연구보고서 2024-21

보건복지분야 비식별화 데이터 생성 및 관리체계 구축 기반 연구

오미애
박성준·안수인·권성훈·송지은·김현규

사람을 생각하는 사람들

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

한국보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS

▌연구진

연구책임자	**오미애**	한국보건사회연구원 연구위원
공동연구진	**박성준**	한국보건사회연구원 부연구위원
	안수인	한국보건사회연구원 전문연구원
	권성훈	건국대학교 응용통계학과 교수
	송지은	한국보건사회연구원 전문원
	김현규	한국보건사회연구원 전문원

연구보고서 2024-21

보건복지분야 비식별화 데이터 생성 및 관리체계 구축 기반 연구

발 행 일 2024년 12월
발 행 인 강 혜 규
발 행 처 한국보건사회연구원
주 소 [30147]세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 http://www.kihasa.re.kr
등 록 1999년 4월 27일(제2015-000007호)
인 쇄 처 고려씨엔피

ISBN 979-11-7252-038-0 [93510]
https://doi.org/10.23060/kihasa.a.2024.21

발|간|사

데이터 분석과 활용의 중요성이 높아짐에 따라 개인정보 보호의 필요성도 강조되고 있다. 개인정보 노출 위험과 관련하여 비식별화 방법론을 통해 데이터를 보호하면서 동시에 데이터를 효과적으로 활용할 수 있는 방법 개발에 대한 요구도 증가하고 있는 상황이다.

비식별화 방법론은 데이터에 포함된 개인정보를 이용하는 환경과 데이터에 포함된 재식별 위험성을 검토하여 이를 다양한 방법으로 적절하게 처리하여 식별이 발생하지 않도록 조치할 수 있는 프로세스를 의미한다.

비식별 처리 방법의 특징은 데이터 도메인과 활용 목적에 따라 방법론의 성능이 민감하게 반응하기 때문에 일반론적으로 접근하기 매우 어렵다는 점이다. 따라서 아직 어떠한 기법도 성능의 측면에서 상대적인 우위에 서 있다고 할 수 없다. 일반화하기 어려운 비식별화 처리 방법의 특성에도 불구하고 비식별화 데이터 생성이 중요한 이유는 개인정보 노출 위험과 데이터 활용 간의 균형을 맞추기 위해서이다.

이 연구는 보건복지분야 데이터 활용도를 제고하기 위해 익명화 수준의 데이터 비식별화 처리 방법을 검토하고 식별 위험을 비교하여 비식별화 데이터의 생성 및 관리체계를 수립함으로써 데이터의 안전한 활용 및 확산 기반을 마련하는 데 목적이 있다. 데이터의 안전성을 확보하고, 개인정보 보호를 위한 일관된 접근 방식을 제공한다면 데이터 보호와 관련된 규제를 준수할 수 있을 뿐만 아니라 데이터를 분석·활용하는 프로세스에 대해 사회적 신뢰를 확보할 수 있을 것이다.

본 보고서의 결과는 우리 연구원의 공식적인 견해가 아니라 연구진의 의견임을 밝혀 둔다.

2024년 12월

한국보건사회연구원 원장직무대행

강 혜 규

목 차

표 목차

그림 목차

요약

1. 연구의 배경 및 목적

현 정부는 '디지털 플랫폼 정부 실현을 위한 모든 데이터의 개방과 연결'이라는 비전하에 디지털 플랫폼 정부의 중점 추진과제로 '국민이 원하는 양질의 데이터 전면 개방 및 활용 촉진'을 위해 데이터 공유 및 활용체계를 확립하고자 한다. 하지만 개인정보가 포함된 데이터를 적절히 보호하지 못하면 개인이 식별되거나 혹은 민감한 정보가 유출되는 등 개인의 프라이버시를 침해하는 다양한 위협을 방지할 수 없다.

데이터 분석과 활용의 중요성이 높아지고 동시에 개인정보 보호의 필요성도 커지고 있기 때문에 비식별화 방법론을 통해 데이터를 보호하면서 동시에 데이터를 효과적으로 활용할 수 있는 방법을 개발할 필요가 있다.

비식별화 방법론은 데이터에 포함된 개인정보를 이용하는 환경과 데이터에 포함된 재식별 위험성을 검토하여 이를 다양한 방법으로 적절하게 처리하여 식별이 발생하지 않도록 조치할 수 있는 프로세스를 의미한다. 이때 처리 방법으로는 데이터를 사용할 수 있는 환경에 대한 제어와 데이터에 대한 처리를 포함하며, 데이터에 대한 처리는 데이터에 포함된 식별정보를 삭제하거나 변수와 레코드를 변형하는 등 데이터 주체의 식별 정보가 데이터로부터 직접적으로 혹은 간접적으로 식별되지 않도록 하고, 민감한 정보가 노출되지 않도록 조치할 수 있는 기법을 포함한다. 비식별화 처리 방법의 특성상 데이터 도메인과 활용 목적에 따라 방법론의 성능이 민감하게 반응하기 때문에 일반론적으로 접근하기 매우 어렵다는 점에도 불구하고, 비식별화 데이터 생성이 중요한 이유는 개인정보 노출 위험과 데이터 활용 간의 균형을 맞추기 위해서이다.

이 연구는 보건복지분야의 데이터 활용도를 제고하기 위해 익명화 수

준의 데이터 비식별화 처리 방법을 검토하고 식별 위험을 비교하여, 비식별화 데이터의 생성 및 관리체계를 수립함으로써 데이터의 안전한 활용 및 확산 기반을 마련하는 데 목적이 있다.

2. 주요 연구 결과

제2장에서는 국내외의 주요 사례를 살펴보는데, 국내 사례로 개인정보 비식별화와 관련된 국내의 5개 가이드라인을 검토하고, 국외 사례로는 미국과 일본의 보건의료분야의 개인정보 비식별화와 관련된 사례를 검토하였다. 국내 사례로 검토한 가이드라인은 금융분야 가명·익명 처리 안내서, 가명정보 처리 가이드라인, 보건의료데이터 활용 가이드라인, 교육분야 가명·익명 정보 처리 가이드라인, 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인이다. 5개 가이드라인의 연혁을 살펴보면, 법 개정과 실무 적용, 가명·익명 처리 사례의 누적 및 시대 변화에 따라 가명 또는 익명 정보 활용 수요자들의 요구에 부응하여 지속적으로 개정되고 있음을 알 수 있다. 5개 가이드라인의 내용을 살펴보면, 크게 통계법과 관련 있는 통계청 가이드라인과 데이터 3법에 관련된 그 외 4개 가이드라인으로 구분할 수 있다. 데이터 3법과 관련된 가이드라인들은 가명처리 또는 익명처리 절차에 따라 처리 방법 및 절차를 설명하고 있는 반면, 통계청 가이드라인에서는 통계 작성 단계별 또는 통계자료 제공 단계별로 업무 절차에 따라 비식별화 방법을 설명하고 있다. 통계청의 가이드라인은 적용 대상이 통계작성기관이기에, 통계작성기관인 한국보건사회연구원도 데이터 제공 시 익명처리 절차 및 비식별화 처리에 대한 기준이 필요함을 시사한다.

비식별화, 가명처리, 익명처리에 대한 정의는 가이드라인별로 상이하게 기술하고 있으나, 맥락은 같다. 통계청 가이드라인은 가명처리와 익명

처리 모두를 포괄하는 '비식별화'를 포괄 범위로 하고 있다. 본 연구에서는 비식별화 데이터 생성을 익명처리에 가까운 비식별화로 정의하여 검토하고자 하였다.

5개 가이드라인에서 소개하는 비식별화 처리 방법은 국제 표준인 ISO/IEC 20889 분류를 따르며, 이 보고서의 제3장 비식별화 방법론 검토도 ISO/IEC 20889에서 제안한 분류 체계에 따라 다양한 비식별화 방법을 정의하고 특징을 검토하고자 하였다.

가명·익명 처리의 절차와 적정성 평가는 통계청 가이드라인의 경우 통계작성기관의 업무 수행 절차를 따르며, 비식별 처리에 대한 적정성 평가는 하지 않는다. 나머지 4개의 가이드라인의 익명처리는 사전 준비 및 익명처리 후 적정성 평가를 하는 절차로 이루어진다. 추가적으로, 식별자(identifier)와 관련한 가이드라인별 용어 정의를 요약하여 제시하였다. 이 보고서는 비식별화 데이터 생성과 관련하여 개인식별 가능 정보(준식별자) 용어를 주로 사용하였다.

국외의 사례는 의료분야를 중심으로 비식별 처리와 관련한 사항을 검토하였다. 미국 HIPAA의 PHI 익명화 방법에 대한 지침은 보호 대상 정보에 대해 비식별 처리를 하는 방법으로 전문가 결정 방식과 세이프 하버 방식을 제시하였다. 세이프 하버 방식은 개인에 대한 이름, 주소, 날짜, 이메일 주소, 사회보장번호 등 18가지의 식별 정보를 모두 제거하고 사용하는 방식이다. 일본 차세대의료기반법 가이드라인은 의료 정보 분류를 식별자, 준식별자, 정적 속성, 반정적 속성, 동적 속성으로 나누어 익명으로 가공하는 방법인데, 제2장에서 그 예시를 제시하였다.

제2장 비식별화 처리 관련 국내 5개 가이드라인과 국외 사례의 비식별 처리 제반 사항을 참고하여 연구원의 상황에 맞게 적용하고자 하였는데, 그에 관한 자세한 부분은 제5장에 기술하였다.

제3장에서는 프라이버시 향상을 위한 데이터 비식별화 용어와 기법 분류에 대한 국제 표준인 ISO/IEC 20889의 비식별화 방법론과 차등 정보보호, 재현 데이터 방법론을 검토하였다.

변수·레코드 수준 비식별화에서 구조적 방법은 데이터의 레코드 구조를 변형하여 제공하는 방법이며 대표적으로 표본추출 방법이 있다. 삭제 방법으로는 식별 변수 삭제, 레코드 삭제, 마스킹이 있으며, 일반화 방법으로는 라운딩 방법, 범주화 방법을 소개하였다. 임의화 방법은 잡음 첨가, 데이터 교환, 부분 총계 방법이 있으며, 실무에서는 통계적으로 다양한 변수의 분포를 고려해야 하는 임의화 방법보다는 삭제 방법이나 일반화 방법에 대한 접근이 더 쉽다고 판단된다.

차등 정보보호는 데이터 분석 시, 개인정보가 노출되지 않도록 보호하는 기술로, 데이터에 노이즈(무작위성을 추가한 값)를 추가하여 개인을 식별할 수 없도록 하면서 데이터의 전체 패턴은 분석할 수 있게 하는 방법이다. 재현 데이터는 원본 데이터의 통계적 특성을 유지하면서 새로운 가상의 데이터를 생성하는 기법으로, 실제 개인정보는 포함되지 않을 수 있다. 개인정보 관련 규제로 데이터의 활용이 제한된 상황에서는 재현 데이터가 좋은 대안이 될 수 있다.

차등 정보보호와 재현 데이터는 상호 보완적으로 사용될 수 있으며, 앞으로도 활용도가 높아질 것으로 예상할 수 있기 때문에 본 연구에서도 방법론적으로 중요하게 다루었다.

마지막으로 비식별화 방법론의 실무 적용을 위해 sdcMico를 사용하여 활용 예시를 구체적으로 제시하였다. 이 부분은 제4장의 노출 위험 분석과도 연관되어 있고, 제5장의 원내 비식별화 가이드라인에도 실무적으로 활용할 수 있는 부분이다.

제4장에서는 한국복지패널, 가족과 출산 조사, 정신질환자의 건강 및

복지서비스 인식 및 이용 경험 조사별로, 실무적으로 활용할 수 있는 개인식별 가능 항목을 범주화 비식별화 처리 방법을 사용해 노출 위험 시나리오 6가지를 구성하였다. 한국복지패널과 가족과 출산 조사는 시나리오 구성 시, 지역 변수의 범주 범위를 확대하였을 경우 노출 위험이 얼마나 증가할 것인지를 검토하였다. 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사는 연구원의 원내 조사자료이므로 어느 정도의 수준까지 비식별화 처리 작업이 필요한지에 대한 판단에 도움을 주고자 시나리오를 구성하였다. 준식별 정보는 인구학적 정보와 개인의 경제상태를 나타내는 정보로 구성하고 각 조사의 민감 정보를 활용하여 k-익명성과 전체 위험도, l-다양성 분석 결과를 제시하였다. 데이터의 비식별화 처리 기법은 실무 레벨에서 적용 가능한 수준인 재범주화, 범주화, 상단코딩, 삭제 방법을 사용하였다.

첫 번째 데이터는 18차 한국복지패널로, 분석 1~3은 지역 변수를 7개 권역으로, 분석 4~6은 지역 변수를 16개 시도로 분류하여 활용하였고, 분석 1에서 분석 3, 분석 4에서 6으로 갈수록 활용 변수를 삭제, 범주화하여 공개 범위를 축소시켰다.

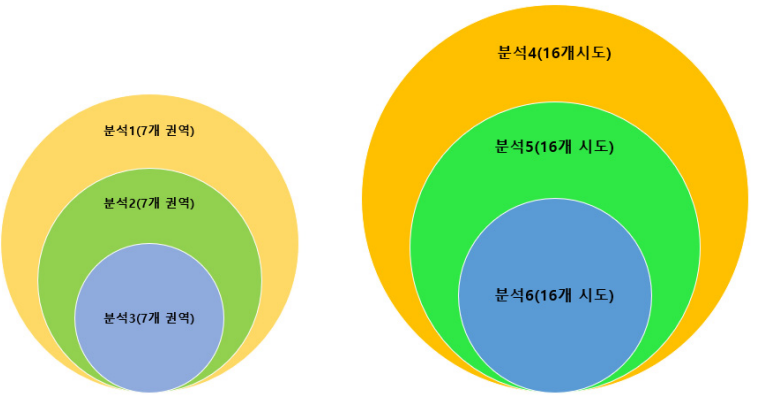
k-익명성의 경우, 분석 1에서 전체 데이터의 약 75%, 분석 3에서는 약 8%, 분석 4에서는 약 83%, 분석 6에서는 약 13%가 3-익명성을 만족하지 않았다.

전체 위험성의 경우, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.36%이고, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 0.03%였다. 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.43%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.06% 정도였다.

두 번째 데이터는 2021년 가족과 출산 조사로, 분석 1~3은 지역 변수

를 3개 권역으로, 분석 4~6은 지역 변수를 17개 시도로 분류하여 활용하였고, 분석 1에서 분석 3, 분석 4에서 6으로 갈수록 활용 변수를 삭제, 범주화하여 공개 범위를 축소시켰다.

[요약 그림 1] 복지패널의 노출 위험 시나리오별 변수 공개 범위 비교



구분 \ 권역	7개 권역		
시나리오	분석 1	분석 2	분석 3
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(교육) (혼인)(취업)(소득)

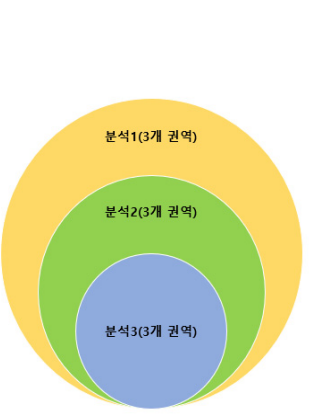
구분 \ 권역	16개 시도		
시나리오	분석 4	분석 5	분석 6
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(교육) (혼인)(취업)(소득)

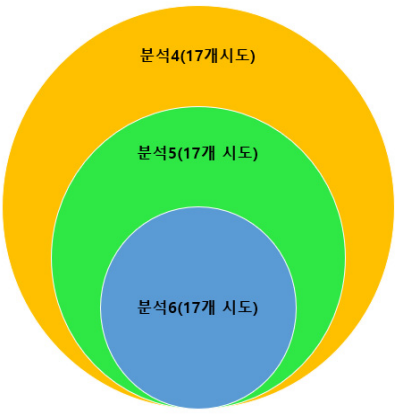
출처: 저자 작성

2021년 가족과 출산 조사를 활용한 k-익명성 분석 결과, 분석 1에서 전체 데이터의 약 40%, 분석 3에서는 약 9%, 분석 4에서는 약 76%, 분

석 6에서는 약 41%가 3-익명성을 만족하지 않았다.

[요약 그림 2] 가족과 출산 조사 노출 위험 시나리오별 변수 공개 범위 비교





구분 \ 권역	3개 권역		
시나리오	분석 1	분석 2	분석 3
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(소득) (혼인)(경제활동상태)(교육)

구분 \ 권역	17개 시도		
시나리오	분석 4	분석 5	분석 6
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(소득) (혼인)(경제활동상태)(교육)

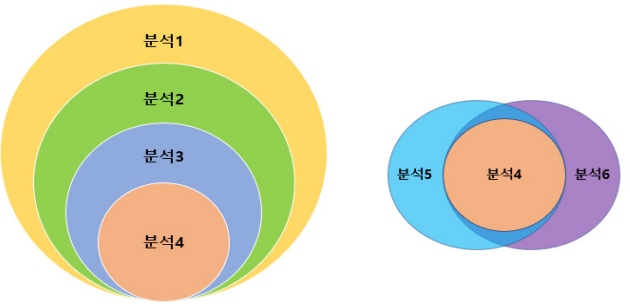
출처: 저자 작성

전체 위험성의 경우, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 46%이고, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 16%였다. 분석 4에서 식별이 될 것으로 예측되는 레코

드 수는 전체의 약 84%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 53%였다.

세 번째 데이터는 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사(2022)이다.

[요약 그림 3] 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 노출 위험 시나리오별 변수 공개 범위 비교



구분 \ 권역	4개의 시나리오 비교			
시나리오	분석 1	분석 2	분석 3	분석 4
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령) (소득)	6 (직종)(연령) (소득) (혼인)(주된 진단명)(교육)	6 (연령)(소득) (교육) -직종, 혼인, 주된 진단명은 삭제-

구분 \ 권역	3개의 시나리오 비교		
시나리오	분석 4	분석 5	분석 6
공개 범위 축소 정도 (변수 삭제 및 범주화)	6 (연령)(소득) (교육) - 직종, 혼인, 주된 진단명은 삭제 -	5 (연령)(소득) (교육)(주된 진단명) - 직종, 혼인은 삭제 -	5 (연령)(소득) (교육)(직종) - 혼인, 주된 진단명은 삭제 -

출처: 저자 작성

정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 분석 1에서 분석 4로 갈수록 활용 변수를 삭제, 범주화하여 공개 범위를 축소시켰다. 분석 5와 분석 6은 분석 4의 활용 변수에서 주된 진단명 범주, 직종 범주를 각각 추가하였을 때의 노출 위험 증가를 보고자 한 것이다.

지역사회 거주 정신질환자 조사의 k-익명성 분석 결과, 분석 1에서 전체 데이터의 약 94%, 분석 3에서는 약 37%, 분석 4에서는 약 18%, 분석 5에서는 약 20%, 분석 6에서는 약 21%가 3-익명성을 만족하지 않았다.

전체 위험성의 경우, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 91%이고, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 39%였다. 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 24%이고, 분석 5와 분석 6에서는 전체의 26% 수준이었다.

이 6가지 시나리오를 통해 복지패널, 가족과 출산 조사 데이터는 지역변수 공개 범위를 시도 단위까지 확대할 경우, k-익명성과 전체 위험성으로 어느 정도 노출 위험도가 증가하는지를 파악할 수 있었다. 전체 위험성의 수준이 한국복지패널과 가족과 출산 조사에서가 다른 이유는 한국복지패널은 일반 가중치를 적용하였고, 가족과 출산 조사는 표본 가중치를 적용하였기 때문이다. 가족과 출산 조사는 일반 가중치를 공개하고 있지 않기 때문에 표본 가중치를 적용한 전체 위험성의 수준이 복지패널보다 높게 나타났다. 이는 어떤 가중치를 적용하느냐도 위험측도에 따라 다른 결과를 제시해줄 수 있음을 의미한다.

지역사회 거주 정신질환자 조사는 원내의 조사자료 중 하나로, 다양한 준식별 정보의 조합으로 노출 위험을 측정하였을 때, 원자료 수준에서는 노출 위험도가 매우 높음을 알 수 있고, 변수별로 범주화, 재범주화가 필요하다는 것을 알 수 있다.

노출 위험 분석으로 k-익명성의 단점을 보완한 l-다양성 분석 결과도 함께 제시하였다.

이 세 데이터 분석을 통해 알 수 있는 것은 데이터의 특성, 준식별 변수의 조합, 노출 위험 측도에 따라 상대적인 평가가 가능할 뿐, 절대적인 기준은 없다는 것이다. 이로써 실제 비식별화 데이터 생성 작업 시, 고려해야 하는 요인들이 적지 않음을 의미하고, 준식별 변수 조합의 수준, 민감정보의 정의에 대해 어느 정도의 범위까지 통일된 양식, 기준을 가지고 갈 수 있느냐에 대한 논의가 필요하다는 것을 알 수 있다. 통계청의 통계작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성(2023)에서도 비식별화 처리 방법에 대한 기준을 강조하기보다, 통계작성기관의 특성에 적합한 방안을 지속적으로 개선·개발할 필요가 있음을 언급하고 있다.

이에 제5장에서는 한국보건사회연구원의 연구보고서 작성 중에 생산된 조사자료의 개인식별 가능 정보를 검토하여 비식별화 처리 현황을 살펴보고, 비식별화 처리 관련 가이드라인을 개발하였다. 그리고 비식별화 데이터의 이관 및 관리 절차를 설정하여 연구원의 데이터 관리체계 안에서 비식별화 데이터 처리에 대한 부분을 상세히 검토하였다.

현재 연구원에서 공개를 전제로 검토하고 있는 조사자료의 경우, 비밀보호 처리 부분에서 명확한 기준 없이 실무자들의 주관적 판단에 의해 데이터를 처리하고 있기 때문에 최소한의 개인식별 정보를 판단하는 기준과 처리에 대한 기준을 설정할 필요가 있다. 이를 위해 2021~2022년 조사자료를 검토한 결과, 개인식별 가능 정보로 사용되는 항목은 성별, 연령, 지역, 최종학력, 혼인상태, 장애, 종교, 가구소득, 경제활동상태, 직종, 거주 형태, 가구 구성, 만성질환, 자산 및 부채, 사회복지제도까지 15개로 요약된다. 동일한 항목일지라도 조사 목적에 따라 항목별 세부 조사내용은 조사별로 상이하기 때문에 해당 변수에 대한 구체적인 처리 방식

을 설정하여 일반화된 처리 원칙을 만들 필요가 있다.

원내의 비식별화 처리 관련 가이드라인을 개발하고, 공개용 조사자료의 사전 검토의견서 양식을 보완하였고, 일관된 데이터 처리를 위해 체크리스트를 새롭게 작성하였다. 기존의 검토의견서 양식은 서술식으로 비식별화 관련 내용을 기술하였다면, 보완된 검토의견서 양식에는 비식별화 처리표를 작성하도록 하여 명확하게 해당 데이터에 대한 개인식별 정보 및 민감 정보를 검토할 수 있도록 하였다.

비식별화 처리 가이드라인에는 비식별화 처리 원칙을 세우고, 비식별화 처리 과정을 단계별로 설정하여, 개인식별 가능 정보의 항목을 어느 정도의 수준까지 범주화할 것인지에 대한 기준표를 작성하였다. 비식별화 처리 절차 중 중요한 부분은 단일 변수의 1차 비식별화 처리이므로 이 부분은 조사자료의 연구보고서에 제시된 범주를 기준으로 비식별화 처리를 하는 것을 원칙으로 하였다. 그리고 이에 대한 모든 사항은 연구책임자가 확인하고 결정하도록 하였다. 원내의 조사자료 비식별화 처리와 관련하여 범주화와 관련된 시나리오, 가공 변수와 관련된 시나리오를 제시하여 비식별화 처리 작업에 대해한 쉽게 이해할 수 있도록 하였다.

비식별화 데이터의 이관 절차는 원내의 조사데이터 관리 절차와 연구데이터 관리 업무 절차와 구분하여 살펴보았다. 비식별화 처리 및 비식별화와 관련한 업무에서 중요한 부분은 제출용 데이터를 생성하기 위한 일련의 과정과 데이터의 보안 범위를 설정하기 위한 검토 및 심의 절차라고 할 수 있다. 비식별화 처리 과정에서 논의가 필요한 경우나 처리 결과에 대해 재검토가 요구되는 경우에는 관련 위원회를 통해 검토를 요청할 수 있으며, 이때 연구자 및 실무자의 의견이 반영될 수 있도록 한다. 이는 기존 데이터 관리지침이나 비식별화 관리지침에 기반하여 다루어지기 어려운 사항들에 한하여 필요한 절차이며, 이때 외부 전문가의 의견 또한 반영될 수 있다.

비식별화 데이터의 처리 과정은 비식별 범위 설정, 비식별 처리 수준 설정, 데이터 제공 의무의 이행 등 많은 요소와 관련되어 있기 때문에 별도의 업무 과정이 설정될 필요가 있다. 비식별 처리의 과정과 비식별화 데이터의 관리를 위해서는 1단계로 범위를 설정(개인식별 가능 민감/일반 정보 구분)하고, 2단계로 처리 예외 사항을 확인하고(삭제/비활용 정보 확인 및 처리), 3단계로 비식별 처리를 하고(비식별화 처리 및 처리 결과에 대한 확인), 4단계로 검증 및 확인하는(검증 절차 확인 및 개인식별 위험성 검토) 단계적 환류체계를 설정하여야 한다.

데이터의 특성별로 검토 과정을 세분화할 필요도 있다. 비식별화 관점에서 데이터의 특성을 삭제 처리 대상, 비식별 처리 대상, 비식별 판단 대상(지침에 따른 범위 설정에 해당되지 않는 경우, 연구자와 담당자 간 협의를 통하여 처리 대상 확정이 불가한 경우를 의미)으로 나누었을 때, 비식별 판단 대상은 관련 위원회 및 전문가의 의견 수렴을 통한 검토가 이루어져야 한다.

비식별화 처리 데이터의 관리는 연구데이터의 관리 단계로 보면 마이크로데이터 제공 단계로 볼 수 있다. 다만, 제공 데이터 가운데 가명정보 결합 요청이 발생한 경우, 데이터 활용 중 데이터의 오류가 확인된 경우, 개인정보 및 민감 정보 관련 이슈가 발생한 경우에는 연구자 및 비식별 처리 담당자가 사후 처리를 진행할 수 있다.

3. 결론 및 시사점

제6장 결론 및 시사점에서는 이 연구의 내용을 요약하고 시사점을 제시하였다. 시사점으로는 원내 비식별화 처리 가이드라인 개발의 의미와 비식별화 데이터의 이관 및 관리 절차 설정의 의미, 데이터 프라이버시

리터러시 역량 강화에 대한 부분을 기술하였다. 비식별화 처리 관련 가이드라인 개발의 의미는 실무자들에게는 일관된 최소한의 업무 지침이며, 연구자들에게는 조사데이터의 개인정보 노출 위험 정도를 알려주는 설명자료라고 할 수 있다. 비식별화 데이터의 이관 및 관리 절차 설정은 한국보건사회연구원의 비식별화 관련 처리 업무의 안착뿐만 아니라 데이터 생산 주체의 안전한 데이터 관리 및 제공 체계를 마련하였다는 점에서 의미가 있다. 비식별화 데이터 사용과 관련된 견해 차이는 비식별화 처리가 된 데이터가 포함된 연구에서 연구 참여자가 개인식별이 가능하다고 생각하는 비율은 연구자 대비 IRB 전문가가 2배, 연구 참여자가 2.6배 더 높았다. 이는 연구자들이 비식별화 처리가 이루어진 데이터 공유에 관한 정책을 수립할 때 조사 참여자들이 인식하는 위험과 피해에 대한 견해를 더 잘 인식하고 고려해야 함을 시사하며, 데이터 프라이버시 리터러시 교육은 필수임을 의미한다. 데이터 프라이버시 리터러시 역량 강화로 개인이 자신의 데이터가 안전하게 보호된다는 확신을 갖는다면 디지털 환경에서의 안전과 신뢰가 높아질 수 있을 것이다.

주요 용어 : 비식별화 데이터, 비식별화 처리 기법, 개인정보 노출 위험, 보건복지분야

제1장

서론

제 1 장 서론

제1절 연구의 배경 및 목적

데이터 경제와 개인정보 디지털 경제가 급성장하면서 데이터의 가치가 높아지고 기업과 정부는 다양한 데이터를 수집·분석하여 혁신적인 서비스를 개발하기 위한 정책을 설계하고 있다. 현 정부는‘디지털 플랫폼 정부 실현을 위한 모든 데이터의 개방과 연결’이라는 비전하에 디지털 플랫폼 정부의 중점 추진과제로 ‘국민이 원하는 양질의 데이터 전면 개방 및 활용 촉진’을 위해 데이터 공유 및 활용체계를 확립하고자 한다. 하지만 개인정보가 포함된 데이터를 적절히 보호하지 못하면 개인이 식별되거나 혹은 민감한 정보가 유출되는 등 개인의 프라이버시를 침해하는 다양한 위협을 방지할 수 없다.

데이터의 유출·해킹 등의 위협이 증가하면서 데이터 보호의 중요성이 크게 부각되고 있으며, 특히 민감한 정보가 포함된 데이터가 보안 위협에 노출되어 심각한 부정적 결과를 초래한 사례가 지속적으로 증가하고 있다. 이러한 문제점을 해결하기 위하여 개인정보 보호에 대한 국제적 규제가 강화되고 있으며 유럽연합의 GDPR, 미국의 CCPA, 한국의 개인정보 보호법 등은 개인정보의 수집, 처리, 보관에 대한 엄격한 기준을 제시하고 있다. 이러한 규제를 준수하기 위해서는 데이터를 비식별화하여 개인정보를 보호할 수 있는 과학적 방법론이 필요하며 이와 관련된 일관성 있는 표준이 요구되고 있다.

분석 변수 측면에서 보면, 조사데이터 변수의 조합이 다양해질수록 속

성 조합의 유일성이 커지면서 특정 개인을 식별할 위험성이 높아진다. 또한 조사데이터에서 유일성이 높지 않다고 해도 이를 집계한 데이터에서 특정 개인이 식별되거나 민감한 속성이 노출될 가능성이 발생한다. 데이터 분석과 활용의 중요성이 높아지고 있지만, 동시에 개인정보 보호의 필요성도 커지고 있으며 비식별화 방법론을 통해 데이터를 보호하면서 동시에 데이터를 효과적으로 활용할 수 있는 방법을 개발할 필요가 있다.

비식별화 방법론은 데이터에 포함된 개인정보를 이용하는 환경과 데이터에 포함된 재식별 위험성을 검토하여 이를 다양한 방법으로 적절하게 처리하여 식별이 발생하지 않도록 조치할 수 있는 프로세스를 의미한다. 이때 처리하는 방법으로는 데이터를 사용할 수 있는 환경에 대한 제어와 데이터에 대한 처리를 포함하며, 데이터에 대한 처리는 데이터에 포함된 식별 정보를 삭제하거나 변수와 레코드를 변형하는 등 데이터 주체의 식별 정보가 데이터로부터 직접적으로 혹은 간접적으로 식별되지 않도록 하고 민감한 정보가 노출되지 않도록 조치할 수 있는 기법을 포함한다.

〈표 1-1〉 익명화와 가명화 비교

특성	익명화(Anonymization)	가명화(Pseudonymization)
복원 가능성	원본 데이터를 복구할 수 없음	처리 기법에 따라 원본 데이터의 복구가 가능
개인식별 가능성	개인 식별과 구별이 불가능	추가 정보가 없는 상태에서는 개인식별이 불가능 개인에 대한 구별은 원칙적으로 가능
데이터 연결성	원본 데이터와의 연결이 완전히 끊김	기법에 따라 원본 데이터와의 연결을 유지할 수 있음
법적 보호	개인 데이터로 간주되지 않음 (GDPR 등에서 규제 완화)	여전히 개인정보이며 개인 데이터 정보에 대한 법적 보호가 필요함
사용 목적	이용자의 환경에 대한 통제가 불가능한 경우 사용(데이터 공개 등)	이용자의 환경에 대한 통제가 가능한 경우 사용 국내에서는 동의 받지 않은 데이터의 특정 목적(과학적 연구, 공익적 기록보존, 통계 작성 등)으로 사용

출처: OpenAI(2024) 및 자문 의견을 참고하여 저자 작성.

가명처리와 익명처리는 가장 대표적인 비식별화 조치이다. 가명처리는 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 것이다. 익명처리는 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없도록 처리하는 것을 의미한다.

비식별 처리 방법의 특징은 데이터 도메인과 활용 목적에 따라 방법론의 성능이 민감하게 반응하기 때문에 일반론적으로 접근하기 매우 어렵다는 점이며, 따라서 아직 어떠한 기법도 성능의 측면에서 상대적인 우위에 서 있다고 할 수 없다. 일반화하기 어려운 비식별화 처리 방법의 특성에도 불구하고 비식별화 데이터 생성이 중요한 이유는 개인정보 노출 위험과 데이터 활용 간의 균형을 맞추기 위해서이다.

이 연구는 보건복지분야의 데이터 활용도를 제고하기 위해 익명화 수준의 데이터 비식별화 처리 방법을 검토하고, 식별 위험을 비교하여, 비식별화 데이터의 생성 및 관리체계를 수립함으로써 데이터의 안전한 활용 및 확산 기반을 마련하는 데 목적이 있다. 데이터의 안전성을 확보하고, 개인정보 보호를 위한 일관된 접근 방식을 제공한다면 데이터 보호와 관련된 규제를 준수할 수 있을 뿐만 아니라 데이터를 분석·활용하는 프로세스에 대한 사회적 신뢰를 확보할 수 있을 것이다.

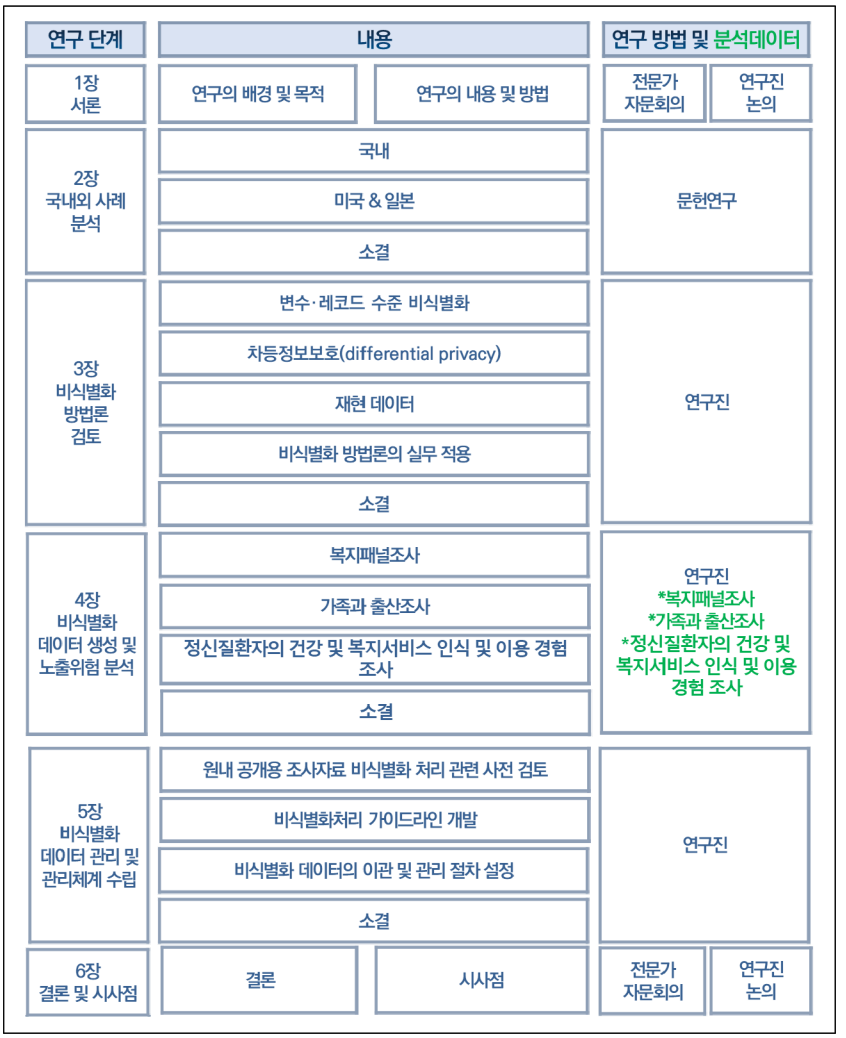
제2절 연구의 내용 및 방법

보고서 「보건복지분야 비식별화 데이터 생성 및 관리체계 구축 기반 연구」의 주요 연구 내용은 다음과 같다. 제2장에서는 국내 개인정보 비식별화와 관련한 5개 가이드라인(금융분야 가명·익명 처리 안내서, 가명정보 처리 가이드라인, 보건의료데이터 활용 가이드라인, 교육분야 가명·익명정보 처리 가이드라인, 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인)과 미국, 일본 사례를 검토하였다. 제3장에서는 데이터 비식별화 용어와 기법 분류에 대한 국제 표준인 ISO/IEC 20889의 비식별화 방법론과 차등 정보보호, 재현 데이터 방법론을 검토하고 비식별화 방법론의 실무 적용을 위해 sdcMico를 사용하여 활용 예시를 구체적으로 제시하였다. 제4장에서는 한국복지패널, 가족과 출산 조사, 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사별로 실무적으로 활용 가능한 개인식별 가능 항목을 범주화 비식별화 처리 방법를 사용해 노출 위험 시나리오 6가지를 구성하여 비교분석하였다. 제5장에서는 한국보건사회연구원 연구보고서 작성 중에 생산된 조사자료의 개인식별 가능 정보를 검토하여 비식별화 처리 현황을 살펴보고 비식별화 처리 관련 가이드라인을 개발하였다. 그리고 비식별화 데이터의 이관 및 관리 절차를 설정하여 연구원의 데이터 관리체계 안에서 비식별화 데이터 처리에 대한 부분을 상세히 검토하였다. 제7장 결론에서는 결론과 시사점을 제시하였다.

이 보고서 작성을 위해 국내외 문헌 연구, 전문가 자문회의, 데이터 분석(R/Python) 등 다양한 방법을 활용하였다.

연구 과정 개요는 다음과 같다.

〔그림 1-1〕 연구 과정 개요



출처: 저자 작성

제2장

국내외 사례 분석

제 2 장 국내외 사례 분석

제1절 국내 사례

국내 사례는 국내의 개인정보 비식별화와 관련한 다음의 5개 가이드라인의 내용을 살펴보고 현재 활용되고 있는 비식별 데이터 생성 기법과 현행 관리체계에 관해 검토하고자 한다.

- 금융분야 가명·익명 처리 안내서
- 가명정보 처리 가이드라인
- 보건의료데이터 활용 가이드라인
- 교육분야 가명·익명 정보 처리 가이드라인
- 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인

1. 국내 개인정보 비식별화 관련 5개 가이드라인 연혁

데이터의 효율적인 활용과 함께 개인정보 보호의 중요성이 커짐에 따라 데이터 산업 활성화 촉진을 위해 국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부 등 관계부처가 합동으로 작성한 '개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내(이하 '비식별 조치 가이드라인')'를 2016년 6월 30일에 발간하였다. 이에 따라, 각 부처가 개인정보 비식별 조치와 관련하여 기존에 발간한 지침, 안내서, 가이드라인 등은 일괄적으로 폐지되고, 2016년 7월 1일부터는 비식별 조치 가이드라인이 적용되었다.

2020년 8월 5일 데이터 3법이 시행되면서 개인정보보호위원회에서는 2020년 9월 24일 '가명정보 처리 가이드라인' 통합본을 발간하였고, 이에 비식별 조치 가이드라인은 폐지되었다. 가명정보 처리 가이드라인은 2021년 10월과 2022년 4월에 개정이 이루어졌으며, 가장 최근의 2024년 2월 개정까지 세 차례에 걸쳐 개정되었다.

「신용정보의 이용 및 보호에 관한 법률(약칭 '신용정보법')」 등 데이터 3법의 개정에 따른 후속 조치로 가명 또는 익명 정보 활용 활성화 및 안전한 활용을 위해 「금융분야 가명·익명 처리 안내서」를 제정하여 발표하였다. 가명·익명 정보의 정의 및 활용 방법 등이 법령에 구체적으로 규정되지 않아 안내서에서는 구체적인 예시 등을 통해 안내하고 있다.

'가명정보 처리 가이드라인' 발표 이후 보건의료 및 교육 분야에서 가명정보의 안전한 활용을 위한 가이드라인들이 마련되었다. 가장 먼저, 보건복지부와 개인정보보호위원회는 '보건의료데이터 활용 가이드라인(이하 '보건의료분야 가이드라인')'을 발표하였다. 보건의료분야 가이드라인에서는 의약품과 의료기기 개발 등을 포함한 과학적 연구를 위해 보건의료데이터가 안전하게 활용될 수 있도록 가명처리 기준과 방법, 절차 등을 제시하였다. 보건의료분야에 뒤이어 교육분야에서도 교육부와 개인정보보호위원회는 '교육분야 가명·익명 정보 처리 가이드라인(이하 '교육분야 가이드라인')'을 작성·발표하였다. 두 가이드라인 모두 '가명정보 처리 가이드라인'에 따른 개인정보 처리 기본원칙을 따르고 있으나 분야별 가이드라인으로서 우선적으로 적용된다.

앞에서 언급한 가이드라인들과 같은 맥락으로, EU의 GDPR과 국내 관련 법 규정에 대응하여 통계작성기관에서는 통계자료 제공을 위해 비식별화 업무 지원의 필요성이 제기되었다. 통계작성기관은 정부부처, 지자체 등의 기관으로 「통계법」에 따라 통계 작성 및 통계 자료 제공 업무를 수행하고 있다. 우리나라의 분산형 통계생산 구조 및 통계 업무 특수성을

고려하여 비식별화를 편리하고 안전하게 수행하기 위한 가이드라인 제작의 필요성이 제기됨에 따라 통계청에서는 '통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인(이하 '통계청 가이드라인')'을 작성하였다. 통계청은 2023년에 이 가이드라인(안)을 국가통계위원회 통계데이터분과위원회의 안건으로 상정하여 심의·의결을 거쳐 통계작성기관을 대상으로 가이드라인 책자를 배포하였다. 그다음 해인 2024년에는 EU의 PEFA tools 2017을 벤치마킹하여 마이크로데이터를 위한 비식별화 세부 기법을 수행하는 가이드라인 프로그램(인터페이스는 엑셀, 실제 분석은 R로 수행)을 개발하는 용역 사업을 수행하는 중이다.

가명정보 처리 가이드라인이 제정된 이후 3차례에 걸쳐 개정이 이루어졌는데 개정 내용을 살펴보면 다음과 같다. 1차 개정에서는 서식 예시와 Q&A 등을 추가하였다. 2차 개정에서는 가명처리 사전 절차인 '위험성 검토'와 '가명처리 방법 및 수준 검토' 방법을 구체적으로 설명하고자 적용 사례, 점검표 등을 제시하였다. 또한, 아우터 결합(outer join) 등 반출 가능한 결합 유형을 시각화하여 제시하였고, 기타 가명정보 처리의 절차별 검토 사항, 필요 자료 등에 대한 참고 사례를 시나리오별로 제시하여 설명하였다. 최근 이루어진 3차 개정에서는 이미지·음성·영상·텍스트 등 비정형 데이터에 대한 가명처리 기준을 제시하였다. 이전까지는 정형 데이터에 대한 가명처리 기준만을 제시하고 있었던 한계점을 보완하고자 비정형 데이터의 가명처리 기준, 기술 및 예시, 가명처리 시나리오 예시 등에 관하여 제시한 것이다. 이는 최근의 인공지능(AI) 기술 확산을 바탕으로 비정형 데이터에 대한 활용 수요 증가 등을 반영한 시대 변화와 시장의 요구에 발맞춘 개정이었다.

국내 개인정보 비식별화 관련 5개 가이드라인의 연혁을 살펴보면, 가명정보 처리 가이드라인뿐 아니라 다른 가이드라인들도 법 개정과 실무 적용, 가명·익명 처리 사례의 누적 및 시대 변화에 따라 가명 또는 익명

정보 활용 수요자들의 요구에 부응하여 지속적으로 개정되고 있음을 알 수 있다. 가이드라인에서는 실무 지원을 위한 가명처리 또는 익명처리 방법 등을 안내하고 있는 만큼 계속된 개정·보완 등이 필요함을 시사한다.

〈표 2-1〉 국내 개인정보 비식별화 관련 가이드라인 연혁

연혁	내용
2016. 6. 30.	관계부처 합동 '개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내' 발간
2020. 8. 5.	데이터 3법 개정·시행 (「개인정보보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」)
2020. 8. 6.	금융위원회·금융감독원 「금융분야 가명·익명 처리 안내서」 제정
2020. 9. 24.	개인정보보호위원회 '가명정보 처리 가이드라인' 발간
2020. 9. 25.	보건복지부·개인정보보호위원회 '보건의료데이터 활용 가이드라인' 발간
2020. 11. 26.	교육부·개인정보보호위원회 '교육분야 가명·익명 정보 처리 가이드라인' 발간
2021. 10. 22.	개인정보보호위원회 '가명정보 처리 가이드라인' 개정(1차)
2022. 1. 7.	금융위원회·금융감독원 「금융분야 가명·익명 처리 안내서」 개정(1차)
2022. 4. 28.	개인정보보호위원회 '가명정보 처리 가이드라인' 개정(2차)
2022. 7. 26.	교육부·개인정보보호위원회 '교육분야 가명·익명 정보 처리 가이드라인' 개정(1차)
2022. 12. 28.	보건복지부·개인정보보호위원회 '보건의료데이터 활용 가이드라인' 개정(1차)
2023. 5. 30.	통계청 통계데이터기획과 '통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인' 작성
2023. 6. 13.	국가통계위원회 통계데이터분과위원회 안건 상정(제2023-07호) '통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성'
2024. 1. 19.	보건복지부·개인정보보호위원회 '보건의료데이터 활용 가이드라인' 개정(2차)
2024. 2. 1. ~	통계청 '통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 프로그램 개발 연구 사업 용역' 진행 중
2024. 2. 4.	개인정보보호위원회 '가명정보 처리 가이드라인' 개정(3차)
2024. 8. 12.	교육부·개인정보보호위원회 '교육분야 가명·익명 정보 처리 가이드라인' 개정(2차)

출처: 개인정보보호위원회. (2024. 2.). 가명정보 처리 가이드라인;
관계부처 합동. (2016. 6. 30.). 개인정보 비식별 조치 가이드라인 –비식별 조치 기준 및 지원·관리체계 안내-;
교육부, 개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인;
보건복지부, 개인정보보호위원회. (2024. 1.). 보건의료데이터 활용 가이드라인;
금융위원회, 금융감독원. (2022. 1.) 금융분야 가명·익명처리 안내서;
통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. 통계데이터분과위원회. 제2023-07호.
각 가이드라인을 참고하여 저자 작성.

2. 국내 개인정보 비식별화 관련 5개 가이드라인의 내용

5개 가이드라인의 내용을 살펴보면, 크게 통계법과 관련 있는 통계청 가이드라인과 데이터 3법에 관련된 그 외 4개 가이드라인으로 구분할 수 있다. 데이터 3법과 관련된 가이드라인들은 가명처리 또는 익명처리 절차에 따라 처리 방법 및 절차를 설명하고 있는 반면, 통계청 가이드라인에서는 통계 작성 단계별 또는 통계자료 제공 단계별로 업무 절차에 따라 비식별화하는 방법을 설명하고 있다. 가명처리 목적에 있어서도 데이터 3법의 법령에서 명시하고 이를 준용하고 있는 가이드라인들과 통계 작성 및 공표, 이용자 요청에 따른 제공 등 업무 수행을 위해 가명처리를 수행하는 통계청 가이드라인으로 구분된다.

