

연구논문

통계적 매칭 방법론을 활용한 다출처 데이터 통합 실증 분석: 한국의료패널조사를 중심으로*

이혜정**

데이터 개방 추세에 따라 데이터가 다량으로 개방되고 있지만 개별적으로 분산되어 제공하고 있어, 여러 개의 데이터를 서로 연계하고 통합한 데이터로의 활용은 저조한 편이다. 이에 다출처 데이터를 통합하는 시도는 곳곳에 분산된 데이터의 활용 가치를 높일 수 있을 것이다.

이 연구에서는 통계적 매칭 방법론을 활용하여 한국의료패널조사와 생활시간조사 데이터를 통합하고, 그 결과물인 통합데이터를 분석하여 주관적 건강 상태와 생활시간 사용의 관계를 살펴보았다. 또한 통계적 매칭 방법론을 사용하여 데이터를 통합하는 경우 고려해야 할 사항에 대해 제시하였다.

주제어: 한국의료패널조사, 생활시간조사, 데이터 통합, 통합데이터, 통계적 매칭 방법론

* 이 글은 이혜정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》(세종: 한국보건사회연구원)의 제5장 “데이터 통합 실증 분석”에 수록된 내용을 바탕으로 재구성하여 작성하였습니다.

** 한국보건사회연구원 부연구위원(snp625@kihasa.re.kr).

I. 서론

윤석열 정부는 2022년 7월 120대 국정과제 중 하나로, 모든 데이터가 연결되는 세계 최고의 디지털플랫폼정부 구현을 목표로 하였다(윤석열 정부 120대 국정과제, 2022.7). 데이터 분석을 통하여 정책효과를 정밀하게 예측하여 국정운영의 과학화를 실현하고, 디지털플랫폼정부 혁신 생태계 조성을 위해 공공데이터를 전면 개방하며 마이데이터를 전산업으로 확산하는 것이다. 또한 민·관이 협업하여 범정부 데이터·서비스의 개방·연계·활용 인프라를 구축하는 것이다. 이러한 정부의 노력으로 국내 데이터 자원은 포털, 플랫폼을 통해 일반인이 자유롭게 사용할 수 있도록 지속적으로 개방 중에 있다.

한편 한국보건사회연구원은 중요한 사회문제를 다루는 여러 가지 다양한 설문조사를 직접 수행하고 있으며 양질의 통계 및 조사데이터(2023년 기준 노인실태조사 외 총 40종)를 생산·배포하고 있다. 한국보건사회연구원은 국가승인통계는 8종, 보건복지부 국가승인통계는 8종, 여성가족부 국가승인통계는 1종, 국가승인통계에 해당하지 않는 미승인통계는 23종으로 구분된다.

이렇듯 데이터 개방 요구도 점차 증가하고 있는 가운데, 데이터 자원은 많아지고 여러 가지 환경적인 요소도 좋아지고 있는 상황이나 다음과 같은 어려움도 있다. 행정데이터의 경우 행정 프로세스나 시스템 프로세스 관리 목적으로 수집되어 결측치나 정보 누락 문제가 종종 발생할 수 있다. 또한 다양한 목적으로 수집되는 조사데이터는 행정데이터에 비해서 표본크기가 작을 뿐 아니라, 패널조사의 경우에는 표본이탈로 인한 대표성 문제가 제기될 수 있다. 이러한 다양한 상황으로 인하여 조사데이터, 공공데이터, 민간데이터 등의 데이터 유형에 관계없이 하나의 데이터로 상세하고 폭넓은 정보를 얻는 것은 어렵다고 볼 수 있다.

종합하여 보면 데이터 개방 추세에 따라 데이터가 다량으로 개방되고 있지만 개별적으로 분산되어 제공하고 있어, 여러 개의 데이터를 서로 연계하고 통합한 데이터로 활용하는 것은 저조한 편이다. 이에 다출처 데이터를 통합하는 시도는 곳곳에 분산된 데이터의 활용 가치를 높일 수 있을 것이다.

다출처 데이터를 통합하는 방법은 자료연계와 통계적 매칭 방법으로 구분할 수

있다. 자료연계는 정확매칭으로 다출처 데이터의 동일 표본에 대해 통합할 수 있는 방법이다. 동일 표본이 아닌 다출처 데이터인 경우에는 통계적 매칭 방법을 활용하여 데이터를 통합해 볼 수 있다. 이 연구에서는 통계적 매칭 방법론을 활용하여 한국의료패널조사와 생활시간조사 데이터를 통합해 보고자 한다. 한국의료패널조사와 생활시간조사 데이터를 통합하는 목적은 다음과 같다.

첫 번째는 다출처 조사데이터에 대한 통합데이터 구축 가능성을 파악하고자 한다. 여러 개의 조사데이터를 활용하여 데이터를 통합한 선행연구는 다음과 같다. 한국미디어패널조사와 스마트폰 과의존 실태조사를 맨허튼 거리 함수에 기반을 둔 최근접 이웃 핫텍 방법을 활용하여 데이터를 통합한 다음에, 개인의 특성과 스마트폰 의존도의 상관관계를 살펴보았다(신지형·오윤석 2018). 한국복지패널조사와 재정패널조사를 맨허튼 거리 함수에 기반을 둔 랜덤 핫텍 방법을 활용하여 데이터를 통합한 후 납세여부와 복지수혜여부에 따른 복지인식에 대한 차이를 분석한 바 있다(오미애·최현수·김용대·이용희·진재현 2014). 생활시간조사와 경제활동인구조사 데이터를 마이크로 매칭 방법(최근접 이웃 핫텍 방법, 랜덤 핫텍 방법, 혼합 방법, 경향 점수 매칭 방법)을 활용하여 데이터를 통합하고 결과를 비교하였다(김대건·박민규 2014). 또한, 사업체기초조사와 국민연금 데이터를 랜덤 핫텍 방법을 활용하여 데이터를 통합한 사례연구도 있다(이영섭·김선웅·안홍엽·임정은·김희경 2009). 이 연구에서는 한국의료패널조사와 생활시간조사 데이터를 통계적 매칭 방법을 활용하여 하나의 데이터로 통합하고자 한다.

두 번째는 통합데이터를 분석하여 주관적 건강 상태와 생활시간 사용의 관계를 살펴보고자 한다. 한국의료패널조사는 주관적 건강 상태에 대한 정보를 가지고 있는 반면에, 생활시간 사용에 대한 정보를 가지고 있지 않아서 분석하기 어려운 상황이다. 그러나 생활시간조사가 생활시간 사용에 대한 정보를 가지고 있어서 데이터를 통합할 수 있다면 이와 관련하여 살펴볼 수 있다. 이때 한국의료패널조사와 생활시간조사는 동일한 표본이 아니므로 통계적 매칭 방법을 활용하여 데이터를 통합하고자 한다. 구축한 통합데이터의 분석결과와 선행연구 결과를 비교하여 통계적 매칭 방법으로 구축한 통합데이터의 활용 가능성도 파악하고자 한다.

이 연구의 구성은 2장에서 통계적 매칭 방법론을 활용한 데이터 통합 방법 절차를 살펴보고, 통합데이터 구축에 활용되는 데이터를 소개하고 공통변수와 유일변수를 선정한다. 3장은 통계적 매칭 방법과 매칭변수를 선정하고, 4장은 통계적 매칭 방법론을 활용하여 구축된 통합데이터의 품질 평가를 실시한다. 5장은 통합데이터

의 분석결과를 살펴보고, 마지막으로 6장에서는 요약 및 결론으로 마무리한다.

II. 데이터 소개 및 공통변수와 유일변수 선정

데이터 통합은 한국의료패널조사를 기준데이터로, 생활시간조사를 제공데이터로 선정하여 실시하였다. 데이터 통합 절차는 <그림 1>과 같이 5단계에 걸쳐 실시하였고 주요 용어에 대한 설명은 <표 1>과 같다. 데이터 통합 절차의 첫 번째 단계는 기준데이터와 제공데이터에 대한 전처리 과정으로, 데이터 클리닝, 공통변수 표준화 등을 실시한다. 두 번째 단계는 데이터 통합을 위한 준비 과정으로, 기준데이터와 제공데이터에서 유일변수와 매칭변수를 선정한다. 그리고 기준데이터와 제공데이터를 통합하기 위해서 통계적 매칭 방법 선정은 모의실험을 실시하여 결정한다. 세 번째 단계는 통합데이터를 생성하고, 네 번째 단계는 통합데이터의 품질 평가로, 통합데이터와 제공데이터의 공통변수 및 유일변수 분포 비교를 통해 실시한다. 마지막 다섯 번째 단계는 통합데이터를 사용하여 연구목적에 대한 분석을 실시한다.

- (단계 1) 기준데이터와 제공데이터에 대한 전처리
 - 데이터 클리닝, 공통변수 표준화 등 실시
- (단계 2) 데이터 통합을 위한 준비
 - 유일변수, 매칭변수 및 통계적 매칭 방법 선정
 - 필요시 기증자 집단 선정
- (단계 3) 통합데이터 생성
- (단계 4) 통합데이터 품질 평가
 - 통합데이터와 제공데이터의 공통변수 및 유일변수 분포 비교
- (단계 5) 통합데이터 분석
 - 연구목적과 관련하여 분석 실시

<그림 1> 데이터 통합 절차

주: 필자 작성.

<표 1> 주요 용어 정리

용어	설명
통합데이터	기준데이터와 제공데이터를 하나의 통합된 데이터로 구축한 결과물.
기준데이터	데이터 통합시 기준이 되는 데이터로, 공통변수와 유일변수를 가지고 있어야 함.
제공데이터	기준데이터에 추가로 제공되는 변수를 포함하는 데이터로, 공통변수와 기준데이터에는 없는 유일변수를 가지고 있어야 함.
유일변수	각 데이터에서만 유일하게 가지는 변수.
공통변수	각 데이터에 공통으로 포함된 변수로, 데이터의 유사성을 판단하는 데 활용됨.
매칭변수	공통변수 중에서 데이터를 통합할 때 기준이 되는 변수.

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.21. <표 1-1>.

데이터 통합 대상인 한국의료패널조사와 생활시간조사에 대한 소개는 다음과 같다. 한국의료패널조사는 보건의료서비스의 대응성·접근성 향상과 효율화를 위한 관련 연구 및 정책 수행에 필요한 기초 정보를 수집하는 데 있다.¹⁾ 2008년부터 한국보건사회연구원과 국민건강보험공단이 공동으로 수행하고 있으며, 2019년까지 1기 패널조사를 완료하였고 2020년부터 2기 패널조사를 매년 실시하고 있다. 조사 내용은 의료서비스 이용, 의료 관련 지출, 의료 이용에 영향을 미치는 사회경제적 요인, 건강 관련 인식 및 행태 등이 있다. 조사 방식은 조사원이 컴퓨터를 이용하여 면접 조사한다. 조사대상은 가구 내 속한 모든 가구원이며 2019년도 연간데이터 기준 6,689가구 내 12,396명이다.

생활시간조사는 개인의 시간 활용과 의식을 파악하여 국민의 생활방식과 삶의 질을 측정하여 가사노동의 경제적 가치 등 다양한 연구에 활용하는 데 목적이 있다. 통계청에서 수행하고 있으며, 조사주기는 5년으로, 첫 조사를 1999년에 시작하여

1) 2기 한국의료패널 2019년 연간데이터 유저가이드(한국보건사회연구원, 국민건강보험, 2022.5)를 참고하여 작성함.

