한국사회정책
Korea Social Policy Review, 제27권 제1호, 2020, pp.5~37.

증거기반 정책연구를 위한 행정자료의 활용*

유 종 성** · 전 병 유*** · 신 광 영**** · 이 도 훈***** · 최 성 수******

요 약

증거 기반 정책(evidence-based policy)을 위한 행정자료(administrative data), 특히 행정 빅데이터의 연계, 통합과 활용은 북유럽 국가들로부터 시작하여 최근 영국, 미국 등에서도 급속도로 확산되고 있다. 한국은 인터넷 강국으로서 전자정부와 공공데이터 개방 등의 국제적 평가에서 최상위권에 위치하지만, 행정 빅데이터를 활용한 사회과학 및 정책 연구는 미흡한 상황이다. 개인과 가구에 대한 소득-자산 빅데이터도 제대로 구축되지 않았고, 행정 자료 간의 연계 통합이나 행정자료와 조사자료(survey data) 간의 연계는 거의 이루어지지 않고 있다. 본고는 증거 기반 정책을 위한 연구에 행정자료의 구축과 활용이 왜 필요한지, 그리고 이와 관련하여 개인정보 보호 문제가 북유럽과 영국, 미국 등지에서는 어떻게 다루어지는지 살펴보고, 한국에의 함의를 검토한다. 우리는 통계청이 국세청, 건강보험공단 등과 협력하여 등록센서스의 패널 구축과 전국민 등록 기반의 고품질 소득-자산 DB 구축 등을 시급히 추진할 것과 함께 학계의 적극적인 이니셔티브 하에 행정자료를 연계, 활용하는 연구 프로젝트들을 한국연구재단과 경제인문사회연구회 등이 적극 지원, 협력할 것을 제안한다.

주요어: 증거기반 정책, 행정자료, 빅데이터, 데이터 연계, 개인정보 보호

* 이 논문은 2018년도 대한민국 교육부와 한국학중앙연구원(한국학진흥사업단)을 통해 한국학 세계화 랩 사업의 지원을 받아 수행된 연구임 (AKS-2018-LAB-1250002)
** 제1저자, 가천대학교 사회정책대학원 교수(youjs0721@gachon.ac.kr)
*** 교신저자, 한신대학교 사회혁신경영대학원 부교수(bycheon@hs.ac.kr)
**** 중앙대학교 사회학과 CAU-Fellow(kyshin20@gmail.com)
***** 연세대학교 사회학과 부교수(dlee2191@yonsei.ac.kr)
****** 연세대학교 사회학과 조교수(s.choi@yonsei.ac.kr)

1. 서론

2017년 가을 데이비드 그러스키(David Grusky) 스탠포드대 교수가 불평등연구회의 초청으로 방한, 연세대학교에서 2회의 연속강연을 했다. 첫날 그는 미국의 "절대적 소득이동"(absolute income mobility)이 감퇴하고 있다는 연구 결과를 발표했다(Chetty et al., 2017). 1940년대 출생 세대가 30세에 부모의 소득을 넘어설 확률이 90% 가량이었는데, 1980년대 출생 세대가 그럴 확률은 50%로 줄어들었다는 것이다. 또한, 경제성장 둔화와 소득분배 악화의 두 요인이 각각 절대적 소득이동의 감퇴에 미친 영향을 분석한 결과 소득분배 악화가 훨씬 더 큰 역할을 했다는 것이다. 이 연구에 사용한 자료는 미국의 인구총조사(Census) 및 인구현황조사(Current Population Survey) 자료와 함께 납세자들이 국세청에 제출한 소득세 신고(tax return) 자료를 연결, 통합한 자료였다. 미국에서는 소득이 있는 개인들이 대부분(아주 작은 규모의 소득자 등은 제외하고) 종합소득세 신고를 하는데, 자녀가 있는 경우 세액공제를 받기 위해 자녀들의 이름과 사회보장번호를 기록한다. 따라서 위의 통합 자료를 가지고 1940년생들과 1980년생들의 30세 때 소득과 그들 부모의 30세 때 소득을 매치하여 1천만이 넘는 부모-자녀 간의 소득 결합 분포를 추정할 수 있었다. 둘째 날 그러스키는 증거기반 정책(evidence-based policy)의 중요성을 강조하면서, 이를 위해 행정 빅데이터의 활용이 긴요함을 역설하였다(유종성, 2019a).

이념 대립으로 흐르기 쉬운 의견기반 정책(opinion-based policy)보다 증거기반 정책의 필요성이 강조되면서 정책효과에 대한 과학적 분석을 위한 데이터가 더 중요해지고 있다. 민주주의는 다수의 의견을 바탕으로 정책을 형성하는 것이라는 점에서 의견기반 정책이 나쁘다고 할 수는 없다. 그러나, 다수의 의견형성 과정이 가급적 정확한 정보와 과학적 분석에 입각한 토론(informed discussion, or deliberation)을 통해 이루어지냐 여부는 민주적 정책결정의 질을 높이는 데 있어서 중요한 문제이다(Commission on Evidence-Based Policymaking, 2017).

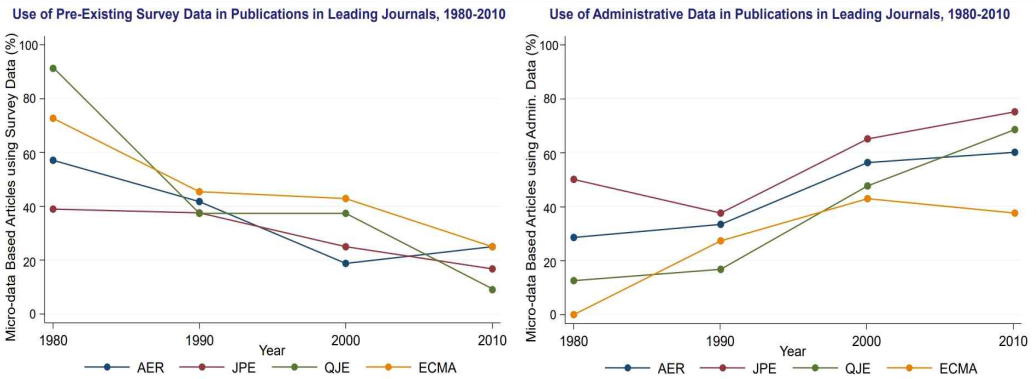
전통적으로 사회과학과 정책연구는 서베이 데이터를 사용했지만, 최근에는 북유럽의 선도하에 영미 등에서도 전 인구 또는 모집단 전체를 포괄하는 행정자료 또는 행정 빅데이터 구축과 행정자료 간의 결합을 통해 증거기반 정책연구의 새로운 지평을 열어가고 있다. 행정 빅데이터는 그 자체가 패널자료이다. 이러한 데이터를 활용해서 과거에는 꿈꾸기 어려웠던 장기간에 걸친 정책효과나 특정 소수집단에 대한 하위그룹 분석이 가능하게 되었다. 또한, 각종 서베이의 응답률이 저하하고 비용이 증가함에 따라 행정자료로 대체하거나 조사자료와 행정자료와의 연계를 통해 조사의 비용(연구자 측)과 부담(응답자 측)을 줄이면서 동시에 자료의 정확도를 높일 수 있다(한은희, 2019; Penner and Dodge, 2019).

본고에서 행정자료라 함은 행정과정에서 구축되는 자료를 말하는데, 행정자료는 많은 경우 해당 행정과 관련된 모집단 전체를 포괄한다. 또 상당수의 행정자료는 국가나 지방정부의 전체 인구, 전체 가구, 또는 전체 사업장 등을 포괄하여 양적으로 볼 때 빅데이터라고 볼 수 있다(Connelly et al., 2016). 다만, 양적으로 어디까지가 빅데이터인지에 대해서 뚜렷하게 구분하기는 어려우므로 행정자료와 행정빅데이터란 용어를 정확하게 구분하지 않고 사용함을 밝혀둔다.

행정자료를 통합, 연계하여 사회과학과 정책연구에 활용하는 다른 국가들의 사례를 보면 크게 북유럽의 하향(top-down)방식과 영미의 상향(bottom-up) 방식으로 구분된다. 북유럽 국가들의 경우 전국민 등록 기반(national register-based) 통합 데이터베이스를 구축, 이를 바탕으로 경제, 사회, 의료 등의 국민 정보를 상시적으로 업데이트하고 이로써 전통적으로 실시되던 설문조사 기반의 센서스를 대체하고 있다. 이렇게 구축·유지되는 데이터는 일정한 기준을 충족하는 연구자, 기업, 일반 시민에게 폭넓게 연구, 분석을 위해 제공된다. 미국, 영국 등의 경우에는 각각 고유의 정책적 목적과 법적 근거를 가지고 다양한 국가 기관에 의해서 수집되는 행정자료를 자격을 갖춘 연구자들이 적법한 절차와 기준에 맞게 연구 목적에 따라 요구할 때마다 국가가 운영 혹은 위임하는 기관이 이런 자료를 활용할 수 있게 한다(신광영 외, 2018).

사실 행정 빅데이터의 구축과 활용에 있어 미국은 북유럽 국가들에 비하면 후발 국가이다. 덴마크, 스웨덴 등 북유럽 국가들에서는 일찍이 전국민 등록기반 행정자료를 이용한 사회과학 연구가 오랜 전통을 가지고 있다. 행정 데이터 간의 통합 또는 행정 데이터와 서베이 데이터를 연계, 통합하여 사회분석에 활용하는 것이 이제 북유럽 국가들뿐만 아니라 유럽연합에서 통계 시스템을 혁신하는 기본 목표가 되었다(Santourian and Petrakos, 2018; Santourian et al., 2018). 행정자료를 중심으로 전통적인 서베이 자료까지 연계하여 경제, 사회정책 평가를 위한 과학적 통계 분석에 활용하는 것은 북유럽 국가들은 물론이고 1990년대 이후 유럽연합에서 통계 시스템을 혁신하는 기본 목표가 되었다. 이는 사회과학 연구에서 전통적인 서베이 자료 중심의 분석으로부터 전체 모집단이나 모집단 일부를 전부 분석하는 방향으로의 전환이 일어나고 있음을 의미한다 한다(Connelly et al., 2016). 최근에는 영국과 미국 등도 증거 기반 정책을 강조하면서 행정 빅데이터의 사회과학 및 정책연구 활용에 많은 노력을 기울이며 상당한 성과를 이룩하고 있다(이건·윤광석, 2016; Penner and Dodge, 2019). 아래 그림은 최근 경제학 톱 저널에 출판되는 논문들에서 전통적인 조사자료를 사용한 논문의 비중이 급격하게 줄어들고 행정자료를 사용한 논문의 비중이 급증해오고 있음을 보여주고 있다.

[그림 1] 경제학 톱 저널에 조사자료와 행정자료를 사용한 논문의 비중 추이, 1980-2010.



주: AER: American Economic Review, JPE: Journal of Political Economy, QJE: Quarterly Journal of Economics, ECMA: Econometrica.
자료: Friedman(2015).

한국은 김대중 정부 이래로 인터넷과 전자정부 분야에 적극적으로 투자하여 인터넷 접속 속도에서 세계 수위를 기록하였으며(Speiser, 2015), 국제연합의 전자정부 평가에서도 항상 최상위권을 유지, 2018년에는 세계 3위를 기록하였다(United Nations, 2018). 공공데이터 개방과 관련해서는 OECD의 열린 정부 데이터 지수(Open, Useful, Reusable Government Data Index; OURdata Index)에서 2015년 이래 계속해서 1위를 차지하고 있다(https://www. oecd.org/gov/digital-government/open-government-data.htm). 이는 한국 정부가 2013년 "공공데이터의 제공 및 이용 활성화에 관한 법률"을 제정하고 공공데이터 포털(https://www.data.go.kr)을 개설하여 일반인에게 공공데이터의 개방을 적극 추진한 결과이다.

최근 한국에서는 제4차 산업혁명과 관련하여 빅데이터와 인공지능에 대한 관심이 높아지고 있고, 인공지능은 빅데이터의 활용 없이는 발전할 수 없음이 인식되고 있다. 이에 따라 정부는 데이터가 제4차 산업혁명의 원유와 같다며 "데이터를 가장 안전하게 잘 쓰는 나라"를 만들겠다고 선언하였다. 이미 공공 빅데이터를 활용한 각종 앱의 설계와 상업적 활용이 이루어져왔고, 지방자치단체에서 스마트 행정을 구현하는 도구로도 사용되고 있다. 정부는 금융, 보건의료 등 데이터의 산업적 활용 요구가 높은 분야에서부터 빅데이터 플랫폼을 구축하여 이를 일반에 공개하고 있으며, 데이터 경제의 발전을 위해 소위 데이터 3법을 입법하였다. 그럼에도 사회과학과 정책연구에 있어서 공공데이터 간의 연계, 통합이나 서베이 데이터와의 연계 활용은 아직 잘 이루어지지 않고 있으며, 북유럽은 물론 영미 등에 비해서도 상당히 뒤지고 있다.

이에 본고에서는 먼저 행정데이터의 구축과 활용에서 앞서가고 있는 덴마크, 스웨덴 등 북유럽국가들, 그리고 최근 적극적으로 행정데이터 구축과 활용을 모색하고 있는 영국과 미국에

서 사회과학과 정책연구에 있어서 행정 빅데이터의 활용이 어떻게 이루어져 왔는지를 살펴보고자 한다. 이어서, 한국에서 공공 데이터간의 통합이나 서베이 데이터와의 연계를 통한 연구가 잘 진척되지 않은 요인들을 고찰, 이를 바탕으로 몇 가지 과제를 제시하고자 한다. 최근 데이터3법(개인정보보호법, 정보통신망법, 신용정보법)의 통과로 행정데이터의 구축과 활용의 가능성이 크게 높아졌다. 이러한 점에서 행정데이터와 관련된 쟁점들과 이에 대한 해결책을 제시하고 있는 해외 사례를 살펴봄으로써 향후 이루어질 한국의 행정데이터 구축과 관련된 함의를 찾고자 한다.

