ICPSR 데이터 재이용 저작물 분석을 통한 사회과학 분야의 지적구조 분석

An Investigation on Intellectual Structure of Social Sciences Research by Analysing the Publications of ICPSR Data Reuse

정 은 경 (EunKyung Chung)*

목 차

초 록

오픈 사이언스 패러다임과 발달된 디지털 정보기술의 영향으로 여러 학문 분야에서 데이터의 공유와 재이용이 활발해지고 있으며, 데이터 중심(data intensive)의 학술 커뮤니티로 변모하고 있다. 본 논문은 사회과학 분야의 대규모 데이터 리파지토리인 Inter-university Consortium for Political and Social Research(ICPSR)에 수록된 데이터를 재이용한 저작물이 구현한 지적구조를 규명하고자 하였다. 이를 위하여 ICPSR 사이트의 2017년 발간된 데이터 재이용 저작물 570건을 분석의 대상으로 하였다. 분석의 과정은 두 단계를 거쳤다. 첫 번째 단계는 총 570건의 저작물에 대해서 저자, 저작물 형태, 저작물 자체의 주제 분석을 수행하였다. 저자를 살펴보면, 미국 대학과 연구기관 소속 연구자가 출현빈도 비중이 높은 것으로 나타났다. 저작물의 형태는 대부분은 학술지였으며, 이를 학술지 주제 분야로 분석하면, 사회과학, 의학, 심리학 분야로 나타났다. 두 번째 단계의 분석은 저작물의 제목에서 추출한 단어를 대상으로 동시출현단어 분석을 수행하여 군집과 네트워크로 시각화하였다. 이러한 결과는 보다 미시적인 주제 분야의 규명을 위해서 수행되었다. 분석결과 총 12군집인 정신건강, 담배영향, 학교/유년기/청년기장애, 청년기 성적위험, 아동부상, 육체활동, 폭력행동, 서베이, 가족역할, 여성, 문제행동, 성별차이로 구성되었음을 밝혔다. 이러한 결과를 종합적으로 살펴보면, ICPSR 데이터의 재이용을 통해 사회과학적 시각으로 의학 주제 분야의 연구가 비중있게 이루어지고 있음을 알 수 있다.

ABSTRACT

Due to the paradigm of open science and advanced digital information technology, data sharing and re-use have been actively conducted and considered data-intensive in a wide variety of disciplines. This study aims to investigate the intellectual structure portrayed by the research products re-using the data sets from ICPSR. For the purpose of this study, a total of 570 research products published in 2017 from the ICPSR site were collected and analyzed in two folds. First, the authors and publications of those research products were analyzed in order to show the trends of research using ICPSR data. Authors tend to be affiliated with university or research institute in the United States. The subject areas of journals are recognized into Social Sciences, Health, and Psychology. In addition, a network with clustering analysis was conducted with using co-word occurrence from the titles of the research products. The results show that there are 12 clusters, mental health, tabocco effect, disorder in school, childhood, and adolescence, sexual risk, child injuries, physical activity, violent behavior, survey, family role, women, problem behavior, gender differences in research areas. The structure portrayed by ICPSR data re-uses demonstrates that substantial number of studies in Medicine have been conducted with a perspective of social sciences.

키워드: ICPSR, 데이터, 데이터 공유, 데이터 재이용, 지적구조, 동시출현단어분석, 네트워크
ICPSR, Data, Data Sharing, Data Re-use, Intellectual Structure, Co-word Analysis, Network

* 이화여자대학교 사회과학대학 문헌정보학과 부교수(echung@ewha.ac.kr)
논문접수일자: 2018년 1월 22일 최초심사일자: 2018년 1월 22일 게재확정일자: 2018년 2월 12일
한국문헌정보학회지, 52(1): 341-357, 2018. [http://dx.doi.org/10.4275/KSLIS.2018.52.1.341]

1. 서 론

오픈 사이언스는 여러 학문 분야의 중요한 패러다임으로 자리매김하고 있다. 오픈 사이언스 패러다임은 일련의 연구 진행 과정을 투명하게 공개하여, 궁극적으로 연구의 투명성, 진실성, 재현가능성을 높이고자 한다. 이러한 패러다임의 영향으로 연구 과정에서 생성되는 데이터에 대한 공유와 재이용이 활발하게 이루어지고 있다. 또한 인터넷 정보기술의 발전으로 인해 대용량 디지털 데이터를 자유롭게 저장하고, 공유하고, 접근하여 이용하는 것이 보편화되었다. 이러한 기술의 발전은 다양한 학문 분야에서 연구데이터의 공유와 재이용으로 연계되어 학술활동의 중요한 현상 중의 하나로 여겨지고 있다(Borgman 2015).

이와 함께 여러 국가에서 연구기금 기관은 연구기금 수혜 연구자에게 연구과정에서 생성된 데이터를 다른 연구자도 접근하여 재이용할 수 있도록 명시적으로 요구하고 있다. 유럽에서는 Open Research Data Pilot (ODP)를 시작으로 하여 연구 데이터의 접근과 이용을 극대화하려는 노력이 이루어지고 있다. 미국에서는 National Science Foundation (NSF)와 National Institutes of Health (NIH)를 중심으로 연구자가 연구기금을 신청할 때 결과물로 생성된 연구데이터의 관리계획을 제출하도록 의무화하였다. 영국에서도 미국과 유사하게 Economic and Social Science Research Council에서 기금 신청 연구자에게 연구 데이터의 공유와 관리 계획 제출을 의무화하였다(Spengler 2012). 이와 같이 연구데이터를 수집하고, 관리하여 다른 연구자가 접근하게 하여 새로운 연구에서 이용하게 하는 것은 유럽, 미국, 영국 등의 국가에서는 국가적인 책임으로 논의되고 있는 상황이다. 따라서 여러 국가에서는 연구 데이터의 공유와 재이용은 동일하거나 유사한 데이터를 수집하는 소모적 연구 활동을 최소화하고, 하나의 연구데이터를 다양한 학문적 시각으로 분석하여 학제적 해석을 제시할 수 있는 정책을 장려하고 있다.

