이달의 초점

사회보장분야 행정데이터 구축·연계와 활용 및 활성화 방안

한국보건사회연구원

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

보건복지포럼 [2023. 11.] DOI: 10.23062/2023.11.2

사회보장 행정데이터 활용사례와 향후 과제

Cases of Using Social Security Administrative Data and Future Tasks

유종성 가천대학교 초빙교수 겸 불평등과사회정책연구소장

행정데이터는 일반적으로 행정 대상이 되는 전체 인구를 포괄하는 빅데이터로서 지역 단위 및 소규모 집단에 대한 연구를 가능하게 한다. 무응답과 기억의 오류 및 거짓 응답으로 인한 문제로부터 자유롭고 자연적으로 패널데이터 구조를 얻게 되는 장점이 있다. 그러나 행정데이터가 서베이데이터보다 부정확한 경우도 있으며, 개별 행정기관의 데이터는 용도가 매우 제한적이어서 다른 행정데이터 또는 서베이데이터와 연계가 이루어져야 큰 효용을 낼 수 있다. 증거 기반 정책을 위해, 특히 지역사회보장계획처럼 지역단위의 정책 평가와 정책 설계를 위해서는 여러 행정데이터와 서베이데이터의 연계가 필수적이다. 이 글에서는 이를 위한 법적, 행정적 개선 과제를 데이터 3법 개정 이후의 경험을 토대로 살펴본다.

1 들어가며

행정데이터는 행정기관이 행정 목적으로 개인이나 가구 등에 대한 정보를 수집하여 관리하는 데이터를 의미한다. 행정데이터는 연구 목적으로 수집된 자료가 아니라는 점에서 서베이자료와 근본적인 차이를 보인다.

사회과학 분야에서의 실증적 연구는 전통적으로 연구 목적을 위해 설계되고 수집된 서베이데이터를 기반으로 이루어졌다. 그러나 덴마크, 스웨덴, 노르웨이, 핀란드와 같은 북유럽 국가들은 상당히 오래 전부터 행정등록 자료를 사회과학과 정책 연구에 활용해 왔다. 미국과 영국을 비롯한 여러 나라도 1990년대 또는 2000년대부터 증거 기반 정책을 위한 연구를 위해 마이크로 행정데이터의 연계 및 활용을 위한 노력을 기울이기 시작했다. 최근 들어서는 빅데이터에 대한 관심과 함께 행정 빅데이터의 구축 및 연구 활용을 촉진하기 위해 여러 나라의 정부와 학계가 많은 자원을 투입하고 있다(Card, Chetty, Feldstein, and Saez, 2010; Connelly,

Playford, Gayle, and Dibben, 2016; Penner & Dodge, 2019; Song and Coleman, 2020; United Nations, 2007).

행정데이터는 서베이데이터에 비해 세 가지 중요한 이점이 있다(Card et al., 2010; 유종성, 2020). 첫째, 행정데이터는 일반적으로 표본 또는 부분 집단보다는 행정 대상이 되는 전체 인구를 포괄하는 빅데이터라는 점이다. 따라서 표본오차를 줄이고 지역 단위 및 소규모 집단이나 희귀한 현상에 대한 연구를 가능하게 한다. 둘째, 서베이데이터에서 흔히 나타나는 무응답과 거짓 응답, 축소 또는 과장 보고, 기억의 오류 등으로 인한 문제로부터 상대적으로 자유로워 고품질의 정보를 제공한다. 특히 설문조사에 대한 응답률 저하 추세와 계층별, 가구 유형별, 연령별 응답률의 차이는 표본의 편향(sample bias)을 키운다. 가령 소득 조사에서 최상위 소득층의 낮은 응답률과 축소 보고 경향, 1인 가구의 과소표집 등은 조사자료의 대표성과 신뢰성에 상당히 큰 문제를 일으키는데, 행정데이터는 이러한 문제로부터 자유롭다. 셋째, 일반적으로 동일한 단위의 대상(개인 또는 가구 등)을 오랫동안 계속해서 다루기 때문에 자연적으로 패널데이터 구조를 얻게 되어 종단 연구를 가능하게 한다. 패널 조사의 경우 비용이 많이 소요될 뿐만 아니라 표본의 유지가 어려운 데 반해 행정자료는 모집단 전수가 유지되는 장점이 있다. 일반적으로 횡단자료만으로는 정책 효과를 포함한 인과관계를 측정하기 어려운데, 행정자료는 모집단 전수에 대한 패널자료를 생성함으로써 이러한 연구를 가능하게 할 수 있다.

그러나 행정데이터가 서베이데이터보다 항상 더 정확한 정보를 제공하는 것은 아니다. 경우에 따라서는 서베이데이터가 보다 정확한 정보를 제공할 수도 있다. 가령 비공식부문에서 일하는 사람의 경우 행정자료에서는 이들의 근로소득이 파악되지 않는다. 이 경우에는 설문조사 자료가 보다 정확한 소득을 파악할 가능성이 크다. 자영업자도 국세청에 보고하는 소득보다 설문조사에 응답하는 소득이 보다 정확할 수 있다. 사적이전소득도 행정자료에는 나타나지 않는다. 또 행정데이터는 기관에 따라 또는 동일 기관 내에서도 행정 목적에 따라 개념 정의를 달리하고(가령 가구에 대한 정의가 기관별로 다름), 정책의 변화에 따라 동일한 변수의 측정 방법이 달라지는 등의 문제가 있다(유종성, 김민혜, 김승연, 유수진, 2021).