최근 2019년 조사를 완료하였다. 가구와 개인 관련 조사항목은 조사원이 면접조사하고, 시간일지는 응답자가 자기기입식(종이)으로 직접 작성한다. 시간일지는 응답자의 하루 24시간 활용을 파악하기 위해서 10분 단위로 2일 동안 작성하도록 되어 있고, 주요 항목은 주행동, 동시행동, 장소 및 이동 수단, 함께한 사람, ICT 기기 사용 등이 있다. 조사대상은 만 10세 이상이며, 조사규모는 2019년도 조사기준 12,388가구 내 26,091명이다.

한국의료패널조사와의 데이터 통합을 위해 생활시간조사는 다음과 같이 고려하였다. 만 19세 이상을 대상으로 하였고 평일에 응답한 생활시간 정보만을 활용하였다. 보통 평일(월~금요일)과 주말(토, 일요일)의 생활시간 소비행태는 다를 것으로 판단했기 때문이다. 또한 공통변수와 유일변수 모두 결측이 아닌 값이 있는 응답자만으로 선정하여 최종 확정된 데이터는 28,247명이다. 한국의료패널조사(기준데이터)의 비율은 생활시간조사(제공데이터)의 35.6%로 나타났다.

한국의료패널조사와 생활시간조사 데이터를 통합하는 목적에 따라 한국의료패널조사와 생활시간조사에서의 유일변수는 <표 2>와 같이 선정하였다. 한국의료패널조사는 주관적 건강상태, 스트레스 인지 정도, 만성질환 유무와 의료비를, 생활시간조사는 활동에 대한 시간(필수시간, 의무시간, 여가시간, 운동시간)과 여가만족도이다. 여기서는 전체 시간량을 살펴보는 것이어서 생활시간조사에서 활동에 대한 시간 변수는 주행동²⁾을 기준으로 생성하였다.³⁾ 필수시간은 잠, 식사 등 개인 유지를 위해서 필요한 필수적인 시간이고, 의무시간은 일, 학습, 가사노동, 이동 등 행동에 의무가 부여된 시간을 의미한다. 여가시간은 필수 및 의무적인 시간 외에 개인이 자유롭게 사용이 가능한 시간으로, 문화 및 여가활동, 교제 및 참여활동, 자원봉사 및 무급연수 등이 해당한다. 여가시간의 세부 항목으로 스포츠 및 레포츠 활동을 한 시간(이하 ‘운동시간’)이 있는데, 이 변수도 생성하여 함께 살펴보았다.

다음으로 공통변수를 선정하였는데, 이때 기준은 두 가지 조사데이터에서 공통으로 포함하고 있는 변수이고 변수의 응답범주가 동일해야 하는 것이다. 변수의 응답범주가 동일하지 않은 경우에는 표준화를 통해 동일한 범주를 가지도록 조정할 필

2) 생활시간조사는 활동에 대한 행동을 주행동, 동시행동으로 나누어서 조사하였음. 주행동의 정의는 10분 간격의 시간대에 수행한 한 가지 행동, 두 가지 이상 행동을 하였을 경우 가장 길거나 주관적으로 주된 행동이라고 판단한 행동임.

3) 『생활시간조사』 통계정보보고서(통계청 2019), 2019년 생활시간조사 결과(통계청 2020), 생활시간조사 행동분류표(통계청 2019)의 내용을 참고하여 작성함.

요가 있다. <표 3>은 두 가지 조사데이터에서 공통변수로 6개, 시·도, 성별, 만 나이, 최종학력, 혼인상태, 경제활동상태 등 6개가 활용 가능한 것으로 나타났다. 시·도, 성별과 나이를 제외한 공통변수의 응답값이 서로 다르게 나타나 표준화 작업을 통해 일치시켰다. 또한 응답 범주가 많은 변수의 경우 시·도는 3개 권역(대도시, 중·소도시, 농·어촌)으로, 혼인상태는 배우자 유무로, 경제활동상태는 3개 형태(임금근로자, 고용주/자영업자/무급가족종사자, 비경제활동인구)로 범주를 통합한 변수를 추가 생성하였다. 한편, 6개의 공통변수는 모두 값을 가지고 있어서 매칭변수를 선정할 때 이중에서 선정할 수 있었다.

<표 2> 한국의료패널조사와 생활시간조사의 유일변수

조사	유일변수	통합데이터 (최종 분석변수)
한국의료 패널조사	주관적 건강상태	- 매우 좋음 - 좋음 - 보통 - 나쁨 - 매우 나쁨
	스트레스 인지 정도	- 대단히 많이 느낀다 - 많이 느끼는 편이다 - 조금 느끼는 편이다 - 거의 느끼지 않는다
	만성질환 유무	- 있음 - 없음
	의료비	0~1,729만 원
	필수시간	190~1,440분
생활시간 조사	의무시간	0~1,170분
	여가시간	0~1,060분
	운동시간	0~690분
	여가만족도	- 매우 만족 - 약간 만족 - 보통 - 약간 불만족 - 매우 불만족
		- 좋음 - 보통 - 나쁨

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.159. <표 5-4>.

<표 3> 한국의료패널조사와 생활시간조사의 공통변수

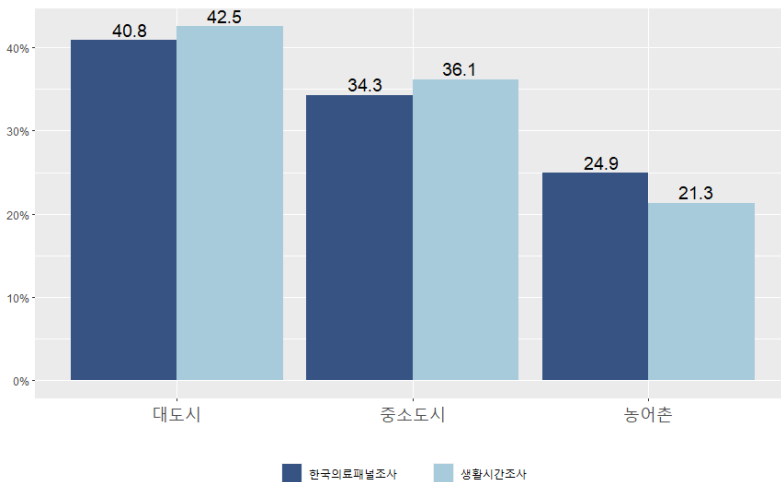
공통변수	한국의료패널조사 (기준데이터)	생활시간조사 (제공데이터)	통합데이터
	- 17개 시·도	- 17개 시·도	- 17개 시·도
시·도 ⁴⁾			*추가 생성 변수 - 대도시 - 중소도시 - 농어촌
성별	- 남자 - 여자	- 남자 - 여자	- 남자 - 여자
만 나이	만 나이 출생연도	- 만 나이	- 만 나이 - 연령대(10세 단위)
최종 학력	- 받지 않음(미취학 포함) - 초등학교 - 중학교 - 고등학교 - 대학교(전문대학 포함) - 대학원	- 받지 않았음(미취학 포함) - 초등학교 - 중학교 - 고등학교 - 대학교(4년제 미만) - 대학교(4년제 이상) - 대학원 석사과정 - 대학원 박사과정	- 무학 - 초등학교 졸업 이하 - 중학교 졸업 - 고등학교 졸업 - 대학교 재학 이상
졸업 유무	- 졸업 - 재학 - 중퇴 - 수료 및 휴학	- 졸업 - 재학 - 수료 - 휴학 - 중퇴	
혼인상태	- 배우자가 있으며, 함께 살고 있음(사실혼 상태 포함) - 배우자가 있으나, 함께 살고 있지 않음(출장 등의 일시적 상태 제외) - 배우자 사망으로 배우자가 없음 - 이혼으로 배우자가 없음 - 결혼한 적 없음 - 응답 거부	- 미혼 - 배우자 있음(동거 포함) - 사별 - 이혼(별거 포함)	- 미혼 - 배우자 있음 - 별거/이혼/사별 *추가 생성 변수 - 배우자 있음 - 배우자 없음

4) 17개 시·도는 서울, 부산, 대구, 인천, 광주, 대전, 울산, 세종, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주임.

경제 활동 상태	경제 활동 참여 상태	- 임금근로자 - 자활근로, 공공근로, 노인 일자리, 희망근로⁵⁾ - 고용주 - 자영업자 - 무급가족종사자 - 비경제활동인구	- 상용종사자 - 임시근로자 - 일용근로자 - 고용원이 있는 자영업자 - 고용원이 없는 자영업자 - 무급가족종사자	- 상용임금근로자 - 임시임금근로자 - 일용임금근로자 - 고용주 - 자영업자 - 무급가족종사자 - 비경제활동인구
	종사상 지위	- 상용직 - 임시직 - 일용직		*추가 생성 변수 - 임금근로자 고용주/자영업자/ 무급가족종사자 - 비경제활동인구

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. pp.159-160. <표 5-5>.

다음은 두 가지 조사의 표준화된 공통변수에 대한 응답값 분포를 막대그래프로 비교해 보았다. <그림 2>를 보면 3개 권역의 경우 한국의료패널조사와 생활시간조사 간 차이는 농·어촌이 3.6%p였으나 3개의 권역별 차이는 대체로 비슷한 편이었다.

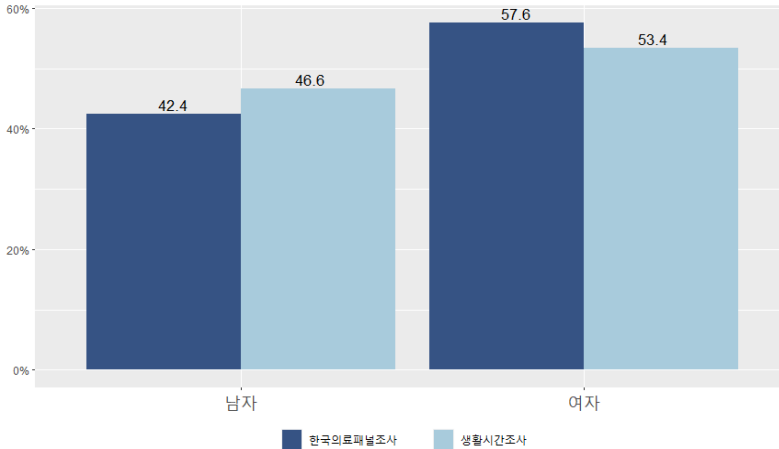


<그림 2> 공통변수 분포: 3개 권역(단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.161. [그림 5-7].

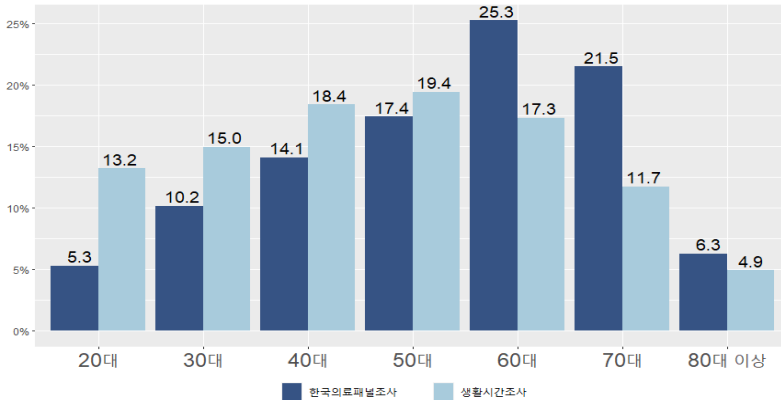
5) 한국의료패널조사(기준데이터)에서 경제활동참여 상태의 경우 자활근로, 공공근로, 노인일자리, 희망근로는 연계 데이터에서 일용임금근로자로 분류하였음.

성별은 두 가지 조사 모두 여성의 비율이 절반 이상으로 더 많으며, 한국의료패널조사의 경우 57.6%로 생활시간조사(53.4%)보다 4.2%p 높았다(<그림 3> 참조).



<그림 3> 공통변수 분포: 성별(단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.162. [그림 5-8].

