보건복지 분야 데이터 통합 연계방안에 대한 고찰

A Study on Data Integration Method for Health and Welfare

오미애 한국보건사회연구원 부연구위원

현재 수많은 분야에서 다양한 조사가 이루어지고 있고 행정자료가 구축되고 있지만 하나의 자료만으로 분석에 필요한 모든 정보를 얻는다는 것은 쉽지 않은 일이다. 단일 자료로부터 얻을 수 있는 정보의 한계는 데이터 매칭(data matching) 방법을 통해 어느 정도 해결가능하며, 기존 통계간의 통합 연계는 데이터의 활용도를 높여줄 뿐만 아니라 별도의 다른 조사를 통해 데이터를 얻는 것보다 시간과 비용을 크게 절감할 수 있다는 점에서 의의가 있다. 이 글에서는 먼저 다양한 데이터 매칭 기법들을 살펴보고 보건복지 연계통계 활용에 대한 호주의 사례를 통해 우리나라 보건복지 데이터 연계방안에 대한 함의를 찾아보고자 한다.

1. 들어가며

현재 수많은 분야에서 전화조사나 면접조사와 같은 설문조사가 이루어지고 있으며 다양한 행정자료가 구축되고 있다. 이렇게 구축된 데이터는 학문적 연구 수행과 더불어 정책수립 및 평가를 위해 다양하게 활용되고 있다. 그러나 이러한 자료들 가운데 어느 하나의 자료만으로 분석에 필요한 모든 정보를 얻는다는 것은 쉽지 않은 일이다. 실제로 통계청의 마이크로데이터 이용에 관한 수요자 설문조사 결과에 따르면, 마이크로데이터 이용 시 가장 어려운 점은 상세한 정보의 부족과 필요한 변수의 부재, 데이터 가공의 어려움, 이용 자료의 부족 등이라고 응답하였으며, 이를 해결하기 위해 담당자 문의, 보조정보 활용, 여타 데이터 연계 및 예측 값 적용을 시도하였다고 응답하였다. 또한, 단일 통계자료의 정보 부족을 해결하기 위해서 통계청이 제공하는 데이터 간 연계를 시도해보았다고 응답한 비율은 44.2%로 높게 나타났다[1]. 즉, 통계청 마이크로데이터 이용자들이 단일 데이터의 제한된 정보를 극복하고자 여러 가지 방법을 활용하고 있다는 것을 알 수 있으며 이와 같은 상황은 보건복지 분야의 데이터 이용자도 마찬가지일 것이다.

데이터 매칭(data matching)은 단일 자료로부

1) 심규호, 박시내(2010). **통계이용 활성화를 위한 2차 자료 생산 · 활용 방안 연구**, 통계청.

터 얻을 수 있는 정보의 한계를 어느 정도 해결할 수 있는 중요한 방법들 중 하나라고 할 수 있다. 데이터 매칭(data matching)이란, 서로 다른 복수의 데이터 파일을 결합하여 보다 풍부한 정보를 제공해 줄 수 있는 하나의 완전한 데이터를 만드는 방법으로 정의될 수 있는데, 이러한 데이터 매칭(data matching)은 레코드 연계(record linkage), 데이터 연계(data linkage), 데이터 통합(data integration)등과 유사한 의미로도 활용된다[2]. 일반적으로 조사 · 구축되는 데이터에는 가구 및 개인 식별번호, 성별, 나이 등 공통적으로 포함되는 항목들이 있다. 이와 같은 공통 항목들을 통해서 다수의 데이터를 통합하면 어떠한 경우는 완전히 같은 사람의 정보를 추가적으로 얻을 수 있으며, 완전히 동일하지 않지만 유사한 사람들이나 집단들로부터 추가 정보를 얻음으로써 데이터의 유용성을 훨씬 높일 수 있다.