또한 각 기관에서 생성된 행정데이터는 다른 행정데이터나 서베이데이터와 연계되지 않으면 그 자체로는 연구 활용 가치가 적은 경우가 대부분이다. 서베이의 경우 연구 목적으로 설계하기 때문에 종속변수에 대한 여러 질문은 물론 설명변수와 통제변수들에 대한 질문이 포함된다. 따라서 이러한 변수들 간의 상관관계를 측정하는 것이 가능하지만, 각 기관의 행정데이터는 그 자체만으로는 큰 활용도를 가지지 못하는 경우가 많다. 따라서 증거 기반의 정책 연구를 위해서는 행정데이터와 서베이데이터의 장단점을 고려하여 보완적인 활용을 해야 한다. 여러 행정데이터 간의 연계, 행정데이터와 서베

이데이터 간의 연계가 데이터의 활용도를 높일 수 있다.

한국은 "ICT는 강국, 데이터는 후진국"(조슬기나, 2020)이라는 기사 제목이 가리키는 바와 같이 전자정부와 공공데이터 개방에서 세계 최상위권 평가를 받고 있음에도 마이크로 행정데이터의 연계와 활용은 뒤져 있다. 2020년 데이터 3법(개인정보보호법, 정보통신망 이용 촉진 및 정보보호 등에 관한 법률, 신용정보의 이용 및 보호에 관한 법률)의 개정 이후 통계와 과학적 연구를 위해서는 정보 주체의 동의 없이 가명정보를 사용한 데이터 간의 결합이 가능하게 되었지만, 아직 정부 부처와 기관들은 마이크로 행정데이터 제공과 데이터 결합에 소극적이다.

이에 이 글에서는 증거 기반 정책을 위해 행정데이터 연계 활용이 가지는 의의를 살펴보고, 필자가 그동안 지역사회 보장 정책의 수립과 평가에 필요한 연구를 위해 행정데이터의 연계 활용을 추진해 온 경험을 토대로 개선 방안을 제시해 보고자 한다.

2 증거 기반 정책을 위한 행정데이터의 연계 활용의 의의

국회 미래연구원(2019)에 의하면 2019년 현재 533개의 법정 중장기 계획이 부처별, 법령별로 수립·집행되고 있으나, 개별 계획의 목표와 추진 전략이 모호하거나 성과지표(performance indicators)가 적합하게 선정되지 않은 경우가 많다. 새로운 중장기 계획 수립 시 기존 계획의 성과평가 및 피드백이 제대로 되지 않고 있다. 특히 적절한 성과지표의 선정과 주기적인 측정은 사업의 효과성과 효율성을 평가하는 데 필수적인 것으로 증거 기반 정책(evidence-based policy)의 기초라고 할 수 있다.

공공부문 성과관리에서 많이 활용되는 논리모형은 투입(input), 활동(activity) 또는 과정(process), 산출(output), 초기 결과(short-term outcome), 장기 결과(long-term outcome) 개념을 사용하여 성과관리의 논리적 흐름을 제시한다. 즉 투입지표 또는 산출지표 위주의 형식적인 성과관리를 넘어 결과 지향적인(results-oriented) 성과관리를 지향한다. 가령 기초생활보장 생계급여 사업의 성과관리를 예로 들면 투입지표(기초생활보장 예산, 담당 복지공무원 인력과 시간의 투입)와 산출지표(생계급여 수급 가구와 인원수, 수급액 등)뿐만 아니라 적합한 결과지표(빈곤율, 빈곤갭, 빈곤탈출율 등)를 선정하여 성과 측정과 평가를 할 필요가 있다. 이를 통해 정책의 효과성(effectiveness)과 효율성(efficiency)을 높이도록 해야 한다. 특히 정책목표 달성에 관한 단기간의 효과를 넘어 중장기적인 효과를 측정·평가하고, 의도한 결과뿐만 아니라 의도하지 않은 결과에 대한 영향까지 측정·평가해야 한다(유종성, 2020; Kristensen, Groszyk and Bühler, 2002).

나아가 정책의 효과성과 효율성에 영향을 미친 조건과 요인들을 밝혀내야 한다. 또한 새로운 정책을 설계할 때 그 기대효과에 대한 시뮬레이션을 정

확하게 하는 것이 중요하다. 이러한 증거 기반 정책 연구를 위해서는 결과지표의 측정, 관련 영향변수 및 통제변수들에 대한 정보를 정확하게 제공하는 중장기 패널데이터의 확보가 무엇보다 중요하다. 그러나 기존 서베이데이터가 이러한 조건을 충족하기는 매우 어렵다.

예를 들어 기초생활보장 생계급여 수급 가구, 인원과 같은 산출지표는 행정의 즉각적인 결과로서 곧바로 측정된다. 그러나 빈곤 탈출 효과와 빈곤의 완화 같은 결과지표의 측정은 그리 쉽지 않다. 소득불평등을 나타내는 지니계수나 상대빈곤율이 사회보장의 결과지표로 사용되고 있으나, 이러한 지표들은 전국 단위에서만 발표될 뿐이다. 기초자치단체는 물론 광역자치단체 수준에서도 지역별 불평등이나 빈곤율과 같은 지표는 측정하지 않는다. 광역 및 기초자치단체들은 사회보장기본법에 따라 4년마다 지역사회보장계획을 수립하고 평가할 때 투입지표와 산출지표 위주의 중기계획만 반복하고 있다.