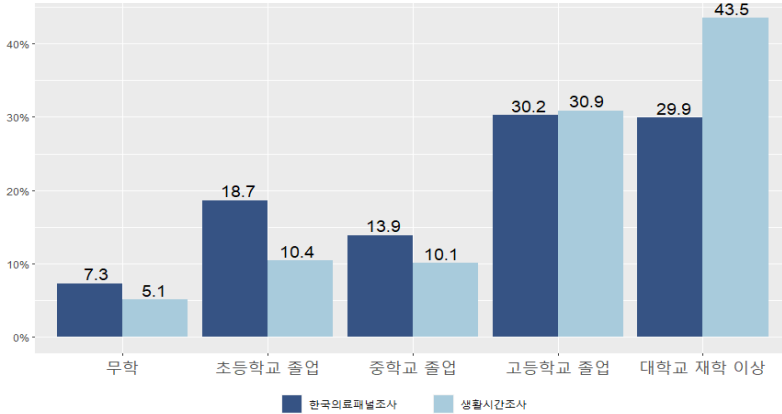


<그림 4> 공통변수 분포: 연령대(단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.163. [그림 5-9].

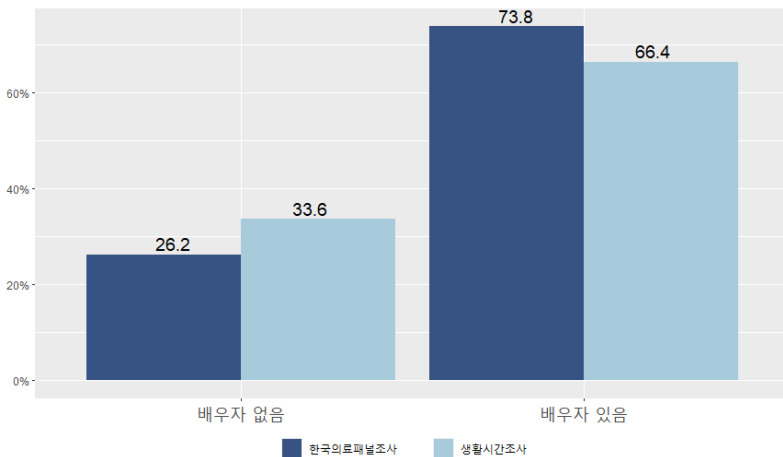
<그림 4>는 연령대 분포에 대한 것으로 한국의료패널조사는 60대 이상의 비율이 생활시간조사에 비해 높은 편이나, 생활시간조사는 50대 이하의 비율이 높은 편으

로 나타났다. <그림 5>는 최종학력에 대한 분포이다. 한국의료패널조사는 초등학교 졸업이 18.7%로 생활시간조사(10.4%)에 비해 높은 편(8.3%p)이고, 생활시간조사는 대학교 재학 이상이 43.5%로 한국의료패널조사(29.9%)에 비해 13.6%p 높았다. 이를 제외한 나머지 최종학력의 분포는 차이가 크지 않다고 볼 수 있다.



<그림 5> 공통변수 분포: 최종학력(단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.163. [그림 5-10].

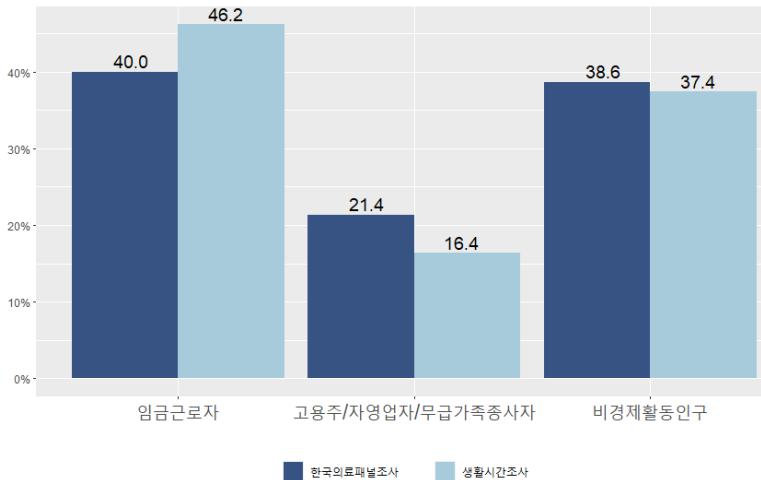


<그림 6> 공통변수 분포: 배우자 유무 (단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.164. [그림 5-11].

<그림 6>은 배우자 유무에 대한 분포로 두 가지 조사 모두 유배우자 비율이 높은 편으로 나타났다(한국의료패널조사 73.8%, 생활시간조사 66.4%).

경제활동 상태를 보면, 두 가지 조사 모두 임금근로자의 비율이 높은 편에 속하였고(한국의료패널조사 40%, 생활시간조사 46.2%), 조사 간 경제활동 상태의 차이는 1.2~6.2%p를 보여 크지 않은 편이라고 볼 수 있다(<그림 7> 참조).



<그림 7> 공통변수 분포: 경제활동 상태(단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.165. [그림 5-12].

III. 통계적 매칭 방법 및 매칭변수 선정

이 장에서는 모의실험을 통해 ‘통계적 매칭 방법’과 공통변수와 유일변수 간 연관성 분석을 통해 ‘매칭변수’를 선정하였다.

먼저 통계적 매칭 방법을 선정하기 위하여 한국의료패널조사 2019년 연간데이터를 사용하여 모의실험을 실시하였다. 모의실험을 위한 연구대상 모집단은 만 19세 이상이면서 사전조사 응답자로 한정하였고, 모의실험에 사용되는 변수를 모두 응답

한 개인만을 추출하여 최종 10,068명으로 확정하였다. 19세 이상이면서 사전조사 응답자로 한정된 이유는 한국의료패널조사 데이터의 유일변수로 만성질환 유무, 주관적 건강상태, 스트레스 인지 정도 등을 활용하므로 동일하게 설정하였다. 또한 모의실험에 사용하는 변수가 결측인 경우는 제외하고 모두 값이 있는 응답자만으로 선정하였다.

확정한 데이터는 단순임의추출(simple random sampling) 방법으로 36% 추출한 개인(3,624명)을 기준데이터로, 나머지 64%(6,444명)는 제공데이터로 나누었다. 보통 제공데이터의 크기는 적어도 기준데이터보다 커야 후보군 활용 측면에서 유용하다는 점과 한국의료패널조사와 생활시간조사 데이터의 표본 규모 비율을 고려하여 결정하였다.

기준데이터의 유일변수는 경제활동 유무와 개인의 연간 근로소득(이하 ‘근로소득’)이고, 제공데이터의 유일변수는 만성질환 유무와 개인의 연간 의료비(이하 ‘의료비’)이다. 공통변수는 연령대(19세 이상부터 10세 단위, 7개 범주), 성별, 최종학력(무학, 초졸, 중졸, 고졸, 대졸 이상), 거주지역(대도시, 중·소도시, 농·어촌), 혼인상태(미혼, 배우자 있음, 별거/이혼/사별)이다.

연구대상 모집단 데이터로부터 기준데이터와 제공데이터를 생성하는 과정을 1,000번 실시하여 모의실험용 데이터를 1,000개 생성하였다.

모의실험에서 고려한 통계적 매칭 방법은 최근접 이웃 핫덱과 랜덤 핫덱 방법이며, 다양한 거리 함수에 따른 데이터 통합에 대한 결과를 살펴보았다. 최근접 이웃 핫덱은 기준데이터와 제공데이터에 있는 매칭변수를 사용하여 관측값의 거리를 계산한 다음에 기준데이터의 개체와 가장 가까운 거리에 있는 제공데이터의 개체를 선택하는 방법이다. 랜덤 핫덱은 제공데이터의 활용 가능한 모든 개체에서 무작위로 하나의 개체를 선택하는 방법이다. 거리 함수는 3가지로 고어(gower), 맨해튼(manhattan), 정확(exact)을 고려하였고, 식은 <표 4>와 같다.

<표 4> 거리 함수

거리 함수	식
고어	$d(i,j)=\frac{1}{p}\sum_{k=1}^pD_{ijk}$ 연속형 변수인 경우에는 $D_{ijk}=1-\frac{ x_{ik}-x_{jk} }{\max(x_k)-\min(x_k)}$, 범주형 변수인 경우에는 $D_{ijk}=\begin{cases}1 & x_{ik}=x_{jk} \\ 0 & x_{ik}\neq x_{jk}\end{cases}$
맨해튼	$d(i,j)=\sum_{k=1}^p x_{ik}-x_{jk} $
정확	$d(i,j)=\sum_{k=1}^pD_{ijk}$ $D_{ijk}=\begin{cases}1 & x_{ik}=x_{jk} \\ 0 & x_{ik}\neq x_{jk}\end{cases}$

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화
를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.147. <표 5-1>.

다음은 모의실험에서 고려된 통계적 매칭 방법으로, <표 5>와 같으며, 6가지이
다. 모의실험 결과표에서의 통계적 매칭 방법에 대한 표기도 함께 정의하였다. 예를
들면 고어 거리 함수에 기반을 둔 최근접 이웃 핫덱 방법은 NND_gow로 나타났다.

<표 5> 모의실험에서 고려된 통계적 매칭 방법

방법	거리 함수	표기
최근접 이웃 핫덱	고어	NND_gow
	맨해튼	NND_man
	정확	NND_exa
랜덤 핫덱	고어	RD_gow
	맨해튼	RD_man
	정확	RD_exa

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화
를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.148. <표 5-2>.

한편 D’Orazio(2022)는 통계패키지 R의 StatMatch 패키지를 사용하여 통계적 매

칭 방법으로 데이터 통합을 할 수 있도록 하고 있으며, 이 모의실험에서도 StatMatch를 활용하였다.

모의실험의 기준데이터에는 제공데이터의 유일변수에 대한 참값을 포함하고 있다. 그래서 통합데이터에 대한 평가는 기준데이터의 참값과 제공데이터의 값을 비교하면 된다. 4가지 평가지표를 통해 비교하는데, 통합데이터의 변수가 이분형 변수이면 오분류율(error rate)로, 연속형이면 편향(bias)과 제공근평균제곱오차(RMSE: Root Mean Square Error), 개체에 대한 제공근평균제곱편차(RMSD: Root Mean Square Deviation)로 평가하였다. 4가지 평가지표에 대한 식은 다음과 같다.

$$\text{오분류율} = E\left(1 - \frac{1}{N} \sum_{i=1}^N I_{ik}\right)$$

$$I_{ik} = \begin{cases} 1 & z_{ik, int} = z_{ik} \\ 0 & z_{ik, int} \neq z_{ik} \end{cases}$$

여기서 $z_{ik, int}$ 는 통합데이터의 값이고, z_{ik} 는 참값이고, I_{ik} 은 지시함수를 의미하며, $i(=1, \dots, N)$ 는 개체수이고, $k(=1, \dots, K)$ 는 모의실험 자료의 개수이다.

$$\text{편향} = E(|\hat{\theta}_{k, int} - \theta_k|)$$

여기서 $\hat{\theta}_{k, int} = \frac{1}{N} \sum_{i=1}^N z_{ik, int}$ 이고, $\theta_k = \frac{1}{N} \sum_{i=1}^N z_{ik}$ 이다.

$$\text{제공근평균제곱오차} = \sqrt{E(\hat{\theta}_{k, int} - \theta_k)^2}$$

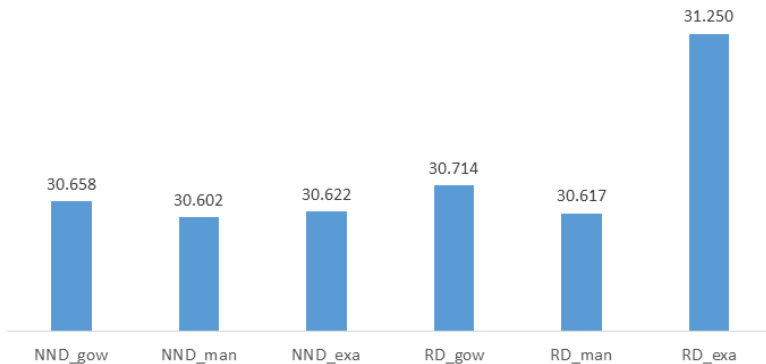
$$\text{개체에 대한 제공근평균제곱편차} = E\left(\sqrt{\frac{1}{N} \sum_{i=1}^N (z_{ik, int} - z_{ik})^2}\right)$$

여기서 $z_{ik, int}$ 는 개체별 통합데이터의 값이고, z_{ik} 는 참값이다.