데이터가 중심이 되는 경향은 학문 분야에 따라서 차이가 있지만, 상당히 보편적인 현상으로 자리를 잡아가고 있다. 최근 연구(Piwowar, Vision and Whitlock 2011)에 따르면, 유전자학(genomics) 분야의 연구에서 2007-2010년의 기간 동안 1,150건의 새로운 논문이 데이터 재이용하여 출간하였다고 밝혔다. 이러한 새로운 연구가 데이터 공유를 통해 가능하게 되었으며, 데이터의 공유는 학술적 기여의 새로운 영역을 열게 되었다. 전통적인 학술지 출판사인 Elsevier나 Springer 등은 데이터에 대한 인용이 참고문헌 목록에 포함되도록 데이터에 관한 정책과 저자에 대한 가이드라인을 제정하였다(Cousijn et al. 2017). 그러나 Silvello(2018)는 여러 학문 분야가 점차 “데이터 중심”으로 발전해가고 있으나, 데이터가 학문 분야에서 어떻게 사용되고 있는지에 대한 실증적인 연구가 부족하다고 지적하였다. 본 연구는 Silvello가 지적한 이러한 간극을 메우기 위한 데이터 중심의 학술 활동에 대한 실증적 규명에 관한 논의라고 할 수 있다. 상대적으로 자연과학이나 공학에 비하여 데이터의 양은 많지 않지만, 데이터의 재이용이 활발한 사회과학 분야를 대상으로 하였다(조재인 2016). 이를 위해서 Inter-university Consortium for Political and Social

Research(ICPSR)[1] 데이터를 재이용하여 형성된 사회과학 분야를 저자와 키워드 네트워크 분석을 통해 규명하고자 한다. 이를 통해서 학문분야의 데이터 재이용이 표현하는 데이터 중심의 학문 현상을 실증적으로 고찰할 수 있다.

2. 관련연구

오픈 사이언스 패러다임에서 데이터의 공유와 재이용에 관한 연구는 대체로 세 그룹으로 구분하여 살펴볼 수 있다. 첫째는 데이터의 공유와 재이용에 미치는 영향을 규명하고자 하는 연구들을 찾아볼 수 있다. 두 번째는 학술지 논문의 인용과 마찬가지로 데이터의 영향력을 측정하는 도구 개발에 대한 시도들이다. 마지막으로는 상대적으로 연구가 미진한 분야로써 데이터의 재이용을 통해 구축되는 학문의 지적구조를 규명하고자 하는 연구들을 찾아볼 수 있다.

첫 번째 연구 그룹으로 데이터의 공유와 재이용에 미치는 영향요인을 규명하려는 시도를 찾아볼 수 있다. Fear(2013)는 ICPSR 데이터를 사용한 저작물과 데이터 이용 연구자를 대상으로 데이터 인용 현황, 데이터 인용 영향력 측정 도구, 데이터 재이용 요인을 밝혔다. 우선 1,500건의 학술지 논문에 실린 사회과학 분야의 데이터 인용 현황을 살펴보면, 대부분의 연구자는 데이터 자체 인용보다는 데이터 생산자의 논문을 인용하는 것으로 나타났다. 그러나 2000-2004년의 기간보다 2005-2012년의 기간에 데이터 인용의 증가를 찾아볼 수 있다. 둘째, 재이용 횟수, 데이터 인용 2차적인 영향력 측정 도구, 다양성, 다운로드 횟수와 같이 4가지 측정 도구를 제안하였다. 셋째로는 사회과학자가 데이터를 재이용하는데 영향을 미치는 요인을 데이터 생산자의 명성, 데이터 처리 상태, 데이터 생산자의 공저자 네트워크의 규모, 데이터 셋과 관련된 출간물의 숫자 등으로 밝혔다. 또한 Faniel, Kriesberg, Yakel(2016)은 ICPSR에서 데이터를 사용한 경험이 있는 연구자를 대상으로 연구데이터 이용 만족에 영향을 미치는 요인을 규명하였다. 데이터 주제적합성, 데이터 완결성, 데이터 접근가능성, 데이터 이용 편리성, 데이터 신뢰성, 데이터 생산자 평판도, 데이터 관련 문서의 품질, 학술지 순위의 8가지 가설을 수립한 후 데이터 이용 만족도와의 연관성을 제시하였다. 총 237건의 서베이 데이터를 분석하여, 학술지 순위와 데이터 주제적합성 가설은 기가되었으며, 나머지 6개의 가설이 받아들여졌다. 사회과학 연구자들은 연구데이터 이용에 있어서 만족도에 영향을 미치는 요인으로 데이터의 완결성, 접근가능성, 이용 편리성, 신뢰성, 생산자 평판도, 관련 문서의 품질이라고 밝혔다. Yoon(2017)은 데이터 재이용에 있어서 신뢰성이 중요하며, 데이터의 품질을 제시하는 표준의 부재를 지적하였다. 이 연구는 반구조화된 인터뷰를 통해 데이터 재이용 과정에서 신뢰를 구축하는 여러 단계를 제시하였다. 연구자가 데이터 재이용을 경험하면서 데이터에 대한 신뢰는 형성되거나, 소멸되거나, 축소되거나, 회복되는 등의 다양한 경험으로 구분될 수 있다는 점을 밝혔다. 이와 함께 Yoon,

1) https://www.icpsr.umich.edu/icpsrweb/

Kim(2017)은 사회과학 분야 연구자의 데이터 재이용 행위를 탐구하고자 하였다. 이를 위해 총 292건의 설문조사 응답을 분석하였다. 분석 결과를 살펴보면, 사회과학자의 데이터 재이용 의도는 주제 분야의 데이터 재이용 규범, 데이터 재이용 대한 태도, 데이터 재이용에 관련된 인지된 노력 등에 영향을 받는다고 밝혔다. 데이터 재이용에 대한 태도는 데이터 재이용 의도에 영향을 미쳐, 데이터의 인지된 유용성과 우려에 간접적으로 영향을 끼친다고 제시하였다. Park, Dietmar(2017)는 데이터 인용이 활발한 유전과 유전형질학 분야의 데이터 공유와 재이용의 특징을 규명하고자 하였다. 이를 위해서 Data Citation Index(DCI)에서 148건의 인용 논문을 탐색하여 데이터 공유와 재이용에 영향을 주는 요인을 규명하였다. 그 요인들은 참고문헌, 본문, 부록의 데이터와 정보, 감사의 글, 연구기금 정보, 저자 정보, 웹 정보원 등으로 나타났다.