기존 통계간의 연계는 데이터 활용도를 높여줄 뿐만 아니라 다른 조사를 통해 데이터를 얻는 것보다 시간과 비용을 크게 절감할 수 있고, 응답자의 부담도 줄일 수 있다는 점에서 의의를 갖는다. 데이터 매칭은 이처럼 다양한 단일 자료로부터 얻을 수 있는 정보의 한계를 해결할 수 있는 방안 중 하나이다. 이 글에서는 먼저 데이터 매칭 기법들을 살펴보고 보건복지 연계통계 활용에 대한 호주의 사례를 통해 우리나라 보건복지 데이터 연계방안에 대한 함의를 찾아보고자 한다.

2. 데이터 매칭 방법론

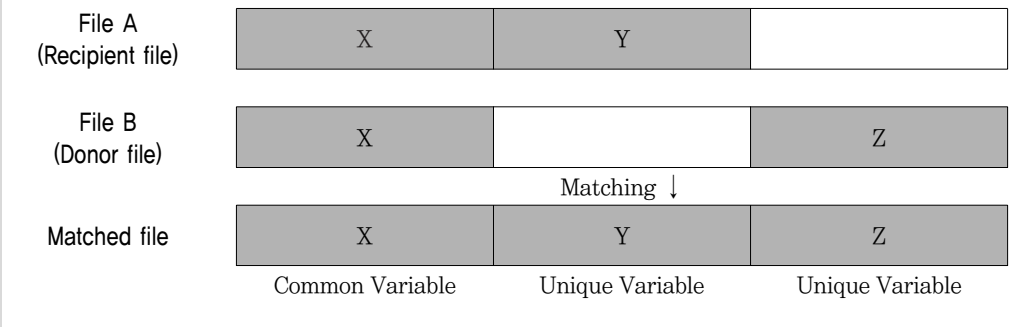
데이터 매칭은 서로 다른 데이터 파일을 결합하여 하나의 데이터 파일을 만드는 방법으로 [그림 1]에서와 같이 서로 다른 경로로 얻어진 두 개의 파일이 있다고 할 때, 파일 A는 변수 X, Y로 구성되어 있고 파일 B는 변수 Y, Z로 구성되어 있다고 가정하자. 파일 A는 기존에 갖고 있는 파일로 기존파일(host file) 또는 수용파일(recipient file)이라고 한다. 파일 B는 통합과정에서 추가적인 정보를 제공하기 위해 사용될 파일로 제공파일(donor file)이라고 한다. 파일 A와 파일 B에 모두 관찰되는 변수 X를 공통변수(common variable)라 하고 파일 A에서만 관찰되는 변수 Y와 파일 B에서만 관찰되는 변수 Z를 유일변수(unique variable)라고 한다. 일반적으로 데이터 매칭을 수행하면 공통변수를 이용하여 파일 B에 있는 변수 Z를 파일 A에 추가하게 되고, 생성된 파일을 결합파일(matched file)이라 한다.

영국의 “National Statistics code of Practice Protocol on Data Matching(2003)”에 의하면 데이터 매칭의 유형을 크게 5가지로 분류하고 있는데, 이것은 정확 매칭(Exact Matching), 판단 매칭(Judgemental Matching), 확률적 매칭(Probability Matching), 통계적 매칭(Statistical Matching), 데이터 연결(Data Linking)로 구분할 수 있다.

먼저 정확 매칭(Exact Matching)부터 살펴보

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

그림 1. 데이터 매칭



자료: 이영섭 외 (2009). 통계조사자료와 행정자료 간의 통계적 매칭기법에 관한 연구, 통계 연구, 14(1), pp.82~98.