2. 행정데이터에 관한 국내외 논의

행정데이터에 대한 국내외 관심은 상대적으로 최근에 시작되었다. 행정데이터는 행정기관이 서비스를 포함한 행정 목적으로 개인이나 가구에 대한 정보를 수집하여 관리하는 데이터를 의미한다(Jones and Elias, 2006). 그러므로 각기 다른 행정기관이 각기 다른 행정 목적에 따라서 각기 다른 개인과 가족에 관한 정보를 축척하였다. 오늘날 행정데이터는 이러한 흩어져있는 데이터를 통합하여 관리하고 사회과학적 목적으로 활용하는 데이터로 빅데이터의 한 종류를 지칭한다.

행정 데이터는 학술적 목적으로 수집된 자료가 아니라는 점에서 서베이 자료와 근본적인 차이를 보인다. 행정데이터와 기존의 데이터의 핵심적인 차이는 기존의 서베이 데이터는 만들어진 데이터이고, 행정데이터는 인위적인 방법으로 수집된 것이 아니라 있는 현실을 데이터로 옮긴 것이다(Connelly et al., 2016: 3-4). 서베이 데이터는 연구자가 조사를 설계하고 자료를 수집하지만, 행정자료는 연구자가 개입하지 않고, 행정 목적으로만 수집한 자료이다. 대체로 행정데이터는 모집단에서 표본을 추출하여 자료를 수집하는 것이 아니라 전체 모집단을 대상으로 자료를 수집한다는 점에서도 조사자료와 차이를 보인다.

행정데이터 분석을 둘러싼 논의들도 다양하게 제시되었다. 행정데이터의 특징은 데이터 규모가 크지만 변수는 대단히 제한되어 있는 점이다(large N, small k). 그러므로 대규모 데이터 분석을 위한 통계 프로그램(Stata MP나 R)과 기존 추론 통계와 행정데이터 통계분석의 차이점에 이르기까지 다양한 이슈가 제기되었다. 무엇보다도 행정데이터의 가장 큰 장점은 증거기반 사회정책을 가능하게 한다는 점이다. 일시적이거나 국지적인 사례 혹은 부분적인 조사자료에 기초한 정책이 아니라 국민 전체의 교육, 노동시장, 건강과 복지에 관한 장기간에 걸친 전체의 분석에 기초한 정책 형성을 가능케 한다. 그 결과 미국 의회는 2016년 머레이-라이언 증거기반

정책 위원회 법(Murray-Ryan Evidence-Based Policy Commission Act)과 2018년 증거기반정책기본법(Foundations for Evidence-Based Policymaking Act)을 통과시켜, 행정데이터 분석에 기초한 정책 입안과 정책평가를 제도적으로 지원하였다(Penner and Dodge, 2019: 2). 증거기반정책 위원회의 최종보고서는 행정자료들의 연계를 위한 전국안전데이터서비스(National Secure Data Service)의 설립을 비롯한 22가지 권고사항을 담았으며, 이에 따라 미연방정부는 여러 부처가 참여하는 연방 데이터전략(Federal Data Strategy; https://strategy.data.gov/)을 설립하고 2020 액션 플랜을 수립하였다.

제3세계의 빈곤퇴치 프로그램에 대한 무작위통제실험 등을 통해 증거기반 정책의 새로운 지평을 연 공로로 2019년도 노벨 경제학상을 수상한 바너지와 듀플로 부부가 주도하는 MIT의 빈곤행동랩(J-PAL)은 서베이와 실험데이터에 행정자료를 연계하여 정책연구를 발전시키고 있다(https://www.povertyactionlab.org). J-PAL은 "행동을 위한 데이터와 실험의 혁신"(Innovations in Data and Experiments for Action; IDEA) 이니셔티브라는 프로젝트를 운영하면서 2020년에 연구와 증거기반정책을 위한 행정자료 활용 핸드북(IDEA Handbook)을 펴낼 예정이다(Cole et al., forthcoming). 앞서 언급한 미국의 "절대적 소득이동"에 대한 연구를 주도한 라지 체티 하버드대학교 교수가 주도하는 기회의 통찰(Opportunity Insights) 프로젝트는 "경제적 사회적 문제들을 해결하기 위한 빅데이터의 활용"이란 과목을 개설하였는데, 강의자료와 동영상을 웹사이트에 공개하고 있다(https://opportunityinsights.org/course/).

증거기반 정책연구를 위한 행정자료, 특히 행정 빅데이터의 활용이 증가하면서 관련된 이슈들에 대한 논의와 연구가 급증하고 있다. 빅데이터 혁명의 일환으로서의 행정빅데이터의 역할 및 기술적 이슈들(Connelly et al., 2016), 증거기반 정책 연구를 촉진하기 위한 행정자료의 구축과 활용(Commission for Evidence-Based Policymaking, 2017; Reamer and Lane, 2018), 행정자료의 연계, 통합과 개인정보 보호의 문제(Harron et al., 2017; National Academies of Science, Engineering and Medicine, 2017), 사회과학과 정책연구에 있어서 행정자료의 활용 활성화 및 개선책(Jones et a., 2019; Penner and Dodge, 2019; Santourian et al., 2018) 등 수많은 논문이 쏟아져나오고 있다. 각국의 상황에 따라 다르게 이루어진 행정자료 활용에 관련된 논의와 연구는 이하 각국에 관한 절에서 더 자세히 소개하고자 한다.

행정데이터에 관한 국내 논의도 최근에 이르러서 시작했다. 특히, 2010년대 정부3.0 논의와 더불어 행정데이터 구축과 활용에 관한 논의들이 등장하기 시작했다(오미애, 2013, 2014, 2019). 오미애(2013)은 다양한 행정데이터를 통합하기 위한 데이터 매칭 방법의 유용성을 소개하고, 호주의 사례를 통해서 행정데이터 구축을 통한 보건복지 서비스에서 시간과 비용의 절감 사례를 소개하고 있다. 오미애(2014)는 한국의 보건복지 분야에서 이루어진 행정데이터를 활

용한 중앙 부처의 사례와 지방자치단체 사례를 소개하고, 행정데이터 활용의 이점과 여기에서 야기되는 이슈와 쟁점들을 소개하였다. 행정데이터를 이용한 거시적 분석과 개인정보보호에 관한 사회적 합의를 바탕으로 행정데이터를 통한 많은 행정비용의 절감과 국민 서비스의 질 개선을 강조하였다.

한국에서 이루어지고 있는 행정데이터 논의는 주로 해외 사례 소개형태로 이루어지고 있다. 해외 사례에 관한 연구들은 스웨덴 사례(신광영, 2017), 스위스, 영국과 스웨덴(오미애최현수, 2015), 북유럽과 영국(신광영최성수김영미, 2018), 미국(김창환이도훈, 2018)을 들 수 있다.

해외 사례 연구는 행정데이터도 국가 간 큰 차이를 보이고 있다는 점을 보여주고, 복지국가가 발달한 북유럽에서 행정데이터 구축과 활용이 일찍부터 체계적으로 활용되기 시작하였고, 영미권에서는 상대적으로 뒤늦게 이루어졌다는 점을 보여준다. 특히 북유럽의 경우, 건강과 복지서비스와 관련하여 행정데이터가 발전하였지만, 영미의 경우는 횡단자료 분석이 지배했기 때문에, 상대적으로 패널데이터의 형식을 지니는 행정데이터는 상대적으로 뒤늦게 관심을 받았다는 점을 밝히고 있다.

또한 법적, 제도적 조건이 미비된 상태에서 행정데이터에 관한 논의들이 이루어지면서, 개인정보보호와 법적 쟁점들이 주로 다루어졌다(전주열, 2016; 채향석, 2017; 최승필, 2017). 법률적인 논의들은 전자정부법 25조 공동정보의 이용에 관한 법률과 개인정보보호법 간의 충돌을 다루고 있다. 행정데이터는 개인정보를 수집한 것이기 때문에, 공공성을 지니고 있기는 하지만, 공적으로 공유되고 있는 정보는 아니라는 것이다. 개인정보의 자기결정권을 전제로 하여, 행정정보의 공동이용을 가능케 하는 개인의 동의, 개인정보 유출 및 누설에 관한 벌칙 강화 등 행정데이터 구축과 활용과 관련된 제도 마련과 거버넌스의 필요성을 제기하고 있다.

3. 북유럽 국가들의 행정 데이터 활용

덴마크, 스웨덴, 핀란드, 노르웨이 등 북유럽 복지국가들은 인구/가구, 소득, 교육, 주거, 고용, 기업, 의료, 복지 등의 행정자료를 통합한 전국민 등록 행정자료를 구축하고, 이를 통계적으로 활용할 수 있도록 한다는 공통점을 가지고 있다. 이는 복지국가가 발전하기 위해서는 양질의 데이터를 사용하여 증거 기반 정책에 활용하는 인프라를 구축하는 것이 중요함을 시사한다.

북유럽의 전국민 등록 기반 행정자료 통합 및 통계적 활용 시스템에 대한 비전은 1960년대 통계학자 및 통계청 인사들의 협력과 토론의 과정에서 형성되었다. 당시 일련의 북유럽 통계학 컨퍼런스에서 새롭게 부상하고 있던 전자 데이터 처리 기술이 가져온 기회를 어떻게 활용할 것

인가에 대한 논의가 그 시발점이 되었다(Thygesen, 2010). 특히 초점은 등록 행정자료를 통합, 활용함으로써 방대하게 소요되는 서베이 데이터 수집과 활용 비용을 혁신적으로 줄일 수 있다는 것이었다. 이는 궁극적으로 서베이 기반 센서스(인구주택총조사) 데이터를 인구, 주거, 교육, 고용, 보건 등의 등록 자료들을 통합한 데이터베이스로 완전히 대체하는 비전으로 발전하였다.

아래 [표 1]은 북유럽 4개국이 다양한 등록 행정자료들을 구축, 통합하여 기존의 서베이 기반 센서스를 대체하는 과정을 보여준다. 이들은 1960년대 중후반에 중앙 인구 등록(Central Population Register; CPR)을 확립, 모든 개인에게 주민 식별번호를 부여해서 이후 다른 등록 자료들을 통합할 수 있는 기반을 마련하였다. 개인 단위, 주거지 단위, 사업체 단위 등록정보들이 개인 식별 주민번호 및 개인들이 속해 있는 공통의 주거지, 사업체 식별번호를 통해 연계되어 통합되었다. 이후 다양한 등록 행정자료들이 생성되고 센서스에 통합된다. 행정자료들에 담긴 모든 개인의 정보들은 지속적으로 수집, 축적되어 종단적 데이터(longitudinal data)가 구축된다(신광영 외, 2018).

[표 1] 북유럽 국가들에서 등록 자료의 생성과 센서스에의 통합년도

	덴마크		핀란드		노르웨이		스웨덴	
	생성	센서스 통합	생성	센서스 통합	생성	센서스 통합	생성	센서스 통합
중앙 인구 등록	1968	1981	1969	1970	1964	1970	1967	1975
사업체 등록	1975	1981	1975	1980	1965	1980	1963	1975
거주	1977	1981	1980	1985	2001	2011	2008	2011
주거환경	1977	1981	1980	1985	2001	2011	2008	2011
교육	1971	1981	1970	1975	1970	1980	1985	1990
고용	1979	1981	1987	1990	1978	2001	1985	1985
가족	1968	1981	1978	1980	1964	1980	1960	1975
가구	1968	1981	1970	1975	2001	2011	2011	2011
소득	1970	1981	1969	1970	1967	1980	1968	1975
총 등록기반 센서스		1981		1990		2011		2011

자료: UNECE(2007). 신광영 외(2018)에서 재인용.

위의 〈표 1〉에서 본 바와 같이 덴마크는 전국민 등록 행정자료를 통합해서 센서스를 완전히 대체한 최초의 국가이다. 1966년 전국민 등록기반 시스템으로 전환하기 위한 통계청법이 제정되어 통계청(Statistics Denmark)에 모든 공공기관이 생성, 보유하는 등록자료에 완전히 접근, 조정, 관리할 수 있는 권한이 주어졌다. 1968년 중앙 인구등록자료가 구축되었는데, 여기에는 개인들의 이름, 주소, 시민권 지위, 가족관계, 개인식별번호(Personal Identification Number)

가 포함되었다. 이후 사업체 및 고용지위 정보(1971년), 연금, 납세 및 임금 정보(1974년) 등 다양한 분야의 등록 자료가 생성, 중앙인구등록 시스템에 통합되었다. 1981년에는 모든 등록행정자료가 통합되어 총 등록기반 센서스가 기존 가구조사 기반 센서스를 완전히 대체하게 되었으며, 이후 덴마크는 더 이상 센서스 조사를 실시하지 않는다(신광영 외, 2018).

스웨덴의 전국민 등록 기반 행정자료 통합, 활용 체제가 구축되는 과정도 1960년대 중반부터 본격적으로 시작되었다. 중앙인구등록 시스템이 1967-68년도에 구축되었고, 모든 종류의 개인정보에 공통으로 사용되는 개인식별번호를 부여하였다. 이후 소득(1968), 교육(1985), 고용(1985) 등의 여러 개인 등록행정자료가 구축되었다. 사업체 등록행정자료(business register)는 이미 1963년부터 구축되어 있었는데, 1975년에는 종합 센서스 시스템에 통합되었다. 그러나 주거지 단위의 등록 시스템, 즉 주택 및 건물, 주거환경, 그리고 가구 등록 자료가 총인구등록에 통합되는 과정이 지연되어 2011년에 이르러서야 통합이 완료되었다. 이로써 기존의 우편조사에 기반한 센서스 조사가 전국민의 등록된 행정자료에 기반한 조사로 대체되었다(신광영, 2017).

