보건복지분야 데이터 연계 필요성 및 활용방안

On the Need for Data Linkage in the Health and Welfare Sectors

오미애 한국보건사회연구원 부연구위원

빅데이터의 3요소는 데이터의 양, 데이터의 증가속도, 데이터의 다양성이며, 이 중 다양성은 제4요소인 새로운 가치를 창출하는데 있어서 매우 중요하다. 다양한 종류의 데이터를 연계하여 새로운 가치를 창출하고 미래를 예측하는 것은 빅데이터 분석의 핵심이며, 정부 3.0 구현을 통해 국가 경쟁력 제고에 기여할 수 있다. 데이터 연계는 데이터의 질을 높이고 효율적인 분석을 가능하게 하며, 새로운 결과를 도출함으로써 다양한 학술적·정책적 함의를 제시할 수 있다. 이 글에서는 데이터 연계에 대한 정의와 데이터 연계 방법론을 살펴보고, 보건복지분야 활용사례를 통해 그 필요성과 활용가치, 활용방안에 대한 함의를 찾아보고자 한다.

1. 들어가며

정부3.0은 공공정보를 개방 · 공유하고, 부처간 칸막이를 없애 소통, 협력함으로써 맞춤형 서비스를 제공하고 일자리 창출과 창조 경제를 지원하는 새로운 정부운영 패러다임을 의미한다. 정부3.0과 빅데이터, 데이터 연계는 서로 유기적인 연관성이 있다. 정부3.0이 제기된 배경에는 IT 환경의 변화로 인한 빅데이터 등장이 있었고, 빅데이터의 활용 부분은 정부 3.0의 추진전략과 밀접한 관련이 있다. 빅데이터의 3요소는 데이터의 양, 데이터의 증가속도, 데이터의 다양성이며, 이 중 다양성은 제4 요소인 새로운 가치를 창출하는 데 있어서 매우 중요하다. 다양한 종류의 데이터를 연계하여 새로운 가치를 창출하고 미래를 예측하는 것은 빅데이터 분석의 핵심이며, 정부 3.0 구현을 통해 국가 경쟁력 제고에 기여할 수 있다. 데이터 연계는 데이터의 질을 높이고 효율적인 분석을 가능하게 하며, 새로운 결과를 도출함으로써 다양한 학술적, 정책적 함의를 제시할 수 있다.

지난해 2월 세상을 떠들썩하게 했던 일명 '송파 세모녀' 사건을 계기로, 복지사각지대에 대한 문제가 지속적으로 제기되고 있다. 이에 따라, 복지 대상자의 효율적 · 상시적 발굴 체계 필요성에 대한 사회적 인식이 확산되고 있으며, 정보시스템과 여러 데이터를 활용한 사각지대 발굴체계 마련의 필요성이 부각되고 있다. 여기에서 데이터 연계를 활용한 사각지대 발굴체계는 기존의 인적 네트워크 활용과 더불어 취약계층을 발굴하고 지원하는 역

할을 할 수 있을 것이다.

이러한 행정데이터의 연계 및 활용은 국정과제로 제시하고 있는 정부3.0을 추진하기 위한 실천과제라고 할 수 있다. 데이터 연계는 행정데이터뿐만 아니라, 조사데이터에서도 적용 가능하다. 이 글에서는 데이터 연계에 대한 정의와 데이터 연계 방법론을 살펴보고, 보건복지분야 활용사례를 통해 그 필요성과 가치, 활용방안에 대한 함의를 찾아보고자 한다.

2. 데이터 연계란[1)]

데이터 연계(data linkage) 또는 데이터 매칭(data matching)이란, 서로 다른 복수의 데이터 파일을 결합하여 보다 풍부한 정보를 제공해 줄 수 있는 하나의 완전한 통합데이터를 만드는 방법으로 정의될 수 있다. 데이터 매칭은 기록(record), 개체 식별(entity resolution), 목표 식별(objective identification), 필드 매칭(field matching), 데이터 통합(data integration), 데이터 퓨젼(data fusion) 등의 개념으로도 사용된다[2)].

이 글에서는 데이터 매칭의 개념을 정확 매칭(exact matching)과 통계적 매칭(statistical matching)으로 나누어서 정의한다.

서로 다른 자료에서 동일한 개체(same individual, family, event or place)를 결합하는 것을 정확 매칭(exact matching) 또는 기록 연계(record linkage)[3)]라고 하며, 서로 다른 자료에서 유사한 개체(similar individual, family, event or place)를 결합하는 것을 통계적 매칭(statistical matching)이라고 한다.