지역별 불평등과 빈곤 지표를 측정하지 못하는 이유는 자료의 결여에 있다. 즉 전국 단위의 불평등과 빈곤 지표는 통계청의 가계금융복지조사(이하 가금복) 자료에 근거해서 산출되는데, 이 자료로 지역별 지표를 생산하기에는 샘플의 수와 대표성이 모자라며 지역별로는 신뢰할 만한 가구 조사 수행이 어렵기 때문이다.

광역 및 기초자치단체의 사회보장 기본계획 수립 및 성과평가를 위한 결과지표의 측정 및 정책 효과에 대한 평가와 분석을 위해서는 기존의 서베이 자료만으로는 불가능하다. 행정데이터의 연계 활용이 필수적으로 요청된다. 불평등과 빈곤 지표는 물론 불평등과 빈곤에 영향을 주는 다른 경제사회적 변수들도 광역 및 기초자치단체 수준에서 측정할 수 있어야 한다. 개별 사회보장 정책들의 빈곤 완화 효과를 평가하기 위해서는 주요 사회보장 정책의 투입 및 산출에 관한 변수들과 함께 개인 및 가구 단위의 소득과 고용 등에 관한 중장기 패널데이터가 필요하다. 사회보장급여의 수급자뿐만 아니라 비수급자도 포함되어야 하며, 빈곤 진입과 탈출의 과정을 상당 기간 추적할 수 있어야 한다.

행정자료가 만능은 아니다. 앞에서 언급한 바와 같이 비공식부문의 소득활동과 사적이전소득은 행정자료에는 나타나지 않는다. 오히려 설문조사 자료에서 이를 구할 수 있다. 자영업자의 사업소득도 국세청에는 축소 신고되는 경향이 있어 설문조사 자료가 보다 정확한 실상을 나타낼 수 있다. 또한 태도변수 등은 설문조사 자료에 의존할 수밖에 없다. 따라서 행정데이터 간의 연계뿐 아니라 행정데이터와 서베이데이터의 연계도 필요한 경우가 많다.

3 지역 단위 소득분배 측정 및 사회보장 정책 효과 평가를 위한 행정데이터 연계 활용의 필요성

증거 기반 정책 수립을 위한 행정데이터와 서베이데이터 연계의 필요성을 실제 소득분배에 관한

연구의 예를 들어 설명하고자 한다. 가천대 불평등과사회정책연구소는 지역별 소득분배 지표를 측정하기 위해 서울연구원과의 협동연구(김승연, 유종성, 최광은, 이혜림, 전병유, 정준호, …, 배세진, 2020; 김승연, 유종성, 박종헌, 전병유, 정준호, 김동진, …, 배세진, 2022)로, 그리고 성남시의 의뢰(유종성, 박종헌, 구정은, 최혜은, 배세진, 2022)로 전 국민 소득·재산 자료를 보유하고 있는 건강보험공단(이하 건보공단)의 자료를 분석하였다. 건보공단 자료 공유 서비스는 일반적으로 소득과 재산 자료를 공개하지 않고 소득의 대리변수로 건강보험료 10분위 정보를 제공한다. 그러나 사회보장급여에 관한 법률 및 개인정보보호법에 근거하여 사회보장기관인 지방자치단체(서울시 및 성남시)의 연구 수행자는 건보공단이 보유한 가명정보의 소득재산 자료를 접근할 수 있다. 다만 건보공단의 데이터분석센터 내에서 분석하고, 원자료가 아닌 분석 결과만을 익명성 심사를 거쳐 반출할 수 있다.

〈표 1〉에 나타난 가금복과 건보공단의 소득자료를 비교해 보면 2019년 귀속 1인당 평균소득(건보공단의 근로소득, 사업소득, 이자배당소득, 기타소득, 공적연금의 계, 가금복의 근로소득, 사업소득, 재산소득, 공적연금의 계)이 건보공단 자료상으로는 1759만 9000원, 가금복 자료상으로는 2114만 2000원으로 건보공단 자료의 소득 파악률이 가금복의 83.2%에 그쳤다. 이는 건보공단의 소득자료에 일용근로소득, 연 1000만 원 이하의 이자배당소득, 분리과세 기타소득 등이 빠져 있고, 비공식부문의 근로소득이 없기 때문으로 보인다.

가구당 소득은 건보공단 자료(주민등록세대 기준)는 3955만 9000원, 가금복 자료(가구 기준)는 5583만 8000원으로 건보의 가구소득이 가금복의

[표 1] 건강보험공단, 가계금융복지조사 및 경제활동인구조사의 2019년 귀속 소득 비교

(단위: 천 원, %)

	건강보험공단(A)	가계금융복지조사(B)	A/B	경제활동인구조사(C)	A/C
1인당 개인소득	17,599	21,142	83.2%		
1인당 근로소득	13,699	14,711	93.1%		
1인당 사업소득	2,901	5,077	57.1%		
1인당 금융소득	284	691	41.1%		
1인당 공적연금소득	716	663	108.0%		
가구소득	39,559	55,838	70.8%		
가구 근로소득	30,789	37,909	81.2%		
15세 이상 중 근로소득자	41.8%			42.3%	98.8%
근로소득자 평균소득	15,606			14,635	106.6%

자료: 김승연 외. (2022). 행정자료를 활용한 서울시의 불평등과 빈곤에 관한 연구(2차 연도). pp. 28~29(표 2-6~7), pp. 32~35(표 2-11~15), pp. 43~44(표 3-1~2).