행정 데이터 통합은 통계청이 각 행정 부처에서 수집한 행정 등록(administrative register) 데이터를 통합하여 통계등록(statistical register) 데이터로 전환하는 과정이다(Wallgren and Wallgren, 2007: 4-9). 가령 통계청은 1990년부터 통합된 개인 수준의 등록행정자료를 활용해 전 국민 개인들의 노동시장과 교육 및 복지 정보를 통계적으로 분석할 수 있는 "교육, 소득과 직업에 관한 시계열 데이터"(LOUISE)를 구축했다. LOUISE는 1990년부터 매년 12월 31일 기준 16세 이상(2010년부터는 15세 이상) 모든 스웨덴 국민의 고용상태, 소득, 직업, 경제활동, 질병, 실업과 퇴직, 사회부조 및 연금(사적, 공적), 출생국, 부모 출생지, 이민, 거주지, 고용지역, 학력 등의 개인정보와 가족원 및 직장의 정보(주소, 산업, 성별/교육수준별 피고용자 수, 연봉정보 등)를 시계열적으로 추적해 제공한다. 2004년 이후에는 노동시장 장기통합 데이터(LISA)로 명칭이 변경되었다.

이는 기존의 서베이를 통한 횡단 데이터나 종단 패널데이터에서는 수집하기가 양적으로, 질적으로 불가능한 것으로서 개인, 가족 및 사업체의 등록행정자료가 개인식별번호를 바탕으로 통합되어 있는 전국민 등록자료가 구축되었기 때문에 가능한 것이다. 통계청은 행정데이터와 서베이 데이터의 연계, 통합 서비스도 제공 한다. 서베이 조사 응답자들로부터 개인정보 사용에 대한 동의를 얻은 후 이들의 개인번호를 통계청으로 보내면, 통계청이 필요한 등록 데이터와 서베이 자료를 통합하여 비실명화한(de-identified) 자료를 자료 신청자에게 보낸다(신광영, 2017).

4. 북유럽 행정자료 구축 과정에서 개인정보 보호에 대한 문제

북유럽 국가들이 행정자료의 통합을 추진하는 과정에서 개인정보의 보호 문제가 중요한 이슈로 제기되었다. 덴마크에서는 1977년부터 통계자료 생산을 위해 개인 식별번호를 사용하는 것의 적법성 여부, 그리고 이른바 빅브라더(Big Brother) 사회 도래의 위험성에 대한 논쟁이 벌어졌다. 이 문제를 둘러싸고 통계청과 데이터 제공을 거부한 5개의 지방정부가 법적 분쟁까지 갔는데, 결국 1981년 통계청이 대법원에서 승소하였다. 이런 과정을 거치면서 행정자료 통합과 통계적 활용에 대한 사회적 논쟁이 정리된 한편, 개인정보 보호에 대한 규제의 강화 역시 병행하여 이루어졌다(신광영 외, 2018).

스웨덴에서는 개인정보의 유출에 따른 프라이버시 침해에 대한 우려가 일찍부터 제기되어 1969년 인구 및 주택센서스에 대한 반대가 있었다. 이에 따라 1973년 스웨덴 의회에서 데이터법(Datalagen)이 통과되어 데이터 감독위원회(Date Inspection Board; DIB)가 세계 최초로 법무부 산하에 설립되었다. DIB는 개인정보를 보호하기 위하여 등록 데이터의 수집, 축적, 사용과 분석에 엄격한 제한을 가하였다. 그러나, 민간 기업이나 사회과학자들은 데이터를 보다 자유롭게 이용하고자 하였다. 1976년 3월 스웨덴 사회과학연구위원회(Swedish Council of Social Science Research)는 “개인정보 보안과 사회과학에서의 데이터 필요”에 관한 심포지엄을 개최하였다. 심포지엄에서 얀손(Janson, 1976: 43-45)은 경험적인 데이터 없이는 스웨덴 사회과학은 철학적이고 훈고학적인 수준에 머물 것이라고 주장하면서, 특히 복잡한 문제를 다루는 종단분석(longitudinal investigation)을 위해 공공 행정데이터가 필요하다고 강조했다. 스웨덴 행정데이터는 이런 점에서 세계에서 가장 좋은 데이터임에도 불구하고 잘 사용이 되지 못하고 있는데, 이는 스웨덴의 사회과학 연구 방법론이 횡단 서베이 자료를 분석하는 연구 방법론에 치우진 미국의 영향을 너무 많이 받은 결과라고 보았다. 스웨덴 사회학 연구가 기여할 수 있는 부분은 바로 이러한 행정 데이터를 이용한 종단 분석에 있다고 주장하였다. 그리고 이를 위해서는 개인식별정보를 활용해 행정 데이터간의 통합, 그리고 행정 데이터와 서베이 데이터를 연계시키는 연구가 절대적으로 필요하다고 보아 경직된 데이터법을 비판했다(신광영, 2017).

이후 스웨덴에서는 개인정보의 보호와 활용의 조화를 위한 노력과 함께 논쟁도 계속되었다. 디지털 정보의 등장과 컴퓨터 사용의 확대로 개인 정보에 대한 위협이 더욱 높아지면서 데이터 감독위원회(DIB)는 데이터 사용을 더 엄격하게 하고자 하였지만, 통계청은 개인정보 보호의 수준을 낮추어야 한다고 주장하였다. 스웨덴의 개인정보 보호정책은 1990년대 들어서는 EU 가입으로 변화를 겪었다. 스웨덴은 유럽연합 회원국들에게 공통적으로 적용되는 〈개인 데이터법〉에 의해 개인정보 중에서도 ‘민감한’(sensitive) 개인정보를 구분하게 되었다. 민감한 개인정보

는 인종과 민족, 정치적 견해, 종교와 철학적 신념, 노조가입 여부, 건강과 성생활에 관한 정보들이다. 민감한 데이터의 분석이 허용되는 경우는 정보를 제공한 당사자가 분석에 명확히 동의하거나 데이터를 직접 분석하여 출판하는 경우, 비영리 기구 내에서의 정보 처리, 고용법에서 요구하는 의무의 완수나 건강과 의료 보장을 위해 필요한 경우로 제한되었다(The Ministry of Justice, 2006: 17). 최근에는 EU 차원에서 『일반 데이터 보호 규제(General Data Protection Regulation)』가 채택, 2018년 5월부터 발효됨에 따라 스웨덴 의회는 새로운 데이터 보호법을 통과, 발효시켰다. 새 법은 개인정보 보호를 섹터별로 구체화하고, 민감한 개인 정보의 가공은 당사자나 고용, 의료보장, 사회보장 등 특정한 기관과 학술적 연구에만 허용하고, 개인 프라이버시가 침해되었을 때의 보상 등을 포함하고 있다(신광영, 2017).

북유럽 국가들은 전국민 등록기반 통합 행정자료를 통계적으로 활용하는 과정에서 개인정보 보호 및 사생활 침해의 위험성을 최소화하는 한편 연구자들에게 마이크로데이터의 접근과 활용을 최대한 허용한다는 두 원칙을 동시에 충족하기 위한 노력을 기울였다. 2000년대 이전에는 노르웨이, 핀란드, 스웨덴 통계청들은 자격이 인증된 연구기관들에 마이크로데이터를 암호화된 테이프, CD 등의 형태로 제한적으로 제공했고, 덴마크는 통계청 내부의 보안 공간에서 연구자들이 직접 접근하는 방식만 허용했다. 그러나, 2000년대 이후부터는 기술 발전으로 인해 보안 수준을 충분히 엄격하게 유지하면서도 연구자들의 마이크로데이터에 대한 원격 접속(remote access)이 가능하게 되었다. 스웨덴 통계청은 2005년 마이크로데이터의 온라인 접근 시스템(Microdata Online Access; MONA)을 도입했다. 다만, '민감한 개인 정보'를 포함한 데이터의 사용은 데이터 감독위원회(DIB)의 허가를 받아야 한다. MONA 시스템은 마이크로데이터에 접속한 장소와 시간, 접속 연구자, 그리고 어떻게 활용되었는지에 대한 모든 정보를 수집, 감독한다. 덴마크의 경우 2000년부터 연구자들을 위한 원격 데이터 접근 시스템(Remote Data Access for Researchers; RDAR)을 구축, 발전시켰다(신광영 외, 2018).

전국민 등록행정자료의 마이크로데이터 활용을 연구자들에게 폭넓게 허용하는 북유럽의 시스템은 전통적 서베이 데이터로는 불가능했던 양질의 사회과학 및 생명, 의료과학 등의 연구를 가능하게 하고 있다. 최근 국제적으로 명망 있는 사회과학 학술지에 실리는 논문들의 연구대상 국가에 스웨덴, 노르웨이, 덴마크, 핀란드 등 북구국가들이 과대 대표되는 데에는 이들의 발전된 복지국가에 대한 관심 못지 않게 행정 빅데이터 활용이 중요한 한 요인이다(신광영 외, 2018). 예를 들어 스웨덴의 행정등록데이터를 이용한 한 연구는 세 세대 내지 네 세대간에 걸쳐 증조부 내지 고조부의 사회경제적 배경이 자손들에게 미치는 영향을 분석하고 있다. 스웨덴 통계청이 2000년에 구축한 다세대 행정등록자료 덕분에 이러한 연구가 가능했다(Hällsten 2014).

5. 영국의 행정자료 활용

영국은 스칸디나비아 국가들보다는 뒤늦게 행정자료의 활용에 나섰지만 2000년대 초반 이후 빠른 속도로 행정자료를 이용한 과학적, 학술적 연구를 지원하는 시스템을 수립했다. 영국은 1960년대 정부 행정기록들의 디지털화를 시작하였으며, 1970년대 후반 이후 정부 부처들이 방대한 행정자료의 연구적 가치를 인지하고 활용하기 시작하였으나 주로 부처 내부 연구팀을 운영하거나 대학 연구팀에게 용역을 주는 등 제한적으로만 활용하였다. 정부기관들이 막대한 예산을 투입하여 구축한 데이터들의 연구를 위한 활용도는 낮았으며, 연구자들이 공공기관에 데이터를 요청하면 공개까지 장기적인 시간이 소요될 뿐만 아니라 일관성 없는 개방원칙에 따라 유사한 행정데이터라 하더라도 공개여부가 부처마다 다르게 결정되는 경우가 비일비재 하였다(Administrative Data Taskforce, 2012).

2000년대에 들어서며 많은 연구자들이 행정 데이터 활용의 중요성을 인식, 정부 부처들에게 분산 관리되는 데이터를 서로 공유하고 효율적으로 제공할 필요성을 촉구했다. 이에 영국 경제사회연구재단(UK Economic and Social Research Council, ESRC)이 2006년 행정데이터 연계 서비스(Administrative Data Liaison Service, ADLS)와 안전데이터서비스(Secure Data Service, SDS)에 재정 지원을 결정하고 2008년부터 서비스를 시작하였다. 그럼에도 부처 간, 지역 간 데이터 공유 및 다른 데이터들 간 연계 과정은 법적인 문제, 정부부처들의 비협조 등으로 여전히 매끄럽지 않았고, 이런 가운데 기술적 실수로 인해 개인정보가 다량 유출, 손실되는 사고들이 발생하면서 정부 부처들이 더욱 주저하게 되어 행정 데이터의 연구 활용은 정체되었다. 다만, 스코틀랜드 지역에서 상대적으로 행정데이터 활용 연구가 진전되었고, 의료과학 쪽에서 선제적으로 국가 보건 서비스의 데이터를 모으고 새로운 법도 만들어 의료 연구에 혁신을 가져오고 있었다.

2010년 경제사회연구재단(ESRC)에 스코틀랜드에서 종단데이터 연계 경험이 풍부한 연구자인 랭랜즈(Sir Alan Langlands)가 이사장으로 부임한 후 폴 보일(Paul Boyle)을 중심으로 행정데이터 태스크포스를 만들었다(신광영 외, 2018; 이건 윤광석, 2016). 2012년 발행된 태스크포스의 보고서는 행정자료 연구 네트워크(Administrative Data Research Network; ADRN)의 구성과 부처 간 갈등과 데이터 공유에 있어 발생되는 문제들을 정리할 새로운 법의 제정을 제안했다. 행정자료 연구네트워크(ADRN)는 네 개의 지역(잉글랜드, 스코틀랜드, 웨일즈, 북아일랜드)에 각각 한 개씩 설치되는 행정자료 연구센터(ADRC)와 에섹스 대학(University of Essex)에 설치된 행정자료 서비스센터(ADS) 및 이사회(Governing board)로 구성되는 독립된 국가 기구이다(Administrative Data Taskforce, 2012). 한편 영국 정부는 2011년부터 국가 데이터 포털

인 'data.gov.uk'를 구축하여 일반인들이 간편하게 온라인으로 행정데이터를 요청할 수 있게 하였다.

영국 내 4개 권역에 설립된 행정데이터연구센터(ADRC)는 정부 각 부처에서 생성되는 방대한 행정데이터가 활용되는 다양한 연구과제들을 주선하며 여러 부처에서 상이한 데이터를 연결, 통합하여 보다 다양한 연구가 가능하도록 제3의 데이터를 창출하는 역할을 하였다. ADRC의 가장 중요한 기술적 지원은 행정데이터의 개인정보를 비식별화(de-identified)하는 작업이라 할 수 있다. 또한 ADRC는 행정데이터가 연구 목적과 무관하게 유출되고 비연구자들과 공유되는 것을 방지하기 위해 연구자에게 보안이 유지되는 데이터 접근 시설(secure access facility)을 제공하는 역할도 한다(Administrative Data Taskforce, 2012).