정확 매칭은 고유식별정보가 존재하여 동일한 개체를 연결하므로 행정데이터 간의 연계나 행정데이터와 조사데이터의 연계에 이용될 수 있다. 통계적 매칭의 경우, 고유식별정보가 존재하지 않아 이와 유사한 개체를 찾아 그 데이터를 결합하므로 조사데이터와 조사데이터의 연계, 행정데이터와 조사데이터의 연계(조사데이터에서 개인에 대한 정확한 정보를 얻지 못할 경우)에서 이용될 수 있다.

데이터 통합 연계 방법에서 정확 매칭과 통계적 매칭 두 경우 모두 해당되는 주요 데이터 연계 단계는 <표 1> 같다.

위의 데이터 연계 프로세스를 통한 통합 데이터는 여러 장점이 있는데, 이는 데이터 연계의 필요성과 직결된다. 첫 번째로, 데이터 연계는 추가적인 정보를 제공해준다. 우리가 가지고 있는 데이터에 추가적인 열(column)정보를 더하여 더 많은 정보를 얻을 수 있게 해주어 더 복잡한 연구 문제를 해결할 수 있게 해준다. 두 번째로, 개인 이력을 통해 종단 분석을 가능하게 해 준다. 서로 다른 시간에 서로 다른 유형의 데이터를 매칭하는 이 메커니즘을 통해 시계열 데이터를 생성하는 것도 가능하다. 세 번째로, 데이터 연계는 설문 조사, 행정 자료의 정확성과 신뢰성을 검사하는 방법이 될 수 있

1) 본 내용은 '오미애 외, (2014). 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안, 한국보건사회연구원'의 일부 내용을 발췌하여 수정 · 보완한 것임.

2) Christen, P.(2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer

3) 상황에 따라서는 기록 연계(record linkage)의 범위에 유사한 개체의 결합을 포함시키기도 함.

표 1. 주요 데이터 연계 단계

순서	프로세스
1	명확하게 정의된 목표 세우기 (Develop clearly defined objectives)
2	법률, 정책, 개인정보보호 및 보안 문제 해결 (Address legal, policy, privacy, and security issues)
3	데이터 제공자 및 데이터 이용자들과의 관계 구축 (Establish relationships with data providers and data users)
4	데이터 출처에 대한 철저한 이해 (Gain a thorough understanding of data sources)
5	매칭방법 결정 (Decide how you will do the matching)
6	정보기술 데이터 저장소 구축 및 처리요구 조건 정의 (Define and build information technology (IT) data storage and processing requirements)
7	원시 데이터 얻기 (Obtain the source data)
8	매칭 수행 (Carry out the matching)
9	매칭의 유효성 검증 및 품질 진단 제공 (Validate the matching and provide quality measures)
10	마이크로데이터 접근권 제공 및 공개된 결과물의 비밀성 (개인 신상에 대한 정보를 알지 못하도록 자료 보호)에 대한 고려 (Consider provision of access to microdata and confidentiality of published outputs)
11	분석 수행 및 결과 공유 (Carry out the analysis and disseminate results)

다. 이는 연구자가 분석하고자 하는 데이터 셋에 대한 정보를 연계된 다른 데이터 셋의 정보와 비교해봄으로써, 자료에 대한 신뢰성을 검사할 수 있다. 네 번째로, 데이터 연계는 조사데이터에서 발생하는 결측치에 대한 정보를 제공해 줄 수 있다. 이는 다양한 데이터 연계 방법론으로 적용 가능하다. 마지막으로, 데이터 연계는 조사 데이터의 응답자 부담을 줄여주고, 설문 조사 비용을 감소시켜 준다. 행정데이터와 조사데이터의 연계는 응답자의 응답에만 의존하는 경우보다 정확한 정보를 얻을 수 있으며 이에 따라 데이터의 품질을 유지하는데 들어가는 시간과 비용을 절감할 수 있는 장점이 있다.

