한다. 또한 향후 10년간의 성과 시뮬레이션 과정에서 건강검진지표의 변수를 현재 수준으로 고정하지 않고 매년 검진지표 변수 간의 영향과 연령에 따라 변화하는 모형을 개발하여 적용한다.

요약하면, 본 연구는 개인 건강 수준을 포괄적이고 종합적으로 평가하는 새로운 건강 평가 지표를 제안하고, 건강관리 프로그램의 성과가 직접적으로 나타나는 건강검진 지표 변화를 활용하여 미래 건강 지표와 의료비 변화를 예측하는 모형을 개발하고자 한다. 인구고령화와 생활습관 악화로 인해 국민의료비 지출이 높아지고 있는 상황에서 만성질환의 사전적 예방 정책의 의료적·경제적 타당성을 검토할 수 있는 모형 활용 사례를 제시하고자 한다.

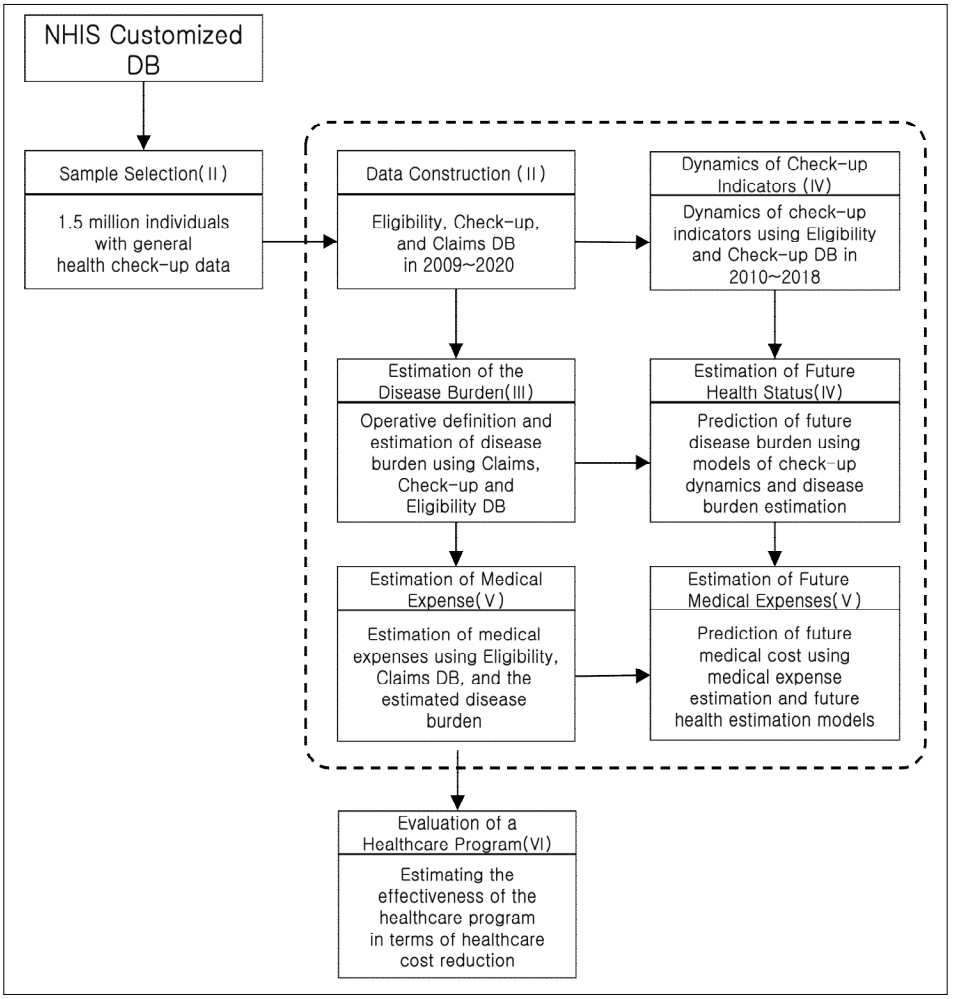
Ⅱ. 연구자료 및 분석 방법

1. 연구 방법 개요

본 연구에서 활용한 자료와 분석방법을 논의하기에 앞서 전체적인 개요를 〈Figure 1〉과 같이 요약하고자 한다. 우선, 분석모형을 개발하기 위해 활용한 자료는 건강보험공단의 맞춤형 DB로, 2010년도 일반건강검진 수검자 150만 명을 추출하여 2002~2020년도의 자격, 진료, 건강검진, 사망원인통계 자료를 연계한 자료이다. 구체적인 자료와 변수에 대해서는 제Ⅱ장의 다음 절에서 설명할 것이다.

다음으로 개인의 포괄적인 건강 수준을 나타내기 위한 지표로서 질병부담을 정의하여 개인 질병부담지수를 산정하는 작업을 진행한다. 2011~2020년도의 진료 DB의 진단코드를 활용하여 개인 질병부담지수를 계산한 뒤, 이를 질병부담 추정모형의 결과변수인 10년 누적 질병부담지수로 만들어, 2010년도의 건강검진지표로 10년 누적 질병부담지수를 추정하는 분석을 수행한다. 추정모형은 기본적으로 벌점회귀분석을 사용하였으며, 다양한 머신러닝 모델들을 활용한 추정 결과를 부수적으로

〈Figure 1〉 An Overview of the Research



제시하여 제Ⅲ장에서 비교할 것이다.

그리고 제Ⅳ장에서는 미래 건강상태의 동태적 변화를 추정하기 위해서 개인의 건강검진지표가 다음 기에 어떻게 변화하는지 추정하는 분석을 시행한다. 그리고 앞 장에서 추정한 질병부담 추정 모형과 검진지표 변화 추정모형을 토대로 추정 질병부담지수가 동태적으로 어떻게 변화하는지 예측한 결과를 제시할 것이다.

한편, 질병부담지수는 개인의 진료 기록을 토대로 산출되기 때문에 의료비와 밀접한 연관성을 가진다. 제Ⅴ장에서는 개인의 의료이용기록을 활용하여 계산된 실제 질병부담지수와 의료비 간 관계를 살펴본 뒤, 제Ⅲ장에서 추정한 질병부담지수를

이용하여 의료비를 예측하는 모형을 개발한다. 그리고 제Ⅳ장에서 질병부담지수가 어떻게 변화하는지 추정한 결과를 활용하여 개인의 미래 의료비가 어떻게 변화하는지도 추정한 결과를 제시하고자 한다.

마지막으로 제Ⅲ~Ⅴ장에서 추정한 모형들을 바탕으로 체중감량 프로그램과 같은 건강관리사업이 개인의 전반적인 건강 수준 개선이나 의료비 감소에 미치는 영향을 정량적으로 분석한 결과가 제Ⅵ장에서 제시될 것이다. 체중 감소 정도에 따라 다른 건강검진 지표가 어느 정도 개선되고, 이에 따라 미래 추정 질병부담지수와 누적 의료비 변화에 미치는 영향에 대한 분석 결과도 함께 제시한다.

2. 연구자료

분석에 활용한 자료는 국민건강보험공단의 맞춤형 DB를 활용하여 구축하였다. 맞춤형 DB는 건강보험공단이 보유한 자료를 연구 등의 목적으로 이용할 수 있도록 수요에 맞게 가공하여 제공하는 데이터로, 본 연구에서는 2010년도 국민건강보험 자격 유지자 중 건강검진을 수검한 약 1,500만 명을 모집단으로 성별, 연령 5세 그룹, 지역(대도시, 중소도시, 농어촌)에 따라 150만 명을 층화추출한 표본을 사용하였다.[6] 자료 기간은 2002~2020년이며 제공된 DB의 종류와 상세한 내역은 〈Table 1〉과 같이 구성되어 있다.

〈Table 1〉 Contents of the NHIS Customized DB

DB	Contents
Eligibility	- Year of birth, gender, region, insurance premium, etc. - (Cause of death Statistics) Year of death, Cause of death
Check-up	- Check-up result: Body Mass Index, waist circumference, blood pressure (systolic, diastolic), fasting blood glucose, total cholesterol, HDL-cholesterol, LDL-cholesterol, triglyceride, urine dipstick test, SGOT, SGPT, γ-GTP, serum creatinine - Questionnaire: family history, smoking habits, alcohol consumption, etc.
Claims	- Table 20: KCD, days of hospital admission, medical expense, etc. - Table 60: prescription details, number of days of administration, etc.

6) 맞춤형 DB는 3개의 DB로 제공받았으며, 모두 국민건강보험공단의 연구 승인(연구관리번호 NHIS-2021-1-434, NHIS-2021-1-498, NHIS-2021-1-499)을 받았다.

자격 DB는 국민건강보험 가입 자격과 관련된 DB로, 개인의 성별과 연령, 보험가입유형, 가구 지출 보험료 등급과 같은 인구·사회학적 변수들을 포함한다. 보험가입유형은 기준년도 시점에서 지역가입자, 지역세대원, 직장가입자, 직장피부양자 여부로 구성되었으며, 보험료는 가구지출 보험료 납입액이다. 생년은 기준 년도 시점 연령으로 활용하였으며, 통계청 사망원인 DB와 연동되어 제공되는 사망일은 추적관찰기간 내 사망하였는지 여부로 활용되었다.

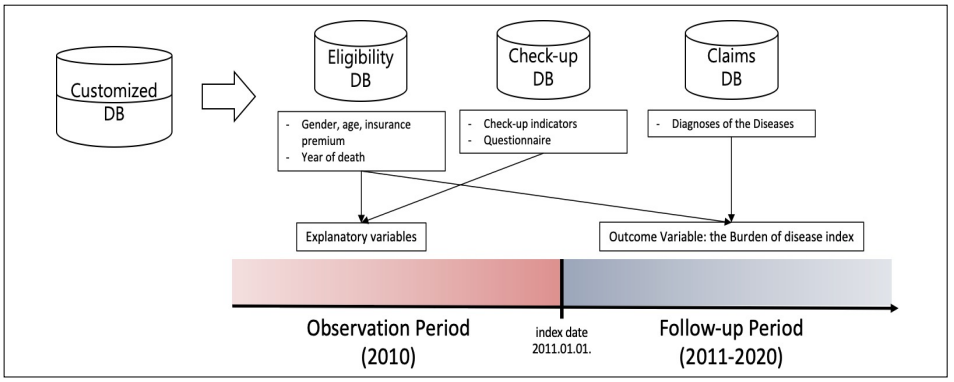
건강검진 DB는 검진 실측 데이터와 수검자 자가 기입 문진 데이터로 구성되며, 본 연구에서는 2010년부터 현재 검진항목까지 공통되는 항목을 추출하여 분석에 활용하였다. 문진은 가족력, 흡연과 음주, 신체활동 빈도 등에 대한 설문 결과 등으로 구성되며, 본 연구에서는 당뇨, 심장병, 뇌졸중, 고혈압, 기타(암 포함)에 대한 가족력 설문 결과, 현재 흡연 여부, 표준잔 기준 음주량 및 음주 빈도, 중·고강도 이상의 신체활동 여부를 변수로 활용하였다. 흡연에서 전자담배 관련 문항, 음주에서 주종별 음주량 및 음주 빈도, 신체활동에서 걷기 운동 일수, 근력운동 여부 문항은 현재 없거나 2010년도 당시에 수집되지 않은 문항이었으므로 분석에서 활용되지 않았다.

3. 분석 방법

제Ⅲ~Ⅵ장에서 사용되는 추정 모형은 다음과 같이 정리할 수 있다. 우선, 제Ⅲ장의 질병부담 추정 모형 개발은 2010년도 성별, 연령과 건강검진지표를 설명변수로 활용하여 설명변수가 2011-2020년에 걸친 10년 누적 질병부담지수에 미치는 영향을 추정하여 구축한다. 추정 과정을 도식으로 요약하면 〈Figure 2〉와 같다. 자격DB에서 성별과 연령 변수를, 검진DB에서 건강검진 및 문진 결과를 활용해서 설명변수를 구성하였으며, 진료DB의 주상병 및 부상병으로 진단받은 상병코드, 자격DB의 사망연도를 활용하여 결과변수인 10년 누적 질병부담지수를 산출하여 분석에 사용하였다.

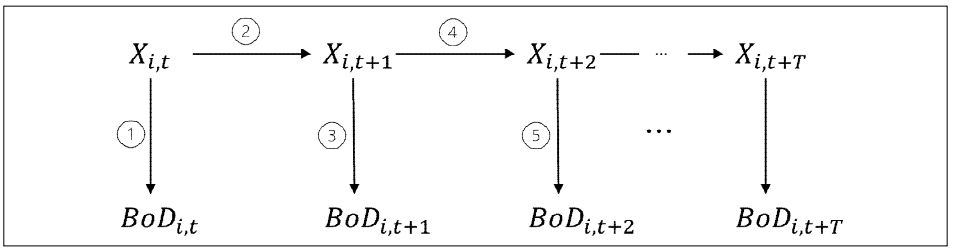
제Ⅳ장의 건강검진지표의 동태적 변화 추정 모형은 이번 기 검진지표로 다음 기 검진지표를 예측하는 모형으로, 자격 DB의 개인 성별, 연령 그리고 건강검진 DB의 연간 건강검진 수검 내용을 사용한다. 또한 추정된 검진지표의 동태적 변화 모형과 제Ⅲ장에서 추정한 질병부담지수 추정 모형을 통해 종합적인 건강상태가 동태

〈Figure 2〉 Research Design of Disease Burden Estimation



적으로 어떻게 변화하는지를 살펴볼 것이다. 이를 도식으로 나타내면 〈Figure 3〉과 같다. t기 개인 i의 건강검진지표가 $X_{i,t}$이고, $X_{i,t}$로 추정한 질병부담지수가 $BoD_{i,t}$라 했을 때, $t+1$기 추정 질병부담지수는 다음 기 검진지표 $X_{i,t+1}$를 추정한 뒤, 추정된 $X_{i,t+1}$를 이용하여 $BoD_{i,t+1}$로 계산된다.

〈Figure 3〉 Dynamics of Check-up Indicators and the Disease Burden

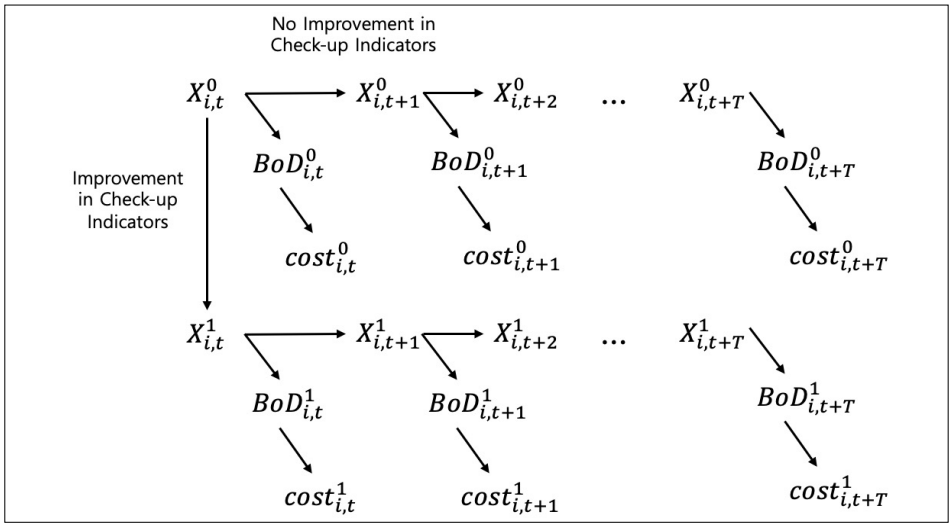


제Ⅴ장에서는 추정한 질병부담지수를 활용하여 의료비를 추정한다. 매기의 의료비는 같은 기에 추정된 질병부담지수를 이용하여 예측하며, 예측 대상 의료비는 건강보험공단을 통해 지출되는 총급여의료비, 의과 및 보건기관 입원의료비, 외래의료비, 그리고 약국조제 의료비이다. 이렇게 개발한 의료비 예측모형은 제Ⅳ장의 건강검진지표의 동태적 변화 및 질병부담지수의 동태적 변화 예측모형과 같이 개인의 장기적인 의료비 지출 변화를 분석하는데 사용될 수 있다.