ADRC의 역할은 기술지원에 국한되지 않고 행정데이터를 요청한 연구 프로젝트의 적합성 여부를 따지는 심의기능도 가지고 있다. 적합성 기준으로는 연구가 비상업적(non-commercial)이어야 하고 공중에 이익(public benefit)을 줄 수 있어야 하는 등이다. 특히 ADRC는 자연 및 이공계의 연구보다는 정책문제를 해결해주기 위한 경제 및 사회과학 연구의 지원에 중점을 두고 있다. ADRN은 새로 제정된 디지털 경제법에 따라 2018년부터 영국 행정자료연구(Administrative Data Research UK; ADR UK)로 변경되었다(https://www.adruk.org). ADR UK는 영국통계청(Offices for National Statistics)와 협력하여 개인 정보를 보호하면서, 보다 국민의 삶을 증진시키는데 도움을 줄 수 있는 연구를 수행하는 연구자들에게 행정데이터 서비스를 제공하는 역할을 하고 있다. ADR UK는 디지털경제법에 근거하여 정보주체의 동의 없이도 개인정보를 처리할 수 있지만, 대중에게 자신들의 개인정보가 어떤 방식으로 사용되고 있는지 공개하는 윤리적 책임을 인정하고 수행한다(한은희, 2019).

ADRN는 5년 활동을 평가하고 18가지 개선 사항을 제안한 보고서를 발행하였는데, 이중 가장 중요한 제안은 ADRN의 "데이터 생성 및 파기' 모델에 관한 것이었다. 즉, 승인된 개별 연구에 대해서 행정데이터를 연계하여 데이터 셋(data set)을 생성하고 분석이 끝난 후에는 파기하는 방식의 기존 모델은 경제적으로 지속 가능하지 않기 때문에, 지속적인 연구가 필요한 연구주제들을 선정하고 통합 데이터 셋을 구축하여 안전한 환경에서 지속으로 재활용하는 것이 바람직하다는 것이다(Jones et al., 2019).

ADR UK는 국립통계자료윤리자문회의(National Statistical Data Ethics Advisory Committee, NSDEAC)와 협력하여, 행정데이터 접근, 사용, 공유가 윤리적이고 공익에 부합하도록 하게 한다. ADR UK는 5가지 안전(five safes, 데이터 안전, 개인 안전, 프로젝트 안전, 장소 안전, 결과물 안전)을 내세우며, 행정데이터 구축, 사용과 분석 결과 공표 과정을 통제하고 있다. 2017년 디지털 경제법(Digital Economy Act 2017) 하에서 행정데이터에 접근이 가능하며, 연구자는 연구 계획

서를 제출하여 NSDEAC의 허가를 받아야 한다. 허락을 받은 연구자는 5년 간 행정데이터 접근을 보장받는다. 허가를 받은 연구자 목록은 영국 통계청 웹사이트에 게시된다.

6. 미국의 행정 빅데이터 구축과 활용

증거기반 정책 수립을 위한 행정 빅데이터의 통합 활용에 있어 미국은 북유럽 국가들에 비하면 후발 국가이지만, 1990년 이후 급속하게 발전시키고 있다. 2016년 미 의회가 초당적 입법으로 설립한 증거기반 정책수립 위원회(Commission on Evidence-Based Policymaking, 2017)는 2017년 10월에 최종 보고서를 발표했는데, 행정정보에 대한 접근의 확대와 사생활 보호가 양립할 수 있음을 강조하고 있다. 이 보고서는 행정 데이터 연계 통합을 촉진시키기 위해 여러가지 구체적인 지침을 제시하고 있으며, 이에 따라 2018년에는 증거기반 정책수립 기본법(Foundations for Evidence-Based Policymaking Act)이 제정되어 그동안 분산적으로 더디게 이루어져온 작업을 체계화하고 연도별 실행계획을 세워 가속도를 내고 있다(Penner and Dodge, 2019).

행정자료 통합은 연방정부와 주정부, 그리고 시정부 차원에서 각각 이루어지고 있다. 연방정부 수준에서는 센서스국(Census Bureau), 국세청(Internal Revenue Services, IRS), 사회보장국(Social Security Adminstration, SSA)을 중심으로 이루어지고 있다(이도훈 김창환, 2018). 행정자료와 서베이 자료를 통합하는 방법으로 한국의 주민등록번호에 해당하는 사회보장번호(Social Security Number)를 익명화한 보호식별키(protected identification key; PIK)를 부여하며, 사회보장번호가 없는 데이터의 경우에는 주소, 이름, 성, 연령 변수 등을 이용해 PIK을 부여하고 있다. 현재 미 연방 센서스국에서 제공하는 서베이-행정자료 통합 데이터는 이 방식에 의해서 이루어지고 있는데, 매칭 성공률이 대체로 80-90%에 이른다고 한다. 서베이 자료와 국세청 소득 자료와의 매칭은 사전에 응답자로부터 승인을 받고 있는데, Opt-out 옵션만을 제공하여 명시적으로 매칭에 반대하는 응답자의 경우에만 국세청 자료와의 통합에서 제외된다. 행정자료와의 통합이 이루어진 케이스와 그렇지 않은 케이스에 체계적인 편향이 있는지를 연구한 결과에 따르면 두 케이스 간에 체계적인 편향이 없는 것으로 보고되었다(Czajka, Mabli and Cody 2008).

행정자료의 통합 이후 연구자들에게 제공할 때에는 개인 식별자를 제거한다. 연구자의 연구 목적 이용 외의 데이터 유출을 막기 위해 행정자료 접근 방식은 연방 통계연구 데이터 센터((Federal Statistical) Research Data Centers; RDC)와 같이 인터넷과 차단된 시스템(secure enclave)에 저장하여 접근을 엄격하게 규제하고 관리한다. 연방 인구조사국은 1994년에 RDC

를 매사추세츠주 케임브리지의 국민경제연구소(National Bureau of Economic Research; NBER)에 열었는데, 현재는 29곳에 RDC를 운영하고 있다. 일반 연구자들은 RDC에 계획서를 제출하고, 승인을 받고, 보안 교육을 받은 후, 인터넷 접근이 차단된 Secure Data Center에서 승인된 연구를 수행한다. 대학 연구자들이 RDC를 거치지 않고 연방정부 자료를 이용한 연구를 수행하는 방법으로는 "정부 부처 간 인력 교환" 프로그램이 있다. 예를 들어 미연방 사회보장국에서 대학 연구자의 연구 시간의 20%를 연방정부 프로젝트를 위해서 지불하는 방식으로 이 경우 대학 연구자는 연방 정부의 임시 공무원 신분을 획득하여 데이터에 접근한다(이도훈 김창환, 2018).

미국에서 행정자료를 이용한 연구가 증가하게 된 이유는 북유럽 국가들의 영향도 있었지만, 서베이 조사에서 소득 등에 대한 무응답의 증가에 있었다. 미국에서 소득불평등 측정과 노동시장 연구의 가장 기본적인 자료가 되는 "현재인구조사(Current Population Survey, CPS)"의 소득 항목 무응답률이 꾸준히 증가하여 1983년에는 14.3%였지만, 2008년에는 33.4%로 증가했다. 또한 복지 혜택에 대한 응답률 역시 꾸준히 감소하여 미국의 대표적 복지 프로그램인 OASDI(Old-age, Survivors, and Disability Insurance) 수혜자의 응답거부율이 1991년 21%에서 2013년에는 35%까지 증가했다(Meyer et al. 2015).

행정자료를 이용한 소득 불평등 연구의 한 주제로 조사자료에서 소득의 측정 오차 정도에 대한 연구가 이루어졌다. 국세청(Internal Revenue Service, IRS)의 소득 신고 행정자료와 센서스국(U.S. Census Bureau)의 서베이 자료를 링크함으로써, 서베이 조사에서 응답자들이 자신의 소득을 얼마나 과소 또는 과대 보고하는지에 대한 평가가 가능해졌다. 소득이 낮은 계층에서는 실제보다 소득을 높여서 서베이에 응답하고, 소득이 높은 계층에서는 실제보다 소득을 낮춰서 응답하는 경향이 나타났다. 이러한 응답 경향은 서베이 조사 결과를 통한 소득불평등 지수가 실제보다 불평등을 과소평가할 위험성이 있음을 시사한다. 한편 서베이의 무작위적 응답 에러는 서베이 소득자료를 통한 소득 불평등이 과대 계상될 수 있는 위험을 내포하고 있다. Kim and Tamborini(2014)에 따르면, 미국의 서베이 자료를 이용한 소득불평등 측정은 위 두 가지가 서로 상쇄 작용을 일으켜 행정 자료를 이용한 소득 불평등 측정과 비교했을 때 신뢰도를 손상시킬 정도의 큰 차이가 없다고 한다.

미국의 사회과학 최고 학술지들에 실리는 논문들 중 서베이 자료만을 활용한 논문의 비중은 줄고 있고 행정 자료 또는 서베이 자료와 행정자료를 통합, 활용한 논문의 비중이 급격히 늘고 있다(Grusky et al. 2019). 현재 행정 자료를 이용한 연구가 가장 활발한 분야는 교육정책과 소득 불평등인데, 교육정책에 대한 연구가 활발한 가장 큰 이유는 각 주정부에서 연구자들에게 교육 행정자료를 제공하고 있기 때문으로 보인다(이도훈 김창환, 2018).

미국의 행정 빅데이터 구축과 통합 활용 사례 중 우리가 벤치마킹할 만한 사례를 간략히 소개한다. 첫째는 미국 국세청(IRS)이 최근 구축한 전국민 개인별 과세자료(full-population administrative tax data)의 패널 데이터(SOI Databank)이다. 최근 미국 국세청은 과거의 국세청 소득통계자료(traditional SOI files)가 세금신고를 하지 않는 저소득층을 누락한 문제를 극복하여 미국인 전인구에 대하여 자세한 과세자료 패널 데이터를 자녀 및 고용주들과 연계할 수 있도록 구축하였다. 즉, 종합소득신고를 하지 않는 저소득자들의 수입과 원천징수 세금도 이들에게 임금 등을 지급한 고용주나 사업자들의 신고 자료에 포함되기 때문에 전산화된 자료들로부터 파악이 가능하다. 이렇게 만들어진 SOI Databank는 모든 미국인에 대하여 1996년부터 2015년까지 20행씩, 총 90억행에다 100개가 넘는 열에 소득과 세금 관련 변수들과 가족 및 고용주 등과의 링크 등을 담고 있는 빅 데이터이다(Chetty et al. 2018).

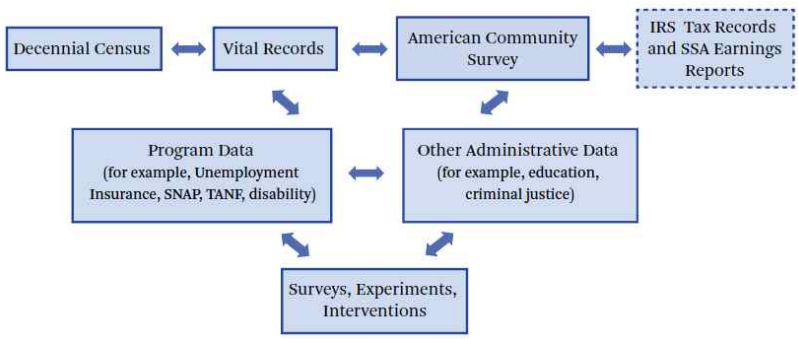
다음으로 현재 진행중인 야심적인 프로젝트는 국세청에서 파악하지 못하는 소득까지 다 포함하는 “포괄적 소득 데이터셋”(Comprehensive Income Dataset)을 구축하는 작업이다(Medalia et al., 2019). 센서스국의 메달리아(C. Medalia)를 비롯한 5명의 연구자들의 공동 프로젝트로서 이들은 과세자료(tax records)와 가계조사 자료(household surveys) 및 복지급여 자료(program participation data) 등을 개인별, 가족별, 가구별로 통합하여 각각의 자료들이 가지는 장점들을 활용하고 단점들을 보완하여 가장 정확한 소득을 측정해내는 것이다. 가계조사 자료로는 현재인구조사(Current Population Survey), 소득과 프로그램참여조사(Survey of Income and Program Participation), 미국공동체 조사(American Community Survey), 소비자지출조사(Consumer Expenditure Survey) 자료들을, 과세자료로는 종합소득신고(Forms 1040), 임금 및 원천징수 세금지급(W-2), 연금 및 보험금 등 지급(1099-R), 근로장려세(EITC) 등 세액공제(tax credits), 실업보험 급여(Unemployment Insurance) 등을, 그리고 연방정부와 주정부의 각종 급여 프로그램 정보들 중 상당한 자료들을 이미 통합했으며, 아직 통합되지 않은 자료들을 순차적으로 포함시키는 계획이다. 각각의 가계조사, 과세자료, 프로그램 자료들은 사회보장번호를 익명화한 개인식별키(personal identification key; PIK)를 포함하고 있어 자료간 통합이 가능하다. 과세자료와 대부분의 복지급여 프로그램 자료들은 99%의 개인들에 대해 PIK이 부여되어 있지만, 서베이 자료의 경우에는 대체로 90-97%의 개인과 가구에만 PIK이 있어 연계가 불가능한 개인들에 대해서는 서베이의 가중치를 낮추는 방법을 사용한다(Medalia et al., 2019). 이미 이 자료를 활용한 연구들이 나오고 있는데, 가령 한 연구는 각종 복지급여 프로그램들의 빈곤감소 효과를 측정한 결과 기존의 연구들이 효과를 과다 측정한 경우와 과소 측정한 경우들이 있음을 밝혀냈다(Meyer and Wu, 2018).

다음으로 야심 찬 프로젝트는 그러스키 등이 인구조사국(Census Bureau)과의 협력 하에 진

행중인 “미국인의 기회 연구”(The American Opportunity Study) 프로젝트이다(Grusky et al., 2019). The American Opportunity Study는 1960년부터 2010년까지의 인구총조사 데이터와 미국공동체조사(American Community Survey) 데이터를 다양한 행정 데이터와 사회보장 데이터 및 서베이와 실험 데이터를 연결하여 지난 70년에 걸친 미국인 전체인구에 대한 대규모 패널 데이터로 전환시키는 작업이다. 이 통합된 패널 행정 빅데이터는 향후 새로운 인구총조사와 미국공동체 조사 자료가 덧붙여짐에 따라 지속적으로 갱신될 것이며, 여기에 과세자료, 근로소득 자료, 복지급여 등 다른 행정자료들과 서베이 및 실험자료 등이 연결될 수 있을 것이다.