5개 가이드라인을 내용을 비교해보면 작성 주체, 발간일, 관련 법령 및 분야뿐 아니라 적용 대상이 서로 상이하다. 개인신용정보, 처리 또는 업무 절차 그리고 개인정보 처리자 또는 제공 받은 자 등이다. 포괄 범위의 경우 대부분 가명처리와 익명처리를 모두 포함하나, 가명정보 처리 가이드라인과 보건의료분야 가이드라인은 가명처리만을 다루고 있다.

적용 우선순위에 있어서는 모든 가이드라인들이 관계 법령과 타 법령을 우선 적용하며, 법령에 특별한 규정이 없는 경우 분야별 가이드라인을 우선 적용하고, 가명정보 처리 가이드라인을 참고한다. 또한, 가명정보 처리 가이드라인은 가이드라인 미준수를 사유로 법적인 처벌을 받지 않음을 명시하고 있다.

〈표 2-2〉 국내 개인정보 비식별화 관련 가이드라인 비교

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료 데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
작성 주체	금융위원회 금융감독원	개인정보 보호위원회	보건복지부 개인정보 보호위원회	교육부 개인정보 보호위원회	통계청
발간일 (제정일)	2020.08.06.	2020.09.24.	2020.09.25.	2020.11.26.	2023.05.30.
관련법령	신용정보법	개인정보보호법			통계법
분야	금융	일반	보건의료	교육	일반
적용 대상	개인 신용정보	개인정보 보호법상의 가명정보 처리	보건의료 데이터를 처리하는 모든 개인 정보 처리자	교육기관의 개인정보 처리자, 교육기관으로부터 정보를 제공받은 자	통계작성기관 업무 절차
적용 대상 데이터	개인신용	개인정보, 개인식별 가능 정보	보건의료 데이터 (보건의료 기본법 제3조 제6호에 따른 보건의료 정보로서 광 또는 전자적 방식으로 처리될 수 있는 것)	교육분야 개인정보	국가 승인통계
적용 우선순위	우선 적용	-	우선 적용	우선 적용	우선 적용
가명처리 목적	1. 통계 작성, 2. 연구, 3. 공익적 기록보존	1. 통계 작성, 2. 과학적 연구, 3. 공익적 기록보존			1. 통계 작성 및 공표 2. 이용자 요청에 따른 제공

<table>
<tr><th></th><th>금융분야
가명·익명
처리 안내서</th><th>가명정보
처리
가이드라인</th><th>보건의료
데이터 활용
가이드라인</th><th>교육분야
가명·익명
정보 처리
가이드라인</th><th>통계 작성
및 통계자료
제공을 위한
비식별화
가이드라인</th></tr>
<tr><td>포괄
범위</td><td>가명처리
익명처리</td><td>가명처리</td><td>가명처리</td><td>가명처리
익명처리</td><td>비식별화
(가명처리
+익명처리)</td></tr>
<tr><td>처리기법</td><td>포함</td><td>포함</td><td>포함</td><td>미포함</td><td>포함</td></tr>
<tr><td>처리기법
분류</td><td colspan="3">ISO/IEC 20889 분류와 유사</td><td>-</td><td>별도 분류</td></tr>
<tr><td>가명처리
절차</td><td colspan="4">1. 사전 준비 → 2. 위험성 검토 → 3. 가명처리 → 4. 적정성 검토 → 5. 사후 관리</td><td rowspan="3">통계작성기관
업무 수행
절차를 따름</td></tr>
<tr><td>결합
절차</td><td>1. 가명처리
및 결합키
생성 후
결합신청
2, 정보 집합
물 결합
3. 가명처리,
익명처리
및 적정성
평가
4. 결합정보
전달</td><td colspan="3">1. 가명처리 절차
2. 결합신청
3. 결합
4. 추가 처리
5. 반출심사
6. 반출 및 활용
7. 사후관리</td></tr>
<tr><td>익명처리
절차</td><td>1. 익명처리
2. 적정성
평가</td><td>-</td><td>-</td><td>1. 사전준비
2. 익명처리
3. 적정성 검토
4. 안전한 관리</td></tr>
<tr><td>가명처리
적정성
평가</td><td>필요 시
실시
결합 시
필수</td><td>적정성
검토위원회
(최소 3인
이상 권고)</td><td>데이터
심의위원회
(구성 요건
有)</td><td>적정성
검토위원회
(구성 요건
有)</td><td>-</td></tr>
<tr><td>익명처리
적정성
평가</td><td>금융위원회
데이터전문
기관에 위탁
가능
(적정성평가
위원회 구성)</td><td>-</td><td>-</td><td>적정성
검토위원회
(구성 요건
有)</td><td>-</td></tr>
<tr><td>결합</td><td>1. 내부결합
: 자체결합</td><td>1. 내부결합
: 자체결합</td><td>1. 내부결합
: 자체결합</td><td>1. 내부결합
: 자체결합</td><td>-</td></tr>
</table>

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료 데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
	2. 외부결합 : 데이터 전문기관 이용	2. 외부결합 : 결합 전문기관 이용	2. 외부결합 : 결합 전문기관 이용	2. 외부결합 : 결합 전문기관 이용	
반출심의	결합 적정성 평가	반출심사 위원회 (구성 요건 有)	반출심사 위원회 (구성 요건 有)	-	1. 제공 시 : 통계자료 제공심의회 2. 반출 시 : 자료반출 평가위원회

출처: 개인정보보호위원회. (2024. 2.). 가명정보 처리 가이드라인;
교육부, 개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인;
보건복지부, 개인정보보호위원회. (2024. 1.). 보건의료데이터 활용 가이드라인;
금융위원회, 금융감독원. (2022. 1.) 금융분야 가명·익명처리 안내서;
통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. 통계데이터분과위원회. 제2023-07호.
각 가이드라인을 참고하여 저자 작성.

3. 5개 가이드라인의 비식별화, 가명처리 및 익명처리에 관한 용어 정의

비식별화, 가명처리 및 익명처리에 관한 용어 정의를 살펴보면, 익명처리를 포괄하지 않는 가이드라인들에서도 익명처리에 대한 용어 정의는 하고 있지 않으나 익명 정보에 대한 용어 정의는 내리고 있음을 알 수 있다. 가명·익명 처리의 대상은 모두 개인(신용)정보이며, 가명·익명 처리된 가명·익명 정보는 개인을 알아볼 수 없게 처리된(비식별화된) 정보로 이를 위한 처리 기법은 동일하다. 가명 정보는 추가 정보가 있다면 식별이 가능해지고, 익명 정보는 추가 정보가 있더라도 식별 불가능한 정보이

다. 특징적인 것은 금융분야 안내서의 경우 익명 정보는 목적 제한 없이 자유롭게 활용할 수 있다고 명시하고 있다.

또한, 통계청 가이드라인은 가명처리와 익명처리 모두를 포괄하는 '비식별화'를 포괄 범위로 하고 있다. 통계청 가이드라인에서 가명처리와 익명처리의 정의는 모두 개인정보법과 신용정보법의 정의를 인용하여 제시하고 있으며, 참고 법령이나 규정 외에도 구체적인 예시를 통해 용어의 개념을 설명하고 있다. 추가로 개념 정리를 통해 가명·익명 처리 및 가명·익명 정보에 대해 구체적으로 설명하고 있다.

〈표 2-3〉 국내 개인정보 비식별화 관련 가이드라인에서의 용어 정의: 비식별화, 가명처리 및 익명처리

용어	정의	해당 가이드라인
비식별화	식별 정보를 암호화하거나 삭제, 총계처리, 재현자료 생성 등을 통해 특정의 개인이나 법인 또는 단체 등을 식별할 수 없도록 하는 조치로, 가명·익명 처리를 모두 포괄하는 개념	통계청 가이드라인
가명처리	개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것 (개인정보보호법 제2조 1호의 2)	가명정보 처리 가이드라인 보건의료분야 가이드라인 교육분야 가이드라인 통계청 가이드라인
	추가 정보(예: 가명정보와 기존 식별자를 연결하는 매핑테이블 등)를 사용하지 아니하고는 특정 개인인 신용정보 주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말하는데, 그 처리 결과가 ① 어떤 신용정보 주체와 다른 신용정보 주체가 구별되는 경우 ② 하나의 정보 집합물에서나 서로 다른 둘 이상의 정보 집합물 간에 어떤 신용정보 주체에 관한 둘 이상의 정보가 연계되거나 연동되는 경우 ③ 위와 유사한 경우로서 대통령령으로 정한 경우의 어느 하나에 해당하는 경우로서 법령에 따라 그 추가 정보를 분리하는 등 특정 개인인 신용정보 주체를 알아볼 수 없도록 개인신용정보를 처리한 경우를 포함한다(신용정보법 제2호제15호).	금융분야 안내서
가명정보	가명처리한 개인신용정보를 말한다. (신용정보법 제2조 제16호)	금융분야 안내서

용어	정의	해당 가이드라인
	가명처리를 거쳐 생성된 정보로서 그 자체로는 특정 개인을 알아볼 수 없도록 처리한 정보	가명정보 처리 가이드라인 보건의료분야 가이드라인
	개인정보를 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보 ※ 가명정보도 개인정보의 범주에 포함	교육분야 가이드라인
	가명처리한 정보	통계청 가이드라인
	(참고) 개인식별 정보 또는 개인식별 가능 정보를 제1호의 2에 따라 가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보(이하 "가명정보"라 한다)	개인정보보호법 제 2조 1호 다목
익명처리	데이터 값 삭제, 가명처리, 총계처리, 범주화 등의 방법으로 개인신용정보의 전부 또는 일부를 삭제하거나 대체함으로써 더 이상 특정 개인인 신용정보 주체를 알아볼 수 없도록 개인신용정보를 처리하는 것을 말한다. (신용정보법 제2조 제17호)	금융분야 안내서
	개인정보의 전부 또는 일부를 데이터 값 삭제, 가명처리, 총계처리, 범주화 등 다양한 기술을 적용함으로써 더 이상 특정 개인을 알아볼 수 없도록 익명 정보로 처리하는 것	교육분야 가이드라인
	시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없도록 처리하는 것	통계청 가이드라인
익명정보	시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보	금융분야 안내서 가명정보 처리 가이드라인 보건의료분야 가이드라인 교육분야 가이드라인
	익명처리한 정보	통계청 가이드라인
	(참고) 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보	개인정보보호법 제 58조의 2

출처: 개인정보보호위원회. (2024. 2.). 가명정보 처리 가이드라인;
개인정보보호법, 법률 제19234호 (2024);
교육부, 개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인;
보건복지부, 개인정보보호위원회. (2024. 1.). 보건의료데이터 활용 가이드라인;
금융위원회, 금융감독원. (2022. 1.) 금융분야 가명·익명처리 안내서;
통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. 통계데이터분과위원회. 제2023-07호.
법률 및 각 가이드라인을 참고하여 저자 작성.

4. 5개 가이드라인의 식별자(identifier)와 관련 용어 정의

〈표 2-3〉에서는 가이드라인별로 정의하고 있는 용어의 목록만 제시하였는데 식별자(identifier)와 관련한 가이드라인별 용어 정의를 〈표 2-4〉에 제시하였다. 데이터 주체란 데이터에 포함된 정보의 주체를 의미하며 데이터 주체는 개인이 될 수도 있고 법인 혹은 기업과 같은 단체가 될 수도 있다. 식별 정보란 데이터 주체 혹은 데이터 주체가 가지는 속성 중에서 데이터로부터 식별(identification)할 수 있는 정보를 의미한다. 본 보고서에서는 개인식별 가능 정보(준식별자) 용어를 주로 사용하였다.

〈표 2-4〉 국내 개인정보 비식별화 관련 가이드라인에서의 용어 정의:
개인신용정보, 개인정보, 식별(정보), 개인식별 정보(식별자), (개인)식별 가능 정보(준식별자 또는 간접 식별자)

용어	정의	해당 가이드라인
개인 신용정보	기업 및 법인에 관한 정보를 제외한 살아 있는 개인에 관한 신용정보로서 다음의 어느 하나에 해당하는 정보를 말한다. (「신용정보법」 제2조 제2호) 1) 해당 정보의 성명, 주민등록번호 및 영상 등을 통하여 특정 개인을 알아볼 수 있는 정보 2) 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 특정 개인을 알아볼 수 있는 정보	금융분야 안내서
식별자	주민등록번호, 이메일 주소, 휴대전화 번호 등과 같이 그 자체로 특정 개인을 직접 식별하는 용도로 사용하는 속성을 말한다.	금융분야 안내서
개인식별 가능 정보	연령, 성별, 거주지역, 국적 등과 같이 해당 정보만으로는 직접적으로 특정 개인을 식별할 수 없지만, 다른 속성과 결합하여 특정 개인의 신원을 전부 또는 일부를 드러낼 수 있는 속성을 말한다.	금융분야 안내서
식별	단독으로 또는 두 개 이상의 속성을 결합하는 등의 방법으로 개인을 알아볼 수 있도록 처리	금융분야 안내서

용어	정의	해당 가이드라인
	하는 것을 말한다.	
개인정보	살아 있는 개인에 관한 정보로서 다음의 정보를 포함함 - 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보 - 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보 ※ 이 경우 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등 개인을 알아보는 데 소요되는 시간, 비용, 기술 등을 합리적으로 고려하여야 함 - 가명처리를 거쳐 생성된 정보로서 그 자체로는 특정 개인을 알아볼 수 없도록 처리한 정보(이하 '가명정보'라 함) ※ 개인정보에 대한 판단 기준은 개인정보 처리자가 보유한 정보 또는 접근 가능한 권한 등 개인정보 처리 상황에 따라 다르게 판단되어야 함	가명정보 처리 가이드라인 보건의료분야 가이드라인 교육분야 가이드라인 통계청 가이드라인
식별 정보	성명, 고유 식별 정보(주민등록번호, 여권번호, 외국인등록번호, 운전면허번호), (개인)휴대전화 번호, (개인)전자우편 주소, 의료기록번호, 건강보험번호 등 식별을 목적으로 생성된 정보	가명정보 처리 가이드라인
	전체 또는 특정 인구 집단 내에서 개인을 고유하게 구별하기 위해 부여한 기호 또는 번호, 기관 내·외에서 개인 간 상호 구별을 위해 부여한 번호, 기호 등을 통칭 ※ (예시) 개인정보 보호 법령 상 고유식별번호(주민등록번호, 여권번호, 운전면허번호, 외국인등록번호), 보험가입자번호, 환자번호, 이름, 웹사이트의 ID, 사원번호 등	보건의료분야 가이드라인
	특정 개인이나 법인 또는 단체 등을 다른 개체와 구별하여 알아볼 수 있는 항목으로 고유·개체·보조 식별 정보 등으로 구분	통계청 가이드라인
개인 식별 정보 (식별자)	고유 식별 정보, 이메일 주소, 휴대전화 번호 등과 같이 그 자체로 특정 개인을 직접 식별하는 용도로 사용하는 정보	교육분야 가이드라인
고유 식별 정보	개인을 고유하게 구별하기 위하여 부여된 식별정보로서 주민등록번호, 여권번호, 운전면허번호, 외국인등록번호	통계청 가이드라인
식별 가능	성별, 연령(나이), 거주 지역, 국적, 직업, 위치	가명정보 처리 가이드라인

용어	정의	해당 가이드라인
정보	정보 등 개인정보 처리자의 입장에서 개인을 알아볼 수 있는* 정보 * 개인을 '알아볼 수 있는지'는 해당 정보를 처리하는 자(정보의 제공 관계에 있어서는 제공받는 자를 포함)를 기준으로 판단하여야 함	
개체 식별 정보 (직접 식별자)	자료 주체를 직접적이고 명확하게 나타내는 항목으로서 성명(사업체 대표자명), 사업자등록번호, 전화번호 등 자료 주체를 직접적으로 알아볼 수 있는 정보	통계청 가이드라인
개인식별 가능 정보 (준식별자 또는 간접 식별자)	연령, 성별, 거주 지역, 국적 등과 같이 해당 정보만으로는 직접적으로 특정 개인을 식별할 수 없지만, 다른 정보와 결합하여 특정 개인을 전부 또는 일부 식별할 수 있는 정보 ※ 개인식별 가능 정보는 개인식별 가능성이 높고 낮음에 따라 가명처리 및 익명처리 수준 등을 달리할 수 있으며, 해당 속성의 개인식별 가능성 여부는 구체적인 사례에 따라 달리 판단 해야 함	교육분야 가이드라인
보조 식별 정보 (간접 식별자)	자료 주체에 대한 직접적인 정보는 없지만 다른 식별 정보와 결합하여서 자료 주체를 식별할 가능성을 증가시키는 정보	통계청 가이드라인

주: 식별 정보 ⊃ 고유 식별 정보, 개체 식별 정보, 보조 식별 정보
출처: 개인정보보호위원회. (2024. 2.). 가명정보 처리 가이드라인;
교육부, 개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인;
보건복지부, 개인정보보호위원회. (2024. 1.). 보건의료데이터 활용 가이드라인;
금융위원회, 금융감독원. (2022. 1.) 금융분야 가명·익명처리 안내서;
통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. 통계데이터분과위원회. 제2023-07호.
각 가이드라인을 참고하여 저자 작성.

5. 5개 가이드라인의 비식별화 처리 기법

교육분야 가이드라인을 제외한 4개 가이드라인에서는 모두 처리 기법에 대한 설명을 제시하고 있다. 4개 가이드라인 중 통계청 가이드라인을 제외한 3개 가이드라인에서는 처리 기법을 국제 표준인 ISO/IEC 20889 분류와 유사하게 개인정보 삭제 기술과 개인정보 일부 또는 전부 대체 기

술로 구분하여 나열하고 있다(재현 데이터, 차등 정보보호 포함). 금융분야 안내서의 경우 k-익명성 모델, l-다양성 모델, t-근접성 모델과 차등 정보보호 기법도 포함하고 있으며, 통계청 가이드라인은 재현자료, 일방향 암호화 방법도 기술하고 있다.

통계청 가이드라인은 처리 기법을 비식별화 방법으로 제시하고 있으며 1) 비식별화 평가 측도, 2) 마이크로데이터 비식별화 방법과 3) 매크로데이터 비식별화 방법으로 구분하여 구체적인 사례와 함께 설명하고 있다. 데이터를 수집하고 활용하는 과정의 측면에서 데이터의 유형을 분류한다면 크게 원시 데이터(raw-data), 마이크로데이터(micro-data), 매크로데이터(macro-data)로 분류할 수 있다. 원시 데이터는 수집된 데이터의 원형을 의미하며 분석용 데이터를 만드는 원천이 되는 데이터이다. 마이크로데이터는 원시 데이터의 수집 과정에서 발생할 수 있는 다양한 오류를 수정하고 분석의 대상 혹은 주체의 정보가 하나의 레코드를 구성하도록 가공하여 정리한 데이터이다. 매크로데이터는 마이크로데이터를 분석 목적에 따라 최대·최소, 합, 평균, 그룹화 등의 기법으로 가공하여 일종의 집계표 형태로 재구성한 데이터를 의미한다.

또한, 자료 특성에 따라 범주형/준연속형/연속형 항목으로 변수를 구분하고 있어 처리 기법별로 적용할 수 있는 항목을 구분하였다.

변수란 데이터를 구성하는 조사 항목을 의미하며 통상의 테이블 형태로 정리된 데이터에서 열 방향의 조사 항목에 해당한다. 변수의 속성은 크게 범주형과 연속형으로 구분할 수 있으며 변수의 속성에 따라 적용할 수 있는 비식별화 방법론이 다를 수 있기 때문에 변수의 속성을 엄밀히 구분할 필요가 있다. 범주형 변수는 성별, 국적, 행정구역 등 가질 수 있는 값이 몇 개의 범주 혹은 속성으로 구분되는 변수로 덧셈과 곱셈 등 산술적 연산이 불가능하며 대소를 구분할 수도 없다. 연속형 변수는 몸무

게, 면적, 소득, 매출액 등 가질 수 있는 값이 실수인 변수로 산술적 연산과 대소 구분이 가능한 변수이다.

마이크로데이터 비식별화 방법은 다른 가이드라인들과 마찬가지로 국제 표준인 ISO/IEC 20889 분류를 따르나 실무적으로 활용하는 방법 위주로 간소화하였다. 처리 기법에 대한 분류만 상이할 뿐 처리 기법은 모두 동일하게 적용 가능하며 교육분야 가이드라인 역시 마찬가지이다.

〈표 2-5〉 (참고) 통계청 비식별화 방법

구분		방법
비식별화 평가 측도	노출 위험 평가 측도	1. 자료 주체의 유일성(표본과 모집단에서의 유일성) 2. 특이정보 3. k-익명성 4. l-다양성 5. 재식별위험 평가측도(연계확률, 초모집단 모형)
	정보손실 평가 측도	1. 범주형 항목(비율) 2. 연속형 항목(주요 집계 통계량, IL1s 등)
마이크로 데이터 비식별화 방법	구조적 방법	표본추출(자료 전체), 총계처리(연속형 항목)
	삭제 방법	항목 삭제(연속형, 범주형), 레코드(행) 삭제(연속형, 범주형), 국소 삭제(연속형, 범주형)
	일반화 방법	일반 반올림(연속형), 랜덤 반올림(연속형, 범주형), 재범주화(연속형, 범주형), 상하단 범주화(연속형, 범주형)
	임의화 방법	잡음 첨가(연속형), 자료 교환(연속형), 부분 총계(연속형)
매크로데이터 비식별화 방법		셀 감추기, 재범주화, 반올림, 구간값 제공, 특이정보 처리

출처: 통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성을 참고하여 저자 작성.

6. 5개 가이드라인의 가명·익명 처리의 절차와 적정성 평가

가명·익명 처리의 절차와 적정성 평가에 관해서는 마찬가지로 통계청 가이드라인과 그 외 4개 가이드라인이 뚜렷하게 구분된다. 4개 가이드라인의 경우 익명처리 절차는 사전 준비 및 익명처리 후 적정성 평가를 하

는 비교적 단순한 절차를 따르며 가명처리의 경우 준비, 위험성 검토, 가명처리, 적정성 평가, 사후관리의 절차로, 결합의 경우 결합, 추가처리, 반출 및 활용, 사후관리로 구성되어 있어 유사하다.

또한 가명정보 처리 가이드라인, 금융분야 안내서, 교육분야 가이드라인의 경우 반출에 관한 설명을 포함하고 있다. 4개 가이드라인 모두 적정성 검토 또는 위험성 검토를 위해 위원회를 구성 요건에 따라 구성해야 하며, 가이드라인(분야)에 따라 더 강화된 구성 요건이 요구되기도 한다. 예를 들어, 보건의료분야와 교육분야 가이드라인에 따르면 가명처리 적정성 검토 시 외부 전문가를 섭외하여 위원회를 구성해야 한다.

가명정보 결합 절차에 관해서는 내외부 결합을 구분하며, 개인정보 처리자 간의 결합인 외부 결합의 경우 기관의 특성과 해당 법령에 따라 데이터 전문기관 또는 결합 전문기관을 통해 결합한다.

통계청 가이드라인에서의 비식별(가명·익명) 처리의 절차는 통계작성기관의 업무 수행 절차를 따르며 비식별 처리에 대한 적정성 평가는 하지 않는다. 그러나 비식별 처리한 통계 자료를 제공할 때, 통계자료 제공심의회의 심의를 통해 최종적인 자료 제공 방법과 항목을 결정한다. 통계자료 제공심의회의 승인을 받은 제공용 자료는 선정된 공개 범위와 방법에 따라서 이용자에게 제공된다. 통계데이터센터 이용자가 분석 결과를 반출하고자 할 때 반출 요청 자료에 대해 검토하고 비식별화 적용하는 등 분석 결과에 대해 점검을 한 후 필요한 경우에는 반출심의위원회의 심의를 통해 자료를 반출할 수 있다. 이용자의 이용 신청에 따른 신청 이력, 비식별화 및 제공 승인 이력 등을 기록하여 관리한다.

이 외에도 가명정보 처리 가이드라인과 보건의료분야 가이드라인의 의 경우 안전성 확보 조치와 관련하여 정보 주체의 권리 보장에 관한 내용을 포함하고 있으며 자세한 내용은 다음의 표에 기술하였다.

〈표 2-6〉 국내 개인정보 비식별화 관련 가이드라인 정리

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
작성 주체	금융위원회 금융감독원	개인정보보호위원회	보건복지부 개인정보보호위원회	교육부 개인정보보호위원회	통계청
발간일 (제정일)	2020년 8월 6일 제정	2020년 9월 24일 발간	2020년 9월 25일 발간	2020년 11월 26일 발간	2023년 5월 30일 작성
관련 법	신용정보법 개인정보보호법	개인정보보호법	개인정보보호법	개인정보보호법	통계법 개인정보보호법
작성 목적	신용정보회사 등이 개인신용정보를 가명처리 또는 익명처리할 때 참고할 수 있는 사항을 안내	가명정보 활용에 필요한 가명정보 처리 목적, 처리 절차 및 방법, 안전조치에 관한 사항 등을 안내하여 안전한 데이터 활용 환경을 마련	보건의료데이터의 분야·유형·목적별 세부 방법과 절차를 제시하여 현장 혼란을 최소화하고, 자료 오남용 방지 처리 과정 전반에 걸쳐 절차 및 거버넌스, 안전조치, 윤리적 사항 등을 정하여 정보 주체의 권익을 보호하고 안전한 개인정보 처리 도모	교육분야의 가명·익명 정보를 처리하는 과정에서 발생할 수 있는 개인정보 오·남용을 방지하고 안전한 가명·익명 정보 처리 방안을 제시하여 교육정보 활용을 통한 효율적 교육정책 수립뿐만 아니라 교육 기관의 안전한 행정 기반 데이터 처리 및 정보 주체의 권익을 보호할 수 있는 안전성을 지원	통계작성기관이 통계법에 근거하여 국가승인통계를 작성 및 공표할 때와 자료 분석 등 이용자 요청에 따라 통계자료를 제공할 때 필요한 비식별화를 지원하기 위해 작성

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
적용 대상	신용정보법 제2조(정의)에 따른 개인신용정보	개인정보보호법 제3장 제3절 가명정보 처리에 관한 특례에 근거한 가명정보 처리	의료기관, 연구자, 기업, 공공기관, 대학교 등 보건의료데이터를 처리하는 모든 개인정보 처리자	교육기관(교육행정기관, 학교 및 교육부장관의 지도감독을 받는 공공기관 및 단체)의 개인정보 처리자와 교육기관으로부터 정보를 제공받은 자	다음과 같은 절차에 적용됨. 1. 국가승인통계 작성 2. 해당 통계 작성을 위하여 수집·취득 또는 사용한 통계자료를 제공용으로 생성 3. 제공용 통계자료를 통계자료 이용신청자에게 제공 4. 통계데이터센터 등에서 각종 통계자료를 분석한 뒤 반출
가명처리 목적	1. 통계 작성 (상업적 목적 포함) 2. 연구 (산업적 연구 포함) 3. 공익적 기록보존	1. 통계 작성 2. 과학적 연구 3. 공익적 기록보존			통계작성기관이 통계법에 근거하여 국가승인통계를 작성 및 공표할 때와 자료 분석 등 이용자 요청에 따라 통계자료를 제공하기 위해

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
내용	1. 개요 : 용어 정의 및 금융분야 가명·익명 처리 일반 등 2, 가명처리 : 절차, 방법. 행위규칙, 보호조치 기준 3. 익명처리 및 적정성 평가: 방법, 적정성 평가 4. 정보 집합물 결합 : 절차, 외부 정보의 결합, 주기·반복적 결합	1. 개요 : 목적, 적용 대상, 용어 정리 2. 가명처리 및 가명정보의 처리 (참고) 비정형데이터 가명처리 기준 3. 가명정보 결합 및 반출 4. 안전성 확보 조치	1. 개요 : 필요성 및 목적, 관련 근거, 적용 범위, 용어 정리 2. 가명처리 3. 가명정보 결합 및 반출 4. 안전성 확보 조치	1. 개요 : 배경 및 목적, 적용 범위, 용어 정리 2, 가명처리 및 가명정보 처리 3. 익명처리 4. 기타	1. 개요 : 목적, 적용 대상, 비식별화 절차, 용어 정리, 개념 정리 2, 비식별화 방법 : 평가 측도, 마이크로데이터/매크로데이터 비식별화 방법 3. 통계 작성 단계별 비식별화 4. 통계자료 제공 단계별 비식별화 5. 보유 정보의 안전한 관리를 위한 보호 조치
용어 정의 목록	개인신용정보, 개인정보. 속성, 개인식별 가능 정보, 식별자, 개인식별 가능 정보, 식별, 정보 집합물, 결합키, 가명처리, 추가 정보, 가명정보, 익명정보, 익명처리, 결합 대상 정보 집합물, 연결키	개인정보, 가명처리, 개인정보파일, 재식별, 결합키, 익명정보, 추가 정보, 개인정보 처리자, 가명정보 처리 시스템, 결합키 연계 정보, 결합대상정보, 결합정보, 반출정보, 반복결합, 반복결합	개인정보, 개인정보 처리자, 개인정보파일, 처리, 가명처리, 정보 주체, 추가 정보, 결합전문기관, 익명정보, 적정성 검토, 재식별, 식별 위험성, 식별 정보, 통계 작성,	개인정보, 개인정보 처리자, 개인식별 정보(식별자), 개인식별 가능 정보(준식별자 또는 간접 식별자), 가명처리, 가명정보, 가명정보 처리자, 가명정보 취급자,	- 식별 정보 자료 주체, 항목/속성, 고유 식별 정보, 개체 식별 정보(직접 식별자), 보조 식별 정보(간접 식별자) - 비식별화 가명처리(가명정보), 익명처리(익명정보) - 각종 통계자료

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
		연결정보, 결합신청자, 결합전문기관, 결합키관리기관, 적정성 검토, 반출 심사, 비정형데이터	공익적 기록보존, 과학적 연구, 결합키, 결합키연계정보, 결합대상정보, 결합정보, 반출정보, 반복결합, 반복결합연결정보, 결합신청자, 결합키관리기관, 폐쇄분석환경, 인간대상연구, 인체유래물, 인체유래물 등, 기관생명윤리위원회, 기관보건의료데이터심의위원회	가명정보처리시스템, 추가 정보, 특이정보, 다른 정보, 익명정보, 익명처리, 익명정보 처리자, 적합성 검토, 적정성 검토, 재식별, 개인정보파일, 결합키, 결합키연계정보, 결합대상 정보, 결합정보, 반출정보, 결합신청자, 결합전문기관, 동질집합, 결합키관리기관	통계자료, 통계기초자료 (microdata), 행정통계자료. 제공용 통계자료, 인가용 통계자료, 공공용 통계자료 - 통계기초자료 제공 서비스 마이크로데이터통합 서비스시스템, 원격접근서비스, 마이크로데이터이용센터, 통계데이터센터 - 통계 작성 기관 국가승인통계, 자료생산부서. 자료관리부서. 자료이관기관 - 범주형 항목, 준연속형 항목, 연속형 항목 - 개체식별 정보, 보조식별 정보, 민감 정보 - 신분 노출(재식별), 속성 노출

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
처리 절차	- 가명처리 1. 사전 준비 2. 가명처리 1) 위험도 측정 2) 가명처리 수준 결정 3) 가명처리 3. 적정성 검토 및 추가 처리 4. 활용 및 사후관리 1) 가명정보 이용·제공·결합 및 사후관리 2) 가명정보 삭제 - 익명처리 1. 익명처리 2. 적정성 평가	- 가명처리 1. 사전 준비 2. 위험성 검토 3. 가명처리 4. 적정성 검토 5. 안전한 관리 - 결합 절차 1. 결합신청 2. 결합 및 추가 처리 3. 반출 및 활용 4. 안전한 관리		- 가명처리 1. 사전 준비 2. 위험성 검토 3. 가명처리 4. 적정성 검토 5. 안전한 관리 - 익명처리 1. 사전준비 2. 익명처리 1) 위험성 검토 2) 익명처리 3. 적정성 검토 4. 안전한 관리	- 통계 작성 단계별 1. 기획 2. 자료 수집 3. 내검 및 정제 4. 통계 작성 5. 공표 및 보관 - 통계자료 제공 단계별 1. 제공용 통계자료 생성 2. 이용 신청에 따른 통계자료 제공 3. 통계데이터센터 분석 결과 반출
처리 기법	- 삭제 기술: 삭제, 부분 삭제, 행 항목 삭제, 로컬 삭제, 마스킹 - 통계 도구: 총계처리, 부분 총계 - 일반화 기술: 일반 라운딩, 랜덤	- 삭제 기술: 삭제, 부분 삭제, 행 항목 삭제, 로컬 삭제, 마스킹 - 통계 도구: 총계처리, 부분 총계 - 일반화(범주화) 기술: 일반 라운딩, 랜덤 라운딩, 제어 라운딩, 상·하단 코딩, 로컬 일반화, 범위 방법, 문자데이터 범주화 - 암호화: 양방향 암호화, 일방향 암호화-암호학적 해시함수, 순서보존 암호화, 형태보존 암호화,		없음	1. 비식별화 평가 측도 - 노출 위험 평가 측도: 자료 주체의 유일성, 특이정보, k-익명성, l-다양성 - 정보손실 측도

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
	라운딩, 제어 라운딩, 상·하단 코딩, 로컬 일반화, 범위 방법, 문자데이터 범주화 - 암호화: 양방향 암호화, 일방향 암호화-암호학적 해시함수, 순서보존 암호화, 형태보존 암호화, 동형 암호화, 다형성 암호화 - 무작위화 기술: 잡음 추가, 순열(치환), 토큰화, (의사)난수생성기 - 프라이버시 보호 모델: k-익명성 모델, l-다양성 모델, t-근접성 모델, 차분 프라이버시 - 기타: 표본추출, 해부화,	동형 암호화, 다형성 암호화 - 무작위화 기술: 잡음 추가, 순열(치환), 토큰화, (의사)난수생성기 - 기타 기술: 표본추출, 해부화, 재현 데이터, 동형비밀분산, 차분프라이버시			2. 마이크로데이터 비식별화 방법 - 구조적 방법: 표본 추출, 총계처리 - 삭제: 항목 삭제, 레코드 삭제, 국소 삭제, 마스킹 - 일반화: 일반 반올림, 랜덤 반올림, 재범주화, 상하단 범주화 - 임의화: 잡음 첨가, 자료교환, 부분 총계 - 재현자료 생성 - 암호화: 일방향 암호화 3. 매크로데이터 비식별화 방법 - 셀 감추기 - 재범주화 - 반올림 - 구간값 제공

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
	재현 데이터, 동형비밀분산				- 특이정보 처리
안전성 관련 조치	- 행위 규칙 1. 추가 정보의 분리 보관 또는 삭제 2. 기술적·관리적·물리적 보안대책 수립·시행 3. 가명처리의 제한 4. 재식별 시 조치 5. 가명처리 기록의 보존 6. 가명처리 관련 사항의 공개 7. 가명정보에 대한 적용 예외 - 보호조치 1. 기술적·물리적 보호조치 2. 관리적 보호조치 3. 보호대책의 준용	- 안전성 확보 조치 1. 관리적 보호조치 2. 기술적 보호조치 3. 물리적 보호조치 4. 정보 주체의 권리보장		- 안전성 확보 조치 1. 관리적 보호조치 2. 기술적 보호조치 3. 물리적 보호조치	- 보유 정보의 안전한 관리를 위한 보호 조치 1. 관리적 보호조치 2. 기술적 보호조치 3. 물리적 보호조치
업무 서식	- 가명처리 기록 - 정보 집합물 결합 신청서	- 통계 작성 계획서 - 과학적 연구 계획서 - 공익적 기록보존	- 가명정보 이용·제공 신청서 - 활용데이터 선정 사유	- 통계 작성 계획서 - 과학적 연구 계획서 - 공익적 기록보존	별도 서식이 아닌 예시 표로 정리

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
	- 익명처리 적정성 평가 신청서 - 정보 집합물 결합 기초 자료 - 익명처리 적정성 평가 기초자료 - 결합정보 관리 환경 및 이행확약서	계획서 - 결합신청서 - 반출신청서 - 내부 관리계획 - 가명정보 이용·제공 신청서 - 개인정보 유형 분류표 - 활용데이터 요구 수준표 - 개인식별 위험성 검토 체크리스트 - 개인식별 위험성 검토 항목별 조치 가이드 - 식별 위험성 검토 결과보고서 - 항목별 가명처리 계획 - 주요 비정형데이터 가명처리 수행 결과 (예시) - 가명정보 처리 기초 자료 명세서 - 가명처리 결과 자체 검증 - 적정성 검토 결과서	및 요구 수준표 - 가명정보에 대한 안전성 확보 조치 계획서 - 가명정보에 대한 안전 조치 의무이행 확약서 - 가명정보 처리 목적 증빙자료 - 식별 위험성 검토 결과보고서 서식 - 항목별 가명처리 계획서 서식 - 가명정보 제공 및 활용 계약서 서식, 부속 합의서 서식 - 비밀유지의무, 이해 상충 서약서 - 가명정보 처리 기초 자료 명세서 - 기관보건의료데이터 심의위원회 검토 결과서(위원용) - 기관보건의료데이터 심의위원회 검토 종합결과서	계획서 - 가명정보 결합 신청서 - 가명정보 반출 신청서 - 위험성 검토, 데이터 구성 관련, 데이터 분포, 우연한 재식별 관련, 가명정보 처리장소 및 형태 관련, 다른 정보 관련, 처리기관의 신뢰도 관련, 재식별 영향도 관련 체크리스트 - 내부관리계획 - 개인정보 처리방침 - 적정성 검토 기초자료 목록 구성표 - 적정성 검토 결과서 (위원용) - 적정성 검토 종합 결과서 - 비밀유지의무 서약서 - 이해상충 공개 서약서 - 가명정보 이용·제공 신청서	

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
		(위원용) - 적정성 검토 종합 결과서 - 비밀유지의무 서약서 - 이해상충 서약서 - 가명정보에 대한 안전조치 이행 확약서 - 가명정보 관리대장 (여러 가명처리 기록 서식) - 추가 정보 관리대장 - 가명 정보 접근 권한 관리대장 - 추가 정보 접근 권한 관리대장 - 가명정보 파기대장 - 추가 정보 파기대장 - 비정형데이터 대상 가명처리 결과에 대한 자체 검증 결과서	- 가명정보 처리 관련 실무 서식 - 가명정보 관리대장, 추가 정보 관리대장 - 가명정보 접근권한 관리대장, 추가 정보 접근 권한 관리대장 - 가명정보 파기 관리대장, 추가 정보 파기 관리대장	- 개인정보 유형 분류표 - 활용 데이터 요구 수준표 - 가명정보 식별 위험성 검토 결과보고서 (일반형) (체크리스트 이용 시) - 가명처리 수준 정의표 - 익명정보 식별 위험성 검토 결과보고서 - 익명처리 수준 정의표 - 가명정보 관리대장 - 익명정보 관리대장 - 가명정보 처리 기초 자료명세서(신청 기관 정보, 데이터 명세) - 원본 데이터 세부 항목별 명세 - 원본 데이터 예시 - 원본 데이터 분포 (가능한 표 또는 그래프 형태와 참조 수치로 표현 권장) - 가명처리된 데이터의	

	금융분야 가명·익명 처리 안내서	가명정보 처리 가이드라인	보건의료데이터 활용 가이드라인	교육분야 가명·익명 정보 처리 가이드라인	통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인
				세부 항목별 명세 - 가명처리된 데이터 예시 - 가명처리된 데이터 분포 - 가명정보 안전조치 이행 확약서 - 교육기관 분류체계별 개인정보 항목별 위험성 목록	

출처: 개인정보보호위원회. (2024. 2.). 가명정보 처리 가이드라인;
교육부, 개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인;
보건복지부, 개인정보보호위원회. (2024. 1.). 보건의료데이터 활용 가이드라인;
금융위원회, 금융감독원. (2022. 1.) 금융분야 가명·익명처리 안내서;
통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. 통계데이터분과위원회. 제2023-07호.
각 가이드라인을 참고하여 저자 작성.

제2절 국외 사례

1. 미국 HIPAA의 PHI 익명화 방법에 대한 지침

1996년 의료 정보 이동 및 책임에 관한 법률(HIPAA)의 개인정보 보호 규칙에 따라 비식별화를 달성하기 위한 가이드로 2012년 11월 발간된 Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act(HIPAA) Privacy Rule(이하 'Deid 지침')에서는 보호 대상 정보에 대해 비식별 처리를 하는 방법으로 다음과 같은 두 가지 방안을 제시하고 있다. 보호 대상 개인정보는 Protected health information(PHI)로 기술하고 있으며 개인정보 중 의료분야의 개인정보로 한정하고 있다.

1) Expert Determination method: 전문가 결정 방식

일반적으로 인정되는 통계 및 과학적 원칙과 정보를 개별적으로 식별할 수 없도록 만드는 방법에 대한 적절한 지식과 경험을 갖춘 전문가가, 데이터의 수신자가 해당 정보를 이용 가능한 다른 정보와 결합하여 해당 정보의 주체인 개인에 대한 식별 위험이 매우 적다고 판단하며, 그 판단에 대한 분석 방법 및 결과를 문서화한다.

2) Safe Harbor method: 세이프 하버 방식

개인에 대한 다음 18가지의 식별 정보를 모두 제거하고 사용하는 방식이다. 단, 데이터의 사용자가 이 정보를 단독으로 사용하거나 다른 정보와 함께 사용하여 정보 주체인 개인을 식별할 수 있는 실체적인 지식을 갖고 있지 않다는 것을 전제로 하고 있다.

그러나 k-익명성을 개발한 하버드 대학의 Latanya Sweeney는 이러한 세이프 하버 방식으로 처리된 데이터의 재식별률을 0.04%로 예측(Sweeney, 2000)하였으며 2017년 후속 연구를 통해 인구통계학적 필드에 추가적인 필드를 더 사용하는 경우 재식별률이 최대 28%까지 올라갈 수 있다고 주장(Sweeney et al., 2017)하였다. 이에, 현재의 세이프 하버 방식을 적용한 대부분의 데이터 공개의 경우 완전 공개의 방식보다는 제한된 공개의 방식으로 사용하는 사례가 대부분이다.

〈표 2-7〉 세이프 하버 방식에서의 개인식별 정보

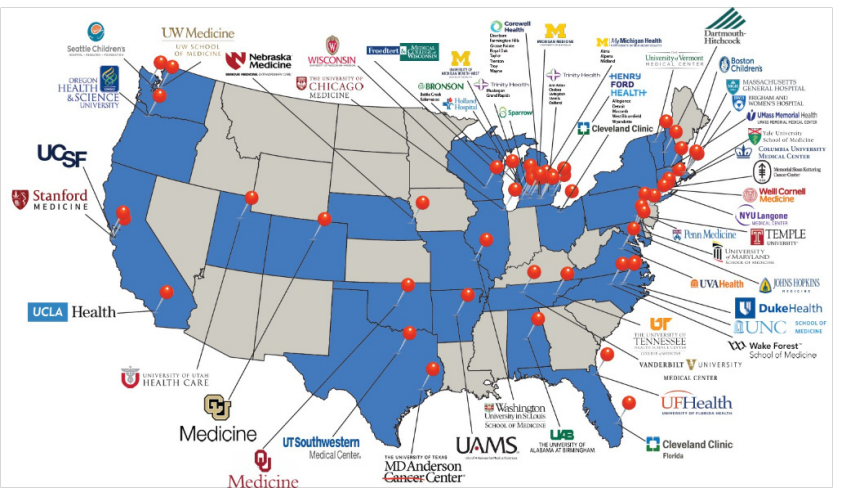
1) 이름 2) 주소 정보 (20,000명 이상의 주소 정보를 사용할 수 있음) 3) 날짜 정보 (년도 단위의 날짜를 사용할 수 있음) 4) 전화번호 5) 팩스번호 6) 이메일 주소 7) 사회보장번호 8) 의료기록번호	9) 건강보험번호 10) 계좌번호 11) 자격취득번호 12) 자동차번호(차량식별번호, 등록번호 등) 13) 각종 장비 식별번호 14) 인터넷 주소(URL 정보) 15) IP 주소 16) 생체정보(지문, 음성 등) 17) 전체 얼굴 사진 및 유사 이미지 18) 기타 특이한 식별 번호 또는 코드

출처: Portability, I., & Act, A. (2012). Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. Washington DC: Human Health Services. pp. 7-8.

세이프 하버 방식을 적용한 사례 중 가장 대표적인 것으로는 미국의 Multicenter Perioperative Outcomes Group(MPOG) 컨소시엄의 사례(https://mpog.org/)가 있다. 수술 환자의 수술 전, 후 치료를 개선하기 위해 2010년 설립된 컨소시엄으로 2024년 현재 70여 개 이상의 의료기관이 참여하고 있다. 참여하는 의료기관에서 수집된 입원 전자의무기록(EHR)은 HIPAA Privacy Rule에 따라 각 기관에서 PHI 제거 등 비식별 처리된 후 미시간 대학교의 중앙 데이터베이스에 저장된다. 이후

MPOG 데이터 사용 승인 후 데이터 신청자에게 제공한다. MPOG 데이터는 다양한 의료기관에서 수집된 수술, 마취, 투약 등의 결과를 분석하여 의료질을 향상하고 수술 전 후 치료의 개선을 하는 것을 목표로 하고 있다.

[그림 2-1] MPOG 가입 병원



출처: Multicenter Perioperative Outcomes Group(MPOG). (2024). 2024. 10. 3. 인출. https://mpog.org/memberhospitals/

2019-2022 표준화 데이터 파일 기준으로 소아(0~17세 환자) 및 성인 사례정보를 제공하고 있다. 〈표 2-8〉과 같은 변수를 제공하고 있다. 비식별 처리 항목은 HIPAA Privacy Rule의 세이프 하버 방식의 비식별 처리 18가지 항목을 포함하여 식별자를 제거한다.

〈표 2-8〉 MPOG 제공 데이터 변수

변수명	설명
기본 환자 정보	- 기본 정보: 나이, 성별, 인종, 민족 - 신체 지표: 체중, 키, BMI(Body Mass Index), 흡연 여부
기관 정보	- 비식별화된 기관 ID, 의과대학 소속 여부 및 병상 규모
수술 전 상태 (Preoperative Conditions)	- 기존 병력: 환자가 가진 기존의 건강 문제 및 병력 (예: 고혈압, 당뇨병, 심장병) - 동반 질환: 현재의 동반 질환 리스트 - ASA 분류: American Society of Anesthesiologists(ASA) Physical Status Classification System 점수 - 수술 전 평가: 혈액 검사 결과, 심전도, 영상 검사 결과 등
수술 및 마취 정보 (Surgical and Anesthesia Details)	- 수술 정보: 수술의 유형 및 명칭, 수술의 주요 목적 및 부위 - 마취 정보: 사용된 마취 유형(전신, 국소, 척추 마취 등) - 시간 정보: 마취 시작 및 종료 시간, 수술 시작 및 종료 시간
수술 중 관리 (Intraoperative Management)	- 모니터링 데이터: 심박수, 혈압, 호흡수, 산소포화도 등 - 투여 약물: 마취제, 진통제, 근이완제, 항생제 등 - 액체 관리: 수액, 혈액 제품, 전해질 투여량 및 시간 - 수술 중 이벤트: 합병증 발생, 혈액 손실량, 기계적 인공환기 사용 여부 등
수술 후 회복 (Postoperative Recovery)	- 회복실 정보: PACU(Post-Anesthesia Care Unit) 체류 시간 회복실에서의 생체 신호 - 합병증: 감염, 출혈, 폐렴 등 주요 수술 후 합병증 - 통증 관리: 통증 정도, 투여된 진통제 종류 및 용량 - 환자 이동: 병실 이동 시간 및 장소, 중환자실(ICU) 이동 여부
임상 결과 (Clinical Outcomes)	- 단기 결과: 병원 내 사망률, 재입원율, 병원 체류 기간 - 장기 결과: 환자의 기능 회복 정도, 장기적인 건강 상태 평가 - 품질 지표: 수술 성공률, 환자 만족도 조사 결과
기타 데이터 (Miscellaneous Data)	- 비용 및 자원 사용: 수술 및 마취 비용, 사용된 의료 자원 및 장비 - 환자 만족도: 환자 만족도 설문 조사 결과 및 분석

출처: Multicenter Perioperative Outcomes Group(MPOG). (2022.12.22.). Standardized Data File – User Guide. pp.6-13을 참고하고 전문가 자문을 거쳐 저자 작성.