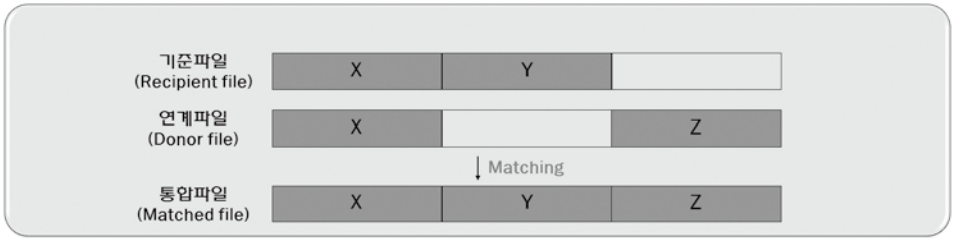
3. 데이터 연계 방법론

데이터 매칭은 서로 다른 데이터 파일을 연계

하여 하나의 통합된 데이터를 생산하는 방법이다. [그림1]에서와 같이 서로 다른 두 개의 데이터가 존재하며 파일 A는 변수 X, Y로, 파일 B는 변수 X, Z로 구성되어 있는 상황에서 파일 A에 파일 B를 연계한다고 가정하면, 파일 A는 기준이 되는 주체파일(host file) 또는 수용파일(recipient file)이라고 하며 이 글에서는 '기준파일'로 정의한다. 파일 B는 통합 연계과정에서 추가적인 정보를 얻기 위해 사용될 파일로 제공파일(donor file)이라고 하며, 여기서는 연계 대상이 되는 데이터라는 의미에서 '연계파일'이라고 정의한다. 파일 A와 파일 B에서 모두 관찰되는 변수 X는 '공통변수(common variable)'이며, 파일 A에서만 관찰되는 변수 Y와 파일 B에서만 관찰되는 변수 Z를 '고유변수(unique variable)'라고 한다. 일반적으로 데이터 매칭은 공통변수를 이용하여 파일 B에 있는 변수 Z를 파일 A에 추가하는 것으로, 공통변수들 중 두 데이터를 연계하는 데 사용하는 핵심변수를 '매칭변수(matching variable)'라고 정의한다.

이렇게 매칭변수로 연계된 하나의 파일을 결합파일(matched file)이라고 하며 여기서는 연계된 결과를 의미하는 '통합파일'이라고 정의한다.

그림 1. 데이터 매칭 개요[4)]



정확 매칭은 고유의 식별자를 이용하여 자료를 연계하기 때문에, 일반적으로 불확실성이 없다. 고유의 식별자는 주민등록번호와 같은 단일변수일 수도 있고, 이름, 생년월일, 성별, 주소 등의 조합으로도 나타낼 수 있는데, 여러 변수의 조합으로 고유 식별자를 이용하는 경우에는 이름 개명, 주소 이전, 입력 오류, 데이터의 생성시기에 따른 차이(나이, 교육수준), 결측 항목 등의 사유로 동일인의 정보가 두 데이터에 다르게 입력되어 있을 수 있다. 이러한 불확실성을 최소화시키며 신뢰할 수 있는 통합데이터를 만들기 위해서는 공통변수들에 오류가 있는 경우에 정확한 정도를 고려한 가중치를 부여하고 확률적으로 데이터를 결합하는 방법을 이용하여 데이터 연계절차를 실시할 수 있다.

개별 단위의 기록을 레코드(record)라고 할 때, 서로 다른 두 데이터의 연계결과는 1종 오류와 2

4) 이영섭 외 (2009) 통계조사자료와 행정자료 간의 통계적 매칭기법에 관한 연구, 통계 연구, 제14권 제1호, 82-98 참고하여 수정 보완한 것임.

종 오류로 표현 가능하다.

서로 다른 두 데이터의 레코드 쌍이 데이터 연계과정을 통해 연계(link)되었다면 이는 같은 개체(사람/사업/사건 등)로 인식한다. 레코드 쌍이 실제 동일한 개체(match)일 때 정확하게 연계(link)되었다면 옳게 맞춘 레코드 쌍(true positives)에 속하며, 레코드 쌍이 실제 다른 개체(non-match)일 때 서로 연계되지 않았다면(non-link) 이 역시 옳게

표 2. 정확 매칭의 데이터 연계 오류

연계 \ 실제	True Match	True Non-Match
Link	True Positive	False Positive Link (1종 오류)
Non-link	False Negative Link (2종 오류)	True Negative

자료: 오미애 외 (2014). 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안, 한국보건사회연구원.

맞춘 레코드 쌍(true negative)에 속한다. 데이터 연계 시 발생되는 오류는 1종 오류(type Ⅰ error)와 2종 오류(type Ⅱ error)가 있다. 1종 오류는 레코드 쌍이 실제 다른 개체(non-match)일 때 같은 개체로 연계된 레코드 쌍(false positive)을 의미하며, 2종 오류는 실제 동일한 개체(match)일 때 부정확하게 연계된 레코드 쌍(false negative)들을 의미한다. 데이터 매칭의 최종적 과정은 성과측정(performance measurement)으로, 데이터 매칭의 질을 측정하는 단계에서 위의 1종 오류와 2종 오류를 계산할 수 있다.