기준데이터에 제공데이터를 통합할 때 매칭변수를 사용하는데, 매칭변수는 공통 변수 중에서 기준데이터와 제공데이터의 유일변수와의 연관성이 높은 것으로 선정한다. 매칭변수 선정 기준은 크래머 V(Cramer's V), 분류와 회귀(CART, Classification

and Regression Tree)를 통해 선정하였다. 크래머 V는 공통변수와 유일변수가 모두 범주형 변수이어야 사용할 수 있다. 분류와 회귀는 연속형 변수와 범주형 변수일 때 모두 적용이 가능하고 변수 분포에 대한 가정을 특정한 분포에 국한하지 않는다(신지형·오윤석 2018. p.58). 매칭변수로 ‘연령대’, ‘성별’, ‘최종학력’을 선정하였다.

매칭변수를 기준으로 기준데이터와 제공데이터를 통합하였고, 4가지 평가지표를 통해 구축된 통합데이터의 결과를 살펴보았다. 먼저, 제공데이터의 유일변수인 만성질환 유무에 대한 오분류율을 보면, 맨해튼 거리 함수에 기반을 둔 최근접 이웃 핫택 방법(NND_man)이 30.602%로 가장 낮았고, 정확 거리 함수에 기반을 둔 랜덤 핫택 방법(RD_exa)이 31.25%로 가장 높게 나타났다(<그림 8> 참조). 정확 거리 함수에 기반을 둔 랜덤 핫택 방법을 제외하고는 비슷한 편이라고 볼 수 있다.

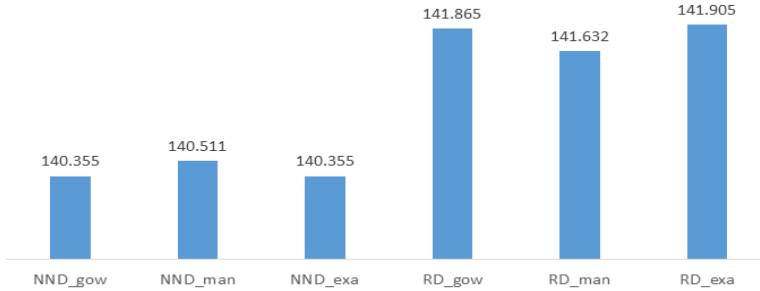


<그림 8> 만성질환 유무에 대한 오분류율 (단위: %)

자료: 이해정, 이기호, 안수인, 임종호, 이상혁, 조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.151. [그림 5-1].

다음으로, 제공데이터의 유일변수인 의료비의 개체에 대한 제공근평균제공편차를 보면 <그림 9>와 같다. 6가지 방법 간 개체에 대한 제공근평균편차의 값이 큰 차이를 가지는 편은 아니었다. 고어 거리 함수에 기반을 둔 최근접 이웃 거리 핫택(NND_gow)과 정확 거리 함수에 기반을 둔 최근접 이웃 거리 핫택(NND_exa)이 140.355로 가장 작았다. 이에 반해 정확 거리 함수에 기반을 둔 랜덤 핫택(RD_exa)이 141.905로 가장 큰 값을 가졌다. 거리 함수와 상관없이 최근접 이웃 거리 핫택 방법이 랜덤 방법에 비해 전반적으로 작게 나타나, 이를 통해 제공데이터의 의료비와

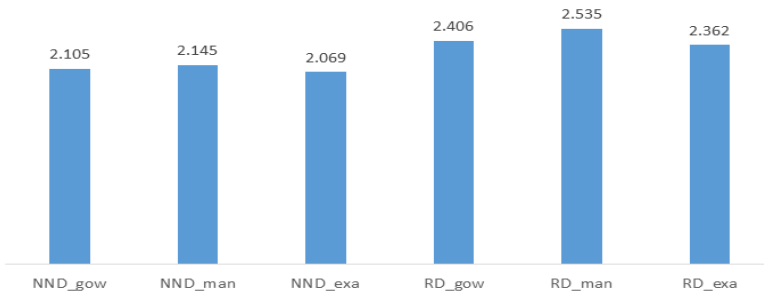
실제값 간의 근사 정도가 높은 편이라고 볼 수 있다.



<그림 9> 의료비의 개체에 대한 제곱근평균제곱편차

자료: 이해정, 이기호, 안수인, 임종호, 이상혁, 조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.152. [그림 5-2].

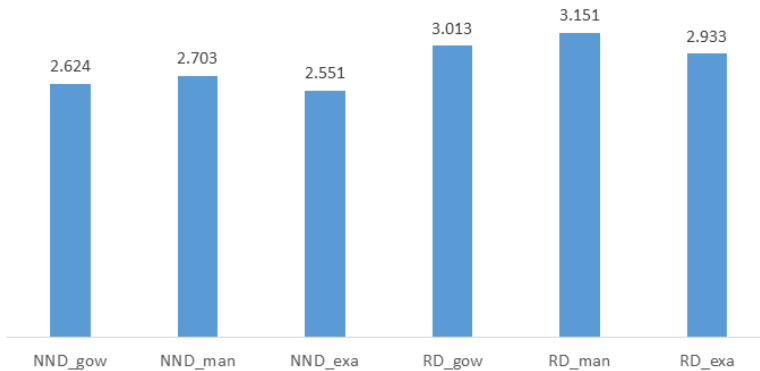
<그림 10>은 의료비의 편향으로, 정확 거리 함수에 기반을 둔 최근접 이웃 거리 핫텍 방법(NND_exa)이 2만 690원으로 가장 작고, 맨해튼 거리 함수에 기반을 둔 랜덤 핫텍(RD_man)이 2만 5,350원으로 가장 큰 값을 가졌다. 하지만 편향 역시 6 가지 방법 간 차이가 크지 않은 편으로 볼 수 있으며 최근접 이웃 거리 핫텍 방법이 랜덤 핫텍 방법에 비해 전반적으로 편향이 작게 나타나 참값 평균에 더 가깝다고 볼 수 있다.



<그림 10> 의료비의 편향 (단위: 만 원)

자료: 이해정, 이기호, 안수인, 임종호, 이상혁, 조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.152. [그림 5-3].

마지막으로, 의료비의 제공근평균제공오차는 정확 거리 함수에 기반을 둔 최근접 이웃 거리 핫텍 방법(NND_exa)이 2.551로 가장 작고, 맨해튼 거리 함수에 기반을 둔 랜덤 핫텍(RD_man)이 3.151로 가장 큰 값을 가졌다(<그림 11> 참조). 제공근평균제공오차도 최근접 이웃 거리 핫텍 방법이 랜덤 핫텍 방법에 비해 전반적으로 작아, 정밀도가 높은 편이어서 평균 추정치를 더 안정적으로 추정한다고 볼 수 있다.



<그림 11> 의료비의 제공근평균제공오차

자료: 이해정, 이기호, 안수인, 임종호, 이상혁, 조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.153. [그림 5-4].

전반적으로 살펴본 결과, 최근접 이웃 핫텍 방법이 랜덤 핫텍 방법에 비해 우수한 편이었으며, 거리 함수에 따른 효과는 크지 않은 편으로 나타났다. 이는 제공데이터의 매칭변수(연령대, 성별, 최종학력)에 따른 기증자가 충분하여 기준데이터의 대상과 동일한 특성을 가진 기증자로 제공되었기 때문이라고 볼 수 있다. 즉, 거리를 구할 때 같은 범주에서 가져오므로 거리 함수에 따른 차이가 없을 수 있다. 그러나 매칭변수가 각 조합(combination)에 따라 충분한 기증자가 없다면 거리 함수에 따라 영향을 줄 수 있다. 이 모의실험에서는 차이가 없었으나, 기준데이터와 제공데이터의 형태가 이와 같지 않다면 다른 결과가 도출될 수 있다. 따라서 데이터 형태에 따라 거리 함수의 영향은 다를 수 있으므로 사전에 모의실험을 통한 확인이 필요하다 고 생각한다. 모의실험에서 가장 좋은 결과를 나타낸 ‘정확 거리 함수에 기반을 둔 최근접 이웃 핫텍 방법’을 최종적으로 통계적 매칭 방법으로 선정하였다.

다음으로 한국의료패널조사와 생활시간조사 데이터를 통합할 때 사용하는 ‘매칭

변수'를 크래머 V 및 분류와 회귀를 통해 선정하였다. 한국의료패널조사의 주관적 건강상태와 공통변수에 대한 크래머 V의 값을 보면 최종학력(0.211)과 연령대(0.185)의 값이 가장 크게 나타났다(<표 6> 참조). 그리고 분류와 회귀의 변수 중요도 결과도 최종학력과 연령대가 높은 것으로 나타났다(<표 7> 참조). 생활시간조사의 생활시간과 공통변수에 대한 분류와 회귀의 변수 중요도 결과를 살펴보면 경제활동 상태와 연령대가 높은 것으로 나타났다(<표 8> 참조). 3개 권역과 배우자 유무는 두 가지 조사에서 모두 유일변수와의 연관 정도가 낮은 편이었다. 이러한 결과를 통해 매칭변수는 '연령대', '최종학력', '경제활동 상태'로 선정하였다.

<표 6> 한국의료패널조사의 주관적 건강상태와 공통변수에 대한 크래머 V

공통변수	크래머 V
3개 권역	0.062
성별	0.102
연령대	0.185
최종학력	0.211
배우자 유무	0.095
경제활동 상태	0.173

자료: 한국보건사회연구원·국민건강보험공단. 2019. 한국의료패널조사 2019년 연간데이터를 분석함.

<표 7> 한국의료패널조사의 주관적 건강상태와 공통변수에 대한 변수 중요도

공통변수	Relative	Importance
최종학력	1.000	15.088
연령대	0.888	13.400
3개 권역	0.672	10.139
배우자 유무	0.606	9.139
경제활동 상태	0.595	8.980
성별	0.492	7.418

자료: 한국보건사회연구원·국민건강보험공단. 2019. 한국의료패널조사 2019년 연간데이터를 분석함.

<표 8> 생활시간조사의 생활시간과 공통변수에 대한 변수 중요도

필수시간		
공통변수	Relative	Importance
경제활동 상태	1.000	4827.2
연령대	0.813	3926.6
최종학력	0.504	2432.8
3개 권역	0.425	2050.2
배우자 유무	0.355	1712.7
성별	0.334	1613.6
의무시간		
공통변수	Relative	Importance
경제활동 상태	1.000	19829.0
연령대	0.620	12300.0
성별	0.283	5612.2
배우자 유무	0.246	4877.0
최종학력	0.223	4427.8
3개 권역	0.164	3247.3
여가시간		
공통변수	Relative	Importance
경제활동 상태	1.000	15077.3
연령대	0.607	9157.8
성별	0.351	5288.2
최종학력	0.240	3618.9
배우자 유무	0.210	3158.8
3개 권역	0.185	2790.9
운동시간		
공통변수	Relative	Importance
경제활동 상태	1.000	2140.4
연령대	0.864	1848.7
성별	0.857	1833.7
최종학력	0.550	1176.8
3개 권역	0.472	1010.8
배우자 유무	0.373	799.1

자료: 통계청. 2019. 2019년 생활시간조사 데이터를 분석함.