〈표 2-9〉 MPOG 비식별 처리 항목

항목	설명
이름	환자 및 가족 구성원, 의료 제공자 및 기타 개인의 이름
주소	주 이하의 모든 지리적 단위(예: 주소, 도시, 우편번호 등)를 제거하며, 초기 세 자리 우편번호가 20,000명 이상을 포함하지 않는 경우 000으로 대체
날짜 정보	연도를 제외한 모든 날짜(예, 생년월일, 입원일, 퇴원일, 사망일)을 제거하며, 90세 이상의 연령에 대해 모든 요소를 단일 범주로 묶음

항목	설명
전화번호	개인의 모든 연락처 번호
팩스번호	개인의 모든 팩스 번호
이메일 주소	개인의 이메일 주소
사회보장번호(SSN)	
의료기록번호	
건강보험번호	건강보험번호 및 가입자 번호
계좌번호	은행 계좌번호 등
인증 번호	라이선스, 인증 및 기타 고유 식별 번호
차량 식별자 및 일련번호	차량 식별 번호 및 차량 등록 번호
디바이스 식별번호	의료기기 및 기타 장치의 일련번호
웹주소(URL)	개인식별이 가능한 모든 웹 주소
IP 주소	
생체 인식 식별자	지문, 망막 스캔, 음성 인식, 유전자 정보 등
얼굴 사진 및 이와 유사한 이미지	얼굴 식별이 가능한 사진 및 이미지
고유 식별자	모든 고유 식별 번호, 코드 또는 속성

출처: Multicenter Perioperative Outcomes Group(MPOG). (2022.12.22.). Standardized Data File – User Guide. pp.6-13을 참고하고 전문가 자문을 거쳐 저자 작성.

〈표 2-10〉 MPOG 추가적 보호조치: 서버의 보호장치

구분	설명
데이터 접근 권한 관리 (Access Control Management)	- 사용자 역할 기반 접근 통제(Role-Based Access Control: RBAC): 사용자(예, 연구자, 관리자, 임상연구 코디네이터 등)에게 필요한 최소한의 데이터 접근 권한만을 부여 - 다단계 인증(Multi-Factor Authentication): 사용자 로그인 시 다단계 인증 요구(예, 비밀번호 외에도 추가적인 인증 수단(모바일 인증 코드) 포함
암호화	- 데이터 전송 암호화: TLS(Transport Layer Security) 프로토콜을 사용하여 네트워크를 통해 전송되는 데이터를 암호화 - 데이터 저장 암호화: AES(Advanced Encryption Standard) 같은 고강도 암호화 알고리즘을 사용하여 데이터베이스에 저장된 데이터를 암호화
Citrix 솔루션	- VPN 접속: Citrix 솔루션을 이용하여 원격으로 MPOG 애플리케이션 사용 가능하며 미시간 대학의 시스템에 VPN으로 연결 - 비밀번호: 미시간 대학의 1레벨 및 2레벨의 비밀번호 사용
통계 가상 서버	- 대부분의 통계 프로그램이 설치된 가상 서버 - 비밀번호: 미시간 대학의 1레벨 및 2레벨의 비밀번호 사용

출처: Multicenter Perioperative Outcomes Group(MPOG). (2024). Security Guidelines for Users. https://mpog.org/securityguidelinesusers/. 2024. 10. 3. 인출을 참고하고 전문가 자문을 거쳐 저자 작성.

〈표 2-11〉 MPOG 데이터 접근 및 사용을 위한 절차

단계	설명
1단계: 연구 제안서 작성 및 제출	연구계획을 MPOG 연구 제안서 양식에 따라 작성
2단계: IRB 심의 및 승인	연구자 소속 기관으로부터 연구계획 제안서 IRB 승인(IRB 승인 후 MPOG 데이터 신청 가능)
3단계: MPOG 데이터검토위원회 심의 및 승인	연구 범위, 데이터 접근 권한, 연구 윤리 준수 조건 등 검토
4단계: 데이터 사용 계약 체결	데이터 사용 조건, 보안 조치, 연구 완료 후 데이터 반환 또는 폐기 조건 명시
5단계: 데이터 제공	데이터 사용 계약(DUA, Data Use Agreement)이 체결된 후, MPOG는 연구자에게 데이터를 제공

출처: Multicenter Perioperative Outcomes Group(MPOG). (2024). Research Proposal Process. https://mpog.org/write-a-research-proposal/. 2024. 10. 3. 인출을 참고하고 전문가 자문을 거쳐 저자 작성.

〈표 2-12〉 MPOG 제공 받는 자의 보호조치 요구사항

구분	설명
컴퓨터 보안	- 하드 드라이브 암호화 - 최신 바이러스 백신 프로그램 설치 - 강력한 비밀번호 설정 - 자동 로그인 금지
파일 보안	- PHI가 포함된 모든 파일 암호화
파일 공유	- Mishare(미시간 대학 파일 공유 시스템)을 통해 전송 시 암호화 전송 - 모든 파일은 4일 동안만 검색 가능
일반 조항 및 지침	- PHI가 포함된 파일은 이메일로 전송 금지 - 휴대용 USB 플래시 드라이브나 휴대용 하드 드라이브에 저장 금지 - 공용 워크스테이션에 파일 저장 금지 - 통계 담당자와 파일 공유 금지 - Dorpbox, Google Docs 또는 기타 온라인 사이트에 게시 금지
Windows 시스템의 경우 BitLocker를 적용하여 데이터 암호화	

출처: Multicenter Perioperative Outcomes Group(MPOG). (2024). Security Guidelines for Users. https://mpog.org/securityguidelinesusers/. 2024. 10. 3. 인출을 참고하고 전문가 자문을 거쳐 저자 작성.

2. 일본 차세대의료기반법 가이드라인

일본에서는 2017년 법률 제28호 「의료분야의 연구 개발에 기여하기 위한 익명 가공 의료 정보에 관한 법률」(약칭 '차세대의료기반법')이 시행됨에 따라 「의료분야의 연구 개발에 기여하기 위해 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인」(약칭 '차세대의료기반법 가이드라인')을 일본 내각부, 문부과학성, 후생노동성과 경제산업성이 공동으로 작성하여 2018년 5월에 제정하였다.

차세대의료기반법은 의료분야의 연구 개발에 기여하기 위한 '익명 가공 의료 정보'와 '가명 가공 의료 정보'에 관하여 국가의 책임, 기본 방침의 수립, 익명 가공 의료 정보 작성 사업을 수행하는 자 및 가명 가공 의료 정보 작성 사업을 수행하는 자의 승인, 의료 정보, 익명 가공 의료 정보, 가명 가공 의료 정보 등의 취급에 관한 규제 등에 관한 내용을 정하고 있다. 이 법은 건강·의료에 관한 첨단 연구 개발 및 신산업 창출을 촉진하고, 건강장수 사회의 형성에 기여하는 것을 목적으로 제정하였다. 차세대의료기반법 가이드라인에서는 차세대의료기반법의 목적에 근거해 구체적인 지침을 제시하고 있다.

차세대의료기반법 가이드라인의 적용 대상은 다음과 같다.

1) 익명 가공 의료 정보 작성 사업을 하는 자(법인에 한함)
2) 가명 가공 의료 정보 작성 사업을 실시하는 자(법인에 한함)
3) 작성 사업자의 위탁(2단계 이상에 걸친 위탁 포함)을 받아
   의료 정보 등을 취급하는 사업을 실시하는 자(법인에 한함)
4) 익명 가공 의료 정보 취급 사업자
   (연결 가능 익명 가공 의료 정보 이용자를 포함)
5) 가명 가공 의료 정보 이용 사업을 하는 자(법인에 한함)
6) 의료 정보 취급 사업자

차세대의료기반법에서 의료 정보란 생존 여부와 상관없이 특정 개인의 병력 및 해당 개인의 심신 상태에 관한 정보로서, 심신의 상태를 이유로 해당 개인 또는 그 자손에 대한 부당한 차별, 편견 및 기타 불이익이 발생하지 않도록 그 취급에 특히 배려가 필요한 것으로서, 정령으로 정하는 기록 등을 포함하는 개인에 관한 정보 중 해당 정보에 포함된 성명, 생년월일 및 기타 기록 등으로 특정 개인을 식별할 수 있는 것(다른 정보와 쉽게 대조할 수 있어 특정 개인을 식별할 수 있는 정보를 포함) 또는 개인식별부호가 포함되는 것을 말한다. 의료 정보에는 사망한 개인에 관한 정보도 포함되는 반면, 개인정보보호법에서의 개인정보는 생존하는 개인에 관한 정보이다. 사망한 개인에 관한 정보는 본인에 대한 차별을 일으킬 수 없어 의료 정보에 대해서는 자손에 대한 부당한 차별을 규정하고 있다.

특정 개인의 병력 및 기타 해당 개인의 심신 상태에 관한 정보란 개인의 과거 병력, 가족력, 약 복용 이력, 신체 소견, 검사 결과, 영상 자료, 치료 방침, PHR(Personal Health Record) 등 개인의 심신 상태에 관한 모든 정보를 포함한 것이다. 공간물 등에 의해 공개되고 있는 정보도 포함되며, 암호화 등에 의한 은닉화 여부와는 상관없다.

의료 정보에 해당하는 사례는 다음과 같다.

1) 병원 또는 진료소가 보유한 의료진, 진료 수가 청구서 또는 건강 검진 결과 또는 보건지도의 결과
2) 약국이 보유한 조제 수가 청구서
3) 지방공공단체가 보유한 건강 검진 결과 또는 보건지도 결과 또는 소아 만성 특정 질병 의료비 지급인정신청서
4) 의료보험자(지방공공단체 포함)가 보유한 진료 수가 청구서 또는 조제 수가 청구서 또는 건강 검진 결과 또는 보건지도의 결과
5) 학교 설치자(지방공공단체 포함)가 보유하는 아동, 학생 등에 대한 건강 검진 결과 또는 보건지도의 결과

〈표 2-13〉 의료 정보 분류의 예

분류	정의	예
식별자	개인에게 직접 묶여 있는 정보	성명, 피보험자 번호 등
준식별자	여러 조합으로 개인식별이 가능한 정보	생년월일, 주소, 소속 조직 등
정적 속성	불변성이 높은 정보	성인의 키, 혈액형, 알레르기, 진찰일 등의 날짜, 장애 등의 외관적인 특징에 관한 정보
반정적 속성	일정 기간 동안 불변성이 있는 정보	체중, 질병, 치료, 투약 등의 정보 등
동적 속성	항상 변화하는 정보	검사치, 식사 및 기타 진료에 관한 정보

출처: 일본 내각부·문부과학성·후생노동성·경제산업성. (2024.4.). 의료분야의 연구 개발에 기여하기 위한 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인(차세대의료기반법 가이드라인). p.79.

익명 가공 의료 정보란 의료 정보의 구분에 따라 해당하는 조치를 통해 특정 개인을 식별할 수 없도록 의료 정보를 가공하여 얻은 개인에 관한 정보로서 해당 의료 정보를 복원할 수 없도록 한 것을 말한다.

- 해당 의료 정보에 포함되는 기록 등의 일부를 삭제하는 것
  (해당 부분의 기록 등을 복원할 수 있는 규칙성을 갖지 않는 방법에 따라 다른 기록 등으로 대체하는 것을 포함)
- 해당 의료 정보에 포함되는 개인식별 코드의 전부를 삭제하는 것
  (해당 개인식별 코드를 복원할 수 있는 규칙성을 갖지 않는 방법으로 다른 설명 등으로 대체하는 것을 포함)

또한, 통계 정보는 복수의 정보로부터 공통 요소에 관한 항목을 추출하여 동일한 분류마다 집계하여 얻어진 정보로 집단의 경향 또는 성질 등을 수량적으로 파악하기 위한 정보이다. 통계 정보는 특정 개인과의 대응 관계가 없는 한 법에서 정의하는 개인에 관한 정보에 해당하지 않아 익명·가명 가공 의료 정보를 승인받아 작성하는 사업자는 작성한 통계 정보를 제3자에게 제공할 수 있다.

〈표 2-14〉 익명 가공의 정의

법률 근거	내용
차세대의료기반법 시행규칙 제18조 제1호	특정의 개인을 식별할 수 있는 기록 등의 삭제
차세대의료기반법 시행규칙 제18조 제2호	개인식별 부호의 삭제
차세대의료기반법 시행규칙 제18조 제3호	정보를 상호 연결하는 부호의 삭제
차세대의료기반법 시행규칙 제18조 제4호	특이한 기술 등의 삭제
차세대의료기반법 시행규칙 제18조 제5호	의료 정보 데이터베이스 등의 성질을 고려한 결과에 근거한 기타 조치 1) 항목 삭제/레코드 삭제/셀 삭제, 2) 일반화, 3) 톱(하단) 코딩, 4) micro aggregation(부분 총계), 5) 데이터 교환(스왑), 6) 노이즈(오차) 부가, 7) 의사 데이터 생성

출처: 일본 내각부·문부과학성·후생노동성·경제산업성. (2024.4.). 의료분야의 연구 개발에 기여하기 위한 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인(차세대의료기반법 가이드라인). pp.69-74.

〈표 2-15〉 의료 정보의 분류에 기초한 익명 가공 방법의 예

분류	익명 가공 방법의 예
식별자	- 삭제 또는 다른 기록 등으로의 비가역적인 대체
준식별자	- k-익명성을 충족시키도록 일반화(생년월일→생년, 주소→도·도·부·현 등) 또는 micro-aggregation(부분 총계) * k값은 제공 데이터 세트의 유용성이 허용되는 범위에서 충분히 큰 값으로 하는 것이 바람직함. - 데이터 항목 삭제 수행 - 의료기관 코드 등은 속성(지리, 규모 등)을 부가하여 특정할 수 없는 형태로 코드 변환
정적 속성 반정적 속성	- 익명 가공의 필요 여부를 검토하고 필요한 경우는 상·하단 코딩, 일반화 또는 micro-aggregation(부분 총계)
동적 속성	- 기본적으로 익명 가공이 불필요하지만 필요한 경우는 상·하단 코딩

출처: 일본 내각부·문부과학성·후생노동성·경제산업성. (2024.4.). 의료분야의 연구 개발에 기여하기 위한 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인(차세대의료기반법 가이드라인). p.80.

가명 가공 의료 정보는 의료 정보의 구분에 따라 정하는 조치를 취해 다른 정보와 조합하지 않는 한 특정 개인을 식별하는 것이 불가능하도록 의료 정보를 가공하여 얻은 개인에 관한 정보이다.

- 해당 의료 정보에 포함되는 기록 등의 일부를 삭제하는 것
  (해당 부분의 기록 등을 복원할 수 있는 규칙성을 갖지 않는 방법에 따라 다른 기록 등으로 대체하는 것을 포함)
- 해당 의료 정보에 포함되는 개인식별 코드의 전부를 삭제하는 것
  (해당 개인식별 코드를 복원할 수 있는 규칙성을 갖지 않는 방법으로 다른 설명 등으로 대체하는 것을 포함)

가명 가공 의료 정보의 작성을 위한 의료 정보의 가공의 기준에 따르면 승인된 가명 가공 의료 정보 작성 사업자는 가명 가공 의료 정보를 작성할 때 다른 정보와 조합하지 않는 한 특정 개인을 식별할 수 없도록 하기 위해 필요한 것으로 주무 성령(우리나라의 주무 부령에 해당)으로 정하는 기준에 따라 의료 정보를 가공해야 한다.

〈표 2-16〉 가명 가공의 정의

법률 근거	내용
차세대의료기반법 시행규칙 제33조 제1호	특정의 개인을 식별할 수 있는 기록 등의 삭제
차세대의료기반법 시행규칙 제33조 제2호	개인식별 부호의 삭제
차세대의료기반법 시행규칙 제33조 제3호	부정하게 이용됨으로써 재산적 피해가 생길 우려가 있는 기록 등의 삭제 예: 신용카드 번호, 송금이나 결제 기능이 있는 웹서비스의 로그인 ID의 비밀번호

출처: 일본 내각부·문부과학성·후생노동성·경제산업성. (2024.4.). 의료분야의 연구 개발에 기여하기 위한 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인(차세대의료기반법 가이드라인). pp.132-135.

〈표 2-17〉 특정의 개인을 식별할 수 있는 기록 등의 삭제에서의 가명 가공 방법의 예

항목	가명 가공 방법의 예
성명	삭제 또는 다른 기록 등으로 비가역적인 대체
주소	주소 삭제
생년월일	변경 없음
화상 정보, 영상 정보	- DICOM 화상 등의 속성 정보, 메타 데이터 등의 부수 정보나 화상·영상 중에 포함되는 텍스트 정보에 대해서는 그 내용에 따라 다른 항목과 마찬가지로 처리 - 얼굴 화상, 영상, 입체를 재구성하여 얼굴 화상을 얻을 수 있고, 해당 화상 1개 또는 몇 개를 조합하여 특정 개인을 식별할 수 있는 화상·영상에 대해서는 1개 또는 조합에 의해 특정 개인이 식별되지 않도록 가공을 실시 - 그 외의 화상 정보·영상 정보에 대해서는 단체 또는 조합으로 특정 개인을 식별할 수 없는 경우 변경 없음
신장, 체중, 혈액형, 알레르기, 진찰일 등의 날짜, 질병·처치·투약 등의 정보, 검사치 등	변경 없음
전자 의료 기록에 포함된 정보 및 기타 텍스트 정보	텍스트 정보의 내용에 따라 다른 항목과 마찬가지로 처리

주: 다른 기록 등을 대체하고자 임시 ID를 붙이는 경우에는 원래의 기록 등을 복원할 수 있는 규칙성을 갖지 않는 방법이어야 한다.
출처: 일본 내각부·문부과학성·후생노동성·경제산업성. (2024.4.). 의료분야의 연구 개발에 기여하기 위한 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인(차세대의료기반법 가이드라인). p.133.

제3절 소결

제2장은 국내 사례로 개인정보 비식별화와 관련된 국내의 5개 가이드라인을 검토하고, 국외 사례로는 미국과 일본의 개인정보 비식별화와 관련된 사례를 검토하였다.

국내 사례로 검토한 가이드라인은 금융분야 가명·익명 처리 안내서, 가

명정보 처리 가이드라인, 보건의료데이터 활용 가이드라인, 교육분야 가명·익명 정보 처리 가이드라인, 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인이다. 5개 가이드라인의 연혁을 살펴보면, 가명정보 처리 가이드라인뿐만 아니라 다른 가이드라인들도 법 개정과 실무 적용, 가명·익명 처리 사례의 누적 및 시대 변화에 따라 가명 또는 익명 정보 활용 수요자들의 요구에 부응하여 지속적으로 개정되고 있음을 알 수 있다. 가이드라인에서는 실무 지원을 위한 가명처리 또는 익명처리 방법 등을 안내하고 있는 만큼 지속해서 개정·보완 등이 필요함을 시사한다.

5개 가이드라인의 내용을 살펴보면, 크게 통계법과 관련 있는 통계청 가이드라인과 데이터 3법과 관련된 그 외 4개 가이드라인으로 구분할 수 있다. 데이터 3법과 관련된 가이드라인들은 가명처리 또는 익명처리 절차에 따라 처리 방법 및 절차를 설명하고 있는 반면, 통계청 가이드라인에서는 통계 작성 단계별 또는 통계자료 제공 단계별로 업무 절차에 따라 비식별화 방법을 설명하고 있다. 통계청의 가이드라인은 적용 대상이 통계작성기관이기에, 통계작성기관인 한국보건사회연구원도 데이터 제공 시 익명처리 절차 및 비식별화 처리에 대한 기준이 필요함을 시사한다.하고 있다.

비식별화, 가명처리, 익명처리에 대한 정의는 가이드라인별로 상이하게 기술하고 있으나, 맥락은 같으며, 통계청 가이드라인은 가명처리와 익명처리 모두가 해당하는 '비식별화'를 포괄 범위로 하고 있다. 본 연구에서는 비식별화 데이터 생성을 익명처리에 가까운 비식별화로 정의하여 검토하고자 하였다.

5개 가이드라인에서 소개하는 비식별화 처리 방법은 국제 표준인 ISO/IEC 20889 분류를 따르며, 이 보고서의 제3장 비식별화 방법론 검토도 ISO/IEC 20889에서 제안한 분류 체계에 따라 다양한 비식별화 방법을 정의하고 특징을 검토한다.

가명·익명 처리의 절차와 적정성 평가는 통계청 가이드라인의 경우 통계작성기관의 업무 수행 절차를 따르며, 비식별 처리에 대한 적정성 평가는 하지 않는다. 나머지 4개의 가이드라인의 익명처리는 사전 준비 및 익명처리 후 적정성 평가를 하는 절차로 이루어진다.

추가로, 식별자(identifier)와 관련한 가이드라인별 용어 정의를 요약하여 제시하였다. 이 보고서는 비식별화 데이터 생성과 관련하여 개인식별 가능 정보(준식별자) 용어를 주로 사용하였다.

국외의 사례는 의료분야를 중심으로 비식별 처리 관련 사항을 검토하였다. 미국 HIPAA의 PHI 익명화 방법에 대한 지침은 보호 대상 정보에 대해 비식별 처리를 하는 방법으로 전문가 결정 방식과 세이프 하버 방식을 제시하였다. 세이프 하버 방식은 개인에 대한 이름, 주소, 날짜, 이메일 주소, 사회보장번호 등 18가지의 식별 정보를 모두 제거하고 사용하는 방식이다. 일본 차세대의료기반법 가이드라인은 의료 정보 분류를 식별자, 준식별자, 정적 속성, 반정적 속성, 동적 속성으로 나누어 익명 가공 방법의 예시를 제시하였다.

비식별화 처리와 관련하여, 원내에도 가이드라인이 필요한 상황으로, 국내 5개 가이드라인과 국외 사례의 비식별 처리 제반 사항을 연구원 상황에 맞게 적용하고자 한다. 자세한 부분은 제5장에 기술하였다.

제3장

비식별화 방법론 검토

제 3 장 비식별화 방법론 검토

이 장에서는 ISO/IEC 20889에서 제안한 분류 체계에 따라 다양한 비식별화 방법을 정의하고 특징을 검토하고자 한다. 데이터의 변수나 레코드에 직접적인 비식별 조치를 취하는 기법의 경우 통계청에서 작성한 '통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인(통계청, 2023)'의 내용을 주로 참고하였으며 가상의 데이터로 예제를 소개하였다. 또한 차등 정보보호 방법론과 재현 데이터 방법론은 박민정 등(2018, 2019), 안성빈 등(2023)과 김지우 등(2023)의 연구에서 사용한 체계와 기호를 수정 없이 그대로 차용하였다.

ISO/IEC 20899는 데이터 마스킹(masking), 가명화, 삭제, 범주화 등 다양한 비식별화를 통해 개인정보를 보호하는 다양한 방법론과 이를 체계적으로 관리하기 위한 분류 표준을 제시하고 있다. 제안된 비식별화 방법론은 데이터의 민감도와 정보보호의 수준에 따라 적절한 기법을 선택하는 것이 중요하다. 본 연구에서는 ISO/IEC 20899에서 제시한 분류 체계에 따라 실무에서 주로 사용되고 있는 기법 중 일부를 소개한다.

〈표 3-1〉 국제표준(ISO/IEC 20889)에 따른 비식별화 방법

구분	비식별화 방법	적용 항목
구조적 방법	표본추출	자료 전체
	총계처리	연속형
삭제	항목 삭제(열 삭제)	연속형, 범주형
	레코드 삭제(행 삭제)	연속형, 범주형
	국소 삭제(셀 삭제)	연속형, 범주형

구분	비식별화 방법		적용 항목
	기호로 표기(masking)		-
일반화	반올림	일반 반올림	연속형
		랜덤 반올림	연속형, 범주형
	범주화	재범주화	연속형, 범주형
		상하단 범주화	연속형, 범주형
임의화	잡음 첨가(noise addition)		연속형
	자료 교환(swapping)		연속형
	부분 총계(microaggregation)		연속형
재현자료 생성	재현자료(synthetic data)		자료 전체
암호화	일방향 암호화		연속형, 범주형

출처: 통계청. (2023). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. p.26.

제1절 변수·레코드 수준 비식별화

1. 구조적 방법: 표본추출

구조적 방법은 마이크로데이터의 레코드 구조를 변형하여 제공하는 방법이며 가장 대표적인 방법은 표본추출 방법이다.

가. 단순임의추출(simple random sampling)

모집단 대상에 고유의 일련번호를 부여하고 부여한 일련번호 중 일부를 동일한 확률로 추출하는 방법으로 가장 단순한 표본추출 방법이며 다른 표본추출 방법의 기본이 된다.

나. 층화추출(stratified random sampling)

모집단 대상을 동질적인 속성을 갖는 층으로 분할(partition)하고 각각의 층에서 단순임의추출로 표본을 추출하는 기법이다. 층화추출은 분석 목적에 맞는 층의 구조를 정확하게 정의하는 것이 매우 중요하며 층의 정의가 적합한 경우 단순임의추출에 비하여 다양한 장점을 가진다. 층화추출은 각각의 층의 크기에 비례하여 표본의 크기를 결정하는 비례층화표본추출과 층의 크기에 무관하게 표본의 크기를 결정하는 비비례층화표본추출로 구분된다.

다. 계통추출

계통추출은 모집단 대상이 가지는 특징에 근거하여 모집단 대상 각각에 순서를 정의할 수 있는 경우에 적용하는 방법으로 모집단 대상을 순서에 따라 나열한 상태에서 동일한 간격으로 표본을 추출하는 방법이다. 계통추출은 추출하는 방법이 매우 쉽고 표본이 모집단 전체에서 고르게 추출되는 장점을 가진다. 하지만 모집단 대상의 특정 속성이 모집단에 부여된 순서에 의존하는 경우 대표성이 결여될 수 있다.

2. 구조적 방법: 총계처리

데이터 레코드의 전체 혹은 일부를 평균, 중앙값, 최빈값, 최대/최소값 등의 대푯값으로 대체하는 비식별화 방법이다.

- 평균: 가장 일반적인 대푯값이나 데이터에 이상치가 존재하는 경우 정보가 왜곡되는 문제점 발생한다.

- 중앙값: 연속형 변수에서 이상치가 존재하는 경우 평균이 가지는 문제점을 보완하기 위하여 사용된다.

- 최빈값: 주로 범주형 변수에 적용하는 대푯값으로 가장 높은 빈도로 나타난 속성으로 데이터를 대체한다.

- 최대/최소값: 가장 큰/작은 값으로 대체하는 방법이며 데이터에서 상/하단의 이상치가 존재하면 정보가 왜곡된다.

〈표 3-2〉 총계처리: 월 소득 변수를 평균으로 대체

성별	연령	혼인	월 소득	주소	월 소득
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길	5,383,992
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12	5,383,992
M	54	Y	10,166,129	대구광역시 중구 동성로 23길	5,383,992
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길	5,383,992
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34	5,383,992
F	56	N	5,925,442	대전광역시 서구 둔산로 88길	5,383,992
F	51	N	5,441,829	울산광역시 남구 삼산로 55길	5,383,992
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길	5,383,992
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길	5,383,992
F	51	Y	11,026,554	강원도 춘천시 소양로 9길	5,383,992

출처: 저자 작성

3. 삭제 방법

가. 식별 변수 삭제

식별 정보를 포함하는 변수 전체를 삭제하는 기법으로 정보를 가장 강력하게 보호할 수 있지만 데이터 활용의 측면에서 가장 비효율적인 기법이다.

〈표 3-3〉 주소 삭제

성별	연령	혼인	월 소득	주소
M	56	Y	4,664,014	
F	52	N	3,166,376	
M	54	Y	10,166,129	
M	59	Y	4,141,649	
M	51	N	3,485,405	
F	56	N	5,925,442	
F	51	N	5,441,829	
F	53	Y	2,570,398	
M	50	Y	3,252,123	
F	51	Y	11,026,554	

출처: 저자 작성

나. 식별 변수 국소 삭제

변수 전체를 삭제하지 않고 일부를 변형하는 기법을 국소 삭제라고 한다.

〈표 3-4〉 주소의 일부를 국소 삭제

성별	연령	혼인	월 소득	주소
M	56	Y	4,664,014	서울특별시
F	52	N	3,166,376	부산광역시
M	54	Y	10,166,129	대구광역시
M	59	Y	4,141,649	인천광역시
M	51	N	3,485,405	경기도
F	56	N	5,925,442	대전광역시
F	51	N	5,441,829	울산광역시
F	53	Y	2,570,398	경상남도
M	50	Y	3,252,123	전라북도
F	51	Y	11,026,554	강원도

출처: 저자 작성

다. 레코드 삭제

이상치 등 식별 가능성이 매우 높은 레코드 전체를 삭제하는 기법이지만 이상치가 포함된 분석이 필요한 경우 분석 결과를 왜곡할 가능성이 높다.

〈표 3-5〉 세 번째와 열 번째 레코드 전체를 삭제

성별	연령	혼인	월 소득	주소
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34
F	56	N	5,925,442	대전광역시 서구 둔산로 88길
F	51	N	5,441,829	울산광역시 남구 삼산로 55길
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길

출처: 저자 작성

라. 마스킹

식별 가능성이 높은 변수 혹은 레코드에 대하여 삭제 대신 일부를 특수기호로 대체하는 기법으로 비식별 처리된 데이터임을 직관적으로 보여주므로 가장 많이 사용되는 기법 중 하나이다.

〈표 3-6〉 성명에서 이름을 기호 *을 사용하여 마스킹

성명	성별	연령	혼인	월 소득	주소
김**	M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길
최**	F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12
박**	M	54	Y	10,166,129	대구광역시 중구 동성로 23길
이**	M	59	Y	4,141,649	인천광역시 남동구 구월로 19길
김**	M	51	N	3,485,405	경기도 성남시 분당구 판교로 34
김**	F	56	N	5,925,442	대전광역시 서구 둔산로 88길
최**	F	51	N	5,441,829	울산광역시 남구 삼산로 55길
박**	F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길
유**	M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길
오**	F	51	Y	11,026,554	강원도 춘천시 소양로 9길

출처: 저자 작성

4. 일반화 방법

가. 라운딩 방법

식별 가능성이 높은 변수에 대하여 라운딩 처리를 통해 식별 가능성을 낮추는 기법으로 정보보호의 수준에 따라 라운딩을 적용할 단위를 결정한다.

〈표 3-7〉 월 소득을 백만 단위로 라운딩

성별	연령	혼인	월 소득	주소	월 소득
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길	4,000,000
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12	3,000,000
M	54	Y	10,166,129	대구광역시 중구 동성로 23길	10,000,000
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길	4,000,000
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34	3,000,000
F	56	N	5,925,442	대전광역시 서구 둔산로 88길	5,000,000
F	51	N	5,441,829	울산광역시 남구 삼산로 55길	5,000,000
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길	2,000,000
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길	3,000,000
F	51	Y	11,026,554	강원도 춘천시 소양로 9길	11,000,000

출처: 저자 작성

나. 범주화 방법

식별 가능성이 높은 연속형 변수를 범주화하여 식별 가능성을 낮추는 기법이며 데이터 활용의 측면에서 유용하나 외부의 데이터와 연계하여 정보를 노출할 가능성이 존재한다.

〈표 3-8〉 연령을 다섯 살 범위로 범주화

성별	연령	혼인	월 소득	주소	연령
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길	55세~60세
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12	50세~54세
M	54	Y	10,166,129	대구광역시 중구 동성로 23길	50세~54세
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길	55세~60세
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34	50세~54세
F	56	N	5,925,442	대전광역시 서구 둔산로 88길	55세~60세
F	51	N	5,441,829	울산광역시 남구 삼산로 55길	50세~54세
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길	50세~54세
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길	50세~54세
F	51	Y	11,026,554	강원도 춘천시 소양로 9길	50세~54세

출처: 저자 작성

5. 임의화 방법

가. 잡음 첨가

연속형 변수의 전체 혹은 일부의 레코드에 잡음(noise)을 더하여 레코드의 값을 대체하는 기법이며 잡음의 크기에 따라 비식별화 수준을 결정할 수 있지만 너무 큰 잡음을 사용할 경우 데이터의 유용성이 낮아지는 단점을 가진다. 최근 활발히 연구되고 있는 차등 정보보호(differential privacy) 기법에 적용되는 핵심 기법이며, 차등 정보보호 기법의 특징은 정보보호의 수준을 하나의 수치로 정량화가 가능하다는 점이다.

〈표 3-9〉 월 소득에 잡음 첨가

성별	연령	혼인	월 소득	주소	월 소득
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길	5,050,890
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12	3,909,509
M	54	Y	10,166,129	대구광역시 중구 동성로 23길	11,135,229
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길	4,580,409
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34	3,182,608
F	56	N	5,925,442	대전광역시 서구 둔산로 88길	6,144,210
F	51	N	5,441,829	울산광역시 남구 삼산로 55길	5,245,309
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길	1,858,430
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길	3,002,632
F	51	Y	11,026,554	강원도 춘천시 소양로 9길	11,994,649

출처: 저자 작성

나. 데이터 교환

식별 가능성이 높은 변수의 값을 특정 순서로 재배열하여 식별 가능성을 낮추는 기법으로 주로 연속형 변수에 적용하는 기법이다. 변수의 값을 특정 기준으로 그룹화하고 각 그룹 안에서 변수들의 값을 임의로 교환하는 방법으로 평균이나 분산 등이 유지되는 장점을 가진다.

〈표 3-10〉 월 소득 레코드를 성별로 재배열한 후 레코드 교환

성별	연령	혼인	월 소득	주소	월 소득
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길	10,166,129
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34	3,252,123
M	54	Y	10,166,129	대구광역시 중구 동성로 23길	4,141,649
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길	3,485,405
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길	4,664,014
F	51	Y	11,026,554	강원도 춘천시 소양로 9길	5,441,829
F	51	N	5,441,829	울산광역시 남구 삼산로 55길	5,925,442
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12	11,026,554
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길	3,166,376
F	56	N	5,925,442	대전광역시 서구 둔산로 88길	2,570,398

출처: 저자 작성

다. 부분 총계

데이터의 레코드들을 특정 기준에 따라 여러 개의 그룹으로 분할하고 각 그룹의 대푯값으로 대체하는 방법이다.

〈표 3-11〉 월 소득을 성별로 부분 총계

성별	연령	혼인	월 소득	주소	월 소득
M	50	Y	3,252,123	전라북도 전주시 덕진구 송천로 77길	5,141,864
M	51	N	3,485,405	경기도 성남시 분당구 판교로 34	5,141,864
M	54	Y	10,166,129	대구광역시 중구 동성로 23길	5,141,864
M	56	Y	4,664,014	서울특별시 강남구 테헤란로 45길	5,141,864
M	59	Y	4,141,649	인천광역시 남동구 구월로 19길	5,141,864
F	51	Y	11,026,554	강원도 춘천시 소양로 9길	5,626,120
F	51	N	5,441,829	울산광역시 남구 삼산로 55길	5,626,120
F	52	N	3,166,376	부산광역시 해운대구 달맞이길 12	5,626,120
F	53	Y	2,570,398	경상남도 창원시 의창구 원이대로 2길	5,626,120
F	56	N	5,925,442	대전광역시 서구 둔산로 88길	5,626,120

출처: 저자 작성

제2절 차등 정보보호(differential privacy)

1. 개념 및 정의

차등 정보보호는 Dwork(2006)이 제안한 방법으로 원본 데이터(original data)를 활용할 목적으로 특정 쿼리(query)를 요청하면 쿼리의 결과물에 적당한 잡음을 첨가하여 비식별 처리하는 방법이다. 차등 정보보호를 적용하면 정보보호의 수준을 하나의 수치적 모수로 결정할 수 있으며 쿼리의 결과물뿐만 아니라 원본 데이터 전체를 비식별 처리하여

재현 데이터(synthtic data)를 만들 수도 있다.

[정의] D와 D'을 레코드가 하나만 다른 두 데이터라고 하자. 만약 주어진 랜덤화 알고리즘(randomized algorithm) K가 적당한 $\alpha > 0$에 대하여 다음과 같은 조건을 만족하면 K를 α-차등 정보보호(α-differential privacy) 랜덤화 알고리즘이라고 한다.

$$P[K(D_1) \in S] \leq e^{\alpha} P[K(D_2) \in S] \quad for\ all\ S \subseteq range(K).$$

[정의] 원본 데이터 $X = (X_1, \ldots, X_n)$와 X를 비식별 처리한 재현 데이터 $Z = (Z_1, \ldots, Z_n)$에 대하여 Z의 X에 대한 조건부 밀도함수(conditional density function)를 $f_{Z|X}$라고 하자. 만약 $f_{Z|X}$가 적당한 양수 α에 대하여 다음과 같은 부등식을 만족하면 $f_{Z|X}$를 X에 대한 α-차등 정보보호 기법(α-differential privacy mechanism)이라고 하고, Z를 X에 대한 α-차등 정보보호 데이터(α-differentially private view)라고 한다.

$$sup_{\|u-v\|_0 = 1} \ sup_z \frac{f_{Z|X}(z|u)}{f_{Z|X}(z|v)} \leq exp(\alpha).$$

위 식에서 $\|u - v\|_0 = \sum_{i=1}^{n} I(u_i \neq v_i)$이며 I는 지시(indicator) 함수이다.

Wasserman과 Zhou(2010)는 차등 정보보호 방법으로 비식별 처리한 데이터가 다음 [정리]와 같은 이론적 성질을 가진다는 것을 증명하였다. 이 정리의 주요 내용은 α-차등 정보보호 데이터를 측도 가능한(measurable) 함수를 사용하여 변환해도 여전히 α-차등 정보보호 데이터가 된다는 것을 의미한다.

[정리] $e = (e_1, \ldots, e_m)$을 밀도함수 f_ϵ를 따르는 임의표본이라고 하자. 만약 e가 원본 데이터 $X = (X_1, \ldots, X_n)$와 독립이면 e와 X를 사용하여 정의한 확률벡터 $Z(X, e)$는 다음과 같은 성질을 가진다.

① $Z(X, e)$가 X에 대한 α-차등 정보보호 데이터이면 임의의 측도 가능한 함수 S에 대하여 $W = S(Z(X, e))$도 X에 대한 α-차등 정보보호 데이터이다.

② $Z(X, e)$가 X에 대한 α-차등 정보보호 데이터이고 f_Z가 $Z(X, e)$의 밀도함수이면 f_Z로부터 추출한 임의표본도 X에 대한 α-차등 정보보호 데이터이다.

Wasserman과 Zhou(2010)는 차등 정보보호 방법에서 상수 α의 의미를 다음 [정리]와 같이 통계적 가설에 대한 검정력의 상한(upper bound)으로 해석한다.

[정리] $Z = (Z_1, \ldots, Z_k)$를 원본 데이터 $X = (X_1, \ldots, X_n)$의 재현 데이터이고 다음 가설에 대한 검정법 T_i, $i \in \{1, \ldots, n\}$의 검정통계량이 Z를 사용하여 정의된다고 하자.

$$H_0 : X_i = u \text{ vs } H_1 : X_i = v,\ u \neq v$$

만약 Z가 X에 대한 α-차등 정보보호 데이터이고 T_i의 유의수준(significance level)이 γ이면 임의의 $i \in \{1, \ldots, n\}$에 대하여 T_i의 검정력(power)은 $\gamma exp(\alpha)$이다.

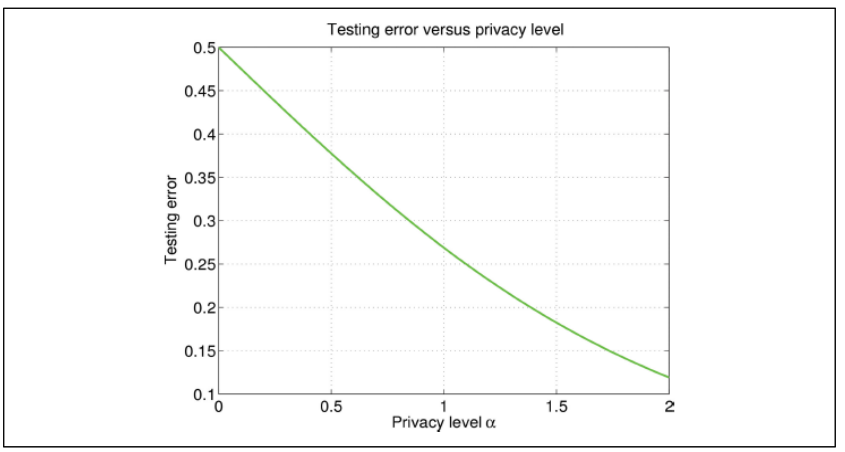
위 정리는 재현 데이터 Z를 사용하여 원본 데이터 X에 포함된 특정 레코드 X_i의 값을 식별(identification)할 수 있는지 통계적으로 검정할 때의 결과를 제시한다. 만약 $\alpha \to 0$인 경우 $\gamma exp(\alpha) \to \gamma$가 되므로 정리에 제시된 검정법의 검정력은 유의수준과 거의 같아진다. 또한 일반적으로

유의수준 γ는 0.1 이하이므로 결국 차등 정보보호가 강해지면, 즉 $\alpha \to 0$이면 재현 데이터를 사용하여 특정 레코드의 값에 관하여 어떠한 판단을 내리기 어려워진다는 것을 의미한다. 이러한 의미에서 α를 차등 정보보호 수준(differential privacy level)이라고 한다.

Duchi et al.(2018)은 α의 의미를 해석하기 위하여 다음과 같은 검정오차(testing error) P_{err}을 사용하였다.

$$P_{err} = \frac{1}{2} inf_{\psi}\{P(\psi(Z) \neq v | X = v) + P(\psi(Z) \neq u | X = u)\}$$

〔그림 3-1〕 차등 정보보호 수준에 따른 검정오차



출처: Duchi et al.(2018)

위 검정오차 P_{err}의 정의에서 $\psi(Z)$는 재현 데이터 Z를 이용한 검정의 결과이며, 첫 번째 항은 제2종의 오류를, 두 번째 항은 제1종의 오류를 의미한다. 따라서 위 그림으로부터 차등 정보보호를 강력하게 적용하면 즉, $\alpha \to 0$이면 $P_{err} \to 1/2$이므로 Z를 사용한 임의의 검정법은 모두 무작위 추측(random guess)이 된다.

2. 차등 정보보호를 적용한 데이터 교란

원본 데이터 X에서 정의된 특정 쿼리의 결과물에 적당한 잡음을 가하는 방법을 데이터 교란(perturbation)이라고 한다. 데이터 교란 방법으로 생성한 데이터가 X의 차등 정보보호 데이터가 되려면 α-차등 정보보호의 정의를 만족시켜야 하며 잡음의 분포와 쿼리의 특징을 고려하여 교란해야 한다.

아래 [예제]는 Dwork et al.(2006)이 제안한 데이터 교란 방법으로 데이터의 합이나 히스토그램에 대하여 α-차등 정보보호를 적용하는 방법을 소개하고 있다.

[예제] 원본 데이터 $X = (X_1, \ldots, X_n)$, $X_i \in \{0,1\}$, $i \in \{1, \ldots, n\}$ 와 데이터 교란에 사용된 오차가 서로 독립이라고 하자. X와 e를 사용하여 다음과 같이 원본 데이터의 합을 교란할 수 있다.

$$Z = \sum_{i=1}^{n} X_i + e$$

이때 오차 e의 밀도함수가 $f_\epsilon(x) \propto exp(-|x-\mu|/\sigma)$, $\sigma = 1/\alpha$ 이면 다음과 같은 부등식을 증명할 수 있으며

$$sup_{\|u-v\|_0 = 1} \; sup_z \frac{F_{Z|X}(z|u)}{F_{Z|X}(z|v)}$$

$$= sup_{a,b \in \{0,1\}} \; exp(\alpha|a-b|) \leq exp(\alpha),$$

이는 위와 같은 데이터 교란이 α-차등 정보보호의 정의를 만족한다는 것을 의미한다.

[예제] 원본 데이터 $X = (X_1, \ldots, X_n)$의 히스토그램은 다음과 같이 표

현할 수 있다.

$$\hat{f}_m(x) = \sum_{j=1}^{m} \frac{C_j}{nh} I(x \in B_j).$$

위 식에서 h, B_j, C_j는 각각 평활모수(bandwidth), 막대(bin), 막대 B_j에서의 데이터의 빈도(frequency)이며, m은 막대의 개수이다. $V = (V_1, \ldots, V_m)$을 이중지수분포에서 추출한 임의표본이라고 하자.

$$g(v) = \frac{\alpha}{4} e^{-\alpha\frac{|v|}{2}}$$

그러면 다음과 같이 교란한 데이터 $D = (D_1, \ldots, D_m)$는 X에 대한 α-차등 정보보호 데이터이다.

$$D_j = C_j + V_j,\ j \in \{1, \ldots, m\}$$

위 예제뿐만 아니라 Wasserman과 Zhou(2010)가 증명한 정리를 적용하면 차등 정보보호 데이터를 생성하는 알고리즘 중간에 적절한 변환을 사용해도 여전히 α-차등 정보보호를 만족한다.

3. 차등 정보보호를 적용한 재현

앞에서 소개한 예제는 Dwork et al.(2006)이 소개하였으며 원본 데이터의 빈도 C_j를 이중지수분포에서 추출한 잡음 V_j로 교란하여 α-차등 정보보호 데이터 D_j를 만드는 방법이다. 하지만 이러한 히스토그램의 빈도가 너무 작으면 잡음을 추가하면서 음의 빈도가 발생할 가능성이 존재한다. 이러한 단점을 개선하기 위하여 Wasserman과 Zhou(2010)는 잡

음을 직접 추가하는 대신 먼저 α -차등 정보보호를 만족하는 재현 데이터를 생성하고, 이 재현 데이터를 사용하여 히스토그램을 만드는 방법을 제안하였다.

$$X_1, \ldots, X_n \xrightarrow[estimation]{density} \hat{p}$$

$$\xrightarrow{privatize} \hat{\boldsymbol{p}}^{*} \xrightarrow{sample} Z_1, \ldots, Z_k$$

[정리] 원본 데이터 $X=(X_1,\ldots,X_n)$, $X_i \in [0,1]$, $i \in \{1, \ldots, n\}$ 에 대하여 다음 밀도함수에서 추출한 임의표본 $Z=(Z_1, \ldots, Z_k)$는 임의의 k에 대하여 X에 대한 α - 차등 정보보호 데이터이다.

$$\tilde{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{q}_j}{h} I(x \in B_j)$$

위 식에서 $\tilde{D}_j = max\{D_j, 0\}$ 이고 $\hat{q}_j = \tilde{D}_j / \sum_{s=1}^{m} \tilde{D}_s$ 이다.

위의 정리에서 소개한 방법의 특징은 차등 정보보호 데이터의 수 k를 임의로 조정할 수 있다는 점이다.