70.8%로 개인소득보다 더 큰 차이가 났다. 이는 건보공단이 주민등록세대를 가구의 단위로 간주하는 점과 아울러 가금복이 1인가구를 과소 대표하기 때문으로 보인다(이승주, 2023). 〈표 2〉가 가리키는 바와 같이 건보공단 자료에서 주민등록세대 기준으로 1인가구의 비중이 39.7%인데, 가금복 자료에서는 1인가구의 비중이 24.9%에 불과하다. 가금복과 같은 가구의 정의를 사용하는 인구주택총조사(이하 인총)의 1인가구 비중 30.2%와 비교해 볼 때 건보공단은 가구의 개념이 달라 1인가구의 비중이 높게 나타나며, 가금복은 조사자의 접근이 어려운 점 등으로 인해 1인가구가 과소 대표된 것으로 보인다. 가금복 자료에서는 1인가구에 속한 인구의 비중은 10.0%로 건보공단 자료의 17.7%는 물론 인총의 12.6%에 비해서도 과소 대표된 것으로 나타났다.

다시 〈표 1〉에서 소득원천별로 건보공단과 가금복 자료를 비교해 보면 근로소득은 건보공단의 1인당 소득이 가금복의 93.1%로 비교적 근접한다. 건보공단의 근로소득이 가금복보다 다소 적은 것은 일용근로소득을 누락하고 있기 때문이다. 그런데 건보공단 자료가 사업소득은 가금복의 57.1%, 금융소득은 가금복의 41.1%밖에 안 된다. 사업소득이 낮게 나타난 것은 사업소득자들이 설문조사에 응답할 때보다 국세청에 소득을 신고할 때 필요경비의 과다 계상으로 소득을 축소 신고하는 경향을 반영할 수 있다. 금융소득의 경우 건보공단은 국세청으로부터 연간 1000만 원 이상 소득자의 자료만 받았기 때문에 소득 파악률이 낮다.

건보공단 소득자료 중 가금복에 비교적 근접하는 소득 포착률을 보이는 근로소득에 대해 경제활동인구조사의 근로소득과 소득분포를 비교해 보았다. 경제활동인구조사는 15세 이상 인구를 대상으로 하므로 건보자료에서 15세 이상 인구의 근로소

[표 2] 건강보험공단, 인구주택총조사, 가계금융복지조사 자료의 2019년 가구원 수별 가구 비중 및 인구 비중

(단위: %)

가구원 수	총가구 중 가구원 수별 비중			총인구 중 가구원 수별 인구 비중		
	건강보험공단	인구총조사	가계금융복지조사	건강보험공단	인구총조사	가계금융복지조사
1	39.7	30.2	24.9	17.7	12.6	10.0
2	22.0	27.8	31.4	19.6	23.3	25.2
3	18.1	20.7	19.4	24.2	26.0	23.3
4	15.3	16.2	19.0	27.2	27.1	30.4
5 이상	4.8	5.0	5.3	11.4	11.0	11.1
전체	100	100	100	100	100	100

자료: 김승연 외. (2022). 행정자료를 활용한 서울시의 불평등과 빈곤에 관한 연구(2차 연도). pp. 24~26(표 2-4~5); 통계청. (2019). 인구총조사; 통계청. (2020). 가계금융복지조사 보도자료. 43쪽.

[표 3] 건강보험공단과 경제활동인구조사 자료의 2019년 근로소득 10분위별 비교

(단위: 천 원, %)

소득분위	국민건강보험공단(A)	경제활동인구조사(B)	A/B
1	3,554	5,871	60.5%
2	9,835	15,111	65.1%
3	16,087	20,748	77.5%
4	21,140	23,682	89.3%
5	25,046	26,359	95.0%
6	30,500	29,729	102.6%
7	37,643	34,417	109.4%
8	47,901	40,245	119.0%
9	64,583	49,266	131.1%
10	116,846	77,211	151.3%

자료: 김승연 외. (2022), 행정자료를 활용한 서울시의 불평등과 빈곤에 관한 연구(2차 연도). p. 45. (표 3-3).

득 자료를 비교해 보면 근로소득자의 비율이 건보는 41.8%, 경활인구조사는 42.3%로 거의 비슷하게 나타난다.

〈표 3〉은 건보공단 자료에서 15세 이상 근로소득자의 10분위별 평균 근로소득과 경제활동인구조사의 임금근로자 10분위별 평균 근로소득을 비교해 보여 준다. 이 표는 두 자료 간 저소득층과 고소득층의 소득 파악에서 매우 큰 차이를 노정한다. 5분위 이하는 경활 자료가 더 높은 소득을 나타냈는데, 그 격차는 저소득 분위일수록 더 커서 1분위의 경우 건보 자료의 근로소득은 경활 자료의 60.5%에 불과하다. 반면 6분위 이상은 건보 자료가 더 높은 소득을 나타냈는데, 그 격차는 고소득 분위일수록 더 커서 10분위의 경우 건보 자료의 근로소득은 경활 자료의 1.5배가 넘게 나타났다. 고소득층일수록 조사자료에 과소 대표되거나 소득의 축소 보고로 행정자료가 보다 정확한 소득 정보를 보여 준다. 저소득층은 일용근로소득과 비공식부문 소득 등 행정자료로 파악되지 않는 소득이 상대적으로 크게 나타남을 시사한다.