마지막으로 제Ⅵ장에서는 앞에서 개발된 모형들의 활용 사례로서 건강관리사업을 정량적으로 평가하는 작업을 수행한다. 건강관리사업을 통해 건강검진지표에 개

선이 있었을 시, 그렇지 않았을 때보다 질병부담지수와 의료비 지출액 차이가 발생하는지 여부를 분석하고자 한다. 분석은 〈Figure 4〉와 같이 요약된다. 현재 건강상태가 $X_{i,t}^{0}$ 인 개인이 건강관리사업에 참여하여 $X_{i,t}^{1}$ 로 검진지표가 개선되는 경우 건강개선이 이뤄지지 않았을 시의 전반적 건강 수준 및 의료비와 건강개선이 이뤄졌을 시의 전반적 건강 수준 및 의료비를 비교하여 건강관리사업의 효과를 정량적으로 평가하는 작업을 수행하는 것이다.

〈Figure 4〉 Evaluation of Healthcare Program



Ⅲ. 질병부담 추정 모형

1. 질병부담지수 산출

개인의 포괄적이고 종합적인 건강 수준을 측정하기 위하여 개인이 연내 진단받은 질환들을 활용하여 질병부담지수를 산출한다. 질병가중치는 사망을 1, 연내 아무런 질환을 진단받지 않은 상태를 0이라 하고, 질환의 중증도에 따라 0에서 1 사이의 값을 가중치로 할당한다. 질병 가중치 산출에는 사망에 미치는 영향, 유관 비용, 전반적인 의료기술 발달 정도 등 사회적인 배경이 영향을 줄 수 있다. 따라서 본 연구에서는 국내를 배경으로 한 연구 결과를 활용하여 질병가중치를 부여하였다

(Ock et al., 2019; Kim, Young-Eun et al., 2020). Kim, Young-Eun et al. (2020)에서는 사망과 건강한 상태를 포함하여 313개 원인 질환에 대해 국내 의료계 전문가[7]를 대상으로 한 온라인 설문조사를 진행하였으며, 이를 바탕으로 국내에서 인식되는 질환별 장애 가중치(disability weight)를 산출하였다. 최종적으로 암종별 병기, 질환 내 중증도 등 주어진 맞춤형 DB만으로는 정의가 어려운 일부 질환[8]들을 제외하고 277개 질환의 가중치만을 활용하였다.

부여된 가중치의 예시는 〈Table 2〉의 I과 같다. 자료 활용을 위하여 질환들에 대해 KCD코드를 활용하여 조작적 정의를 추가하였으며, 여기서는 진료 DB에서 해당 KCD코드를 진단받은 경우 원인 질환에 걸린 것으로 간주하였다. 예컨대 I20-I25에 해당하는 상병코드로 진단받은 표본의 경우 허혈성 심질환(ischemic heart disease)으로 진단받은 것으로 보아 0.703의 가중치가 부여된다. 가중치는 질환이 건강 수준에 미치는 영향에 비례하는데, 1에 가까울수록 중증도의 질환이며 0에 가까울수록 경증의 질환을 의미한다. 예컨대 식도암(Oesophageal cancer)의 경우 중증도의 질환으로 1에 가까운 0.870의 가중치를 가지는 반면, 비교적 경증질환인 편두통(migraine)의 경우 0에 가까운 0.189의 값을 가진다. 기타 심근병증(other cardiomyopathy)과 같이 제외코드가 있는 원인 질환에 대해서는, 해당 KCD코드를 진단받은 바 있는 표본 중 제외코드를 받은 표본을 제외하고 가중값 0.714를 부여하였다.

암, 허혈성 뇌졸중(ischemic stroke)이나 만성 폐쇄성 폐질환(COPD)처럼 질환별로 병기(stage)나 중증도가 구분된 경우가 일부 존재하는데, COPD와 같이 KCD로 구분이 가능한 경우 문제가 없으나 암이나 허혈성 뇌졸중과 같이 KCD만으로는 중증도를 식별할 수 없는 경우가 있었다. 이 경우 가중치의 평균을 내어 부여하는 방식을 따랐다. 가령 I63 코드를 받은 허혈성 뇌졸중 환자의 경우 0.733의 가중치를 부여하였다.[9] 이렇게 선정된 질환별 가중치들의 통계량은 〈Table 2〉의 II와 같다.

---

7) 전문가 패널은 의사와 의대생(Group 1), 간호사와 한의사(Group 2)로 나뉘는데, 본 연구에서는 Group 1을 대상으로 한 결과를 사용하였다.

8) 총 2개 질환이 제외되었는데 G6PD trait은 원인 질환에 해당하는 상병코드가 존재하지 않았으며 caries of deciduous teeth는 소아/청소년기의 질환이라 본 연구의 분석 표본인 20세 이상 성인에게 해당되지 않는 질환이었다.

9) 허혈성 뇌졸중 외에 자료만으로 중증도 분류가 되지 않는 질환들은 1~4기로 구분된 유방암, 자궁경부암, 대장암, 간암, 전립선암, 위암, 갑상선암, 폐암, 알코올성 간경화 및 간질환,

평균적으로 가중치는 0.474의 값을 가지며, 질환 277개에 완전히 건강한 상태와 사망한 상태를 포함하여 총 279개 가중치를 분석에 활용하였다.

〈Table 2〉 Operational Definitions of the Diseases and Weights (Partial Excerpt) and Summary Statistics of the Weights

I. Operational definitions of diseases and weights			
Cause of Disease	KCD of the disease	Exception	Weight
Ischaemic heart disease	I20-I25		0.703
Ischemic stroke (mild)	I63		0.560
Ischemic stroke (moderate)	I63		0.797
Ischemic stroke (severe)	I63		0.843
Hemorrhagic stroke	I60-I62		0.800
Hypertensive heart disease	I11-I15		0.474
Myocarditis	I40, I41, I51.4		0.663
Alcoholic cardiomyopathy	I42.6		0.649
Other cardiomyopathy	I42	I42.6	0.714
Atrial fibrillation and flutter	I48		0.549
Peripheral vascular disease	I73		0.449
Endocarditis	I33, I38		0.690
COPD (mild)	J44.00, J44.10, J44.80, J44.90		0.474
COPD (moderate)	J44.01, J44.11, J44.81, J44.91		0.658
COPD (severe)	J44.02, J44.12, J44.82, J44.92		0.753
Silicosis	J62		0.666
Asbestosis	J61		0.653
Coal workers pneumoconiosis	J60		0.658
Other pneumoconiosis	J63-J65		0.582
Asthma	J45-J46		0.409
Migraine	G43		0.189
Oesophageal cancer	C15		0.870
Full Health	-		0
Death	-		1

II. Summary statistics of weights	
Mean	0.474
Standard deviation	0.206
Minimum	0
Q1	0.309
Q2	0.488
Q3	0.652
Maximum	1
Number of Diseases	279

Note: Cause of Diseases and their weights are from Kim, et al. (2020).

요통, 골관절염이 있다.

질환별 가중치가 선정되면, 같은 해에 여러 질환을 동시에 진단받는 경우를 고려하여 식 (1)과 같이 진단받은 질환들의 가중치를 활용하여 개인의 연내 질병부담을 계산한다.

$$BD_{i,t} = 1 - \prod_{d \in D_{i,t}} (1 - DW_d) \tag{1}$$

여기서 $D_{i,t}$는 개인 i가 t년도에 진단받은 질환들의 집합이며, DW_d는 질환 d에 대해 부여된 가중치이다. 진단받은 질환의 가중치를 식 (1)과 같이 변형하여 모두 곱하는 방식으로 산출한 개인의 연내 질병부담은 매년 0~1 사이의 값을 가지며, 중증도가 높은 질환을 많이 진단받을수록 1에 가까운 값을 갖게 되나 사망 시 부여되는 1을 넘지는 않게 된다.

분석에서 종속변수로 활용할 추적관찰기간 10년에 걸친 개인의 누적 질병부담은 아래 식 (2)와 같이 시간할인된 연간 질병부담의 합산으로 계산하였다. 여기서 시간할인율 δ는 다른 연구들과 같이 3%로 적용하였다.[10]

$$BoD_i = \sum_{t=2011}^{2020} BD_{i,t}(1-\delta)^{t-2010} \tag{2}$$

추적관찰기간 종료 전에 사망하는 경우, 추적관찰기간 종료까지 모두 연간 질병부담을 1로 두었다. 예컨대 2015년도에 사망한 표본의 경우 2015년도 외에도 2016~2020년까지 모두 연간 질병부담이 1이다. 2015년도 사망자의 경우 2016~2020년도의 실제 질병부담은 존재하지 않으나 조기 사망자의 질병부담이 추적관찰기간 내 생존한 표본이나 비교적 후기에 생존한 표본에 비해 역전되지 않고 높게 측정될 수 있도록 이러한 방식을 택하였다. 이렇게 계산하면 개인의 시간할인된 10년 누적 질병부담은 최소 0에서 최대 약 8.495의 값을 가진다. 10년 누적 질병부담이 0인 경우 추적관찰기간에 걸쳐 진단받은 질환이 없다는 것을 의미하며, 누적 질병부담이 최댓값인 8.495이라면 추적관찰기간 개시 시점(2011년)에 사망한 표본이다.

10) 질병부담의 개인화 및 시간할인을 활용한 누적질병부담지수 산출과 관련된 보다 자세한 논의는 Zhang(2020)를 참고하였다.

2. 건강검진지표를 활용한 질병부담지수 예측

질병부담지수 추정을 위해 설명변수로 사용한 2010년도 건강검진지표와 2011~2020년 10년 누적 질병부담지수의 기초통계량은 〈Table 3〉과 같다.

〈Table 3〉 Summary Statistics for Disease Burden Estimation

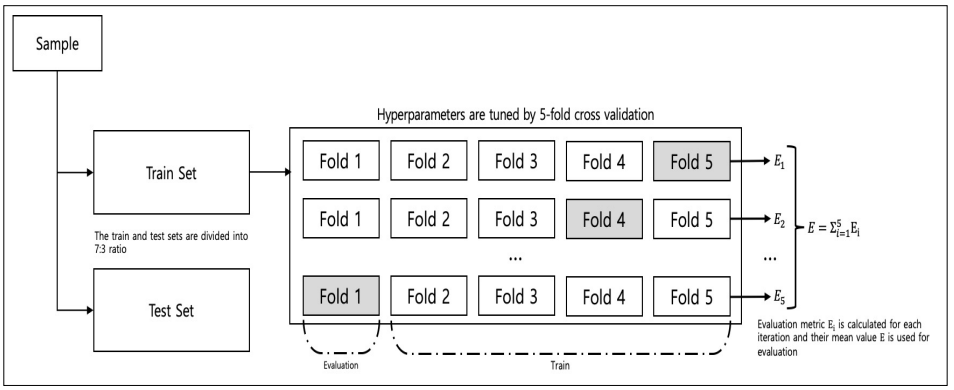
Variable	Mean	Variable	Mean
Outcome variable		Explanatory variables	
10-year cumulative Disease Burden	2.382	Proteinuria: Weak Positive	0.022
	(1.881)		(0.147)
		Proteinuria: Positive	0.024
Explanatory variables			(0.153)
Gender (1 if male)	0.492	SGOT (IU/L)	24.547
	(0.500)		(15.628)
Age	45.297	SGPT (IU/L)	23.920
	(15.536)		(21.152)
Body mass index (kg/m2)	23.412	γ-GTP (IU/L)	33.668
	(3.360)		(46.055)
Waist circumference (cm)	79.107	Serum creatinine (mg/dL)	1.014
	(9.531)		(1.083)
Systolic blood pressure (mmHg)	121.152	Risky drinking	0.662
	(15.063)		(0.473)
Diastolic blood pressure (mmHg)	75.381	Exercise	0.502
	(10.030)		(0.500)
Fasting blood glucose (mg/dL)	96.235	Hepatitis B	0.040
	(22.738)		(0.195)
Hemoglobin (g/dL)	13.830	Family history: Stroke	0.054
	(1.630)		(0.225)
Total cholesterol (mg/dL)	191.701	Family history: Heart Disease	0.031
	(36.525)		(0.174)
HDL cholesterol (mg/dL)	56.221	Family history: Hypertension	0.115
	(20.378)		(0.319)
Triglyceride (mg/dL)	123.859	Family history: Diabetes	0.090
	(82.430)		(0.286)
Obs. No.	1453990		

Note: Standard deviations in parentheses. Abbreviation SGOT refers to Serum glutamic oxaloacetic transaminase, SGPT to Serum glutamate pyruvate transaminase, γ-GTP to gamma-glutamyl transferase. Risky drinking refers to consuming alcohol on five or more days per week or consuming more than seven drinks for men and five drinks for women in a single occasion. Exercise refers to performing moderate or vigorous exercise weekly.

원표본은 150만 명이나 검진지표가 이상치이거나 기준년도 2010년도에 검진을 받은 뒤 사망한 표본을 제외하여 관측치 수는 1,453,990명이 되었다. 종속변수인 10년 누적 질병부담지수는 평균이 약 2.382로 추정되며, 평균 연령은 약 45세였다. 검진지표들의 평균은 대부분 정상 구간에 속해있었으나, 예외적으로 수축기혈압(≥120), 고밀도콜레스테롤(≤60)이 평균적으로 주의군에 속해있었다. 이 외에 주 5회 이상 음주하거나 1회 음주량이 7잔(남성) 또는 5잔(여성) 이상인 경우로 정의된 고위험음주군이 66%였으며, 현재 흡연하는 표본은 약 24.5%, 주 1회 이상 중등도 이상의 운동을 실행하는 표본이 약 50.2%였다. 가족력의 경우 고혈압과 당뇨 유병자 가족을 둔 표본이 각각 11.5%와 9%로 높았으며, 뇌졸중(5%), 심장병(3.1%)은 비교적 낮은 수준이었다.