아래 〈그림 2〉는 이 데이터 연결 프로젝트의 설계도를 보여주고 있다. 예를 들어 위에서 언급한 PIK가 모든 인구총조사와 미국공동체조사에 포함된 개인에게 부여된다는 전제 하에, 각 개인의 출생 및 사망 정보는 Vital Records에서, 사회보장 프로그램(예: 실업 보험) 참여 및 수당액 정보는 각 Program Data에서, 교육 성취 관련 정보는 주 단위 교육행정 데이터에서 추출하여 변수화할 수 있다. 이들 정보가 패널 데이터로 결합된다는 것은 한 개인의 생애단계에 걸친 세대내 이동성 뿐만 아니라 그 개인의 부모 및 자녀와 연결됨에 따라 세대간 이동성을 다양한 측면에서—즉, 소득, 교육, 사회보장 프로그램 참여 등—파악할 수 있게 해줌을 의미한다. 더구나 이렇게 구성된 대규모 패널 데이터가 서베이나 실험 데이터와 연계될 경우, 행정 데이터에서 변수화하기 어려운 시기, 태도, 정책 등에 따른 고유한 차원들을 측정하여 변수로 추가한 분석이 가능해진다. 따라서 이 통합 패널 빅데이터는 세대내 및 세대간 이동성을 포함한 장기간의 노동시장 과정과 인구학적 변화과정을 분석할 수 있도록 할 것이며, 다른 데이터들을 추가로 연결함에 따라 프로그램 및 실험의 장기 효과에 대한 연구를 가능하게 할 것이다.

[그림 2] The American Opportunity Study의 설계도



자료: Grusky et al.(2019)

7. 한국의 행정자료 통합 활용 연구의 현황과 활성화를 위한 과제

한국은 전자정부와 공공데이터 개방에 있어서 세계 최상위권으로 평가받고 있다(United Nations, 2018). 정부는 공공데이터 포털(www.data.go.kr)을 통해 일반인에게 공공데이터를 적극적으로 개방하고 있으며 금융, 보건의료 등 데이터의 산업적 활용 요구가 높은 분야에서 빅데이터 플랫폼을 구축, 일반에 공개하고 있다. 최근에는 이른바 '데이터 3법' 입법을 통해 데이터의 개방 및 공개 의제를 더욱 적극적으로 추진하고 있다. 그럼에도 사회과학과 정책연구에 있어서 행정 빅데이터의 통합 활용은 매우 더디게 이루어지고 있는 실정이다. 국책연구기관의 일부 프로젝트 외에 행정 데이터의 정책연구 활용은 매우 제한적으로만 이루어지고 있으며 특히 이를 가능하게 하는 시스템 구축은 아직 요원한 실정이다. 본 장에서는 현재 한국에서 행정자료 통합 활용의 현황을 간략하게 정리, 평가하고 현재 행정 데이터의 연구적 잠재성에 대한 미진한 활용 상황을 개선하기 위한 방향을 정리하고자 한다. 특히 이 과정에서 해결해야 할 가장 핵심적 과제라고 할 수 있는 개인정보 보호에 대한 우려에 대해 논하고 행정 빅데이터 인프라와 함께 데이터 거버넌스 시스템 구축을 제안하고자 한다.

현재 행정 데이터를 근거기반 정책적 관점에서 연계 통합하여 활용하는 것은 주로 통계청과 건강보험공단을 통해서 이루어져오고 있다. 통계청과 건강보험공단 외에도 행정 빅데이터를 연계하는 발전적 시도들이 교육부 자료 등에서 이루어지고 있기는 하지만 전반적으로 이상의 노력들이 조율, 통합되기보다는 산발적으로 이루어지고 있는 실정이다.

외국의 사례에서 보듯이 행정자료의 구축과 통합 활용에 가장 중요한 역할을 할 것으로 기대되는 공공부문 주체는 통계청이다. 한국의 경우, 통계청이 다양한 방식으로 행정 빅데이터 구축을 추진하고 있지만 행정 빅데이터 통합 시스템 구축을 핵심 의제로 삼고 이 과정을 주도하는 모습은 보이지 않는다. 현재 통계청이 추진해 온 사업은 등록 센서스 도입 및 구축(2015년 이후), 가계금융복지조사의 행정자료 연계 보완(2017년 이후), 그리고 2018년부터 이루어진 통계빅데이터센터(http://data.kostat.go.kr/) 개설이다.

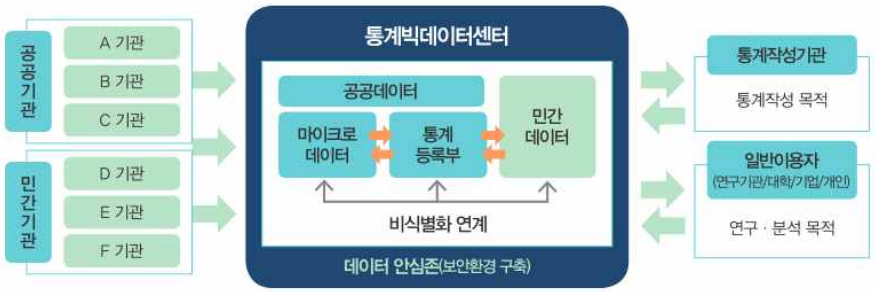
등록 센서스의 경우, 통계청은 2015년부터 기존 전 가구 방문조사 방식의 인구주택총조사를 주민등록부, 건축물대장 등 행정등록자료를 활용해 현장조사 없이 자료를 생산하는 등록 센서스로 대체하기 시작했다. 이에 따라 2015년에는 행정자료로 파악하기 어려운 항목들만 20% 표본 통해 현장 조사하였다. 북유럽 국가들처럼 행정등록자료로 현장조사를 완전 대체하는 수준에 아직 이르지는 못했지만, 5년마다 실시하는 센서스 외에 2016년부터는 핵심항목들에 대해 행정자료만으로 등록 센서스를 매년 구축하기 시작했다. 북유럽 국가들과 같이 등록 센서스를 통해 동일한 가구와 개인을 시계열적으로 추적할 수 있는 패널 시스템 구축은 아직 이루어지지

못한 상황이다. 통계청은 한국인구학회에 등록 센서스 패널 구축 방안 등에 대한 연구용역을 맡기기도 했으나(한국인구학회, 2017), 아직 구체적으로 패널 구축에 관한 계획이 발표된 것은 없다. 등록 센서스의 정책 연구 측면에서 가장 큰 잠재력이 패널 조사와 비교할 수 없는 양질의 패널 데이터를 구축하고 이를 다양한 자료들 및 정책 실험 연구들과 연계하여 활용할 수 있다는 점을 고려할 때, 현재의 등록 센서스 발전 상황은 아직 아쉬운 점이 있다고 할 수 있다.

통계청이 행정 빅데이터를 성공적으로 활용하고 있는 다른 분야는 가계금융복지조사와 국세청 세금 행정자료의 연계이다. 통계청이 소득분배 지표 생산을 위해 활용했던 가계동향조사의 경우 조사 설계상 고소득층의 과소 대표 문제가 심각하고 불평등 지표의 과소 추정의 문제가 오랜 동안 지적되어 왔기 때문에, 이를 가계금융복지조사로 대체했다. 가계금융복지조사는 특히 2017년부터 가구/개인 소득 정보를 국세청 등의 행정자료와 연계함으로써 자가 보고로부터 오는 오류와 편향을 보완하기 시작했다(본 기획특집 중 김낙년의 논문 참고). 이는 행정자료 통합 활용과 관련, 가장 의미 있는 성과라고 평가할 수 있다. 특히 가계조사에서 피하기 어려운 고소득층의 소득 과소 보고하는 경향 문제를 해소하는데 있어 중요한 진전이라고 할 수 있겠다.

끝으로 통계청이 2018년부터 개설한 통계빅데이터센터를 들 수 있다. 기존에 통계청이 제공해 오던 마이크로데이터 제공에 더해 새로이 개설된 통계빅데이터센터는 외부와 차단된 공간에서 데이터를 자유롭게 결합 및 분석할 수 있는 '데이터 안심존'을 통해 개인정보 유출 우려없이 공공데이터와 민간데이터의 연계·융합을 지원하는 이른바 "보안 환경이 구축된 '데이터 분석 플랫폼'"을 지향한다(그림 3). 그러나 이러한 지향과는 달리 통계빅데이터센터가 제공하는 자료는 사회과학 및 정책 연구 관점에서 볼 때 여전히 매우 제한적이다. 빅데이터센터가 포괄하는 행정 빅데이터는 등록센서스(인구, 가구, 주택DB), 사업장 기초DB(4대보험 및 법인) 및 기업등록부, 농림어업 DB 등이며, 그 외에 한국데이터산업진흥원이 민간기관(SKT)에서 구입하여 제공하는 40여 종의 민간데이터를 제공하고 있다. 일부 기관들과 통계청이 MOU를 체결하고 데이터 연결을 시도하고 있지만(예를 들어, 코리아크레딧뷰로(KCB)와의 MOU를 통해 통계청 기업등록부 및 등록센서스 자료와 KCB 대출 잔액 및 연체 잔액 등을 연계하여 개인기업 부채 분석을 실시하고 사업환경지표를 개발), 이러한 사례는 드물고 실제로 연구자들이 과학 및 정책 연구에 통계빅데이터센터의 자료를 활용하는 성과는 아직 부진한 상황이다. 그 이유는 빅데이터센터가 통합, 연계하여 제공하는 행정 데이터 자료 범위 및 정보적 부가가치가 매우 제한적인 것에 반해 데이터 접근 및 활용에 대한 조건은 매우 엄격하고 실효성이 낮기 때문이다.

[그림 3] 통계청의 통계빅데이터센터 개념도



자료: 통계청(2018). http://sti.kostat.go.kr/window/2018b/main/2018_winter_16.html

통계청 이상으로 행정 빅데이터 구축 및 통합 활용에 가장 앞섰던 분야는 의외로 '민감한' 개인정보인 건강관련 정보를 다루는 보건의료분야로 보인다. 국민건강보험공단, 건강보험심사평가원, 질병관리본부, 국립 암센터 등 네 개 기관의 보건의료 데이터를 결합한 '보건의료 빅데이터 플랫폼'(hcdl.mohw.go.kr)이 2019년 9월에 개통되었고, 약 2년간 시범사업이 진행되고 있다(한은희, 2019).

전 국민의 소득 및 재산에 대한 통합 연계 빅데이터를 구축에 앞선 곳도 국세청이나 통계청이 아닌 국민건강보험공단이다. 건강보험공단은 국민건강보험법에 따라 보험료 부과를 위해 국세청 등 여러 부처로부터 소득과 재산 관련 자료들을 제공받고 있다. 국세청 종합소득세, 건강보험공단 직장가입자 연말정산 자료, 직장가입자 퇴직중간정산 자료와 5대 공적연금 자료를 활용하여 소득을 파악하고 있고, 행안부의 재산세 자료와 국세청의 종합부동산세, 국토교통부의 자동차 자료 등으로 자산을 파악하고 있다(전병유, 2018). 여기에는 금융소득과 금융자산 등 일부 소득과 자산 정보는 누락되어 있지만, 다양한 정보원 간 불일치하는 소득개념과 정보 입수 시기 등을 조화시켜 제공한다. 가령, 소득정보 중 종합과세 대상이 아닌 2천만원 미만의 금융소득, 종합소득 신고에서 빠진 연말정산 사업소득과 기타소득, 종합소득신고나 연말정산 데이터에 포함되어 있는 비소비지출(세금, 사회보장기여금 등)에 관한 정보, 그리고 연금소득 이외에 고용보험이나 기초연금과 같은 공적이전 소득 등은 포괄하지 못하고 있다. 이와 같은 불완전성에도 불구하고 현재로서는 한국의 전국민의 소득-자산에 대해 구축된 가장 진전된 패널 데이터(2002년 이후)라고 평가할 수 있다. 특히, 건강보험공단의 소득-자산 데이터베이스는 부모 세대와 자녀 세대를 연결하는 세대 간 데이터베이스를 내부적으로 구축하여 단순히 가구 소득-자산 분포 정보를 활용하는 것을 넘어 소득 및 자산이 세대 간 어떻게 전이되는지 그리고 그 양상이 시계열적으로 어떻게 변화하는지를 파악할 수 있는 빅데이터 자료원을 보유하고 있다(계봉오 외 2019).

이처럼 민감한 건강정보를 다루는 보건의료분야가 소득, 재산 등 비민감 정보를 다루는 일반 사회보장 분야보다 빅데이터 구축에 앞선 것은 아이러니컬하다. 2016년부터 2019년까지 건보공단의 건강정보 자료제공 심의위원회를 통해 2,500건 이상의 자료가 제공되어 보건의료 분야의 연구에 활용되었다고 한다. 경제학이나 사회복지 정책 연구에서 행정 빅데이터의 활용이 거의 전무했던 것과 대조가 된다. 이는 박근혜 정부와 문재인 정부의 공공데이터 개방정책이 주로 제4차 산업혁명을 내세운 기업 측의 요구에 부응하여 형성되었고, 특히 보건의료 산업의 요구가 중요한 동력으로 작용한 것으로 추측된다. 또한, 사회정책과 사회과학 연구자들이 미국이나 영국 등과는 달리 행정자료 활용을 적극적으로 요구하거나 시도하지 않았음을 보여주는 것이라고도 할 수 있다.