두 번째는 데이터의 영향력을 평가하기 위하여 측정 도구를 탐색하는 연구들을 찾아볼 수 있다. 우선 Mooney(2011)는 학술지 논문에서 데이터 인용이 활발하게 이루어지고 있지 않다고 지적하며, ICPSR의 논문을 대상으로 저자의 데이터 인용 현황을 규명하였다. 총 49건을 대상으로 하였으며, 61%(30건)의 논문이 사용한 데이터를 논문에서 인용하지 않았다고 밝혔다. 19건의 논문 중에서 14건(29%)은 참고문헌에서 데이터셋에 대한 공식적인 형태의 인용을 사용한 것으로 나타났다. 나머지 5건은 데이터셋을 인용하지 않고, 데이터와 관련된 저작물의 인용을 사용한 것으로 나타났다. 이러한 결과는 데이터에 대한 명확한 인용이 이루어지지 않고 있어서, 데이터의 공유와 재이용에 대한 장애요인으로 볼 수 있다. He, Han(2017)은 데이터의 영향력을 평가하기 위하여 Dryad 데이터 리파지토리의 데이터 이용 횟수와 관련 논문의 인용 횟수의 상관관계를 비교하였다. 분석결과는 데이터의 이용 횟수와 관련된 논문의 인용 횟수 사이에 상당히 긍정적 상관관계를 보여준다고 밝혔다. 또한 Ingwersen, Chavan(2011)은 데이터의 공유와 재사용을 촉진시키기 위해서는 적정한 학문적 인정 메카니즘이 수반되어야 한다고 제시하였다. 저자는 생물학 분야에서 적정한 학문적 인정의 중요한 요소로 Data Usage Index(DUI)를 제안하였다. DUI는 총 14개의 측정도구를 포함하며 데이터셋의 5가지 측면(지리적 위치, 시간, 주제, 규모, 사용)에 근거하여 가중 사용 측정 도구를 제시하였다.

세 번째는 데이터의 재이용을 통하여 데이터 중심의 학문 분야의 특성을 규명하고자 하는 시도들을 찾아볼 수 있다. 조재인(2016)은 연구데이터 인용색인 데이터베이스인 DCI(Data Citation Index)를 분석하여 연구데이터의 구축 현황, 주제 분야, 고인용 연구데이터의 특성을 규명하였다. 우선 구축 현황을 살펴보면, 2015년 3월 DCI 검색을 통해 살펴본 결과는 총 79개 분야 3,379,301건의 데이터가 수록된 것으로 나타났다. 주제 분야별로 살펴보면, 유전학 분야가 1,772,377건으로 가장 많았으며, 생화학/분자 생물학이 1,355,128건으로 2위를 차지했다. 또한 상위 인용 순으로 500건의 고인용 연구데이터를 분석한 결과, 경제학, 사회학, 인구학, 건강관리/정책학 분야로 나타났으며, 연구데이터의 형식은 설문조사이다. 고인용 연구데이터가

수록된 리파지토리를 살펴보면, Data Archive, ICPSR이 전체의 85% 이상을 차지한 것으로 나타났다. 또한 최형욱, 정은경(2017)은 연구데이터의 공유와 재사용이 점점 증가하는 상황에서 연구데이터 재사용이 표현하는 학문의 특성을 규명하고자 하였다. 이를 위해서 일반적 학술저널의 논문 키워드 기반의 사회학 분야의 지적구조와 DCI(Data Citation Index) 수록 사회학 분야의 연구데이터 키워드를 분석한 사회학 분야의 지적구조를 네트워크로 표현하여 비교분석을 수행하였다. 분석결과, 연구데이터의 재사용을 통하여 본 사회학 분야는 의학 분야의 접목을 특징적으로 살펴볼 수 있으며, 그중에서도 공중보건과 심리학 분야가 특징적으로 나타났다. 본 연구는 세 번째 연구 경향에 속한다고 볼 수 있으며, Silvello(2018)가 지적한 바와 같이 데이터 중심의 학문을 정의하고 규명하기 위해서 실증적으로 데이터의 재이용을 통해 구축된 지적구조의 탐색이 지속적으로 수행될 필요가 있다.

3. 데이터 수집과 분석

3.1 데이터 수집

ICPSR은 Data Archive와 함께 데이터 인용이 가장 많이 되는 데이터 리파지토리이다(조재인 2016). 1962년 미국 미시건대학이 대학 간 컨소시움을 구성한 것이 시작이 되었으며 미국선거연구 데이터 수집과 공유를 목적으로 하였다. 현재는 정치학 분야뿐만 아니라 사회과학 분야 전반으로 확대되었다(Swanberg 2017). ICPSR 데이터를 사용한 저작물의 정보를 수집하기 위해서 ICPSR 사이트에 접속하였다. 해당 사이트의 “Bibliography of Data-Related Literature” 메뉴를 사용하여 2017년 12월 4일부터 8일까지 5일 동안 외부 반출 기능(csv형태)을 사용하여 수집하였다. 본 데이터의 수집기간에 ICPSR 사이트에서는 총 71,809건의 저작물이 ICPSR 데이터를 사용하여 출간되었다고 제시하였다(ICPSR, n.d.). 이 중에서 본 연구의 대상으로 삼은 가장 최근에 발간된 2017년 저작물 총 570건이다. 570건의 논문 정보는 ICPSR 사이트에서 제공하는 바와 같이, 저자명, 논문제목, 저널명, 권, 호, 시작페이지, 종료페이지, 연도로 구성되었다.

ICPSR 데이터를 이용하여 출간한 출간물의 형태는 2017년 12월 4일에서 8일까지 데이터 수집 기간에 ICPSR 홈페이지(ICPSR, n.d.)에서 2008년도 저작물 분석을 통해 제시한 바와 같이 학술지 논문이 가장 많은 비중을 차지하는 것으로 나타났다.