분석 결과는 우선 머신러닝 모형들을 적용한 결과의 평가지표들을 제시한 뒤, 선형 모형 추정 결과를 기준으로 추후 논의를 이어가고자 한다. 우선 머신러닝 분석을 진행하기 위해 자료를 분할하고 훈련하는 과정은 〈Figure 5〉와 같이 요약된다.

〈Figure 5〉 Model Training and Cross-Validation



〈Figure 5〉에서 표본은 7:3 비중으로 임의로 분할하여 각각 모형을 훈련하는 용도의 훈련용 표본(train-set)과 훈련된 모형으로 최종 평가지표를 산출하는 검증용 표본(test-set)으로 나누었다. 훈련용 표본은 모형 훈련 과정에서 모형이 자료에 과적합되지 않도록 k-fold 교차검증(k-fold cross validation) 방식으로 모형 훈련에 사용되었다. 구체적으로 훈련용 표본을 k개로 나눈 뒤, k-1개를 학습용 데이터로 1개를 평가용으로 활용하며, 이 방법을 k회 시행하여 k개의 평가지표를 얻는 방식이

다. 본 연구에서는 k=5로 두어 훈련 과정을 진행하였다.

건강검진지표를 활용하여 10년 누적 질병부담지수를 추정하는 모형은 OLS와 Lasso, 머신러닝 모델을 모두 활용하였으며, 모형별 비교 결과는 아래 〈Table 4〉와 같다.

〈Table 4〉 Evaluation Metrics of the Machine Learning Models

Model	R-Squared	MAE	MSE
LGBM	0.378	0.185	0.084
Lasso	0.354	0.198	0.090
Ridge	0.351	0.200	0.090
OLS	0.361	0.189	0.089
Random Forest	0.311	0.231	0.101

Note: Abbreviation LGBM represents Light Gradient Boosting, MAE refers to the mean absolute error, MSE to the mean squared error.

모형의 성능을 평가하기 위한 지표로 R-squared, MAE, MSE를 사용하였으며, 모든 지표에서 LGBM(Light Gradient Boosting Machine)[11]의 추정 정확도가 가장 높았으며 Random Forest의 정확도가 가장 낮게 추정되었다. 추정과정에서 변수를 선택하는 Lasso와 같은 모형들이 있음을 고려하여 MAE, MSE를 주로 모형 성능 평가지표로 고려하였을 때 전반적으로 평가지표의 값에 큰 차이가 나지는 않았으며, 대체로 LGBM 외에 OLS, Lasso, Ridge, Random Forest 순으로 정확한 예측 모형이 구축된 것으로 보인다. 다만, Random Forest는 모형 훈련 과정에서 과적합을 방지하기 위해 조정해야 할 초매개변수의 수가 많고 적당한 값을 찾는데 오랜 시간이 걸린다는 점에서 위의 비교 결과에 대해 지나치게 일반화하는 해석은 지양할 필요가 있다.

이하부터는 Lasso 모델을 활용하여 추정한 결과를 중심으로 논의하고자 한다. Lasso는 선형 추정 모형으로 계수의 부호와 크기를 통한 결과 해석이 보다 용이하며, 다른 머신러닝 모델들에 비해 예측력이 크게 부족하지도 않기 때문이다. Ridge나 OLS와 같은 다른 선형회귀모형도 비슷한 정확도를 보였으나, 일부 추정된 계수의 값이 유의하지 않거나 의학 이론과 다른 방향으로 추정되는 경우가 있어, 건강

11) LGBM 모형과 활용에 관해서는 Ke(2017)를 참고하였다.

관리의 효과를 추정하는데 있어 문제가 생길 수 있기 때문이다.[12]

〈Table 5〉 Disease Burden Estimation Result

Variable	Coef.	Variable	Coef.
Gender	-0.511***	γ-GTP	0.001***
	(0.003)		(0.000)
Age	0.063***	Serum creatinine	0.014***
	(0.000)		(0.001)
Body Mass Index: Low-weight	-0.135***	Proteinuria: Weak Positive	0.130***
	(0.005)		(0.009)
Body Mass Index: overweight	0.003***	Proteinuria: Positive	0.440***
	(0.001)		(0.008)
Waist circumference	0.031***	Risky alcohol consumption	0.056***
	(0.001)		(0.003)
Blood pressure	0.002***	Family history: Stroke	0.016***
	(0.000)		(0.006)
Fasting blood glucose(mg/dL)	0.008***	Family history: Heart Disease	0.026***
	(0.000)		(0.007)
Hemoglobin(g/dL)	-0.128***	Family history: Hypertension	0.160***
	(0.003)		(0.004)
Total cholesterol(mg/dL)	0.004***	Family history: Diabetes	0.020***
	(0.000)		(0.005)
Triglyceride	0.003***	Exercise	-0.022***
	(0.000)		(0.003)
HDL cholesterol	-0.007***	Hepatitis B	0.197***
	(0.000)		(0.006)
SGOT, SGPT	0.001***	Constant	-0.486***
	(0.000)		(0.005)
Obs. No.	1453990		
adj. R-sq	0.352		

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

Lasso로 추정한 검진지표들의 계수는 〈Table 5〉에서 제시하였다. 남성의 경우 여성에 비해 미래 질병부담지수가 낮게 추정되는데, 남성의 만성질환 노출 위험이

12) 마찬가지로 LGBM, 랜덤 포레스트 등도 의학적으로 중요한 변수가 활용되었는지, 변수의 한계적 효과의 크기와 방향이 불명확하다는 점에서 추가적인 활용은 Lasso 모형의 결과를 따랐다.

낮음을 의미하지만 남성이 상대적으로 의료이용 빈도가 낮기 때문에 질병부담지수의 기초가 되는 질환 진단 가능성이 낮은 영향일수도 있다. 본 연구는 각 요인이 질병부담지수에 미치는 인과적인 관계를 추정하는 것이 목적이 아니며, 요인과 질병부담지수 간의 상관성을 추정하고 예측력을 극대화하는데 목적을 두고 있음에 유의할 필요가 있다.

요단백의 경우 음성에 비해 약양성인 표본이 약 0.13, 양성인 표본이 약 0.44 더 높은 질병부담지수를 갖는 것으로 추정되었다. 이 외에 다른 위험요인이 증가함에 따라 질병부담지수는 증가하는 것으로 분석되었다. 다만, 저체중과 고밀도콜레스테롤, 혈색소의 경우 일반적으로 더 높은 값을 갖는 것을 건강한 것으로 간주한다. 이들 지표의 경우 계수가 음의 값을 가졌다는 것은 따라서 해당 지표가 낮아짐에 따라 질병부담이 점차 증가하는 것으로 분석된 것이다. 1주에 중고강도 운동을 수행하는지 여부 역시 마찬가지 이유에서 음의 계수값을 갖도록 추정되었다.

Ⅳ. 검진지표와 질병부담지수의 동태적 변화 추정

1. 검진지표의 동태적 변화 추정

건강검진지표 동태적 변화는 식 (3)과 같이 추정하였다.[13] 통상적으로 비사무직 근로자가 1년, 사무직 근로자가 2년 주기로 일반건강검진을 받게 되어있다는 점을 고려하여 2년 후의 건강검진지표를 예측하는 모델을 개발하였으며, 활용 변수는 이번 기 건강검진결과와 연령이다. 예측 대상 변수는 체질량지수, 허리둘레, 혈압(수축기혈압, 이완기혈압), 혈색소, 공복혈당, 콜레스테롤(총콜레스테롤, 중성지방, 고밀도콜레스테롤), 간수치(혈청지오티, 혈청지피티, 감마지티피), 혈청크레아티닌이며, 검진DB에 존재하지만 예측대상 변수가 아닌 변수들은 시간이 지남에도 일정하게 유지된다고 가정하였다.[14] 예측대상 변수가 아닌 변수들은 기간 간 변화가 거의 관측되지 않아 분석에서 제외되었다. 종속변수는 이번 기 건강검진지표($v_{i,t}$)와 다음 기 건강검진지표의 차분값($\Delta v_{i,t+2} = v_{i,t+2} - v_{i,t}$)으로 변환하였으며, 주요한

13) 건강검진지표의 예측과 관련해서는 Vuik et al. (2019; 2021)을 참고하였다.

14) 생활습관 문진 항목, 가족력, 요단백, B형간염 바이러스 보유 여부 등이 일정하다고 가정되었다.

설명변수로 흡연, 음주, 운동 등 문진으로 응답한 생활습관 변수들과 종속변수 v를 제외한 나머지 개인의 다른 건강검진지표($X_{i,t}^{-v}$)를 사용하였다. 연령($age_{i,t}$)은 2차식까지 활용하여 통제하였으며, 성별에 따라 검진지표의 변동 추이나 전반적인 수준이 상이하다는 점을 고려하여 모든 모형은 성별로 표본을 나누어 추정하였다.[15]

$$\Delta v_{i,t+2} = X_{i,t}^{-v} B + \gamma_1 age_{i,t} + \gamma_2 age_{i,t}^2 + \epsilon_{i,t} \tag{3}$$

마지막으로 일부 검진지표의 경우 일시적인 사유, 혹은 측정상의 오차 등으로 인해 값이 크게 변동하는 경우가 존재한다. 따라서 이하부터는 콜레스테롤(총콜레스테롤, 고밀도콜레스테롤, 중성지방), 공복혈당의 경우 극단적인 이상치를 제외하고 분석한 결과를 제시한다. 극단적 이상치는 검진지표가 $3Q+3IQR$을 초과하는 경우로 정의하였고, 이에 해당하는 경우 분석 표본에서 제거하였다(선정연 외, 2019).

추정을 위해 사용한 자료는 2010~2019년 연간 건강검진 자료이다. 분석 모형 특성상 자료에서 2년 주기로 검진기록이 있는 표본들만이 분석에 사용되었으며, 해당 표본 내에서 건강검진지표의 분포는 〈Table 3〉의 검진지표 기초통계량 분포와 유사하다.

여성과 남성 표본 각각에 대해 분석한 결과는 〈Table 6〉~〈Table 9〉와 같다. 여성 표본에서의 흡연, 남성 표본에서의 흡연과 고위험 음주의 경우 대부분의 경우 다음 기 검진수치를 높이는 것으로 추정되었으며, 특히 흡연은 다음 기 검진수치 악화를 더 크게 야기하는 것으로 추정되었다. 신체활동의 경우 반대로 다음 기 수치를 낮추는 것으로 추정되었다. 검진지표로는 체질량지수, 혈색소 등이 다음 기 검진지표 추정에 사용되는 빈도가 높았다.

15) 건강검진지표의 동태적 변화 추정모형은 일반적으로 건강관리를 하지 않았을 때의 동태적 변화를 추정하는 것이 목적이다. 따라서 검진지표가 질환의심군으로 판정받는 경우 분석 표본에서 제외하였다. 해당 표본의 경우 의원 방문이나 투약, 식단 조절과 같은 자체적인 건강관리 등 자료를 통해 측정할 수 없는 방식으로 건강관리행태가 이뤄질 수 있기 때문이다.

〈Table 6〉 Estimated Changes in Check-up Indicators of Female Samples

	Body mass index	Waist circumfer-ence	Systolic blood pressure	Diastolic blood pressure	Hemoglobin	Fasting blood glucose	Total cholesterol
Body mass index			0.099***	0.038***		0.083***	
			(0.005)	(0.002)		(0.005)	
Waist circumference			0.001			0.036***	
			(0.002)			(0.002)	
Systolic blood pressure							
Diastolic blood pressure	0.001***	0.004***			-0.001***		
	(0.000)	(0.000)			(0.000)		
Hemoglobin						-0.154***	-0.647***
						(0.007)	(0.014)
Fasting blood glucose					-0.001***		
					(0.000)		
Total cholesterol					-0.001***		
					(0.000)		
HDL cholesterol						-0.015***	-0.096***
						(0.001)	(0.001)
Triglyceride						0.003***	
						(0.000)	
SGOT					-0.000***	0.002	
					(0.000)	(0.001)	
SGPT					-0.000	0.015***	
					(0.001)	(0.001)	
γ-GTP			0.001		-0.000***	0.003***	
			(0.000)		(0.000)	(0.001)	
Serun creatinine	0.005***	0.083***					0.028
	(0.001)	(0.006)					(0.031)
Smoking	0.090***	0.192***	0.075	0.130***		0.543***	1.464***
	(0.006)	(0.027)	(0.061)	(0.044)		(0.063)	(0.142)
Risky drinking	0.004*					0.048**	
	(0.002)					(0.022)	
Exercise					0.006***		
					(0.002)		
Age	0.011***	0.023***	-0.122***	-0.021***	-0.023***	-0.058***	0.900***
	(0.000)	(0.001)	(0.003)	(0.002)	(0.000)	(0.004)	(0.008)
Age^2	-0.000***	-0.000***	0.002***	0.000***	0.001***	-0.010***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Obs. No.	1644613	1703274	1565453	1601167	1669912	1627315	1515771
adj. R-sq	0.021	0.010	0.030	0.009	0.008	0.026	0.039

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

〈Table 7〉 Estimated Changes in Check-up Indicators of Female Samples (continued)

	HDL cholesterol	Triglyceride	SGOT	SGPT	γ-GTP	Serum creatinine
Body mass index		0.192***	0.051***	0.060***	0.091***	
		(0.021)	(0.003)	(0.003)	(0.005)	
Waist circumference		0.059***			0.008***	
		(0.008)			(0.002)	
Systolic blood pressure						
Diastolic blood pressure					0.002***	
					(0.001)	
Hemoglobin	-0.263***	-0.479***	-0.134***		-0.233***	-0.015***
	(0.009)	(0.027)	(0.006)		(0.007)	(0.000)
Fasting blood glucose			0.000			
			(0.001)			
Total cholesterol	-0.039***					
	(0.000)					
HDL cholesterol			-0.010***	-0.016***	-0.007***	-0.002***
			(0.000)	(0.001)	(0.001)	(0.000)
Triglyceride	-0.012***		0.001***		0.002***	
	(0.000)		(0.000)		(0.000)	
SGOT	-0.008***					
	(0.002)					
SGPT	-0.007***					0.000***
	(0.002)					(0.000)
γ-GTP	-0.001	0.041***				0.000*
	(0.001)	(0.002)				(0.000)
Serun creatinine	-0.831***	0.188***				
	(0.020)	(0.053)				
Smoking		4.680***	0.547***	0.409***	1.332***	0.020***
		(0.246)	(0.055)	(0.069)	(0.063)	(0.004)
Risky drinking						
Exercise						-0.003*
						(0.001)
Age	-0.255***	0.154***	0.089***	0.055***	0.079***	0.008***
	(0.005)	(0.014)	(0.003)	(0.003)	(0.004)	(0.000)
Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Obs. No.	1612767	1573964	1672079	1659609	1604890	1703274
adj. R-sq	0.009	0.025	0.011	0.009	0.017	0.007