또 하나 언급되어야 할 한계는 여전히 정교한 정책 및 학술적 연구를 위한 정보의 접근성이 제한적이라는 점이다. 건강보험공단은 현재 연구계획 심사와 안전 공간 접속 제한을 통해 연구 DB를 외부 연구자들에게 제공하고 있지만, 건강보험 자료공유서비스(https://nhiss.nhis.or.kr/)를 통해 일반적으로 제공되는 변수의 목록과 정보는 상당히 제한적이다. 즉, 건보공단의 전국민 소득-재산 DB에 포함된 여러 변수들이 자료공유서비스 사이트에는 나타나지 않고 보험료 10분위 변수만이 제공되고 있어 소득과 재산에 관한 보다 상세한 자료를 접근하려면 특별한 절차를 필요로 한다.[1)]

인구, 소득 및 재산, 보건 및 의료 정보의 활용에 비해 교육과 같은 그 외 사회 부문에서의 행정 데이터 활용은 미진한 상황이다. 그러나 현재 교육통계 부문에서도 행정 데이터의 연계 활용이 일정 정도의 진전이 있었다. 대표적인 예는 2018년부터 고등교육기관 졸업자 통계를 국민건강보험공단, 국세청, 고용노동부, 병무청, 산업인력공단 등의 행정 데이터와 연계하기 시작한 것이다(교육부·한국교육개발원 2018). 그 이전까지는 개별 대학들로부터 보고되는 정보를 취합하는 것을 바탕으로 취업 상태 및 고용 지위 통계를 작성하던 것에서 행정 빅데이터와 연계를 통해 보다 정확한 대졸자들의 고용 지위 상태를 파악할 수 있게 되었고 더불어 이전에는 조사할 수 없었던 대졸자들의 초임 급여 정보도 파악하기 시작했다. 이는 학력 수준, 더 나아가 더 상세한 출신학교 및 전공 등에 따라 청년들의 노동시장 성과가 어떻게 차별적으로 나타나는지 정교하게 근거를 파악하고 논의할 수 있는 기초를 제공하는 것이라는 점에서 중요한 진전이라고 평가할 수 있다. 그러나 가능성에 비해 한계 역시 뚜렷하다. 첫째, 초임 급여/소득

1) 전국민 소득-재산DB를 이용하려면 건강보험 자료공유서비스(https://nhiss.nhis.or.kr/)의 절차를 따라 연구계획에 대한 소속기관 IRB 승인과 함께 건보공단 자료제공심의위원회의 승인을 요하며, 공단본부의 빅데이터실 내에서만 자료접근과 분석이 가능하며 분석결과만을 승인을 받아 반출할 수 있다. 타기관 자료와의 연계는 법적인 근거(사회보장급여 이용제공과 수급자 발굴에 관한 법률 등)에 따라 해당 국가기관이나 지자체 등의 승인을 요한다.

에 대한 마이크로데이터는 전혀 제공되지 않고 있다. 현재는 전공별 및 학위수준별(학부, 석사, 박사) 당해 졸업생들의 소득 분포를 5개의 100만원 단위 월 소득 집단 범주로 분류하여 제공하고 있다. 둘째, 졸업생 청년들의 노동시장 경력의 생애 궤적을 추적할 수 있는 패널 데이터 구축은 아직 이루어지지 않은 상태이다. 졸업 직후의 고용 정보 및 초임 급여 정보만 일회성으로 연계하고 있다는 한계를 가지고 있다. 사회정책적 차원의 중요성에 비해 현재 청년들의 초기 노동시장 성과를 중장기적으로 추적하는 데이터가 사실상 전무하다는 점(이수빈·최성수 2020)을 생각할 때, 행정 빅데이터 연계를 활용한 패널 데이터 구축은 매우 필요한 과제라 생각된다.

이상 간략하게 소개한 통계청의 행정 빅데이터 통합의 노력은 이전에 비하면 상당한 진보라고 할 수 평가할 수 있다. 그러나 여전히 많은 부분에서 노력에 비해 실효성과 접근성이 떨어지고 따라서 방대하면서도 정확한 행정 빅데이터의 잠재력을 충분히 활용하여 근거기반 정책연구 활성화를 가져오기에는 아직 부족한 상황으로 보인다. 이러한 한계들이 기인하는 원인들과 그 원인들을 극복하고 나아가야 하는 방향에 대해 크게 네 가지 측면에서 논의 및 제언하고자 한다.

첫째, 전 국민 등록 데이터베이스를 바탕으로 다양한 행정 데이터(예를 들어, 소득 및 자산, 교육, 보건/의료, 고용, 주거 등)를 통합하는 종합적인 통계 데이터베이스를 구축, 활용하는 방향으로 발전될 필요가 있다. 현재는 행정 빅데이터를 바탕으로 구축되고 있는 다양한 연계 자료들이 통합적인 기반 아래 이루어지기보다는 산발적으로 이루어지고 있다. 보다 종합적이고 다차원적인 근거기반 정책 연구가 이루어지기 어려운 상황이다. 가령 현재의 등록 센서스로는 북유럽의 사례에서 볼 수 있었던 것처럼 여러 행정자료 및 서베이 조사 자료를 연계 통합시킬 수 있는 전국민 행정 빅데이터 데이터베이스의 모체이자 기본적 플랫폼을 제공하는 수준을 기대하기 어렵다. 다양한 공공기관들이 수집하는 수십 개의 서베이 데이터들이 현재처럼 산발적으로 이뤄지는 대신 센서스나 행정자료를 공통 기반으로 이루어진다면 비용절감, 표본 대표성 및 자료의 정확성을 획기적으로 제고할 수 있다. 응답 기피, 거주자 접근성의 저하 등으로 서베이 비용이 지속적으로 증가할 수밖에 없는 상황에서 행정 정보와의 적극적 연계는 응답 부담을 줄이고 여러 자료원들의 상호보완을 통해 개별 조사들의 효율성과 효과성을 동시에 제고할 수 있는 방법이다.

예를 들어, 조세와 복지정책의 소득 재분배 효과를 조세-급여 미시 모의실험으로 분석하고자 할 때, 연구자들은 주로 재정패널 서베이 자료를 사용한다. 그러나 재정패널 표본은 전 인구 및 가구 대표성이 취약하고 소득의 과대 및 과소보고 문제로 정확성이 많이 떨어지는 등의 문제를 가지고 있다. 그렇다면 현재 가계금융복지조사에서 이루어지고 있는 행정 빅데이터와의 연계를 재정패널로 확대하는 방안을 고려하지 못할 이유가 없다. 이 경우 행정 데이터와 재정

패널을 개별적으로 연계하는 것이 아니라 등록 센서스 같은 전 국민 등록 자료를 바탕으로 통합 데이터베이스를 구축한 후 재정패널 외 다양한 조사자료들을 그 위에 연계시키는 방안을 생각해 볼 수 있다.

가계금융복지조사의 경우 이전에 많은 문제를 노정했던 가계동향조사보다 진일보한 자료이고 무엇보다 조세 자료와의 연결을 통해 성공적인 행정 빅데이터 활용의 모범적 성취 사례를 보여주고 있다. 그러나 가계금융복지조사는 여전히 표본 조사이기 때문에 행정 데이터에 비해 규모가 작고 따라서 최고 소득층과 최저 소득층의 과소 표집 문제를 극복하기 어렵다. 이에 대한 대안은 전 인구 및 가구의 등록기반 소득-재산 데이터베이스의 구축이다. 센서스 데이터를 기반으로 국세청 등에서 제공하는 관련 행정자료들을 통합하고, 행정자료가 파악하지 못하는 부분(예를 들어, 자영업자 사업소득)에 대해서는 서베이 자료로 보완하는 방식이다. 일차적으로 국세청이 과세자료를 기반으로 미국 국세청의 SOI Databank 같이 개인별 소득 데이터베이스를 본격적으로 구축하는 것이 필요하다.

이를 위해서는 핵심적 국가 통계자료 생산자로서 국세청의 역할 제고가 필수적이다. 현재 국세청이 생산하는 국세 통계는 각종 분리과세와 분류과세에 대한 통계를 개별적으로 작성할 뿐 이를 개인별로 합산하여 보고하지는 않고 있다.[2] 또한 마이크로데이터의 공유 및 제공을 극도로 꺼리는 경향이 강하다. 가령 국민건강보험공단에 보험료 산정을 위한 소득-자산 파악에 필요한 과세자료를 제공해주는 것, 가계금융복지조사에 연계해주는 것, 그리고 조세재정연구원에 일부 제공해주는 등 법적으로 명시되어 의무적으로 제공해야 하는 경우 외에는 연구 목적은 물론 다른 정부 부처(예를 들어, 통계청)의 자료 구축 노력에 대한 협력도 소극적이다. 즉, 국세청이 스스로 조세 정보의 정책적 잠재성을 적극적으로 실현시키기 위한 시도가 미진하다는 것이다. 국세청이 전 인구 및 가구의 등록자료에 기반하여 과세자료에 근거한 소득 데이터베이스를 구축할 수 있다면, 여기에 복지수급 및 사회보험 자료 등을 결합하여 미국의 포괄적 소득 데이터셋(Comprehensive Income Dataset) 구축과 같은 작업이 가능할 수 있다. 최근 영국과 핀란드가 도입한 실시간(real time information) 소득정보 시스템 구축(최현수 외, 2018; European Platform Undeclared Work, 2019) 또한 가능할 것이다.

한편 정부는 빅데이터 기반 정책결정과 연구지원으로 사회보장의 과학화를 달성한다는 목

2) 종합소득에 대한 과세는 누진세가 적용되는 종합과세와 일용직 근로소득과 일정규모 이하의 이자나 배당 등을 분리 과세 형태로 원천 징수하는 분리과세로 되어 있다. 한편, 퇴직소득과 양도소득은 분류과세로 별도의 세율을 적용하여 과세하고 있다. 개인의 전체 소득을 파악하려면 종합과세 대상 소득만 아니라 분리과세 및 분류과세 대상 소득을 모두 개인별로 통합해야 한다. 과세자료를 활용해 소득을 온전히 파악하는 데 있어 또 다른 한계는 법에 열거된 소득이 아니면 파악하지 않는 열거주의 과세 원칙, 수입이나 비용 등에서 세무조정을 거치는 점, 과세제외소득과 비과세소득 등이다. 온전한 개인소득 데이터베이스를 구축하기 위해서는 서베이 자료를 활용하여 과세 자료를 보완할 필요가 있다.

표 하에 2019년부터 2021년까지 3년간 총사업비 3,560억원을 들여 차세대 사회보장 정보시스템을 구축, 2022년부터 전면 개방한다는 계획이지다(보건복지부, 2019). 이 엄청난 예산이 투입되는 사업 계획에 전국민 등록기반 소득과 자산을 바탕으로 하는 빅데이터의 구축은 누락되어 있다. 과거처럼 복지 수급자 정보를 통합하는 체계로는 사각지대가 여전히 누락되어 포용적 복지를 구현하기 위한 정보 인프라로서의 역할을 제대로 할 수가 없다. 바람직한 대안은 앞서 언급한 "전국민 등록기반 소득-자산 실시간 정보시스템"을 사회보장정보원이 통계청, 국세청 등과 함께 구축하는 것이다.

"전 국민 등록기반 소득-자산 실시간 정보시스템"이 도입된다면 조세와 복지 행정에 새로운 혁신을 기대할 수 있게 되며 정책연구에도 새로운 기회가 열릴 것으로 기대할 수 있다. 가령, 보건복지부와 사회보장정보원 복지 정보 시스템인 '복지로'(https://bokjiro.go.kr)의 경우 현재는 복지 프로그램 수급자격 및 예상 수급액 추정 프로그램에 각자가 자신의 기억에 따라 소득, 재산 등의 정보를 입력하도록 하고 있다. 하지만 전 국민 등록기반 시스템을 구축한다면 국세청 홈택스에서처럼 정부가 파악하고 있는 정보가 자동으로 연동되어 수급자격 및 예상 수급액을 바로 정확하게 알아보는 것이 가능해진다. 자산 심사에 필요한 복지행정의 인력과 시간 역시 크게 절감할 수 있다. 복지 사각지대의 파악을 통해 찾아가는 복지서비스의 제공은 물론 부정 및 중복수급 파악, 방지에도 큰 효과를 기대할 수 있다. 이렇게 종합적이고 정확한 데이터의 구축을 통해 정교한 조세-급여 마이크로 시뮬레이션 모델링을 하는 것도 역시 가능해진다.

물론 전 국민 기반 통합 데이터베이스 구축과 함께 중요한 것은 이 데이터의 접근성을 높이는 것이다. 지자체 복지담당 공무원 등의 행정을 지원하는 용도로만 제한되었던 기존 사회보장정보시스템이나 공단 내부적으로만 허용되었던 건강보험 빅데이터 등에서 알 수 있듯 현재 접근, 사용체계에서는 행정 빅데이터의 정책적, 과학적 활용 잠재력이 상당히 제한되어 있다. 궁극적으로는 외부 연구자들도 보안 조건 아래서 데이터에 접근, 양질의 정책 근거를 생산할 수 있게 하는 방법을 제고할 필요가 있다. 전 국민의 고용과 소득, 질병, 실업과 퇴직, 공적부조와 연금, 학력 등의 개인정보와 가족원 및 직장의 정보를 종합적으로 그리고 시계열적으로 추적하는 자료를 구축해 제공하는 스웨덴의 노동시장 장기통합데이터(LISA)은 현재 한국의 행정 빅데이터 활용이 벤치마킹해야 할 바람직한 모델을 보여준다.

둘째, 행정 빅데이터 활용 연계 통합 시스템을 구축하는데 있어서 극복해야 할 가장 중요한 도전은 개인정보 보안과 관련된 법적 규제 및 사회적 인식이다. 구미의 경험에서도 드러나는 특징은 행정 데이터의 통합 과정에서 가장 큰 도전은 개인정보 문제의 사회적, 제도적 해법을 찾는 것이었다. 특히 개인정보의 비식별화와 보안 문제 등 기술적인 측면보다 법적, 행정적인 문제 그리고 시민들의 인식 측면이 관건이었다고 볼 수 있다. 빅데이터는 행정 데이터나 민간

데이터 모두 내부자들에 의한 남용, 외부자에 의한 해킹 위험으로부터 자유로울 수 없다. 따라서 그런 위협에 대응하기 위한 기술적 대응과 데이터 거버넌스를 체계적으로 구축하는 것은 행정 빅데이터 활용의 필수 조건이다. 다행스러운 것은 이런 기본적인 조건이 만족된다면 사회과학과 정책연구에 행정 빅데이터를 연계하여 활용하는 것이 개인정보 보호에 실질적으로 위험이 더해질 가능성은 사실상 거의 없을 것으로 예상된다(Penner and Dodge, 2019).