저소득층에서 비공식부문의 근로소득 등으로 인해 행정자료의 소득 파악이 조사자료보다 낮게 나타나는 것은 서울시 안심소득 실험 자료를 통해서도 확인된다. 〈표 4〉는 서울시 안심소득 실험 1차 연도(2022년)에 참여한 기준 중위소득 50% 이하의 1523가구에 대해 조회한 공적 자료와 이들에게 직접 설문조사를 통해 응답받은 자료상에 나타난 근로·사업소득과 가처분총소득(재산소득, 기타소득, 사적 및 공적이전소득 포함)을 비교해 보여 준다.

〈표 4〉를 보면 중위소득 50% 이하의 저소득 가

구의 모든 유형에서 행정자료에 나타난 소득보다 설문조사 응답 소득이 더 높게 나타났다. 〈표 3〉과 달리 공적 근로소득 유무 및 기초생활보장 수급 여부를 기준으로 4가지 가구 유형을 분류하여 근로·사업소득과 가처분총소득이 각각 행정자료와 조사자료에 어떻게 나타났는지를 보여 준다. 행정자료에 근로·사업소득이 있는 가구(근로 가구)보다 없는 가구(비근로 가구)에서, 그리고 기초생활보장 수급 가구보다 비수급 가구에서 근로·사업소득 및 가처분총소득의 차이가 더 크게 나타났다. 4가지 가구 유형 중에서는 '근로-수급 가구'(행정자료에 근로·사업소득이 있는 기초생활보장 수급 가구)에서 행정자료와 설문조사 자료 사이에 근로·사업소득 및 가처분총소득의 차이가 가장 작게 나타났으며, '비근로-비수급 가구'(행정자료에 근로·사업소득이 없는 기초생활보장 비수급 가구)에서 이러한 차이가 가장 크게 나타났다.

공식 근로·사업소득이 없는 가구들이 설문조사에서 근로·사업소득이 있다고 응답한 것은 비공식부문에서 근로소득을 벌었거나 사업소득을 0으로 신고한 경우로 생각된다. 수급 가구보다 비수급 가구에서 행정자료와 서베이자료 간에 더 큰 소득 차이가 나타나는 이유는 수급 가구들은 행정자료에 소득 파악이 더 높게 이루어져 있거나 설문조사 응답 시 비공식적인 소득의 노출을 꺼리는 데 비해 비수급 가구들은 행정자료에 소득 파악이 덜 되어 있거나 설문조사 응답 시 상대적으로 더 거리낌없이 비공식적인 소득을 노출하기 때문이 아닌가 추측된다.

행정자료와 조사자료 사이에 소득계층별로, 그리고 저소득층 내에서도 기초생활보장 수급 여부

[표 4] 서울시 안심소득 실험 참여 가구의 공적 자료 및 설문조사 응답상의 근로·사업소득과 가처분총소득 비교

(단위: 월 천 원)

	근로·사업소득			가처분총소득		
	공적 자료	설문 응답	차이	공적 자료	설문 응답	차이
전체 참여 가구	513	739	226	847	1,418	571
수급 가구	485	577	92	1,061	1,263	202
비수급 가구	530	838	308	717	1,513	796
근로 가구	956	1,106	150	1,209	1,643	434
수급	1,013	1,021	8	1,458	1,600	142
비수급	927	1,148	221	1,074	1,664	590
비근로 가구	0	314	314	435	1,159	724
수급	0	168	168	696	954	258
비수급	0	423	423	241	1,311	1,070

자료: 정은하. (2022). 서울시 안심소득 시범사업 추진현황. 한국재정정책학회 추계학술대회 자료집. pp. 36~37.

및 공식 근로소득 유무에 따라 소득의 차이가 상당히 다른 패턴을 보이는 것은 행정자료나 조사자료 중 어느 하나에만 의존할 경우 소득분배의 파악이 편향될 수 있으며, 정책 평가나 정책 설계에서 오류를 범하기 쉽다는 점을 시사한다. 그 한 예로 서울시가 지난 2020년 봄 코로나19 재난지원금을 중위소득 이하의 가구에 지급하는 정책을 실시할 때 예상 신청 및 지급 인원에 비해 실제 신청 및 지급 인원이 약 두 배에 이른 것을 들 수 있다. 당시 서울시는 서베이자료에 기초하여 기준 중위소득 이하 가구의 수를 추정하였는데, 저소득층에 대해서는 행정자료가 서베이데이터에 비해 소득을 낮게 파악하고 있기 때문에 이러한 문제가 발생한 것으로 보인다.