면, 주민등록번호처럼 ID를 나타낼 수 있는 변수가 자료에 공통으로 있고 변수의 값이 완전히 일치하는 경우에 데이터를 결합하는 방법이다. 공통인 변수에 측정오차가 없다는 전제 하에, 동일한 개체에 대한 정보들을 정확하게 얻을 수 있고 이상적으로 데이터 매칭을 할 수 있다는 장점이 있다. 하지만 사람과 관련된 자료의 경우에는 개인의 사생활(privacy)과 비밀(confidentiality)의 법적 제도적 규제로 실제 데이터를 이용 할 경우 정확 매칭을 수행하는 데 한계가 있을 수 있다.

다음으로, 판단 매칭(Judgemental Matching)은 공통인 변수들 사이에 정확하게 일치하는 주요 변수는 없지만 데이터에 대해 잘 알고 있는 경우 또는 컴퓨터의 도움으로 적절하다고 판단되는 케이스를 결합하는 방법을 의미한다.

한편, 확률적 매칭(Probability Matching)은 정확 매칭의 경우에서 공통변수들에 오류가 있는 경우 정확한 정도에 따라서 가중치를 부여하고 확률적으로 데이터를 결합하는 방법을 의미하며, 통계적 매칭(Statistical Matching)은 공통인 변수에 개인별로 식별 가능한 변수가 없을 때 수행하는 방법이다.

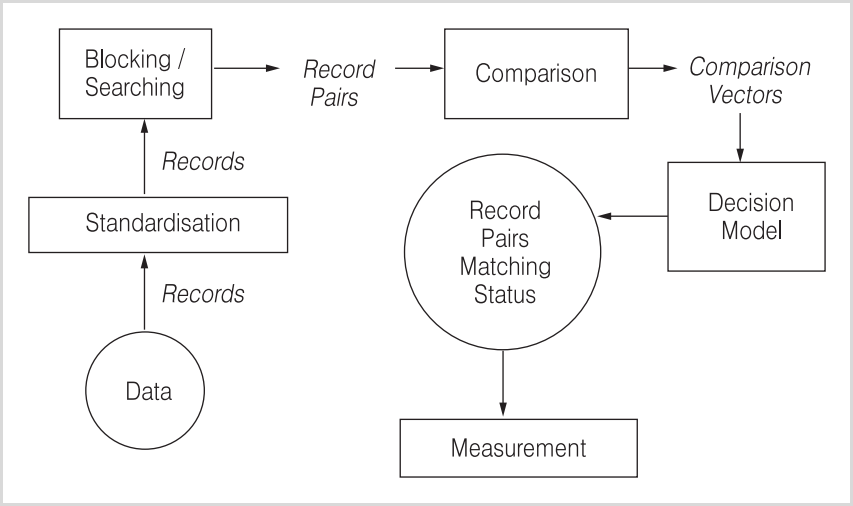
마지막으로, 데이터 연결(Data Linking)은 앞서 언급한 방법과는 조금 다르게, 둘 이상의 데이터 파일에서 변수들 간의 연관성을 만들어내는 것과 관련이 있다. 이 연관성은 또 다른 데이터 파일과도 연결 가능하게 하며 동시에 자료를 업데이트하는데 쓰일 수 있다.

데이터 매칭을 적절하게 수행하기 위해서는 위의 다양한 매칭 기법을 혼용하여 수행할 수 있다. 데이터 매칭 시스템에는 여러 가지가 있지만 여기서는 TAILOR 시스템[3]을 중심으로 데이터 매칭의 전반적인 과정을 간략하게 설명하고자 한다.

우선 데이터 표준화(standardisation)는 통합된 데이터의 질과 관계있는 중요한 사전작업이다.

3) Elfeky, M. G., Verykios, V. S., & Elmagarmid, A. K.(2002). TAILOR: A record linkage toolbox. In Data Engineering, 2002. Proceedings. 18th International Conference on (pp.17~28). IEEE.