3.2 데이터 분석

본 연구는 2017년도 한 해 동안 ICPSR 데이터를 사용하여 출간한 저작물을 분석의 대상으로 삼았다. 데이터 분석은 세 단계를 거쳐 이루어졌다. 첫째는 저작물의 저자에 대한 분석이다. 저자의 소속 분석을 위해서는 개별 논문을 검색하여 저자의 소속 정보를 수집하였다. ICPSR 데이터를 재이용하여 새로운 연구를 수행한 저자들의 특성을 규명하였다. 둘째는 저작물이 출간된 저작물의 형태에 대한 분석이다. 특히 가장 비중이 큰 학술지에 대해서 주제 분석을 수행

하였다. 학술지의 주제를 규명하기 위해서 가장 많은 학술지와 학술대회 발표집을 수록하고 있는 Scopus를 검색하여 학술지/학술대회 발표집의 주제 분야를 수집하였다. 저작물이 출간된 학술지/학술대회 발표집의 주제 분야를 통해 ICPSR 데이터의 재이용이 형성하는 지식의 구조를 규명할 수 있다. 세 번째는 저작물의 제목에서 추출된 단어를 활용하여 동시출현분석을 수행하였다. 제목 추출 단어 동시출현분석 행렬은 PNNC 군집분석과 네트워크로 시각화하였다. 분석의 과정을 살펴보면, 우선 제목에서 단어를 추출하고, 불용어 제거를 제거하며, 단복수형 일치시키는 작업을 수행하였다. 군집분석과 네트워크 분석을 위해 이재윤이 개발한 소프트웨어인 semi.exe[2] 와 COOC ver 0.4[3]를 사용하였다. 분석된 결과를 네트워크로 시각화하기 위해서 NodeXL이 사용되었다.

4. 결과분석

4.1 저자 특성

우선 저작물의 저자를 살펴보면, ICPSR 데이터를 사용한 저작물 570건에 대해서 공동저자로 인해 총 1,561명의 저자가 나타났다. 〈표 1〉에서 살펴볼 수 있는 바와 같이 저자의 출현빈도별로 분석하면, 7회 출현한 저자는 1명에 불과하며, 6회는 4명, 5회는 2명, 4회는 8명으로 각각 나타났다. 특정 저자의 집중화보다는 다양한 저자가 대체로 1회 논문의 저자로 나타났다.

〈표 1〉 저자출현횟수와 분포

저자출현횟수	저자수	출현횟수 %
7	1	0.1%
6	4	0.3%
5	2	0.1%
4	8	0.5%
3	32	2.0%
2	118	7.6%
1	1,396	89.4%
합계	1,561	100%

이 중에서 4회 이상 나타난 상위에 나타난 저자에 대해서 소속기관과 전공 학문 분야를 저자의 소속기관을 통해 살펴보면, 〈표 2〉와 같다. 우선 소속기관으로 살펴보면, 15명의 상위에 출현하는 연구자들은 모두 미국의 대학이나 연구기관의 소속되어 있는 것으로 나타났다. 둘째, 학문분야를 살펴보면, 대체적으로 건강과학 관련 분야라고 볼 수 있다. 특히 건강 행위 분야(Health Behavior), 암, 부상, 담배, 약물 등과 관련된 학문 분야로 나타났다. 셋째, 15명의 연구자의 소속을 찾아보는데 있어서도 중복된 기관(The Schroeder Institute for Tobacco Research and Policy Studies, Center for Injury Research and Policy)이 나타났다. 이는 동일한 기관에 소속된 연구자들이 공동으로 연구한 결과를 활발하게 발표한 것으로 볼 수 있다. 이러한 결과는 데이터의 재이용은 특정 학문 분야가 다른 분야에 비교하여 활발하게 이루어지기 때문이라고 볼 수 있다.

2) 입력데이터 생성을 위한 응용프로그램
3) 행렬생성을 위한 응용프로그램

〈표 2〉 4회 이상 출현 연구자의 소속기관과 전공 학문 분야

번호	저자명	횟수	소속기관	학문 분야
1	Stanton, Cassandra A.	7	Brown University	Psychiatry and Human behavior
2	Borek, Nicolette	6	Food and Drug Administration	Center for Tobacco Products
3	Bansal-Travers, Maansi	6	Roswell Park Cancer Institute	Department of Health Behavior
4	Pearson, Jennifer L.	6	The Schroeder Institute for Tobacco Research and Policy Studies	The Schroeder Institute for Tobacco Research and Policy Studies
5	Hyland, Andrew	6	The Schroeder Institute for Tobacco Research and Policy Studies	The Schroeder Institute for Tobacco Research and Policy Studies
6	Bhandari, Prem	5	University of Michigan	Population Studies Center
7	Smith, Gary A.	5	The Ohio State University	Center for Injury Research and Policy
8	Conway, Kevin P.	4	National Institute on Drug Abuse	National Institute on Drug Abuse
9	Pierce, John P.	4	University of California, San Diego	UC San Diego Moores Cancer Center
10	Hsu, Hui-Chin	4	University of Georgia	Department of Human Development and Family Science
11	Young-DeMarco, Linda	4	University of Michigan	Institute for Social Research/Survey Research Center
12	Xie, B.	4	Claremont Graduate University	School of Community and Global Health
13	Green, Victoria R.	4	National Institutes of Health	Division of Epidemiology, Services, and Prevention Research
14	Chounthirath, Thiphalak	4	The Research Institute of Nationwide Children's Hospital	Center for Injury Research and Policy
15	Feng, X.	4	The Ohio State University	Department of Human Sciences

4.2 저작물의 출판 특성

ICPSR 데이터를 재이용한 저작물의 형태를 살펴보면, 〈표 3〉과 같다. 학술지, 학술대회 발표집, 단행본, 기타로 구분되었다. 이 중에서 가장 높은 비중을 차지하는 저작물 형태는 학술지(91.8%)이며, 학술대회 발표집이 7.2%로 나타났다. 4건(0.7%)의 단행본과 통상적인 저작물 형태로 구분이 불가한 사례가 2건(0.4%)으로 나타났다.

이 중에서 가장 비중이 높은 학술지에 대하여 살펴보면 [부록 1]과 같이 출현 빈도가 높은 학술지를 찾아볼 수 있다. 이러한 학술지 전체에 대하여 SCOPUS 주제 분야명이 부여되었으며, 이를 추출하여 분석하였다. SCOPUS의 주제 분야명은 27분야의 대분류와 하위 분야로 구성되어 있으며, 복수의 주제를 수록하는 학술지에 대해서는 복수의 주제 분야명이 부여된다.