서로 다른 두 데이터에 대한 연계과정은 우선, 데이터 표준화(standardisation)로 시작한다. 이는 통합된 데이터의 질과 관계있는 중요한 사전작업이다. 여기에는 데이터클리닝 작업도 포함되어 있으며 데이터 사이의 형식을 맞추고, 정보의 표현과 인코딩을 동일하게 하는 과정이다. 그 다음 과정은 블로킹 기법(blocking)으로 대량의 레코드 쌍 비교 수를 줄이기 위해 사용되는 방법이다. 하나 또는 둘 이상의 레코드 속성의 조합으로 데이터베이스를 몇 개의 블록으로 나눈다. 레코드 속성의 조합 내에서 같은 값을 갖는 레코드들은 동일한 블록에 속하며 후보 레코드 쌍들은 다양한 방법들로 비교(comparison)가 이루어진다. 다음으로, 결정 모델(decision model)은 레코드 쌍들이 매치(match), 비매치(non-match), 매치 가능(possible match) 중에 어디에 속하는지 결정해주는 단계로 이 과정에 매칭 가중치 벡터를 이용하여 분류한다. 레코드 쌍이 매치 되었을 때, 연계된 레코드 쌍이 실제로 동일한 개체라면 옳게 분류된 상황이지만 연계된 레코드 쌍이 실제 다른 개체라면 1종 오류가 발생한다. 마찬가지로, 레코드 쌍이 비매치되었을 때, 연계되지 않은 레코드 쌍이 실제로 다른 개체라면 옳게 분류된 상황이고 실제로 동일한 개체라면 2종 오류가 발생한다. 매칭 가능한 레코드는 블로킹 단계로 돌아가서 이 과정을 반복한다. 연계된 통합

데이터의 질에 대한 측정은 1종 오류와 2종 오류로 나타낼 수 있으며 오류를 최소화하기 위해서는 결정 모델의 성능이 중요하다.

통계적 매칭은 둘 이상의 데이터 결합에서 고유식별정보가 없을 때 사용할 수 있는 방법으로, 유사한 개체와의 결합으로 새로운 데이터를 제공하는데 그 목적이 있으며, 무응답 대체(imputation) 방법론을 이용할 수 있다. 통계적 매칭에 필요한 기본가정은 다음과 같다.

우선, 연계하고자 하는 여러 가지 데이터의 표본은 동일한 모집단에서 조사되었다고 가정한다. Rässler(2002)[5]에 기술된 매칭 룰에 따르면, 두 데이터의 샘플은 공통변수들의 평균에서 유의하게 차이가 나서는 안 된다. 두 번째 가정은, 공통변수인 X(예를 들어, 연령)가 주어져 있을 때, 두 자료의 고유변수 Y(예를 들어, 소득)와 Z(예를 들어, 부채)는 조건부 독립(conditional independence)을 만족해야 한다. 조건부 독립에 대한 가정이 필요한 이유는 기준파일과 연계파일을 결합한 (X,Y, Z)의 결합확률분포인 $\hat{f}_{X,Y,Z}(x,y,z)$를 추정하기 위해서이다.

$$\begin{aligned}\hat{f}_{X,Y,Z}(x,y,z) &= f_{Z,X}(z,x)\, f_{Y|X}(y|x) \\ &= f_{Z|X}(z|x)\, f_X(x)\, f_{Y|X}(y|x) \\ &= f_{Z|X}(z|x)\, f_{Y,X}(y,x)\end{aligned}$$

조건부 매칭 분포는 다음의 수식과 같이 표현된다.

$$\hat{f}_{Z,Y|X}(z,y|x) = f_{Z,X}(z|x)\, f_{Y|X}(y|x)$$

1980년대에 연구된 통계적 매칭의 고전적인 방법은 두 가지로 비제한적 매칭방법(unconstrained matching)과 제한적 매칭방법(constrained matching)이 있다(Rodgers,1984[6]). 비제한적 매칭은 두 데이터를 연계하는 데에 아무런 제한을 두지 않는 방법이다. 따라서 비제한적 매칭은 연계된 파일에서 Z^B의 평균과 표준편차가 달라진다. 비제한적 매칭은 파일 B에서 중복을 허용하여 A 레코드(record)와 가장 가까운 값을 찾는다는 장점이 있지만, Z변수와 관련된 추정량의 샘플 분산을 크게 하는 단점이 있다. 이러한 단점에도 불구하고 비제한적 매칭방법은 많이 쓰인다.

제한적 매칭방법은 두 자료를 연계할 때 B 레코드(record)를 모두 사용하여 연계하는 방법으로 연계된 결합파일의 가중치는 다음과 같은 제약조건을 갖는다.

$$\sum_{j=1}^{m} w_{ij} = w_i \ (for\ i = 1,\cdots,n),\ \sum_{i=1}^{n} w_{ij} = w_j \ (for\ j = 1,\cdots,m)$$

w_i는 파일 A(기준 파일)의 가중치를 의미하고, w_j는 파일 B(연계 파일) 가중치를 의미하며, w_{ij}는 두 자료를 연계한 결합파일의 조정된 가중치를 의미한다.

제한적 매칭방법은 마이크로데이터의 가중치를 이용하여 제약조건을 정의하고 있는데, 기본적으로 데이터 연계를 위한 마이크로데이터들은 동일

5) Rässler R. (2002). Statistical Matching : a Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches, Springer.

6) Rodgers, W.L. (1984) An evaluation of statistical matching. Journal of Business and Economic Statistics 2, 91-102.