그림 2. 데이터 매칭(data matching) 시스템의 정보 흐름도



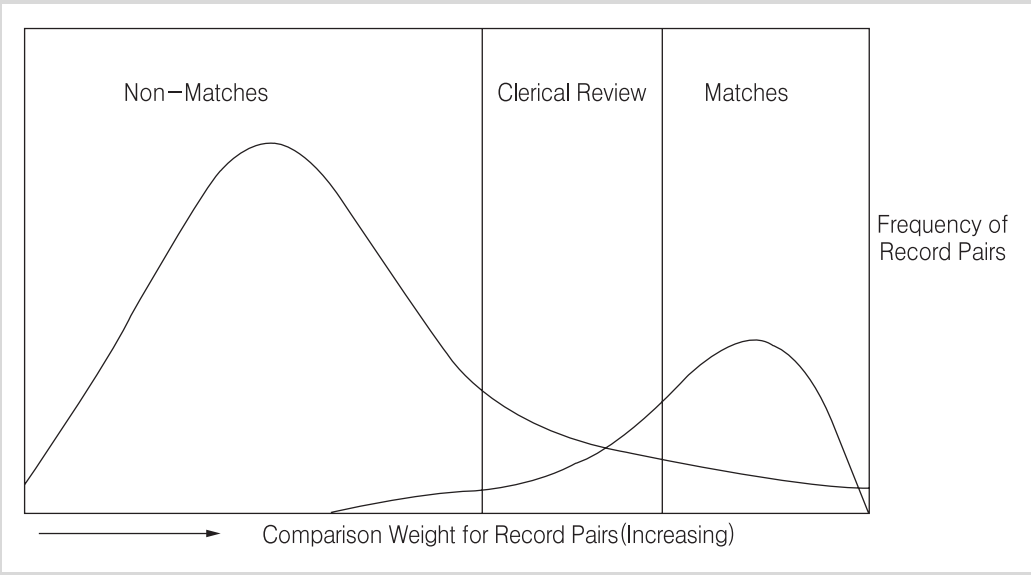
자료: Gu, Lifang, et al. (2003). Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences Technical Report, 3, 83.

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. [그림3]에서 확인할 수 있는 바와 같이 비매치의 최빈값은 매치의 최빈값보다 훨씬 크다. 최빈값 사이의 분류 정도는 데이터 연결 작업의 어려운 수준 정도를 의미한다. 마지막으로 데이터 매칭의 최종적 과정은 성과측정(performance measurement)으로, 데이터 매칭의 질을 측정하는 단계를 의미하며 정확하게 연계된 레코드 쌍 수(true positives), 부정확하게 연계된 레코드 쌍 수(type Ⅰ error), 정확하게 연계되지 않은 레코드 쌍 수(true negatives), 부정확하게 연계되지 않은 레코드 쌍 수(type Ⅱ error)를 측정한다.