〈표 3〉 저작물 형태 분포 현황

저작물 형태	횟수	%
학술지	523	91.8
학술대회 발표집	41	7.2
단행본	4	0.7
기타	2	0.4
합계	570	100

〈표 4〉 SCOPUS 대주제 분야별 학술지 분포 현황

SCOPUS 주제 대분류	출현횟수	%
Social Sciences	321	33.0
Medicine	307	31.6
Psychology	170	17.5
Arts and Humanities	38	3.9
Economics, Econometrics and Finance	21	2.2
Nursing	21	2.2
Biochemistry, Genetics and Molecular Biology	15	1.5
Computer Science	15	1.5
Agricultural and Biological Sciences	13	1.3
Environmental Science	11	1.1
Neuroscience	8	0.8
Pharmacology, Toxicology, and Pharmaceutics	8	0.8
Immunology and Microbiology	8	0.8
Health Professions	5	0.5
Engineering	4	0.4
Business, Management, and Accounting	4	0.4
Energy	1	0.1
Decision Sciences	1	0.1
Earth and Planetary Sciences	1	0.1
Multidisciplinary	1	0.1
Mathematics	0	0.0
Chemistry	0	0.0
Material Science	0	0.0
Physics and Astronomy	0	0.0
Chemical Engineering	0	0.0
Dentistry	0	0.0
Veterinary	0	0.0
합계	973*	100.0

* 학술지에 복수로 부여된 주제 분야명으로 인한 합계임

〈표 4〉에서 살펴볼 수 있는 바와 같이 사회과학(Social Sciences)분야가 33%, 의학(Medicine)은 31.6%, 심리학(Psychology) 분야가 17.5%를 차지하는 것으로 나타났다. 의학 분야가 비중있게 나타난 특징을 살펴볼 수 있다. 이러한 결과는 의학 분야 연구에서 인구통계학적 의미를 규명하는 관점에서 사용된 것으로 볼 수 있다.

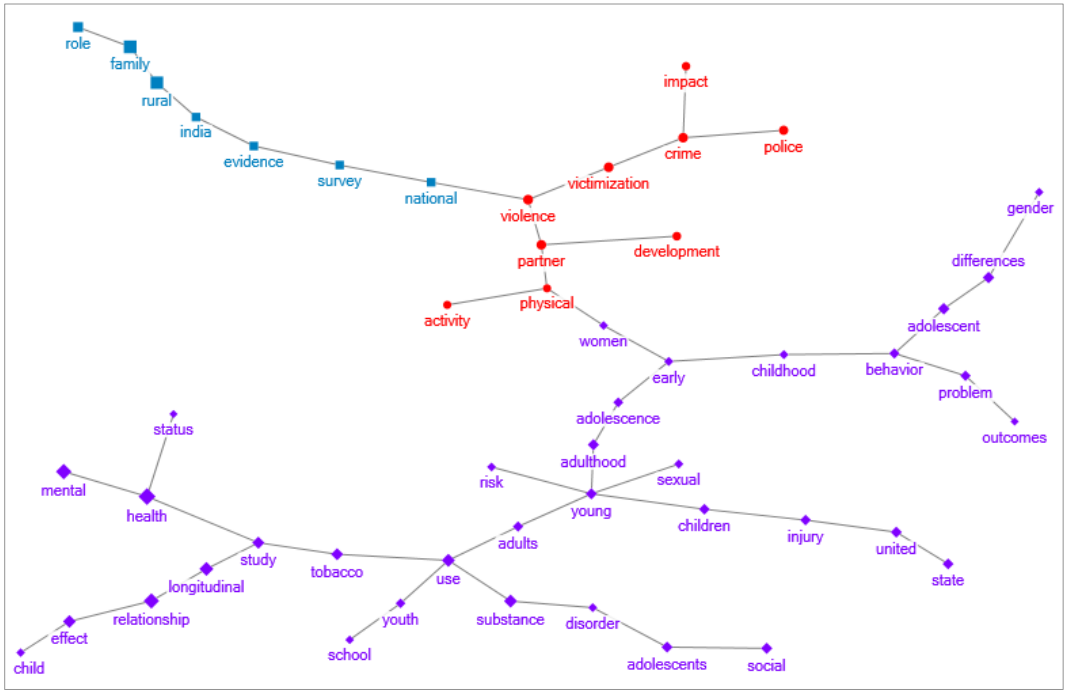
규모가 큰 세 분야를 자세히 살펴보기 위해서 세부 분야와 함께 분석하였다. 〈표 5〉에서 살펴볼 수 있는 바와 같이 사회과학 분야에서 주된 학술지 세부 주제 분야는 법학(Law), 사회과학과 정치학(Sociology and Political Science), 교육학(Education), 건강과학(Health(Social Science)), 사회과학(기타)(Sociology(miscellaneous)), 인구학(Demography), 인류학(Anthropology) 등으로 나타났다. 의학 분야의 상대적으로 많이 출현한 세부 주제 분야는 공중보건학/환경 및 직업 건강학(Public Health, Environmental & Occupational Health), 정신과학 및 정신건강(Psychiatry and Mental Health), 병리학 및 범죄의학(Pathology and Forensic Medicine), 일반의학(General Medicine), 소아과학/주산과학 및 아동 건강학(Pediatrics, Perinatology and Child Health), 의학(기타)(Medicine (miscellaneous)) 등이 나타났다. 심리학 세부 주제 분야에는 발달 및 교육 심리학(Developmental and Educational Psychology), 사회심리학(Social Psychology) 임상심리학(Clinical Psychology), 응용심리학(Applied Psychology), 일반심리학(General Psychology) 등이 비중있게 나타났다.