한 모집단으로부터 조사된 자료이므로 두 조사의 가중치 합은 동일해야 함을 내포하고 있다[7]. 제한적 매칭방법에서는 거리 함수(d_{ij})와 제약조건을 이용한 목적함수를 다음과 같이 정의하고 있으며, 목적함수를 최소로 하기 위해 결합파일에서 가중치 w_{ij}를 재조정한다[8].

$$\min\left\{\sum_{i=1}^{n}\sum_{j=1}^{m}(d_{ij}\times w_{ij})\right\}$$

제한적 매칭방법은 파일 B의 모든 샘플이 다 사용되고, 결합파일에서 Z^B분포가 파일 B와 동일하게 유지된다는 장점이 있는 반면에, 연결과정에서 파일 A의 레코드(record)에 가장 가까운 값이 연계되지 않을 수 있고, 제약조건의 목적함수를 최소화하기 위해 computation time이 길어지는 단점이 있다.

4. 보건복지분야 데이터 연계 활용사례

데이터 연계 유형은 기준파일과 연계파일의 데이터 특성(조사데이터 vs. 행정데이터) 및 매칭 방법(정확 매칭 vs. 통계적 매칭)에 따라 <표 3>과 같이 8가지로 구분할 수 있다.

여기에서는 위의 유형에 따른 여러 데이터 연계 활용사례를 살펴봄으로써 데이터 연계의 활용가치를 알아보고자 한다.

1) 조사데이터와 행정데이터의 연계

한국의료패널은 한국보건사회연구원에서 2008년도부터 건강보험공단과 컨소시엄으로 시작된 조사로, 개인의 의료이용 결정요인(사회경제적요

표 3. 데이터 특성 및 매칭 방법에 따른 8가지 유형

매칭 방법	데이터 특성	
	기준파일	연계파일
Exact Matching	행정데이터	행정데이터
	행정데이터	조사데이터
	조사데이터	행정데이터
	조사데이터	조사데이터
Statistical Matching	행정데이터	행정데이터
	행정데이터	조사데이터
	조사데이터	행정데이터
	조사데이터	조사데이터

자료: 오미애 외 (2014). 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안, 한국보건사회연구원.

7) Paass, G. (1985) Statistical record linkage methodology: state of the art and future prospects. In Bulletin of the International Statistical Institute, Proceedings of the 45th Session, Vol. LI, Book 2. Voorburg, Netherlands: ISI.

8) Barr, R.S. and Turner, J.S. (1981). Microdata file merging through large-scale network technology. Mathematical Programming Study, Volume 15, pp. 1-22.

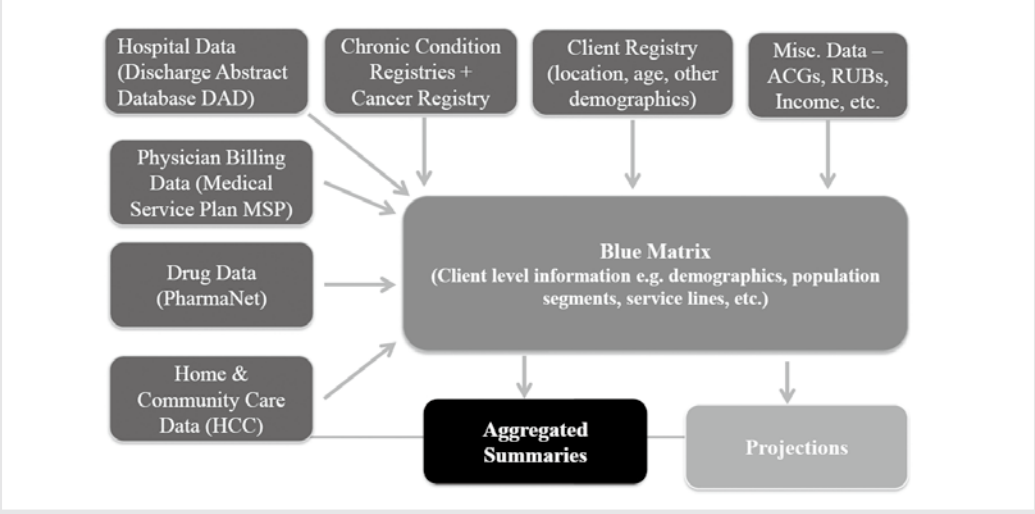
인, 건강의식, 건강행태 등)과 보건의료서비스 이용, 의료비 지출/재원을 종합적으로 살펴볼 수 있는 보건의료 관련 패널조사이다. 조사설계 당시, 건강보험공단 행정DB 연계를 염두에 두고, 고유식별정보를 수집한 후, 패널조사 내용을 보완하여 정확한 의료이용 내역을 파악하기 위해 건강보험DB를 연계 · 활용하였다.