3. 데이터 매칭 활용사례: 호주

호주에는 데이터 연계 시스템인 the Westem Australian Data Linkage System(WADLS)이 1995

그림 3. 매칭가중치의 히스토그램

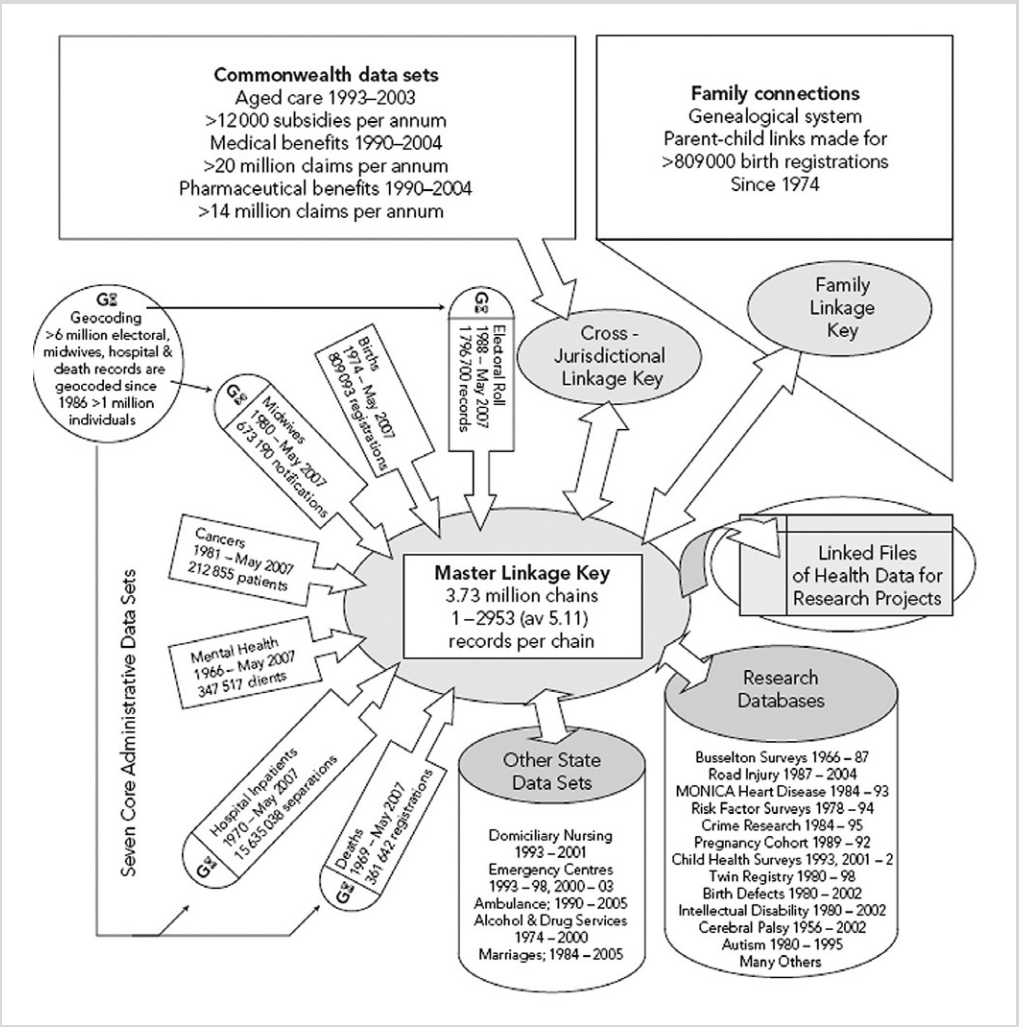


자료: Gu, Lifang, et al. (2003). Record linkage: Current practice and future directions. CSIRO Mathematical and Information Sciences Technical Report, 3, 83.

년부터 있었으며 이는 2007년도까지 400개가 넘는 연구들과 250개의 학술논문 등의 연구 성과를 가능케 했다[4]. 데이터 매칭 연구 분야에서 선두적 위치에 있는 호주는 인구보건과 임상자료의 연계를 통해 보건 영역에서 데이터 매칭의 역할을 확실히 보여주고 있다. 데이터 매칭의 연구 활용은 다음과 같다. 특정조건을 가진 환자의 병원치료(병원 이용자료)와 병원치료의 패턴과 비용(병원 비용자료)을 연결해서 연구를 할 수 있고, 의료서비스 이용자료와 사망 · 질병 명단 자료와 연계시켜 의료 서비스의 효과를 분석할 수 있다. 또, 건강 데이터와 다른 영역의 데이터(지역사회 치료 서비스, 교육, 노인 간호)의 통합으로 건강과 사회적 요인과의 관계를 살펴볼 수 있다. 특정 지역단위 환경에 노출된 개인에 대한 자료와 건강에 대한 자료가 연계되면 환경적 요인과 건강간의 관계도 연구할 수 있다. 이처럼 연계 데이터 활용은 보건정책 관련 연구의 질을 높이고, 연계된 데이터 분석을 통해 각 분야(임상, 보건, 환경 등)의 협력관계를 이끌어 낼 수 있다. [그림4]를 보면 WADLS에 무수히 많은 데이터 셋들이 포함되어 있고 연구자가 알고자 하는 목적에 따라서 다양하게 데이터가 연계될 수 있음을 알 수 있다.