그리고 블록화 기법을 사용하였는데, 이는 컴퓨터 계산속도와 메모리 저장에 있어서 효율적이기 때문이다. 블록화 변수는 ‘성별’로 선정하여 성별이 같은 집단에서만 기부 개체를 선택하였다.

IV. 통합데이터 품질 평가

한국의료패널조사와 생활시간조사 데이터를 통합한 후 통합데이터에 대한 품질 평가를 실시하였다. 통합데이터의 품질 평가는 통합데이터에서의 공통변수 분포 및 통합데이터와 제공데이터(생활시간조사)의 유일변수 분포 비교를 통해 이루어졌다.

먼저 통합데이터에서의 공통변수 분포는 제공데이터와 일치하였다(<부도 1>~<부도 6>참조). 매칭변수로 사용된 연령대, 최종학력, 경제활동 상태의 경우 2개 데이터에 대한 분포도 정확하게 일치하는 것으로 나타났다. 기증자 집단으로 활용된 성별 역시 동일하였다. 반면에 배우자 유무와 3개 권역은 두 가지 데이터의 분포에서 약간 차이가 났으나, 데이터 통합을 실시하기 전에 비해 범주별 차이가 줄어들음을 확인할 수 있다.

다음은 통합데이터와 제공데이터(생활시간조사)의 유일변수 분포를 비교·분석한 것으로, 상자그림과 확률밀도함수그림을 통해 살펴보았다. 먼저, 필수시간의 경우 통합데이터와 제공데이터의 분포가 전반적으로 유사한 편으로 나타났다. 통합데이터의 평균은 690분(11시간 30분)이고, 제공데이터는 682분(11시간 22분)으로 8분 적었고 표준편차는 비슷하였다(<부도 7> 참조).

의무시간의 경우 통합데이터의 분포가 450분(7시간 30분) 이하 부분이 제공데이터에 비해 두텁게 나타났고, 중위수는 낮은 편이었다. 이는 한국의료패널조사의 응답자 연령이 높은 편에 기인한 것으로 보인다. 통합데이터의 평균은 456분(7시간 36분)이고, 제공데이터는 482분(8시간 2분)으로 26분 적었으나 표준편차는 유사하였다(<부도 8> 참조).

여가시간의 경우 통합데이터의 분포가 250분(4시간 10분) 이상 부분이 제공데이터에 비해 약간 두텁게 나타났다. 상자그림을 보면 통합데이터는 제공데이터에 비

해 중위수는 큰 편이었고, 구간의 길이는 긴 편이었으며, 특이치는 적은 편에 속하였다. 앞에서 살펴본 의무시간처럼 한국의료패널조사의 응답자 연령이 높은 것 때문이라고 생각한다. 통합데이터의 평균은 294분(4시간 54분)이고, 제공데이터는 275분(4시간 35분)으로 19분 많았으나 표준편차 및 중위수는 유사하였다(<부도 9> 참조).

운동시간의 경우 통합데이터와 제공데이터에 대한 확률밀도함수는 매우 비슷하게 나타났다. 상자그림을 보면 통합데이터의 중위수는 0분으로 제공데이터와 동일하였으며, 구간의 길이는 유사한 편이었고, 특이치는 적은 편에 속하였다. 통합데이터의 평균은 34분이고, 제공데이터는 32분으로, 2분 차이로 크지 않았다(<부도 10> 참조).

V. 통합데이터 분석결과

주관적 건강상태에 따라 시간 사용 소비행태의 차이가 존재하는지에 대해 일원 배치 분산분석방법을 사용하여 분석하였다.

보통 건강행태는 성별과 연령별로 다르게 나타나므로, 주요 분석을 실시하기 전에 성별과 연령대별을 고려하여 시간 사용 소비에 차이가 있는지를 살펴보았다. 성별은 남자와 여자로, 연령대는 20, 30대/40, 50대/60대 이상으로 구분하였다. 성별과 연령대의 교호작용 효과가 통계적으로 유의한 결과를 보여 성별에 따라 연령 차이가 존재한다고 볼 수 있었다(<표 9> 참조).

그래서 성별에 따라 연령대별 차이가 있는지 살펴보았다. <부표 1>을 보면 필수 시간, 의무시간, 여가시간 및 운동시간 사용은 모두 유의수준 5%에서 20, 30대/40, 50대/60대 이상 남자 간에 통계적으로 유의한 차이가 있는 것으로 나타났다. 또한 20, 30대/40, 50대/60대 이상 여성도 같은 결과를 보였다.

<표 9> 시간 사용에 대한 성별과 연령대별 집단 차이 결과

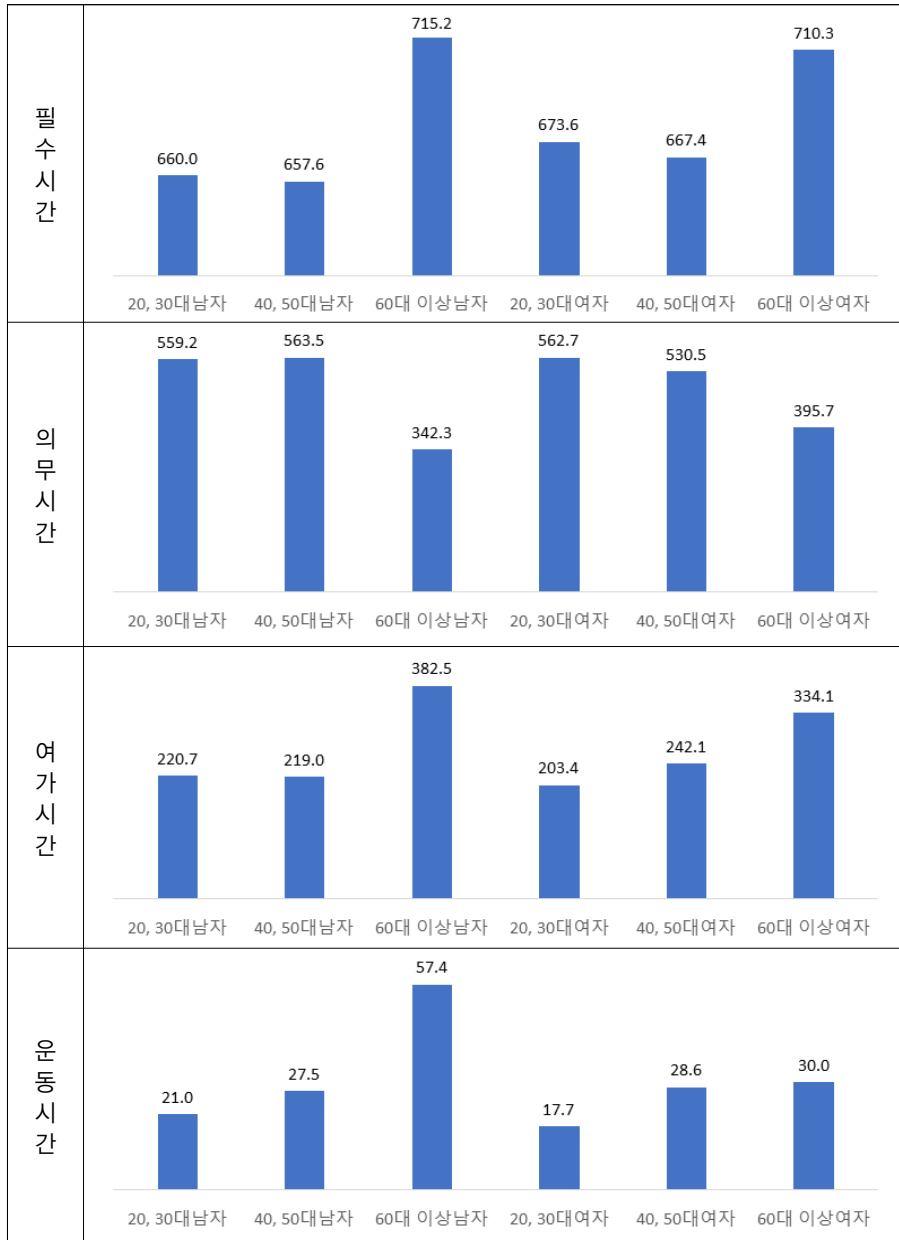
	필수시간		의무시간		여가시간		운동시간	
	추정치	표준 오차	추정치	표준 오차	추정치	표준 오차	추정치	표준 오차
절편	710.264***	1.998	395.657***	3.404	5.649***	0.014	1.661***	0.037
성별								
남자	.908	3.041	-53.406***	5.182	0.116***	0.022	0.836***	0.057
연령대								
20, 30대	-36.695***	4.139	167.011***	7.052	-0.618***	0.030	-0.625***	0.077
40, 50대	-42.906***	3.253	134.870***	5.542	-0.401***	0.023	-0.106***	0.061
성별 * 연령대								
남자*20, 30대	-18.429***	6.442	49.949***	10.976	.003	0.046	-0.752***	0.121
남자*40, 50대	-14.711***	4.993	86.367***	8.506	-0.197***	0.036	-0.965***	0.093

주: 1) ()은 유의수준을 나타내며 *는 10%이고, **는 5%이고, ***는 1%를 의미함.

2) 여가시간과 운동시간은 로그변환하여 분석한 결과를 나타냄.

자료: 한국보건사회연구원·국민건강보험공단. 2019. 한국의료패널조사 2019년 연간데이터를 사용한 모의실험용 데이터를 분석함.

<그림 12>를 보면, 평균 필수시간의 경우 성별에 상관없이 연령대가 높아질수록 평균 필수시간 사용이 많아지는 경향을 보였다. 40, 50대 남자가 필수시간을 658분(10시간 58분) 사용하여 가장 적은 편이었고, 60대 이상 남자가 715분(11시간 55분)으로 가장 많았다. 평균 의무시간 사용의 경우 40, 50대 남자와 20, 30대 여자가 564분(9시간 24분) 및 563분(9시간 23분)으로 많은 편에 속하였고, 60대 이상 남자가 342분(5시간 42분)으로 가장 적은 편이었다. 다음은 평균 여가시간으로 60대 이상 남자가 383분으로 가장 많이 사용하였고, 20, 30대 여자(203분, 3시간 23분), 40, 50대 남자(219분, 3시간 39분), 20, 30대 남자(221분, 3시간 41분)가 적게 사용하는 것으로 나타났다. 마지막으로 운동시간 사용의 경우 60대 이상 남자가 57분으로 가장 많은 편에, 20, 30대 여자가 18분으로 가장 적은 편에 속하였다. 성별과 관계없이 연령대가 높아질수록 운동하는 시간이 많아지는 것으로 나타났다.



<그림 12> 성별과 연령대별 집단의 생활시간 평균 비교 - 필수, 의무, 여가, 운동시간
(단위: 분)

자료: 이해정, 이기호, 안수인, 임종호, 이상혁, 조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. pp.176-178. [그림 5-24]~ [그림 5-27]을 재구성함.