이상의 사례들은 소득분배의 정확한 측정에 서베이데이터와 행정데이터가 각각 가지는 한계, 행정데이터 간의 연계 및 행정데이터와 서베이데이터의 연계가 중요함을 보여 준다. 전국 단위의 소득분배 측정은 서베이데이터의 한계를 행정데이터로 보완한 가계금융복지조사가 비교적 정확한 자료를 제공하고 있으나, 지역 단위의 소득분배 측정에는 행정데이터의 사용이 필수적이다. 다만 국세청이나 건보공단의 소득 데이터는 지역별 분석이 가능한 장점에도 불구하고 이상에서 밝힌 바와 같이 저소득층의 소득 파악 미비 등 한계도 있으므로 타 행정데이터 및 서베이데이터와의 연계가 필요하다.

이에 필자가 책임을 맡고 있는 가천대 불평등과 사회정책연구소는 건보공단의 소득재산 자료를 이용하는 것을 넘어 이 데이터를 다른 행정데이터 및 서베이데이터와 연계하는 프로젝트를 추진 중이다. 건보 자료를 가금복과 연계하는 프로젝트는 가금복에 없는 건강보험 및 장기요양보험 급여와 같은 현물급여 정보를 보완하여 건강보험 급여 등이 소득계층별로, 성별 및 연령별로, 그리고 직장 및 지역가입자별로 어떻게 배분되는지를 파악할 수 있게 할 것이다. 당초 건보의 소득재산 자료까지 가금복 자료와 연계하여 사업소득과 금융소득 등의 큰 격차와 소득계층에 따라 조사자료와 행정자료의 차이에 관한 다른 패턴을 분석하고자 하는 계획도 있었으나, 국세기본법의 과세정보 비밀 유지 및 목적 외 사용 금지 조항에 대한 지나치게 경직된 해석에 따라 건보공단의 소득재산 자료를 데이터 결합을 위해 외부로 반출할 수 없다고 하여 이 계획은 무산되었다.

그러나 건보의 소득·재산 자료를 개인 단위가 아닌 천분위 자료로 변환하면 더 이상 개인정보도 아니고 국세기본법상의 과세정보도 아닌 익명의 통계자료가 되므로 소득·재산의 천분위 자료를 사회보장정보원의 사회보장 수급 자료와 연계하는 프로젝트를 전라북도의 용역으로 추진 중이다. 이를 위해 보건복지부의 8개 과와 주거급여를 담당하는 국토교통부, 한부모가족을 담당하는 여성가족부, 교육급여를 담당하는 교육부로부터 지난 10년간의 데이터 사용 승인을 받았으며, 데이터 결합을 위한 절차를 밟고 있다. 사회보장위원회가 구축한 사회보장통합행정데이터에 비해서는 포괄하는 사회보장

사업의 항목 수는 적다(이현주, 오욱찬, 이윤경, 이원진, 성재민, 이길제, …, 이병재, 2020). 그러나 횡단면의 단년도 데이터가 아닌 10년간의 패널데이터로 연계하여 사회보장의 사각지대 파악과 사회보장 정책들의 효과에 대한 연구를 할 계획이다. 이처럼 행정데이터 간에 연계가 이루어져 패널데이터가 구축되면 이전의 설문조사 자료나 단일 행정자료로는 불가능했던 많은 연구가 가능해질 것으로 기대된다.

4 나가며

설문조사 자료나 행정자료나 개인정보의 보호 필요성은 마찬가지로 중요하다. 그런데 행정자료는 여러 자료의 결합을 요청받는 경우가 많기 때문에 데이터 결합 과정에서 추가적으로 개인정보 보호에 관한 장치가 필요하다. 또한 설문조사와 달리 모집단 전수를 포괄하는 빅데이터인 경우가 많아 개인정보 보호에 보다 세심한 주의가 필요하다. 따라서 개인정보 보호에 대한 안전장치를 이중 삼중으로 설치할 필요가 있다. 과거 스웨덴 등 북유럽 국가들에서도 데이터 활용과 개인정보 보호를 둘러싼 논쟁이 있었다. 개인정보 보호를 철저히 하면서 데이터를 연계하고 가명 처리된 데이터를 연구자들이 안전하게 사용할 수 있는 방법들이 발전되어 왔다. 영국에서 기술적 실수로 인해 개인정보가 다량 유출, 손실되는 사고가 발생한 적이 있긴 하지만(유종성, 전병유, 신광영, 이도훈, 최성수, 2020), 아직까지 연구자들이 비실명화된 마이크로 행정데이터를 사용하는 과정에서 개인정보의 침해가 보고된 사례는 극히 드문 것으로 보인다(Penner and Dodge, 2019).

한국은 북유럽 국가들처럼 모든 개인에게 고유의 주민등록번호가 부여되기 때문에 행정자료의 연계를 쉽게 할 수 있는 조건을 구비하고 있지만, 주민등록번호를 결합키 작성에 사용하지 못하도록 하는 법적 제약과 함께 마이크로 데이터 제공에 대한 정부 및 공공기관들의 소극적인 자세와 관행으로 진전이 더딘 상황이다. 북유럽 국가들은 모든 개인에게 부여되는 주민번호(PIN)를 재식별이 불가능하게 암호화한 연계키를 사용하여 개인 단위 데이터를 연계하며, 주소 코드를 이용하여 가구를 구성하고, 기업아이디번호, 사업체아이디번호 등을 사용하여 행정데이터 간의 연계를 하여 통계용 행정등록부들을 구축한다. 미국은 데이터 결합 시 사회보장번호가 있는 경우에는 이를 우선적으로 사용하고 사회보장번호가 없는 경우에는 이름, 주소 등 다른 정보를 이용하여 확률 매칭을 한다(Wagner & Lane, 2014). 결합키 생성 과정에서는 개인식별정보 외에 다른 개인정보들에는 접근되지 않으므로 주민등록번호를 암호화해 결합키를 생성하는 것이 현재와 같이 성명, 성별, 생년월일을 사용해서 결합키를 생성하는 것보다 개인정보 유출의 위험이 크다고 볼 수 없다. 적어도 공공의 이익을 위한 정책연구를 위한 가명정보 결합에는 결합키 생성에 주민등록번호를 사용하는 것을 금지할 이유가 없다고