현재 한국에서 행정 빅데이터 공개 및 활용에 있어 더 핵심적인 논란은 익명정보가 아닌 가명정보를 활용한 데이터를 공익적인 학술 및 정책연구 외 이윤 추구를 목적으로 하는 산업적, 상업적 이용에도 허용할 것인가, 허용한다면 어떠한 범위와 조건과 제한 속에서 이루어져야 하는가에 대한 것이다. 익명정보는 외부 데이터를 결합하더라도 개인 식별을 할 수 없는 정보를 의미하며 따라서 개인정보보호법에 적용대상이 아니다. 마이크로데이터가 아닌 마이크로데이터를 바탕으로 작성된 집단 수준 기술 통계 정보를 보여주는 경우라고 할 수 있다. 가명정보는 개인정보를 비식별화하지만 외부 데이터를 연계할 경우 특정 개인을 식별할 수 있을 가능성이 있는 정보를 의미한다. 따라서 가명정보는 개인정보 보호 대상이다(이대희, 2017; 전승재·권현영, 2018). 본고에서 논하는 행정 빅데이터의 연계, 통합 활용 전망은 마이크로데이터의 활용을 전제로 하고 있기 때문에 가명정보가 정책 및 학술 연구를 위해 활용될 수 있도록 법적, 제도적으로 허용하는 문제가 그 핵심에 있다고 할 수 있다. 그런 점에서 이러한 가능성을 법적으로 열어준 데이터 3법은 행정 빅데이터 활용 시스템 구축에 있어서는 중요한 진전이라고 할 수 있다.[3] 그러나 동시에 이 점이 시민들의 데이터 3법에 대한 우려의 핵심이기도 하다. 데이터 3법 개정안에 따르면 가명정보의 이용 가능 범위는 "데이터를 기반으로 한 새로운 기술·제품·서비스의 개발, 산업 목적을 포함하는 과학연구, 시장조사, 상업 목적의 통계작성, 공익 기록보존 등"으로 폭넓게 규정된다. 시민단체들이 우려하는 바는 공익적, 과학적 연구를 매우 폭넓게 정의함으로써 기업과 개인의 사익 추구를 위해서도 활용될 수 있는 가능성이 법적으로 허용될 수 있다는 것이다(오병일, 2019; 이상윤, 2019). 반면 정부에 따르면 이 개정안은 동시에 개인정보의 개념과 책임성을 더 명확하게 하고 개인정보 보호를 위한 거버넌스를 더 실효성 있게 이루어질 수 있도록 체계화하는 내용도 함께 담겨 있기 때문에 오히려 개인정보 보호를 제고하는 측면이 함께 있다(문화체육관광부 2020).

---

3) 데이터 3법은 개인정보 보호법, 정보통신망법, 신용정보법 등 데이터 활용에 관련된 내용을 담고 있는 세 개의 법률을 통칭한다. 이전에는 이들 데이터 관련 법들이 개인정보 식별의 위험성이 극히 낮은 정보까지 개인정보 보호대상으로 해석하고 정보활용을 매우 예외적으로만 허용하게 함으로써 활발한 데이터 활용을 제한했던 경향이 강했다. 지난 2020년 1월 9일 통과한 데이터 3법 개정안은 가명 정보 개념을 도입함으로써 데이터 활용의 활성화를 가능하게 하는 법적 요건을 제공하고, 동시에 실질적인 개인정보 거버넌스가 효율적으로 이루어질 수 있도록 체계를 제공하는 동시에 개인정보 처리의 책임성을 강화하는 내용을 담고 있다(문화체육관광부, 2020. "정책위키: 데이터 3법." http://www.korea.kr/special/policyCurationView.do?newsId=148867915).

데이터 3법 개정안이 통과되었지만 개인정보의 비식별화에 대한 구체적인 기술적 규제는 시행령 등 하위 규칙으로 위임되어 있는 상태이다. 비식별화 과정과 자료의 이용, 보관 등에 대한 지나친 규제는 지양하면서 정보주체의 동의없이 자료를 이용할 수 있는 허용범위에 대한 명확한 규정이 마련되어야 할 필요가 있다. 가령 정부가 2016년에 발표한 개인정보 비식별조치 가이드라인을 보면 행정자료의 결합시 주민등록번호를 사용하여 임시 대체키를 만드는 것은 위법소지가 있다고 판단했다. 주민등록번호를 직접 사용하는 것도 아니고 주민등록번호를 사용해 임시 대체키를 만드는 것까지도 불허하는 것은 행정자료의 통합 활용을 제한하고 데이터 활용 자체를 매우 비효율적으로 만드는 과잉규제적 측면이 강하다. 연구 후에 사용된 자료를 즉시 파기하도록 강제하는 것은 재정적으로 비효율적이고, 연구 검증 및 추가적인 후속연구 측면에서도 바람직한 접근이 아니다. 기술적, 행정적 절차를 통해 실질적인 개인정보 유출의 위험이 극히 낮게 관리하면서도 데이터의 효율적, 효과적 활용이 충분히 가능하며 이를 실질적으로 실현하고 있는 나라들이 있다는 점을 고려할 때 지나치게 위험 회피적이기만 한 현재의 상황은 개선될 필요가 있다. 개정법의 시행령에서 이러한 문제들이 합리적으로 규정됨으로써 개인정보 유출에 대한 우려를 불식하면서도 비식별화한 행정 자료를 학술 연구와 정책 평가 목적으로 효과적으로 활용할 수 있는 방법이 정착될 수 있어야 할 것이다.

셋째, 행정 빅데이터 기반의 학술과 정책 연구가 실질적으로 가능하려면 행정 빅데이터 구축과 활용을 위한 거버넌스 시스템의 구축이 필요하다. 앞서 살펴본 바와 같이 행정 빅데이터 거버넌스 시스템은 정부 주도의 북유럽 모델과 정부와 대학이 협력하는 미국 및 영국 모델이 있다. 한국의 경우 정부 주도 모델로 범 정부 차원의 거버넌스를 구축하되 여기에 학계와 민간이 긴밀하게 결합되는 방식을 생각해 볼 수 있다. 한국은 북유럽 국가들처럼 전 국민에게 출생시부터 주민등록번호가 주어지기 때문에 행정 빅데이터의 통합이 정책적 의지만 있으면 비교적 용이한 조건을 갖추고 있다(이성균 외, 2018). 북유럽 국가들처럼 인구주택총조사를 행정자료에 의한 등록센서스로 완전 대체하고 등록센서스의 패널을 구축하며, 전국민 등록기반 행정자료 통합을 추진할 수 있는 기술적, 행정적 기반은 상당히 갖추어져 있다. 현재 부족한 것은 이러한 기술적, 행정적 잠재성을 현실화시키기 위한 거버넌스 시스템이 필요하다는 것이다.

가령 북유럽의 경우 모든 공공기관이 생성하는 등록 자료들을 완전하게 접근하고 조정, 관리할 권한을 통계청에게 부여하여 행정 빅데이터 통합 활용을 촉진하는 작업을 추진할 수 있었다. 한국의 경우, 현재 유명무실한 통계청 산하의 국가통계위원회의 위상을 높여 국세 자료와 사회보험 자료, 기타 행정통계, 서베이 자료들의 연계와 통합, 공개와 활용 범위에 관한 권한과 책임을 부여하는 방식의 거버넌스 구조를 고려해 볼 수 있다. 구체적인 거버넌스 구조와 내용은 한국과 북유럽 행정의 차이를 감안하여 보다 유연하게 설계할 수 있을 것이다. 이러한 거버

넌스 구조 설계 구상에 있어서 앞서 살펴본 북유럽의 경험과 함께 유럽연합 각국의 법적 제도적 환경, 모범적 관행에 관한 유럽 통계시스템(European Statistical System) 보고서를 참고할 만하다(Santourian and Petrakas, 2018; Santourian et al., 2018).

한편으로는 영미식으로 학계의 적극적인 이니셔티브 하에 행정자료를 통합 활용하는 연구 프로젝트들을 연구재단과 경제인문사회연구회 등이 지원, 협력하는 방안을 고려해 볼 수도 있다. 가령 조세-급여 모델 구축을 위한 프로젝트, 교육과 고용에 관한 장기 패널데이터 구축과 이를 통한 교육과 불평등, 세대간 이동에 관한 연구 프로젝트 등을 생각해 볼 수 있다. 광역 및 기초 지자체 수준에서 소득 불평등과 빈곤에 대한 지표들을 측정하고 각종 복지 프로그램의 효과를 추정하는 프로젝트도 필요한 상황이다. 현재는 이러한 지표가 없기 때문에 지자체별 지역사회보장계획을 보면 소득 불평등이나 빈곤의 축소와 같은 정책목표 설정이나 성과 평가가 불가능한 상황이다. 중앙 정부 중심(예를 들어, 통계청 중심)의 거버넌스 시스템 구축과 학계와의 협력을 중심으로 하는 시스템 구축 중 반드시 택일할 필요는 없다. 한국의 상황에 맞게 두 모델의 장점을 유연하게 혼합하는 접근이 바람직하다. 한국의 경우 데이터 수집 및 구축에 있어서 학계의 경험과 역량이 취약한 편인 반면 데이터의 활용과 정책 근거 생산에 있어서 정부와 학계의 협력이 활발하기 때문에 북유럽과 비슷한 정부 중심의 행정 빅데이터 거버넌스 구조를 구축하면서 이 과정에서 학계와의 적극적인 협력 구조를 제도화하는 방안을 생각할 수 있겠다.

넷째, 개인의 소득과 조세정보의 전면적 공개 필요성이 제기된다. 한국에서 질 좋은 행정 데이터의 구축과 통합 활용에 있어 가장 큰 걸림돌이 되는 것은 개인정보 보호를 이유로 국세청 등이 소득, 자산 등의 데이터 구축과 활용을 꺼리는 데 있다. 개인정보보호법에 의하면 건강정보는 민감정보로 취급되지만 소득과 세금 등은 그 자체로 민감한 정보가 아닌데 이는 한국과 북유럽 모두 마찬가지이다. 그럼에도 민감한 정보가 포함된 보건의료 빅데이터는 일찍부터 구축되고 상당히 활용되어온 반면, 소득-자산-조세의 경우 데이터 구축이 매우 미진한 현실이다. 앞서 언급했듯이 건강보험공단 데이터 중에도 민감한 건강정보를 활용한 연구는 많이 이루어졌지만, 소득-재산 데이터베이스의 활용은 극도로 제한되어왔다. 이런 상황은 단지 개인정보에 대한 법적인 제한으로만 설명하기 어렵고 문화적, 인식적 차원의 요인이 있다고 생각된다.

이러한 걸림돌을 극복하는 방안으로 한국에서도 북유럽 국가들처럼 개인의 소득과 조세정보를 전면적으로 공개하는 방안을 검토할 필요가 있다(유종성, 2019b). 조세정보를 개인정보로 엄격히 보호하는 대부분의 국가들과 달리 스웨덴, 노르웨이, 핀란드, 아이슬란드 등 북유럽 국가들은 이미 19세기 중반부터 개인과 기업의 조세 정보를 공개해 왔다. 누구나 지방 세무서나 시청을 방문해서 다른 사람들의 조세 정보를 열람할 수 있다. 노르웨이의 경우 2001년부터 인터넷 상에서 타인의 소득과 납세액을 쉽게 검색할 수 있다. 스웨덴에서는 각 지역별로 매년 조

세달력(tax calendar)을 발간하여 과거 한국의 전화번호부와 비슷한 형식으로 알파벳 순으로 이름, 주소와 함께 근로소득(earned income), 불로소득(unearned income), 결정세액을 기록, 공개하고 있으며 기업에 대해서도 소득과 세액을 볼 수 있다.

북유럽 국가들처럼 투명한 사회, 사회적 신뢰가 높은 사회, 높은 수준의 세금을 요구하는 높은 수준의 복지국가를 이루고자 한다면, 한국에서도 모든 소득자의 종합소득 신고를 의무화하고 조세 정보를 공개하는 방안을 적극적으로 논의해 볼 필요가 있다고 생각된다. 공직자를 포함, 김영란법 적용 대상자부터 우선 공개하는 방안을 검토할 수도 있다. 재산정보 공개의 경우 기존의 공직자 재산등록 및 공개제도에서 출발하여 공개 대상자를 점차 확대하는 방향으로 도입을 하는 것도 가능하다. 개인별 소득, 재산과 과세정보에 대한 폭넓은 공개는 질 좋은 행정자료의 구축과 통합 활용을 촉진하는데 필요한 과정이기도 하며, 더 나아가서는 한국 사회의 신뢰와 투명성을 높이는데 일조할 것이라 생각된다.

8. 결론

다양한 사회문제를 해결하기 위한 정책적 대응이 더욱 중요해지고 있다. 민주화의 상당한 성취에도 불구하고, 정책이 없는 정치가 낳은 폐해는 정치에 대한 불신과 혐오감으로 이어지고 있다. 이념만 내세우는 정치는 소모적인 정쟁으로 이어져 정치불신을 더 심화시키고 있다. 조세와 복지정책의 효과를 데이터를 기반으로 엄밀하게 분석할 수 없는 경우, 정책적 논의는 증거에 기반을 둔 논의가 아니라 이념적 논의로 흐르기 쉽고, 이는 정치적, 사회적 갈등을 증폭시키게 된다.