4. 차등 정보보호 히스토그램

히스토그램은 데이터를 분석하여 정리하는 가장 단순한 기법 중 하나이지만 가장 자주 사용되는 기법이기도 하다. 히스토그램을 비식별 처리하는 방법은 크게 두 가지로 구분할 수 있다. 하나는 히스토그램의 빈도를 교란하는 방법이며, 다른 하나는 원본 데이터를 직접 비식별화하고 이를 사용하여 히스토그램을 만드는 것이다. 이 소절에서는 먼저 기본 히스

토그램을 소개하고 이를 비식별 처리한 다양한 차등 정보보호 히스토그램을 소개한다.

가. 기본 히스토그램

원본 데이터를 $X = (X_1, \ldots, X_n)$가 $X_i \in \chi = R^d$, $i \leq n$, $d \geq 1$이라고 하자. X의 기본 히스토그램은 비모수 추정법을 이용한 밀도함수의 하나이며 다음과 같이 정의된다.

$$\hat{f}_m(x) = \sum_{j=1}^{m} \frac{C_j}{nh^d} I(x \in B_j) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h^d} I(x \in B_j)$$

위 식에서 h, B_j, C_j는 각각 히스토그램의 평활모수, 막대, 막대 B_j에서의 데이터의 빈도이며, m은 막대의 개수이고, $\hat{p}_j = C_j/n$이다. 기본 히스토그램의 작성법은 다음과 같다.

[작성법] 기본 히스토그램 작성법

(1) 히스토그램 막대의 개수 m을 결정

(2) 다음을 만족하는 m개의 막대(bin) B_j를 결정

$B_j \cap B_{j'} = \varnothing$, $\forall j \neq j'$, $\cup_{j=1}^{m} B_j = R^d$

(3) 막대에서 원본 데이터의 빈도(frequency) C_j, $j \leq m$ 계산

$C_j = \sum_{i=1}^{n} I(X_i \in B_j)$, $j \leq m$

(4) 막대의 부피 $V_j = Vol(B_j)$, $j \leq m$를 사용하여 히스토그램을 작성

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j}I(x \in B_j),\ x \in R^d$$

표본의 수 n이 무한히 증가하면, 위 기본 히스토그램이 X의 밀도함수로 수렴하는 일치(consistent) 추정량임이 잘 알려져 있다.

나. 교란 히스토그램

교란 히스토그램은 Dwork(2006)가 제안한 α-차등 정보보호 히스토그램이며 원본 데이터의 히스토그램을 교란하여 작성하는 히스토그램을 교란 히스토그램이고 한다. Dwork(2006)의 교란 히스토그램은 다음과 같이 이중지수분포에서 생성한 잡음으로 교란한다.

$$f(x) = (2|b|)^{-1}e^{-|x|/|b|},\ x \in R$$

[작성법] 교란 히스토그램 작성법

(1) 원본 데이터를 사용하여 기본 히스토그램을 작성

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j}I(x \in B_j),\ x \in R^d$$

(2) 정보보호 수준 $\alpha > 0$를 결정

(3) 산포모수가 $b = 2/\alpha$인 이중지수 분포에서 임의표본 $W_j,\ j \le m$을 생성

(4) 교란빈도 $D_j = C_j + W_j,\ j \le m$을 사용하여 히스토그램을 작성

$$\hat{f}_m^{perturbed}(x) = \frac{1}{D}\sum_{j=1}^{m}\frac{D_j}{V_j}I(x \in B_j)$$

$$x \in R^d,\ D = \sum_{j=1}^{m}D_j$$

Dwork(2006)는 교란 히스토그램이 α-차등 정보보호 히스토그램임을 증명하였다. 이 기법은 잡음을 생성하여 더하는 단순한 기법이므로 구현이 쉽고 정보보호 과정이 직관적이라는 장점이 있다.

다. 보정된 교란 히스토그램

보정된(modified) 교란 히스토그램은 Wasserman과 Zhou(2010)가 제안한 α-차등 정보보호 히스토그램으로 교란 히스토그램에서 음수로 교란된 빈도를 강제로 0으로 대체하여 단점을 개선한 히스토그램이다.

[작성법] 보정된 교란 히스토그램 작성법

(1) 원본 데이터를 사용하여 기본 히스토그램을 작성

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j} I(x \in B_j),\ x \in R^d.$$

(2) 정보보호 수준 $\alpha > 0$를 결정

(3) 산포모수가 $b = 2/\alpha$ 인 이중지수 분포에서 임의표본 $W_j,\ j \le m$을 생성

(4) 보정된 교란빈도 $M_j = max\{C_j + W_j,\ 0\},\ j \le m$을 사용하여 히스토그램을 작성

$$\hat{f}_m^{modified}(x) = \frac{1}{M}\sum_{j=1}^{m}\frac{M_j}{V_j} I(x \in B_j)$$

$$x \in R^d,\ M = \sum_{j=1}^{m} M_j$$

Wasserman과 Zhou(2010)는 보정된 교란 히스토그램이 α-차등 정보보호 데이터임을 증명하였다. 기본 히스토그램에서 0인 빈도가 0이 아

닌 빈도로 바뀔 수 있으며 역의 상황도 발생할 수 있다.

라. 재현된 교란 히스토그램

Wasserman과 Zhou(2010)는 재현 데이터를 생성하고 생성된 재현 데이터로부터 히스토그램을 작성하는 기법을 소개하였으며 이러한 히스토그램을 재현된 교란 히스토그램이라고 한다.

[작성법] 재현된 교란 히스토그램 작성법

(1) 보정된 교란 히스토그램을 작성

$$\hat{f}_m^{modified}(x) = \frac{1}{M}\sum_{j=1}^{m}\frac{M_j}{V_j}I(x \in B_j)$$

$$x \in R^d,\ M = \sum_{j=1}^{m} M_j$$

(2) 재현 데이터의 개수 k를 결정

(3) 보정된 교란 히스토그램에서 재현 데이터 $Z_i,\ i \le k$를 임의 생성

(4) 재현 데이터를 사용하여 히스토그램을 작성

$$\hat{f}_m^{syn.perturbed}(x) = \frac{1}{R}\sum_{j=1}^{m}\frac{R_j}{V_j}I(x \in B_j)$$

$$x \in R^d,\ R_j = \sum_{i=1}^{k} I(Z_i \in B_j),\ R = \sum_{j=1}^{m} R_j$$

Wasserman과 Zhou(2010)는 재현된 교란 히스토그램이 α-차등 정보보호 히스토그램임을 증명하였다. 재현된 교란 히스토그램은 α-차등 정보보호 재현 데이터를 사용하며, 특히 재현 데이터를 원본 데이터보다 더 많이 생성할 수도 있다는 장점이 있다.

마. 재현된 평활 히스토그램

재현된 평활 히스토그램은 Wasserman과 Zhou(2010)가 제안한 히스토그램으로 빈도를 직접 교란하지 않고 평활(smoothed) 히스토그램을 사용하여 비식별 히스토그램을 만드는 기법이다.

[작성법] 원본 데이터를 사용하여 기본 히스토그램을 작성

$$\hat{f}_m^{original}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{C_j}{V_j} I(x \in B_j),\ x \in R^d$$

(2) 평활수준 $\delta > 0$을 결정한다.

(3) 평활 히스토그램에서 재현 데이터 Z_i, $i \le k$를 임의 생성

$$\hat{f}_m^{\delta}(x) = (1-\delta)\hat{f}_m^{original}(x) + \delta / V$$

$$x \in R^d,\ V = \sum_{j=1}^{m} V_j$$

(4) 재현 데이터를 사용하여 히스토그램을 작성

$$\hat{f}_m^{syn.smoothed}(x) = \frac{1}{R}\sum_{j=1}^{m}\frac{R_j}{V_j} I(x \in B_j)$$

$$x \in R^d,\ R_j = \sum_{i=1}^{k} I(Z_i \in B_j),\ R = \sum_{j=1}^{m} R_j$$

Wasserman과 Zhou(2010)는 특정 조건을 만족하면 평활 히스토그램이 α-차등 정보보호 히스토그램임을 증명하였다. 재현된 평활 히스토그램은 정보보호 수준 α가 반영되지 않은 것처럼 보이지만 평활수준 δ와, 재현 데이터의 수 k, 막대의 개수 m등이 정해지면 α의 값을 계산할 수 있다.

제3절 재현 데이터

1. 개요

재현 데이터는 기존의 비식별화 기법과 차별성을 가지는 새로운 개념의 비식별화 기법이며 통계적으로는 데이터의 분포에서 새로운 표본을 추출하는 방법으로 해석할 수 있다. 재현 데이터를 사용한 비식별 처리는 원본 데이터와 분포적 성질이 유사한 가상의 데이터를 생성하는 것을 의미한다.

재현 데이터는 원본 데이터와 동일하거나 유사한 특성을 가진 데이터를 직접 생성하는 기법이므로 조사나 관측이 불가능한 상황에서 가상 시나리오에 따라 실험을 하는 것과 동일한 효과를 가진다.

재현 데이터는 정보를 보호한다는 의미뿐만 아니라 유용한 데이터를 생성하는 유효한 방법론으로 자리 잡고 있으며 다양한 목적으로 활용할 수 있다.

- 데이터 프라이버시 보호: 민감한 데이터를 직접 다루지 않고도, 데이터 분석 및 연구를 수행할 수 있으며, 특히 개인정보 보호가 중요한 분야에서, 재현 데이터를 통해 개인의 정보를 보호하면서도 연구가 가능하다.
- 데이터 접근성 향상: 원본 데이터에 접근할 수 없는 경우에도, 재현 데이터를 사용하여 분석을 진행할 수 있다. 예를 들어, 비용이나 시간 등의 이유로 데이터 수집이 어려운 상황에서 재현 데이터는 대안으로 활용할 수 있다.
- 모형 검증 및 성능 평가: 재현 데이터는 모형이나 알고리즘의 성능을 평가하는 데 사용될 수 있으며 다양한 시나리오에서 모형의 정확성과 안정성을 평가할 수 있다.
- 데이터 손실 복구: 원본 데이터가 손실되거나 손상된 경우, 재현 데

이터를 이용해 원본 데이터의 특성을 일부 복구하는 것이 가능하다.

재현 데이터의 가치는 원본 데이터의 분포와 얼마나 유사한지 그리고 정보보호가 얼마나 안전하게 이루어지고 있는지로 결정된다. 이를 재현 데이터의 유용성과 위험성이라고 하며 일반적으로 서로 상충하는 관계를 가진다. 이외에도 분석 목적에 맞는 데이터를 가상으로 생성한다는 의미에서 활용의 유용성과 위험성 평가 측도로 사용하기도 한다.

- 활용의 유용성 평가 측도
  1) 통계적 유사성: 재현 데이터는 원본 데이터와 통계적으로 유사한 특성을 가져야 하며 데이터 분포, 상관관계, 평균 및 분산 등 통계적 지표에서 원본 데이터와 유사한 결과를 도출할 수 있어야 한다.
  2) 유연성: 다양한 분석 목적에 따라 재현 데이터를 생성할 수 있어야 하며 이를 통해 특정 연구 질문에 맞는 데이터를 만들고, 이를 분석에 활용할 수 있어야 한다.
- 위험성 평가 측도

  비식별성: 재현 데이터는 원본 데이터의 민감한 정보가 드러나지 않도록 구성되어야 하며 개별 데이터를 식별할 수 없도록 익명화가 철저히 이루어져야 한다.

2. 순차회귀 재현 데이터

순차회귀(sequential regression) 재현 데이터는 회귀모형을 순차적으로 적용하여 재현 데이터를 생성하는 방법이다. 순차회귀 재현 데이터는 원본 데이터를 구성하는 변수 X_1, ..., X_p의 결합분포를 순차회귀 모형을 이용하여 추정하여 생성하며 재현 데이터를 생성할 때도 추정된 회귀모형을 이용하여 순차적으로 생성한다.

순차회귀 재현 데이터를 생성하는 절차는 다음과 같다.

(1) 원본 데이터를 구성하는 변수의 결합(joint) 밀도함수를 다음과 같이 순차적 형태로 분해하고

$f(x_1,\ldots,x_p) = f_1(x_1)f_2(x_2|x_1)\cdots f_p(x_p|x_1,\ldots,x_{p-1})$

각각의 조건부(conditional) 밀도함수를 순차적으로 추정한다.

$f_j(x_j|x_1,\ldots,x_{j-1})$

이때 분해의 순서는 사용자가 데이터의 특징에 맞게 결정해야 하며 분해의 순서가 다르면 재현 데이터도 다르게 생성된다.

(2) 첫 번째 변수 X_1의 밀도함수 $f_1(x_1)$을 X_1의 관측값 x_{11}, x_{21}, ..., x_{n1}을 사용하여 경험(empirical) 밀도함수로 추정한다.

$$\hat{f_1}(x_1) = \frac{1}{n}\sum_i I(x = x_{i1})$$

(3) 조건부 밀도함수 $f_j(x_j|x_1,\ldots,x_{j-1})$를 추정할 때는 X_j의 분포의 성질을 고려하여 적절한 회귀모형을 선택하여 추정한다. 예를 들어 X_j가 연속형 분포를 가지는 경우 다음과 같이 선형회귀 모형을 사용하고

$$X_j = \beta_0 + \sum_{k=1}^{j-1}\beta_k X_k + \epsilon,$$

X_j가 범주형 분포를 가지면 다음과 같은 다중 로지스틱 회귀모형을 사용한다.

$$P(X_j = m|X_1 = x_1,\ldots,X_{j-1} = x_{j-1}) = \frac{\exp(v_m(x_1,\ldots,x_{j-1}))}{\sum_l \exp(v_l(x_1,\ldots,x_{j-1}))},$$

$$v_m(x_1,\ldots,x_{j-1}) = \beta_{m0} + \sum_{k=1}^{j-1}\beta_{mk}x_k,$$

이외에도 CART(Breiman, 2017)와 같은 비선형 회귀모형을 사

용할 수도 있다.

(4) 추정된 조건부 밀도함수로부터 조건부 밀도함수를 추정한 순서와 동일한 순서로 단순임의추출법을 사용하여 재현 데이터를 생성한다.

$$y_{11}, \ldots, y_{n1} \sim \hat{f}_1(x_1) = \frac{1}{n}\sum_i I(x = x_{i1})$$

$$y_{1l}, \ldots, y_{nl} \sim \hat{f}_j(x_j | x_1, \ldots, x_{j-1}),\ l = 1, \ldots, j-1.$$

3. 베이지안 재현 데이터

Murray와 Reiter(2016)는 연속형 변수와 범주형 변수가 모두 포함된 데이터에서 베이지안 기법을 사용하여 재현 데이터를 생성하는 방법을 제안하였다. 제안된 방법에서 사후 분포를 추정하기 위한 가정은 다음과 같다.

- 범주형 변수는 혼합 다항 분포(mixture multinomial distributions)를 따른다.
- 연속형 변수는 범주형 변수를 설명변수로 사용하는 혼합(mixture) 회귀모형 따른다.
- 혼합 분포에서 사용되는 혼합 요소(mixture component)와 두 종류의 변수가 가지는 의존성을 설명하는 요소는 모두 디리클레 과정(Dirichlet process)을 따른다.

제안된 베이지안 재현 데이터를 생성하는 방법은 순차회귀를 이용한 생성 방법보다 계산 시간이 오래 걸리는 단점이 있지만 사용자가 사전에 결정해야 하는 조율 모수(tuning parameter)가 없기 때문에 재현 과정이 좀 더 일관성을 가진다. 또한 원본 데이터로부터 재현 데이터가 생성되는 과정을 모형에 기반하여 설명할 수 있다는 장점이 있다.

4. 딥러닝 기반 재현 데이터 생성

가. 적대적 학습

심층신경망을 사용한 재현 데이터 생성 방법론에서 가장 자주 사용되는 모형은 다음과 같이 잠재변수(latent variable)를 사용하는 생성 모형이다.

$$X|Z = g(Z;\theta),\ Z \sim N(0_d, I_d)$$

위 식에서 $Z \in R^d$는 잠재벡터, $X \in R^p$는 관측벡터이고, $g(\cdot, \theta)$는 모수가 θ인 생성자(generator)이며, 0_d와 1_d는 d차원 0-벡터와 항등행렬(identity matrix)이다.

위 잠재변수 생성 모형의 특징은 잠재벡터가 관측벡터와 분포적(distributional) 관계가 아니라 결정적(deterministic) 관계를 갖는다는 점이다. 또한 $d < p$인 경우 X의 밀도함수가 존재하지 않으므로 생성자의 모수 θ를 최대우도 추정법(maximum likelihood estimation)을 사용하여 추정할 수 없다.

적대적 학습(adversarial training)은 생성 모형을 학습하기 위하여 생성자와 함께 모수가 η인 구분자 $d(\cdot;\eta)$(dicriminator)를 생성 모형에 도입하고 생성자와 구분자가 서로 적대적 관계를 가지도록 학습하는 기법이다. 여기에서 적대적 관계란 구분자가 원본 데이터와 재현 데이터를 최대한 정확하게 분류하고, 생성자가 원본 데이터와 최대한 유사한 데이터를 생성하는 관계를 의미한다.

적대적 학습은 재현 데이터와 원본 데이터의 분포 사이의 거리를 정의하는 방식에 따라 구분할 수 있으며 GAN(generative adversarial net-

work, Goodfellow 등, 2014), f-GAN(f-diverbence GAN, Nowozin 등, 2016), WGAN(Wasserstein GAN, Arjovsky 등, 2017) 등이 대표적인 적대적 학습 방법이다.

Goodfellow 등(2014)은 이러한 적대적 관계를 학습하기 위하여 다음과 같은 최적화 문제를 제안하였다.

$$\min_{\theta}\max_{\eta} L_{gan},$$
$$L_{gan} = E_{X\sim P_X}[\log d(X;\eta)] + E_{Z\sim N(0_d, I_d)}[\log(1-d(g(Z;\theta);\eta))]$$

GAN은 위 최적화 문제에서 손실함수 L_{gan}을 사용하여 구분자 $d(\cdot,\eta)$가 관측벡터 X와 재현 벡터 $g(Z,\theta)$를 최대한 정확하게 분류하고 생성 모형 $g(\cdot;\theta)$가 X와 가장 유사한 $g(Z,\theta)$를 생성하도록, 즉 두 모형이 적대적 관계를 가지도록 학습한다.

한편 X와 $g(Z;\theta)$가 모두 밀도함수를 가지면 주어진 생성자 $g(\cdot;\theta)$에 대하여 손실함수 L_{gan}을 최대로 하는 최적(optimal) 구분자는 다음과 같다.

$$d^{*}(x) = \frac{p_X(x)}{p_X(x)+p_\theta(x)}$$

위 식에서 p_X는 X의 밀도함수, p_θ는 $g(Z;\theta)$의 밀도함수이다. 이제 최적 구분자를 최적화 문제에 대입하고 정리하면 다음의 결과를 얻는다.

$$\min_{\theta} KL\left[p_X \| \frac{p_X+p_\theta}{2}\right] + KL\left[p_\theta \| \frac{p_X+p_\theta}{2}\right]$$
$$-\log 4 = 2JS(p_X \| p_\theta) - \log 4.$$

위 식에서 KL은 쿨백-라이블러(KL, Kullback과 Leibler, 1951) 발산

(divergence)과 젠슨-샤넌(JS, Jensen-Shannon, Lin, 1991) 발산을 의미한다. 위 등식으로부터 GAN이 관측벡터의 밀도함수와 JS 발산을 최소로 하는 생성자의 밀도함수를 추정하는 기법임을 확인할 수 있다.

나. Wasserstein GAN

GAN을 최적화의 측면에서 해석하려면 밀도함수의 존재성이 필수적이다. 따라서 관측벡터가 특이 분포(singular distribution)를 따르면, 즉 관측벡터의 밀도함수가 존재하지 않으면 GAN의 최적화 과정을 해석하기 어렵다. 또한 특이 분포를 따르는 데이터를 사용하여 GAN을 학습하면 관측벡터의 일부 속성만 제한적으로 생성하는 문제점이 발생하며 이를 모드 붕괴(mode collapse) 문제라고 한다.

이러한 모드 붕괴 문제를 해결하기 위하여 Arjovsky 등(2017)은 밀도함수 대신 분포 함수의 Wasserstein-1 거리를 최소화하는 방향으로 GAN을 개선하였다. Wasserstein-1 거리는 공간 R^p에서 정의된 두 분포 P, Q의 거리이며 다음과 같이 정의된다.

$$W(P,Q) = inf_{\gamma \in \Pi(P,Q)} E_{(U,V) \sim \gamma} \| U - V \|_2$$

위 식에서 $\Pi(P,Q)$는 $R^p \times R^p$에서 정의된 분포 중 주변(marginal) 분포가 각각 P와 Q인 모든 분포의 집합을 의미하며 $\| \cdot \|_2$는 유클리디안 놈(Euclidean norm)이다.

Villani(2009)는 다음과 같이 Wasserstein-1 거리의 정의에서 최적화 문제를 쌍대 문제(dual problem)로 치환하여 재정의하였다.

$$W(P,Q) = sup_{\|f\|_L \leq 1} [E_{U \sim P}[f(U)] - E_{V \sim Q}[f(V)]]$$

위 식에서 $\|f\|_L$은 주어진 함수 $f : R^p \rightarrow R$의 Lipschitz 상수이다.

위 쌍대 문제로부터 두 분포 P와 Q 사이의 Wasserstein-1 거리는 1-Lipschitz 조건을 만족하는 함수 f를 사용하여 변환한 분포의 기댓값의 차이의 최대값과 동일하다는 것을 확인할 수 있다.

Arjovsky 등(2017)은 이러한 관계를 이용하여 GAN의 손실함수를 개선하여 다음과 같이 WGAN(Wasserstein GAN)을 제안하였다.

$$\min_{\theta}\max_{\eta} L_{wgan}$$

$$\text{subject to } \|d(\cdot;\eta)\|_L \leq 1$$

$$L_{wgan} = E_{X \sim P_X}[d(X;\eta)] - E_{Z \sim N(0_d, I_d)}[d(g(Z;\theta);\eta)]$$

WGAN은 GAN과 마찬가지로 생성자와 구분자가 서로 적대적 관계를 가지도록 학습한다. 또한 구분자에 대한 1-Lipschitz 제약 조건을 반영하기 위하여 모수 η의 크기를 사전에 정해진 기준값 이하로 절삭(clipping)하는 학습 기법을 제안하였다.

한편 WGAN은 구분자에 대한 1-Lipschitz 조건을 정확하게 만족하는 해를 찾지 못한다는 단점이 있다. Gulrajani 등(2017)은 최적 구분자의 경사(gradient)가 항상 1이 된다는 성질을 증명하고 이러한 사실을 이용하여 WGAN의 절삭 과정을 개선하고 WGAN-GP(Wasserstein GAN with gradient penalty)를 제안하였다.

$$\min_{\theta}\max_{\eta} L_{wgan-gp}$$

$$L_{wgan-gp} = E_{X \sim P_X}[d(X;\eta)] - E_{Z \sim N(0_d, I_d)}[d(g(Z;\theta);\eta)]$$

$$+ \lambda E_{T \sim Unif[0,1]} E_{\tilde{X} \sim \tilde{P}_{\theta,T}}[\|\nabla_{\tilde{X}} d(\tilde{X};\eta)\|_2 - 1]^2.$$

위 식에서 $\tilde{P}_{\theta,t} = tP_X + (1-t)P_\theta$, $t \in [0,1]$이고, $\lambda \in [0,\infty)$는 구분 모

형의 경사의 크기를 조정하는 조율 모수이다.

다. MedGAN

Choi 등(2017)은 전자 건강기록(electronic health record, EHR) 데이터를 재현하기 위하여 기존의 심층 신경망 생성 기법을 개선하였다. EHR 데이터는 성별 혹은 나이와 같은 환자 개인의 정보와 질병의 진단 결과, 발병의 횟수, 완치 여부와 병원 방문 여부 등 다양한 질병 관련 정보를 포함하고 있는 데이터이다. MedGAN은 이진(binary) 범주형 변수가 포함된 데이를 생성할 수 있는 기법이며, 다범주 변수는 원핫 인코딩(one-hot encoding) 처리하여 이진 범주형 변수로 변환하여 생성 기법을 적용한다.

MedGAN은 범주형 변수를 임베딩(embedding)하는 저차원 임베딩 공간과 관측벡터 공간을 연결하는 두 개의 오토 인코더(autoencoders, AE) 모형을 활용하며, 생성자, 구분자, 오토 인코더에 대하여 DNN(deep neural network) 모형을 사용한다.

$$X|Z = dec(g(Z;\theta);\xi) \quad Z \sim N(0_d, I_d)$$

위 생성 모형에서 각 모형의 역할은 다음과 간다.

- 생성자 $g(\cdot;\theta): R^d \rightarrow R^c$는 d차원 잠재벡터 Z를 사용하여 c차원 임베딩 공간의 재현 벡터 $g(Z;\theta)$를 생성한다
- 인코더 $enc(\cdot;\phi): R^p \rightarrow R^c$은 p차원 관측벡터 X를 c차원 임베딩 공간으로 변환한다. ϕ는 인코더 모형의 모수이다.
- 디코더 $dec(\cdot;\xi): R^p \rightarrow R^c$은 생성된 재현 벡터를 원래의 관측 공간으로 변환한다. ξ는 디코더 모형의 모수이다.

MedGAN에서 모형의 학습 단계는 다음과 같다. 먼저 인코더와 디코더 재구성 손실함수(reconstruction loss function)를 최소화하는 방향으로 학습한다.

$$\min_{\phi,\xi} E_{X\sim P_X}[\sum_{X_j\in X} l_j(X_j, dec(enc(X_j;\phi);\xi)]$$

위 식에서 l_j는 X_j에 대응하는 재구성 손실함수이며 X_j가 수치형인 경우 제곱 손실함수를 사용하고, 이진형인 경우 교차-엔트로피(cross-entropy, CE) 손실함수를 사용한다.

MedGAN은 성능을 향상시키기 위하여 배치 표준화(batch normalization)와 제외 연결(skip connection) 등의 기법을 추가로 사용하며, 시그모이드(sigmoid) 함수를 사용하여 임베딩 공간에서 만들어지는 벡터의 원소가 0과 1 사이의 값을 가지도록 변형하여 사용하므로, 학습 과정에서 관측 데이터에 대한 특별한 전처리를 필요로 하지 않는다는 장점이 있다.

마지막으로 AE를 학습한 후 다음과 같은 최적화 과정을 통해 생성자와 구분자를 학습한다.

$$\min_{\theta,\xi}\max_{\eta} E_{X\sim \hat{P}_X}[\log d(X;\eta)]$$
$$+ E_{Z\sim N(0_d, I_d)}[\log(1-d(dec(g(Z;\theta);\xi);\eta))]$$

위 최적화 과정에서 디코더, 생성자, 구분자의 모수만 학습하며 인코더의 모수 ϕ는 고정한다.

라. TableGAN

TableGAN(Park 등, 2018)은 CNN(convolution neural network)을 사용한 구분자와 DCNN(deep CNN)을 사용한 생성자 그리고 CNN을 사용한 분류기(classifier)를 추가로 활용하는 심층 재현 데이터 생성 기법이다.

TableGAN은 수치형과 범주형 변수가 혼합된 테이블 데이터에 적용할 수 있으며, 수치형 변수의 크기를 정해진 방법으로 스케일링(scaling)하고, 범주형 변수의 경우 가변수(dummy variable)로 변환하는 등 간단한 전처리가 필요한 기법이다. 또한 CNN과 DCNN이 행렬 형태의 이미지 데이터를 입력값으로 사용하므로 테이블 데이터를 행렬로 변환하는 전처리 작업을 추가로 수행한다.

TableGAN에서 p차원의 입력 변수를 $k \times k$ 형태의 정방 행렬로 변환할 때, 변수의 원소 사이의 순서는 고려하지는 않으며, 정방 행렬에 원소가 존재하지 않을 경우 zero padding을 이용하여 대체한다. 역으로 TableGAN의 생성자가 생성한 정방 행렬 형태의 데이터를 다시 원래의 벡터 형식으로 변환해주는 후처리 과정을 도입한다.

TableGAN은 위와 같은 전처리 과정 이외에도 테이블 데이터의 분포 학습에 적합하도록 GAN의 손실함수와 함께 두 개의 보조 손실함수를 조합하여 모형을 학습한다.

첫 번째 보조 손실함수는 관측벡터와 재현 벡터 사이의 통계적 유사성(statistical similarity)을 유지하기 위하여 사용하는 정보 손실(information loss) 함수이다. 정보 손실함수는 관측벡터와 재현 벡터가 구분 모형의 최상위 은닉층 $h(\cdot\,;\eta)$을 통해 변형될 때 다음과 같이 평균과 표준편차의 차이에 힌지 손실(hinge loss) 함수를 적용한다.

$$L_{info} = \max(0, L_{mean} - \delta_{mean}) + \max(0, L_{sd} - \delta_{sd})$$

$$L_{mean} = \| E_{X \sim P_X}[h(X;\eta))] - E_{Z \sim N(0_d, I_d)}[h(g(Z;\theta);\eta)] \|_2$$

$$L_{sd} = \| SD_{X \sim P_X}[h(X;\eta))] - SD_{Z \sim N(0_d, I_d)}[h(g(Z;\theta);\eta)] \|_2$$

위 식에서 δ_{mean}과 δ_{sd}는 조율 모수이며 원본 데이터와 재현 데이터 사잉의 통계적 유사성의 정도를 조절한다.

두 번째 보조 손실함수는 분류 손실(classification loss) 함수이며 모수가 ζ인 분류기 $cl(\cdot;\zeta): R^{p-1} \rightarrow R$을 학습하여 관측벡터와 재현 벡터가 의미론적 일치성(semantic integrity)을 확보할 수 있도록 제어한다.

분류기 모형의 경우 사전에 주어진 인덱스 $c \in \{1, \ldots, p\}$에 대하여 X에서 X_c를 제외한 X_{-c}를 입력값으로 사용하여 X_c를 예측하고, 동시에 $g_{-j}(Z;\theta)$를 입력값으로 사용하여 $g_j(Z;\theta)$를 예측하며, 전체적으로 두 예측 오차가 작아지도록 학습한다. 따라서 분류 손실함수는 TableGAN의 생성 모형이 관측벡터에서 발생할 수 없는 데이터의 조합을 생성하지 못하도록 방지하는 보조적 역할을 수행한다.

$$L_{class}^{cl} = E_{X \sim P_X} \| X_j - cl(X_{-j};\zeta) \|$$

$$L = E_{Z \sim N(0_d, I_d)} \| g_j(Z;\theta) - cl(g_{-j}(Z;\theta);\zeta) \|$$

이상을 요약하면 TableGAN의 손실함수는 다음과 같이 GAN의 손실함수와 두 보조 손실함수를 모두 포함하며 학습에 필요한 최적화 과정은 다음과 같다.

$$\min_\theta L_{gan} + L_{info} + L_{class}^g$$

$$\max_\eta L_{gan}$$

$$\min_\zeta L_{class}^{cl}.$$

마. CWGAN

CWGAN은 데이터에 범주의 비율이 매우 불균형한 변수가 포함되어 비율이 작은 범주가 재현되지 않는 모드 붕괴 문제를 해결하기 위해 제안된 심층 생성 기법이다. CWGAN(conditional Wasserstein GAN, Engelmann과 Lessmann, 2021)의 학습 과정은 WGAN의 학습 과정과 매우 유사하지만 데이터의 내부 혹은 외부에 포함된 정보를 사용하여 조건 변수(condition variable)를 정의하고 조건 변수를 사용하여 불균형 범주로 인해 발생하는 문제를 해결하는 기법이다.

CWGAN은 분류(classification) 모형에서 종속 변수의 범주가 매우 불균형한 경우에 사용하는 오버 샘플링(oversampling) 기법으로 제안되었지만 동일한 원리를 적용하여 수치형과 범주형 변수가 혼합된 테이블 데이터를 재현하는 것이 가능하다.

CWGAN은 생성자 $g(\cdot, Y;\theta)$와 구분자 $d(\cdot, Y;\eta)$는 사전에 정의한 조건 변수 Y에 대한 조건부 형태로 정의한다. 또한 조건부 재현 벡터 $g(Z, Y;\theta)$와 조건 변수 Y의 분포를 비교하기 위하여 모수가 γ인 이진 분류기 $ac(\cdot;\gamma): R^h \rightarrow R$을 사용한다.

$$
\begin{aligned}
&\min_{\theta,\gamma}\max_{\eta} E_{X\sim P_{X|Y}}[d(X, Y;\eta)] \\
&- E_{Z\sim N(0_d, I_d)}[d(g(Z, Y;\theta), Y;\eta)] \\
&\quad + \lambda_{GP} E_{T\sim Unif[0,1]} E_{\tilde{X}\sim \tilde{P}_{\theta,T}}[\|\nabla_{\tilde{X}} d(\tilde{X}, Y;\eta)\|_2 - 1]^2 \\
&+ \lambda_{AC} E_{Z\sim N(0_d, I_d)}[CE(Y, ac(g(Z, Y;\theta);\gamma))]
\end{aligned}
$$

위 식에서 $P_{X|Y}$는 $X|Y$의 조건부 분포, CE는 교차 엔트로피(cross entropy) 손실함수, λ_{GP}와 λ_{AC}는 양수인 조율 모수이다.

바. CTGAN

CTGAN(Xu 등, 2019)은 TableGAN과 마찬가지로 테이블 데이를 재현하기 위해 개발된 생성 기법이며 수치형 변수에 대하여 다양하고 정밀한 전처리 과정을 적용하여 TableGAN을 개선한 생성 기법이다. CTGAN은 수치형 변수의 분포를 가우시안 혼합 분포(Gaussian mixture model)로 가정하고 이를 전처리하기 위하여 MSN(mode-specific normalization) 기법을 사용한다.

CTGAN에서 MSN 기법을 사용하여 수치형 변수를 전처리하고 범주형 변수의 전처리 결과와 통합하는 과정은 다음과 같다.

(1) 가우시안 혼합 분포의 분포 모수와 군집의 개수를 변분 베이지안(variational Bayesian) (Bishop, Nasrabadi, 2006) 기법으로 추정한다.

$$p_{X_j}(x) = \sum_{k=1}^{m} \pi_k \phi(x; \mu_k, \sigma_k^2)$$

위 식에서 $\phi(\cdot\,; \mu, \sigma^2)$는 평균이 μ, 분산이 σ^2인 정규 분포의 밀도함수이며 m은 군집의 개수이다.

(2) 베이즈 정리를 이용해 j번째 변수의 i번째 관측값 x_{ij}가 속할 확률이 가장 높은 군집 k^* 확인한다.

$$k^* = \arg max_k \pi_k \phi(x_{ij}; \mu_k, \sigma_k^2),$$

(3) 관측값 x_{ij}를 다음과 같이 α_{ij}로 정규화(normalization)하고 k^*번째 원소만 1이고 나머지가 0인 이진(binary) 벡터 $\beta_{ij} \in R^m$를 사용하여 x_{ij}가 속한 군집의 정보를 저장한다.

$$\alpha_{ij} = \frac{x_{ij} - \mu_{k^*}}{4\sigma_{k^*}}, \; \beta_{ij} = (0, \ldots, 1, \ldots, 0)^T.$$

(4) 범주형 변수를 원-핫 인코딩하고 연속형 변수를 처리한 결과와 하나의 벡터로 연결한다.

CTGAN은 원본 데이터 X에서 불균형한 범주를 가지는 관측 변수 X_{imb}를 경험 분포로부터 생성한 후 X_{imb}의 원핫 인코딩 벡터 Y_{imb}를 조건으로 사용하여 조건부 생성자 $g(\cdot, Y_{imb};\theta)$와 조건부 구분자 $d(\cdot, Y_{imb};\eta)$를 WGAN-GP의 손실함수를 사용하여 학습하고 나머지 변수 X_{-imb}를 생성한다.

CTGAN은 모드 붕괴 문제를 해결하고 학습이 좀 더 안정적으로 이루어질 수 있도록 구분자가 1개 이상의 데이터를 입력받을 수 있도록 개선(PacGAN, Lin 등, 2018)하였으며, 조건부 재현 벡터 $g(Z, Y_{imb};\theta)$와 X_{-imb}가 최대한 일치하도록 CE 손실함수를 추가로 사용한다.

사. CTAB-GAN

CTAB-GAN(Zhao 등, 2021)은 데이터에 포함된 특정 수치형 변수의 분포가 매우 긴 꼬리를 가지는 경우에 사용하기 위하여 제안된 생성 기법이며 TableGAN과 CTGAN의 장점을 모두 적용한 생성 기법이다. 또한 CTAB-GAN은 수치형 값과 범주형 값을 동시에 가지는 혼합형 변수가 포함된 데이터에도 적용 가능하다는 장점이 있다.

CTAB-GAN에서 각 유형의 변수를 전처리하는 방법은 다음과 같다.

- 범주형 변수는 원핫 인코딩 벡터로 변환
- 일반 수치형 변수는 CTGAN의 MSN 기법을 적용
- 꼬리가 긴 수치형 변수는 다음과 같이 변환 후 MSN 적용

$$\tilde{x}_{ij} = \begin{cases} \log x_{ij}, & l_j > 0 \\ \log(x_{ij} - l_j + \epsilon), & l_j \leq 0 \end{cases}$$

위 식에서 x_{ij}는 변수 X_j의 i번째 관측값, l_j는 X_j의 하한 값, ϵ은 적당히 작은 양수

- 혼합형 변수는 수치형과 범주형 전처리 기법을 각각 적용

CTAB-GAN과 TableGAN은 벡터로 전처리된 데이터를 정방 행렬로 변형하며, 불가능한 조합의 데이터 생성을 방지하기 위하여 분류 손실함수를 사용하며, 변수의 통계적 특성을 유지하기 위하여 정보 손실함수를 사용한다는 공통점이 있다. 하지만 정보 손실함수에서 힌지 손실함수 대신 평균과 표준편차를 더한 새로운 정보 손실함수를 사용하는 것이 CTAB-GAN의 특징 중 하나이다.

CTAB-GAN과 CTGAN은 조건부 생성자와 조건부 구분자를 사용하며, 조건부 생성자를 학습하기 위한 손실함수를 추가로 사용한다는 공통점이 있다. 하지만 CTGAN과 달리 모든 변수에 대하여 training-by-sampling 기법을 적용하여 변수 사이의 상관관계를 더욱 잘 재현할 수 있다는 장점이 있다.

아. TVAE

Xu 등(2019)은 CTGAN과 함께 TVAE(Table VAE) 기법을 동시에 제안하였다. TVAE는 Kingma 와 Welling(2013)이 제안한 VAE(variational auto-encoder) 기법을 생성 모형에 적용한 우도(likelihood) 기반의 생성 기법이다.

VAE는 잠재벡터에 대한 관측벡터의 조건부 분포를 다음과 같이 모수가 θ인 조건부 확률을 사용하여 정의한 모형이다.

$$Z \sim N(0_d, I_d)$$

$$X|Z \sim p_{X|Z}(x|z;\theta) = \prod_{j=1}^{p} p_{X_j|Z}(x_{j|z};\theta)$$

TVAE는 GAN과 달리 잠재벡터와 관측벡터를 확률적 (stochastic) 관계를 가지도록 정의한다. 또한 X_j의 조건부 밀도함수 $p_{X_j|Z}$는 X_j가 수치형일 경우 정규 분포를, X_j가 이산형일 경우 다항 분포를 사용한다.

TVAE는 관측 벡터의 우도 함수를 최적화하여 학습하기 때문에 우도 함수가 계산이 어렵거나 최적화 시간이 오래 걸리는 경우 로그 우도 함수의 하한(ELBO, evidence lower bound)을 최대화하는 방식으로 학습한다.

$$\log p_{X(x;\theta)} = \log \int p_{X|V}(x|v;\theta) p_{V(v)dv}$$

$$\geq \int \log\left(\frac{p_{X,V}(x,v;\theta)}{q_{V|X}(v|x;\phi)}\right) q_{V|X}(v|x;\phi)dv$$

$$=: ELBO(\theta,\phi;x)$$

위 식에서 $q_{V|X}(\cdot|x;\phi)$는 ϕ를 모수로 가지는 변분 밀도함수 (variational density function)이다. VAE는 ELBO를 최대화하는 모수 θ와 ϕ를 추정하고, 이를 활용하여 다음의 밀도함수로부터 재현 데이터를 생성한다.

$$p_{X,Z}(x,z;\theta) = p_{X|Z}(x|z;\theta)p_Z(z)$$

5. 심층 재현 데이터 생성 기법

대부분의 적대적 학습 기법은 이미지 데이터를 생성하기 위해 개발되었다. 따라서 테이블 형태의 데이터를 생성하려면 적대적 학습 기법을 테

이블 데이터의 특징에 맞도록 적절하게 개선해야 할 필요가 있다.

- 범주형 변수의 처리: 이미지 데이터는 항상 수치형 변수로 구성되지만 테이블 데이터는 수치형 변수와 범주형 변수가 혼합되어 구성된다. 따라서 범주형 변수를 처리하기 위한 새로운 접근법이 필요하다.
- 변수 사이의 관계 반영: 이미지 데이터와 달리 테이블 데이터에는 개별 변수의 특성 혹은 변수와 변수 사이의 의미론적, 형식론적 관계가 반드시 포함되어 있다. 개별 변수의 경우 대부분 자체적인 범위, 최대와 최소, 속성의 제한과 같은 제약이 있고, 변수와 변수 사이의 상관관계 혹은 의미론적 제약이 복잡하게 얽혀 있기 때문에 테이블 데이터를 생성하려면 이러한 제약적 관계에 대하여 합리적으로 접근해야 한다.
- 불균형 분포의 문제: Salimans 등(2017)의 연구 결과에 따르면 이미지 데이터를 구성하는 변수의 주변 분포는 대부분 좌우 대칭이며 비교적 단순한 형태라는 것이다. 반면에 테이블 데이터에 포함된 수치형 변수의 경우 매우 높은 왜도(skewness)를 가지거나 하나 이상의 봉우리를 가지는 복잡한 분포를 보이는 경우가 많다. 또한 특정 속성이 데이터 전체에서 높은 비율을 차지하여 변수가 전체적으로 불균형한 분포를 보이는 경우도 빈번하다.

최근 테이블 데이터를 생성하는 적대적 생성 기법이 다양하게 제안되었다. 이러한 기법은 대부분 위와 같은 이미지 데이터와 테이블 데이터가 가지는 근본적인 차이점에 주목하고 이를 반영하여 기존의 생성 기법을 개선하고 있다.

6. 재현 데이터의 평가 지표

재현 데이터는 원본 데이터의 통계적 특성을 보존하면서 동시에 개인정보를 보호하기 위해 가상으로 생성된 데이터이다. 따라서 재현 데이터의 평가는 재현 데이터의 유용성과 위험성의 측면을 동시에 고려해야 한다. 본 연구에서는 재현 데이터 평가 지표 중 자주 사용되는 대표적인 지표를 몇 가지 소개한다.

가. 성향 점수

성향 점수(propensity score)는 Rosenbaum과 Rubin(1983)이 처음 제안하였으며 처리 효과(treatment effect)를 추정하는 통계적 기법으로 소개되었다. 이후 Woo 등(2009)은 성향 점수를 비식별화 데이터(masked data)의 유용성을 평가하는 지표로 사용하였으며 이후 성향 점수를 활용한 다양한 평가 지표가 제안되고 있다.

일반적으로 어떤 이진(binary) 확률변수 T의 확률변수 X에 대한 조건부 확률 $P(T=1|X)$를 성향 점수라고 한다. 성향 점수의 정의에서 처리 효과 T를 재현 데이터인지 재현 데이터가 아닌지의 여부로 정의하면 성향 점수를 사용하여 재현 데이터의 유용성을 평가할 수 있으며 그 절차는 다음과 같다.

(1) 원본 데이터 D_o와 재현 데이터 D_s를 레코드 방향으로 병합하여 병합 데이터 $D=(D_o, D_s)$를 구성하고, 병합된 레코드가 재현 데이터인지 아닌지를 나타내는 이진 변수 T 를 생성한다. 이때 레코드가 재현 데이터의 레코드이면 $T=1$이고, 원본 데이터의 레코드이면 $T=0$이다.

(2) D가 독립변수이고 T가 종속변수인 이진 분류 모형을 사용하여 X가 재현 레코드일 확률 $\hat{P}(T=1|X)$를 추정한다.

(3) D의 i 번째 레코드 X_i에 대하여 $\hat{p}_i = \hat{P}(T=1|X_i)$가 D에서 데이터의 비율 $c = \frac{n_s}{n_s + n_o}$와 가까울수록 X_i의 유용성이 높다고 판단한다.

Snoke 등(2018)은 재현 데이터를 활용, 성향 점수를 요약하여 성향 점수 평균제곱 오차(pMSE, propensity mean squared eror)를 정의하였으며, 이를 사용하면 재현 데이터 전체의 유용성을 측정하는 평가 지표로 적용할 수 있다고 제안하였다.

$$pMSE = \frac{1}{n_s + n_o} \sum_{i=1}^{n_s+n_o} (\hat{p}_i - c)^2$$

$pMSE$는 원본 데이터와 재현 데이터를 사용하여 적합한 이진 분류 모형의 분류 성능이 약할수록 작은 값을 가지므로 $pMSE$가 작을수록 재현 데이터의 유용성이 높다고 판단한다.

적절한 $pMSE$의 크기는 분석의 목적, 데이터 고유의 특징, 재현 데이터의 비율, 분류 모형의 종류 등 다양한 요소에 따라 다르다는 것이 일반적으로 알려진 사실이다. 이러한 단점을 개선하기 위하여 Snoke 등(2018)은 $pMSE$의 크기에 대한 절대적인 기준을 만들기 위해 $pMSE$의 이론적 분포를 유도하고 $pMSE$-비율(ratio)과 표준화(standardized)-$pMSE$를 제안하였다.

로지스틱 회귀모형을 사용하는 경우 $pMSE$는 적절한 가정하에 다음과 같은 카이제곱분포를 따른다.

$$pMSE \sim \frac{(1-c)^2 c}{n_s + n_o} \chi^2_{r-1}$$

위 식에서 r은 독립변수의 개수이다. 위 분포를 사용하면 다음과 같이 $pMSE$의 평균과 표준편차를 얻을 수 있다.

$$\mu_{null} = E(pMSE) = \frac{(1-c)^2 c}{n_s + n_o}(r-1)$$

$$sd_{null} = sd(pMSE) = \frac{(1-c)^2 c}{n_{s+n_o}}\sqrt{2(r-1)}$$

$pMSE$-비율과 표준화-$pMSE$를 다음과 같이 정한다.

$$pMSE-ratio = \frac{pMSE}{\mu_{null}}$$

$$standardized-pMSE = \frac{pMSE-\mu_{null}}{sd_{null}}$$

$pMSE$-비율이 1에 가까울수록 원본 데이터와 재현 데이터의 구분이 어려워지므로 $pMSE$의 값이 1에 가까울수록 재현 데이터의 유용성이 높다고 판단한다. 마찬가지로 표준화-$pMSE$의 값이 0에 가까울수록 재현 데이터의 유용성이 높다고 판단한다.

나. 분포 사이의 거리

재현 데이터의 유용성을 평가하는 측도 중 확률분포 사이의 거리를 사용하는 방법이 다양하게 제안되어 있다. 가장 대표적인 측도는 KL 발산과 Wasserstein 거리(Villani, 2009)를 이용하는 기법이다. 데이터에 포함된 변수가 많은 경우 변수 전체의 결합분포를 추정하는 것은 어려우므

로 주변분포를 추정하여 합을 이용하는 방식으로 근사한다. 두 밀도함수 f와 g에 대하여 KL 발산과 r-Wasserstein 거리는 각각 다음과 같다.

$$KL(f\|g) = \int_{-\infty}^{\infty} f(x)\log\frac{f(x)}{g(x)}dx$$

$$W_r(f,g) = \left(\int_0^1 |F_f^{-1}(t) - F_g^{-1}(t)|^r dt\right)^{1/r}$$

위 식에서 F_f와 F_g는 f와 g의 분포함수(distribution function)이다.