본다. 또한 서베이데이터에서는 주민등록번호를 수집하기 어려우므로 성명, 성별, 생년월일 외에 주소와 전화번호 등 다양한 정보의 조합을 결합키로 생성하여 결합률을 높이고 결합의 정확성을 높일 필요가 있다.

다음으로 부처 간의 데이터 칸막이를 해소하여 마이크로 행정데이터의 연계를 적극적으로 추진하는 정부의 종합적 노력이 필요하다. 북유럽 국가들과 미국, 영국은 물론 캐나다, 호주, 뉴질랜드 등이 행정자료 구축과 함께 승인된 연구자들에게 행정자료 간 연계, 행정자료와 서베이데이터 연계 서비스를 적극적으로 제공한다(유종성, 2022). 유럽연합(EU) 차원에서도 각국에 행정데이터의 적극적 활용을 권장하고 있다. 최근 미국은 기존의 공공데이터 포털인 Data.gov 외에 증거기반정책기본법에 따라 행정데이터를 비롯하여 연구용 마이크로 데이터의 통합 검색 및 신청 포털인 Research Data Gov.org를 개설하였다.

이제 데이터 3법 통과 후 지난 3~4년간 사회과학과 정책 연구에 가명정보 결합을 통한 행정데이터의 연계 활용이 얼마나 잘 이루어지고 있는지, 법적인 걸림돌과 행정적, 관행적인 문제는 무엇인지 점검하고 개선책을 강구해야 할 때이다. 필자의 경험으로는 행정데이터의 연계를 위해 데이터 보유기관으로부터 사용 승인을 받는 것부터 시작해 데이터 결합 신청 및 결합 과정에 너무 많은 어려움이 있고 시간이 오래 걸린다. 또한 국세청의 행정데이터를 이용하거나 국세청 자료에 기초한 건보공단의 소득자료를 다른 데이터와 연계하는 일이 국세기본법의 비밀 유지 및 목적 외 사용 금지 조항의 경직된 해석 때문에 지나치게 제한되기도 하였다. 다행히도 최근 디지털플랫폼정부위원회가 부처 간 데이터 칸막이를 해소하기 위해 국세기본법 등 개별 법령의 비밀 유지 및 목적 외 사용 금지 조항에 대한 정비 및 특별법 제정을 추진한다고 한다. 디지털 플랫폼 정부를 표방하는 현 정부 아래에서 증거 기반 정책을 위한 행정데이터의 연계 활용에 획기적 진전이 이루어지기를 기대한다.

우리도 북유럽 국가들처럼 통계청이 중심이 되어 행정데이터의 표준화, 여러 행정데이터를 연계하여 다양한 통계와 연구 목적의 행정데이터를 구축할 필요가 있다. 최근 통계청이 성인 전수에 대해 공사연금의 가입 및 수급 실태를 경제활동 등의 행정자료와 결합한 연금통계 DB를 구축한 것은 좋은 사례이다. 나아가 통계청의 각종 통계등록부를 비롯한 통계데이터센터(SDC: Statistics Data Center) 보유 자료는 물론 마이크로 데이터 통합 서비스(MDIS: Microdata Integrated Service)를 통해 제공하는 서베이데이터에도 개인별로 재식별이 불가능한 고유통계번호를 부여하여 보유 데이터들 간의 연계를 쉽게 할 필요가 있다. 향후 각종 설문조사는 인구가구등록부 또는 센서스를 기반으로 표본을 추출하여 조사 항목 중 상당 부분을 행정자료로 대체하거나 보완할 수 있도록 하여 조사자의 응답 부담과 조사 비용을 줄이면서 조사의 정확도를 높이는 방안을 모색해야 한다.

끝으로 사회보장위원회의 사회보장통합행정데이터 구축은 반가운 일이지만 횡단면의 표본 자료로서 가명정보 결합으로 인한 표본 상실의 문제를 안고 있다(이현주 외, 2020). 사회보장기본법의 재개정을 통해 통계청과의 협력으로 전수 패널데이터를 구축하도록 하고, 연구용 제공 시에는 연구 목적에 필요한 제한된 수준의 샘플과 제한된 변수 데이터만을 제공하되 타 행정데이터나 서베이데이터와의 연계를 통한 연구를 적극 지원해야 한다. 또한 미국처럼 모든 행정데이터와 연구용 마이크로 데이터의 통합 검색 및 신청 포털 개설을 서두를 필요가 있다.