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

〈Table 8〉 Estimated Changes in Check-up Indicators of Male Samples

	Body mass index	Waist Circumfer-ence	Systolic blood pressure	Diastolic blood pressure	Hemoglobin	Fasting blood glucose	Total cholesterol
Body mass index			0.099***	0.051***		0.110***	
			(0.003)	(0.002)		(0.007)	
Waist circumference						0.032***	0.000
						(0.003)	(0.003)
Systolic blood pressure	0.003***	0.008***					0.033***
	(0.000)	(0.000)					(0.002)
Diastolic blood pressure					-0.002***		
					(0.000)		
Hemoglobin						-0.265***	-0.347***
						(0.008)	(0.015)
Fasting blood glucose					-0.001***		
					(0.000)		
Total cholesterol					-0.002***		
					(0.000)		
HDL cholesterol					0.000	-0.024***	-0.045***
					(0.000)	(0.001)	(0.001)
Triglyceride					0.000***		
					(0.000)		
SGOT					-0.000***	0.009***	
					(0.000)	(0.001)	
SGPT						0.014***	
						(0.001)	
γ-GTP			0.003***	0.002***	-0.000***	0.004***	
			(0.000)	(0.000)	(0.000)	(0.000)	
Serun creatinine	0.010***				-0.006***		0.113***
	(0.001)				(0.001)		(0.022)
Smoking	0.093***	0.282***	0.209***	0.102***	-0.014***	0.961***	1.296***
	(0.002)	(0.007)	(0.020)	(0.015)	(0.001)	(0.025)	(0.044)
Risky drinking	0.019***	0.047***	0.122***	0.102***		0.166***	0.351***
	(0.002)	(0.007)	(0.020)	(0.014)		(0.024)	(0.043)
Exercise	-0.014***		-0.023				-0.032
	(0.002)		(0.020)				(0.043)
Age	0.001***	0.001	-0.100***	-0.006***	-0.030***	0.036***	0.441***
	(0.000)	(0.001)	(0.003)	(0.002)	(0.000)	(0.005)	(0.008)
Age^2	-0.000***	-0.000***	0.002***	0.000	0.000***	-0.000	-0.006***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Obs. No.	1943073	2034108	1812705	1826523	2012101	1879209	1826202
adj. R-sq	0.044	0.019	0.026	0.018	0.012	0.039	0.024

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

〈Table 9〉 Estimated Changes in Check-up Indicators of Male Samples (continued)

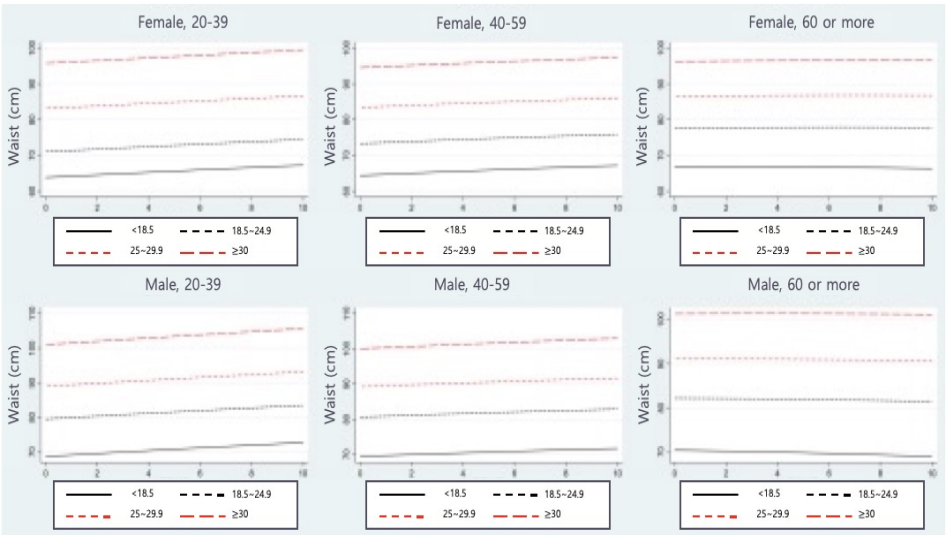
	HDL cholesterol	Triglyceride	SGOT	SGPT	γ-GTP	Serum creatinine
Body mass index		0.644***	0.036***	0.071***		0.001***
		(0.031)	(0.006)	(0.008)		(0.000)
Waist circumference		0.115***	0.002	0.015***	0.020***	
		(0.012)	(0.002)	(0.003)	(0.002)	
Systolic blood pressure		0.006	0.003***	0.020***	0.037***	
		(0.006)	(0.001)	(0.001)	(0.001)	
Diastolic blood pressure	-0.012***	0.002				
	(0.001)	(0.009)				
Hemoglobin	0.069***	-0.903***			-0.191***	-0.014***
	(0.009)	(0.041)			(0.013)	(0.000)
Fasting blood glucose	-0.006***		0.004***			
	(0.001)		(0.000)			
Total cholesterol	-0.024***	0.026***				
	(0.000)	(0.002)				
HDL cholesterol		-0.005				-0.001***
		(0.004)				(0.000)
Triglyceride						0.000
						(0.000)
SGOT	-0.016***	0.005				
	(0.001)	(0.004)				
SGPT						0.000***
						(0.000)
γ-GTP	-0.003***	0.063***	0.003***			0.000***
	(0.000)	(0.001)	(0.000)			(0.000)
Serun creatinine	-0.307***	0.138***				
	(0.015)	(0.057)				
Smoking	-0.122***	7.673***	0.501***	0.790***	2.560***	0.019***
	(0.028)	(0.114)	(0.021)	(0.028)	(0.037)	(0.002)
Risky drinking	-0.082***	2.248***	0.132***	0.131***	0.893***	0.002
	(0.027)	(0.110)	(0.020)	(0.028)	(0.036)	(0.002)
Exercise	0.108***		0.071***			-0.023***
	(0.028)		(0.021)			(0.002)
Age	-0.229***	0.012	-0.015***	-0.091***	0.008	0.004***
	(0.006)	(0.022)	(0.004)	(0.005)	(0.007)	(0.000)
Age^2	0.002***	-0.003***	0.000	0.000***	-0.001***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Obs. No.	1750376	1622630	1947398	1911766	1771063	2034108
adj. R-sq	0.007	0.069	0.016	0.017	0.031	0.007

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

2. 질병부담지수의 동태적인 변화 추정

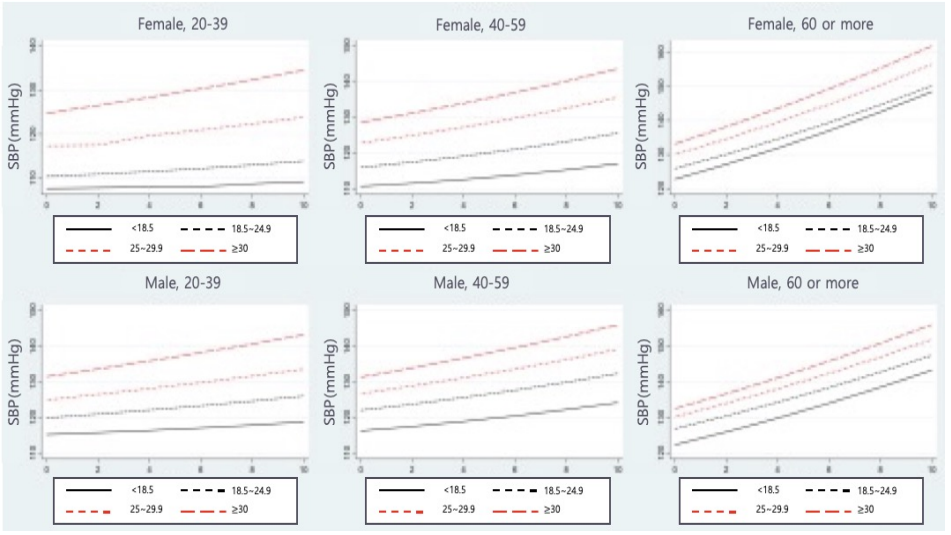
제Ⅲ장에서 추정한 질병부담지수 추정모형과 위에서 추정한 건강검진지표의 동태적 변화 추정 모형을 활용하면, 질병부담지수의 동태적 변화를 다음과 같이 추정할 수 있다. 구체적으로, 건강검진의 동태적 변화 추정모형에 사용한 자료에 있는 표본들의 2년 뒤 건강검진지표들의 값을 예측하고, 예측된 건강검진지표들을 이용하여 질병부담지수를 예측하는 과정을 거쳤다. 이 과정을 5회 반복하여 10년 뒤의 건강검진지표와 질병부담지수까지 추정하였으며, 추정한 결과는 〈Figure 6〉~〈Figure 8〉과 같이 10년에 걸친 변화의 궤적으로 나타내었다. 아래 그림들은 성·연령·체질량지수 그룹에 따른 평균적인 변화의 추이이다. 체질량지수 그룹은 0기에 체질량지수가 저체중(BMI<18.5), 정상(18.5-25), 과체중(25-29.9), 비만(≥30)인지 여부로 구분하여 해당 구간에 속하는 표본들의 건강검진지표의 동태적 변화를 여 추정한 후 그룹별 평균값을 나타낸 결과이다. 연령대, 성별에 따라 검진지표나 질병부담지수의 변화 추이가 상이할 수 있음을 고려하여 연령대(20-39세, 40-59세, 60세 이상)와 성별로 결과를 제시한다. 체질량지수는 측정이 용이하며, 다른 검진지표들과 높은 상관성을 가지기 때문에 건강상태를 구분하는 용도로 사용하였다.

〈Figure 6〉 10-year Simulated Results of Waist Circumference



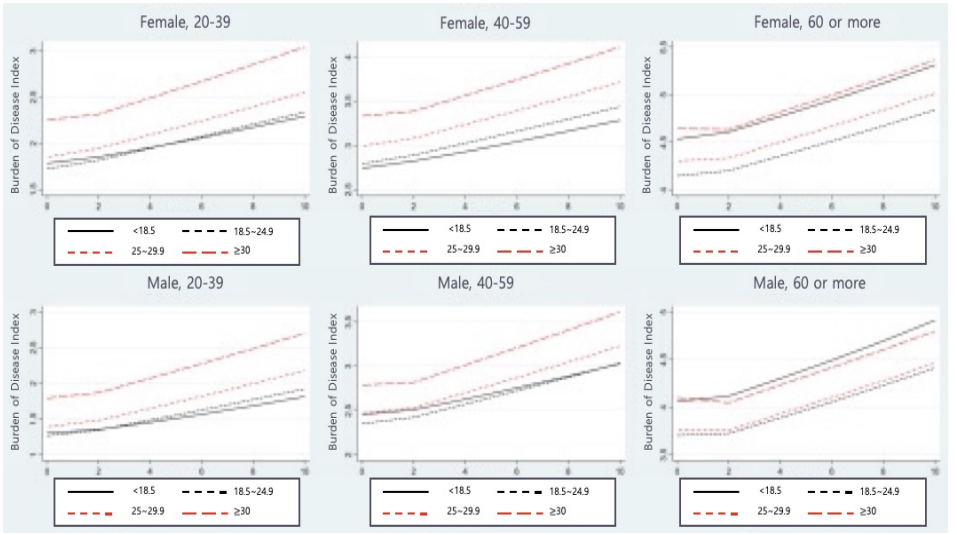
〈Figure 6〉는 허리둘레의 변화를 나타낸 것이다. 초기 건강 수준에 따라 허리둘레 수준이 시간이 지남에 따라 수렴하지 않고 그 격차가 유지되는 양상이 나타난다. 한편 〈Figure 7〉의 수축기혈압에서는 시점 경과에 따라 전반적으로 혈압이 증가하는 추이가 발견된다. 여기서는 초기 건강수준에 따라 혈압 수준이 발산하는 양상이 나타나며, 발산의 정도는 저연령대에서 비교적 강하고 증가의 추이는 고령 표본에서 가장 크게 나타났다. 다른 검진지표들의 경우에도 이와 비슷한 양상을 보인다.

〈Figure 7〉 10-year Simulated Results of Systolic Blood Pressure



이렇게 추정한 미래 검진지표를 활용하여 미래 질병부담지수의 변화를 예측한 결과는 〈Figure 8〉과 같다. 대체로 저체중 표본과 정상체중 표본의 질병부담지수 변화의 추이는 60세 미만에서는 거의 차이가 나타나지 않다가, 60세 이상 고령자의 경우 저체중 상태가 전반적인 건강 상태에 미치는 악영향이 과체중인 경우보다 더 심한 것으로 나타났다. 그 외 연령대에서는 과체중과 비만 표본의 질병부담지수가 가장 높게 추정되었으며, 건강 수준의 격차는 시간이 지남에 따라 점차 발산한다는 것을 확인할 수 있다.

〈Figure 8〉 10-year Simulated Results of the Burden of Disease Index



V. 의료비 추정 모형

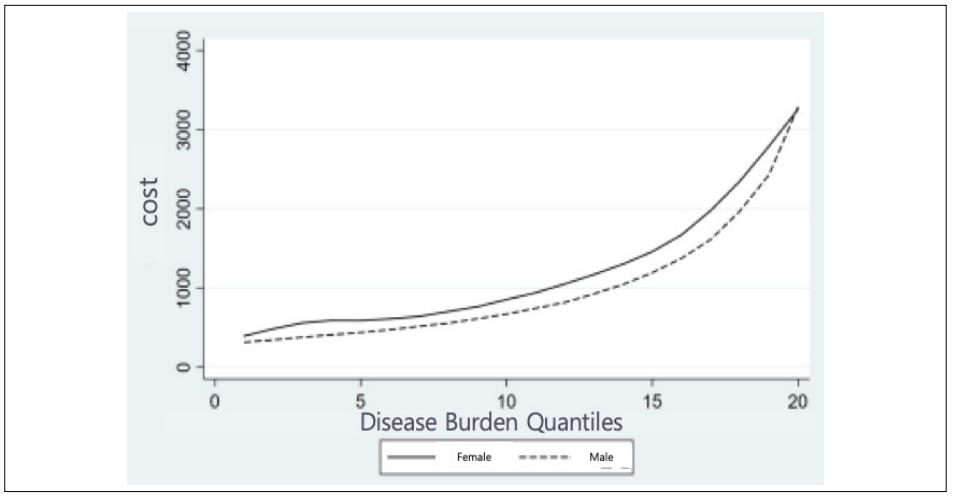
1. 질병부담지수와 의료비의 관계

질병부담지수는 개인이 실제로 진단받은 질환을 사용하여 산정되므로 반드시 의료이용을 수반한다는 점, 질환의 중증도가 질병부담지수와 의료비 부담 모두를 증가시킨다는 점에서 의료비와 밀접한 연관성을 가진다고 볼 수 있다. 본 연구에서 의료비는 건강보험공단 부담금과 개인 부담금의 합인 심결요양급여비용 총액을 사용했는데, 질환의 중증도나 사회경제적인 부담을 더 정확히 반영하기 위함이다. 실제 질병부담지수와 의료비 간 관계를 보면 〈Figure 9〉와 같다.

〈Figure 9〉는 개인의 실제 질병부담지수를 계산하여 20분위로 나눈 뒤, 분위별 평균 의료비 지출액을 나타낸 것이다. 남성의 경우 전반적으로 의료비 지출액이 동일한 질병부담지수 분위에 속하더라도 여성에 비해 더 적은 의료비를 지출하고 있음을 알 수 있다. 가령 여성 표본에서 질병부담이 가장 낮은 수준인 하위 5%는 여성의 경우 연간 약 39만 원, 남성의 경우 약 32만 원 의료비 지출이 있으며, 질병부담이 높은(건강수준이 가장 낮은) 상위 5%의 경우 여성의 경우 연간 약 326만 원, 남성의 경우 연간 약 329만 원의 의료비 지출액이 있다. 실제 질병부담이 높을수록

의료비 지출액은 증가하며, 증가 속도 역시 더 빨라지는 것이 확인된다.