4.3 제목 키워드 네트워크

ICPSR의 데이터를 재이용한 저작물 570건의 제목에서 단어를 추출하여 동시출현단어 분석을 수행하였다. 출현횟수 15회를 기준으로

〈표 5〉 SCOPUS 세부 주제 분야별 학술지 분포 현황

대분류	세부 분류	출현 횟수	%
Social Sciences	Law	77	23.7
	Sociology and Political Science	52	16.0
	Education	36	11.1
	Health (social science)	32	9.8
	Social Sciences (miscellaneous)	32	9.8
	Demography	23	7.1
	Anthropology	18	5.5
	Gender Studies	9	2.8
	General Social Sciences	9	2.8
	Life-span and Life-course Studies	8	2.5
	Cultural Studies	6	1.8
	Geography, Planning and Development	6	1.8
	Development	4	1.2
	Political Science and International Relations	4	1.2
	Communication	2	0.6
	Urban Studies	2	0.6
	Human Factors and Ergonomics	1	0.3
	Linguistics and Language	1	0.3
	Public Administration	1	0.3
	Safety Research	1	0.3
	Transportation	1	0.3
	소 계		100
Medicine	Public Health, Environmental & Occupational Health	62	20.2
	Psychiatry and Mental Health	52	16.9
	Pathology and Forensic Medicine	46	15.0
	General Medicine	36	11.7
	Pediatrics, Perinatology and Child Health	23	7.5
	Medicine (miscellaneous)	16	5.2
	Geriatrics and Gerontology	15	4.9
	Epidemiology	12	3.9
	Health Policy	10	3.3
	Orthopedics and Sports Medicine	6	2.0
	Surgery	5	1.6
	Emergency Medicine	3	1.0
	Otorhinolaryngology	3	1.0
	Pharmacology (medical)	3	1.0
	Endocrinology, Diabetes and Metabolism	2	0.7
	Infectious Diseases	2	0.7
	Neurology (clinical)	2	0.7
	Obstetrics and Gynecology	2	0.7
	Cardiology and Cardiovascular Medicine	1	0.3
	Dermatology	1	0.3
	Health Informatics	1	0.3
	Microbiology (medical)	1	0.3
	Ophthalmology	1	0.3
	Physiology (medical)	1	0.3
	Rehabilitation	1	0.3
	소 계		100
Psychology	Developmental and Educational Psychology	50	29.4
	Social Psychology	46	27.1
	Clinical Psychology	31	18.2
	Applied Psychology	19	11.2
	General Psychology	14	8.2
	Experimental and Cognitive Psychology	5	2.9
	Psychology (miscellaneous)	5	2.9
	소 계	170	100

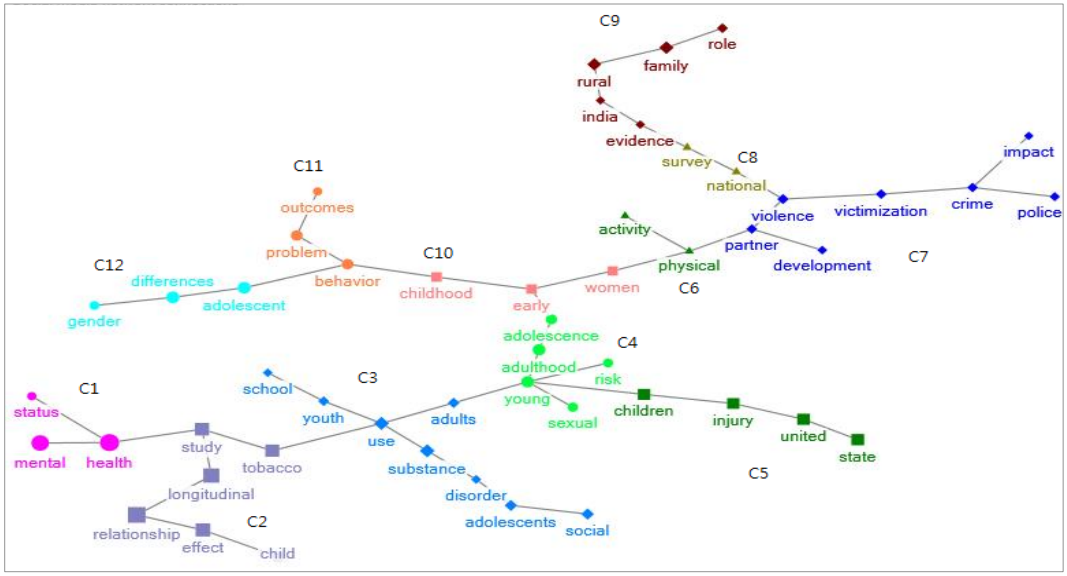


〈그림 1〉 저작물 제목 키워드 동시출현단어 네트워크(3개 군집)

하여 총 51건의 단어를 대상으로 분석하였다([부록 2] 참조). 노드의 크기는 출현횟수 기반하여 상대적으로 표현하였다(〈그림 1〉 참조).

다이아몬드 노드로 표현된 대규모 군집, 원노드로 표현된 가운데 군집, 사각형으로 표현된 왼쪽 군집으로 구분되었다. 가장 규모가 큰 군집은 건강(Health), 관계(Relationship), 청년세대(Young), 행위(Behavior), 담배(Tobacco) 등이 수록되어 있다. 이 군집은 35개의 단어 노드들로 구성이 되어 있으며, 건강, 연령, 성별, 국가, 유해물질 등을 광범위하게 다루었다고 볼 수 있다. 이와 함께 연결되어 있는 원노드 군집은 총 9개 단어 노드로 구성되었다. 육체적, 폭력, 피해, 범위, 경찰 등의 단어 노드로 구성되어 있으며 범죄와 관련 있는 세부 주제로 볼 수 있다. 범죄 군집과 연결되어 있는 사각형 노드 군집은 총 7개의 단어 노드로 구성되어 있다. 이 군집은 가족이나 지역에 관한 국가 단위의 서베이를 세부 주제로 다루고 있다고 볼 수 있다.

보다 세밀한 주제 분야를 파악하기 위해서 12개의 세부 군집으로 나타낸 〈그림 2〉와 〈표 6〉에서 살펴볼 수 있다. 총 12개 군집은 추출 대상이 된 저작물의 제목을 검토하여 정신건강(C1), 담배영향(C2), 학교/유년기/청년기장애(C3), 청년기 성적위험(C4), 아동부상(C5), 육체활동(C6), 폭력행동(C7), 서베이(C8), 가족역할(C9), 여성(C10), 문제행동(C11), 성별차이(C12)로 명명될 수 있다. 우선 가장 중심에 위치한 C4의 청년기 성적위험 군집은 C3의



〈그림 2〉 저작물 제목 키워드 동시출현단어 네트워크(12개 군집)

〈표 6〉 12개 군집명과 단어

군집	군집명	단어	군집	군집명	단어
C1	정신건강	status	C6	육체활동	physical
		mental			activity
		health	C7	폭력행동	partner
C2	담배 영향	study			development
		tobacco			violence
		longitudinal			victimization
		relationship			impact
		effect			police
		child	C8	서베이	national
C3	학교/유년기/ 청년기장애	school			survey
		youth	C9	가족역할	role
		use			india
		adults			evidence
		substance			rural
		disorder			family
		adolescents	C10	여성	early
		social			childhood
C4	청년기 성적위험	adolescence			women
		adulthood	C11	문제행동	outcomes
		young			problem
		risk			behavior
		sexual	C12	성별차이	gender
C5	아동부상	children			differences
		injuries			adolescent
		united			
		states			

학교/유년기/청년기장애 군집과 C10의 여성 군집에 연결되어 있다. C3의 학교/유년기/청년기장애 군집은 C2 담배영향 군집에 직접적으로 연결되었으며, C2 군집의 단어들은 출현빈도가 상대적으로 높은 것을 알 수 있다. C2 담배영향 군집은 C1 군집인 정신건강에 연결되어 있으며, 이 군집 역시 출현빈도의 비중이 높은 것으로 나타났다. 또한 C10의 여성 군집은 왼쪽으로 C11 문제행동 군집과 직접적으로 연결되어 있으며, C11 군집은 C12 성별차이 군집과 연결되어 있다. C10의 여성 군집은 오른쪽으로 C6 군집인 육체활동과 직접적으로 연결되어 있다. 육체활동 군집은 C7인 폭력행동 군집과 연결되어 있으며, 이 군집은 또한 서베이 군집과 연결되어 있다. 마지막으로 서베이 군집인 C8은 가족역할 군집과 연결되어 있다.