한국의료패널은 조사연도 1년 전 의료이용에 대한 회상자료이며, 비급여부분에 대한 자료수집이 가능하다. 한편, 건강보험공단은 의료기관을 이용한 환자의 급여 신청자료이므로 정확한 의료이용 내역이 파악 가능하지만 비급여에 대한 정보는 알 수 없다. 따라서 2013년도에 한국의료패널과 건강보험공단의 자료연계 작업을 실시하여 급여부분에서의 회상자료에 대한 부정확성을 행정자료로 보완하고자 하였다. 한국의료패널조사가 기준파일이고 이에 대한 보완으로 건강보험공단의 자료를 사용했으므로 이는 정확 매칭에 의한 '조사데이터(기준파일) + 행정데이터(연계파일)' 연계의 사례에 속한다고 할 수 있다.

2) 행정데이터와 행정데이터의 연계

정확 매칭을 이용한 대표적인 행정데이터 간 연계사례는 Blue Matrix이다. 캐나다 BC 보건부의 Health system planning and Innovation division에서 행정자료들의 연계를 통해 보건정책의 실효성을 높이고자 Blue matrix DB를 만들었다. Blue matrix 프로젝트는 하나의 큰 데이터베이스에 모든 정보를 담고자 하는 것이 아니라, 각각의 행정자료에서 주요 정보만을 가져와서 서로 다른 행정자료의 연계로 종합적이면서 세밀한 BC Health care system을 구축함으로써 BC 국민의 건강에

그림 2. Blue Matrix 데이터 흐름도



자료: 2012/13 Blue Matrix 5.0

대한 이해를 돕고 의료계획 및 치료 · 예방 개선 관련 정책 수립의 근거확립에 대한 기초자료를 제공하는데 그 목적이 있다.

환자(Patient)중심의 DB인 Blue matrix는 건강상태를 반영한 인구집단구분에 따른 건강 욕구, 건강 조건, 사람들이 사용하는 모든 서비스를 한눈에 살펴 볼 수 있다. Blue matrix를 만들기 위한 행정자료는 Physician billings, Hospitalizations, Prescription drugs, Inter-Rai assessments, Home & Integrated Care programs, Emergency Department visits, Eligibility for program 등이 있으며 데이터 연계는 고유식별정보(unique client identifier)를 이용한 정확매칭방법을 활용하였다. 데이터의 생성시기를 기준회계년도로 정하여 연계 데이터 간 정보의 일치성을 확보하였다.

Blue Matrix는 건강상태, 인구학적 정보, 의료서비스의 소비를 고려하여 인구집단을 구분하였다. 인구집단은 보건부에서의 업무 정의를 바탕으로 개인별 건강상태를 반영하여 나누었으며 복합질환을 가진 환자의 경우 가장 우선되는 질환으로 분류하였다.

보건부의 행정데이터 정보를 기본으로 위에서 분류한 인구집단별 보건 예산을 추정한 결과는 [그림 3]과 같다. BC 인구의 34%가 건강한 편이며, 29% 정도가 경증의 복합만성질환이 있는 것으로 나타났다. 중증의 복합만성질환자에게 전체 보건예산의 20%가 지원되며, 장기요양도 21%의 비중을 차지한다. BC 인구 1인당 지원되는 보건예산은 2,210$ 로, 건강한 집단과 경증의 복합만성질환이 있는 집단을 제외하고는 1인당 평균 지원예산을 넘는 것을 알 수 있다. 이러한 유의미한 분석은 BC 인구의 건강상태를 파악하고, 정책적으로

그림 3. Blue Matrix 인구집단별 health care 예산별 분포

BC Age Groups: All Health System Matrix 5.0 Estimate of the distribution of $10.5 Billion of Selected Publicly Funded Health Services used by BC Residents in 2012/13 Population Segments Assigned to their highest health care need in the year	People	Estimate of Publicly Funded Health Care, 2012/13 (Million of Dollars)							
		Primary Care (GP and Professional Home Nursing)	Physicians' Specialty Care, Surgery, Labs and Diagnostics	PharmaCare	Hospital Care (Inpatient & Day Surgery, not including Physicians)	Emergency Department Physican and Hospital Costs	Supports for Daily Living	Residential Care	Total Cost (Millions)
PS01 Non User	13.8%								
PS02 Healthy	34.2%	1.4%	1.8%	0.1%	0.0%	1.5%			4.8%
PS03 Adult Major Age 18+	2.4%	0.2%	1.1%	0.4%	1.9%	0.3%			4.0%
PS04 Child and Youth Major <18 years	0.9%	0.1%	0.4%	0.0%	1.6%	0.2%			2.3%
PS05 Low Complex Chronic Conditions	29.0%	2.2%	4.9%	1.8%	4.1%	1.7%			14.8%
PS06 Medium Complex Chronic Conditions	8.6%	1.5%	3.1%	2.0%	4.4%	0.8%			11.7%
PS07 Mental Health and Substance Use	1.5%	0.2%	1.0%	1.2%	2.0%	0.3%			4.7%
PS08 Maternity and Healthy Newborns	2.3%	0.3%	1.4%	0.0%	1.8%	0.3%			3.9%
PS09 Frail In The Community	0.3%	0.2%	0.2%	0.3%	1.0%	0.1%	1.4%	0.1%	3.1%
PS10 High Complex Chronic Conditions	4.7%	1.6%	3.1%	2.4%	9.8%	0.9%	2.2%	0.1%	20.1%
PS11 Cancer	1.2%	0.3%	1.0%	0.3%	2.9%	0.2%	0.1%	0.0%	4.8%
PS12 Frail In Care (In Residential Care)	0.8%	0.3%	0.4%	0.4%	3.7%	0.2%	0.4%	15.9%	21.2%
PS13 End Of Life	0.3%	0.3%	0.4%	0.3%	2.8%	0.2%	0.3%	0.5%	4.7%
All Population Segments	100.0%	8.4%	18.8%	9.2%	36.0%	6.7%	4.4%	16.5%	$10,500