4) Holman, C. D'Arcy J., et al.(2008). A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Australian Health Review, 32(4), pp.766~777.

그림 4. 호주 the Western Australian Data Linkage System의 범위(2007년)



자료: Holman, C. D'Arcy J., et al.(2008). A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Australian Health Review, 32(4), pp.766~777.

이러한 데이터 연계 사례는 복지 분야에서도 다양하게 찾아볼 수 있는데 호주 정부(Australian Government)와 호주보건사회연구원(Australian Institute of Health and Welfare)이 발간한 Data Linkage Series 13호에 의하면 사회적 배제의 위험에 있는 아동 및 청소년에 관한 연구(Children and young people at risk of social exclusion)는 세 가지의 데이터베이스를 연계 활용하여 중요한 결과를 제공한다. 지원시설 보조 프로그램(Supported Accommodation Assistance

Program), 청소년 사법감독(Juvenile justice supervision), 아동보호 고시(Child protection notifications)자료의 결합을 통해 아동학대 및 방임, 노숙자 및 범죄행위 사이의 관계를 살펴보고자 하였다. 그 결과 세 가지 데이터베이스 중 어느 하나에 속한 사람은 일반 인구보다 다른 두 데이터베이스의 영역에 포함될 가능성이 높았다. 그리고 아동보호 기록이 있는 청소년은 어린 나이에 청소년 사법감독 하에 들어간다는 것과 구금형별을 마친 여성 청소년이 노숙자가 될 가능성이 높다는 결과를 도출할 수 있었다.

4. 데이터 통합 연계 시 고려사항

실제로 보건복지 분야에서 다양한 데이터를 활용할 때 필요한 변수를 모두 포함하고 있는 데이터 파일은 발견하기 쉽지 않다. 이러한 단일 자료로부터 얻을 수 있는 정보의 한계를 보완하기 위해서 다양한 데이터 매칭기법들을 활용하여 필요한 변수를 만들어내고 통합된 데이터로부터 추가 정보를 얻어내는 것이 데이터 통합 연계방안의 핵심이라고 할 수 있다. 데이터 매칭은 효율적 분석을 가능하게 하고 데이터의 질을 높일 수 있으며 새로운 결과를 도출하여 다양한 학술적 · 정책적 함의를 제시할 수 있는 기반을 제공할 수 있다. 특히, 최근 다양한 분야에서 제기되고 있는 다양한 통계조사자료와 정부 및 공공기관 행정자료의 통합 연계 및 활용은, 박근혜 정부가 국정과제로 제시하고 있는 정부 3.0 시대를 실현하기 위해서 추진되어야 하는 구체적인 실천 사례가 될 것이다.