5. 논의 및 결론

오픈 사이언스 패러다임과 디지털 정보기술의 발전으로 인해 데이터 중심의 학술 활동은 활발하게 이루어지고 있다. 데이터 중심의 학술활동에 있어서 가장 근간이 되는 것은 데이터의 공유와 재이용을 통한 새로운 시각에서의 연구라고 볼 수 있다. 본 연구는 ICPSR 데이터를 이용하여 학술활동을 수행한 결과물인 2017년도 발간 저작물 570건에 대하여 분석을 수행하였다. ICPSR 리파지토리는 1962년 정치학 분야 데이터 수집에서 시작하였지만, 점차 분야를 넓혀 사회과학 분야의 데이터를 수집하여 연구자가 직접 접근하여 이용할 수 있는 대표적인 리파지토리이며, 활발하게 이용되고 있다.

본 연구는 ICPSR 데이터 이용에 대한 실증적 접근방식의 연구로써 데이터 재이용이 보여주는 지적구조를 규명하고자 하였다. 이를 위해서 첫째, 저자의 특성과 저작물이 게재된 학술지의 주제 분야를 분석하였다. 우선 저자의 소속기관과 학문 분야를 살펴보면, 건강과학 분야에 집중된 것을 찾아볼 수 있다. 또한 학술지 주제 분야를 살펴보면, SCOPUS 전체 27개의 대주제 분야 중에서 사회과학, 의학, 심리학 분야의 학술지가 비중이 가장 큰 것으로 나타났다. 둘째, 보다 미시적으로 살펴보기 위해서 저작물의 제목에서 추출한 단어를 대상으로 한 동시출현단어분석을 수행하였다. 동시출현단어분석은 PNNC 군집과 네트워크로 시각화되었다. 최종 12개의 군집으로 구성되었으며, 정신건강(C1), 담배영향(C2), 학교/유년기/청년기장애(C3), 청년기 성적위험(C4), 아동부상(C5), 육체활동(C6), 폭력행동(C7), 서베이(C8), 가족역할(C9), 여성(C10), 문제행동(C11), 성별차이(C12)이다. 이러한 세부 군집의 주제 분야와 학술지 주제 분야에서 밝혀진 바와 같이, 사회과학, 의학, 심리학의 강세를 찾아볼 수 있으며, 이 중에서도 의학 분야가 특기할 만한 현상이다. 이러한 결과에 대한 해석으로 저작물에 사용된 데이터를 살펴볼 수 있다. ICPSR 홈페이지에서 제시하는 가장 다운로드를 많이 받은 데이터의 상위 8건을 살펴보면(ICPSR, n.d.), National Health and Nutrition Examination Survey(1위), National Longitudinal Study of Adolescent to Adult Health(2위), National Health and Nutrition Examination Survey(3위), Uniform Crime Reporting Program Data(4위), Maternal Lifestyle Study in Four Sites

in the United States(5위), Youth Development Study(6위), Fragile Families and Child Wellbeing Study(7위), National Health and Nutrition Examination Survey(8위)로 나타났다. 이들 데이터의 주제를 보면, 주로 건강, 범죄, 가족, 청소년 관련 분야의 대규모 데이터이다. 또한 DCI를 사용하여 사회학 분야의 데이터 키워드를 분석한 기존의 연구(최형욱, 정은경 2017)에서도 의학 분야의 두드러짐 현상을 유사하게 찾아볼 수 있다. 이러한 결과를 통해 데이터의 재이용은 DDC 학문분류체계나 Web of Science의 Social Science Index의 주제 분야(subject category) 등과 비교하면, 전통적인 사회과학 학문 분야와 상당히 다른 지적구조를 형성한 것으로 볼 수 있다.

본 연구의 결과는 데이터 중심의 학술 활동 패러다임에 있어서 중요한 시사점을 제공할 수 있다. 데이터의 이용 가능성과 데이터가 담고 있는 주제 분야의 활발한 연구가 밀접하게 연계될 수 있다는 점이다. 특정 학문 분야의 풍부한 이용 가능한 데이터는 연구자들에게 여러 관점의 다양한 연구를 수행할 수 있는 기회를 주게 되며, 해당 학문 분야의 성장과 발전으로 이어질 수 있다. 반면에 공개적으로 이용 가능한 데이터가 풍부하지 않은 학문 분야의 상대적인 위축으로도 연결될 수 있다. 따라서 데이터 중심의 학술활동 패러다임에서 균형 있는 학문의 발전과 성장을 위해서는 개별 학문 분야에서 지속적으로 데이터의 생산, 공유, 재이용의 선순환이 이루어질 수 있도록 하는 것이 중요하다고 볼 수 있다. 또한 본 연구의 결과와 관련하여 후속 연구는 크게 두 방향으로 논의할 수 있다. 첫째는 본 연구의 결과에서 보여준 사회과학, 의학, 심리학의 비중을 보다 심도 깊게 탐구하기 위해서 데이터 자체에 대한 분석이 수행되는 것이 바람직하다. 현재의 결과에서는 개별 저작물에서 사용된 데이터가 사회과학과 접목하여 사용된 것으로 볼 수 있다. 그러나 이러한 해석은 제한적이기 때문에 직접 사용된 데이터의 규명과 함께 논의될 필요가 있다. 두 번째 연구의 방향은 대상 범위의 확대를 고려할 수 있다. 이러한 결과는 본 연구의 데이터가 2017년 1년간의 발간된 논문만을 대상으로 하였기 때문에 제한적인 결과일 수 있다. 따라서 연구의 대상이 되는 데이터의 범위를 확장하여 보편적인 결과를 도출하는 것이 바람직하다.