다. 신뢰구간 중첩 지표

원본 데이터에 대한 분석 목적이 정해지면 원본 데이터로 분석한 결과와 재현 데이터로 분석한 결과를 비교하는 것이 재현 데이터의 유용성을 평가하는 지표(Karr 등, 2006; Drechsler 와 Reiter, 2009; Snoke 등, 2018)가 될 수 있다. 가장 대표적인 예는 재현 데이터를 사용하여 추정한 회귀계수와 원본 데이터를 사용하여 추정한 회귀계수를 비교하여 분석 결과가 얼마나 차이가 나는지 비교하는 것이다.

Snoke 등(2018)은 추정된 회귀계수의 신뢰구간이 얼마나 중첩되는지 파악하는 기법을 제안하였으며 이를 신뢰구간 중첩 지표라고 한다. 신뢰구간 중첩 지표를 구하는 절차는 다음과 같다.

(1) 원본 데이터와 재현 데이터를 각각 사용하여 회귀계수 β_j의 95% 신뢰구간의 중첩도 IO_j를 구한다.

$$IO_j = 0.5\left[\frac{\min(u_{o,j},u_{s,j}) - \max(l_{o,j},l_{s,j})}{u_{o,j} - l_{o,j}} + \frac{\min(u_{o,j},u_{s,j}) - \max(l_{o,j},l_{s,j})}{u_{s,j} - l_{s,j}}\right],\ j = 0,1,2,\ldots,p,$$

단 위 식에서 p는 독립변수의 개수이고, $(l_{o,j}, u_{o,j})$와 $(l_{s,j}, u_{s,j})$는 각각 원본 데이터와 재현 데이터를 사용하여 추정한 95% 신뢰구간이다.

(2) 개별 회귀계수의 중첩도의 평균을 사용하여 요약한다.

$$IO = \frac{1}{p+1}\sum_{j=0}^{p} IO_j$$

IO의 값이 클수록 두 데이터를 사용한 분석 결과가 유사하다는 것을 의미하므로 재현 데이터의 유용성이 크다고 판단한다. 참고로 IO의 최댓값은 1이며 음수도 가능하다.

라. α-정밀도, β-재현율, 독창성 점수

앞에서 설명한 성향점수, 분포 사이의 거리, 신뢰구간 중첩 지표는 masked data에서 이미 활용되고 있는 방법으로 재현 데이터의 평가 측도로 새롭게 등장한 개념은 아니다. 재현 데이터의 평가 측도로 새롭게 나온 개념이 α-정밀도, β-재현율, 독창성 점수라고 할 수 있다. Alaa 등(2022)은 재현 데이터가 원본 데이터에 가까운 현실성 있는 데이터를 재현하는지, 원본 데이터의 다양성을 충분히 반영하는지, 원본 데이터에 포함되지 않는 독창적인 레코드를 재현하는지를 평가하는 측도를 제안하였다.

α-정밀도(precision)는 현실성에 대한 평가 측도이며 다음과 같이 정의한다.

$$P_\alpha := P(x_s \in S_o^\alpha),\ \alpha \in [0,1]$$

위 식에서 x_s는 재현 데이터의 레코드이며 S_o^α는 원본 데이터 D_o에서

계산한 경험 확률이 α인 영역 중 크기가 가장 작은 영역을 의미한다. 대부분의 데이터에서 S_o^α는 분포의 최빈값(mode)을 포함하는 영역이 된다.

β-재현율(recall)은 재현 데이터의 다양성에 대한 평가 측도이며 다음과 같이 정의된다.

$$R_\beta := P(x_o \in S_s^\beta),\ \beta \in [0,1]$$

위 식에서 x_o는 원본 데이터의 레코드이며, S_s^β는 재현 데이터 D_s에서 계산한 경험 확률이 α인 영역 중 크기가 가장 작은 영역을 의미한다. 따라서 β-재현율은 α-정밀도와 데이터에 대하여 대칭인 관계를 가진다.

결국 두 측도는 원본 데이터와 재현 데이터가 서로의 레코드를 포함하는 정도를 평가하는 측도이며 α와 β의 값에 따라 α-정밀도 곡선과 β-재현율 곡선이 45^o를 이루면 재현 데이터의 유용성이 높다고 판단한다.

위 두 측도를 사용하여 P_α와 R_β의 평균절대편차(mean absolute deviation)를 다음과 같이 정의할 수 있다.

$$\Delta P_\alpha := \int_0^1 |P_\alpha - \alpha| d\alpha,\ \Delta R_\beta := \int_0^1 |R_\beta - \beta| d\beta$$

평균절대편차는 모두 0과 0.5 사이에서 값을 가지며 0에 가까울수록 재현 데이터의 유용성이 높다고 판단한다. 또한 다음과 같이 변환하면 평균절대편차가 0과 1 사이의 값을 가지고, 1에 가까울수록 재현 데이터의 유용성이 높다고 판단할 수 있다.

$$IP_\alpha = 1 - 2\Delta P_\alpha,\ IR_\beta = 1 - 2\Delta R_\beta$$

독창성 점수(authenticity score)는 재현 데이터의 독창성에 대한 평가 측도이며 다음과 같이 정의된다.

$$P_s = A \times P'_s + (1 - A) \times \delta_{o,\epsilon},\ 0 \leq A \leq 1$$

위 식에서 계수 P'_s은 원본 데이터에 포함되지 않은 재현 데이터의 레코드로 추정한 분포이고 $\delta_{o,\epsilon}$은 원본 데이터의 경험 분포에 $N(0,\epsilon^2)$에서 추출한 잡음을 더한 분포이다. 또한 두 분포를 혼합할 때 P'_s의 계수 A를 독창성 점수라고 정의한다. 즉 독창성 점수는 재현 데이터의 분포 중에서 원본 데이터에 포함되지 않은 레코드로 구성된 분포가 차지하는 비율로 해석할 수 있으며 독창성 점수 A의 크기가 1에 가까울수록 재현 데이터가 독창적이고 따라서 재현 데이터의 위험성이 낮다고 판단한다.

마. 거리 기반 측도

Park 등(2018), Shokri 등(2017), Zhao 등(2021)은 재현 데이터의 레코드와 원본 데이터의 레코드 사이의 거리를 계산하여 재현 데이터의 유용성을 평가하는 기법을 제안하였다.

DCR(distance to the closest record) (check, Park 등, 2018)은 재현 데이터의 레코드 중에서 가장 가까운 원본 데이터의 레코드 사이의 거리를 의미하며 다음과 같이 정의한다.

$$d_i^* = \min_{j \leq n} |r_i - c_j|_2,\ i = 1, \ldots, n$$

위 식에서 r_i와 c_j는 각각 재현 데이터와 원본 데이터의 레코드를 표준화한 레코드이다. Park 등(2018)은 d_i^*가 작을수록 재현 데이터 레코드 r_i의 위험성이 높다고 판단하며, $d_i^*,\ i \leq n$의 평균이 작고 표준편차가 클수록 재현 데이터의 위험성이 높다고 판단한다. 평균과 표준편차 이외에도 위치와 산포에 대한 통계량을 적절히 사용(Zhao 등, 2021)하면 다양

한 위험성 측도를 정의할 수 있다.

바. 멤버쉽 추론 공격

멤버쉽 추론 공격(Membership Inference Attack) (MIA, Shokri 등, 2017)은 원본 데이터로 학습한 기계학습 모형이 원본 데이터의 특정 레코드를 사용하였는지 여부를 추론하는 공격이다. 공격자는 모형의 출력을 다양한 방법으로 분석하여 특정 레코드가 해당 모형의 학습에 사용되었는지 추론하여 해당 레코드를 식별하려고 한다.

Park 등(2018)은 MIA를 tableGAN의 생성 모형을 공격하기 위해 사용하였으며 절차는 다음과 같다.

(1) 대상(target) 모형 설정 및 생성(shadow) 모형 학습: 민감한 레코드를 포함하는 원본 데이터로 학습된 모형을 대상 모형으로 설정하고 대상 모형의 출력값 등 외부의 정보를 활용하여 대상 모형과 유사한 여러 개의 생성 모형을 학습한다. TableGAN의 경우 생성자를 사용하여 여러 개의 재현 데이터를 확보할 수 있다.

(2) 멤버십 데이터 구성 및 공격(attack) 모형 학습: 생성 모형을 학습할 때 사용한 데이터와 학습에 사용되지 않은 임의의 데이터에 대하여 생성 모형의 분류 결과를 저장하고 저장된 분류 결과를 사용하여 공격 모형을 학습한다.

(3) 대상 모형에 대한 공격 수행 및 평가: 공격 모형에 새로운 레코드를 입력하여 이 레코드가 대상 모형의 학습에 사용되었는지 여부를 결정하고 그 결과를 평가한다.

멤버쉽 추론 공격의 측도는 분류 지도학습(supervised learning) 모형을 평가하는 측도를 대부분 사용할 수 있으며 이러한 측도를 재현 데이

터의 측면에서 해석하면 다음과 같다.

- 정확도(Accuracy): 공격자가 모형이 학습한 레코드에 대한 멤버십을 얼마나 정확하게 추론하는지를 나타내는 지표이며 높은 정확도를 가지면 공격이 성공적이라고 판단한다.
- 정밀도(Precision): 모형이 잘못 분류한 레코드 중에서 실제로 학습에 사용된 레코드로 판별된 경우의 비율을 의미하며 공격자가 예측한 멤버십이 얼마나 정확한지를 측정한다.
- 재현율(Recall): 학습에 사용된 레코드를 올바르게 분류한 비율을 의미하며 공격자가 학습에 사용된 레코드를 얼마나 잘 찾는지를 측정한다.
- F1 스코어(F1 Score): 정밀도와 재현율의 조화 평균으로, 두 측도를 모두 고려하여 공격 성능을 평가한다.
- AUC-ROC(Area Under the ROC Curve): ROC 곡선은 민감도와 특이도(1-재현율)를 비교한 곡선이며 AUC는 ROC 아래의 면적을 의미한다. AUC 값이 1에 가까울수록 공격이 성공적이라고 판단한다.

제4절 비식별화 방법론의 실무 적용

이 절에서는 2021년도 가족과 출산 조사 데이터를 사용하여 앞에서 소개한 비식별화 방법론 중 실무에 적용 가능한 일부 방법론을 검토한다. 본 연구에는 원 데이터에 포함된 변수 중 아래의 7개 변수만 사용하였으며 R 패키지 sdcMicro(Templ, 2017)를 사용하였다.

〈표 3-12〉 연구에 사용된 변수의 리스트

변수명	설명
id	응답자 id
area	지역(동부/읍면부)
age	(응답자) 나이
job	(응답자) 경제활동 상태
income	합계소득(월평균)(만 원)
debt	부채(만 원)
weight	표본 가중치

출처: 한국보건사회연구원. (2021). 가족과 출산조사 코딩북.
https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

1. sdcMicro 개요

sdcMicro는 sdc 객체(object)를 사용하므로 데이터에 포함된 변수와 레코드 그리고 비식별 처리 전·후의 변화를 쉽게 확인할 수 있어 매우 편리하다.

sam.df는 data.frame 형식으로 저장된 가족과 출산 데이터이며, createSdcObj 함수를 사용하여 sdc라는 이름의 sdc 객체를 생성하였다. 범주형 식별 변수(keyVars)는 area, sex, age, job이고, 수치형 변수(nymVars)는 income, debt이다. 수치형 변수는 모두 민감 변수(sensibleVar)로 지정하였다.

[그림 3-2] sdcMicro 객체 생성

```
〉 sdc.obj = createSdcObj(dat = sam.df,
keyVars = c("area","sex","age","job"),
numVars = c("income","debt"),
sensibleVar = c("income","debt"),
weightVar = c("weight"),
hhId = c("id"))
```

출처: 저자 작성

[그림 3-3] sdcMicro 객체의 하부 객체

```
〉 slotNames(sdc.obj)

[1] "origData" "keyVars" "pramVars" "numVars"
[5] "ghostVars" "weightVar" "hhId" "strataVar"
[9] "sensibleVar" "manipKeyVars“ "manipPramVars”
"manipNumVars"
[13] "manipGhostVars" "manipStrataVar“
"originalRisk" "risk"
[17] "utility" "pram" "localSuppression" "options"
[21] "additionalResults" "set" "prev“ "deletedVars”

@origData: 원본 데이터, @keyVars: 범주형 변수 인덱스(index),
@pramVars: pram(post randomization method) 변수 인덱스
@numVars: 연속형 변수 인덱스, @weightVar: 샘플링 가중치,
@hhId: 군집 변수 인덱스, @strataVar: 층화 변수,
@sensibleVar: 민감 정보를 포함하는 변수, @manip***: 비식별 처리된 변수,
@originalRisk: 원본 데이터의 리스크, @utility: 데이터 유용성 정보,
@pram: pram 변수 정보, @localSuppression: 국소 감추기 정보
```

출처: 저자 작성

〈표 3-13〉 sdcMicro 패키지의 주요 함수

함수명	기능
freqCalc()	표본 빈도와 모집단 빈도
suda2()	부분 데이터에서의 빈도
ldiversity()	L-다양성
measureRisk()	레코드, 데이터 위험도
LLmodGlobalRisk()	로그선형모형을 이용한 위험도
dRisk()	연속형 변수에 대한 위험도
dRiskRMD()	연속형 변수에 대한 rmd 위험도
dUtility()	데이터 유용성
globalRecode()	범주형 변수 익명화

함수명	기능
localSupp()	변수 국소 감추기
localSuppression()	K-익명성 확보를 위한 국소 감추기
pram()	pram을 이용한 교환
topBottomCoding()	상위 하위 레코드 코딩
addNoise()	잡음 첨가
rankSwapp()	순위 바꾸기
mafast()	국소 통합(대용량 데이터)
microaggregation()	국소 통합
shuffle()	순서 바꾸기
get.sdcMicroObj()	sdc 객체로부터 직접 슬롯을 리턴
undolast()	마지막 수정된 sdc 객체로 되돌리기
extractManipData()	비식별 처리된 데이터 추출
calcRisks()	sdc 객체의 노출 위험 계산
varToFactor()	factor 변수로 변환
varToNumeric()	factor 변수를 numeric 변수로 변환

출처: 저자 작성

2. 식별 위험 추정

sdcMicro는 범주형 식별 변수들의 범주를 사용하여 식별 위험을 추정하거나 연속형 변수 사이의 거리를 사용하여 식별 위험을 추정한다.

가. 유일성을 사용한 위험 추정

범주형 변수의 범주를 사용하여 식별 위험을 추정하는 방법은, 데이터에 포함된 특정 조합의 레코드가 모집단에서도 유일한 레코드일 확률을 추정하여 식별 위험을 추정하는 방법이다. sdcMicro는 유일성 기반 식별 위험을 음이항 모형을 사용(Benedetti, Franconi, 1998)하여 추정한다.

indivisual 하부 항목은 데이터에 포함된 모든 레코드의 식별 위험을 제공한다. 결과에서 risk는 음이항 모형을 이용해 추정한 레코드별 위험도, fk는 범주형 식별 변수의 속성이 동일한 레코드의 빈도 수, Fk는 음이항 모형을 사용하여 추정한 모집단의 빈도수이다. 또한 global$risk 하부 항목은 hier_risk의 평균값으로 데이터 전체의 식별 위험을 나타낸다.

[그림 3-4] sdcMicro를 이용한 식별 위험 추정

```
〉 head(sdc.obj@risk$individual)

              risk  fk      Fk    hier_risk
[1,] 0.06548292 19 15.064   0.3992140
[2,] 0.08100867 14 12.217   1.1259620
[3,] 0.46068796  3   1.756   0.9514993
[4,] 0.23191095  4   4.416   0.9799354
[5,] 0.25602791  2   5.087   0.9396981
[6,] 0.22918258  3   5.045   0.9935719

〉 sdc.obj@risk$global$risk

[1] 0.2364321
```

출처: 저자 작성

나. k-익명성 추정

데이터의 각 레코드에 대하여 범주형 식별 변수의 속성 조합과 동일한 조합을 가지는 레코드가 k개 이상 존재하면 해당 레코드가 k-익명성(k-anonymity) (Sweeney, 2002)을 만족한다고 본다. k-익명성을 사용하면 특정 레코드의 식별 위험을 직관적으로 파악할 수 있기 때문에 k-익명성은 비식별 실무에서 가장 많이 사용되는 개념이다.

k-익명성을 만족하면 모든 레코드가 자기 자신과 구별되지 않는 k개 이상의 레코드를 가지므로 k가 커질수록 레코드의 식별 가능성이 더 낮아진다. k의 값은 데이터의 종류와 분석 목적 등에 따라 다르지만 보통 3 이상으로 정한다.

k-익명성을 확보하는 방법은 매우 다양하며 범주형 변수를 상위 범주로 재분류(global recording) (Sweeney, 2002)하거나 식별 위험이 높은 레코드를 마스킹(masking) 혹은 삭제(suppression)하는 방법으로 확보할 수 있다.

[그림 3-5] sdcMicro를 이용한 k-익명성 추정

```
〉 print(sdc.obj, type="kAnon")

Infos on 2/3-Anonymity:
Number of observations violating
  - 2-anonymity: 70 (7.000%)
  - 3-anonymity: 138 (13.800%)
  - 5-anonymity: 281 (28.100%)
```

출처: 저자 작성

위 결과에서 데이터의 7%가 2-익명성을 만족하지 않으며, 약 28%가 5-익명성을 만족하지 않는다는 것을 확인할 수 있다.

다. l-다양성 추정

l-다양성(l-diversity) (Machanavajjhala 등, 2007)은 k-익명성의 단점을 보완한 개념이며 식별 변수의 속성 조합에 대한 동질성 공격과 외부 지식이나 데이터를 활용한 공격에 대한 식별 위험을 정의하는 측도이다.

데이터에 포함된 범주형 식별 변수의 각 속성 조합에 대하여 민감 변수의 속성 조합이 l개 이상이면 해당 조합이 l-다양성을 만족한다고 한다.

[그림 3-6] sdcMicro를 이용한 l-다양성 추정

```
〉 ldiv = ldiversity(
sdc.obj@origData,keyVar=c("sex","job"),
ldiv_index=c("income","debt"))

〉 ldiv

L-Diversity Measures
income_Distinct_Ldiversity debt_Distinct_Ldiversity
Min.   :10.0               Min.   : 6.0
1st Qu. :161.0             1st Qu. :52.0
Median :248.0              Median :52.0
Mean   :215.3              Mean   :48.8
3rd Qu. :263.0             3rd Qu.: 53.0
Max.   :263.0              Max.   : 53.0
```

출처: 저자 작성

위 예제에서 속성 조합을 구성하기 위한 식별 변수는 sex와 job을 사용하였으며, 민감 변수 income과 debt에 대한 l-다양성을 추정하여 그 분포를 보여주고 있다.

라. 전체 위험도 추정

전체 위험도(global risk)는 데이터 전체의 식별 위험을 추정하는 방법으로 여러 개의 비식별 처리된 데이터가 있을 경우 가장 적합한 비식별 데이터를 선택할 때 사용할 수 있다. 전체 위험도는 레코드의 개별 위험

도의 평균으로 추정하는 방법과 로그-선형 모형을 사용하여 추정(Skinner, Holmes, 1998)하는 방법이 있다.

[그림 3-7] sdcMicro를 이용한 전체 위험도 추정 – 유일성

```
〉 print(sdc.obj,"risk")

Risk measures:
Number of observations with higher risk than the main
part of the data: 223
Expected number of re-identifications: 236.43 (23.64%)
```

출처: 저자 작성

위의 예제에서 식별 위험이 높은 레코드는 223개로 나타났으며. 식별이 될 것으로 예측되는 레코드의 수는 전체의 약 23.64%이다.

[그림 3-8] sdcMicro를 이용한 전체 위험도 추정 – 로그-선형모형

```
〉 sdc.obj =
 modRisk(sdc.obj,form=~sex+job+age+area)

〉 get.sdcMicroObj(sdc.obj,"risk")$model

The estimated model (using method 'default') was:
        ~ sex + job + age + area

global risk-measures:
        Risk-Measure 1: 301.876 (30187.637 %)
        Risk-Measure 2: 30.364 (3036.434 %)
```

출처: 저자 작성

주어진 변수 조합이 만들어내는 교차 분할표의 K개 셀에 대해 F_k를 k 번째 셀의 모집단 개체 수, f_k를 표본 개체 수라고 하면, 로그-선형 모형에서는 표본에서 유일한 개체가 모집단에서 유일할 확률(Risk-Measure 1), 모집단 개체 수 역수의 기댓값(Risk-Measure 2)이라고 노출 위험 측도를 정의한다.

$$Risk-measure1 : P(F_k = 1 | f_k = 1)$$
$$Risk-measure2 : E[1/F_k | f_k = 1]$$

위 예제에서 sex, job, age, area를 로그-선형 모형의 독립변수로 사용하였으며, Risk-Measure 1과 2 두 가지 방법으로 추정한 전체 위험도는 각각 301.876과 30.364이다.

3. 데이터 비식별 처리

가. 전반적 재코딩

전반적 재코딩(global recoding)은 특정 범주형 변수의 속성을 상위 범주로 재분류하는 비식별 기법이다. 전반적 재코딩은 범주형 변수와 수치형 변수에 모두 사용할 수 있다.

[그림 3-9] sdcMicro를 이용한 전반적 재코딩

```
> age.breaks = c(20,30,40,50)

> sdc.obj = globalRecode(sdc.obj,
column="age",breaks=age.breaks,
labels=LETTERS[1:(length(age.breaks)-1)])

> head(data.frame(
old.age=sdc.obj@origData[,c("age")],
          new.age=sdc.obj@manipKeyVars[,c("age")]))

  old.age new.age
1      38       B
2      42       C
3      44       C
4      29       A
5      22       A
6      52    <NA>
```

출처: 저자 작성

위 예에서 전반적 재코딩을 적용한 변수는 age이고 20에서 50까지 10단위로 묶어 범주형 변수로 변환하였다. 지정된 범위를 벗어나는 레코드는 NA로 처리된다.

전반적 레코딩은 레코드의 값을 대체하므로 식별 위험과 정보 손실이 발생한다. 다음 예제는 k-익명성과 대표적인 정보 손실 측도인 IL-1의 변화를 보여준다.

[그림 3-10] sdcMicro를 이용한 전반적 재코딩 후 k-익명성

```
〉 print(sdc.obj, type="kAnon")

Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 4 (0.400%) | in original data: 70 (7.000%)
  - 3-anonymity: 4 (0.400%) | in original data: 138 (13.800%)
  - 5-anonymity: 15 (1.500%) | in original data: 281 (28.100%)
```

출처: 저자 작성

sdcMicro에서 제공하는 IL_1은 특성의 집계 수준에서 데이터 품질의 손실 정도를 측정하는 데 사용된다.

$$IL_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{|\, x_i - x_i^{'} \,|}{\sigma_i}$$

여기서

n : 데이터 세트의 총 개수(총 레코드 수)

x_i : 원본 데이터 세트의 i번째 값

$x^{'}_i$: 비식별화된 데이터 세트의 i번째 값

σ_i : 원본 데이터의 해당 변수에 대한 표준편차

이다.

〔그림 3-11〕 sdcMicro를 이용한 전반적 재코딩 후 $IL1$

```
〉 print(sdc.obj,"numrisk")

Numerical key variables: income, debt

Disclosure risk (~100.00% in original data):
  modified data: [0.00%; 100.00%]

Current Information Loss in modified data
                              (0.00% in original data):
  IL1: 0.00
  Difference of Eigenvalues: 0.000%
```

출처: 저자 작성

나. 국소 감추기

감추기(suppression)는 레코드 혹은 속성이 나타나지 않도록 레코드를 변형하는 기법이며 식별 위험을 확실히 낮출 수 있으나 정보 손실이 많다는 단점이 있다. 감추기를 적용할 때는 위험도가 높은 속성 조합과 레코드를 적절히 선택하는 것이 중요하며, 따라서 데이터에 포함된 변수에 대하여 사전에 민감한 정도 등에 따라 순서를 결정하는 것이 합리적이다.

감추기 기법 중 가장 자주 사용되는 기법은 국소 감추기(local suppression)이다. 국소 감추기는 식별 위험이 높은 속성 조합을 가지는 레코드를 선별하고 해당 레코드의 속성 중 하나를 감추는 기법이다. sdcMicro는 국소 감추기 결과 모든 레코드가 k-익명성을 만족하도록 처리한다.

[그림 3-12] sdcMicro를 이용한 국소 감추기

```
> key.imp = sample(1:length(sdc.obj@keyVars))

> sdc.obj = localSuppression(sdc.obj,k=5,
importance=key.imp)

> head(sdc.obj@manipKeyVars)

   area  sex   age  job
1     1    2    38    1
2     1    2    42    1
3    NA    2    44    1
4    NA    2    29    3
5     1    1    22   NA
6     1    1    NA    1

> print(sdc.obj, type="kAnon")

Infos on 2/3-Anonymity:
Number of observations violating
- 2-anonymity: 0 (0.000%) | in original data: 70 (7.000%)
- 3-anonymity: 0 (0.000%) | in original data: 138 (13.800%)
- 5-anonymity: 0 (0.000%) | in original data: 281 (28.100%)
```

출처: 저자 작성

다. 국소 통합

국소 통합(micro-aggregation)은 레코드를 최소 m개 이상의 레코드를 가지는 그룹으로 분할하고 각 그룹의 레코드들을 각 그룹의 대푯값으로 대체하는 기법이다. 레코드를 묶어 그룹을 만들 때 하나의 변수를 사용하거나 여러 변수의 조합을 사용할 수도 있다.

[그림 3-13] sdcMicro를 이용한 국소 통합

```
〉 sdc.obj = microaggregation(sdc.obj,
variables=c("income","debt"),method="simple",aggr=2)

〉 head(data.frame(sdc.obj@origData[,c("income","debt")],
sdc.obj@manipNumVars))

      income  debt     income.1        debt.1
1017     821 49000   793.2862  42031.034
8004     650  6000   793.2862  42031.034
4775     505  7000   516.2262   9245.249
10369    520 10000   516.2262   9245.249
13218     96  2000   376.3818   3262.981
9725     762  5000   376.3818   3262.981
```

출처: 저자 작성

위의 예제에서 income, debt 변수에 대하여 국소 통합을 적용하였다. aggr=2 옵션은 레코드 수가 2가 되도록 그룹을 결정하는 옵션이며, 그룹을 결정한 후 대푯값을 결정하는 옵션은 simple을 사용하였다. 국소 통합 결과 처음 두 레코드가 동일한 값으로 대체된 것을 확인할 수 있다.

sdcMicro는 대용량 데이터에 사용할 수 있는 국소 통합 함수를 따로 제공하고 있다. 아래 예제는 income, debt 변수에 대하여 aggr=4 옵션을 사용하여 sex 변수를 기준으로 그룹을 결정하고 디폴트 옵션인 simple을 적용한 결과를 보여준다.

[그림 3-14] sdcMicro를 이용한 국소 통합: mafast

```
〉 sdc.obj = mafast(obj=sdc.obj,
variables=c('income','debt'),by="sex",aggr=4)

〉 head(data.frame(sam.df$sex,
sdc.obj@origData[,c("income","debt")],
sdc.obj@manipNumVars))

     sam.df.sex income  debt income.1    debt.1
1017          2    821 49000 793.5727 40027.742
8004          2    650  6000 793.5727 45470.076
4775          2    505  7000 515.0675  9253.609
10369         2    520 10000 516.8663  9253.609
13218         1     96  2000 375.2737  3247.502
9725          1    762  5000 375.2737  3312.370
```

출처: 저자 작성

라. 잡음 추가

잡음 추가는 연속형 변수에 평균이 0인 잡음을 추가하여 비식별 처리하는 기법이다. sdcMicro는 상관(correlated) 잡음 추가와 비상관(uncorrelated) 잡음 추가의 두 가지 기법을 제공한다. 상관 잡음 추가는 비식별 처리할 변수들의 표본 공분산 행렬을 추정하고 이에 비례라는 공분산 구조를 가지는 분포에서 잡음을 생성하는 기법이다. 비상관 잡음 추가는 원본 데이터의 표본평균과 표본분산을 평균과 분산으로 가지는 정규분포에서 잡음을 생성하는 기법이다.

〔그림 3-15〕 sdcMicro를 이용한 잡음 추가

```
> sdc.obj = addNoise(sdc.obj,
variables=c('income','debt'),
noise=5,method="correlated")

> head(
data.frame(sdc.obj@origData[,c("income","debt")],
sdc.obj@manipNumVars))

      income  debt income.1     debt.1
1017     821 49000 809.8494 46837.9682
8004     650  6000 880.3443 41028.9144
4775     505  7000 611.7718 14170.4766
10369    520 10000 572.6896 13687.9556
13218     96  2000 457.3052  2191.2132
9725     762  5000 320.3058   992.4173
```

출처: 저자 작성

위 예제에서 잡음을 추가할 변수는 income, debt이고 obj는 sdcMicro 객체뿐만 아니라 matrix와 data frame도 가능하다. noise=5는 잡음을 추가하는 레코드의 비율이며, method=correlated는 상관 잡음 추가를 의미하는 옵션이다. 이외에도 additive, correlated2, restr, ROMM, outdect 등의 옵션이 있다.

제5절 소결

제3장에서는 프라이버시 향상을 위한 데이터 비식별화 용어와 기법 분류에 대한 국제 표준인 ISO/IEC 20889의 비식별화 방법론과 차등 정보

보호, 재현 데이터 방법론을 검토하였다.

변수·레코드 수준 비식별화에서 구조적 방법은 데이터의 레코드 구조를 변형하여 제공하는 방법이며 대표적으로 표본추출 방법이 있다. 삭제 방법으로는 식별 변수 삭제, 레코드 삭제, 마스킹이 있으며, 일반화 방법으로는 라운딩 방법, 범주화 방법을 소개하였다. 임의화 방법은 잡음 첨가, 데이터 교환, 부분 총계 방법이 있으며, 실무에서는 통계적으로 다양한 변수의 분포를 고려해야 하는 임의화 방법보다는 삭제 방법이나 일반화 방법에 대한 접근이 더 쉽다고 판단된다.

차등 정보보호는 데이터 분석 시 개인정보가 노출되지 않도록 보호하는 기술로, 데이터에 노이즈(무작위성을 추가한 값)를 추가하여 개인을 식별할 수 없도록 하면서 데이터의 전체 패턴은 분석할 수 있게 하는 방법이다. Apple은 iOS와 macOS에서 차등 프라이버시 기술을 적용하여 사용자의 활동 정보를 익명으로 수집하고, 그 데이터를 통해 제품의 성능을 개선시킨다(unite.ai, 2022.12.09.). 차등 정보보호는 프라이버시 손실을 정량화할 수 있고, 이를 분석하여 다양한 알고리즘을 개발할 수 있다는 장점이 있지만, 주어진 쿼리에 대해 그 복잡성에 따라 α 값을 다르게 설정해야 하는 등의 한계점도 존재한다.

재현 데이터는 원본 데이터의 통계적 특성을 유지하면서 새로운 가상의 데이터를 생성하는 기법으로, 실제 개인정보는 포함되어 있지 않도록 생성할 수 있다. 개인정보 관련 규제로 데이터의 활용이 제한된 상황에서는 재현 데이터가 좋은 대안이 될 수 있다. 재현 데이터는 통계청, 신용정보원 등에서도 연구를 진행하고 있으며, 스위스 보험회사인 Die Mobiliar는 재현 데이터로 익명화한 뒤 소비자 개인정보를 보호하면서도 효과적인 마케팅 전략을 수립한 사례가 있다(김현태 외, 2023).

〈표 3-14〉 차등 정보보호와 재현 데이터 비교

	차등 정보보호 (Differential Privacy)	재현 데이터 (Synthetic Data)
목적	개별 데이터의 프라이버시 보호	데이터 분석 및 테스트를 위한 안전한 데이터 제공
작동 방식	노이즈 추가를 통해 데이터나 결과를 보호	원본 데이터를 기반으로 통계적 특성을 모방한 인공 데이터 생성
데이터 형태	실제 데이터에 노이즈를 추가한 변형된 데이터 또는 분석 결과	원본과 유사하지만 실제 개인정보를 포함하지 않는 새로운 데이터
프라이버시 보장 방식	수학적 보장을 통해 개별 데이터의 존재 여부를 숨김	실제 개인정보를 포함하지 않음으로써 간접적으로 보호
적용 범위	실시간 데이터 접근, 쿼리 응답 등	데이터 공유, 테스트, 연구 등
장점	강력한 수학적 프라이버시 보장, 다양한 분석에 적용 가능	데이터 공유가 용이하고, 원본 데이터의 민감한 정보를 포함하지 않음
단점	노이즈 추가로 인해 데이터의 정확성이 약간 저하될 수 있음	완벽하게 원본 데이터와 동일한 분석 결과를 보장하지 않을 수 있음

출처: OpenAI(2024)

차등 정보보호와 재현 데이터는 상호 보완적으로 사용될 수 있으며, 앞으로도 활용도가 높아질 것으로 예상할 수 있기 때문에 본 연구에서도 방법론적으로 중요하게 다루었다.

마지막으로 비식별화 방법론의 실무 적용을 위해 sdcMico를 사용하여 활용 예시를 구체적으로 제시하였다. 이 부분은 제4장의 노출 위험 분석과도 연관되어 있고, 제5장의 원내 비식별화 가이드라인에도 활용될 수 있는 부분이다.

제4장

비식별화 데이터 생성 및 노출 위험 분석

제 4 장 비식별화 데이터 생성 및 노출 위험 분석

이 장에서는 한국복지패널(2023), 가족과 출산 조사(2021), 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사(2022)를 분석하여 비식별화 데이터를 생성하고 개인 노출 위험을 측정하여 비교해보고자 한다. 한국복지패널과 가족과 출산 조사는 국가 승인 통계로 현재 지역 구분을 7개 권역, 3개 권역으로 각각 제공하고 있는데, 지역 구분을 시도 레벨까지 공개하였을 경우 개인 노출 위험은 어느 정도로 증가하는지에 대한 분석 결과도 함께 제시하여 시도 공표 가능성도 살펴보았다. 분석 변수는 준식별 변수인 공통 변수와 민감 변수로 정의하였다. 공통변수는 다른 데이터와의 결합을 염두에 둘 때 기준이 될 수 있는 변수로 인구사회학적 항목들을 공통변수로 보았고, 민감 변수는 조사 결과로 나올 수 있는 항목으로 공격자가 알고 싶은 변수로 정하였다. 노출 위험 분석은 R 패키지의 sdcMicro를 사용하였다.

제1절 한국복지패널

1. 분석 개요

2006년에 구축한 한국복지패널은 빈곤층, 근로빈곤층(working poor), 차상위층(near poor)의 가구 형태, 소득 수준, 취업 상태가 급격히 변화하는 상황에서 이러한 계층의 규모 및 생활실태 변화를 동태적으

로 파악함으로써 정책 형성에 이바지함과 동시에 정책 지원에 따른 효과성을 높이고자 하는 목적하에 2023년 제18차 조사를 완료하였다.

복지패널 조사는 국가승인통계로 7개 권역, 5개 권역별로 지역을 구분하여 공개하고 있다.

〈표 4-1〉 복지패널 활용 변수

변수 구분	변수명	변수 설명	문항 내용
공통 변수 (key 변수)	h18_reg7	7개 권역별 지역구분	1. 서울, 2. 수도권(인천/경기), 3. 부산/경남/울산, 4. 대구/경북, 5. 대전/충남/세종, 6. 강원/충북, 7. 광주/전남/전북/제주도
	reg16	16개 시도	(비공개 변수) (세종은 충남에 포함)
	h18_pid	개인 패널ID	
	h18_g3	성별	1. 남, 2. 여
	h18_g4	태어난 연도	년
	h18_g6 h18_g7	교육수준	1. 미취학(만 7세 미만), 2. 무학(만 7세 이상), 3. 초등학교, 4. 중학교, 5. 고등학교, 6. 전문대학, 7. 대학교, 8. 대학원(석사), 9. 대학원(박사)
			0. 비해당, 1. 재학, 2. 휴학, 3. 중퇴, 4. 수료, 5. 졸업
	h18_g8	장애 종류	0. 비해당(비장애인), 1. 지체장애, 2. 뇌병변장애, 3. 시각장애, 4. 청각장애, 5. 언어장애, 6. 정신지체(지적장애), 7. 발달장애(자폐성장애), 8. 정신장애, 9. 신장장애, 10. 심장장애, 11. 호흡기장애. 12. 간장애, 13. 안면장애. 14. 장루, 요루 장애, 15. 간질장애(뇌전증장애), 16. 비등록 장애인(보훈처 등록 장애인 포함)
	h18_g9	장애 정도	0. 비해당(비장애인), 1. 장애 정도가 심한 장애인, 2. 장애 정도가 심하지 않은 장애인, 3. 비등록 장애인(보훈처 등록 장애인 포함)
	h18_g10	혼인상태	0. 비해당(18세 미만) 1. 유배우, 2. 사별, 3. 이혼, 4. 별거, 5. 미혼(18세 이상, 미혼모 포함), 6. 기타(사망 등)
	g4_group	연령 범주	10세 단위
	marrst	혼인상태	1. 미혼, 2. 기혼(사실혼, 이혼, 사별 포함)

변수 구분	변수명	변수 설명	문항 내용
		범주	
	edu_g	교육수준 범주	1. 고등학교 졸업 이하, 2. 대학 졸업(2~3년제, 전문대, 4년제 이상), 3. 대학원 졸업(석사, 박사)
민감 변수	h18_eco2	근로능력 정도	0. 만 14세 이하, 1. 근로 가능. 2. 단순근로 가능(집에서 돈벌이를 할 수 있는 정도), 3. 단순근로 미약자(집안일만 가능), 4. 근로능력 없어 경제활동을 하지 않음(집안일도 불가능)
	h18_eco4	주된 경제활동 참여 상태	1. 상용직 임금근로자, 2. 임시직 임금근로자, 3. 일용직 임금근로자. 4. 자활근로, 공공근로, 노인일자리, 5. 고용주, 6. 자영업자, 7. 무급 가족종사자, 8. 실업자(지난 4주간 적극적으로 구직활동을 함), 9. 비경제활동인구
	empl	취업 여부	1. 취업, 2. 비취업
	h18_eco8	업종	농업, 임업 등 중분류(2자리)
	h18_eco9	직종	관리자, 전문가 및 관련 종사자 등 소분류(4자리)
	h1807_9	총생활비	월평균 지출액(단위: 만 원)
	h18_hc_n_all	균등화소득에 따른 가구 구분	1. 일반가구, 2. 저소득층 가구
	h18_din	가처분소득	소수점 4자리(단위: 만 원)
	h18_cin	경상소득	소수점 4자리(단위: 만 원)
	h1801_11aq2	생계급여 수급 형태1	0. 해당 없음, 1. 일반 수급 가구, 2. 조건부 수급 가구, 3. 특례가구
	h1801_11aq3	생계급여 수급 형태2	0. 해당 없음, 1. 가구원 전부 수급, 2. 가구원 중 일부 수급
	h1801_11aq5	의료급여 수급 형태1	0. 해당 없음, 1. 의료급여 1종, 2. 의료급여 2종, 3. 국가유공자 무료진료
	h1801_11aq6	의료급여 수급 형태2	0. 해당 없음, 1. 가구원 전체 의료급여, 2. 가구원 일부 의료급여
	h1801_11aq8	주거급여 수급 형태	0. 해당 없음, 1. 임차급여(특례 포함), 2. 수선유지 급여(특례 포함)
	h1801_11aq10	교육급여 수급자 수	명(0명은 해당 없음)

주: 연령 범주, 혼인상태 범주, 교육수준 범주, 취업 여부 변수는 본 분석을 위해 생성한 변수임.
출처: 한국보건사회연구원. (2023). 한국복지패널조사 코딩북.
https://www.koweps.re.kr:442/data/book/list.do

18차 복지패널 분석 대상자 수는 15,931명이고, 통합표본의 일반 가중치를 적용하여 분석하였다.

2. 18차 한국복지패널 노출 위험 분석

노출 위험 시나리오는 활용 변수를 다르게 설정하여 6개의 분석을 실시하였다. 분석 1은 원본 데이터로 7개 권역 변수, 성별, 생년, 교육 정도, 장애 정도, 혼인상태, 주된 경제활동 참여 형태, 직종, 총생활비, 경상소득 변수를 활용한 것이고, 분석 2는 분석 1의 변수에서 직종 변수를 제외하고 연령, 경상소득을 범주형 변수로 비식별화 처리를 하였다. 분석 3은 분석 2의 변수에서 교육수준, 혼인상태, 경제활동상태를 범주형 변수로 비식별화 처리를 하여 노출 위험도를 측정하였다. 분석 4~6은 현재 7개 권역으로 공개하고 있는 지역 변수를 16개 시도 변수(세종시는 충남에 포함)로 공개하였을 시, 노출 위험을 측정하기 위해 7개 권역별 변수를 16개 시도 변수로 바꾸어 분석하였다. 18차 복지패널의 노출 위험 측도는 k-익명성과 전체 위험도(global risk)로 측정하였다.

〈표 4-2〉 18차 복지패널 노출 위험 시나리오 활용 변수

구분	활용 변수
분석 1	7개 권역(reg7), 성별(g3), **생년(g4)**, 교육 정도(g6), 장애 정도(g9), 혼인상태(g10), 주된 경제활동참여 상태(eco4), 직종(eco9), 총생활비(h1807_9), 경상소득(cin)
분석 2	7개 권역(reg7), 성별(g3), **연령 범주(g4_group)**, 교육 정도(g6), 장애 정도(g9), 혼인상태(g10), 주된 경제활동참여 상태(eco4), 총생활비(h1807_9), **균등화소득에 따른 가구 구분(hc_n_all)**
분석 3	7개 권역(reg7), 성별(g3), **연령 범주(g4_group)**, **교육수준_범주(edu_g),** 장애 정도(g9), **혼인상태(marrst)**, **취업 여부(empl),** 총생활비(h1807_9), **균등화소득에 따른 가구 구분(hc_n_all)**

구분	활용 변수
분석 4	16개 시도(reg16), 성별(g3), **생년(g4)**, 교육 정도(g6), 장애 정도(g9), 혼인상태(g10), 주된 경제활동참여 상태(eco4), 직종(eco9), 총생활비(h1807_9), 경상소득(cin)
분석 5	16개 시도(reg16), 성별(g3), **연령 범주(g4_group)**, 교육 정도(g6), 장애 정도(g9), 혼인상태(g10), 주된 경제활동참여상태(eco4), 총생활비(h1807_9), **균등화소득에 따른 가구 구분(hc_n_all)**
분석 6	16개 시도(reg16), 성별(g3), **연령 범주(g4_group)**, **교육수준_범주(edu_g)**, 장애 정도(g9), **혼인상태(marrst)**, **취업 여부(empl)**, 총생활비(h1807_9), **균등화소득에 따른 가구 구분(hc_n_all)**

출처: 한국보건사회연구원. (2023). 한국복지패널조사 코딩북.
https://www.koweps.re.kr:442/data/book/list.do

분석 1은 데이터의 약 64%가 2-익명성을 만족하지 않으며, 약 84%가 5-익명성을 만족하지 않는다. 직종 변수를 제외하고 연령 및 가구소득을 범주화 변수로 처리한 분석 2는 데이터의 약 20%가 2-익명성을 만족하지 않으며, 약 43%가 5-익명성을 만족하지 않는다. 직종 변수를 제외하고 연령, 가구소득, 교육수준, 혼인상태, 경제활동상태를 범주화 변수로 처리한 분석 3은 데이터의 약 4%가 2-익명성을 만족하지 않으며, 약 13%가 5-익명성을 만족하지 않는다.

전체 위험도는 데이터 전체의 식별 위험을 추정하는 방법으로, 분석 1에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.36%이고, 분석 3에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.03%이다.

〈표 4-3〉 18차 복지패널 노출 위험도 추정 1

노출 위험	분석 1	분석 2	분석 3
2-anonymity (Number of observations violating)	10228 (64.20%)	3298 (20.70%)	708 (4.44%)
3-anonymity (Number of observations	12020 (75.45%)	4984 (31.29%)	1248 (7.83%)

노출 위험	분석 1	분석 2	분석 3
violating)			
5-anonymity (Number of observations violating)	13465 (84.52%)	6938 (43.55%)	2090 (13.12%)
global risk (Expected number of re-identifications)	58 (0.36%)	23 (0.14%)	5 (0.03%)

출처: 한국보건사회연구원. (2023). 한국복지패널조사 데이터.
https://www.koweps.re.kr:442/data/data/list.do

분석 4는 데이터의 약 72%가 2-익명성을 만족하지 않으며, 약 92%가 5-익명성을 만족하지 않는다. 분석 5는 데이터의 약 30%가 2-익명성을 만족하지 않으며, 약 59%가 5-익명성을 만족하지 않는다. 분석 6은 데이터의 약 8%가 2-익명성을 만족하지 않으며, 약 21%가 5-익명성을 만족하지 않는다. 7개 지역 구분에서 16개 시도로 공개 범위를 확대했을 시에는 k-익명성의 노출 위험 수준이 매우 높아지는 것을 알 수 있다.

〈표 4-4〉 18차 복지패널 노출 위험도 추정 2

노출 위험	분석 4	분석 5	분석 6
2-anonymity (Number of observations violating)	11502 (72.19%)	4806 (30.17%)	1242 (7.79%)
3-anonymity (Number of observations violating)	13240 (83.11%)	6994 (43.90%)	2108 (13.23%)
5-anonymity (Number of observations violating)	14704 (92.29%)	9515 (59.73%)	3343 (20.98%)
global risk (Expected number of re-identifications)	68 (0.43%)	34 (0.21%)	9 (0.06%)

출처: 한국보건사회연구원. (2023). 한국복지패널조사 데이터.
https://www.koweps.re.kr:442/data/data/list.do

전체 위험도를 보면 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.43%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.06% 정도이다.

l-Diversity Measures 결과는 데이터 세트의 민감한 정보를 보호하기 위해 다양한 수준의 익명성을 평가하는 데 사용하며, 앞에서 언급한 바와 같이 k-익명성을 보완할 수 있다. l-다양성은 데이터에 포함된 민감한 값들이 충분히 다양하게 분포되어 있는지를 나타내며, 이를 통해 재식별 위험을 줄인다. 여기에서는 Distinct l-Diversity(서로 다른 민감한 값의 개수)를 사용하여 각 그룹에서 민감한 속성의 값들이 얼마나 다양한지를 계산한다. 한 그룹에 포함된 민감한 정보가 모두 동일하다면, 그 그룹은 l-다양성이 낮다. 높은 l-다양성 값은 해당 그룹의 민감한 정보가 충분히 다양하게 분포되어 있음을 의미하며, 익명성 수준이 높다고 볼 수 있다. 반면에, 낮은 l-다양성 값은 해당 그룹의 민감한 정보가 재식별될 위험이 크다는 것을 의미한다.