이렇듯 성별에 따라 연령대별 생활시간 사용 행태는 다른 것으로 나타나, 주관적 건강상태에 따른 생활시간 사용 행태를 성별과 연령대를 고려하여 6개 집단으로 구분하여 각 집단별로 살펴보았다(<표 10> 참조). 먼저 평균 필수시간의 경우 60대 이상 남자와 60대 이상 여자의 경우 주관적 건강상태가 좋음, 보통, 나쁨에 따라 유의수준 1%에서 통계적으로 유의한 차이를 보였다. 주관적 건강상태가 좋음에서 나쁨으로 갈수록 평균 필수시간 사용이 많아지는 것으로 나타났으며 다른 집단에 비해 사용 시간 증가폭도 큰 편이었다. 60대 이상 남자의 경우 주관적 건강상태가 ‘좋음’인 집단은 708분(11시간 48분)인 반면에, ‘나쁨’인 집단은 31분 많은 739분(12시간 19분)을 사용하였다.

다음 평균 의무시간의 경우 40, 50대 남자, 50대 이상 남자, 60대 이상 여자의 경우 주관적 건강상태가 좋음, 보통, 나쁨에 따라 유의수준 1%에서 통계적으로 유의한 차이를 보였다. 주관적 건강상태가 좋음에서 나쁨으로 갈수록 평균 의무시간 사용이 적어지는 편에 속하였고 다른 집단에 비해 사용 시간 감소폭도 큰 편이었다. 40, 50대 이상 남자의 경우 주관적 건강상태가 ‘좋음’인 집단은 573분(9시간 33분)인 반면에, ‘나쁨’인 집단은 1시간 33분 적은 480분(8시간)을 사용하였다. 보통, 중·장년층의 경우 만성질환 발병 등 건강 문제가 발생하기 시작하는 시기이다. 즉 건강 문제가 발생하여 주관적 건강 상태가 좋지 않은 것으로 인지하는 집단이 질병 치료를 위해, 또는 질병으로 인한 활동 제한으로 일이나 가사 등의 활동에 사용하는 시간이 감소하여 이러한 결과가 나타난 것으로 보인다. 한편 40, 50대 여자는 유의수준 10%에서 통계적으로 유의한 차이를 보였다.

평균 여가시간의 경우 40, 50대 남자, 60대 이상 남자, 60대 이상 여자의 경우 주관적 건강상태가 좋음, 보통, 나쁨에 따라 유의수준 1%에서 통계적으로 유의한 차이를 보였다. 앞에서 살펴본 의무시간과 반대의 경향을 나타냈다. 이는 의무시간이 줄어든 만큼 여가시간을 보내는 비중이 높아진 결과로 볼 수 있다. 40, 50대 이상 남자의 경우 주관적 건강상태가 ‘좋음’인 집단은 211분(3시간 31분)인 반면에, ‘나쁨’인 집단은 1시간 23분 많은 294분(4시간 54분)을 사용하였다.

마지막으로 평균 운동시간의 경우 40, 50대 남자만이 주관적 건강상태가 좋음, 보통, 나쁨에 따라 유의수준 1%에서 통계적으로 유의한 차이가 있었다. 앞에서 살펴본 의무시간과 반대의 경향을 나타냈다. 40, 50대 이상 남자의 경우 주관적 건강상태가 ‘좋음’인 집단은 29분인 반면에, ‘나쁨’인 집단은 9분 많은 38분을 사용하였다. 이를 통해 주관적 건강상태가 좋지 않은 집단이 건강관리를 위해 운동시간을

<표 10> 주관적 건강상태에 따른 집단별 생활시간에 대한 평균 - 필수, 의무, 여가, 운동 시간

(단위: 분)

성별	연령대	주관적 건강상태	필수시간			의무시간			여가시간			운동시간		
			평균	표준편차	$Pr > F$	평균	표준편차	$Pr > F$	평균	표준편차	$Pr > F$	평균	표준편차	$Pr > F$
남자	20, 30대	좋음	657.1	99.1	0.7752	555.7	189.2	0.8738	227.2	152.1	0.5953	21.8	45.6	0.9357
		보통	662.9	100.3		563.7	176.3		213.4	146.0		19.5	39.7	
		나쁨	662.3	112.3		557.1	203.8		220.6	153.1		23.7	46.0	
	40, 50대	좋음	655.8	99.0	0.5482	573.1	160.7	<.0001	211.1	130.5	<.0001	29.4	52.2	0.0088
		보통	657.1	102.6		573.7	154.1		209.3	125.6		23.8	48.6	
		나쁨	666.4	123.7		480.1	224.2		293.5	187.7		38.0	58.5	
	60대 이상	좋음	707.9	116.4	<.0001	364.5	227.0	<.0001	367.7	190.0	<.0001	55.1	74.7	0.0616
		보통	710.7	113.3		351.7	217.6		377.4	183.6		57.3	72.7	
		나쁨	738.6	119.7		280.1	204.3		421.3	179.9		61.8	73.9	
여자	20, 30대	좋음	671.9	111.6	0.9181	560.6	161.7	0.9085	207.5	127.8	0.1111	16.6	38.0	0.8611
		보통	674.9	104.2		563.0	174.3		201.3	138.7		17.9	39.2	
		나쁨	674.1	102.9		568.8	159.7		197.1	133.9		20.7	45.7	
	40, 50대	좋음	666.1	103.5	0.5909	528.9	170.5	0.0697	245.0	147.2	0.2019	30.1	51.7	0.7890
		보통	666.7	103.9		536.9	169.9		236.4	142.1		27.3	45.9	
		나쁨	674.0	107.9		508.1	167.6		257.9	137.0		29.9	47.0	
	60대 이상	좋음	698.0	106.9	<.0001	422.9	184.7	<.0001	319.1	160.8	<.0001	32.4	52.1	0.6787
		보통	706.0	113.9		412.8	183.0		321.2	161.7		29.9	46.8	
		나쁨	725.4	118.3		352.0	179.6		362.6	163.4		28.3	45.4	

주: 여가시간과 운동시간의 경우 분산분석시 로그변환하여 분석한 결과를 나타냄. 참고로 크루스칼-왈리스 검정으로 분석하면 여가시간에서 40, 50대 여자가 유의수준 5%에서 통계적으로 유의한 차이를 가졌으나, 다른 집단은 동일한 결과를 보였음.

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. pp.179-183.

<표 5-11> ~ <표 5-14>를 재구성함.

더 많이 사용한 것으로 추정된다. 남자의 경우 다른 연령층도 주관적 건강상태가 좋지 않은 집단이 좋은 집단보다 운동시간의 사용이 많은 편으로 나타났다. 그러나 40대 이상 여자의 경우 통계적으로 유의성은 없으나 주관적 건강상태가 나쁠수록 운동시간이 감소하는 경향이 나타났다. 건강상태 인식과 운동시간 사용의 관계에서 남녀 간에 다른 경향이 나타나는 것으로 보였다. 한편 60대 이상 남자는 유의수준 10%에서 통계적으로 유의한 차이를 보였다.

VI. 요약 및 결론

한국의료패널조사와 생활시간조사 데이터를 정확 거리 함수에 기반을 둔 최근접 이웃 핫택 방법으로, 매칭변수는 연령대, 최종학력, 경제활동 상태를 활용하여 데이터 통합을 실시하였다. 통합데이터 분석의 주요 결과를 보면, 성별에 따른 연령대별 생활시간(필수, 의무, 여가, 운동시간) 사용 행태는 통계적으로 유의한 차이가 있는 것으로 나타났다. 주관적 건강상태와 생활시간 사용 간의 관계에 대한 결과는 다음과 같다. 주관적 건강상태가 좋음, 보통, 나쁨에 따라 60대 이상 남녀는 평균 필수, 의무, 여가시간이 통계적으로 유의한 차이를 보였다. 40, 50대 남자의 경우 주관적 건강상태에 따라 평균 의무, 여가, 운동시간이 통계적으로 유의한 차이를 보였다. 한편 40, 50대 여자는 평균 의무시간이 통계적으로 유의한 차이가 있었다. 보통, 중·장년층의 경우 만성질환 발병 등 건강 문제가 발생하기 시작하는 시기이다. 즉 건강 문제가 발생하여 주관적 건강 상태가 좋지 않은 것으로 인지하는 집단이 질병 치료를 위해, 또는 질병으로 인한 활동 제약으로 일이나 가사 등의 활동에 사용하는 시간이 감소하여 이러한 결과가 나타난 것으로 보인다. 한편 주관적 건강상태가 좋지 않은 집단이 건강관리를 위해 운동시간을 더 많이 사용한 것으로 추정된다. 남자의 경우 다른 연령층도 주관적 건강상태가 좋지 않은 집단이 좋은 집단보다 운동시간의 사용이 많은 편으로 나타났다. 그러나 40대 이상 여자의 경우 통계적으로 유의성은 없으나 주관적 건강상태가 나쁠수록 운동시간이 감소하는 경향이 나타나 건강상태 인식과 운동시간 사용의 관계에서 남녀 간에 서로 다른 경향이 나타나는 것으로 보였다.

또한 주관적 건강상태와 생활시간 사용과의 관계를 분석한 연구결과는 선행연구

와 유사한 결과를 가졌다. 예를 들면 근로시간이 증가할수록 주관적 건강상태가 좋지 않다는 연구(유혜림 2018, 김현규 외 2018; Spiegelaele & Piasna 2017; Bannai & Tamakoshi 2014), 고령자의 운동 참여 집단과 비참여 집단 간 주관적 건강상태는 참여 집단이 높다는 연구(남연희·남지란 2011, 이정숙·이인수 2005, 김남진 2000) 등의 선행연구 결과가 있다. 이를 통해 통계적 매칭 방법을 이용한 데이터 통합의 활용 가능성을 확인하였다. 보통 국민의 건강상태와 건강행동을 파악하기 위해 가장 많이 쓰는 방법은 설문조사이다. 그러나 응답자의 피로도, 거부감 등과 같은 이유로 설문조사를 통해서만은 제한된 정보만을 파악할 수 있다. 그런데 이 연구에서 생활시간조사를 활용한 것처럼, 다양한 데이터와의 통합을 통해 건강 상태, 건강행동에 영향을 미치는 것으로 알려진 여러 정보를 추가적으로 얻을 수 있다. 향후, 보다 다양한 데이터를 통합한다면 기존 보건 분야에서 고려가 미흡했던 여러 가지 정보를 사용할 수 있게 되어 국민의 건강에 영향을 미치는 여러 요인을 파악하고 이에 대한 대응책을 마련해 볼 수 있다고 생각한다.

한편 통계적 매칭 방법은 여러 가지 장점이 있다. 기존에 있는 데이터를 통합할 때 통합하는 데이터 간 오버랩을 필요로 하지 않기 때문에 기존 데이터에 바로 적용할 수 있다는 점, 개인정보보호 및 비밀유지의 가치를 보호할 수 있다는 점, 그리고 유사한 분포를 유지하는 대상들을 통합하기 때문에 통합데이터에 대한 분석 결과를 신뢰할 수 있다는 점 등이 있다. 통계적 매칭 방법은 자료연계처럼 정확한 정보를 주는 것은 아니지만 유사한 분포를 유지하는 대상들을 통합하기 때문에 통합데이터에 대한 분석결과도 신뢰할 수 있다(박근화 외 2018, p.180). 이러한 측면에서 보면, 통계적 매칭 방법은 데이터를 통합할 때 매우 유용하며 활용 가능성도 높다고 볼 수 있다. 그러나 통합데이터를 구축하기는 쉽지 않은 편이며, 통계적 매칭의 정확성을 고려해야 할 것이다. 이를 위해 통계적 매칭 방법을 활용하여 조사데이터를 통합할 때 고려해야 할 사항은 다음과 같이 4가지로 볼 수 있다.

첫 번째, 기준데이터와 제공데이터의 연구모집단 대상이 동일해야 한다. 각각의 조사는 조사목적에 따라 조사 응답자 대상이 다르게 구성되기 때문에 조사데이터를 통합할 때 일치시키는 과정이 필요하다. 한국의료패널조사에서는 19세 이상에게만 질문하는 항목을 유일변수 중 하나로 선정하였다. 그래서 한국의료패널조사의 연령 조건을 생활시간조사에서도 동일하게 맞춰주었다.