참 고 문 헌

[1] 이재윤. COOC ver 0.4 프로그램 [cited 2018. 1. 15.]

[2] 이재윤. semi.exe 프로그램 [cited 2018. 1. 10.]

[3] 조재인. 2016. Data Citation Index를 기반으로 한 연구데이터 인용에 관한 연구. 『한국문헌정보학회지』, 50(1): 189-207.

[4] 최형욱, 정은경. 2017. 사회학 분야의 연구데이터 특성과 지적구조 규명에 관한 연구. 『정보관리학회지』, 34(3): 109-124.

[5] Borgman, C. L. 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, MA: MIT Press.

[6] Cousijn, H. et al. 2017. "A Data Citation Roadmap for Scientific Publishers." *bioRxiv*. [online] [cited 2018. 1. 16.] 〈https://www.biorxiv.org/content/early/2017/01/19/100784〉

[7] Faniel, I. M., Kriesberg, A. and Yakel, E. 2016. "Social Scientists' Satisfaction with Data Reuse." *Journal of the Association for Information Science and Technology*, 67(6): 1404-1416.

[8] He, L. and Han, Z. 2017. "Do Usage Counts of Scientific Data Make Sense? An Investigation of the Dryad Repository." *Library Hi Tech*, 35(2): 332-342.

[9] ICPSR, n.d. *Inter-University Consortium for Political and Social Research*. [online] [cited 2018. 1. 3.] 〈https://www.icpsr.umich.edu/icpsrweb/〉

[10] Mooney, H. 2011. "Citing Data Sources in the Social Sciences: Do Authors Do It?." *Learned Publishing*, 24(2): 99-108.

[11] Park, H. and Wolfram, D. 2017. "An Examination of Research Data Sharing and Re-use: Implications for Data Citation Practice." *Scientometrics*, 111(1): 443-461.

[12] Piwowar, H. A., Vision, T. J. and Whitlock, M. C. 2011. "Data Archiving is a Good Investment." *Nature*, 473: 285-285.

[13] Silvello, G. 2018. "Theory and Practice of Data Citation." *Journal of the Association for Information Science and Technology*, 69(1): 6-20.

[14] Social Media Research Foundation. *NodeXL Program*. [cited 2018. 1. 10.] 〈https://www.smrfoundation.org/nodexl/〉

[15] Swanberg, S. M. 2017. "Inter-University Consortium for Political and Social Research (ICPSR)." *Journal of the Medical Library Association: JMLA*, 105(1): 106-107.

[16] Yoon, A. 2017. "Data Reusers' Trust Development." *Journal of the Association for Information Science and Technology*, 68(4): 946-956.

[17] Yoon, A. and Kim, Y. 2017. "Social Scientists' Data Reuse Behaviors: Exploring the Roles of Attitudinal Beliefs, Attitudes, Norms, and Data Repositories." *Library & Information Science Research*, 39(3): 224-233.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Lee, Jae-Yun. COOC ver 0.4 Software. [cited 2018. 1. 15.]

[2] Lee, Jae-Yun. semi.exe Software. [cited 2018. 1. 10.]

[3] Cho, Jane. 2016. "Study about Research Data Citation Based on DCI (Data Citation Index)." *Journal of the Korean Library and Information Science Society*, 50(1): 189-207.

[4] Choi, Hyung Wook and Chung, EunKyung. 2017. "An Investigation on Characteristics and Intellectual Structure of Sociology by Analyzing Cited Data." *Journal of the Korean Society for Information Management*, 34(3): 109-124.

[부록 1] 5회 이상 출현한 학술지 목록과 주제 분야

저작물명	주제분야(Scopus)	출현횟수
Policing	Social Sciences: Law Medicine: Pathology and Forensic Medicine	11
Journal of Youth and Adolescence	Social Sciences: Education Psychology: Social Psychology Social Sciences: Social Sciences (miscellaneous) Psychology: Developmental and Educational Psychology	10
Crime and Delinquency	Social Sciences: Law Medicine: Pathology and Forensic Medicine	8
Journal of Quantitative Criminology	Social Sciences: Law Medicine: Pathology and Forensic Medicine	8
Social Science Quarterly	Social Sciences: General Social Sciences	8
Journal of Marriage and Family	Social Sciences: Anthropology Social Sciences: Social Sciences (miscellaneous) Arts and Humanities: Arts and Humanities (miscellaneous)	7
PLOS One		7
Biennial Meeting of the Society for Research in Child Development	Agricultural and Biological Sciences: General Agricultural and Biological Sciences Medicine: General Medicine Biochemistry, Genetics and Molecular Biology: General Biochemistry, Genetics and Molecular Biology	6
Journal of Adolescent Health	Medicine: Pediatrics, Perinatology and Child Health Medicine: Public Health, Environmental and Occupational Health Medicine: Psychiatry and Mental Health	6
Justice Quarterly	Social Sciences: Law Medicine: Pathology and Forensic Medicine	6
Social Science and Medicine	Social Sciences: Health (social science) Arts and Humanities: History and Philosophy of Science	6
American Journal of Public Health	Medicine: Public Health, Environmental and Occupational Health	5
Demography	Social Sciences: Demography	5
Deviant Behavior	Social Sciences: Law Social Sciences: Sociology and Political Science Psychology: Clinical Psychology Psychology: Social Psychology	5
Journal of Criminal Justice	Social Sciences: Law	5
Preventive Medicine	Medicine: Public Health, Environmental and Occupational Health Medicine: Epidemiology	5

[부록 2] 동시출현단어 분석을 위한 제목 키워드

단어	출현 횟수	단어	출현 횟수	단어	출현 횟수
health	63	role	27	school	19
relationship	53	sexual	27	differences	18
use	52	police	26	gender	18
effect	43	united	25	substance	18
family	38	childhood	24	development	17
study	38	risk	24	india	17
adolescent	36	victimization	24	national	17
young	35	adolescents	22	partner	17
social	34	women	22	evidence	16
state	33	adolescence	21	impact	16
children	32	child	21	outcomes	16
adults	30	early	21	physical	16
behavior	30	injury	21	status	16
adulthood	28	tobacco	21	activity	15
violence	28	disorder	19	problem	15
youth	28	longitudinal	19	rural	15
crime	27	mental	19	survey	15