향후, 보건복지 분야에서 다양한 데이터 매칭 기법을 활용한 데이터 통합 연계 시 반드시 고려해야 하는 사항이 몇 가지 있다. 첫 번째 고려사항은 개인정보보호에 관련된 이슈이다. 데이터 통합 연계는 신중하게 접근하여 해결해야 하는 개인정보 보호의 윤리적 · 법적 문제를 발생시킨다. 호주의 경우, 연방 개인정보 보호위원회 (The Office of the Federal Privacy Commissioner)가 개인정보 보호의 지속적 발전에 중요한 역할을 하고 있다. 연방 개인정보 보호위원회는 개인정보 보호법 및 관련 법령에 따라 그들의 보호받을 권리를 개인에게 조언해주며 개인정보 보호기준 및 실천과 관련된 모범사례를 발굴하여 홍보하기도 한다. 또한, 개인정보 보호법 및 관련 법규를 준수하는 방법에 대한 조언을 제공해주며 개인정보 보호 관련사항에 대하여 제안된 법령을 검토하는 등의 중요한 역할을 담당하고 있다. 우리나라도 데이터 통합 연계와 관련된 개인정보 보호를 위한 협의체를 두어 부처 간의 정보의 교류와 정부의 정책 추진을 효율적으로 지원할 수 있어야 한다. 이와 관련하여, 개인정보를 최대한 보호하고 합법적으로 정책 수립 및 이와 관련된 연구 목적을 위해 그들의 정보를 이용하고, 여러 기관에 흩어져 있는 다양한 데이터의 통합 연계방법에 대한 프로세스를 설명하기 위한 프로토콜이 필요하다. 이와 같은 프로토콜은 개인정보보호 문제를 해결하고 미래의 통계적 연계 프로젝트를 지원하기 위한 실천 가능한 틀을 제공하기 위해서 개발되어야 한다. 예를 들어, 데이터 결합 이전에 임상실험이나 진단의 기타 정보에서 동일한 이름 및 주소 등의 식별번호를 암호화하고

구분시키는 것이다.

두 번째로 통합 연계에 의해 생산된 데이터의 신뢰성 등에 대한 품질관리 방안의 마련이 필요하다. 데이터의 품질은 데이터 사용과 관련된 적합성으로 정의되기도 하며 통합 연계된 데이터 값의 정확도, 완전성, 시의성, 일관성[5]으로 설명될 수 있다. 이러한 데이터의 품질을 평가하는 것은 실제로 매우 어려운 작업이지만, 일반적으로 데이터 품질은 품질 속성이나 특성으로 정의된다. 데이터 품질 향상을 위한 알고리즘 구현 및 설계를 통해서 통합 연계된 데이터에 대한 신뢰성을 높이기 위한 작업은 이탈리아 등의 연구 프로젝트에서 그 사례를 찾아볼 수 있으며, 이와 관련된 지속적인 연구가 필요하다.

마지막으로, 대용량 데이터와의 연계 시 자료분석방법과 시스템, 인력 등의 기반시설이 지원되어야 한다. 정부 및 공공기관에서 빅 데이터 구축 및 활용을 위한 다양한 사업들을 추진하고 있는 상황에서, 대용량 데이터와의 연계는 빅데이터와 직결될 수 있으며 다양한 데이터 매칭에 대한 분석을 수행할 수 있는 기반을 구축하여 적극적으로 활용할 수 있어야 할 것이다. 그 밖에도 데이터 통합 연계방법에 의한 자료의 생산 및 활용 등 활성화를 위해 정부 및 공공기관에 구축된 다양한 행정자료에 대한 접근을 용이하게 지원해주어야 하며, 이를 담당하는 별도의 기관을 설치하여 법적 근거를 바탕으로 체계적으로 관리 및 지원이 이루어질 수 있도록 해야 한다.

요컨대, 데이터 통합 연계는 그 자체가 어떠한 결과를 제공해준다고 할 수는 없을지라도 향후 학술 연구를 위한 통계 분석뿐만 아니라 보건복지 분야 및 범정부적으로 정책수립 및 평가에 있어서 매우 중요한 툴로 활용할 수 있는 가능성이 무한하다고 할 수 있다. 다양한 데이터 매칭기법들을 활용하여 통합 연계된 자료로부터 얻은 추가 정보로 변수들 간의 관계를 새롭게 바라볼 수 있는 시각이 중요하다. 데이터 통합 연계는 신규 통계의 발굴과도 밀접하게 관련되어 있으므로 보건복지 분야뿐만 아니라 관련 분야에 생산되는 다양한 조사자료 및 행정자료의 통합 연계방안에 대한 지속적 관심이 필요하다고 할 수 있다. 보건복지

5) Bertolazzi, P., et al.(2003). Automatic record matching in cooperative information systems. In Proceedings of the ICDT(Vol. 3).