두 번째, 기준데이터와 제공데이터의 작성주기와 조사시점 등의 확인이 필요하다. 조사에 따라 작성주기와 조사시점이 다르므로 조사데이터 간 통합을 할 때 조사시점

을 동일하게 맞추는 것이 좋다. 실증분석에서 조사시점은 2019년으로 일치하였다.

세 번째, 공통변수를 다양화 및 표준화해 주는 과정이 필요하다. 기준데이터와 제공데이터를 통합할 때 중요한 역할을 하는 매칭변수는 공통변수 중에서 유일변수와 연관성이 높은 변수로 선정해야 한다. 그래서 공통변수가 다양하고 많을수록 최적의 매칭변수를 선정할 수 있다. 또한 기준데이터와 제공데이터의 공통변수에 대해 표준화를 실시하여 동일한 형태를 가지도록 구성해야 한다. 실증분석에서는 공통변수가 6개였는데, 최종학력의 경우 교육 수준과 졸업 유무에 대해 생활시간조사가 더 세세한 범주를 가지고 있었다. 그래서 한국의료패널조사의 범주로 일치시켰다. 한편, 2개 데이터에서 가구소득 문항이 있었는데 시점이 같지 않아서 공통변수로 활용할 수 없는 경우도 있었다. 이러한 측면에서 보면 조사의 목적이 다르더라도 공통변수로 활용할 수 있는 문항(예: 인구사회학적, 소득 등의 변수)은 될 수 있으면 유사한 형태로 구성하는 방안을 생각해 볼 수 있다.

마지막으로, 데이터 통합에 영향을 줄 수 있는 매칭 방법 및 매칭변수 선정과 관련한 사전적 검토의 중요성이다. 통계적 매칭 방법을 활용한 데이터 통합은 매칭 방법과 매칭변수 선정에 따라 통합데이터의 품질이 다를 수 있으므로, 사전 검토를 통하여 최적의 조합을 마련해야 한다. 모의실험 결과를 통하여 그 중요성을 확인한 바 있다.

다출처 조사데이터 통합은 여러 가지 장점이 있다. 새로운 조사를 계획하고 수행하는 데 필요한 시간과 새로운 조사의 수행을 위해 발생하는 금전적 비용도 줄일 수 있다. 또한 다양한 정보를 수집하기 위하여 많은 설문 문항을 포함함으로써 발생할 수 있는 새로운 조사데이터에 대한 품질 저하의 위험도 크지 않은 편이다. 응답자 측면에서는 추가 조사로 발생하는 응답 부담 가중을 줄일 수 있다. 그러므로 다출처 조사데이터 통합을 통해 조사데이터의 활용도를 높일 수 있다고 생각한다.

참고문헌

- 김남진. 2000. “규칙적인 운동참여가 고령자의 주관적 건강상태 및 사회적 활동에 미치는 영향.” 《한국체육학회지》 39(1): 149-158.
- 김대건·박민규. 2014. “마이크로 매칭 방안 비교 연구 - 경제활동인구조사와 생활시간조사 자료에 적용.” *Journal of the Korean Data Analysis Society* 16(5B): 2393-2408.
- 김현규·서유리·조교영. 2018. “제조업 근로자의 근무환경이 건강상태에 미치는 영향.”

- 《한국데이터정보과학지》 29(6): 1555-1563.
- 남연희·남지란. 2011. “노인의 주관적인 건강상태에 영향을 미치는 요인에 관한 연구.” 《한국가족복지학》 16(4): 145-162.
- 박근화·장훈·한정임·김정림·김희경·김광섭. 2018. 《문화·체육·관광 데이터 연계를 통한 빅데이터 생산 및 활용방안 연구》. 한국문화관광연구원.
- 신지형, 오윤석. 2018. 《방송미디어 분야 자료의 통합연계 및 활용 방안 연구》. 정보통신정책연구원.
- 오미애·최현수·김용대·이용희·진재현. 2014. 《보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안》. 한국보건사회연구원.
- 유혜림. 2018. “근로시간이 주관적 건강상태에 미치는 영향: 내생적 순서형 프로빗 모형을 중심으로” 《노동정책연구》 18(1): 71-102. 한국노동연구원.
- 윤석열 정부 120대 국정과제. 2022.7. 대한민국정책브리핑.
<https://www.korea.kr/archive/expDocView.do?docId=40075>. 2022.9.27. 인출.
- 이영섭·김선웅·안홍엽·임경은·김희경. 2009. “통계조사자료와 행정자료 간의 통계적 매칭기법에 관한 연구.” 《통계연구》 14(1): 82-98.
- 이정숙·이인수. 2005. “노년기 건강관리 행동과 사회 경제 요인이 건강 상태에 미치는 영향.” 《노인복지연구》 27(봄): 231-254. 한국노인복지학회.
- 이혜정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. 한국보건사회연구원.
- 통계청. 2019. 생활시간조사 행동분류표.
- 통계청. 2019.11. 『생활시간조사』 통계정보보고서.
- 통계청. 2020. 2019년 생활시간조사 결과.
- 한국보건사회연구원·국민건강보험공단. 2022. 2기 한국의료패널 2019년 연간데이터 유저가이드.
- Bannai, A. and A. Tamakoshi. 2014. “The Association between Long Working Hours and Health: A Systematic Review of Epidemiological Evidence.” *Scandinavian Journal of Work, Environment & Health* 5-18.
- De Spiegelaere, S. and A. Piasna. 2017. *The Why and How of Working Time Reduction*. European Trade Union Institute.
- D'Orazio, M. 2022. Package ‘StatMatch’.
<https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf>.

<접수 2024.07.07; 수정 2024.08.06; 게재확정 2024.09.02>

Empirical Analysis and Implications of Multi-source Data Integration Using Statistical Matching Methodology: Based on the Korea Health Panel Survey

Hyejung Lee

(Korea Institute for Health and Social Affairs)

In line with the data open trend, a large amount of data is being opened. However, they are individually distributed and provided so that use of the multiple data is not well linked and integrated among themselves.

The purpose of this study is to examine the linkage and integration of multi-source data dispersed in various places in order to increase the utilization value, and to prepare a plan to solve the difficulties and to activate them. Data integration and analysis were conducted by using real data. The Korea Health Panel Survey data and Life Time Use Survey data were integrated by using the nearest neighborhood hot deck method based on the exact distance function. The purpose of integrating the two data was to identify the possibility of creating integrated data on multi-source surveys, and to analyze the relationship between health behavior and life time use.

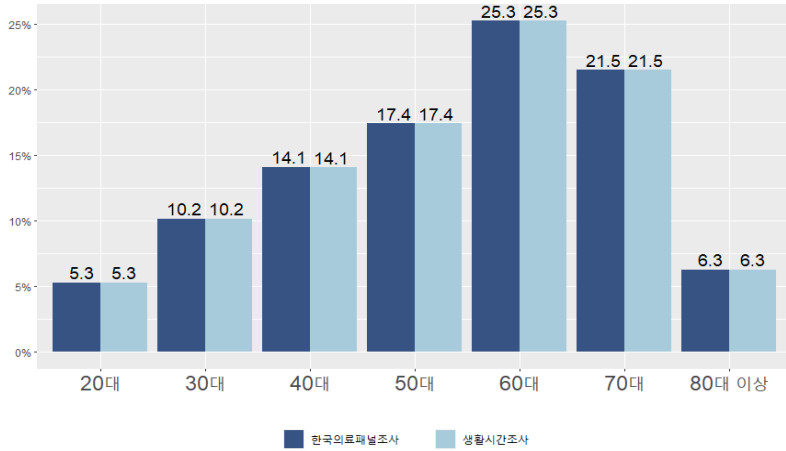
Key words: Korea Health Panel Survey, Life Time Use Survey, data integration, integrated dataset, statistical matching

부표

<부표 1> 성별과 연령대별 집단의 시간 사용에 대한 기초통계					(단위: 분)
시간	성별	연령대	평균	표준편차	$Pr > F$
필수	남자	20, 30대	660.0	100.8	<.0001
		40, 50대	657.6	103.7	
		60대 이상	715.2	116.2	
	여자	20, 30대	673.6	107.1	<.0001
		40, 50대	667.4	104.2	
		60대 이상	710.3	114.1	
의무	남자	20, 30대	559.2	185.1	<.0001
		40, 50대	563.5	167.8	
		60대 이상	342.3	220.7	
	여자	20, 30대	562.7	167.4	<.0001
		40, 50대	530.5	170.0	
		60대 이상	395.7	184.8	
여가	남자	20, 30대	220.7	149.6	<.0001
		40, 50대	219.0	137.7	
		60대 이상	382.5	186.2	
	여자	20, 30대	203.4	133.6	<.0001
		40, 50대	242.1	143.5	
		60대 이상	334.1	163.1	
운동	남자	20, 30대	21.0	43.2	<.0001
		40, 50대	27.5	51.3	
		60대 이상	57.4	73.7	
	여자	20, 30대	17.7	39.4	<.0001
		40, 50대	28.6	48.2	
		60대 이상	30.0	47.8	

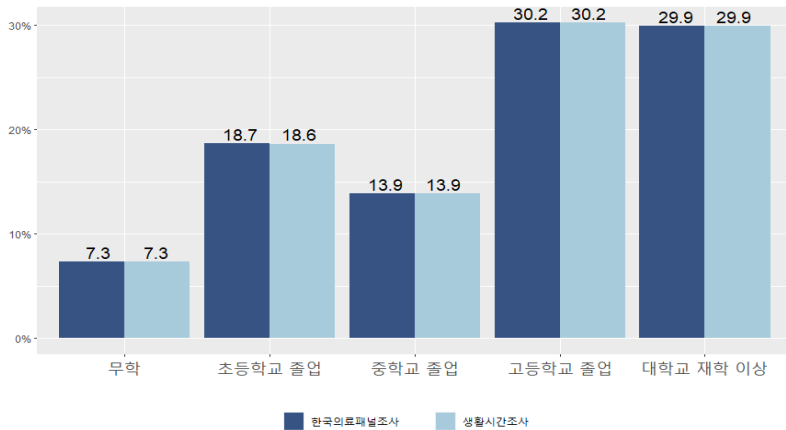
자료: 이해정, 이기호, 안수인, 임종호, 이상혁, 조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. pp.175-176. <표 5-10>을 재구성함.

부도



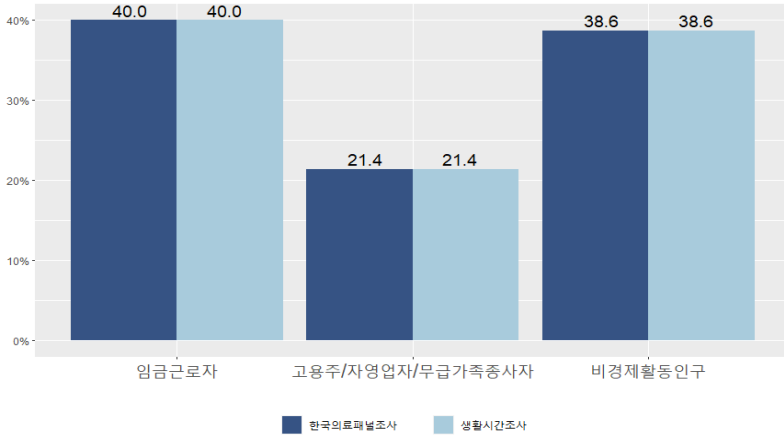
<부도 1> 통합데이터에서의 공통변수 분포: 연령대 (단위: %)

자료: 이해정·이기호·안수인·임중호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.166. [그림 5-13].