〈Figure 9〉 Average Medical Cost by 20-Quantiles of Disease Burden



2. 질병부담지수가 의료비에 미치는 영향 분석 결과

앞서 검진기록의 변수가 미래 질병부담지수에 미치는 영향 분석에서는 인과적 관계보다는 예측력을 높이는데 집중하였지만, 질병부담지수가 의료비에 미치는 영향은 인과적 관계의 추정이 중요하다. 하지만 건강보험공단 빅데이터의 변수들의 유형이 의료적 변수로 제한적이므로 요인변수의 통제를 통해 인과성을 추정하는 것은 불가능하다. 대안적인 방법으로 아래 분석에서는 식 (4)와 같이 고정효과 패널 회귀분석을 활용하였다. 완벽하게 인과성을 추정할 수는 없지만 개인단위에서 질병부담지수가 개선될 때 의료비 절감 정도를 추정하는데 어느 정도 신뢰적인 방법으로 판단된다.

주요 설명변수는 앞 장에서 건강검진지표를 이용하여 추정된 질병부담지수 $(\widehat{BoD_{i,t}})$이며, 질병부담지수가 총체적인 건강상태를 반영하는 지표로 정의되었기 때문에 다른 건강수준과 유관한 요인은 사용하지 않았다. 즉 건강과 관련된 요인들은 질병부담지수 채널을 통해서만 의료비에 영향을 준다고 가정하였다. 이밖에 연령$(age_{i,t})$, 작년 의료비, 가구 지출 보험료$(x_{i,t})$를 통제변수로 활용하였다. 작년 의료비는 동일 건강상태에 대해서라도 개인의 특성에 따라 상이할 수 있는 의료이

용패턴을 통제하기 위해 사용되었다. 연령과 가구지출 보험료는 개인 및 가구의 사회경제적 요인들을 통제하기 위하여 사용되었다.

모든 의료비는 2020년도 보건의료부문 소비자물가지수를 적용하여 실질가치화하였고, 분석대상 의료비($c_{i,t}$)는 총 의료비, 입원 의료비, 외래 의료비, 약국 의료비로 분류하였다. 맞춤형 DB에는 치과 및 한방 의료비 항목도 포함되지만, 질병부담지수가 주로 병의원 방문을 통해 계산되며 검진지표가 주로 예측하고 관리하려는 대상 질환 대부분(예: 심뇌혈관질환)은 병의원 방문을 통한 치료가 주된 의료이용 행태가 되므로 분석 대상 의료비를 의과 및 보건기관 관련 의료비로 한정하였다. 한편, 성별, 연령별로 의료이용 패턴이 상이하다는 점을 고려하여, 모든 모형은 성별·연령 20세 그룹에 따라 표본을 나누어 분석하였다.

$$c_{i,t} = \alpha + \beta \widehat{BoD}_{i,t} + \gamma age_{i,t} + \delta c_{i,t-1} + \eta x_{i,t} + \mu_i + \epsilon_{i,t} \quad (4)$$

추정 결과는 아래 〈Table 10〉~〈Table 12〉와 같다. 우선, 총 의료비 기준으로 20-39세 남녀의 경우 추정 질병부담지수 1단위 증가에 따라 여성의 경우 약 9만 원, 남성의 경우 약 8만 원이 증가하는 것으로 추정되었다. 보다 고연령대에서 추정 질병부담지수의 한계효과는 40-59세 표본의 경우 여성 약 20만 원, 남성 약 16만 원, 60세 이상 표본에 대해서는 여성 약 35만 원, 남성 약 50만 원으로 추정되었다. 연령이 증가함에 따라 추정 질병부담지수의 한계효과가 커지며, 60세 미만 표본에서는 여성의 한계효과가 더 큰 것에 비해 60세 이상 표본에서는 남성의 한계효과가 더 크게 추정되었다. 연령 그룹이 고령자 그룹으로 갈수록 연령의 한계효과는 남녀 모두에서 점차 증가하였으며, 증가 속도는 남성 그룹에서 더 크게 나타났다.

전 기 의료비의 경우 의료비 항목에 따라 추정 결과가 상이하였는데, 평소의 의약품 복용이나 외래방문과 같이 어느 정도 습관에 영향을 받는 의료비의 경우 작년 의료비의 계수값이 크게, 입원과 같이 일회적인 사고나 예측하지 못한 중증질환이 발병함으로써 발생하는 의료비의 경우 작년 의료비의 계수값이 작거나 유의성이 떨어지는 방향으로 추정이 이뤄졌다.

〈Table 10〉 Medical cost prediction result (Aged 20~39)

		Total cost	Inpatient cost	Outpatient cost	Prescription cost
Female	Estimated burden of disease index	87300***	41870**	31658***	12944***
		(22480.413)	(21255.293)	(4769.057)	(1785.181)
	Age	34739***	13535***	16593***	1530***
		(1641.317)	(1511.149)	(534.480)	(187.942)
	Last year's cost	0.057***	-0.067***	0.261***	0.392***
		(0.010)	(0.009)	(0.028)	(0.044)
	Insurance fee	0.102**	0.112***	0.020	-0.006
		(0.045)	(0.039)	(0.014)	(0.006)
	Constant	-760973***	-349596***	-379361***	-1312***
		(18872.286)	(15384.370)	(8335.347)	(3132.733)
	Obs. No.	903515	903515	903515	903515
	R-sq. between	0.028	0.005	0.350	0.636
	R-sq. within	0.017	0.007	0.078	0.122
	R-sq. overall	0.024	0.000	0.234	0.500
Male	Estimated burden of disease index	79689***	32177***	22270***	21097***
		(7709.675)	(6020.054)	(2870.001)	(1354.329)
	Age	19146***	3888***	8393***	1817***
		(983.710)	(550.019)	(524.278)	(270.534)
	Last year's cost	0.152***	-0.026*	0.438***	0.597***
		(0.035)	(0.016)	(0.040)	(0.048)
	Insurance fee	0.097***	0.044**	0.032**	0.011**
		(0.030)	(0.021)	(0.014)	(0.005)
	Constant	-389877***	-68901***	-184401***	-41626***
		(17495.628)	(11472.760)	(8972.769)	(4665.751)
	Obs. No.	1217211	1217211	1217211	1217211
	R-sq. between	0.470	0.015	0.920	0.839
	R-sq. within	0.029	0.001	0.173	0.345
	R-sq. overall	0.378	0.003	0.827	0.707

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

〈Table 11〉 Medical cost prediction result (Aged 40~59)

		Total cost	Inpatient cost	Outpatient cost	Prescription cost
Female	Estimated burden of disease index	204923***	126056***	27442***	41946***
		(15901.040)	(11900.600)	(8562.668)	(2681.569)
	Age	49686***	12748***	23520***	7303***
		(1674.604)	(823.227)	(1523.106)	(585.321)
	Last year's cost	0.206***	0.075***	0.364***	0.499***
		(0.020)	(0.008)	(0.049)	(0.048)
	Insurance fee	0.026	0.047*	0.008	-0.011
		(0.038)	(0.028)	(0.016)	(0.008)
	Constant	-2224037***	-760542***	-929173***	-321915***
		(47512.286)	(23937.195)	(48704.835)	(20301.361)
	Obs. No.	1445847	1445847	1445847	1445847
	R-sq. between	0.222	0.035	0.509	0.736
	R-sq. within	0.046	0.007	0.103	0.181
	R-sq. overall	0.167	0.024	0.384	0.560
Male	Estimated burden of disease index	160295***	93245***	32613***	28787***
		(12956.483)	(10956.884)	(4010.858)	(2218.495)
	Age	52202***	16894***	15816***	9499***
		(1217.263)	(955.295)	(836.843)	(461.809)
	Last year's cost	-0.077*	-0.021*	-0.027**	-0.010*
		(0.040)	(0.033)	(0.012)	(0.007)
	Insurance fee	0.174***	0.046***	0.491***	0.493***
		(0.001)	(0.007)	(0.036)	(0.027)
	Constant	-2188242***	-828144***	-649737***	-369009***
		(39063.445)	(30134.022)	(27479.961)	(15981.480)
	Obs. No.	1619958	1619958	1619958	1619958
	R-sq. between	0.218	0.029	0.751	0.749
	R-sq. within	0.036	0.004	0.205	0.225
	R-sq. overall	0.158	0.016	0.618	0.632

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

〈Table 12〉 Medical cost prediction result (Aged 60 or more)

		Total cost	Inpatient cost	Outpatient cost	Prescription cost
Female	Estimated burden of disease index	353785***	276298***	23886***	45445***
		(24294.148)	(22297.407)	(5832.149)	(3504.084)
	Age	121285***	55698***	35357***	17472***
		(2311.338)	(2040.041)	(874.882)	(713.426)
	Last year's cost	0.190***	0.119***	0.400***	0.489***
		(0.007)	(0.008)	(0.020)	(0.029)
	Insurance fee	0.062	0.097	0.009	-0.037***
		(0.080)	(0.071)	(0.022)	(0.013)
	Constant	-7997568***	-4433985***	-2013975***	-974809***
		(110733.770)	(97125.082)	(43041.474)	(32741.148)
	Obs. No.	740608	740608	740608	740608
	R-sq. between	0.108	0.045	0.404	0.607
	R-sq. within	0.055	0.019	0.166	0.198
	R-sq. overall	0.092	0.030	0.344	0.531
Male	Estimated burden of disease index	495315***	370366***	86634***	41716***
		(28717.855)	(24990.950)	(9349.345)	(4402.445)
	Age	148467***	71495***	30887***	21587***
		(2690.717)	(2564.114)	(6349.692)	(893.973)
	Last year's cost	0.205***	0.045*	0.599***	0.529***
		(0.003)	(0.026)	(0.122)	(0.026)
	Insurance fee	-0.089	-0.047	-0.013	-0.023
		(0.088)	(0.075)	(0.030)	(0.015)
	Constant	-1.012e+07***	-5594269***	-2060147***	-1229653***
		(133004.864)	(133277.708)	(351380.874)	(41449.516)
	Obs. No.	681755	681755	681755	681755
	R-sq. between	0.116	0.017	0.589	0.650
	R-sq. within	0.089	0.015	0.285	0.227
	R-sq. overall	0.118	0.013	0.505	0.581

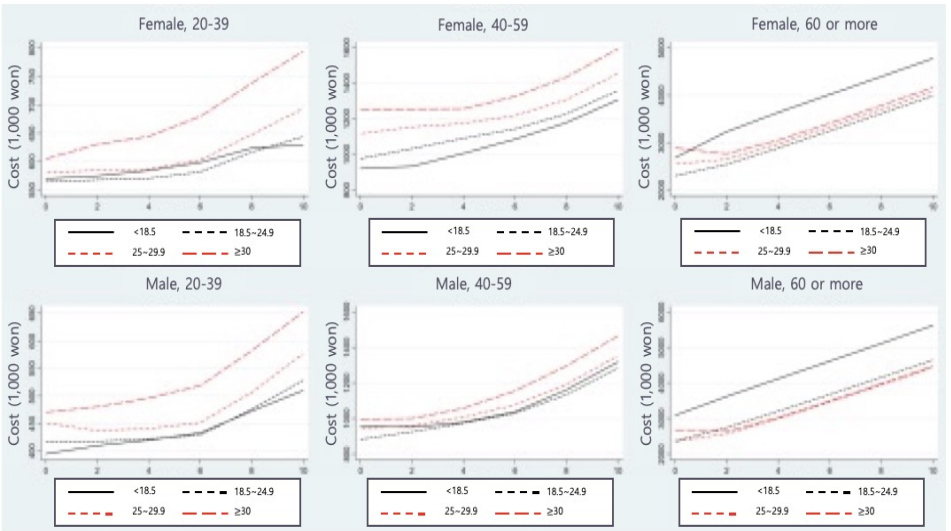
Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

3. 미래 의료비의 동태적 변화 추정

제Ⅲ장, 제Ⅳ장 2절과 3절에서 추정한 결과를 바탕으로 미래 의료비의 궤적을 동일한 방식으로 추정하였다. 구체적으로 제Ⅳ장 3절에서 추정한 질병부담지수를 활용하여 미래 의료비를 〈Figure 10〉과 같이 성·연령 20대 그룹·체질량지수 수준에 따라 나타내었다.

전반적으로 20~59세까지는 저체중과 정상 표본 간 격차가 불분명하지만 60세 이상 표본의 경우 저체중인 경우 의료비가 가장 높게 추정되었으며, 이는 앞서 질병부담 변화 추정 결과와 유사한 양상이다. 또 59세 이하까지는 남성 의료비 지출 수준이 여성에 비해 전반적으로 낮은 수준에 머물지만, 60세 이상이 되면서 미래 의료비 지출 수준이 여성을 상회하는 것으로 나타난다. 또한 의료비 지출수준은 초기 건강수준에 따라 그 격차가 좁혀지지 않고 유지되거나 발산하는 경향이 나타났다. 특히 20~39세 여성 표본의 경우 초기에는 거의 격차가 나타나지 않으나 경과시점이 지남에 따라 정상, 과체중, 비만 표본 간 격차가 점차 벌어지는 것을 확인할 수 있다.

〈Figure 10〉 10-year Simulated Results of Medical Cost



Ⅵ. 모형의 활용

1. 체질량지수 개선이 다른 검진지표에 미치는 영향 분석

제Ⅵ장에서는 앞에서의 논의를 바탕으로 건강관리사업의 효과성을 평가한다. 건강관리사업은 참가자의 체중, 허리둘레 등 주요 신체계측지표를 개선을 통해 주요 건강검진지표의 개선을 유도하는 방식으로 통상 이뤄진다.[16] 본 연구에서는 체중감량을 주요 목표로 하는 건강관리사업이 있다고 가정한 뒤, 체중감량 정도에 따른 주요 검진지표의 개선 정도를 추정하여 건강관리의 효과를 분석하였다.

보다 구체적으로, 체중감량을 통해 주요 건강검진지표의 개선이 있고, 이를 통해 추정 질병부담지수의 개선이 추정되며, 의료비 경감액이 추정되도록 구성하였다. 체중감량을 통해 개선되는 건강검진지표로는 허리둘레, 수축기혈압, 이완기혈압, 공복혈당, 총콜레스테롤, 고밀도콜레스테롤, 저밀도콜레스테롤, 중성지방을 선택하였다. 이들 지표는 대사증후군 여부를 결정짓는 주요한 건강검진지표로 사용되고 있으며, 체중감량과도 밀접한 연관성을 지니고 있다고 판단되어 분석에 활용하였다.