자료: 2012/13 Blue Matrix 5.0

어느 인구집단에 예산을 더 투입해야 하는지에 대한 근거로 활용된다.

3) 조사데이터와 조사데이터의 연계

조사데이터와 조사데이터의 연계는 대부분 개인식별정보가 없기 때문에 통계적 매칭 방법을 통하여 통합 데이터를 구축한다. 조사데이터 간 연계 사례로, 한국복지패널과 재정패널을 들 수 있다[9]. 한국복지패널은 한국보건사회연구원과 서울대학교 사회복지연구소가 2006년부터 공동 수행하는 조사로 7,000여 가구를 대상으로 패널을 구축하였으며, 표본추출 시 중위소득 60% 미만 저소득층에 전체 표본의 50%를 할당하였다. 한국조세재정연구원이 조사 · 구축하는 재정패널조사는 우리나라의 조세 및 재정정책 분석 및 연구에 활용할 수 있도록 기획된 패널조사이며, 2008년부터 5,000여 가구를 대상으로 패널을 구축하였다. 두 데이터는 연구목적에 맞게 특정 집단을 과대표집하였기 때문에 두 데이터가 완전히 동일한 분포를 갖고 있지는 않지만, 모집단은 인구주택총조사 자료로 동일하다.

이 연구에서는 통계적 매칭 방법을 이용하여 한국복지패널과 재정패널을 연계한 통합 데이터를 구축하고 납세 및 복지수급 여부에 따른 복지인식을 비교 분석함으로써 조세 및 복지정책 마련을 위한 근거자료를 제공하고자 하였다. 연구의 분석 결과, 납세 및 복지수급 여부를 기준으로 구분한 4개 집단 중에, 면세점 이상에 해당하여 실질적으로 납세를 하면서 동시에 최근 복지 확대로 인하여 복지혜택을 받는 집단의 경우, 납세는 하지만 복지혜택을 받지 못하고 있는 집단에 비해 보편적 복지의 확대를 선호하는 경향이 상대적으로 높게 나타났다. 또한, 이처럼 납세자와 복지수급자의 이중적 지위를 가지고 있는 집단에서 현재 우리나라의 중간계층이 부담하고 있는 세금의 수준이 낮다는 인식이 상대적으로 높았으며, 복지 확대를 위한 증세 필요성에 대한 부정적인 인식 역시 실질적으로 납세를 하면서도 복지혜택을 받지 못하는 집단에 비해 낮은 것으로 분석되었다.

5. 데이터 연계 활용방안

데이터 연계는 동일한 개체에 대한 서로 다른 데이터를 결합함으로써 다양한 정보를 얻을 수 있고, 새로운 조사를 실시하는 것보다는 비용이 훨씬 적게 든다는 데 장점이 있다. 한번 데이터가 결합되면 프로젝트를 수행하는 데 걸리는 시간을 줄일 수 있고, 기초 데이터 수집 시간이 크게 감소한다. 데이터 연계는 학계, 산업, 정부 기관뿐만 아니라 지역사회 연구 분야에서 활용될 수 있으며 실제로 모니터링, 감시, 분석 평가 등에 유용하게 쓰이고 있다. 또한, 현재의 데이터가 미래에도 쓰일 수 있다는 생각을 사람들에게 심어줌으로써, 데이터의 품질을 향상시킬 수 있으며 데이터의 완결성을 높일 수 있다. 이러한 데이터 연계 활용가치를 증대시키기 위해 뉴질랜드 통계청에서 발간한 "Data Integration Manual"에서는 데이터 연계 시 고려해야할 12가지 원칙을 규정하고 있는데 이는 다음과 같다.