참고문헌

국회미래연구원. (2019), 정부 중장기계획 메타평가 실시 방안 연구. **국회미래연구원 연구보고서 19-15.**

김승연, 유종성, 최광은, 이혜림, 전병유, 정준호, 김민혜, 이승주, 유수진, 박종헌, 배세진. (2020). **행정자료를 활용한 서울의 불평등과 빈곤연구(1차연도): 건강보험데이터를 활용한 서울시의 소득 및 자산 분배.** 서울연 2020-CR-15. http://www.kirl.re.kr/bbs/board.php?tbl=ga_bbs32 에서 2023. 9. 20. 인출.

김승연, 유종성, 박종헌, 전병유, 정준호, 김동진, 서준영, 최혜은, 이혜림, 배세진. (2022). **행정자료를 활용한 서울의 불평등과 빈곤에 관한 연구(2차연도).** 서울연 2021-CR-04. http://www.kirl.re.kr/bbs/board.php?tbl=ga_bbs32 에서 2023. 9. 20. 인출.

유종성. (2020). 행정빅데이터 활용 성과관리체계 운영 방안 검토. 이채진, 정혜진, 김지원, 유종성(편). 정부 중장기계획 평가방안 연구. **국회미래연구원,** 95-136.

유종성. (2022). **증거기반 정책을 위한 행정데이터 활용의 해외동향과 시사점.** 사회보장위원회 제출, 미출판 논문.

유종성, 전병유, 신광영, 이도훈, 최성수. (2020). 증거기반 정책연구를 위한 행정자료의 활용. **한국사회정책.** 27(1). 5-37

유종성, 김민혜, 김승연, 유수진. (2021). 소득분배 연구를 위한 건보공단 빅데이터의 의의와 한계: 서울시 사례연구를 중심으로. **한국사회정책,** 28(3), 75-105.

유종성, 박종헌, 구정은, 최혜은, 배세진. (2022). **성남시 불평등과 빈곤에 관한 연구.** 성남: 가천대 불평등과사회정책연구소.

이승주. (2023). 비동거 가구원 보정이 소득불평등지표 변화에 미치는 영향에 관한 연구: 가계금융·복지조사 자료를 중심으로. **한국사회복지조사연구, 78,** pp. 33-58.

이현주, 오욱찬, 이윤경, 이원진, 성재민, 이길제, …, 이병재. (2020). **사회보장정책 효과성 분석을 위한 행정데이터 연계·활용 방안**. 세종: 한국보건사회연구원.

정은하. (2022). 서울시 안심소득 시범사업 추진현황. **한국재정정책학회 추계학술대회 자료집.** 대전: 한국재정정책학회.

조슬기나. (2020). "ICT는 강국, 데이터는 후진국." **아시아경제.** 2020. 6. 18.

통계청(2019). 2019년 인구주택총조사 결과-등록센서스 방식, 보도자료.

통계청(2020). 2020년 가계금융복지조사 결과, 보도

자료.

Card, David, Raj Chetty, Martin S. Feldstein, and Emmanuel Saez. (2010). "Expanding Access to Administrative Data for Research in the United States." SSRN Scholarly Paper ID 1888586, Social Science Research Network, Rochester, NY.

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. (2016). "The role of administrative data in the big data revolution in social science research." *Social Science Research* 59:1–12.

Kristensen, Jens Kromann, Walter S. Groszyk and Bernd Bühler. (2002). "Outcome-focused Management and Budgeting", *OECD Journal on Budgeting,* 1(4), pp. 7–34.

Penner, A. M. & Dodge, K. A. (2019). Using Administrative Data for Social Science and Policy. *RSF: The Russell Sage Foundation Journal of the Social Sciences. 5*(3). 1–18.

Song, Xi and Thomas S. Coleman. forthcoming. (2020). "Using Administrative Big Data to Solve Problems in Social Science and Policy Research." *Global Social Security Review* 14: 5–15.

United Nations. (2007). *Register-Based Statistics in the Nordic Countries—Review of Best Practices with Focus on Population and Social Statistics.* Methodological Guidelines. UN. http://digitallibrary.un.org/record/609979. 2023. 9. 25. 인출.

Wagner, D., & Lane, M. (2014). *The person identification validation system (PVS): applying the Center for Administrative Records Research and Applications'(CARRA) record linkage software.* CARRA Working Papers 2014-01. Center for Economic Studies, US Census Bureau.

Wallgren, Anders and Britt Wallgren. (2022). *Register-based Statistics: Registers and the National Statistical System.* Third Edition. John Wiley & Sons.

Wan, W.-Y., Williams, L., Lee, E., & Lu, L. (2023). *Longitudinal literacy and numeracy in Australia (LLANIA) dataset: Technical report.* Australian Education Research Organisation. https://www.edresearch.edu.au/resources/longitudinal-literacy-and-numeracy-australia-llania-dataset-technical-report 2023. 9. 25. 인출.

Cases of Using Social Security Administrative Data and Future Tasks

You, Jong-sung

(Gachon University)

Administrative data is a type of big data that typically covers the entire population subject to administration, allowing for research at regional and small-group levels. It has the advantage of being free from non-response, memory errors, and false responses and of naturally obtaining panel data structures. However, there are cases where administrative data is less accurate than survey data, and the use of data from particular administrative agencies is very limited, so it can be of considerable utility only when linked with other administrative data or survey data.

For evidence-based policies, especially for policy evaluation and design at the regional level, such as regional social security plans, it is crucial to link various administrative data and survey data. Based on the experience of the past few years since the enactment of the three data laws, legal and administrative improvement measures should be sought.