분석모형은 개인 고정효과를 활용한 모형으로 아래 식 (5)와 같다. 체중은 개인의 신장이나 성별 등에 영향을 받기 때문에 체중 대신 주요 설명변수로 체질량지수를 선택하였으며, 모형은 성별로 달리 분석되었다. 체질량지수나 주요 검진지표는 연령에 따라 영향을 받을 수 있기 때문에 연령의 2차식까지 통제변수로 활용되었다. 마지막으로 체질량지수 구간에 따라 체중감량의 한계효과가 상이할 것으로 보아 체질량지수 구간에 따라 범주화($Obesity_c$)하여 교호항을 만들었다. 체질량지수의 범주는 정상(<23), 과체중(23-24.9), 비만(25-29.9), 고도비만(≥30)으로 분류하였다.

$$v_{i,t} = \sum_c 1\{BMI_{i,t} \in Obesity_c\}\beta_c BMI_{i,t} + \gamma_1 age_{i,t} + \gamma_2 age_{i,t}^2 + \delta_i + \epsilon_{i,t} \quad (5)$$

16) 건강관리 프로그램의 효과성 평가는 박미숙 외(2017), 정혜선 외(2014), Kim, Meelim et al. (2020), Kim, Youngin et al. (2020), Toro-Ramos et al. (2017) 등에서 이루어졌다.

분석 결과는 아래 〈Table 13〉~〈Table 14〉와 같다. 분석 결과 체질량지수 증가에 따라 검진지표들이 악화하며, 악화의 정도는 고도비만 이상 구간을 제외하면 비만도가 높아짐에 따라 심화되는 것으로 나타났다. 다만 고도비만 이상인 표본은 그 수가 많지 않고, 건강행태에서도 다른 표본과 상이한 패턴을 보일 수 있다는 점에서 해석에 주의를 요한다. 또 공복혈당 외 모든 검진지표에서 남성 표본의 체질량지수 한 단위 변화에 따른 검진지표 변동 폭이 더 크게 추정되었다.

〈Table 13〉 Estimated changes in major check-up indicators according to changes in body mass index

	Waist circumference		Systolic blood pressure		Diastolic blood pressure		Fasting blood glucose	
	Female	Male	Female	Male	Female	Male	Female	Male
BMI<23	1.873***	1.961***	0.795***	1.245***	0.352***	0.756***	0.446***	0.285***
	(0.008)	(0.011)	(0.010)	(0.011)	(0.007)	(0.008)	(0.014)	(0.020)
23-24.9	1.885***	1.970***	0.809***	1.249***	0.362***	0.757***	0.445***	0.287***
	(0.008)	(0.010)	(0.009)	(0.011)	(0.006)	(0.008)	(0.013)	(0.019)
25-29.9	1.886***	1.976***	0.829***	1.256***	0.381***	0.766***	0.461***	0.294***
	(0.007)	(0.009)	(0.009)	(0.010)	(0.006)	(0.007)	(0.012)	(0.018)
≥30	1.864***	1.974***	0.846***	1.252***	0.405***	0.775***	0.502***	0.301***
	(0.007)	(0.009)	(0.008)	(0.010)	(0.006)	(0.007)	(0.012)	(0.018)
Age	0.157***	0.295***	0.014	-0.103***	0.328***	0.410***	0.311***	1.235***
	(0.004)	(0.004)	(0.010)	(0.009)	(0.007)	(0.007)	(0.013)	(0.016)
Age^2	-0.000***	-0.002***	0.002***	0.002***	-0.003***	-0.005***	0.002***	-0.005***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	25.400***	26.651***	93.022***	94.771***	56.296***	50.845***	63.397***	46.667***
	(0.161)	(0.183)	(0.277)	(0.280)	(0.203)	(0.203)	(0.355)	(0.421)
Obs. No.	3176479	3546115	3176479	3546115	3176479	3546115	3176479	3546115
R-sq. between	0.787	0.768	0.358	0.104	0.150	0.072	0.118	0.064
R-sq. within	0.235	0.322	0.017	0.018	0.008	0.017	0.019	0.021
R-sq. overall	0.690	0.692	0.241	0.071	0.100	0.052	0.087	0.052

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

〈Table 14〉 Estimated changes in major check-up indicators according to changes in body mass index (continued)

	Total cholesterol		HDL cholesterol		LDL cholesterol		Triglyceride	
	Female	Male	Female	Male	Female	Male	Female	Male
BMI<23	2.441***	3.186***	-0.809***	-1.152***	1.876***	2.006***	7.093***	12.641***
	(0.026)	(0.028)	(0.013)	(0.012)	(0.023)	(0.026)	(0.051)	(0.095)
23-24.9	2.498***	3.251***	-0.818***	-1.173***	1.915***	2.055***	7.235***	12.805***
	(0.024)	(0.026)	(0.012)	(0.011)	(0.022)	(0.024)	(0.048)	(0.089)
25-29.9	2.499***	3.268***	-0.822***	-1.184***	1.899***	2.043***	7.349***	13.043***
	(0.023)	(0.025)	(0.011)	(0.011)	(0.020)	(0.023)	(0.046)	(0.085)
≥30	2.375***	3.178***	-0.811***	-1.159***	1.789***	1.947***	7.228***	12.995***
	(0.022)	(0.024)	(0.011)	(0.010)	(0.019)	(0.022)	(0.044)	(0.082)
Age	4.493***	4.426***	0.726***	0.305***	3.352***	3.596***	1.336***	3.565***
	(0.023)	(0.021)	(0.014)	(0.011)	(0.022)	(0.022)	(0.040)	(0.059)
Age^2	-0.041***	-0.047***	-0.004***	-0.002***	-0.033***	-0.037***	-0.013***	-0.044***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Constant	25.087***	18.394***	55.204***	72.495***	-8.074***	-16.978***	-90.702***	-229.030***
	(0.666)	(0.651)	(0.414)	(0.316)	(0.632)	(0.650)	(1.263)	(2.074)
Obs. No.	3176479	3546115	3176479	3546115	3176479	3546115	3176479	3546115
R-sq. between	0.084	0.062	0.007	0.047	0.051	0.034	0.107	0.098
R-sq. within	0.039	0.058	0.006	0.011	0.022	0.027	0.038	0.052
R-sq. overall	0.068	0.054	0.001	0.036	0.048	0.029	0.088	0.079

Note: Standard errors in parentheses. * p<.1, ** p<.05, *** p<.01.

2. 건강관리 프로그램의 효과 추정

건강관리 프로그램의 효과를 추정하기 위하여 맞춤형 DB 2016년도 건강검진 수검 표본을 추출하여 분석에 활용하였다. 건강관리 프로그램은 체중 감량을 통해 대사증후군 위험요인을 관리하는 사업이라고 했을 때, 〈Table 15〉는 대사증후군의 정의와 프로그램에 참여할 수 있는 표본의 분포를 보여준다. 〈Table 15〉의 검진지표별 기준 중 하나라도 해당하면 프로그램 참여 대상이며, 참여 대상군 중 기준 1~2개 이상 해당하는 경우 주의군, 3개 이상 해당하는 경우 위험군으로 정의한다.

해당 정의는 건강보험공단의 대사증후군 관리사업의 대사증후군 정의를 활용하였다. 대사증후군 정의를 위해 활용하는 지표는 허리둘레, 혈압, 공복혈당, 고밀도콜레스테롤, 중성지방이다. 분석 표본에서는 이들 지표 중 공복혈당이나 혈압이 기준치를 초과하는 경우가 가장 높은 빈도를 보였으며, 고밀도콜레스테롤이 기준치 미만인 경우가 가장 적었다. 전반적으로 모든 지표가 정상 범위 내로 측정된 정상군의 비중은 전체의 약 27%였으며, 1개 이상의 지표가 기준에 해당하여 프로그램에 참여할 수 있는 표본은 73%이며 이 중 과반인 51%가 대사증후군 주의군, 22%가 위험군으로 분류되었다.

〈Table 15〉 Diagnostic Criteria for Metabolic Syndrome

	Indicators and Stages	Criteria	Proportion of samples(%)
Check-up Indicators	Waist Circumference(cm)	Male≥90, Female≥85	23.85%
	Blood Pressure(mmHg)	SBP≥130 or DBP≥85	36.86%
	Fasting Blood Glucose(mg/dL)	≥100	37.35%
	HDL cholesterol(mg/dL)	Male<40, Female<50	22.45%
	Triglyceride(mg/dL)	≥150	28.02%
Stages	Normal	None of the above criteria	27.08%
	Pre-metabolic syndrome	Meets 1~2 of the above criteria	51.04%
	Metabolic syndrome	Meets 3 or more of the above criteria	21.87%

건강관리 프로그램은 전술한 바와 같이 체중감량을 통해 이뤄지며, 사업 참가자들이 10%의 체중을 감량하였다고 가정하자. 또 분석 편의상 해당 사업은 일회성 사업이라 가정하였다. 즉, 참여 표본들은 1회에 한해서 즉시적인 10%의 체중감량이 있고, 그 이후로는 평소와 같은 정도로만 건강관리 행태를 보인다고 가정하였다.

프로그램 참여를 통한 질병부담지수 및 의료비 절감 효과를 추정한 결과는 〈Table 16〉와 같이 나타낼 수 있다. 구체적으로 사업 효과는 다음 과정을 통해 추정되었다. 우선 체중 10% 감량으로 다른 건강검진지표가 개선되는 효과는 제Ⅵ장 1절의 분석결과를 활용하여 추정하였다. 이렇게 개선된 건강검진지표를 제Ⅳ장의 건강검진지표 동태적 변화 추정 모형에 대입하여 향후 10년 검진지표의 변화를 추정하였고, 이를 이용하여 추정 질병부담지수 및 추정 의료비를 구하였다. 한편 주의군이나 위험군임에도 프로그램에 참여하지 않는 대조군을 가정하여 마찬가지 과

정을 통해 검진지표 및 질병부담지수, 의료비의 동태적 변화를 추정하여 비교하였다.

분석 결과 불참한 표본에 비해 참여 표본의 질병부담지수에 개선이 있었으며, 주의군에 비해 위험군일수록 개선의 효과가 크게 나타났다. 개선의 효과는 남성 표본에서 전반적으로 더 크게 추정되었으며, 남녀 모두 저연령대일수록 프로그램의 참여로 인한 건강 개선의 정도가 더 높았다. 대사증후군 위험군인 경우 남녀 20-39세 표본이 각각 19.1%, 18.1%으로 가장 높은 개선 효과를 보였으며, 효과가 가장 낮

〈Table 16〉 Cumulative Changes in Estimated Medical Expenses According to Participation in the Healthcare Program

			BoD improvement rate	Non-Participants (A)	Participants (B)	Difference (C=A-B)	Rate (C/A)
Female	20~39	Normal	-	7,291,056	-	-	-
		PreMet	6.5%	7,752,581	7,422,308	330,273	3.8%
		Met	18.1%	9,167,022	7,987,432	1,179,590	11.8%
	40~59	Normal	-	13,521,470	-	-	-
		PreMet	3.6%	15,194,386	14,683,081	511,305	3.0%
		Met	11.4%	17,660,875	15,819,607	1,841,268	9.9%
	≥60	Normal	-	28,990,555	-	-	-
		PreMet	1.8%	33,193,948	32,715,463	478,485	1.5%
		Met	7.4%	37,401,353	35,420,830	1,980,523	5.6%
Male	20~39	Normal	-	5,751,887	-	-	-
		PreMet	9.0%	6,503,542	6,114,021	389,520	5.4%
		Met	19.1%	7,836,051	6,763,286	1,072,765	12.6%
	40~59	Normal	-	12,609,650	-	-	-
		PreMet	5.3%	14,349,652	13,719,127	630,525	3.9%
		Met	13.9%	16,596,127	14,619,846	1,976,281	10.9%
	≥60	Normal	-	34,340,305	-	-	-
		PreMet	2.7%	37,332,724	36,419,242	913,482	2.5%
		Met	9.9%	40,788,866	37,303,273	3,485,593	8.9%

Note: The improvement rate of the disease burden index refers to the cumulative sum of the estimated dynamic changes in the disease burden index, indicating the extent of improvement. Medical expenses represent the cumulative sum of the estimated medical expenses over 10 years. The presence of metabolic syndrome and age group are defined based on the program's starting point. Abbreviation PreMet refers to pre-metabolic syndrome patients, Met to metabolic syndrome patients.

았던 그룹은 여성과 남성 모두 각각 60세 이상 주의군으로, 각각 2.7%, 1.8%의 건강 개선 효과가 추정되었다. 고령자의 경우 대사증후군 위험요인 관리를 통한 경증 질환에 대한 건강관리보다는 보다 중증도 위험요인을 통제하고 예방하는 사업이 보다 효과적인 반면, 60세 미만 연령대에서는 체중감량을 통한 대사증후군 지표 관리와 같은 비교적 경증 질환에 대한 건강관리가 효과적일 수 있음을 암시한다.

개선된 질병부담지수를 활용하여 의료비 절감액을 추정한 결과도 마찬가지의 양상을 보인다. 남녀 모두 위험군일수록, 저연령일수록 개선 비율이 가장 크게 추정되었다. 특히 남성 20-39세 대사증후군 위험군의 경우 의료비 절감 효과가 12.6%로 가장 크게 추정되었다. 한편 고령 표본의 경우 절감의 비율 자체는 크지 않은 편이었으나 절감액은 가장 높은 수준으로 추정되었다. 60세 이상의 경우 여성 위험군 표본의 경우 1인당 약 198만 원, 남성 위험군 표본의 경우 인당 약 348만 원으로 절감되는 액수가 추정되었다.

Ⅶ. 강건성 분석

1. 질병부담지수의 하방 편의 발생 가능성

본 연구에서 개발한 질병부담지수의 성·연령별 분포를 보면, 75세 미만에서는 남성 표본의 질병부담이 여성에 비해 낮은 수준에 머물러있음이 확인된다. 고연령대에서는 남성 사망률이 높아짐에 따라 남성의 질병부담이 증가하는 것으로 보이나, 75세 미만에서 남성 표본의 질병부담이 여성에 비해 낮게 추정되는 것에 하방편의(downward bias)가 존재하는지 두 가지 측면에서 살펴보았다.

우선, 자료 한계상 질병부담지수는 개인이 의료기관에서 진단받은 코드를 통해 측정되는데, 의료기관 방문에는 의료적 필요성 외에도 사회경제적 배경 등 다양한 요인들이 영향을 미칠 수 있어 질병부담지수 산출에 편의가 발생할 수 있다. 가령 근로형태나 지리적인 이유로 의료기관 방문이 어려운 표본은 경증 질환으로 의료기관 방문을 하는 경우가 상대적으로 드물 수 있다.[17]

이렇게 의료기관 방문이 건강 외적 요인으로 인해 영향을 받는 효과가 존재하는

17) 이밖에 정신과 질환 등 사회문화적 인식으로 인해 의료이용이 실제 의료적, 사회적 필요성에 비해 과소하게 일어나는 질환들이 존재할 수 있다.