9) 최현수, 오미애 (2015), 데이터 연계 방법론을 활용한 납세 및 복지수급 여부에 따른 복지인식 비교 분석, vol 17, No.4, 자료분석학회지

- 데이터 연계가 새로운 통계를 생성하거나 기존 통계의 질을 향상시킬 수 있는 경우에만 수행한다.
- 데이터를 연계해서 경비를 줄이거나 질을 높이거나 규제를 줄이는 등의 가치가 있는 경우에만 데이터 연계를 고려한다.
- 데이터연계의 이득이 개인정보보호의 위험에 비하여 월등히 높은지를 고려한다.
- 원자료의 순수성을 해치지 않는 범위에서 데이터 연계를 고려한다.
- 데이터 제공자가 데이터 연계를 반대하는 경우 연계하지 않는다.
- 데이터 연계는 반드시 데이터를 제공한 모든 기관은 동의를 얻는다.
- 연계된 데이터는 공식 통계나 관련 연구에만 사용한다.
- 필요 이상으로 데이터나 변수를 연계하지 않는다.
- 연계 데이터는 다른 데이터와는 따로 관리한다.
- 이름이나 주소 등의 개인식별자는 데이터 연계 시에만 사용하고 파기한다.
- 연계된 데이터에는 외부기관에서 사용하고 있는 개인식별자(예: 아이디 등)등을 포함하지 않는다.
- 데이터 연계는 공개적으로 진행한다.

데이터 연계 활용방안으로는 위의 12가지 원칙을 바탕으로, 개인정보보호문제, 데이터의 신뢰성 등에 대한 품질관리 방안, 변화될 미래에도 지속적으로 사용할 수 있도록 자료를 연계하고 제공하는 안전한 통계시스템 구축에 대한 적극적인 노력이 필요하다. 그리고, 무엇보다 중요한 것은 데이터 연계에 대한 필요성을 공감하는 사회적 분위기가 뒷받침 되어야 한다.

데이터 연계는 정책을 수행할 때 기반이 되는 중요한 증거를 생성할 수 있고, 더 나아가 사회현상을 여러 가지 면에서 입체적으로 파악할 수 있는 중요한 정보를 제공할 수 있는 강력한 도구이다. 이러한 데이터 연계 작업이 많은 부문에서 체계적이고 활발하게 이루어지려면 통계 자료나 정보를 많이 보유한 국가기관이 선도적인 역할을 수행하고 성공 사례를 만들어 기술과 방법을 민간부문에 전파해야 한다.

작년 2월의 '송파 세모녀' 사건과 관련하여 사회보장정보시스템과 여러 데이터를 활용한 사각지대 발굴체계 마련을 위하여 「사회보장급여의 이용 · 제공 및 수급권자 발굴에 관한 법률」이 올해 7월부터 시행되었다. 이 법은 「사회보장기본법」에 따른 사회보장급여의 이용 및 제공에 관한 기준과 절차 등 기본적 사항을 규정하고 지원을 받지 못하는 지원대상자를 발굴하여 지원함으로써 사회보장급여를 필요로 하는 사람의 인간다운 생활을 할 권리를 최대한 보장하고, 사회보장급여가 공정하고 효과적으로 제공되도록 하며, 사회보장제도가 지역사회에서 통합적으로 시행될 수 있도록 그 기반을 구축하는 것을 목적으로 한다. 이 법률에 의해 단전, 단가스, 단수 가구, 학교생활기록 정보 중 담당교원이 위기상황에 처하여 있다고 판단한 학생의 가구정보, 보험료를 6개월 이상 체납한 사람의 가구정보 등 그 밖에 지원대상자의 발굴을 위하여 필요한 정보로서 대통령령으로 정하는 정보를 연계하여 취약계층 사각지대 발굴체계를 구축할

수 있는 근거가 마련되었다. 복지 사각지대에 놓인 지원대상자들의 타 기관 정보를 사회보장시스템과 연계하기 위해서는 타 기관과 충분한 협의를 한 뒤에 연계 가능한 정보범위를 정하고, 연계정보의 레이아웃을 확정하며, 어떤 형태로 연계정보를 제공받을 수 있는지 등 연계기반 조건에 대한 충분한 검토도 필요할 것이다. 이를 잘 활용한다면 보건복지분야에서 선도적인 데이터 연계의 성공적인 사례가 될 것이다.

국가기관이 선제적으로 자료를 결합하고 이용하여 좋은 정보를 제공하면 국민들이 정보의 생산과 활용에 대한 중요성을 인식하여 데이터 연계의 필요성을 공감하는 사회적 분위기가 조성될 것이며, 각 분야에서 유용한 정보가 생산되고 국민 경제발전, 복지 증진 등 긍정적인 효과가 나타날 수 있을 것으로 기대한다.