사회정책을 선진적으로 발전시킨 북유럽 국가들에서는 증거기반 정책 논의를 꾸준히 발전시켰다. 북유럽 국가들의 노동정책과 사회정책은 추상적인 사회민주주의 이념이 아니라 실사구시 차원에서 이루어진 실질적인 효과를 발휘하는 정책이라는 점에서 공통적인 특징이 있다. 철저하게 현실에서 출발하는 정책적 논의를 통해서 현실에서 현실을 개혁하는 수단과 방법을 찾는다는 점에서 철학적인 접근이 아니라 사회과학적인 접근이다. 단지, 평등과 연대, 성장과 분배, 포용과 통합의 가치를 실현하기 위한 실사구시적 접근이라는 점에서 특징을 찾을 수 있다.

그 중심에는 행정 빅데이터가 있다. 공공기관들이 수집 및 관리하고 있는 각종 데이터를 통합하여, 정책 논의와 학술적 연구에 활용하는 행정 빅데이터는 행정비용의 절감과 과학적 논의의 발전을 동시에 도모할 수 있는 새로운 흐름으로 등장하고 있다. 데이터 부족으로 인하여 관념적인 논쟁으로 흐를 수 밖에 없는 정책 담론이 경험적인 증거와 검증을 통한 논의로 발전될

수 있었던 것은 논쟁을 검증 가능케 하는 경험적 자료와 분석에 있다. 북유럽 국가들의 선진적인 정책들은 이러한 분석과 검증을 통한 증거기반 정책의 발전에 기반을 두고 있다. 이제 유럽연합과 미국도 이러한 흐름에 적극적으로 동참하고 있으며, 행정 빅데이터가 오늘날의 사회과학 흐름을 바꿔놓고 있다. 서베이 데이터가 지닌 한계를 극복하고, 보다 실체적 현실에 접근하기 위한 사회과학적 흐름으로 행정 빅데이터의 활용이 점차 필수적이 되고 있다.

한국은 행정 빅데이터를 구축하고 활용할 수 있는 행정적, 기술적 기반은 상당히 갖춰져 있다고 생각된다. 행정 빅데이터 연계 구축에 핵심적인 요소인 개인식별은 주민등록번호 정보를 활용할 수 있으며 행정 빅데이터 구축에 필요한 비식별 조치 또한 선진적인 디지털 기술을 이용하여 해결 가능하다. 한국이 서베이 데이터를 이용한 사회과학 연구와 정책 형성에는 뒤늦었지만, 행정 빅데이터를 이용한 연구와 정책에 있어서는 오히려 다른 국가들을 선도할 수 있는 역량이 충분히 갖춰져 있는 셈이다. 새로운 흐름을 만들 수 있는 정부의 인식 전환과 정책 전환이 필요한 시점이다. 아울러 학계에서도 증거기반 정책을 위한 연구에 행정 빅데이터를 활용하기 위한 노력을 기울일 필요가 있다고 본다.

▌참고문헌 ▐

계봉오, 황선재, 최율 (2019). 행정자료를 이용한 세대 간 분위소득 이동 분석. 정해식 외(편). 소득불평등 심화의 원인과 정책적 대응 효과 연구 2. 서울: 경제인문사회연구회 협동연구총서 18-34-01.

교육부, 한국교육개발원 (2018). 고등교육기관 졸업자 취업통계연보. 진천: 한국교육개발원.

김창환, 이도훈 (2018). 미국 행정자료 통합 사례. 강신욱 외(편). 소득불평등 심화의 원인과 정책적 대응 효과 연구. 서울: 경제인문사회연구회 협동연구총서 18-05-01.

보건복지부 차세대 사회보장정보시스템 구축 추진단 (2019). 사회보장 정보전달체계 개편 기본방향.

문화체육관광부 (2020). 정책위키: 데이터 3법. http://www.korea.kr/special/policyCurationView.do?newsId=148867915

신광영 (2017). 스웨덴의 행정 데이터 통합과 활용에 관한 연구. 스칸디나비아 연구. *20.* 83-108.

신광영, 최성수, 김영미 (2018). 유럽 국가들의 행정자료 통합 및 활용. 강신욱 외(편). 소득불평등 심화의 원인과 정책적 대응 효과 연구. 서울: 경제인문사회연구회 협동연구총서 18-05-01.

오미애 (2013). 보건복지 분야 데이터 통합 연계방안에 대한 고찰. 보건복지포럼. *203.* 34-41.

오미애 (2014). 정부3.0과 빅데이터: 보건복지 사례를 중심으로, 보건복지 Issues& Focus. *230.* 1-8.

오미애 (2019). 보건복지정책에서의 빅데이터 활용 전략과 과제. 보건복지포럼. *274.* 29-40.

오병일 (2019). 빅데이터 시대, 개인정보 보호체계의 강화가 필요하다. 월간 복지동향. *251.* 20-27.

유종성 (2019a). 복지국가를 위한 행정 빅데이터 구축과 조세정보 공개의 필요성. 월간 복지동향. *251.* 28-36.

유종성 (2019b). 북유럽 복지국가의 비결은 소득공개. 랩2050.

이건, 윤광석 (2016). 영국의 ADRN 사례를 적용한 공공데이터 개방 및 활용 촉진을 위한 거버넌스 구축 모색. 한국행정학회 학술발표논문집. 638-660.

이대희 (2017). 개인정보 보호 및 활용 방안으로서의 가명·비식별정보 개념의 연구. 정보법학. *21(3).* 217-251.

이도훈, 김창환 (2018). 미국 행정자료 통합 사례. 강신욱 외(편). 소득불평등 심화의 원인과 정책적 대응 효과 연구. 서울: 경제인문사회연구회 협동연구총서 18-05-01.

이상윤 (2019). 보건의료 빅데이터 사업, 국민 건강을 위한 것인가, 기업 이윤을 위한 것 인가. 월간 복지동향. *251.* 5-11.

이성균, 계봉오, 최율, 황선재 (2018). 한국 행정자료 통합. 강신욱 외(편). 소득불평등 심화의 원인과 정책적 대응 효과 연구. 서울: 경제인문사회연구회 협동연구총서 18-05-01.

이수빈, 최성수 (2020). 한국 대학들의 사회이동 성적표: 경제적 지위의 세대 간 이동과 유지에서 대학이 하는 역할. 한국사회학. *54(1).* 181-240.

전병유 (2018). 소득 관련 통계 인프라 실태 파악과 개선방안. 정해식 외(편). 소득불평등 심화의 원인과 정책적 대응 효과 연구 2. 서울: 경제인문사회연구회 협동연구총서 18-34-01.

전승재, 권현영 (2018). 개인정보, 가명정보, 익명정보에 관한 4개국 법제 비교분석 정보법학. *22(3).* 183-218.

전주열 (2016). 프랑스 공공데이터 법률제정 최근 동향. 외법논집. *40(4).* 63-80.

채향석 (2017). 빅데이터 시대의 개인정보보호 자율규제 활성화 방안. 고려법학. *85.* 41-80.

최승필 (2017). 행정정보의 공동이용에 대한 법적 쟁점의 검토. 행정법연구. *51.* 109-130.

최현수, 백승호, 진재현, 고금지 (2018). 4차산업혁명 대응을 위한 데이터 주도의 혁신적 포용 사회안전망 개편방안. 세종: 한국보건사회연구원 연구보고서 2018-46.

한국인구학회 (2017). 인구주택총조사 자료 활용성 증대를 위한 심층연구: 최종보고서. 통계청.

한은희 (2019). 사회보장 통합 행정데이터 구축 및 활용의 쟁점과 전망. 사회보장정보원 사회보장정보 이슈리포트 제12호.

Administrative Data Taskforce (2012). The UK Administrative Data Research Network: Improving Access for Research and Policy. Report from the Administrative Data Taskforce.

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R. & Narang, J. (2017). The fading American dream: Trends in absolute income mobility since 1940. Science. *356.* 398-406.

Chetty, R., Friedman, J. N., Saez, E. & Yagan, D. (2018). The SOI Databank: A case study in leveraging administrative data in support of evidence-based policymaking. Statistical Journal of the IAOS. *34.* 99-103. DOI: 10.3233/SJI-170418.

Commission on Evidence-Based Policymaking (2017). The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking.

Connelly, R., Playford, C. J., Gayle, V. & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. Social Science Research. *59.* 1-12.

Czajka, J. L., Mabli, J. & Cody, S. (2008). Sample Loss and Survey Bias in Estimates of Social Security Beneficiaries: A Tale of Two Surveys. (Final Report, contract no. 0600-01-60121) Mathematica Policy Research.

European Platform Undeclared Work (2019). The Incomes Register: Finland.

Friedman, J. N. (2015). The Value of Administrative Data in Policy-Relevant Economic Research. https://www.aeaweb.org/about-aea/committees/ economic-statistics/administrative-data

Grusky, D., Hout, M., Smeeding, T. M. & Snipp, C. M. (2019). The American Opportunity Study: A New Infrastructure for Monitoring Outcomes, Evaluating Policy, and Advancing Basic Science. RSF: The Russell Sage Foundation Journal of the Social Sciences. *5(2).* 20-39.

Hällsten, M. (2014). Inequality Across Three and Four Generations in Egalitarian Sweden: 1st and 2nd Cousin Correlations in Socio-Economic Outcomes. Research in Social Stratification and Mobility. *35(1).* 19-33.

Janson, G. (1976). Longitudinal Studies and Their Need for Data. Proceedings of symposium on Personal Integrity and the need for data in the social sciences, Swedish ed. by Dalenius, Tore and Aners Klevmarken. Council for Social Science Research. 43-48.

Jones, K. H., Heys, S., Tingay, K. S., Jackson, P. & Dibben, C. (2019). The Good, the Bad, the Clunky: Improving the Use of Administrative Data for Research. International Journal of Population Data Science. *4(1).* 03.

Jones, P. & Elias, P. (2006). Administrative data as a research resource: A selected audit. Report to the UK Economic and Social Research Council. Coventry: University of Warwick.

Kim, C. H. & Tamborini, C. R. (2014). Response Errors in Earnings: An Analysis of the Survey of Income and Program Participation Matched with Administrative Data. Sociological Research & Methods. *43(1).* 39-72.

Medalia, C., Meyer, B., O'Hara, A. & Wu, D. (2019). Linking Survey and Administrative Data to

Measure Income, Inequality, and Mobility. International Journal of Population Data Science. *4(1).* 04.

Meyer, B. D. & Wu, D. (2018). The Poverty Reduction of Social Security and Means-Tested Transfers. ILR Review. *71(5).* 1106-1153.

Meyer, B., Mok, W. K. C. & Sullivan, J. X. (2015). Household Surveys in Crisis. Journal of Economic Perspectives. *29(4).* 199-226.

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.

Santourian, A. & Petrakos, M. (2018). Analysis of the legal and institutional environment in the EU Member States and EFTA Countries. ESS.VIP ADMIN, Work Package 1. Access to and development of administrative data sources. https://ec.europa.eu/eurostat/cros/system/files/admin-wp1.1_analysis_legal_institutional_environment_final.pdf

Santourian, A., Kitromillidou, S. & Famarkis, G. (2018). Good practices in accessing, using and contributing to the management of administrative data. ESS.VIP ADMIN, Work Package 1. Access to and development of administrative data sources. https://ec.europa.eu/eurostat/cros/system/files/admin-wp1.2_good_ practices_final.pdf

Speiser, M. (2015). The 10 countries with the world's fastest internet speeds. Business Insider (May 18, 2015). https://www.businessinsider.com/fastest-internet-connection-speeds-2015-5?r=UK

The Ministry of Justice (2006). Personal Data Protection. Stockholm: The Ministry of Justice, Sweden.

Thygesen, L. (2010). The Importance of the Archive Statistical Idea for the Development of Social Statistics and Population and Housing Censuses in Denmark. Nordic Statistical conference, Copenhagen. http://www.dst.dk/ extranet/staticsites/Nordic2010/pdf/bf7d6701-5b9f-4888-adc2-a45ce8debf87.pdf

UNECE (2007). Register-based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics. United Nations.

United Nations (2018). United Nations e-government Survey 2018: Gearing e-government to Support Transformation Towards Sustainable and Resilient Societies. https://publicadministration.un.org/en/research/un-e-government -surveys

Wallgren, A. & Wallgren, B. (2007). *Registered-based Statistics: Administrative Date for Statistical Purposes.* Chichester: Wiley & Sons.

Abstract

Using Administrative Data for Evidence-Based Policy Research

Jong-sung You* · Byung You Cheon** · Kwang-Yeong Shin*** · Dohoon Lee**** · Seongsoo Choi*****

The use of administrative data, or big administrative data, for evidence-based policy research has been rapidly spreading, from Nordic countries at the beginning to other countries in Europe and North America. Although Korea ranks high in international assessment of e-government and open access to government data, the use of administrative data in social sciences and policy research has been very limited. Linking multiple sources of administrative data and survey data and constructing a comprehensive high-quality population data is extremely difficult and hardly achieved. In this paper, we discuss the value of utilizing administrative data for evidence-based policymaking and social science research. We examine how the issue of privacy protection has been addressed in relation to the use of linked administrative data in Scandinavia, the UK, and the US. In conclusion, we urge Statistics Korea to construct a panel dataset of register-based Census and a full-population register-based quality dataset of incomes and assets in collaboration with National Tax Service and other public agencies. Social scientists should take the initiative in launching large research projects using linked administrative data, and National Research Foundation of Korea and National Research Council for Economics, Humanities and Social Sciences need to actively promote such projects.

Keywords: evidence-based policy, administrative data, big data, data linkage, privacy protection

◈ 2020. 2. 1. 접수 / 2020. 3. 7. 1차수정 / 2020. 3. 11. 게재확정

* First author, Professor, Graduate School of Social Policy, Gachon University(jsyou0721@gachon.ac.kr)
** Corresponding author, Associate Professor, Graduate School of Social Innovation Management, Hanshin University (bycheon @hs.ac.kr)
*** CAU-Fellow, Department of Sociology, Chung-Ang University(kyshin20@gmail.com)
**** Associate Professor, Department of Sociology, Yonsei University(dlee2191@yonsei.ac.kr)
***** Assistant Professor, Department of Sociology, Yonsei University(s.choi@yonsei.ac.kr)