지 확인하기 위해 다음과 같은 작업을 수행하였다. 우선 중증질환은 의료적 필요성 외 요인에 의해 의료기관 방문이 제한받는 경향이 적다고 보아 중증질환만을 활용하여 질병부담지수를 재산출한 뒤 성별 편차가 발생하는지를 확인하였다. 여기서 중증질환은 질병 가중치가 0.8 이상인 질환, 0.75 이상인 질환으로 정의하여 해당 상병 진단만으로 질병부담지수를 산출하여 성 및 연령별 분포가 어떻게 달라지는지 보았다. 단 중증질환만으로 질병부담지수를 산출하는 경우 사망한 경우는 제외하였는데, 사망은 의료기관 방문과 무관히 정의될 수 있기 때문이다. 〈Table 17〉은 중증질환 정의에 사용된 질환과 정의, 질병 가중치의 목록이다.

〈Table 17〉 List of Severe Diseases and weights

#	Disease	ICD	Weight
1	Pancreatic cancer	C25	0.929
2	Brain and nervous system cancer	C70-72	0.882
3	Oesophageal cancer	C15	0.870
4	Larynx cancer	C32	0.848
5	Nasopharynx cancer	C11	0.847
6	Acute myeloid leukaemia	C92.0, C92.6, C92.8, C94.0, C94.2	0.830
7	Acute lymphoid leukaemia	C91.0	0.827
8	Other leukaemia	C91-C95	0.823
9	Ovarian cancer	C56	0.821
10	Gallbladder and biliary tract cancer	C23-C24	0.816
11	Malignant skin melanoma	C43	0.807
12	Hemorrhagic stroke	I60-I62	0.800
13	Bladder cancer	C67	0.787
14	Other pharynx cancer	C09-C10, C12-C14	0.777
15	Ebola virus disease	A98.4	0.774
16	Testicular cancer	C62	0.772
17	Kidney cancer	C64-66	0.771
18	Mesothelioma	C45	0.766
19	Chronic myeloid leukaemia	C92.1, C92.2	0.764
20	HIV/AIDS resulting in other diseases	B20-B24	0.756
21	Chronic obstructive pulmonary disease (severe)	J44.02, J44.12, J44.82, J44.92	0.753
22	Chronic lymphoid leukaemia	C91.1	0.752

Note: The selected diseases were limited to those with a weight of 0.75 or higher. Other leukaemia excludes patients diagnosed with ICD C91.0, C91.1, C92.0-C92.2, C92.6, C92.8, C94.0, C94.2, and HIV/AIDS resulting in other diseases excludes patients diagnosed with B20.0, U84.30, U84.31.

〈Table 18〉 Average disease burden index for overall and severe diseases by gender and age group

Age Group	All		DW≥.75		DW≥.80	
	Female	Male	Female	Male	Female	Male
20~24	1.690	1.006	0.001	0.001	0.001	0.001
	(0.843)	(0.776)	(0.032)	(0.034)	(0.029)	(0.026)
25~29	1.730	1.085	0.002	0.001	0.002	0.001
	(0.875)	(0.776)	(0.039)	(0.033)	(0.035)	(0.025)
30~34	1.770	1.227	0.002	0.002	0.002	0.001
	(0.981)	(0.868)	(0.038)	(0.045)	(0.036)	(0.031)
35~39	1.814	1.366	0.003	0.003	0.002	0.002
	(1.069)	(1.007)	(0.042)	(0.049)	(0.039)	(0.037)
40~44	2.064	1.596	0.004	0.004	0.003	0.003
	(1.196)	(1.186)	(0.057)	(0.062)	(0.050)	(0.048)
45~49	2.469	1.928	0.006	0.006	0.005	0.004
	(1.362)	(1.408)	(0.071)	(0.077)	(0.065)	(0.058)
50~54	2.895	2.361	0.007	0.010	0.006	0.006
	(1.519)	(1.642)	(0.081)	(0.095)	(0.074)	(0.076)
55~59	3.336	2.869	0.010	0.014	0.008	0.009
	(1.669)	(1.839)	(0.093)	(0.114)	(0.086)	(0.088)
60~64	3.872	3.467	0.011	0.019	0.009	0.011
	(1.816)	(2.117)	(0.100)	(0.126)	(0.089)	(0.094)
65~69	4.359	4.093	0.013	0.024	0.010	0.012
	(1.966)	(2.384)	(0.102)	(0.138)	(0.090)	(0.099)
70~74	4.651	4.603	0.013	0.026	0.010	0.012
	(2.093)	(2.687)	(0.100)	(0.143)	(0.089)	(0.091)
75~79	4.697	5.112	0.011	0.028	0.008	0.012
	(2.302)	(3.058)	(0.083)	(0.134)	(0.073)	(0.088)
≥80	4.863	5.528	0.009	0.026	0.006	0.008
	(3.043)	(3.475)	(0.068)	(0.140)	(0.053)	(0.071)
Total	2.668	2.087	0.006	0.007	0.005	0.004
	(1.805)	(1.912)	(0.068)	(0.080)	(0.061)	(0.059)
Obs. No.	738,455	715,535	705,293	671,825	705,293	671,825

Note: Standard deviations in parentheses. DW≥0.75 and DW≥0.80 represents the disease burden index derived from diseases with a disease weight (DW) of 0.75 or higher and 0.80 or higher, respectively. When calculating the average value of the disease burden index for severe diseases, samples of deceased individuals were excluded.

중증질환만을 사용하여 산출한 중증질환 질병부담지수와 전체 질환을 사용하여 산출한 질병부담지수의 분포를 성·연령그룹별로 나타낸 결과는 〈Table 18〉과 같다.

전체 질환이 아닌 중증질환만을 사용하여 만든 질병부담지수의 경우, 낮은 연령대에서는 성별에 따른 격차가 거의 발생하지 않으며, 고령자 표본에서 남성 질병부담이 여성 질병부담을 넘어서는 경향을 보인다. 결과적으로 질병부담지수 산출과 관련하여 의료이용 경향으로 인한 편의는 주로 경증질환에서 기인한다고 볼 수 있으며, 건강상태 및 전반적인 삶에 질에 보다 큰 영향을 미치는 중증질환 및 사망률의 경우 해당 편의가 발생하지 않는다는 점에서 전체 질환 대상 질병부담지수를 분석에 활용하는 것에 큰 무리가 없다고 보인다.

두 번째로 저연령대에서 발생한 성별 질병부담지수의 격차는 여성의 임신 및 출산, 산후기 질환으로 인한 질병부담에서 기인했을 수 있다. 〈Table 19〉는 질병가중치 산정 대상 질환들 중 이러한 모성(maternal) 질환들의 목록과 가중치이며, 해당 질환들만을 사용하여 산출한 여성들의 연령대별 평균 질병부담지수는 20~24세에서 0.088, 25~29세에서 0.119, 30~34세에서 0.075이었다. 연령대가 낮을수록 전반적인 질병 부담이 낮다는 점을 고려했을 때 모성 질환들로 인한 여성 질병 부담 증가는 의료기관 방문의 편의에서 비롯된 것이 아닌, 성별 질병 부담 분포 차이를 유발한 원인 중 하나로 지적될 수 있을 것이다.

〈Table 19〉 Diseases for Pregnancy, Childbirth and the Puerperium

#	Disease	KCD	Weight
1	Maternal haemorrhage	O20, O46, O67, O72	0.599
2	Maternal sepsis and other pregnancy related infections	O75.3, O85	0.643
3	Maternal hypertensive disorders	O10-O16	0.410
4	Maternal obstructed labour and uterine rupture	O71.0, O71.1	0.668
5	Maternal abortion, miscarriage, and ectopic pregnancy	O00-O08	0.379
6	Other maternal disorders	O21-O29	0.387

Note: Selected diseases are those can be defined only by Claims DB 20 table KCD codes.

2. 장기적 건강수준의 영향

건강검진지표나 의료비를 추정할 때 현재의 건강상태 외에도 과거의 건강상태가

장기적으로 영향을 미칠 수 있다. 과거의 위험요인들이 미래 건강검진지표 및 현재 의료비에 영향을 미치는지 여부를 분석하기 위해 기존의 모형에 시차변수를 추가하여 분석한 결과는 이하와 같다. 우선 검진지표의 동태적 변화 추정을 위해서 식 (6)을 활용하여 분석하였다.

$$\Delta v_{i,t+2} = X_{i,t}^{-v} B_1 + X_{i,t-2}^{-v} B_2 + \gamma_1 age_{i,t} + \gamma_2 age_{i,t}^2 + \epsilon_{i,t} \quad (6)$$

기존 모형에서와 마찬가지로 시차변수는 2년 전 변수를($X_{i,t-2}^{-v}$) 활용하였다. 매년 검진받는 표본을 분석 대상으로 상정하는 경우 분석 표본이 비사무직군으로 한정되어 편의가 발생할 수 있으며, 표본 수도 크게 감소하기 때문이다.

〈Table 20〉 Comparison of Female BMI Estimation Results

	(1)	(2)	Time
Waist Circumference	-	0.005***	t-2
	-	(0.000)	
Diastolic blood pressure	0.002***	0.001***	t-2
	(0.000)	(0.000)	
Hemoglobin	-	-0.023***	t-2
	-	(0.001)	
Serum creatinine	0.016***	0.007***	t
	(0.002)	(0.002)	
Proteinuria	0.043***	0.032***	t-2
	(0.007)	(0.007)	
Smoking	0.095***	0.069***	-
	(0.009)	(0.009)	
Risky Drinking	0.010***	-	-
	(0.003)	-	
Age	0.009***	-0.012***	-
	(0.000)	(0.001)	
Age^2	-0.000***	0.000***	-
	(0.000)	(0.000)	
Obs. No.	911025	911025	
Adj. R-squared	0.025	0.028	
RMSE	1.265	1.263	

Note: Standard errors in parentheses. (1) represents the results estimated using only the screening indicators from the current year, while (2) represents the results estimated using both the screening indicators from the current year and two years prior. The notation "t-2" indicates that Model (2) includes the screening indicators from two years prior, while "t" represents the screening indicators from the current year. *p<.1, **p<.05, ***p<.01.

분석 결과 중 여성 표본 체질량지수 추정 결과를 〈Table 20〉과 같이 나타내었다. 모형 (1)은 당해 연도 검진결과만을 사용하여 다음 기(2년 뒤) 체질량지수 차분값을 추정한 결과이며, (2)는 당해 연도와 2년 전 검진결과 모두를 사용하여 추정한 결과이다. 다만, 설문 항목인 흡연, 음주, 신체활동 변수는 당해 연도 설문 결과만을 활용하였다. 모든 분석은 2년 전 검진결과가 존재하는 표본으로 진행하였다. 분석 결과 전반적으로 일부 변수를 제외하면 기존 모형에서 선택된 위험요인이 선택되는 경향이 있었으나 허리둘레, 혈색소 등이 새롭게 선택되었고, 음주여부는 추정계수의 유의성이 떨어져 분석에서 제외되었다. 이밖에 Adjusted R-squared나 RMSE와 같이 모형의 전반적인 예측력을 비교했을 때 큰 개선이 나타나지는 않았다. 다만 해당 모형을 추정하기 위해 격년으로 3회 연속 검진받은 표본을 추출하는 과정에서 표본 수가 상당수 감소하였고, 연속으로 검진을 수검한 표본과 그렇지 않은 표본 간 편의가 존재할 가능성이 있어 위 결과를 해석함에 있어 주의를 요한다.

의료비의 경우에도 마찬가지로 기존 모형에 시차항을 추가하여 식 (7)과 같이 추정하였다. 추정 질병부담지수의 시차변수로 $\widehat{BoD_{i,t-2}}$가 추가되었으며, 그 외의 통제변수는 연령, 가구 보험료, 1년 전 의료비, 개인고정효과로 기존 모형과 같다.

$$c_{i,t} = \alpha + \beta_1 \widehat{BoD_{i,t}} + \beta_2 \widehat{BoD_{i,t-2}} + \gamma age_{i,t} + \delta c_{i,t-1} + \eta x_{i,t} + \mu_i + \epsilon_{i,t} \quad (7)$$

추정 결과는 〈Table 21〉과 같다. 본고에서는 총의료비 추정 결과에 대해서 논의하며, 항목별 의료비(입원, 외래, 약국조제 의료비)의 추정 결과는 총의료비 추정 결과와 유사하였다. 우선 시차항을 추가하면서 표본의 탈락이 일어나 기존 결과와 계수값이 상이해졌으며, 시차항 없는 모형에 비해 전반적인 예측력이 크게 개선되지는 않았다. 또, 추가된 시차항의 계수는 대부분의 경우 양의 값을 가졌으며, 이번 기 추정 질병부담지수의 계수보다 크기가 작았다. 즉 과거의 높은 질병부담이 장기적으로 현재 의료비 부담에 영향을 미치지만, 그 크기가 현재의 질병부담에 비해 작다고 볼 수 있다.

다만 앞서와 마찬가지로 연달아 검진받은 개인을 추출하는 과정에서 분석 표본의 수가 상당히 감소하여 그 과정에서 편의가 발생했을 수 있고, 연령과 질병부담, 질병부담의 시차항 등 변수들 간 높은 상관관계로 인한 공선성 문제를 고려하면 위의 결과는 제한적으로 해석할 필요가 있다고 보인다.

〈Table 21〉 Medical Cost Estimation Results Using the Lagged Variable

		Aged 20-39		Aged 40-59		Aged 60 or more	
		(1)	(2)	(1)	(2)	(1)	(2)
Female	Estimated burden of disease index	59706**	62013**	192812***	221783***	347587***	390353***
		(23740)	(24141)	(22006)	(22614)	(36339)	(38076)
	Lagged term		13359		130323***		147526***
			(25377)		(23444)		(39246)
	Age	77058***	76014***	85748***	75519***	193508***	179368***
		(3198)	(3763)	(2516)	(3117)	(4814)	(6109)
	Last year's cost	-0.119***	-0.119***	0.124***	0.123***	0.136***	0.136***
		(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
	Insurance fee	0.552***	0.552***	-0.022	-0.023	0.343**	0.342**
		(0.166)	(0.166)	(0.092)	(0.092)	(0.142)	(0.142)
	Constant	-2007682***	-1999271***	-3860000***	-3791138***	-1.265e+07***	-1.246e+07***
		(83586)	(85100)	(103368)	(104102)	(279578)	(284366)
	Obs. No.	168321	168321	477819	477819	299110	299110
	R-sq. between	0.023	0.023	0.019	0.019	0.037	0.037
	R-sq. within	0.015	0.015	0.106	0.108	0.062	0.064
	R-sq. overall	0.005	0.005	0.084	0.086	0.055	0.057
Male	Estimated burden of disease index	70151***	78964***	129015***	151572***	470721***	539270***
		(13066)	(13166)	(15047)	(15304)	(39982)	(41564)
	Lagged term		773212***		129493***		258643***
			(14306)		(16095)		(42921)
	Age	35812***	29354***	81273***	70171***	241203***	217127***
		(1853)	(2204)	(2139)	(2546)	(5554)	(6842)
	Last year's cost	0.021***	0.021***	0.078***	0.078***	0.160***	0.160***
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
	Insurance fee	0.163	0.166	0.029	0.030	-0.326*	-0.325*
		(0.106)	(0.106)	(0.084)	(0.084)	(0.167)	(0.167)
	Constant	-8887537***	-777801***	-3460164***	-3262072***	-97396***	-1.566e+07***
		(49142)	(53254)	(89001)	(92335)	(325789)	(334968)
	Obs. No.	315891	315891	572054	572054	278805	278805
	R-sq. between	0.006	0.007	0.014	0.014	0.072	0.072
	R-sq. within	0.037	0.038	0.074	0.079	0.085	0.089
	R-sq. overall	0.037	0.038	0.060	0.064	0.084	0.088

Note: Standard errors in parentheses. (1) refers to the original model, and (2) refers to the model using the lagged variable. * p<.1, ** p<.05, *** p<.01.