복지패널 데이터에 포함된 범주형 식별 변수의 속성 조합은 2개로 7개 권역별 지역 변수를 사용하였을 때와 16개 시도 지역 변수를 사용하였을 때 l-다양성의 추정을 비교하고자 하였다. 첫 번째 조합은 7개 권역(reg7), 성별(g3), 연령 범주(g4_g), 주된 경제활동 참여 상태(eco4)이고, 두 번째 조합은 16개 시도(reg16), 성별(g3), 연령 범주(g4_g), 주된 경제활동참여상태(eco4)이다. 분석에 활용한 민감 변수는 총생활비(h1807_9), 경상소득(cin), 생계급여 수급 형태1(h1801_11aq2)이며 세 민감 변수의 l-다양성을 추정한 분포는 다음 표와 같다. 16개 시도별 지역 변수를 속성 조합으로 사용하였을 때는 7개 권역별 지역 변수를 사용하였을 때보다 낮은 l-다양성 값을 보여주고 있고, 이는 해당 그룹의 민감한 정보가 재식별될 위험이 상대적으로 높다는 것을 의미한다.

〈표 4-5〉 18차 복지패널 노출 위험도 l-다양성 추정

	7개 권역별 지역 변수 사용			16개 시도 지역 변수 사용		
	총생활비 (h1807_9)	경상소득 (cin)	생계급여 수급 형태 (h1801_11aq2)	총생활비 (h1807_9)	경상소득 (cin)	생계급여 수급 형태 (h1801_11aq2)
최소	1	1	1	1	1	1
1분위	16	16	1	7	7	1
중앙값	37	38	2	17	18	1
평균	45	51	1	25	27	1
3분위	70	75	2	36	37	2
최대	137	207	3	105	118	3

출처: 한국보건사회연구원. (2023). 한국복지패널조사 데이터.
https://www.koweps.re.kr:442/data/data/list.do

제2절 가족과 출산 조사

1. 분석 개요

가족과 출산 조사는 국내에서 유일하게 출산과 결혼 행동의 이력을 체계적으로 파악하는 대표적인 조사로, 주요 목적은 결혼과 출산 등 인구학적 행동을 중심으로 개인의 생애과정과 가족 경로의 변화를 관찰할 수 있는 자료를 수집하는 것이다(박종서 외, 2021).[1] '2021년도 가족과 출산 조사'(이하 가족과 출산 조사)는 그동안 매 3년 주기로 실시하던 '전국 출산력 및 가족 보건·복지실태 조사'(이하 출산력 조사)의 새로운 이름이다.

가족과 출산 조사의 조사표 내용 구성은 〈표 4-6〉과 같다.

2021년도 가족과 출산 조사에서 노출 위험을 측정하기 위해 검토한 변

1) 요약 7페이지 인용

수는 인구학적 속성을 지닌 준식별 변수(공통변수, keyvar)와 소득, 부채 등과 같은 민감 변수이다. 분석 대상자 수는 14,538명이며, 가중치는 표본가중치를 적용하여 분석하였다.

〈표 4-6〉 2021년도 가족과 출산 조사 조사표 내용 구성

구분	주요 내용
가구 및 가구원 사항	- 이름, 관계, 성, 연령, 교육, 혼인상태, 취업 여부, 종교, 국적, 동거 여부 - 주거, 소득, 지출, 자산, 부채
부모와의 관계	- 부모 연령, 학력, 생존 여부, 직업, 경제 상황 - 세대 간 자원의 교환 및 세대관계
동거와 결혼	- 유배우자: 결혼 시점, 동거 시점, 혼인신고 시점, 결혼식 시점, 결혼 기대 척도, 결혼 이력 - 동거 중: 시작 시기, 동거 기대 척도, 결혼이행, 차별 - 미혼(기혼싱글): 결혼 의향, 결혼 기대 척도, 동거 의향 등 - 이혼 경험자: 자녀 유무, 양육비 지급 사항
임신 출산 건강	- 피임: 생식계 건강, 대처방안, 피임 인지도, 경험, 방법, 이유, 의사결정 - 임신 출산: 임신 횟수, 계획 여부, 결과, 총 출생아 수, 결혼 당시 출산계획, 향후 출산 의향, 의사결정 척도 - 난임(불임): 경험, 검사, 원인, 시술 경험, 결과, 기간 등
산전 산후 관리	- 산전, 분만: 진찰 장소, 횟수, 초진 시기, 불편사항, 분만 장소, 자연분만 여부, 분만 시기와 체중 - 산후관리: 산후진찰 경험, 산후조리 장소, 산후우울 경험 및 치료 - 수유: 방법, 정보취득, 모유 여부 및 계획, 이력
자녀 양육	- 분담: 육아와 가사 시간 분담, 만족도 - 미취학돌봄: 희망 돌봄 유형, 현재 유형, 주체, 만족도, 어려움 - 취학돌봄: 희망 돌봄 유형, 현재 유형, 주체, 만족도, 어려움 - 양육비: 돌봄기관, 공/사교육비, 돌봄인력 비용, 기타
일	- 본인과 배우자 현 취업, 항목별 시간량과 적정성, 생애사건 시 취업 이력
성장기와 주거 이동	- (15세 당시) 거주지역, 동거 부모, 분거 경험, 15세 때 경제 형편 - 주거 독립, 자립 인식, 결혼 전후 주거
가치관과 인식	- 성역할 태도와 가치, 자녀 출산 태도, 사회 신뢰

출처: 박종서 외. (2021). 2021년도 가족과 출산 조사-(구)전국 출산력 및 가족보건복지 실태조사. p.19.

〈표 4-7〉 2021년도 가족과 출산 조사 활용 변수

변수 구분	변수명	변수 설명	문항 내용
공통 변수 (key 변수)	area1	지역(동부/읍면부)	① 동부, ② 읍면부
	area2	지역(대도시/ 중소도시/농어촌)	① 대도시, ② 중소도시, ③ 소도시
	sido	17개 시도	(비공개 변수)
	hpid	가구ID+가구원번호	
	sex	성별	① 남, ② 여
	birthy	생년(양력기준)	
	edu1 edu2	교육 정도 교육상태	⓪ 미취학, ① 무학. ② 초등학교, ③ 중학교, ④ 고등학교, ⑤ 대학(2~3년제, 전문대 포함). ⑥ 대학교(4년제 이상), ⑦ 대학원(석사). ⑧ 대학원(박사)
			① 졸업, ② 수료, ③중퇴, ④ 재학 휴학
	marr	혼인상태	① 미혼, ② 배우자 있음, ③ 이혼, ④ 별거, ⑤ 사별
	age_g	연령 범주 (5세 구간)	① 19~24세. ② 25~29세. ③ 30~34세, ④ 35~39세, ⑤ 40~44세. ⑥ 45~49세, ⑦ 50세 이상
	marrst	혼인상태	① 미혼, ② 기혼(사실혼, 이혼, 사별 포함)
	edu_g	교육수준_범주	④ 고등학교 졸업 이하, ⑤ 대학 졸업(2~3년제, 전문대, 4년제 이상), ⑦ 대학원 졸업(석사, 박사)
민감 변수	job	경제활동 상태	① 취업, ② 실업, ③ 비경제활동
	h0405	합계소득(월평균)	단위: 만 원
	h0408	부채	단위: 만 원
	c07	총 임신 횟수	
	c10	난임 여부	① 있다 ☞ 문항 10-1, ② 없다 ☞ 문항 12, ③ 비해당 ☞ 문항 12
	f0104	직종	① 관리자, ② 전문가 및 관련 종사자, ③ 사무 종사자, ④ 서비스 종사자, ⑤ 판매 종사자. ⑥ 농림어업 숙련 종사자. ⑦ 기능원 및 관련 기능종사자, ⑧ 장치·기계조작 및 조립 종사자, ⑨ 단순노무종사자, ⑩ 군인
	f0105	업종	① 농림어업, ② 광업·제조업, ③ 건설업, ④ 도매 및 소매업, ⑤ 숙박 및 음식점업, ⑥ 금융 및 보험업, ⑦ 교육서비스업,

변수 구분	변수명	변수 설명	문항 내용
			⑧ 전기, 가스, 증기 및 공기조절 공급업, ⑨ 운수 및 창고업, 정보통신업, ⑩ 전문, 과학 및 기술서비스업, ⑪ 공공행정, 국방 및 사회보장행정, ⑫ 보건업 및 사회복지서비스업, ⑬ 사업시설관리, 사업지원 및 임대서비스업, ⑭ 기타(부동산, 수도·하수·폐기물 처리, 예술 등)
	f0106	직장 유형	① 정부기관(공무원 및 군인, 국공립 교사 등), ② 정부외공공기관(정부투자 · 출자기관, 정부출연기관, 정부보조위탁기관, 자회사, 재출연기관 등), ③ 민간 대기업(300인 이상), ④ 민간 중기업(50~299인), ⑤ 민간 소기업(5~49인), ⑥ 개인사업체(5인 미만), ⑦기타
	tinc_g	가구소득_범주	① 60% 미만, ② 60~80% 미만, ③ 80~100% 미만, ④ 100~120% 미만. ⑤ 120~140% 미만, ⑥ 140~160% 미만, ⑦ 160% 이상
	empl	취업 여부	① 취업, ② 비취업

출처: 한국보건사회연구원. (2021). 가족과 출산조사 코딩북.
https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

2. 2021년도 가족과 출산 조사 노출 위험 분석

노출 위험 시나리오는 활용 변수를 다르게 설정하여 6개의 분석을 실시하였다. 분석 1은 원본 데이터로 지역 변수, 성별, 생년, 교육수준, 혼인상태, 직종, 월평균 소득, 부채 변수를 활용한 것이고, 분석 2는 분석 1의 변수에서 직종 변수를 제외하고 연령, 가구소득을 범주형 변수로 비식별화 처리를 하였다. 분석 3은 분석 2의 변수에서 교육수준, 혼인상태, 경제활동상태를 범주형 변수로 비식별화 처리를 하여 노출 위험도를 측정하였다.[2)]

2) 분석 2, 분석 3의 범주화 변수들은 2021년 가족과 출산 조사에서 제공하고 있는 변수로,

분석 4~6은 현재 3개 권역으로 공개하고 있는 지역 변수를 17개 시도 변수로 공개하였을 때 발생할 노출 위험을 측정하기 위해 3개 권역별 변수를 17개 시도 변수로 바꾸어 분석한 것이다. 2021년도 가족과 출산 조사의 노출 위험 측도는 k-익명성과 전체 위험도(global risk)로 측정하였다.

〈표 4-8〉 2021년도 가족과 출산 조사 노출 위험 시나리오 활용 변수

분석	활용 변수
분석 1	동부/읍면부(area1), 대도시/중소도시/농어촌(area2), 성별(sex), **생년(birthy)**, 교육 정도(edu1), 혼인상태(marr), 경제활동상태(job), **직종(f0104)**, **월평균 소득(h0405)**, 부채(h0408)
분석 2	동부/읍면부(area1), 대도시/중소도시/농어촌(area2), 성별(sex), **연령 범주(age_g)**, 교육 정도(edu1), 혼인상태(marr), 경제활동상태(job), **가구소득_범주(tinc_g)**, 부채(h0408)
분석 3	동부/읍면부(area1), 대도시/중소도시/농어촌(area2), 성별(sex), **연령 범주(age_g)**, **교육수준_범주(edu_g)**, **혼인상태(marrst)**, **취업 여부(empl)**, **가구소득_범주(tinc_g)**, 부채(h0408)
분석 4	동부/읍면부(area1), 17개 시도(sido), 성별(sex), **생년(birthy)**, 교육 정도(edu1), 혼인상태(marr), 경제활동상태(job), **직종(f0104)**, **월평균 소득(h0405)**, 부채(h0408)
분석 5	동부/읍면부(area1), 17개 시도(sido), 성별(sex), **연령 범주(age_g)**, 교육 정도(edu1), 혼인상태(marr), 경제활동상태(job), **가구소득_범주(tinc_g)**, 부채(h0408)
분석 6	동부/읍면부(area1), 17개 시도(sido), 성별(sex), **연령 범주(age_g)**, **교육수준_범주(edu_g)**, **혼인상태(marrst)**, **취업 여부(empl)**, **가구소득_범주(tinc_g)**, 부채(h0408)

출처: 한국보건사회연구원. (2021). 가족과 출산조사 코딩북. https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

분석 1은 데이터의 약 25%가 2-익명성을 만족하지 않으며, 약 56%가 5-익명성을 만족하지 않는다. 직종 변수를 제외하고 연령 및 가구소득을 범주화 변수로 처리한 분석 2는 데이터의 약 10%가 2-익명성을 만족하

노출 위험도 측정을 위해 범주화된 변수 활용 시 원 변수는 분석에 포함시키지 않았음.

지 않으며, 약 31%가 5-익명성을 만족하지 않는다. 직종 변수를 제외하고 연령, 가구소득, 교육수준, 혼인상태, 경제활동상태를 범주화 변수로 처리한 분석 3은 데이터의 약 4%가 2-익명성을 만족하지 않으며, 약 15%가 5-익명성을 만족하지 않는다.

전체 위험도는, 데이터 전체의 식별 위험을 추정하는 방법으로, 분석 1에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 46%임에 반해, 분석 3에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 15% 정도로 줄어드는 것을 알 수 있다.

〈표 4-9〉 2021년도 가족과 출산 조사 노출 위험도 추정 1

노출 위험	분석 1	분석 2	분석 3
2-anonymity (Number of observations violating)	3666 (25.22%)	1583 (10.89%)	667 (4.59%)
3-anonymity (Number of observations violating)	5790 (39.82%)	2729 (18.77%)	1289 (8.87%)
5-anonymity (Number of observations violating)	8181 (56.27%)	4520 (31.09%)	2472 (17.00%)
global risk (Expected number of re-identifications)	6714 (46.18%)	3899 (26.82%)	2320 (15.96%)

출처: 한국보건사회연구원. (2021). 가족과 출산조사 마이크로데이터. https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

분석 4는 데이터의 약 56%가 2-익명성을 만족하지 않으며, 약 89%가 5-익명성을 만족하지 않는다. 분석 5는 데이터의 약 36%가 2-익명성을 만족하지 않으며, 약 78%가 5-익명성을 만족하지 않는다. 분석 6은 데이터의 약 23%가 2-익명성을 만족하지 않으며, 약 65%가 5-익명성을 만족하지 않는다.

전체 위험도를 보면 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 83%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 52% 정도이다. 분석 1과 분석 4를 비교하였을 때 3개 권역별 변수에서 17개 시도 변수로 공개 범위를 확대할 경우, 노출 위험이 46%에서 83%로 증가하는 것을 알 수 있다.

〈표 4-10〉 2021년도 가족과 출산 조사 노출 위험도 추정 2

노출 위험	분석 4	분석 5	분석 6
2-anonymity (Number of observations violating)	8260 (56.81%)	5321 (36.60%)	3402 (23.40%)
3-anonymity (Number of observations violating)	11052 (76.02%)	8277 (56.93%)	6056 (41.65%)
5-anonymity (Number of observations violating)	13067 (89.88%)	11373 (78.23%)	9480 (65.21%)
global risk (Expected number of re-identifications)	12158 (83.63%)	9711 (66.80%)	7688 (52.89%)

출처: 한국보건사회연구원. (2021). 가족과 출산조사 마이크로데이터.
https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

l-다양성 추정으로는 Distinct *l*-Diversity(서로 다른 민감한 값의 개수)를 사용하여 각 그룹에서 민감한 속성의 값들이 얼마나 다양한지를 계산한다. 가족과 출산 조사 데이터에 포함된 범주형 식별 변수의 속성 조합은 2개로 3개 권역별 지역 변수를 사용하였을 때와 17개 시도 지역 변수를 사용하였을 때 *l*-다양성의 추정을 비교하고자 하였다. 첫 번째 조합은 동부/읍면부(area1), 대도시/중소도시/농어촌(area2), 성별(sex), 연령 범주(age_g), 경제활동상태(job)이고, 두 번째 조합은 동부/읍면부(area1), 17개 시도(sido), 성별(sex), 연령 범주(age_g), 경제활동상태

(job)이다. 분석에 활용한 민감 변수는 월평균 소득(h0405), 부채(h0408), 총 임신 횟수(c07)이며 세 민감 변수의 l-다양성을 추정한 분포는 다음 표와 같다. 17개 시도별 지역 변수를 속성 조합으로 사용하였을 때는 3개 권역별 지역 변수를 사용하였을 때보다 낮은 l-다양성 값을 보여주고 있고, 이는 해당 그룹의 민감한 정보가 재식별될 위험이 상대적으로 높다는 것을 의미한다.

〈표 4-11〉 2021년도 가족과 출산 조사 노출 위험도 l-다양성 추정

	3개 권역별 지역 변수 사용			17개 시도 지역 변수 사용		
	월평균 소득(h0405)	부채 (h0408)	총 임신 횟수(c07)	월평균 소득(h0405)	부채 (h0408)	총 임신 횟수(c07)
최소	1	1	1	1	1	1
1분위	97	31	1	17	9	1
중앙값	195	53	4	31	14	2
평균	180	47	4	41	17	3
3분위	265	66	7	54	23	5
최대	319	78	9	137	51	8

출처: 한국보건사회연구원. (2021). 가족과 출산조사 마이크로데이터.
https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

제3절 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사

1. 분석 개요

정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사(2022)는 “사회정신건강연구센터 운영 지역사회 거주 정신질환자의 건강증진 및 복지서비스 지원 방안(전진아 외, 2022)” 연구에서 진행된 조사이다. 이

연구의 목적은 정신질환자가 가지는 건강 상태, 건강 관리 및 복지서비스 이용 경험, 욕구를 파악하여 지역사회에 거주하는 정신질환자의 건강 및 복지서비스 현황과 수요를 파악하고 지원 전략을 도출하는 것이다. 설문조사 대상은 정신질환자와 가족(혹은 보호자)이며, 조사 시점은 2022년 9월 26일부터 10월 31일까지, 정신질환자들은 우편조사, 가족(혹은 보호자)은 우편조사와 온라인 웹 조사를 병행하여 진행하였다. 설문조사표는 "정신질환자(정신장애인)의 건강 및 복지서비스 인식 및 이용 경험조사"와 "정신질환자(정신장애인) 가족 보호자의 건강 및 복지서비스 인식 및 이용 경험 조사" 두 가지이며, 본 연구에서는 정신질환자 당사자의 조사 결과를 활용하여 개인식별 노출 위험을 살펴보았다.

〈표 4-12〉 지역사회 거주 정신질환자의 설문조사 주요 내용

구분		조사 항목	
일반적 특성	인구학적 특성	• 거주지역 • 성별 • 연령 • 학력	• 결혼상태 • 동거가족 • 주된 의사결정자
	사회경제적 특성	• 소득 • 주관적 경제적 상태	• 의료보장 유형
	장애 특성	• 장애등록 여부 • 장애등록 과정	• 장애등록하지 않은 이유 • 관심 설문 영역
건강수준 및 삶의 만족도	건강수준	• 주관적 건강수준 (신체, 정신) • 신체이미지 • 흡연 • 음주 • 운동 • 영양 • 눈 건강	• 만성질환 여부 • 우울감 • 만성질환 자기효능감 • 의료이용 경험 • 미충족의료 경험 • 약 복용 • 신체 건강 관련 지원 및 서비스(필요도, 이용 경험, 도움 정도)
	삶의 만족도	• 삶의 만족도	
	일상생활 기능 정도	• 일상생활 기능 정도	• 일상생활 도움 여부

구분			조사 항목	
서비스 이용 경험	전반적인 서비스 이용 경험		• 기관 및 서비스 인지 여부 • 기관 및 서비스 참여 여부 • 기관 및 서비스 이용 시 만족하지 못한 이유	• 서비스 이용 어려운 점 • 서비스 이용 어려움 정도 • 서비스 이용 제약 이유
	고용 및 소득보장 관련 지원 및 서비스		• 현재 근로 여부 • 직종 • 근로시간 • 급여 • 근로 중 애로사항 • 취업 유지 가능성	• 근로하고 있지 않는 이유 • 취업으로 얻는 이득 • 고용 및 직업 관련 제도나 서비스의 필요 정도, 이용 경험, 도움 정도
	자립 지원 관련 지원 및 서비스		• 자립생활에 대한 감정 • 자립에 대한 우려 수준 • 주거 관련 정보 이해 정도	• 자립 관련 제도나 서비스의 필요 정도, 이용 경험, 도움 정도
	기타 영역 서비스	교육	• 교육 관련 제도나 서비스의 필요 정도, 이용 경험, 도움 정도	• 향후 참여하고 싶은 평생교육 프로그램 영역 • 평생교육 프로그램에 참여하기 위해 필요한 지원
		사회 활동, 문화, 여가	• 지난 1주일간 참여한 문화 및 여가 활동 • 참여를 희망하는 문화 및 여가 활동	• 문화 및 여가활동 만족도 • 문화 및 여가활동을 만족스럽게 보내지 못하는 이유
기타			• 현재 생활에 대한 만족도 • 지역사회 거주 가능성	

출처: 전진아 외. (2022). 사회정신건강연구센터 운영: 지역사회 거주 정신질환자의 건강증진 및 복지서비스 지원 방안. 한국보건사회연구원. p.24.

분석 대상자 수는 652명이며, 가중치는 없이 분석하였다.

〈표 4-13〉 지역사회 거주 정신질환자 조사 활용 변수

변수 구분	변수명	변수 설명	문항 내용
공통 변수 (key 변수)	A2	성별	① 남성, ② 여성
	A3_1	생년월일(년)	1942~2002
	A3_NEW	생년월일(년) 범주화	① 20~29세, ② 30~39세, ③ 40~49세, ④ 50~59세, ⑤ 60세 이상
	A4_1 A4_2	학교 졸업상태	① 무학, ② 초등학교, ③ 중학교, ④ 고등학교, ⑤ 대학(2~3년제, 전문대 포함),

변수 구분	변수명	변수 설명	문항 내용
			⑥ 대학교(4년제 이상), ⑦ 대학원 이상
			① 재학, ②중퇴, ③졸업, ④휴학
	A4_NEW	교육수준 범주	① 중학교 졸업 이하, ② 고등학교 졸업, ③ 대학교 졸업 이상, ④ 모름/무응답
	A5	결혼상태	① 미혼, ② 결혼, ③ 동거(함께 사는 배우자 있음), ④ 별거, ⑤ 이혼, ⑥ 사별, ⑦ 기타
	A5_NEW	결혼상태 범주	① 미혼, ② 결혼/동거. ③ 별거, 이혼, 사별, ④ 기타
민감 변수	A8	가구 월 소득	① 100만 원 미만, ② 100-199만 원, ~ ⑩ 900만 원 이상, ⑪ 잘 모르겠다
	A8_NEW	가구 월 소득 재범주화	① 100만 원 미만, ② 100~199만 원, ③ 200~299만 원, ④ 300~399만 원, ⑤ 400~499만 원, ⑥ 500만 원 이상, ⑦ 잘 모르겠다
	A12	주된 진단명	① 조현병(정신분열증), ② 양극성정동장애(조울증), ~, ⑨ 기타, ⑩ 진단받지 않았음
	A12_NEW	주된 진단명 범주	① 진단받음, ② 진단받지 않았음
	A11	의료보장제도	① 의료급여, ② 건강보험, ③ 모름
	C7_1	직종	① 관리자, ② 사무 종사자, ③ 판매 종사자, ④ 기능원 관련 기능 종사자, ⑤ 단순노무 종사자, ⑥ 전문가 및 관련 종사자, ⑦ 서비스 종사자, ⑧ 농림, 어업 숙련 종사자, ⑨ 장치, 기계 조작 및 조립 종사자, ⑩ 군인, ⑪ 기타
	C7_1_NEW	직종 범주화	① 비생산직: (① 관리자, ② 사무 종사자. ③ 판매 종사자, ⑥ 전문가 및 관련 종사자, ⑦ 서비스 종사자) ② 생산직: (④ 기능원 관련 기능 종사자, ⑤ 단순노무 종사자, ⑧ 농림, 어업 숙련 종사자, ⑨ 장치, 기계 조작 및 조립 종사자) ③ 기타:(⑩ 군인, ⑪ 기타)

주: 변수명에서 _NEW인 변수는 비식별화 작업을 위해 생성한 변수임.

출처: 한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 코딩북(내부 자료).

2. 지역사회 거주 정신질환자 조사의 노출 위험 분석

노출 위험 시나리오는 활용 변수를 다르게 설정하여 3개의 분석을 실시하였다. 분석 1은 원본 데이터로 성별, 생년, 교육수준, 혼인상태, 직종, 월평균 소득, 주된 진단명 변수를 활용한 것이고, 분석 2는 분석 1의 변수에서 연령, 가구소득, 직종을 범주형 변수로 비식별화 처리를 하였다. 분석 3은 분석 2의 변수에서 교육수준, 혼인상태, 주된 진단명을 범주형 변수로 비식별화 처리를 하여 노출 위험도를 측정하였다.

분석 4는 노출 위험을 검토하는 변수 조합을 좁혀서 성별(A2), 연령 범주((A3_NEW), 교육수준 범주(A4_NEW), 가구 월 소득 범주(A8_NEW)로 활용 변수를 구성하였다. 분석 5는 분석 4의 변수에서 주된 진단명 범주(A12_NEW)를 포함하였고, 분석 6에서는 분석 4의 변수에서 직종 범주(C7_1_NEW)를 포함하였다.

지역사회 거주 정신질환자 조사의 노출 위험 측도는 k-익명성과 전체 위험도(global risk)로 측정하였다.

〈표 4-14〉 지역사회 거주 정신질환자 조사의 노출 위험 시나리오 활용 변수

분석	활용 변수
분석 1	성별(A2), **생년(A3_1)**, 교육수준(A4_1), 혼인상태(A5), 직종(C7_1), 가구 월 소득(A8), 주된 진단명(A12)
분석 2	성별(A2), **연령 범주((A3_NEW)**, 교육수준(A4_1), 혼인상태(A5), **직종 범주(C7_1_NEW)**, **가구 월 소득 범주(A8_NEW)**, 주된 진단명(A12)
분석 3	성별(A2), **연령 범주((A3_NEW)**, **교육수준 범주(A4_NEW)**, **혼인상태 범주(A5_NEW)**, **직종 범주(C7_1_NEW)**, **가구 월 소득 범주(A8_NEW)**, **주된 진단명 범주(A12_NEW)**
분석 4	성별(A2), **연령 범주((A3_NEW)**, **교육수준 범주(A4_NEW)**, **가구 월 소득 범주(A8_NEW)**
분석 5	성별(A2), **연령 범주((A3_NEW)**, **교육수준 범주(A4_NEW)**, **가구 월 소득 범주(A8_NEW)**, **주된 진단명 범주(A12_NEW)**
분석 6	성별(A2), **연령 범주((A3_NEW)**, **교육수준 범주(A4_NEW)**, **가구 월 소득 범주(A8_NEW)**, **직종 범주(C7_1_NEW)**

출처: 한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 코딩북(내부 자료).

분석 1은 데이터의 약 83%가 2-익명성을 만족하지 않으며, 약 99%가 5-익명성을 만족하지 않는다. 연령 및 가구소득, 직종을 범주화(재범주화) 변수로 처리한 분석 2는 데이터의 약 44%가 2-익명성을 만족하지 않으며, 약 77%가 5-익명성을 만족하지 않는다. 연령, 가구소득, 교육수준, 혼인상태, 직종, 주된 진단명을 범주화(재범주화) 변수로 처리한 분석 3은 데이터의 약 21%가 2-익명성을 만족하지 않으며, 약 49%가 5-익명성을 만족하지 않는다.

전체 위험도는, 데이터 전체의 식별 위험을 추정하는 방법으로, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 90%임에 반해, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 39% 정도로 줄어드는 것을 알 수 있다.

분석 4는 데이터의 약 8%가 2-익명성을 만족하지 않으며, 약 33%가 5-익명성을 만족하지 않는다. 분석 5는 데이터의 약 10%가 2-익명성을 만족하지 않으며, 약 36%가 5-익명성을 만족하지 않는다. 분석 6은 데이터의 약 9%가 2-익명성을 만족하지 않으며, 약 36%가 5-익명성을 만족하지 않는다.

〈표 4-15〉 지역사회 거주 정신질환자 조사 노출 위험도 추정 1

노출 위험	분석 1	분석 2	분석 3
2-anonymity (Number of observations violating)	543 (83.28%)	293 (44.94%)	140 (21.47%)
3-anonymity (Number of observations violating)	617 (94.63%)	413 (63.34%)	243 (37.27%)
5-anonymity (Number of observations violating)	647 (99.23%)	506 (77.60%)	322 (49.38%)
global risk (Expected number of re-identifications)	590 (90.54%)	400 (61.43%)	253 (38.88%)

출처: 한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 마이크로데이터(내부 자료).

전체 위험도를 보면 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 24%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 26% 정도이다. 분석 1과 분석 4를 비교하였을 때 노출 위험을 검토하는 변수 조합을 줄이고, 변수를 범주화 또는 재범주화하였을 경우, 노출 위험이 약 90%에서 약 24%로 감소함을 알 수 있다.

l-다양성 추정으로는 Distinct l-Diversity(서로 다른 민감한 값의 개수)를 사용하여 각 그룹에서 민감한 속성의 값들이 얼마나 다양한지를 계산한다. 지역사회 거주 정신질환자 조사 데이터에 포함된 범주형 식별 변수의 속성 조합은 2개로, 조합을 다르게 하였을 때의 l-다양성의 추정을 비교하고자 하였다. 첫 번째 조합은 성별(A2), 연령 범주(A3_NEW), 가구 월 소득(A8)이고, 두 번째 조합은 성별(A2), 연령 범주(A3_NEW), 교육수준(A4_1)이다. 분석에 활용한 민감 변수는 주된 진단명(A12), 의료보장제도(A11), 직종(C7_1)이며 세 민감 변수의 l-다양성을 추정한 분포는 〈표 4-17〉과 같다. 속성 조합 변수를 어떻게 구성하느냐에 따라 민감 변수의 l-다양성을 추정한 분포는 달라질 수 있다. 첫 번째 조합은 가구 월 소득 변수, 두 번째 조합은 교육수준 변수가 활용되었는데 여기에서는 두 조합의 l-다양성 값이 크게 차이가 나지 않음을 알 수 있다.

〈표 4-16〉 지역사회 거주 정신질환자 조사 노출 위험도 추정 2

노출 위험	분석 4	분석 5	분석 6
2-anonymity (Number of observations violating)	55 (8.43%)	67 (10.27%)	64 (9.81%)
3-anonymity (Number of observations violating)	119 (18.25%)	131 (20.09%)	134 (20.55%)
5-anonymity (Number of observations violating)	218 (33.43%)	235 (36.04%)	241 (36.96%)
global risk (Expected number of re-identifications)	158 (24.23%)	169 (25.92%)	170 (26.18%)

출처: 한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 마이크로데이터(내부 자료).

〈표 4-17〉 지역사회 거주 정신질환자 조사 노출 위험도 l-다양성 추정

	성별(A2), 연령 범주(A3_NEW), 가구 월 소득(A8) 변수 조합 사용			성별(A2), 연령 범주(A3_NEW), 교육수준(A4_1) 변수 조합 사용		
	주된 진단명 (A12)	의료보장제도(A11)	직종 (C7_1)	주된 진단명 (A12)	의료보장제도(A11)	직종 (C7_1)
최소	1	1	1	1	1	1
1분위	3	2	7	3	3	11
중앙값	4	3	13	5	3	18
평균	4	3	18	5	3	20
3분위	5	3	28	6	3	26
최대	7	3	45	7	3	46

출처: 한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 마이크로데이터(내부 자료).

제4절 소결

개인정보 노출 위험 시나리오는 어떠한 준식별 정보를 조합하여 활용하느냐, 어떠한 비식별화 처리 기법을 사용하느냐에 따라 수십, 수백 가지의 시나리오를 구성할 수 있다. 이 장에서는 한국복지패널, 가족과 출산 조사, 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사별로, 실무적으로 활용할 수 있는 개인식별 가능 항목을 범주화 비식별화 처리 방법으로 노출 위험 시나리오를 6가지로 구성해보았다. 한국복지패널과 가족과 출산 조사는 시나리오 구성 시, 지역 변수의 범주 범위를 확대하였을 경우 노출 위험이 얼마나 증가할 것인지를 검토하였다. 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사는 연구원의 원내 조사자료이므로 어느 정도의 수준까지 비식별화 처리 작업이 필요한지에 대한 판단에 도움을 주고자 시나리오를 구성하였다. 준식별 정보는 인구학적

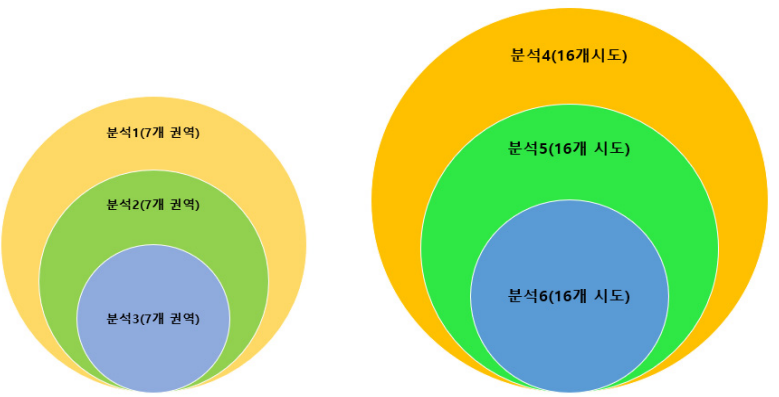
정보와 개인의 경제상태를 나타내는 정보로 구성하고 각 조사의 민감 정보를 활용하여 k-익명성과 전체 위험도, l-다양성 분석 결과를 제시하였다. 데이터의 비식별화 처리 기법은 실무 레벨에서 적용 가능한 수준인 재범주화, 범주화, 상단코딩, 삭제 방법을 사용하였다.

첫 번째 데이터는 18차 한국복지패널로, 분석 1~3은 지역 변수를 7개 권역으로, 분석 4~6은 지역 변수를 16개 시도로 구분하여 활용하였고, 분석 1에서 분석 3, 분석 4에서 6으로 갈수록 활용 변수를 삭제, 범주화하여 공개 범위를 축소시켰다.

k-익명성의 경우, 분석 1에서 전체 데이터의 약 75%, 분석 3에서는 약 8%, 분석 4에서는 약 83%, 분석 6에서는 약 13%가 3-익명성을 만족하지 않았다.

전체 위험성의 경우, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 0.36%이고, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 0.03%였다. 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.43%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 0.06% 정도였다.

〔그림 4-1〕 복지패널의 노출 위험 시나리오별 변수 공개 범위 비교



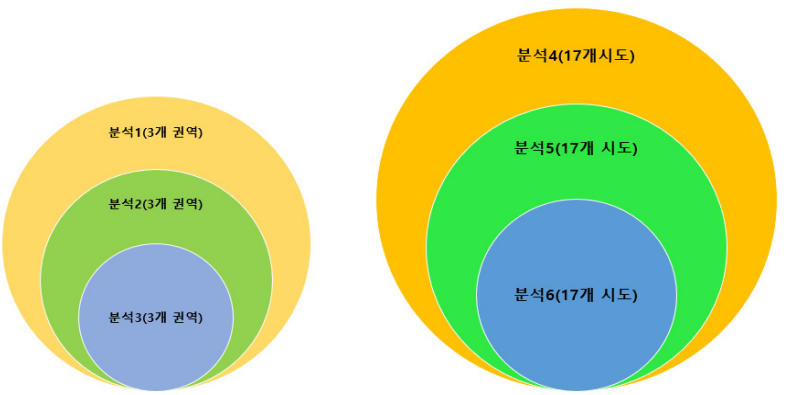
구분 \ 권역	7개 권역		
시나리오	분석 1	분석 2	분석 3
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(교육) (혼인)(취업)(소득)

구분 \ 권역	16개 시도		
시나리오	분석 4	분석 5	분석 6
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(교육) (혼인)(취업)(소득)

출처: 저자 작성

두 번째 데이터는 2021년 가족과 출산 조사로, 분석 1~3은 지역 변수를 3개 권역으로, 분석 4~6은 지역 변수를 17개 시도로 구분하여 활용하였고, 분석 1에서 분석 3, 분석 4에서 6으로 갈수록 활용 변수를 삭제, 범주화하여 공개 범위를 축소시켰다.

[그림 4-2] 가족과 출산 조사의 노출 위험 시나리오별 변수 공개 범위 비교



구분 \ 권역	3개 권역		
시나리오	분석 1	분석 2	분석 3
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(소득) (혼인)(경제활동상태)(교육)

구분 \ 권역	17개 시도		
시나리오	분석 4	분석 5	분석 6
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령)(소득)	6 (직종)(연령)(소득) (혼인)(경제활동상태)(교육)

출처: 저자 작성

2021년 가족과 출산 조사 k-익명성 분석 결과, 분석 1에서 전체 데이터의 약 40%, 분석 3에서는 약 9%, 분석 4에서는 약 76%, 분석 6에서는 약 41%가 3-익명성을 만족하지 않았다.

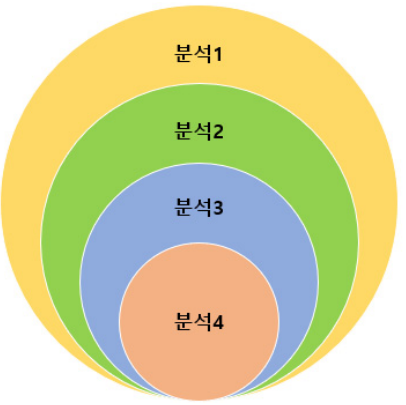
전체 위험성의 경우, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 46%이고, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수는 전체의 약 16%였다. 분석 4에서 식별이 될 것으로 예측되는 레코드 수가 전체의 약 84%이고, 분석 6에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 53% 정도였다.

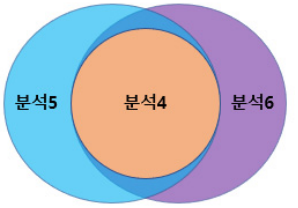
세 번째 데이터는 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사(2022)로, 분석 1에서 분석 4로 갈수록 활용 변수를 삭제, 범주화하여 공개 범위를 축소시켰다. 분석 5와 분석 6은 분석 4의 활용 변수에서 주된 진단명 범주, 직종 범주를 각각 추가하였을 때의 노출 위험 증가를 보고자 하였다.

지역사회 거주 정신질환자 조사의 k-익명성 분석 결과, 분석 1에서 전체 데이터의 약 94%, 분석 3에서는 약 37%, 분석 4에서는 약 18%, 분석 5에서는 약 20%, 분석 6에서는 약 21%가 3-익명성을 만족하지 않았다.

전체 위험성의 경우, 분석 1에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 91%이고, 분석 3에서는 식별이 될 것으로 예측되는 레코드 수가 전체의 약 39%였다. 분석 4에서 식별이 될 것으로 예측되는 레코드 수는 전체의 약 24%이고, 분석 5, 분석 6에서는 전체의 26% 수준이었다.

〔그림 4-3〕 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 노출 위험 시나리오별 변수 공개 범위 비교





구분 \ 권역	4개의 시나리오 비교			
시나리오	분석 1	분석 2	분석 3	분석 4
공개 범위 축소 정도 (변수 삭제 및 범주화)	1 (생년)	3 (직종)(연령) (소득)	6 (직종)(연령) (소득) (혼인)(주된 진단명)(교육)	6 (연령)(소득) (교육) – 직종, 혼인, 주된 진단명은 삭제 –

구분 \ 권역	3개의 시나리오 비교		
시나리오	분석 4	분석 5	분석 6
공개 범위 축소 정도 (변수 삭제 및 범주화)	6 (연령)(소득) (교육) – 직종, 혼인, 주된 진단명은 삭제 –	5 (연령)(소득) (교육)(주된 진단명) – 직종, 혼인은 삭제 –	5 (연령)(소득) (교육)(직종) – 혼인, 주된 진단명은 삭제 –

출처: 저자 작성

이 6가지 시나리오를 통해 복지패널, 가족과 출산 조사 데이터는 지역 변수 공개 범위를 시도 단위까지 확대할 경우, k-익명성과 전체 위험성으로 노출 위험도가 어느 수준까지 증가하는지를 파악할 수 있었다. 전체

위험성의 수준이 한국복지패널과 가족과 출산 조사가 다른 이유는 한국복지패널은 일반 가중치를 적용하였고, 가족과 출산 조사는 표본 가중치를 적용하였기 때문이다. 가족과 출산 조사는 일반 가중치를 공개하고 있지 않기 때문에 표본 가중치를 적용한 전체 위험성의 수준이 복지패널보다 높게 나타났다. 이는 어떤 가중치를 적용하느냐도 위험측도에 따라 다른 결과를 제시해줄 수 있음을 의미한다.

지역사회 거주 정신질환자 조사는 원내의 조사자료 중 하나로, 다양한 준식별 정보의 조합으로 노출 위험을 측정하였을 때, 원자료 수준에서는 노출 위험도가 매우 높음을 알 수 있고, 변수별로 범주화, 재범주화가 필요하다는 것을 알 수 있다.

노출 위험 분석으로 k-익명성의 단점을 보완한 l-다양성 분석 결과도 함께 제시하였다.

이 세 데이터 분석을 통해 알 수 있는 것은 데이터의 특성, 준식별 변수의 조합, 노출 위험 측도에 따라 상대적인 평가가 가능할 뿐, 절대적인 기준은 없다는 것이다. 이는 실제 비식별화 데이터 생성 작업 시, 고려해야 하는 요인들이 적지 않음을 의미하고, 준식별 변수 조합의 수준, 민감 정보의 정의에 대해 어느 정도의 범위까지 통일된 양식, 기준을 가지고 갈 수 있느냐에 대한 논의가 필요하다는 것을 알 수 있다. 통계청의 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성(2023)에서도 비식별화 처리 방법에 대한 기준보다도, 통계작성기관의 특성에 적합한 방안을 지속적으로 개선·개발할 필요성을 언급하고 있다. 이에 제5장에서는 한국보건사회연구원의 조사자료 비식별화 처리 현황을 살펴보고 비식별화 처리 가이드라인 개발, 비식별화 데이터의 이관 및 관리 절차 설정을 다루고자 한다.

제5장

데이터 비식별화 처리 가이드라인 개발 및 관리체계 수립

제 5 장 데이터 비식별화 처리 가이드라인 개발 및 관리체계 수립

제1절 원내 공개용 조사자료 비식별화 처리 관련 사전 검토

공공데이터의 수요가 늘어나고, 개인정보 보호에 대한 중요성이 부각됨에 따라 공공데이터의 비식별화 처리의 중요성이 점차 강조되고 있다. 한국보건사회연구원은 연구보고서 작성 중 생산된 조사자료(마이크로데이터)를 민간 활용 활성화를 위해 공개하고 있다. 하지만 현재 원내에는 체계적인 데이터 공개 처리 가이드라인이 없어 개발이 필요한 상황이다. 특히, 개인정보 보호법을 준수하고 데이터 활용성을 저해하지 않는 수준으로 데이터를 처리할 수 있는 가이드라인이 필요하다. 따라서 이 절에서는 가이드라인을 작성하기에 앞서 공개용 조사자료 처리 현황을 검토하고자 한다. 이에 현재 데이터조사관리팀과 연구자가 협의하는 공개용 조사자료 처리 수준을 확인하고, 더불어 체계적인 원내 비식별화 처리를 위해 원내 조사자료를 이용하여 개인식별 가능 정보의 현황을 파악하고자 한다.

1. 기존 공개용 조사자료 처리 현황

기존에 공개용 조사자료를 처리하는 일은 마이크로데이터, 조사표, 코드북 간의 매칭 부분과 마이크로데이터 비밀보호 처리 부분으로 나누어져 있다. 해당 처리 내용과 관련된 부분을 가이드라인에서 발췌하였고,

그에 따른 실무자들의 처리 내역은 아래 표와 같다.

〈표 5-1〉 기존 공개용 조사자료 사전 검토 가이드라인

① 마이크로데이터, 조사표, 코드북 간 매칭

- 조사표 문항에 있으나 코드북 또는 마이크로데이터에 없는 항목 리스트를 작성하여 추가 가능 여부를 연구책임자에게 문의
- 조사표의 보기 문항과 코드북의 일치성 확인
- 마이크로데이터에 연구진에서 가공하여 만든 변수는 마지막 열로 보내고, 코드북에 관련 정보를 작성한 후 연구책임자에게 컨펌 요청
- 가중치 변수는 조사표 없으므로 존재 여부를 별도 확인하고, SPSS 파일의 경우 마이크로데이터에 가중치 적용이 디폴트로 지정되어 있는지 체크

② 마이크로데이터 비밀보호 처리

- 마이크로데이터에 개인을 직접 식별할 수 있는 변수 제거 검토
- 마이크로데이터 수준에서 개인을 직접 식별할 수 있는 변수를 제거한 후에는 변수 간 조합(예를 들어 제주도 57세 여성 군인)을 통해 개인을 식별할 수 있는 경우를 제거해야 하는데, 이는 아래 매크로데이터 단계에서 해결하는 것이 현실적임
- 매크로데이터 기준 개인정보 노출 위험 파악: 개별 변수 빈도표 작성

출처: 한국보건사회연구원 정보통계연구실. (2022.4.26). 제공용 조사자료 사전검토 가이드라인(내부자료)에서 일부 발췌

〈표 5-2〉 가이드라인에 따른 실제 처리 내역

검수 작업	처리 내용
마이크로데이터, 조사표, 코드북 간 매칭	- 조사표에 따른 마이크로데이터 및 코드북 불일치 확인 · 문항(변수) · 보기 항목(변수값) - 지시 로직 확인 - 연속형 변수 오류값 확인 - 가중치 변수 누락 확인
마이크로데이터 비밀보호 처리	- 조사구 및 응답자 정보 제공 여부 확인 - 개인식별 변수 삭제 - 개인식별 가능 변수 처리 - 낮은 빈도 확인 · 교차분석 진행

출처: 저자 작성

가이드라인에 따른 실무자들의 처리 내역을 정리하여 살펴본 결과, 조사표와 데이터 매칭 부분에서는 실무자들 간의 내검 범위가 상이하였다. 그리고 마이크로데이터 비밀보호 처리 부분에서는 고유 식별 정보 및 준식별 정보 같은 개인식별 가능 변수에 대한 명확한 기준 없이 실무자들의 주관적인 판단에 의하여 데이터를 처리하고 있었다. 조사 목적에 따라 개인식별 가능 변수와 민감 변수의 구분은 다를 수 있고, 그에 따라 비식별화 처리 방식 또한 다르지만, 원내에서 사용할 최소한의 개인식별 정보를 판단하는 기준과 처리에 대한 기준을 설정하는 것이 필요할 것으로 판단된다.

2. 조사자료에서 개인식별 가능 항목 검토

가. 조사별 개인식별 가능 항목 현황

원내 조사별 개인식별 가능 항목 현황을 살펴보기 위하여 2021~2022년 조사자료 20건을 검토하였다. 원내 조사에서 개인식별 가능 정보로 사용되는 항목은 성별, 연령, 지역, 최종학력, 혼인상태, 장애, 종교, 가구소득, 경제활동상태, 직종, 거주 형태, 가구 구성, 만성질환, 자산 및 부채, 사회복지제도까지 15개로 요약된다. 조사별로 해당 변수의 유무를 파악하였고, 대부분의 조사에서 성별, 연령, 지역, 최종학력이 사용되었음을 확인하였다. 가구소득, 혼인상태, 경제활동상태, 직종, 사회복지제도, 거주 형태, 만성질환, 종교, 자산 및 부채, 장애 순으로 나타났다. 원내 20건의 조사자료를 이용했기 때문에 대표성의 문제는 있을 수 있지만, 원내 조사의 개인식별 가능 항목 현황을 파악하기에는 무리가 없다고 판단된다.