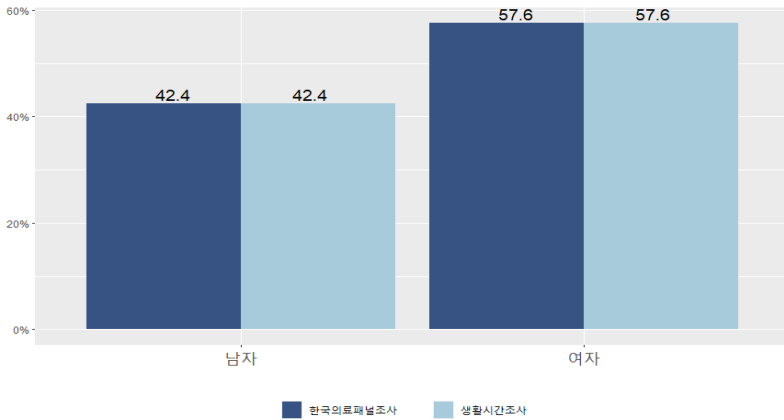
<부도 2> 통합데이터에서의 공통변수 분포: 최종학력(단위: %)

자료: 이해정·이기호·안수인·임중호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.167. [그림 5-14].



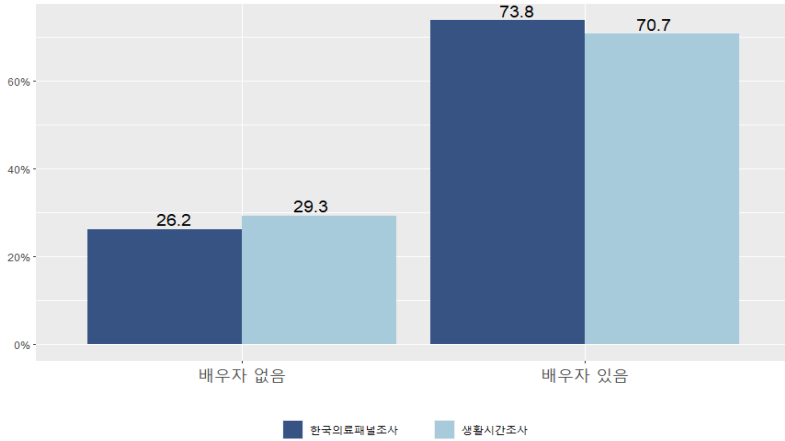
<부도 3> 통합데이터에서의 공통변수 분포: 경제활동 상태 (단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.167. [그림 5-15].



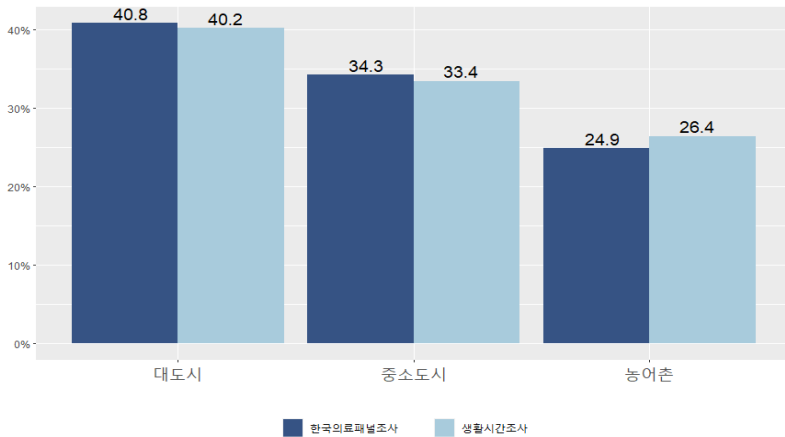
<부도 4> 통합데이터에서의 공통변수 분포: 성별 (단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.168. [그림 5-16].



<부도 5> 통합데이터에서의 공통변수 분포: 배우자 유무 (단위: %)

자료: 이해정·이기호·안수인·임중호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.168. [그림 5-17].

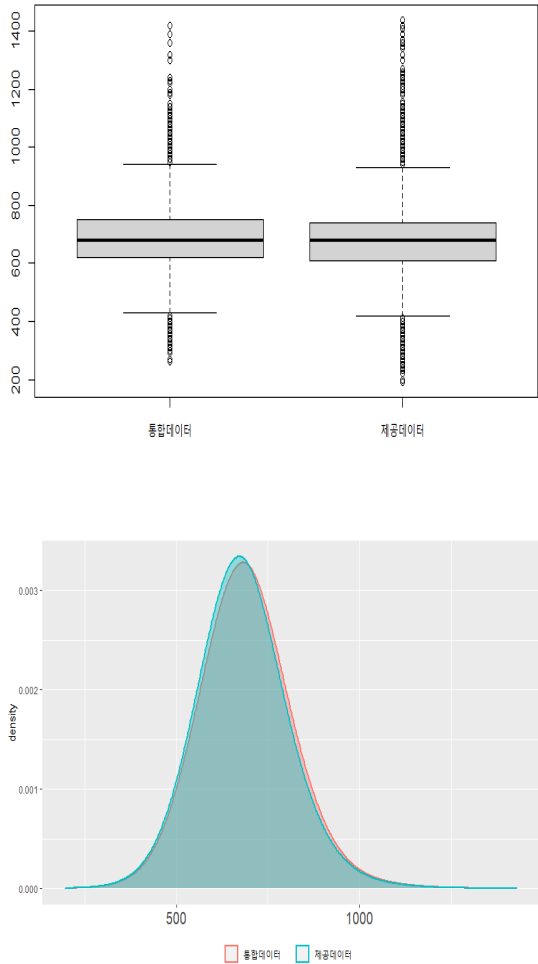


<부도 6> 통합데이터에서의 공통변수 분포: 3개 권역 (단위: %)

자료: 이해정·이기호·안수인·임중호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.169. [그림 5-18].

(단위: 분)

	평균	표준편차
통합데이터	690.05	112.66
제공데이터	682.02	111.36

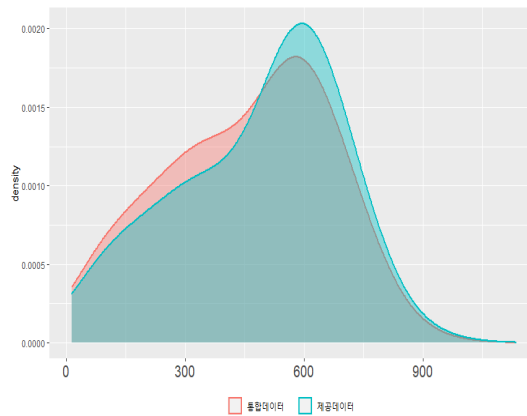
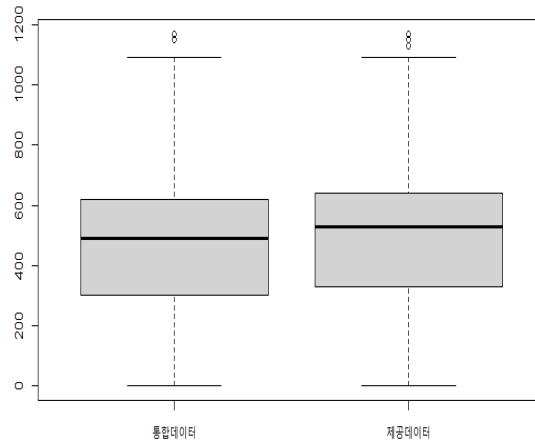


<부도 7> 통합데이터와 제공데이터의 필수시간 분포 비교 (단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.169. <표 5-6> 및 p.170. [그림 5-19]을 재구성함.

(단위: 분)

	평균	표준편차
통합데이터	455.88	208.50
제공데이터	482.73	208.10

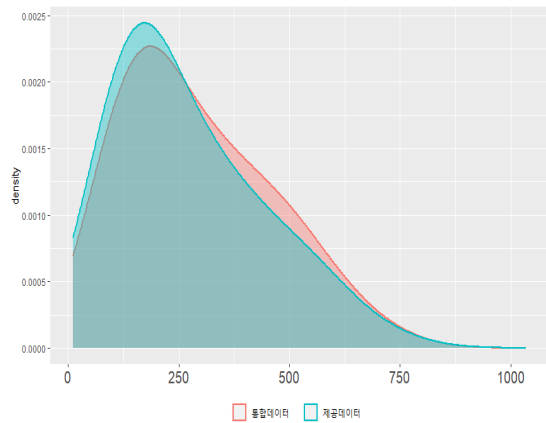
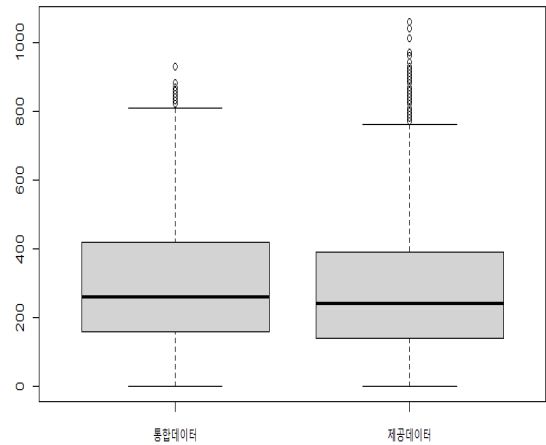


<부도 8> 통합데이터와 제공데이터의 의무시간 분포 비교 (단위: %)

자료: 이해정·이기호·안수인·임중호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.170. <표 5-7> 및 p.171. [그림 5-20]을 재구성함.

(단위: 분)

	평균	표준편차
통합데이터	294.02	172.82
제공데이터	275.21	172.62

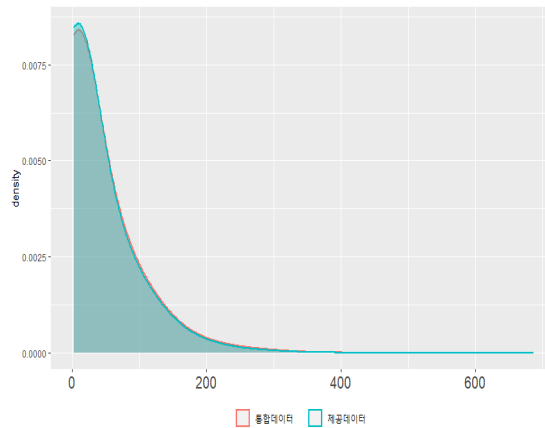
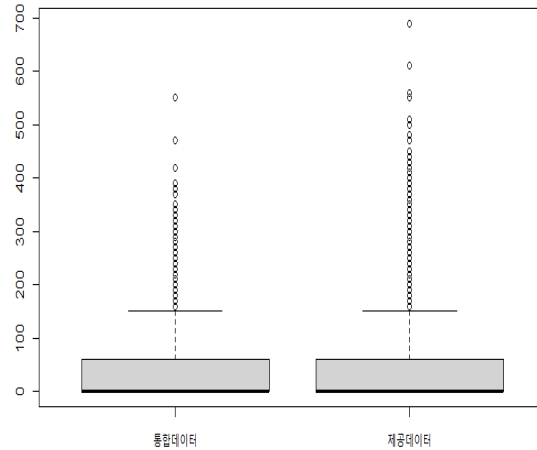


<부도 9> 통합데이터와 제공데이터의 여가시간 분포 비교 (단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.172. <표 5-8> 및 [그림 5-21]을 재구성함.

(단위: 분)

	평균	표준편차
통합데이터	34.00	56.01
제공데이터	31.79	54.35



<부도 10> 통합데이터와 제공데이터의 운동시간 분포 비교 (단위: %)

자료: 이해정·이기호·안수인·임종호·이상혁·조용찬. 2022. 《보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구》. p.173. <표 5-9> 및 [그림 5-22]을 재구성함.