Ⅷ. 요약 및 결론

개인의 건강 수준을 측정하고 건강수준 악화 또는 개선으로 인한 개인적, 사회적 비용을 측정하는 것은 만성질환 관리에 있어서 효율적인 의사선택을 위한 주요한 근거자료가 될 수 있다. 본 연구에서는 대표성있는 빅데이터인 건강보험공단 맞춤형 DB를 활용하여 개인 건강 수준을 비교적 경증 질환에서 중증질환, 사망까지 고려하는 총체적 건강 지표로서 질병부담지수를 개발하고, 미래의 누적 질병부담지수를 건강검진지표로 예측하는 작업을 수행하였으며, 누적 질병부담지수로 의료비를 추정하여 질병부담이 높은 의료비와 밀접한 상관성을 갖는다는 것을 보였다. 그리고 검진지표의 동태적 변화 추정을 통해 질병부담과, 질병부담 변화로 인한 의료비의 추이를 보였다. 마지막으로 가상의 건강관리 프로그램 하에서 프로그램 참여를 통한 건강 개선이 질병부담 및 의료비 절감을 가져올 수 있다는 것을 정량적으로 제시할 수 있었다.

연구의 의의와 발전방향에 논의하기 앞서 본 연구의 한계를 몇 가지 언급하고 의의와 발전방안을 논의하고자 한다. 우선, 가중치 산정 과정의 한계가 지적될 수 있겠다. 예컨대 제공받은 맞춤형 DB만으로는 암과 뇌졸중 등 일부 주요 질환자들에 대해서 중증도를 식별할 수 없었다. 추가 테이블이나 외부 자료 등을 활용하여 건강 상태를 보다 정확히 나타내는 지표가 만들어질 수 있을 것이라 보인다.[18]

다양한 머신러닝 모형을 활용하였으나, 모형의 해석과 설명력 측면에서 가장 정확도가 높은 LGBM을 활용하지 못했다는 점 역시 추후 보완할 사항 중 하나로 지적해야 할 것이다. Shapley value 등 변수 중요도를 계산하는 방법론 등을 활용하여 후속 연구에서는 모형 적합도가 높으면서도 해석 가능한 건강 평가 모형을 개발하고자 한다.

의료비 추정 역시 급여 항목에만 한정되어 진행되었다는 한계점이 있는데, 건강보험공단 맞춤형 DB에는 비급여 진료비가 제공되지 않기 때문이다. 추후 연구에서는 비급여 진료비나, 간접의료비, 질환으로 인한 소득 상실 등 비용 측면에서 보다 다각적인 분석이 추가되어 본 연구 결과를 보완할 수 있을 것이라 기대한다.

---

18) 건강보험공단 진료DB 30테이블의 수가코드를 사용하여 치료행위를 분석하여 암의 병기 구분을 조작적으로 정의하거나, 국립암센터 자료 등 병기 구분이 제공되는 자료의 활용을 통해 가중치를 보다 세밀히 나눌 수 있을 것이다.

이러한 한계점에도 불구하고 본 연구에서 개발된 종합적 건강지표 추정 모델, 건강 수준의 동태적 변화 분석 모델, 의료비 예측모델은 개인의 포괄적 건강수준을 측정하고 건강수준의 개선이 사회적 비용 절감으로 이어질 수 있음을 정량적으로 보였다는 의의가 있다. 개발 모형들을 통해 건강관리사업이 어느 집단에 더 효과적일 수 있는지, 사업의 중장기적 비용과 편익이 어떻게 변화하는지 사전에 파악하고 더 효율적인 사업이 무엇인지 선택하는데 있어 근거가 될 수 있을 것으로 기대된다.

■ 참 고 문 헌

1. 건강보험공단・건강보험심사평가원, 『2021 건강보험통계연보』, 2022.

   (Translated in English) National Health Insurance Service, *2021 National Health Insurance Statistical Yearbook*, Seoul, Korea: National Health Insurance Service; 2022.

2. 건강보험공단, 『대사증후군 관리 서비스로 건강한 생활 습관 만드세요!』, 〔웹사이트〕, (2022. 12. 15.) URL:https://www.nhis.or.kr/static/alim/paper/oldpaper/202102/sub/s04_01.html.

   (Translated in English) National Health Insurance Service, *Create a Healthy Lifestyle with Metabolic Syndrome Management Service!* 〔Website〕, (2023, Jan) https://www.nhis.or.kr/static/alim/paper/oldpaper/ 202102/sub/s04_01.html.

3. 건강보험심사평가원, 『급여의약품・치료재료청구현황』, 2022.

   (Translated in English) Health Insurance Review & Assessment Service, *Reimbursment of Drug and Materials for Medical Treatment Status*, 2022.

4. 김준경・김준일, "건강보험제도의 도입과 발전과정: 정치경제적 배경과 거시경제적 고찰," 『한국경제포럼』, 제14권, 제3호, 2021, pp. 1-46.

   (Translated in English) Kim, Joon-Kyung and Jun Il Kim, "Development of Korea's National Health Insurance System: Political Economy and Macroeconomic Considerations," *The Korean Economic Forum*, Vol. 14, No. 3, 2021, pp. 1-46.

5. 박미숙・박윤숙・김선영・박수진・설혜민・우선옥・조수경・임도선, "서울시 대사증후군 관리사업의 현황 및 효과," 『대한공공의학회지』, 제1권, 제1호, 2017, pp. 25-39.

   (Translated in English) Park, MiSuk, et al. "Introduction and Effectiveness of The Seoul Metabolic Syndrome Management," *Public Health Affairs*, Vol. 1, No. 1, 2017, pp. 25-39.

6. 보건복지부, 『2020 국민보건계정』, 2021.

(Translated in English) Ministry of Health and Welfare, *Korean National Health Accounts in 2020*, 2021.

7. _____, "스스로 건강관리, 이제 국가가 지원합니다," 보건복지부 보도자료, 2021.

(Translated in English) Ministry of Health and Welfare, Self-Health Management: Now Supported by the Government [Press release], 2021.

8. 선정연 · 김기영 · 김진휘, 『이상치 탐색을 위한 통계적 방법과 활용 방안』, 건강보험심사평가원, 2019.

(Translated in English) Seon, J. Y., K. Kim, and J. Kim, *Statistical Methods for Outlier Detection and Their Applications*, The Health Insurance Review and Assessment Service, 2019.

9. 정혜선 · 이복임 · 권영현 · 민규리 · 명수영, "U-Health 프로그램을 이용한 직장인 대사증후군 관리사업의 효과," 『한국직업건강간호학회지』, 제23권, 제1호, 2014, pp. 47-54.

(Translated in English) Jung, H. S., B. Lee, Y. H. Kwon, K. R. Min, and S. Y. Myung, "The Effects of U-health Program on Metabolic Syndrome of Workers," *Korean Journal of Occupational Health Nursing*, Vol. 23, No. 1, 2014, pp. 47-54.

10. 질병관리청, 『2021 만성질환 현황과 이슈』, 2021.

(Translated in English) Jung EK, *2021 Chronic Disease Status and Issues; Chronic Disease Fact Book*, Cheongju: Korea Centers for Disease Control and Prevention, 2021.

11. 통계청, 『2021 고령자 통계』, 2021.

(Translated in English) Statistics Korea, *2021 Statistics on the Aged*, Daejeon, Korea: Statistics Korea, 2021.

12. _____, 『2021년 사망원인통계 결과』, 2022.

(Translated in English) Korean Statistical Information Service, *Annual Report on the Cause of Death Statistics*, Daejeon, Korea: Statistics Korea, 2022.

13. 홍석철 외, 『당뇨환자의 당뇨합병증 발생률과 사망률 예측 및 의료비 추정 모형 개발』, 보험개발원, 2017.

(Translated in English) Hong, S. C., et al., *Development of a Predictive Model for the Incidence and Mortality Rates of Diabetes Complications and Estimation of Medical Expenses in Diabetes Patients*, Korea Insurance Development Institute, 2017.

14. 홍석철 외, 『고혈압 · 고지혈증 환자의 합병증 발생률과 사망률 예측 및 의료비 추정 모형 개발』, 보험개발원, 2018.

(Translated in English) Hong, S. C., et al., *Development of a Predictive Model for the Incidence and Mortality Rates of Complications and Estimation of Medical Expenses in Patients with Hypertension and Hyperlipidemia*, Korea Insurance Development Institute, 2018.

15. 홍석철 외, 『주요 만성질환 예측모형 및 건강관리 시스템 개발』, 보험개발원, 2019.

(Translated in English) Hong, S. C., et al., *Development of a Predictive Model and Health Management System for Major Chronic Diseases*, Korea Insurance Development Institute, 2019.

16. 홍석철, "인구고령화와 국민건강 추이," 서울대 금융경제연구원 발표자료, 2021.
(Translated in English) Hong, S. C., et al., Population Aging and Trends in National Health, Seoul National University Institute for Research in Finance and Economics, 2021.
17. Choi, Yongok, "Impact of Longevity Risks on the Korean Government: Proposing a New Mortality Forecasting Model," *Korean Economic Review*, Vol. 36, No. 1, 2020, pp. 201-225.
18. Fleisher, Lee A., et al., "2014 ACC/AHA Guideline on Perioperative Cardiovascular Evaluation and Management of Patients Undergoing Noncardiac Surgery: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *Journal of the American College of Cardiology*, Vol. 64, No. 22, 2014, e77-e137.
19. Hippisley-Cox, Julia, Carol Coupland, and Peter Brindle, "Development and Validation of QRISK3 Risk Prediction Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study," *BMJ*, 2017, 357, j2099.
20. Kannel William B., D. McGee, and T. Gordon, "A General Cardiovascular Risk Profile: the Framingham Study," *American Journal of Cardiology*, Vol. 38, No. 1, 1976, pp. 46-51.
21. Ke, Guolin, et al., "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," NIPS, 2017.
22. Kim, Meelim, et al., "Multidimensional Cognitive Behavioral Therapy for Obesity Applied by Psychologists using a Digital Platform: Open-label Randomized Controlled Trial," *JMIR mHealth and uHealth*, Vol. 8, No. 4, 2020, e14817.
23. Kim, Young-Eun, et al., "Updating Disability Weights for Measurement of Healthy Life Expectancy and Disability-adjusted Life Year in Korea," *Journal of Korean Medical Science*, Vol. 35, No. 27, 2020, e219.
24. Kim, Youngin, Bumjo Oh, and Hyun-Young Shin, "Effect of mHealth with Offline Antiobesity Treatment in a Community-based Weight Management Program: Cross-sectional Study," *JMIR mHealth and uHealth*, Vol. 8, No. 1, 2020, e13273.
25. Murray, Christopher J., "Quantifying the burden of disease: the technical basis for disability-adjusted life years," Bulletin of the World health Organization, Vol. 72, No. 3, 1994, pp. 429-445.
26. Ock, Minsu, et al., "Disability Weights Measurement for 289 Causes of Disease Considering Disease Severity in Korea," *Journal of Korean Medical Science*, Vol. 34, Suppl 1, 2019.
27. OECD, "OECD Health Statistics 2022," 2022.
28. Sassi, Franco, "Calculating QALYs, Comparing QALY and DALY Calculations," *Health Policy and Planning*, Vol. 21, No. 5, 2006, pp. 402-408.
29. Toro-Ramos, Tatiana, et al., "Effectiveness of a Smartphone Application for the Management of Metabolic Syndrome Components Focusing on Weight Loss: A Preliminary Study," *Metabolic Syndrome and Related Disorders*, Vol. 15, No. 9, 2017, pp. 465-473.

30. Vos, Theo, et al., "Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990-2019: A Systematic Analysis for the Global Burden of Disease Study 2019," *The Lancet*, Vol. 396, No. 10258, 2020, pp. 1204-1222.

31. Vuik, S., A. Lerouge, Y. Guillemette, A. Feigl, and A. Aldea, "The Economic Burden of Obesity," in The Heavy Burden of Obesity: The Economics of Prevention, *OECD Publishing*, Paris, 2019, pp. 74-100.

32. Vuik, Sabine, and Michele Cecchini, "Modelling Life Trajectories of Body-mass Index," *OECD Health Working Papers*, No. 132, OECD Publishing, Paris, 2021.

33. World Health Organization, "WHO Methods and Data Sources for Global Burden of Disease Estimates 2000-2019," Global Health Estimates Technical Paper WHO/DDI/DNA/GHE, 2020.

34. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB, "Probability of Stroke: A Risk Profile from the Framingham Study," *Stroke*, Vol. 22, No. 3, 1991, pp. 312-318.

35. Zhang, P., M. Woodward, J. Shen, and Y. Wu, "24 Individual Disability-Adjusted Life Year: A Summary Health Outcome Indicator Used for Prospective Studies," in Handbook of Disease Burden and Quality of Life Measures, *Springer*, New York, NY, 2010, pp. 425-436.

Predicting the Future Disease Burden and Medical Cost: Using National Health Insurance Big Data*

Sok Chul Hong** · Sangyon Lee*** · Sehyun Kim**** · Sunghyun Jun*****

Abstract

The aging population and the rising prevalence of chronic diseases have led to a surge in disease burden and medical expenses, prompting an urgent need for proactive healthcare, while the evaluation of medical and economic feasibility remains insufficient. This study proposes a methodology for evaluating the medical and economic feasibility of healthcare using National Health Insurance Service database and big data analysis techniques. A disease burden index, which considers the severity of various diseases, is proposed, and a 10-year cumulative index is estimated using health checkup indicators. The impact of healthcare on medical cost reduction is illustrated through the alleviation of the disease burden index with an example of a healthcare program. Finally, the importance of health care is emphasized from the perspective of health care policy and financial sustainability of NHIS.

Key Words: National Health Insurance Service (NHIS), big data, burden of disease, medical expense, machine learning
JEL Classification: I10, I18

Received: April 18, 2023. Revised: May 9, 2023. Accepted: May 30, 2023.

* This research was supported by the National Health Insurance Service and the Health Finance Center at Seoul National University. The study had Seoul National University institutional review board approval (IRB No. E2105/001-008). This paper was presented in 2023 Korea's Allied Economic Associations Annual Meeting special session of the Korean Journal of Economic Studies (Applied Economic Research using Big Data, Unstructured Data, and AI).

** First Author, Professor, Department of Economics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea, Phone: +82-2-880-6360, e-mail: sokchul.hong@snu.ac.kr

*** Second Author, M. A. Student, Department of Economics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea, e-mail: stepano1992@snu.ac.kr

**** Third Author, Ph.D. Student, Department of Economics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea, e-mail: iocking2001@snu.ac.kr

***** Fourth Author, Ph.D. Student, Department of Economics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea, e-mail: jeon6884@snu.ac.kr