〈표 5-3〉 조사별 개인식별 가능 항목

조사명	성별	연령	지역	최종 학력	혼인 상태	장애	종교	가구 소득	경제 활동 상태	직종	거주 형태	가구 구성	만성 질환	자산 및 부채	사회 복지 제도
디지털헬스 접근성 및 개인역량요인에 대한 설문조사	1	1	1	1				1		1		1			
코로나19 유행 기간 중 미충족 의료 및 의료이용에 대한 조사(일반인)	1	1	1	1	1			1	1				1		1
헬스커뮤니케이션에 대한 현황과 인식 파악을 위한 설문조사	1	1	1	1	1								1		
보건복지정책과 기술 간 융합체계 구축 필요성에 대한 인식조사(일반국민)	1	1	1	1				1		1					
한국인의 사회적 문제 경험과 인식 조사	1	1	1	1	1		1	1	1	1	1	1		1	1
복지분야 사각지대 및 부적정 지출축소 방안마련 인식조사	1	1	1	1				1				1			1
2020 나눔 실태 및 인식현황조사	1	1	1	1	1		1	1	1	1					
한국인의 행복과 삶의 질 실태조사	1	1	1	1	1	1		1	1	1	1	1	1		
한국사회 분배인식 조사	1	1	1	1	1			1	1			1			
사회보장수요 및 지출부담 수준에 관한 인식조사	1	1	1	1	1			1	1			1			1
사회경제적 위기와 사회통합 실태조사	1	1	1	1	1			1	1	1	1	1			
사회 참여, 자본, 인식 조사	1	1	1	1	1			1				1		1	
코로나19 유행 기간 중 미충족 의료 및 의료이용에 대한 조사(만성질환자)	1	1	1	1	1			1	1				1		1
생활 사건 및 트라우마 경험 조사	1	1	1	1	1										
보건복지정책과 기술 간 융합체계 구축 필요성에 대한 인식조사(전문가)	1	1													
외국인유학생 복지실태조사	1	1	1	1							1				
한국 사회 수용성에 대한 이주민 인식 조사	1	1	1	1	1		1	1	1	1					
코로나 시기 초등학생 돌봄 실태 및 정책 요구도 조사	1	1	1	1	1			1	1	1		1			1
노인돌봄인력 처우개선을 위한 실태조사	1	1	1						1	1					
긴급 지원 실태 및 인식 조사	1	1	1	1	1			1	1	1	1	1			
	20	20	19	18	14	1	3	15	12	10	5	10	4	2	6

출처: 저자 작성

나. 조사별 항목 세부 내용 검토

동일한 항목일지라도 조사 목적에 따라 항목별 세부 조사 내용은 조사별로 상이한 것으로 파악되었다. 혼인상태 항목 및 직업 항목을 살펴보면, 혼인상태는 배우자의 여부만 조사하는 경우도 있지만 미혼, 유배우, 별거, 사별, 이혼과 같이 더 상세한 상태를 담는 경우도 있었다. 직업항목은 한국표준직업분류 대분류에 해당하는 10개 항목으로 조사하거나, 조사 목적에 따라 특정 직업군에 대하여 상세하게 조사하였다. 이는 연구자 또는 조사 목적에 따라 개인식별 가능 정보의 범주를 다르게 설정할 수 있고, 조작적 정의 또한 다를 수 있어 표준화가 어렵다는 점을 시사한다.

〈표 5-4〉 조사별 혼인상태 항목 세부 내용

조사명	혼인상태
코로나19 유행 기간 중 미충족 의료 및 의료이용에 대한 조사(일반인 대상)	① 배우자가 있으며, 함께 살고 있다(사실혼 상태 포함), ② 배우자가 있으나, 함께 살고 있지 않다(출장 등의 일시적 상태 제외), ③ 배우자가 없다
코로나19 유행 기간 중 미충족 의료 및 의료이용에 대한 조사(만성질환자 대상)	
헬스커뮤니케이션에 대한 현황과 인식 파악을 위한 설문조사	① 미혼, ② 기혼(유배우), ③ 이혼/별거. ④ 사별
한국인의 사회적 문제 경험과 인식 조사	① 배우자 있음(사실혼 포함), ② 별거, ③ 사별, ④ 이혼, ⑤ 미혼
2020 나눔 실태 및 인식현황조사	① 미혼, ② 기혼. ③ 이혼/사별
한국인의 행복과 삶의 질 실태조사	① 유배우, ② 별거. ③ 사별. ④ 이혼. ⑤ 미혼(미혼 부/모 포함)
한국사회 분배인식 조사	① 미혼, ② 기혼(이혼, 사별 포함)
사회경제적 위기와 사회통합 실태조사	① 유배우(사실혼 포함), ② 별거, ③ 사별, ④ 이혼, ⑤ 미혼(미혼 부·모 포함)
사회 참여, 자본, 인식 조사	① 유배우(사실혼 포함), ② 무배우(미혼, 비혼, 사별, 별거, 이혼)
생활 사건 및 트라우마 경험 조사	① 미혼, ② 유배우, ③ 이혼, 별거, 사별

조사명	혼인상태
한국 사회 수용성에 대한 이주민 인식 조사	① 배우자 없음, ② 배우자 있음
코로나 시기 초등학생 돌봄 실태 및 정책 요구도 조사	① 미혼, ② 기혼(사실혼 포함), ③ 기타(별거, 이혼, 사별)
긴급 지원 실태 및 인식 조사	① 미혼, ② 배우자 있음, ③ 이혼/사별

출처: 저자 작성, 항목 사례가 없는 조사는 제외하였음.

〈표 5-5〉 조사별 직업 항목 세부 내용

조사명	직업
디지털헬스 접근성 및 개인역량요인에 대한 설문조사	① 농업/임업/축산/어업, ② 자영업, ③ 판매/영업/서비스직. ④ 생산/기능/단순노무직, ⑤ 사무/관리/전문직, ⑥ 학생/재수생, ⑦ 군인(직업군인), ⑧ 전업주부, ⑨ 무직/퇴직/기타, ⑩ 모름/무응답
보건복지정책과 기술 간 융합체계 구축 필요성에 대한 인식조사(일반국민)	① 관리자, ② 전문가 및 관련 종사자, ③ 사무 종사자, ④ 서비스 종사자, ⑤ 판매 종사자, ⑥ 농림어업 숙련 종사자, ⑦ 기능원 및 관련 기능 종사자, ⑧ 장치·기계 조작 및 조립 종사자, ⑨ 단순노무 종사자, ⑩ 무직, 학생, 주부
2020 나눔 실태 및 인식현황조사	① 농업/수산업/축산업, ② 자영업, ③ 판매/서비스직, ④ 기능/숙련공, ⑤ 일반 작업직, ⑥ 사무/기술직, ⑦ 경영/관리직(사무관/부장 이상), ⑧ 전문/자유직(변호사/의사/건축사/교수/예술가), ⑨ 가정주부, ⑩ 학생, ⑪ 무직, ⑫ 기타
한국인의 행복과 삶의 질 실태조사	⓪ 비해당, ① 관리자, ② 전문가 및 관련 종사자, ③ 사무 종사자, ④ 서비스 종사자, ⑤ 판매 종사자. ⑥ 농림어업 숙련 종사자, ⑦ 기능원 및 관련 기능 조사자. ⑧ 장치/기계 조작 및 조립 종사자, ⑨ 단순노무 종사자, ⑩ 군인
사회경제적 위기와 사회통합 실태조사	① 관리자, ② 전문가 및 관련 종사자, ③ 사무 종사자, ④ 서비스 종사자, ⑤ 판매 종사자, ⑥ 농림어업 숙련 종사자. ⑦ 기능원 및 관련 기능 종사자, ⑧ 장치·기계 조작 및 조립 종사자. ⑨ 단순노무 종사자, ⑩ 군인
사회 참여, 자본, 인식 조사	특수고용 혹은 종속적 자영자 해당 여부 ① 예, ② 아니오

조사명	직업
한국 사회 수용성에 대한 이주민 인식 조사	① 공장노동자, ② 건설노동자, ③ 어업 종사자, ④ 음식점 종업원, ⑤ 간병인, ⑥ 가사 관련 단순 노무자(가정부, 파출부, 보육사 등), ⑦ 기타 단순 노무자(모텔 청소 등 육체노동 종사자), ⑧ 기타 서비스 종사자(관광가이드 등), ⑨ 판매 종사자(가게 운영, 세일즈맨, 보험설계사 등), ⑩ 준전문직 종사자(학원 강사, 유치원/학교 교사 등), ⑪ 사무 종사자(일반 행정사무 등, ⑫ 전문가 및 관련 종사자 (대학교수, 변호사, 의사, 약사, 간호사, 엔지니어, 통번역사, 컴퓨터 프로그래머 등), ⑬ 임직원 및 관리자(고급공무원, 교장, 기업체 임원 등), ⑭ 농축산업 종사자, ⑮ 기타
노인돌봄인력 처우개선을 위한 실태조사	① 요양보호사, ② 사회복지사, ③ 간호(조무)사, ④ 물리(작업)치료사
긴급 지원 실태 및 인식 조사	① 농림어업, 광업, ② 제조업, ③ 전기, 환경, 건설업, ④ 도소매, 운수, 숙박음식점업, ⑤ 출판 영상, 금융보험, 부동산업, ⑥ 과학기술, 사업지원, 임대업, ⑦ 서비스업, ⑧ 그 외

출처: 저자 작성, 항목 사례가 없는 조사는 제외하였음.

3. 조사자료 관련 세부 지침의 필요성

현재까지 원내 「조사자료 관리지침」에 따라 연구자는 조사표, 코드북, 조사데이터, 변수 구성표 등 양적 조사에 관한 조사자료를 연구사업 종료 이후 이관하고 있다. 하지만 해당 지침에는 이관할 조사자료에 대한 형식이나 제공범위 등 조사자료에 대한 구체적인 사항은 담고 있지 않아, 최소한 코드북, 자료 내검 범위, 이관 데이터 형태 세 가지 항목에 대해서는 구체적인 지침이 필요하다.

먼저, 이관된 코드북의 사례를 살펴보고자 한다. 첫 번째 사례는 문항 설명이나 보기 항목에 대한 설명 없이 변수만 나열되어 있는 경우이다. 해당 사례는 〈표 5-6〉에서 확인할 수 있다.

〈표 5-6〉 코드북 구성 – 사례 1

• 이관 받은 코드북 구성

온라인 조사용	DP용(데이터 헤더 교체 필수)	
Variable Labels	Variable Labels	
IDX	IDX	설문 접속 순서
VC_ID	VC_ID	응답자 아이디
A1o1	A1_1	'A1. 귀하가 거주하시는 지역은 어디입니까? - 시/도'

↓

• 실무자가 재작성한 코드북 구성

변수명	변수 설명	보기 문항 내용
A1_1	A1. 귀하가 거주하시는 지역은 어디입니까? - 시/도	1. 서울특별시 2. 부산광역시 … 16. 제주도

출처: 저자 작성

이러한 코드북 구성 형태는 데이터를 작성한 당사자만 알 수 있는 내용으로 구성되어 있기 때문에 이용자에게는 편의성이 낮은 수준으로 판단된다. 따라서 실무자는 연구자의 동의하에 코드북을 재작성하고 있고, 해당 코드북은 변수명, 변수 설명, 보기 문항으로 구성된다.

두 번째 사례는 변수값과 변수 구성을 각각 상이한 sheet에 구성한 사례이다. 이 경우는 하나의 변수 정보를 확인하려면 각각의 sheet를 살펴보아야 하므로 사례 1과 마찬가지로 이용자 편의성이 낮은 수준으로 판단된다. 따라서 실무자는 이용자의 편의성을 고려하여 하나의 통합된 자료로 코드북을 재작성하여 제공하고 있다.

〈표 5-7〉 코드북 구성 – 사례 2

• 이관 받은 코드북 구성

변수값 sheet		변수 구성 sheet	
변수명	보기 문항 내용	변수명	변수 설명
A5_2	1 '재학' 2 '중퇴' 3 '졸업(수료 포함)' 4 '휴학'	A5_2	'5_2. 귀하의 졸업상태는 어디에 해당하십니까?

↓

• 실무자가 재작성한 코드북 구성

통합 sheet		
변수명	변수 설명	보기 문항 내용
A5_2	5_2. 귀하의 졸업상태는 어디에 해당하십니까?	1 '재학' 2 '중퇴' 3 '졸업(수료 포함)' 4 '휴학'

출처: 저자 작성

이처럼 현재 이관된 코드북은 이용자에게 제공하기에 부적합하거나 이용 시 효율성이 떨어지므로 연구원 차원의 표준화된 공개용 코드북 양식을 제작하는 것도 고려해 볼 필요가 있다.

다음으로 구체적인 지침이 필요한 부분은 '자료 내검'의 범위이다. 이관된 데이터는 최종보고서 작성을 위해 활용된 자료임을 전제로 '자료 내검'이 완료되었다고 판단할 수 있다. 하지만 공개 자료 처리를 위해 데이터를 검토하여 보면 '자료 내검'이 적절하지 않은 경우도 있었다. 공개용 조사자료 사전 검토 단계에서는 보고서 발간 이후에 조사자료를 이관받기 때문에 조사 내용상의 내검 규칙에 관한 검토는 어려운 상황이므로, 조사설계상의 논리성(문항 지시로직, 결측값, 무응답, 범주값 외의 응답값의 존재 등)에 대해서만 사전 검토 항목으로 다룰 수 있다. 통계자료의 품질과 관련된 완전성과 일관성 확보와 더불어 데이터 이용자의 편의성 제고, 개인정보보호법의 준수를 위하여 추가적인 검토는 필요할 것으로

판단된다.

마지막으로 이관해야 하는 조사데이터가 원시 데이터 또는 마이크로데이터인지 규정화되어 있지 않아, 이관된 데이터 형태는 제각각인 상황이다. 해당 데이터에는 조사구 정보 및 개인정보가 포함되어 있는 경우, 분석을 위한 가공 변수를 포함하지 않은 경우도 있어 별도의 처리가 필요한 상황이다. 이러한 점은 공개용 조사자료 처리 시 실무자의 업무부담을 가중시키는 요인으로 작용한다. 따라서 이관할 데이터를 명확하게 정의할 필요가 있다.

4. 가이드라인 관련 개선사항 도출

앞에서 기존의 공개용 조사자료 사전 검토 가이드라인에 따른 공개용 조사자료 처리 수준을 확인하였고, 조사자료의 개인식별 가능 정보 현황을 살펴보았다. 위의 내용에서 현재 공개용 조사자료 처리에 대한 한계점을 토대로 도출할 수 있는 개선사항을 제시하고자 한다.

우선, 데이터 비식별화 처리를 위해서는 원내의 기준을 마련하는 것이 필요하다. 실무자들이 명확한 기준 없이 주관적으로 판단해야 한다는 점은 기존의 가이드라인에서 비밀보호 처리 부분의 한계점이라고 생각된다. 따라서 조사별 개인식별 가능 항목 현황을 토대로 원내 조사자료의 개인식별 가능 정보 추출을 위한 최소한의 기준을 만들어야 한다. 개인식별 가능 정보의 추출뿐만 아니라 그에 대한 처리 기준도 필요하다. 해당 변수에 대한 구체적인 처리 방식을 설정하여 일반화된 처리 원칙을 만들 필요가 있다. 더 나아가 앞으로 진행해야 할 원내 공개용 조사자료 처리를 통해 개인식별 정보의 기준과 처리 방식에 대한 사후적인 검토를 이어가야 할 것이다.

또한, 공개용 조사자료를 사전 검토할 때 실무자들 간의 일관성 있는 처리 방식이 필요하다. 기존의 가이드라인 기준으로 동일한 조사자료를 각각의 실무자들이 처리하는 경우, 결과가 다르게 도출되는 상황이다. 따라서 공개용 조사자료 처리 체크리스트를 새롭게 작성하여 일관성 있게 조사자료를 검토하기 위한 도구로 활용할 필요가 있다. 이러한 부분을 고려하여 원내 비식별화 가이드라인을 개발하였고, 제2절에서 자세한 내용을 다루고자 한다.

제2절 원내 비식별화 처리 관련 가이드라인 개발

이 절에서는 공개용 조사자료 비식별화 처리와 관련하여 사전 검토를 통하여 보완이 필요한 사항을 전반적으로 살펴보고, 원내 비식별화 처리 가이드라인을 제시하고자 한다. 우선, 공개용 조사자료 비식별화 처리 가이드라인을 작성하면서 변경된 검토의견서와 추가된 체크리스트를 먼저 제시하였다.

1. 검토의견서 및 체크리스트

공개용 조사자료 비식별화 처리 관련 가이드라인을 검토하면서 공개용 조사자료 사전 검토의견서 양식을 변경하였다. 기존 대비 변경안에서 가장 크게 수정한 사항은 마이크로데이터 비밀보호 처리를 위한 작성 방식을 변경한 점이다. 기존의 검토의견서에서는 실무자가 서술식으로 기재하여 해당 내용을 작성했던 반면, 변경안에서는 개인식별 정보 관련 표를 삽입함으로써 좀 더 구체적이고 명확하게 해당 데이터에 대한 개인식별

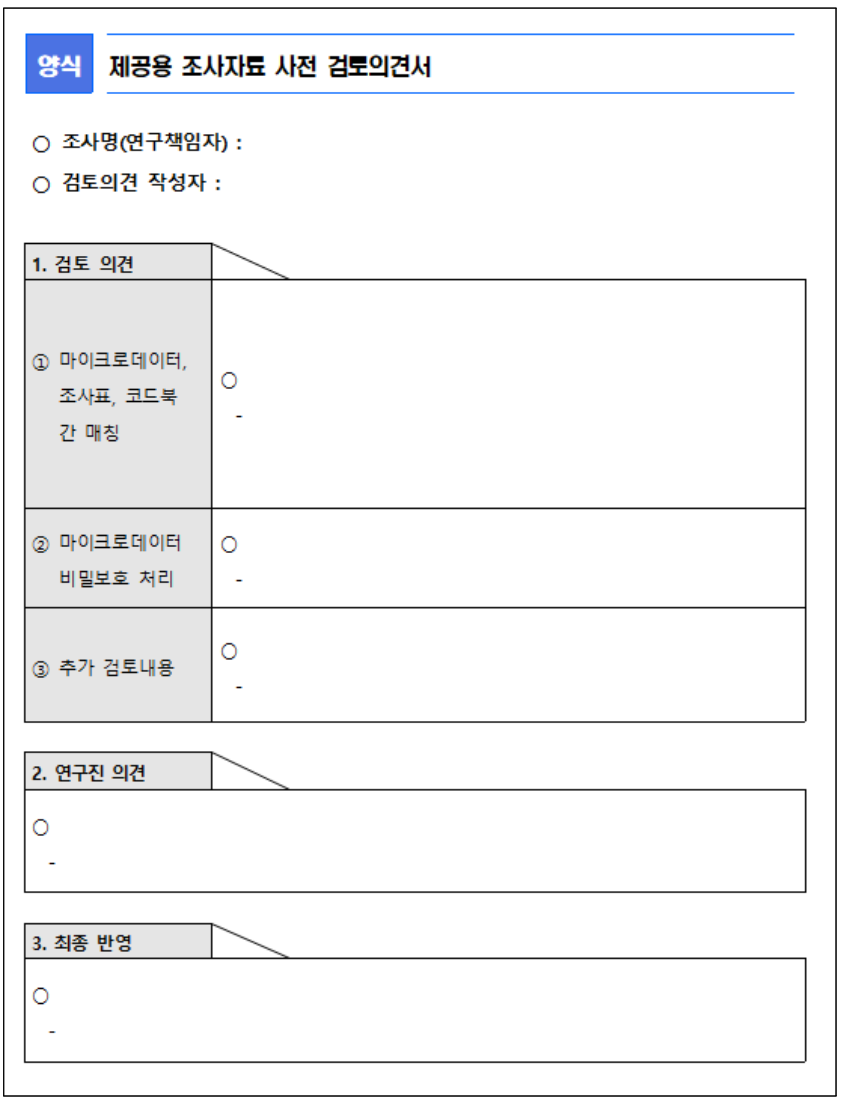
정보 및 민감 정보를 검토할 수 있도록 구성하였다. 개인식별 가능 변수와 비식별화 처리 전 항목 요약, 처리 방식, 비식별화 처리 후 내역 세부를 포함하였고, 이를 통해 실무자는 물론 연구자가 개인식별 정보를 면밀하게 살펴볼 수 있게 하였다. 또한, 일관된 데이터 처리를 위하여 체크리스트를 새롭게 추가하였다. 체크리스트를 통해 실무자가 기본적으로 살펴보아야 할 내용을 담아 해당 업무를 처리할 때 놓치지 않도록 구성하였다.

〈표 5-8〉 가이드라인 비교표

구분		기존	변경(안)
전체	용어	✓제공용 조사자료 사전 검토의견서	✓공개용 조사자료 사전 검토의견서
	검토의견서	✓서술식으로 작성	✓서술식과 비식별화 처리표 작성 병행
	체크리스트	-	✓조사표와 데이터의 매칭 및 비식별화 처리 체크리스트 생성
조사표와의 매칭	용어	✓마이크로데이터, 조사표, 코드북 간 매칭	✓마이크로데이터, 조사표, 코드북 간 매칭
	처리 방식	✓조사표와 데이터의 일치 여부 확인	✓조사표와 데이터의 일치 여부 확인
		✓기존 자료 내검 범위 불확실 - 지시로직, 연속형 변수 오류값 확인	✓자료 내검 범위 확정 - 지시로직, 연속형 변수 오류값 확인 - 순위형, 중복응답, 합계 문항 로직 확인, 텍스트 변수 처리
		✓이관된 조사자료를 파일 형식으로 공개	✓csv 파일 형식으로 제공
비식별화 처리	용어	✓마이크로데이터 비밀보호 처리	✓마이크로데이터 비식별화 처리
	처리 방식	✓개인식별 정보 및 파라데이터 삭제	✓개인식별 정보 및 파라데이터 삭제
		-	✓개인식별 가능 정보 탐색 및 분류
		✓빈도를 통한 비식별화 처리	✓비식별화 절차에 따른 처리
		✓연구진에게 비식별화 처리 방안 요청	✓비식별화 처리 적용 기준을 설정하여 연구진에게 제안

출처: 저자 작성

〔그림 5-1〕 제공용 조사자료 사전 검토의견서(기존)



양식 제공용 조사자료 사전 검토의견서

○ 조사명(연구책임자) :

○ 검토의견 작성자 :

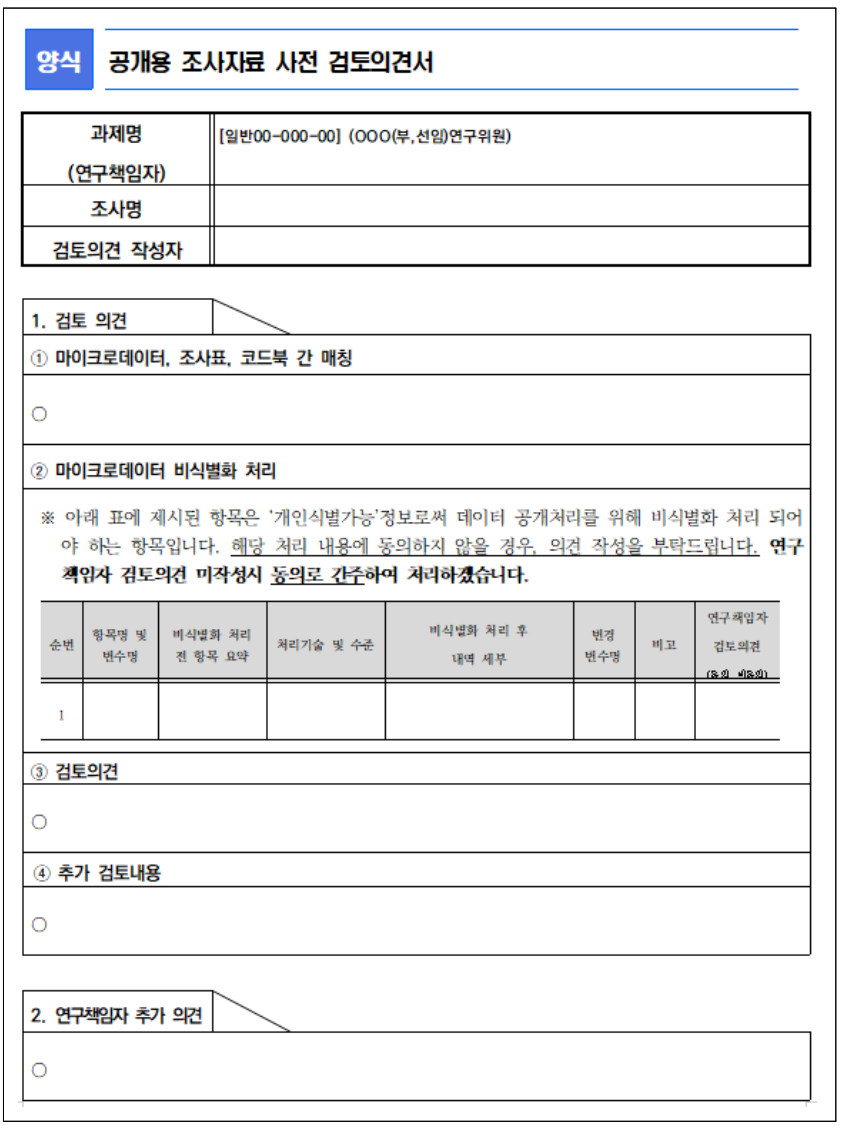
1. 검토 의견	
① 마이크로데이터, 조사표, 코드북 간 매칭	○ -
② 마이크로데이터 비밀보호 처리	○ -
③ 추가 검토내용	○ -

2. 연구진 의견
○ -

3. 최종 반영
○ -

출처: 저자 작성

[그림 5-2] 공개용 조사자료 사전 검토의견서 변경(안)



양식 **공개용 조사자료 사전 검토의견서**

과제명 (연구책임자)	[일반00-000-00] (OOO(부,선임)연구위원)
조사명	
검토의견 작성자	

1. 검토 의견

① 마이크로데이터, 조사표, 코드북 간 매칭

○

② 마이크로데이터 비식별화 처리

※ 아래 표에 제시된 항목은 '개인식별가능'정보로써 데이터 공개처리를 위해 비식별화 처리 되어야 하는 항목입니다. 해당 처리 내용에 동의하지 않을 경우, 의견 작성을 부탁드립니다. **연구책임자 검토의견 미작성시 동의로 간주하여 처리하겠습니다.**

순번	항목명 및 변수명	비식별화 처리 전 항목 요약	처리기술 및 수준	비식별화 처리 후 내역 세부	변경 변수명	비고	연구책임자 검토의견 (동의 비동의)
1							

③ 검토의견

○

④ 추가 검토내용

○

2. 연구책임자 추가 의견

○

출처: 저자 작성

〔그림 5-3〕 공개용 조사자료 처리 체크리스트 개발

공개용 조사자료 처리 체크리스트

	확인사항	여부
	• **이관된 자료가 데이터, 코드북, 조사표로 구성되어있는가?**	
마이크로데이터·조사표·코드북 간 매칭	• **제출된 조사표, 코드북에 오타는 없는가?**	
	• **조사표의 응답 대상자와 데이터의 응답 대상자가 동일하게 구성되어있는가?** ✓ 만 19세 이상 응답자인데 응답자의 연령이 18세인 경우 등	
	• **조사표와 동일하게 데이터, 코드북이 구성되어있는가?** ✓ 조사표의 순서와 데이터, 코드북의 순서가 일치하는지 확인 ✓ 누락되거나, 추가된 조사 항목이 있는지 확인 ✓ 불일치하는 변수(ex.분석을 위해 작성된 가공변수)가 데이터 및 코드북에 있는지 확인 ✓ 코드북의 변수명과 데이터의 변수명이 일치하는지 확인	
	• **조사표 문항의 보기와 데이터, 코드북이 동일하게 구성되어있는가?** ✓ 보기값 항목 및 내용 불일치 확인 ✓ 척도 의미 불일치 확인	
	• **문항 별 지시로직에 따른 빈도가 맞게 구성되어있는가?** ✓ 응답 빈도의 오류 확인	
	• **연속형 변수 응답 문항에 오류값은 없는가?** ✓ 연속형 변수 응답값 중 확실한 범위를 알 수 있는 조사대상자의 생년, 월 등이 범위값을 벗어난 값이 있는지 확인	
	• **순위형 응답 문항에 응답값이 동일한 오류는 없는가?** ✓ 1,2순위 응답값이 동일한지 확인	
	• **중복(복수)응답 문항에 응답값이 동일한 오류는 없는가?**	
	• **합계(비율) 응답 문항에 비율합이 100%가 맞는가?**	
	• **가중치 변수가 있는가?** ✓ 데이터에 가중치 변수가 없는 경우, 가중치 누락 여부를 확인	
마이크로데이터 비밀보호 처리	• **개인식별 변수가 데이터에 있는가?** ✓ 개인식별 변수(이름, 연락처, 주민등록번호, 운전면허 번호 등) 삭제	
	• **표본정보, 파라데이터가 코드북 및 데이터에 있는가?** ✓ 조사구(집계구)정보, 조사자료 수집과정에서 부수적으로 얻게되는 데이터가 코드북 및 데이터에 있다면 삭제	
	• **조사표 문항(코드북, 데이터)에 개인식별 가능 변수가 있는가?** ✓ 개인식별 가능 변수[연령(생년, 생월), 지역(지역상세정보, 17개 시도 등), 최종학력, 혼인상태, 소득, 직업(직종) 등]가 있는지 확인	
	• **변수 조합을 통해 개인을 식별할 수 있는 경우가 있는가?(빈도 5이하 확인)** ✓ 연령 및 거주지 변수와 특정 변수의 조합을 통해 개인 식별 가능성이 있는 경우가 있는지 확인	
	• **비식별화 처리 가이드라인에 따라 작업을 수행하였는가?**	

출처: 저자 작성

2. 비식별화 처리 가이드라인

데이터 비식별화 처리의 가장 이상적인 결과는 개인정보 노출 위험성은 낮추고 이용자의 편의성을 확보하는 것이다. 따라서 데이터를 외부로 공개할 때 적절한 비식별화 처리를 하여 개인식별 가능성을 차단시키면서 데이터의 활용 가능성은 높여야 한다. 하지만 데이터를 공개하면 변수 간의 연결 가능성과 그에 따른 추론 가능성 측면에서 취약성이 생긴다. 따라서 이를 최대한 배제하기 위해서 여러 가지 개인정보 보호 모델이 사용되고 있다. 이를 통해 조사 데이터의 품질 제고와 이용자의 편의성까지 고려한다면 최적의 데이터라 할 수 있다. 하지만 현실적으로 원내의 조사자료는 표본 수가 적고 변수 범주의 세분화 정도에 따라 개인정보 보호 모델을 적용하기가 어려운 상황이다. 지역과 직종(직업)처럼 범주 항목이 많을수록, 조합하는 변수가 많을수록 단일 레코드 발생 가능성이 높아지고, 이에 따라 개인식별 가능성 또한 높아진다. 즉, 범주의 항목 수와 조합하는 변수 수에 비례하여 개인이 식별될 가능성은 높아진다. 따라서 원내 조사자료에서는 최신의 개인정보 보호 모델을 적용하기 어렵고, 만약 개인정보 보호 모델을 적용하더라도 실무에서 적용 가능한 k-익명성 수준의 노출 위험도를 검토해야 한다.

이와 같은 한계점이 있으므로, 현시점의 비식별화 처리 가이드라인에서는 데이터의 활용도를 크게 저해하지 않는 수준에서 범주화 기준을 포괄적으로 설정할 필요가 있다고 판단된다. 다만, 가이드라인을 통해 비식별화 처리를 하더라고 연구 목적에 따라 반드시 공개되어야 하는 보기 항목이 있을 수 있으므로 단일 레코드에 대한 문제는 여전히 남아 있다. 이에 대한 기술적인 문제는 추가적으로 보완하여 비식별화 처리에 대한 수준을 점차 확대해 나갈 계획이다.

가. 개요

원내의 조사자료를 공개할 때 특정 개인을 식별할 수 없도록 비식별화 처리를 한 후 이를 공개하고자 한다.

본 가이드라인은 원내 조사자료의 비식별화 처리 방안을 제시하여 개인정보 보호뿐만 아니라 비식별화 처리에 대한 이해도를 높이고 활용을 돕기 위해 작성하였다. 가이드라인에서 사용하는 용어는 다음과 같이 정의하였다.

〈표 5-9〉 원내 비식별화 처리 가이드라인 용어 정리

용어	정리
개인식별 정보	- 특정 개인을 직접 식별할 수 있는 정보
개인식별 가능 정보	- 연령, 성별, 거주지 등 특정 개인을 직접적으로 식별할 수 없지만, 다른 정보와 조합하면 특정 개인을 식별할 수 있는 정보
범주화	- 기존 범주형 변수를 재범주화하거나, 연속형 변수에 대한 범위를 범주화하는 처리 방식

출처: 저자 작성

나. 비식별화 처리 원칙

기본적으로 개인식별 정보, 파라데이터, 생월, 생일은 완전 삭제하고, 개인식별 가능 정보는 비식별화 처리를 하여 공개하는 것을 원칙으로 한다. 단, 다음 표의 '조건부 삭제' 항목은 원내의 조사 특성을 고려하여 삭제를 원칙으로 하되 연구책임자의 요청에 따라 공개가 가능하다.

〈표 5-10〉 비식별화 처리 기본 원칙에 따른 삭제 변수

구분	항목
완전 삭제	- 개인식별 정보 ※ 개인식별 정보: 이름, 주민등록번호 등
	- 파라데이터 ※ 조사구 정보, ID 등
	- 생월, 생일
조건부 삭제	- 직종 또는 직업, 산업분류
	- 종교
	- 텍스트 문항
※ '조건부 삭제'의 경우, 연구자의 요청에 따라 공개 가능함	

출처: 저자 작성

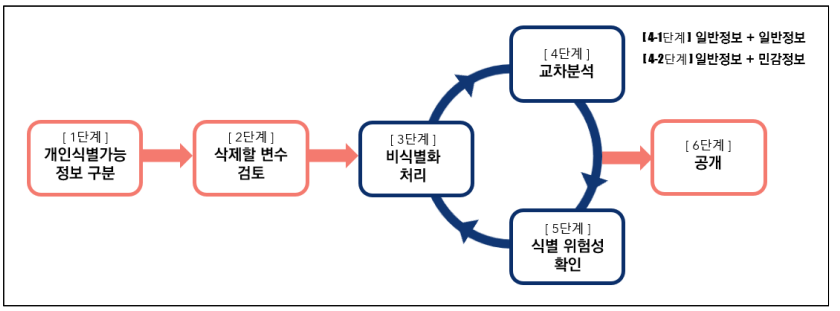
다. 비식별화 처리 절차

〈표 5-11〉 원내 비식별화 처리 절차

비식별화 처리 과정(안)	
1단계	- 개인식별이 가능한 일반정보와 개인식별이 가능한 민감 정보로 구분함. ※ 일반정보: 인구사회학적 특성 정보 ※ 민감 정보: 조사의 목적 및 특성에 따른 민감한 정보
2단계	- 삭제할 변수를 검토함. ※ 개인식별 정보, 파라데이터, 직종 또는 직업, 산업분류 텍스트 변수 등
3단계	- **비식별화 처리** **※ 단일 변수의 1차 비식별화 처리가 가장 중요하며, 보고서에 제시된 범주를 기준으로 비식별화 처리함. (단, 단일 빈도가 5 미만인 경우에도 비식별화 처리할 수 있음.)** **※ 분석에서 사용되지 않은 변수는 비식별화 처리 수준(부록)을 참고하여 처리함.** **※ 단, 연구책임자의 판단에 따라 처리 방식이 변경될 수 있음.**
4단계	- 개인식별 가능 정보 간의 교차분석을 통해 추가 비식별화 처리가 필요한 개인식별 가능 변수를 탐색함. ※ 단, 교차분석 진행 시 1차적으로 비식별화 처리된 변수를 이용함.
4-1단계	- 교차분석 = 개인식별 가능 일반정보 + 개인식별 가능 일반정보
4-2단계	- 교차분석 = 개인식별 가능 일반정보 + 개인식별 가능 민감 정보
5단계	- 식별 위험성을 확인함. ※ 최종적으로 **개인식별 위험성은 연구책임자가 확인하고 결정**함.
6단계	- 공개

출처: 저자 작성

〔그림 5-4〕 비식별화 처리 과정



출처: 저자 작성

라. 비식별화 처리 적용 기준

원내 비식별화 처리를 적용하기 위한 개인식별 가능 정보의 항목은 성별, 연령, 지역, 최종학력, 가구소득, 혼인상태, 직종, 거주 형태, 사회복지제도, 만성질환, 장애, 경제활동상태, 가구 구성, 자산 및 부채, 종교로 구성된다. 단, 원내 비식별화 처리 적용 기준표 이외의 내용은 〈부표 1-1〉을 참고하여 처리한다.

〈표 5-12〉 원내 비식별화 처리 적용 기준표

항목	분류	비식별화 처리	비고
성별	일반	현행 유지	
연령	일반	- 10세 단위 범주화	
지역	일반	- 단계별 비식별화 처리 (1단계) 보고서의 분석 기준으로 비식별화 처리 (2단계) 권역별 처리 (1안) 7개 권역(서울, 인천/경기, 대전/충청/세종, 광주/전라, 대구/경북, 부산/울산/경남, 강원/제주) (2안) 7개 권역(서울, 경기/인천, 강원권, 충청권, 대구/경북권, 부산/경남권, 호남/제주권) ※ 17개 시도와 동부/읍면부가 조합되었을 때 민감해질 수 있음 ※ 또한, 세종과 제주의 모집단 수가 상대적으로 작음	

항목	분류	비식별화 처리	비고
최종 학력	일반	- 현행 유지 ※ 낮은 연령이지만, 학력이 높은 경우 민감해질 수 있음(단, 조사 연령에 10대가 포함될 경우)	
가구 소득	일반	- 범주화(100만 원 단위)	500만 원 이상으로 처리
혼인 상태	일반	- 현행 유지 ※ 다만, 혼인상태와 자녀 유무가 조합되었을 때 민감해질 수 있어 주의를 요함(미혼모, 미혼부의 식별 가능성 생김)	
직종 또는 직업	일반	- 삭제 권고 - 연구책임자의 요청 시 공개(재범주화 처리 가능성 있음) ※ 직종(직업) 중 특히 군인은 보기 항목으로 명시되어 있으므로 개인식별 가능성이 생김	
거주 형태	일반	- 현행 유지 ※ 임대아파트와 관련 정책이 조합되었을 때 민감해질 수 있음	
사회 복지 제도	민감	- 연구책임자와 논의 필요	교차분석 시 사회복지 제도는 맞춤형 급여만 진행
만성 질환 별 유무	민감	- 연구책임자와 논의 필요	
장애	민감	- 연구책임자와 논의 필요 ※ 장애 종류 및 중증도와 관련 정책이 조합되었을 때 민감해질 수 있음	
경제 활동 상태	일반	- 단계별 비식별화 처리 (1단계) 보고서의 분석 기준으로 비식별화 처리 (2단계) 임금근로자 = 상용직, 임시직, 일용직 근로자 비임금근로자 = 고용주, 자영업자, 무급 가족 종사자 (조사에 따라) 실업자, 비경제활동인구, 기타로 구분	
가구 구성	일반	- 가구원 수는 단계별 비식별화 처리 (1단계) 보고서의 분석 기준으로 비식별화 처리 (2단계) 빈도 확인 후 범주화 처리 - 가구 형태는 연구책임자와 논의 필요	

항목	분류	비식별화 처리	비고
자산 및 부채	일반	- 단계별 비식별화 처리 (1단계) 보고서의 분석 기준으로 비식별화 처리 (2단계) 금액 규모 확인 후 범주화 처리	
종교	일반	- 단계별 비식별화 처리 (1단계) 보고서의 분석 기준으로 비식별화 처리 (2단계) 삭제	

출처: 저자 작성

마. 원내 조사자료 비식별화 처리를 위한 시나리오

원내 조사자료의 비식별화 처리 과정에 대해 두 가지 시나리오를 가정하여 설명하고자 한다. 첫 번째로 범주화와 관련된 시나리오이다. 범주화는 비식별화 처리 시 가장 먼저 고려할 수 있는 기본적인 방법으로 다른 방법에 비해 실무자가 비교적 쉽게 구현할 수 있는 매우 효과적인 방법이라 할 수 있다. 항목이 가진 정보의 미세한 정도를 조정하여 정보의 정확성을 감소시키는 대신 개인정보 노출의 위험을 줄일 수 있다는 장점이 있다. 범주화 방법은 크게 연속형 변수를 범주화하는 방법, 범주화 변수를 재범주화하여 재범주하는 방법, 상단 또는 하단의 특정 값을 범주화하여 상하단 범주화 처리를 하는 방법 세 가지로 분류할 수 있다.

〈표 5-13〉 실무 적용을 위한 비식별화 처리 시나리오 1에서 비식별화 처리 전 원자료는 지역, 지역 상세, 지역 구분, 연령, 성별, 직종, 직업, 혼인상태, 가구원 수, 월평균 근로소득과 같이 원내 조사에서 많이 조사하고 있는 항목으로 구성되어 있다. 그러나 이러한 정보들은 항목 수준에 따라 개인식별 가능성이 높을 수 있어 조사 항목의 수준을 면밀히 살펴볼 필요가 있다.

〈표 5-13〉 실무 적용을 위한 비식별화 처리 시나리오 1 - 범주화

비식별화 처리 전 원자료

ID	지역	지역 상세	지역 구분	연령	성별	직종	직업	혼인 상태	가구원 수 (명)	월평균 근로소득 (만 원)
1	서울	서울	도곡동	50	남자	전문가	의사	유배우	2	5,000
2	세종특별자치시	세종	반곡동	45	남자	관리자	고위 공무원	미혼	1	500
3	제주 특별 자치도	제주시	노형동	40	남자	농림어업 숙련 종사자	어부	유배우	7	1000
4	강원 특별 자치도	원주시	중앙동	60	여자	비경제활동	무직	미혼	1	0
5	제주 특별 자치도	제주시	노형동	45	여자	군인	장교	사별	2	350

↓

비식별화 처리 후 자료

ID	지역	지역 상세	지역 구분	연령	성별	직종	직업	혼인 상태	가구원 수 (명)	월평균 근로소득 (만 원)
1	수도권	삭제	삭제	50~59세	남자	전문가	삭제	유배우	2인	500만 원 이상
2	비수도권	삭제	삭제	40~49세	남자	관리자	삭제	미혼	1인	500만 원 이상
3	비수도권	삭제	삭제	40~49세	남자	기타	삭제	유배우	5인 이상	500만 원 이상
4	비수도권	삭제	삭제	60세 이상	여자	비경제활동	삭제	미혼	1인	100만 원 미만
5	비수도권	삭제	삭제	40~49세	여자	기타	삭제	사별	2인	300~400 만 원 이하

출처: 저자 작성

‘비식별화 처리 전 원자료’에서 1번 레코드는 의사이면서 월평균 5,000만 원의 고소득자, 5번 레코드는 제주에 거주하며 군인인 여성이라는 개체의 특이성으로 인해 개인식별 위험의 정도가 다른 개체에 비해 높다. 따라서 거주지역이나 직업, 근로소득 정보는 개인식별 가능 정보이기 때문에 비식별화 처리가 반드시 필요한 항목으로 분류되어야 할 것이다.

‘비식별화 처리 후 자료’를 살펴보면, 범주형 항목인 17개 시도를 수도권/비수도권으로 범주를 재조정하여 재범주화된 항목으로 처리하고, 지역에 대한 상세정보를 삭제하였다. 직종은 레코드가 적거나 비슷한 직종을 결합하여 재범주화된 항목으로 처리하고, 상세 직업 내용은 삭제하여 개인식별 가능 위험성을 낮추었다. 또한, 연속형 수치로 있던 연령 항목을 10세 단위로 구간화하여 범주형 항목으로 처리하였다. 마지막으로 가구원 수와 월평균 근로소득은 특정 값 이상 또는 이하 값은 해당 레코드가 적을 수 있어 개인정보 노출 위험이 크기 때문에 상하단 범주화 처리를 통해 개인정보 노출의 위험을 줄이는 방식을 택하였다.

〈표 5-14〉 실무 적용을 위한 비식별화 처리 시나리오 2 – 가공 변수

비식별화 처리 전 원자료

ID	직종	맞춤형 급여 수급 여부					키	몸무게
		생계급여	의료급여	주거급여	교육급여	생계급여 수급액 (만 원)		
1	전문가	아니오	아니오	아니오	아니오	비해당	210	80
2	관리자	예	예	예	아니오	388	160	45
3	전문가	아니오	아니오	아니오	아니오	비해당	160	45
4	비경제활동	예	예	예	아니오	388	180	50
5	군인	아니오	아니오	아니오	아니오	비해당	175	60

↓

비식별화 처리 후 자료

ID	직종	맞춤형 급여 수급 여부	생계급여 수급액 (만 원)	BMI
1	전문가	아니오	삭제	18.1
2	관리자	예	삭제	17.6
3	전문가	아니오	삭제	17.6
4	비경제활동	예	삭제	15.4
5	군인	아니오	삭제	19.6

출처: 저자 작성

두 번째로 가공 변수와 관련된 시나리오를 살펴보고자 한다. 앞에서 설명한 범주화는 항목 조정으로 변수에 변화를 주는 방법이라면, 가공 변수는 변수를 새로이 생성하여 새로운 정보를 생성한다는 점에서 그 차이가 있다.

〈표 5-14〉 실무 적용을 위한 비식별화 처리 시나리오 2의 '비식별화 처리 전 원자료'를 살펴보면 맞춤형 급여 수급 여부는 생계, 의료, 주거, 교육 별로 수급 여부를 알 수 있다. 만약 해당 정보가 민감 정보라고 판단되는 조사에서는 비식별화가 필요하고, 급여 종류 중 한 개 이상 수급하는 경우에는 맞춤형 급여 수급 여부로 가공하여 비식별화 처리를 한다. 추가로 생계급여 수급액 항목을 삭제해야 생계급여 수급 여부를 파악할 수 없게 된다.

몸무게처럼 변동성이 있는 정보에 비해서 키는 변동성이 적으므로 비식별화가 필요한 항목이다. 예를 들어, 키가 210cm인 레코드는 다른 변수들과의 조합을 고려했을 때 개인식별 가능성을 배제할 수 없으므로 비식별화 처리가 필요하다. 데이터에서 키만 있는 경우에는 상하단 범주화 방식을 고려하여 키를 범주화 형태로 비식별화 처리를 진행한다. 하지만 몸무게도 함께 있는 데이터라면 BMI 지수로 가공하여 비식별화 처리를 한다.

바. 원내 조사자료 검토의견서 작성 사례

제시한 원내 비식별화 처리 가이드라인에 따라 생성된 검토의견서와 체크리스트를 실무에 적용해 보았고, 추가로 k-익명성 수준의 노출 위험도의 적용 가능성을 검토해 보았다. 원내 조사자료인 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사(2022)에 대한 검토의견서 내용 중

일부 발췌한 부분은 다음과 같다. 첫째, 마이크로데이터, 조사표, 코드북 간 매칭 부분에서는 '문항의 순서 확인', '데이터와 코드북의 변수명 불일치 확인', '데이터 보기 항목 확인' 등 전반적으로 이용자의 편의성을 제고하기 위한 자료의 직관성을 확보할 수 있었다. 또한 '연속형 변수의 응답 범위 확인', '중복응답 로직 오류' 등 조사표 설계 시 발생할 수 있는 오류를 최종적으로 검토함으로써 데이터의 품질을 향상시킬 수 있었다. 추가로 '텍스트 문항 제공 여부 확인'을 통해 개인 및 기관에 대한 정보 노출 여부에 대해 점검할 수 있었다. 둘째, 마이크로데이터 비식별화 처리 부분에서는 개인식별 정보와 민감 정보를 선별하고, 각 변수에 대한 비식별화 처리 전 항목 요약, 처리 방식, 비식별화 처리 후 세부 내역을 포함하여 이전의 서술식으로 기재한 것보다 더 구체적이고 명확하게 해당 데이터에 대한 개인식별 정보 및 민감 정보를 검토할 수 있었다. 또한, 원내 비식별화 처리 가이드라인에 따른 처리 방안을 제안함으로써 실무자들은 주관적인 판단을 줄일 수 있었다. 셋째, 검토의견 부분에는 k-익명성을 통해 노출 위험도를 측정하였다. 이를 검토하기 위하여 성별, 연령, 최종학력, 결혼상태, 장애, 직종을 활용 변수로 구성하였고, 비식별화 처리 전과 후의 변수를 이용한 두 가지 시나리오를 통해 k-익명성의 수치를 비교하였다. 비식별화 처리 전 시나리오에서는 데이터의 약 65%가 2-익명성을 만족하지 않았으며, 약 96%가 5-익명성을 만족하지 않았다. 반면 비식별화 처리 후 시나리오에서는 데이터의 약 9%가 2-익명성을 만족하지 않았으며, 약 31%가 5-익명성을 만족하지 않았다. 따라서 k-익명성에 따른 노출 위험이 감소하였다는 것을 확인할 수 있었다.

위와 같이 제시한 원내 비식별화 처리 가이드라인을 원내 조사자료에 적용한다면 개인정보 보호와 데이터 유용성 간의 균형을 맞추면서 실무자들 간의 일관된 비식별화 처리가 가능할 것이다.

〈표 5-15〉 원내 조사자료 검토의견서 사례

1. 검토의견

③ 검토의견

○ 개인식별 가능 변수는 위와 같이 비식별화 처리를 하고, 텍스트 변수는 삭제를 제안하였습니다.

○ k-익명성을 통한 노출 위험을 파악하기 위하여 성별, 연령, 최종학력, 결혼상태, 장애, 직종을 활용 변수로 구성하였습니다. 비식별화 처리 전 시나리오에서는 데이터의 약 65%가 2-익명성을 만족하지 않았으며, 약 96%가 5-익명성을 만족하지 않았습니다. 반면 비식별화 처리 후 시나리오에서는 데이터의 약 9%가 2-익명성을 만족하지 않았으며, 약 31%가 5-익명성을 만족하지 않았습니다.

노출 위험	비식별화 처리 전	비식별화 처리 후
2-anonymity (Number of observations violating)	425 (65.18%)	57 (8.74%)
3-anonymity (Number of observations violating)	551 (84.51%)	95 (14.57%)
5-anonymity (Number of observations violating)	627 (96.17%)	200 (30.68%)

○ 추가로 비식별화 처리된 변수들의 교차분석을 진행하였을 때, 일부 변수에서 단일 레코드가 발생하였습니다. **다만, 개인식별 정보가 1차 비식별화(지역 변수 삭제, 연령 범주화, 소득 범주화 등)되어** 추가 비식별화 처리 없이 제공하려고 합니다. 그럼에도 추가 처리가 필요하시면 요청 바랍니다.

출처: 한국보건사회연구원. (2024). 원내 조사자료 검토의견서에서 일부 발췌했음.

제3절 비식별화 데이터의 이관 및 관리 절차 설정

본 절에서는 비식별화 처리 또는 비식별화 관련 업무가 연구데이터의 이관 및 관리 체계와 어떠한 관련성을 가지는지를 검토한다. 또한 조사데이터의 이관 또는 연구데이터의 관리 차원에서 비식별화 업무 절차가 추진될 시에 필요한 세부 절차에 대하여 논의[3]한다. 이를 위하여 기존의 조사데이터 관리 절차와 가명정보 관련 절차, 연구데이터 관리 체계 등과 비식별화 처리 간의 관계를 확인하였다.

1. 비식별화 데이터 이관 절차

가. 조사데이터 관리 절차 내 검토

비식별화 처리가 요구되는 데이터는 주로 한국보건사회연구원에서 생산한 조사자료이며, 이에 따라 원내 조사데이터 관리 절차 내에서 비식별화 데이터의 처리가 검토될 필요가 있다. 기존 조사데이터의 관리 절차는 〔그림 5-5〕와 같이 총 8단계의 절차가 있으며, 이 가운데 데이터의 이관과 관련된 절차는 (5단계)데이터 검토, (6단계)데이터 제출, (7단계)데이터 보관, (8단계)데이터 제공 단계이다.

비식별화 데이터의 관리는 조사데이터가 비식별화되는 일련의 과정과 비식별화 처리가 된 데이터의 이관을 포함한다. 5단계인 데이터의 검토 과정에서는 실질적인 비식별화 처리가 검토되지 않지만 연구진의 데이터

3) 해당 절의 내용은 관련 지침 및 규정의 적용 등 비식별화 처리 절차가 실질적으로 운영될 경우를 상정하여 이때 필요한 데이터 관리 체계를 제안하는 것이다. 또한 조사데이터 및 연구데이터의 관리 절차는 본 원에서 수행된 선행연구의 내용을 바탕으로 작성하였다. 이에 따라 본 연구가 수행된 시점에서의 실제 관련 업무 수행 내용 및 절차와 다소 차이가 있다.

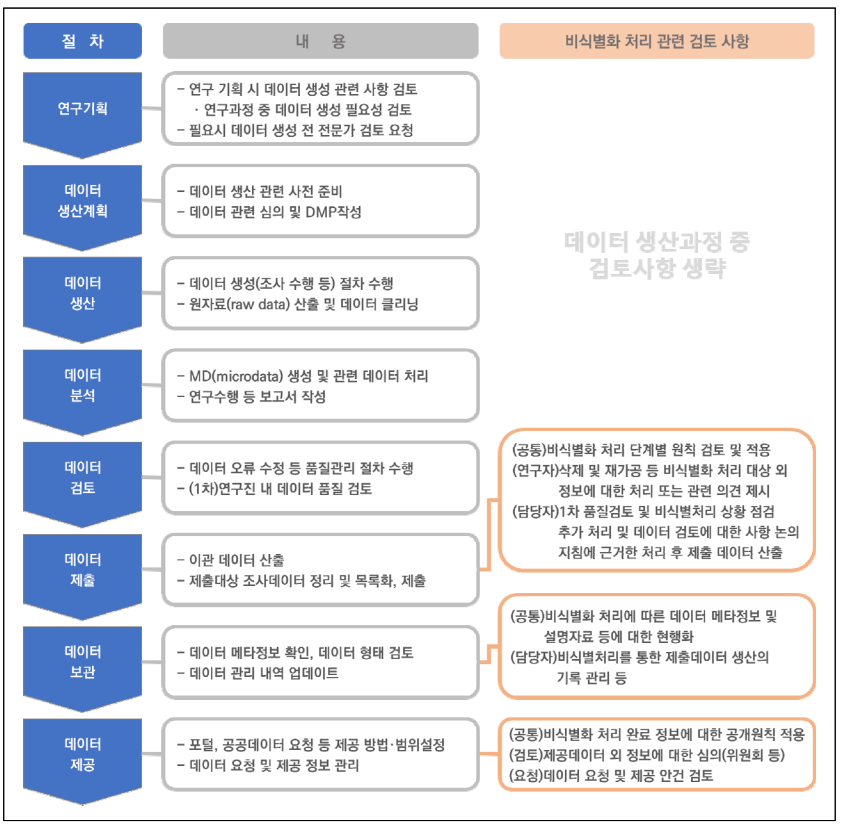
오류 검토 및 품질관리 절차에서 개인식별 가능한 정보 또는 민감 정보에 대해 연구 목적에 근거하여 비식별화 처리가 이루어진다. 이때 연구진에서는 비식별화 대상 정보에 대한 구분과 민감 정보에 대한 처리 범위를 설정하게 된다.

다음으로 6단계인 연구데이터 제출 단계에서는 연구자가 비식별 처리 단계에 대한 검토를 진행하여 비식별화 처리가 필요한 정보의 범위를 설정하게 된다. 예를 들어 삭제가 필요한 개인정보, 연구 목적에 부합하는 비식별화 처리, 비식별화 처리 지원이 요구되는 정보, 그 밖에 비식별화 관련 연구책임자의 의견 수렴 등이 이 단계에서 이루어질 수 있다. 이때 비식별화 관련 업무를 수행하는 실무자는 1차적으로 연구진에서 수행한 비식별화 결과를 점검하고 추가적인 비식별화 처리를 수행한다. 이때 비식별화 지침에 근거한 비식별화 처리 단계를 기준으로 데이터 처리를 진행하고 그 결과를 연구자에게 공유한다.

연구데이터 보관 단계에서는 원시자료뿐만 아니라 메타정보를 포함한 데이터 설명자료 등을 이관하고 관리 내역을 기록한다. 이때 비식별화 처리에 대한 설명자료를 현행화하고, 비식별화 처리된 내역과 그 사유에 대하여 기록 관리한다. 담당자는 1차 비식별화 처리에 대한 관리뿐만 아니라 이후 내검을 통한 추가적인 수정 결과 등을 관리한다.

데이터 제공 단계에서는 데이터의 제공 범위를 설정하게 된다. 기본적으로 조사자료관리지침에 해당되는 조사데이터 가운데 비식별화 처리를 통한 익명데이터는 공개 대상으로 설정된다. 다만 최종적인 공개처리 이전에 관련 위원회의 검토를 통하여 해당 익명데이터가 포털 등을 통하여 공개되는 것이 적절한지를 확인한다. 이후 데이터 수요에 대응하기 위한 절차는 기존 데이터 제공 절차를 따르게 된다.

〔그림 5-5〕 기존 데이터 관리 절차 기준 비식별화 처리 검토 사항



주: 박성준 외(2023)의 도식화 결과를 일부 수정하여 관련 내용 추가 및 수정하여 작성했음.
출처: 저자 작성

나. 연구데이터 관리 업무 절차 내 적용

비식별화 대상이 되는 데이터는 연구데이터에 해당된다. 특히 조사를 통하여 수집된 데이터가 아닌 경우에도 개인정보나 민감 정보에 대한 검토가 요구된다. 이러한 관점에서 비식별화 처리의 검토 범위는 연구데이터를 확대될 수 있다. 아래의 내용에서는 연구데이터 관리 절차 가운데

비식별화 처리 관련 절차가 해당되는 부분을 구체적으로 검토하였다.

연구데이터의 관리 절차[4] 가운데 제출용 데이터의 처리, 이관 신청 및 보안 범위 설정, 데이터 이관 처리가 비식별화 처리 및 데이터 이관과 관련된다. 특히 제출용 데이터를 생성하기 위한 일련의 과정과 데이터의 보안 범위를 설정하기 위한 검토 및 심의 절차 중에 비식별화 처리 및 비식별화 관련 업무 절차가 매우 중요하게 다루어진다.

먼저, 제출용 데이터가 '익명처리'를 요구할 경우에, 제출용 데이터의 처리는 익명처리를 의미하게 된다. 비식별화를 통한 익명데이터 생산은 연구자와 실무 담당자 간의 비식별화 처리 절차를 통하여 이루어진다. 먼저, 연구자는 비식별화 처리 기준에 따라 제출용 데이터의 구성을 설정한다. 실무 담당자는 연구자가 보내온 제출용 데이터 구성을 검토하며, 추가적인 비식별화 처리 및 비식별화 처리에 대한 검토를 수행한다. 이 과정에서 연구자와 담당자가 비식별화 처리 범위, 처리 방법, 처리 결과 등에 대하여 논의하게 된다.

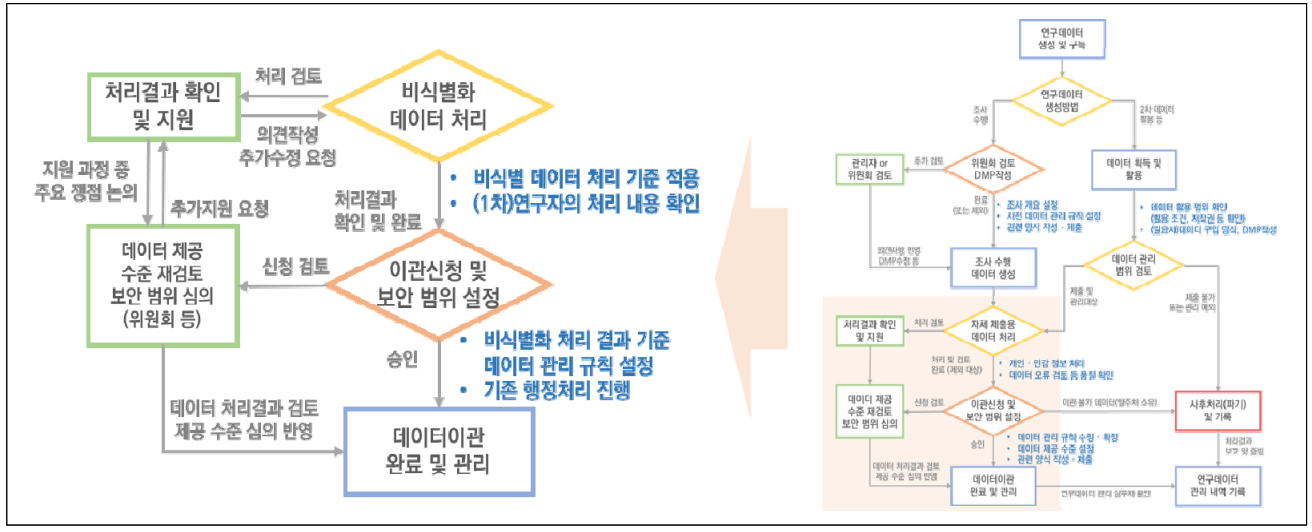
일반적으로 비식별화 처리 과정을 통하여 데이터의 제공 범위가 설정되고, 그에 따라 익명처리가 이루어지지만, 비식별화 처리 과정에서 발생하는 쟁점 사항은 관련 위원회를 통하여 검토 및 심의가 이루어질 필요가 있다. 검토 및 심의할 사항은 연구 목적 또는 데이터 제공 의무에서 벗어난 데이터의 처리나 비식별화 범위 설정 여부가 불명확한 정보의 처리, 그 밖에 데이터의 관리 측면에서 데이터 관련 전문가의 의견이 요구되는 사항 등이 해당된다. 비식별화 처리 과정에서 논의가 필요한 경우나 처리 결과에 대해 재검토가 요구되는 경우에는 관련 위원회에 검토를 요청할 수 있으며, 이때 연구자 및 담당자의 의견이 반영될 수 있도록 한다. 이는 기존 데이터 관리지침이나 비식별화 관리지침에 기반하여 다루어지기 어

4) 연구데이터의 관리 절차에 대한 세부적인 내용은 박성준 외(2023) 참고

려운 사항들에 한하여 필요한 절차이며, 이때 외부 전문가의 의견 또한 반영될 수 있다.

이상의 과정을 통하여 데이터의 비식별화 처리가 완료되고 관련 위원회를 통한 데이터 제공이 승인된 경우에는 기존 데이터 이관 및 보관 절차를 따르게 된다. 또한 데이터의 이관에 요구되는 각종 행정처리를 완료하되 연구데이터의 관리와 관련한 사항(데이터의 요청, 추가적인 내검 결과 반영, 가명정보 활용 요청 등)은 지속해서 관리되어야 한다.

[그림 5-6] 연구데이터 관리 업무 절차 중 비식별 데이터 관련 절차 적용



주: 박성준 외(2023)의 도식화 결과를 일부(음영 표기) 수정하여 관련 내용 추가 및 수정하여 작성했음.
출처: 저자 작성

다. 가명정보 처리 절차와의 관계

데이터에 대한 비식별화 처리를 통하여 익명데이터와 가명데이터가 생산될 수 있다. 본 연구에서는 비식별화를 통한 익명데이터 생성에 초점이 맞추어져 있으므로 가명처리 후 데이터 관리[5]에 대한 사항은 생략한다.

데이터의 신청 내용에 따라 각 데이터는 비식별화 과정의 수준을 결정하게 되며, 이때 익명처리가 필요한 경우에는 데이터 제공 범위 설정과 함께 익명처리의 범위 및 수준을 결정하게 된다. 데이터의 신청 내용은 연구자의 의견 반영을 통하여 제공 가능한 익명 데이터 세트가 구성된다. 다만, 비식별화 처리가 필요한 경우에는 연구자의 의견에 따른 세트 구성 외에 담당자의 비식별화 처리 결과가 반영된 결과가 추가적으로 반영되어야 한다.

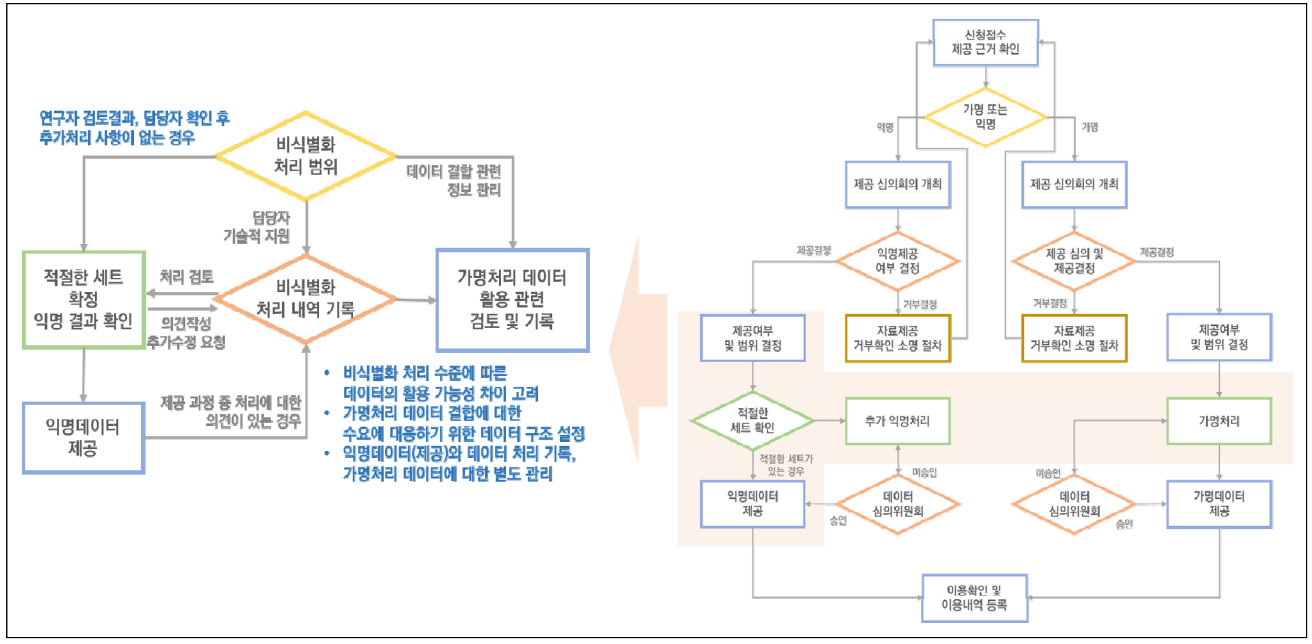
비식별화 처리 담당자는 관련 지침에 따라 비식별화를 수행하고 그 결과는 연구자와 협의한 다음에 확정한다. 연구자가 추가 요청을 하는 경우에는 익명처리 결과를 재검토할 수 있으며, 익명처리된 데이터는 데이터 제공 절차에 따라 데이터 신청자에게 제공된다.

그러나 가명정보의 활용을 위해서는 비식별화 처리 전의 데이터 또는 데이터 결합이 가능한 수준으로 가명처리가 되도록 데이터를 관리해야 하며, 이용기관이 가명정보에 대한 보호조치를 적절하게 하고 있는지에 대해서도 판단해야 한다. 이에 따라 비식별화 처리 담당자는 가명처리 데이터 활용에 법제적으로 요구되는 사항을 고려해야 한다. 예를 들어 연구목적에 따라 관리가 요구되는 원자료의 세트 검토, 가명처리 후 데이터 결합에 요구되는 정보의 생성 및 관리, 이상의 절차에서 발생한 업무처리 내용, 근거, 결과의 기록 등을 지켜야 한다. 다만, 이상의 과정은 가명정보 처리 관련 절차에 따른 역할 설정에 따라 재조정될 수 있다.

---

5) 가명처리 데이터의 관리 절차는 이혜정 외(2022) 참고

[그림 5-7] 가명 및 익명 정보 처리 절차 중 비식별화 처리 절차의 적용



주 1: 이혜정 외(2022)의 도식화 결과를 일부(음영 표기) 수정하여 관련 내용 추가 및 수정하여 작성했음.
주 2 :연구수행 시점에서는 가명정보 처리 절차의 적용을 검토하고 있으므로 실제 관련 업무 수행 내용 및 절차와 다를 수 있음.
출처: 저자 작성

2. 비식별화 데이터 관리 절차의 세분화

이상의 내용에서는 데이터 이관 절차 중 비식별화 관련 업무의 적용을 중심으로 논의하였다. 그러나 비식별화 데이터의 처리 과정을 별도로 살펴보았을 때 별도의 업무 과정이 설정될 필요가 있다. 특히 비식별화 데이터의 생산 및 관리에 요구되는 사항이 복합적(비식별 범위 설정, 비식별화 처리 수준 설정, 데이터 제공 의무의 이행 등)이라는 점을 고려하면, 이에 대한 구체적인 사항이 검토될 필요가 있다.

가. 비식별화 데이터 관리를 위한 단계적 환류체계

비식별화 처리의 과정과 비식별화 데이터의 관리를 위해서는 아래의 단계를 거치게 되며, 단계별로 연구자와 실무 담당자 간의 미시 환류 절차가 발생할 수 있다. 먼저, 비식별화 처리 범위를 설정하는 단계에서는 개인식별 가능 정보와 민감 정보 등을 구분하게 된다.

먼저, 비식별화 처리 범위 설정 과정에서 연구자는 비식별화 관련 지침에 따른 기준을 검토하게 되고, 1차적으로 이를 반영한 범위를 설정하게 된다. 한편, 담당자는 연구자가 설정한 범위를 검토하고 비식별화 처리가 요구되는 데이터의 처리 근거를 확인한다. 이러한 환류 과정에서 합의된 처리 결과, 데이터 세트가 연구 목적 또는 데이터 제공 의무에 부합하는지, 개인정보 보호 의무를 충실히 이행하고 있는지를 검토하게 된다.

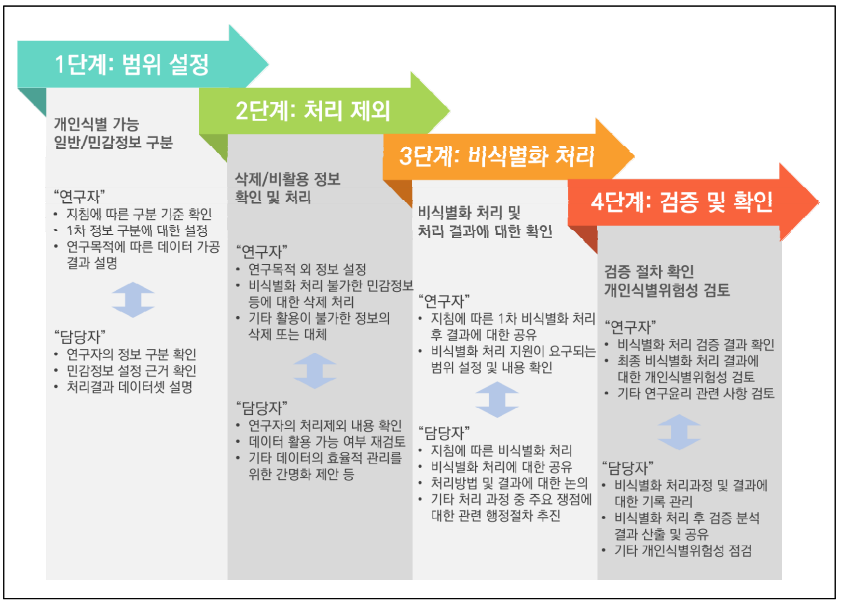
1단계에서 검토된 처리 범위 설정 결과를 바탕으로 개인식별 정보 및 민감 정보의 삭제와 비활용 정보에 대한 처리 등을 결정한다. 연구자는 연구 목적 외 수집된 데이터를 어떻게 처리할지 결정하고 비식별 처리가 불가한 민감 정보에 대한 삭제 처리를 요청할 수 있다. 담당자는 정보 삭

제 범위를 확인하고 연구자의 요청사항이 데이터 관리 목적에 부합하는지를 검토한다. 이때 담당자는 데이터의 효율적 관리를 위한 데이터 세트 구성을 연구자에게 제안할 수 있다. 삭제 및 비활용 정보에 대한 처리 내용과 결과는 기록되어야 하며, 데이터의 제공 시 연구자가 검토하고, 관련 위원회의 심의를 거치게 된다.

다음 단계에서는 비식별화 처리가 이루어지며, 비식별화 처리 지침에 따른 처리 단계가 실질적으로 이행된다. 연구자는 지침에 근거하여 비식별화 처리를 할 수 있으며, 기술적 사항에 대하여 담당자에게 비식별화 처리를 요청할 수 있다. 다만, 연구자 또한 비식별화를 통한 원자료의 수정 내역을 기록해야 하며, 이는 비식별화 담당자에게 공유되어야 한다. 만약 연구자의 비식별화 처리가 충분하지 않거나 적절하지 않은 경우, 담당자가 수정 요청이나 추가적인 처리 등을 할 수 있다. 한편, 담당자는 지침에 따른 비식별화 처리를 수행하고 그 결과를 연구자와 공유한다. 지침에서 벗어난 비식별화 처리 사항의 경우에는 연구자에게 해당 쟁점에 대하여 설명하고 이를 처리하기 위한 논의를 한다. 만약 연구자와 담당자 간 환류를 통하여 처리할 수 없는 사항이 발생할 경우에는 관련 위원회를 통한 자문 또는 심의를 받을 수 있다.

검증 및 확인 단계에서는 비식별화 결과에 대한 검증이 이루어진다. 연구자는 비식별화 처리 범위, 내용, 결과에 대하여 검토해야 할 의무가 있으며, 최종적으로 처리된 데이터가 데이터의 활용 목적, 개인정보 보호, 연구윤리 의무 등에 저촉되지 않는지를 확인한다. 한편, 담당자는 비식별화 처리 범위, 내용, 결과를 기록하고 관리해야 할 의무가 있으며, 비식별화 처리 후 검증 분석 결과를 연구자에게 공유하여야 한다. 특히 데이터 제공 과정 중에 발생할 수 있는 개인정보 보호 이슈 등에 대응하기 위하여 각 데이터별 비식별화 처리 결과를 저장·관리하여야 한다.

〔그림 5-8〕 단계적 환류체계 설정



출처: 저자 작성

나. 데이터 특성별 검토 과정 세분화

데이터의 비식별화 처리 과정 전에 해당 데이터에 대한 포괄적 검토가 선행된다. 예를 들어 해당 데이터가 조사자료관리지침의 관리 대상에 포함되는지, 한국보건사회연구원의 관리 대상에 해당되는지, 데이터 내 정보 중 비식별화 처리 등의 데이터 가공 절차가 필요한 사항이 있는지 등이 이에 해당된다. 만약 데이터 이관 대상에 해당되는 경우에는 한국보건사회연구원의 데이터 관리 절차에 따라 연구자는 기술적·행정적 절차를 추진하게 된다. 한편, 담당자는 이관 대상 여부를 확인하고 해당 데이터 안에 비식별화 처리 또는 품질관리를 위한 처리 등이 필요한지 여부를 확인한다. 이관 대상 데이터의 경우, 데이터 보관 및 제공을 위한 절차에 따

라 처리되며, 미이관 대상의 경우에는 데이터의 파기나 외부 이관 등의 절차를 추진하게 된다.

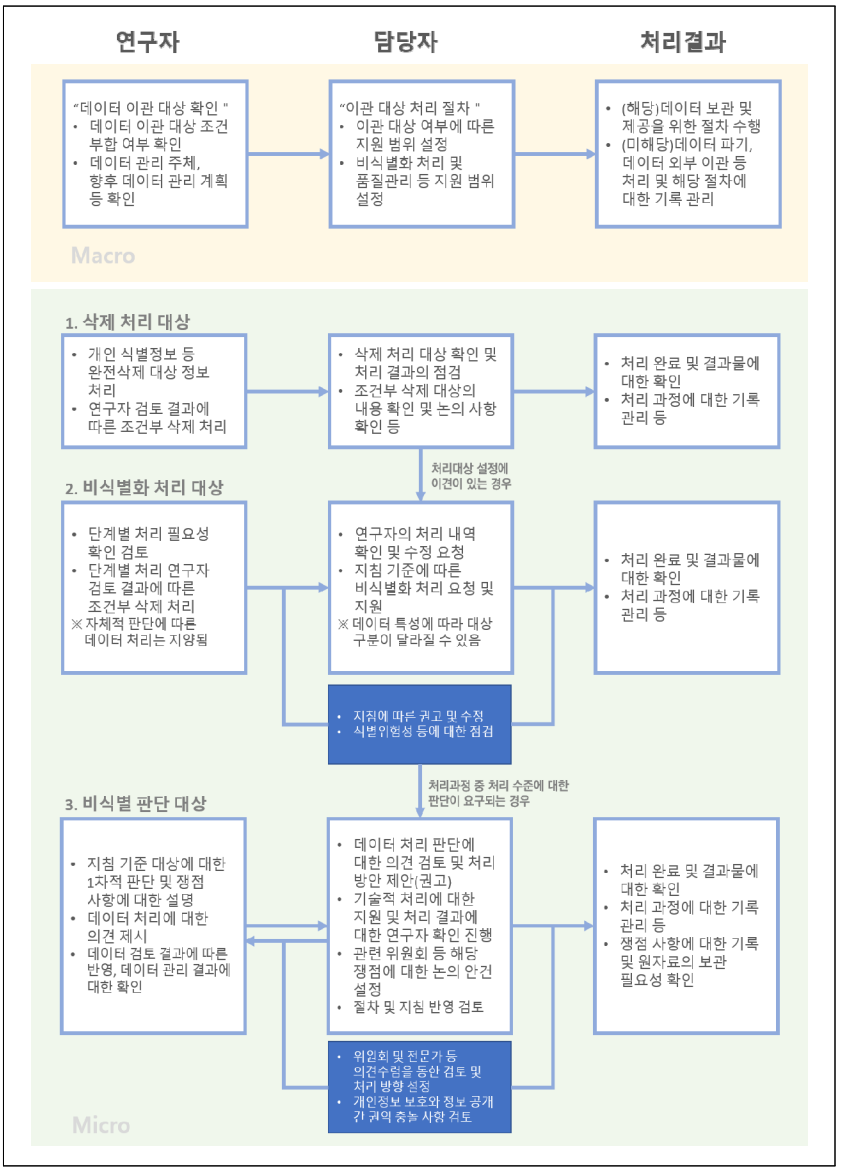
위 단계에서 이관 대상에 포함되는 경우, 담당자의 검토를 통하여 비식별화 처리가 요구되는 정보가 포함되는 경우에는 실질적인 비식별화 처리 절차를 추진하게 된다. 단계별 환류 체계에서 설명한 바와 같이 비식별화 처리 대상 정보는 크게 삭제 처리 대상과 비식별화 처리 대상으로 구분될 수 있으며, 그 밖에 지침 기준을 통하여 처리 여부 또는 방법을 확정할 수 없는 경우는 기존 절차와 다른 접근을 시도하게 된다.

먼저, 삭제 처리 대상으로 볼 수 있는 개인식별 정보나 비식별 처리가 불가한 민감 정보, 연구 목적에 해당되지 않는 비활용 정보는 연구자와 담당자의 검토에 따라 삭제 처리하게 된다. 다만, 삭제 처리 기준에 부합하지만 연구 목적의 이행을 위하여 비식별화가 필요한 경우에는 담당자와의 협의를 통하여 비식별화 방법을 논의할 수 있다.

비식별화 처리 대상은 지침에 근거한 비식별화 처리 절차 및 단계에 따라 처리 또는 관리된다. 1차로 연구자의 검토 결과에 따라 비식별화 처리가 이루어질 수 있으나 그 범위 및 방법은 공유되어야 하고, 자체적 판단에 따른 데이터 처리는 최소화할 필요가 있다. 한편 담당자는 연구자의 처리 내역 또는 처리 요청 사항을 검토하고 그 결과에 따라 기술적인 처리를 지원한다. 기본적으로 비식별화 처리 절차에 따르되 데이터의 세부 정보에 따라 처리 방식이 달리 적용될 수 있음을 가정해야 한다(조사데이터 기준, 각 변수마다의 특성에 따라 처리 방식을 달리 적용). 지침의 따른 처리가 적절하지 않은 경우, 지침의 기준에 벗어나는 경우, 지침을 통한 처리가 연구 목적 또는 데이터 제공 목적에 부합하지 않는 경우, 그 밖에 연구자의 판단에 따라 지침 외 처리 절차가 요구되는 경우에는 비식별 판단 대상으로 설정할 수 있다.

비식별 판단 대상은 지침에 따른 범위 설정에 해당되지 않는 경우, 연구자와 담당자 간 협의를 통하여 처리 대상 확정이 불가한 경우를 일컫는다. 이는 비식별 처리 대상 선정과 1차적 처리 과정 중에 발생하는 쟁점에 근거하여 설정된다. 연구자는 지침에 따른 데이터 처리(삭제 또는 비식별화 처리) 이후에 여전히 처리되지 않는 데이터에 대한 검토를 하게 된다. 담당자는 비식별 데이터 처리와 관련된 의견을 연구자에게 제안할 수 있으나 연구자와 담당자 간 이견이 발생하는 경우 관련 위원회(조사자료관리위원회, 데이터심의위원회, 연구윤리위원회 등)를 통하여 데이터의 처리와 관련된 자문 의견을 받을 수 있다(자문 요청 대상은 위원회 소속 내·외부위원이 될 수 있다). 이상의 절차가 추진되는 경우, 연구자와 담당자는 해당 안건에 대하여 설명할 의무를 가지며, 특히 담당자는 자문 결과에 대한 관리와 반영 계획 수립 등을 수행해야 한다. 또한 데이터 관리 담당자 또는 관련 위원회의 간사는 이상의 절차가 추진되는 일련의 과정에 대하여 지원할 필요가 있다.

[그림 5-9] 데이터 특성별 검토 과정 세분화



출처: 저자 작성

다. 비식별화 처리 데이터의 관리 구분

데이터의 성격과 비식별화 수준 등에 따라 처리 데이터의 관리 방향이 다르게 설정된다. 이를 연구데이터 관리 수준에 적용하여 설명하면 다음과 같다(연구데이터 관리 수준에 대한 사항은 〔그림 5-10〕 참고).

비식별 처리 대상은 기본적으로 관리 대상에 해당되는 경우에 한한다. 따라서 비식별 처리의 수준과 관계없이 데이터의 관리 주체가 한국보건사회연구원인 경우(위탁 관리 포함)에 비식별화 처리 대상으로 설정된다. 이 가운데 비공개 데이터의 경우에는 개인식별 정보 및 민감 정보를 적절하게 처리하지 못하는 경우에 해당되므로 데이터 관리의 효율성을 고려하여 개인정보 및 비활용 정보를 삭제하고, 최소한의 비식별화 처리를 수행해야 한다. 단, 데이터가 비공개 관리 대상으로 설정되기 위해서는 관련 위원회의 심의가 필수적이다.

한국보건사회연구원에서 생산한 조사데이터를 익명처리할 경우에는 마이크로데이터 제공 대상에 포함될 수 있다. 이 경우는 비식별화 처리를 익명 수준으로 수행하고 개인정보 및 민감 정보를 충분히 처리한 한 결과로 볼 수 있다. 이에 해당되는 데이터는 앞에서 제시한 비식별과 관련 기술적·행정적 절차가 완료된 경우에 한하며, 처리 결과에 대한 기록 관리가 완료된 경우를 가정한다. 마이크로데이터 제공 단계에 해당되는 데이터는 데이터 제공 담당자의 관리 범위에 속한다. 다만 마이크로데이터 제공 데이터 가운데 가명정보결합 요청이 발생한 경우, 데이터 활용 중 데이터의 오류가 확인된 경우, 개인정보 및 민감 정보 관련 이슈가 발생한 경우에는 연구자 및 비식별화 처리 담당자가 사후 처리를 진행할 수 있다. 이때 제공 데이터의 수정 내역에 대하여 데이터 제공 담당자는 기록, 관리해야 하며, 세부적인 기술적 처리는 비식별화 처리 담당자가 기록, 관리해야 한다.

[그림 5-10] 연구데이터 관리 범위 설정 및 세분화

출처: 박성준, 이기호, 허혜옥, 심현보. (2023). 보건복지 연구데이터의 통합적 관리 및 활용 방안 연구. 한국보건사회연구원.

제4절 소결

제5장에서는 한국보건사회연구원 연구보고서 작성 중에 생산된 조사자료에서 개인식별이 가능한 정보는 무엇인지 검토한 뒤, 비식별화 처리 현황을 살펴보고, 비식별화 처리 관련 가이드라인을 개발하였다. 그리고 비식별화 처리를 위한 절차가 실제 데이터 관리 업무에 적용될 경우를 가정하여, 연구원의 데이터 관리 체계 내에서 비식별화 데이터의 이관 및 관리 절차가 어떻게 반영될 수 있을지를 상세히 검토하였다.

현재 연구원에서 공개를 전제로 검토하고 있는 조사자료의 경우, 비밀보호 처리 부분에서 명확한 기준 없이 실무자들의 주관적 판단에 의해 데이터를 처리하고 있기 때문에 최소한의 개인식별 정보를 판단하는 기준과 처리에 대한 기준을 설정할 필요가 있다. 이를 위해 2021~2022년 조사자료를 검토한 결과, 개인식별 가능 정보로 사용되는 항목은 성별, 연령, 지역, 최종학력, 혼인상태, 장애, 종교, 가구소득, 경제활동상태, 직종, 거주 형태, 가구 구성, 만성질환, 자산 및 부채, 사회복지제도까지 15개로 요약된다. 동일한 항목일지라도 조사 목적에 따라 항목별 세부 조사내용은 조사별로 상이하기 때문에 해당 변수에 대한 구체적인 처리 방식을 설정하여 일반화된 처리 원칙을 만들 필요가 있다.

원내의 비식별화 처리 관련 가이드라을 개발하고, 공개용 조사자료의 사전 검토의견서 양식을 보완하였고, 일관된 데이터 처리를 위해 체크리스트를 새롭게 작성하였다. 기존의 검토의견서는 서술식으로 비식별화 관련 내용을 기술하였다면, 보완된 검토의견서 양식에는 비식별화 처리표를 작성하도록 하여 명확하게 해당 데이터에 대한 개인식별 정보 및 민감 정보를 검토할 수 있도록 하였다.

비식별화 처리 가이드라인에는 비식별화 처리 원칙을 세우고, 비식별

화 처리 과정을 단계별로 설정하여, 개인식별 가능 정보의 항목을 어느 정도의 수준까지 범주화할 것인지에 대한 기준표를 작성하였다. 비식별화 처리 절차 중 중요한 부분은 단일 변수의 1차 비식별화 처리이므로, 이 부분은 조사자료의 연구보고서에 제시된 범주를 기준으로 비식별화 처리를 하는 것을 원칙으로 하였다. 그리고 이에 대한 모든 사항은 연구책임자가 확인하고 결정하도록 하였다. 원내의 조사자료 비식별화 처리 관련하여 범주화와 관련된 시나리오, 가공 변수와 관련된 시나리오를 제시하여 비식별화 처리 작업에 대해 쉽게 이해할 수 있도록 하였다.

비식별화 데이터의 이관 절차는 원내의 조사데이터 관리 절차와 연구데이터 관리 업무 절차와 구분하여 살펴보았다. 비식별화 처리 및 비식별화와 관련한 업무에서 중요한 부분은 제출용 데이터를 생성하기 위한 일련의 과정과 데이터의 보안 범위를 설정하기 위한 검토 및 심의 절차라고 할 수 있다. 비식별화 처리 과정에서 논의가 필요한 경우나 처리 결과에 대해 재검토가 요구되는 경우에는 관련 위원회에 검토를 요청할 수 있으며, 이때 연구자 및 실무자의 의견이 반영될 수 있도록 한다. 이는 기존 데이터 관리지침이나 비식별화 관리지침에 기반하여 다루어지기 어려운 사항들에 한하여 필요한 절차이며, 이때 외부 전문가의 의견 또한 반영될 수 있다.

비식별화 데이터의 처리 과정은 비식별 범위 설정, 비식별화 처리 수준 설정, 데이터 제공 의무의 이행 등 많은 요소와 관련되어 있기 때문에 별도의 업무 과정이 설정될 필요가 있다. 비식별화 처리의 과정과 비식별화 데이터의 관리를 위해서는 1단계 범위를 설정(개인식별 가능 민감/일반 정보 구분)하고, 2단계 처리 예외 사항을 확인하고(삭제/비활용 정보 확인 및 처리), 3단계 비식별화 처리를 하고(비식별화 처리 및 처리 결과에 대한 확인), 4단계 검증 및 확인하는(검증 절차 확인 및 개인식별 위험성

검토) 단계적 환류체계를 설정하여야 한다.

데이터의 특성별로 검토 과정을 세분화할 필요도 있다. 비식별화 관점에서 데이터의 특성을 삭제 처리 대상, 비식별화 처리 대상, 비식별 판단 대상(지침에 따른 범위 설정에 해당되지 않는 경우, 연구자와 담당자 간 협의를 통하여 처리 대상 확정이 불가한 경우를 의미)으로 나누었을 때, 비식별 판단 대상은 관련 위원회 및 전문가의 의견 수렴을 통한 검토가 이루어져야 한다.

비식별화 처리를 한 데이터의 관리는 연구데이터의 관리단계로 보면 마이크로데이터 제공 단계로 볼 수 있다. 다만, 제공 데이터 가운데 가명정보 결합 요청이 발생한 경우, 데이터 활용 중 데이터의 오류가 확인된 경우, 개인정보 및 민감 정보 관련 이슈가 발생한 경우에는 연구자 및 비식별화 처리 담당자가 사후 처리를 진행할 수 있다.

제6장

결론 및 시사점

제6장 결론 및 시사점

제1절 결론

이 연구는 보건복지분야의 데이터 활용도를 제고하기 위해 익명화 수준의 데이터 비식별화 처리 방법을 검토하고, 3개의 데이터로 개인식별 위험을 비교하여, 비식별화 데이터의 생성 및 관리체계를 수립함으로써 데이터의 안전한 활용 및 확산 기반을 마련하고자 하였다.

제2장에서는 국내의 개인정보 비식별화와 관련된 금융분야 가명·익명처리 안내서, 가명정보 처리 가이드라인, 보건의료데이터 활용 가이드라인, 교육분야 가명·익명 정보 처리 가이드라인, 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인과 미국, 일본의 의료분야 비식별화 관련 사례를 검토하였다. 국내의 가이드라인은 법 개정과 실무 적용, 가명·익명 처리 사례의 누적 및 시대 변화에 따라 가명 또는 익명 정보를 활용하려는 수요자들의 요구에 부응하여 지속적으로 개정되고 있다. 이는 통계작성기관인 한국보건사회연구원도 데이터 제공 시 익명처리 절차 및 비식별화 처리에 대한 기준이 필요하다는 것을 시사한다.

제3장에서는 국제 표준인 ISO/IEC 20889의 비식별화 방법론과 차등정보보호, 재현 데이터 방법론을 중심으로 검토하였고, 활용 예시를 통해 이해를 돕고자 하였다. 실무적으로는 ISO/IEC 20889의 비식별화 방법론의 활용이 높겠으나, 차등 정보보호와 재현 데이터는 앞으로도 주목받는 비식별화 처리 기술이기 때문에 방법론적으로 중요하게 다루었다. 그리고 비식별화 방법론의 실무 적용을 위해 sdcMico를 사용하여 활용 예

시를 구체적으로 제시하였다.

제4장에서는 제3장의 비식별화 방법론을 활용하여 3개의 데이터에 대한 개인정보 노출 위험을 분석하였다. 한국복지패널, 가족과 출산 조사, 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사별로, 실무적으로 활용할 수 있는 개인식별 가능 항목을 범주화 비식별화 처리 방법을 사용해 노출 위험 시나리오 6가지를 구성하였다. 한국복지패널과 가족과 출산 조사는 시나리오 구성 시, 지역 변수의 범주 범위를 확대하였을 경우 노출 위험이 얼마나 증가할 것인지를 검토하였다. 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사는 연구원의 원내 조사자료이므로 어느 정도의 수준까지 비식별화 처리 작업이 필요한지에 대한 판단에 도움을 주고자 시나리오를 구성하였다. 6가지 시나리오를 통해 복지패널, 가족과 출산 조사 데이터는 지역 변수 공개 범위를 시도 단위까지 확대할 경우, k-익명성과 전체 위험성으로 어느 정도 노출 위험도가 증가하는지를 파악할 수 있었다. 한국복지패널과 가족과 출산 조사 같은 국가 승인 통계의 경우에는 개인정보와 관련된 변수의 범주 수준을 결정하기에 앞서, 해당 수준에서 생산되는 통계의 신뢰도를 우선 검토해야 할 필요성이 있다. 상대표준오차가 기준값보다 큰 경우, 범주 수준을 축소하고 이후 노출 위험도를 분석하는 과정이 필요하다. 지역사회 거주 정신질환자 조사는 원내의 조사자료 중 하나로, 다양한 준식별 정보의 조합으로 노출 위험을 측정하였을 때, 원자료 수준에서는 노출 위험도가 매우 높음을 알 수 있고, 변수별로 범주화, 재범주화가 필요하다는 것을 알 수 있었다. 향후 본 연구에서 다룬 노출 위험 시나리오를 더 확대하여 비식별화가 어려운 고위험 데이터의 경우 데이터 접근 제한 조치나 다층 보호 체계를 추가로 도입하는 방안에 대한 검토도 필요하다. 특정 데이터에 대해 접근 레벨을 제한하는 다중 인증 방식 도입과 복지 분야 외 유사 고위험 데이터를 포함한 다양한 시나리오를 향후 연구에 포함한다면 더 폭넓은 적용

성을 확보할 수 있을 것이다.

제5장에서는 한국보건사회연구원의 연구보고서 작성 중에 생산된 조사자료에서 개인식별이 가능한 정보는 무엇인지 검토한 뒤, 비식별화 처리 현황을 살펴보고 비식별화 처리 관련 가이드라인을 개발하였다. 비식별화 처리 가이드라인에는 비식별화 처리 원칙을 세우고, 비식별화 처리 과정을 단계별로 설정하여, 개인식별 가능 정보의 항목을 어느 정도의 수준까지 범주화할 것인지에 대한 기준표를 작성하였다. 비식별화 처리 절차 중 중요한 부분은 단일 변수의 1차 비식별화 처리이므로, 이 부분은 조사자료의 연구보고서에 제시된 범주를 기준으로 비식별화 처리를 하는 것을 원칙으로 하였다. 그리고 비식별화 데이터의 이관 및 관리 절차를 설정하여 연구원의 데이터 관리체계 안에서 비식별화 데이터 처리에 대한 부분을 상세히 검토하였다.

본 연구에서는 다루지 않았지만 비식별화 후에도 재식별 위험을 줄일 수 있도록 데이터 보안 및 보호 기술을 추가로 적용하는 방법도 활용할 수 있다. 데이터 접근 제어, 암호화, 익명화된 데이터의 복구 불가능성 유지 등의 조치를 단계별로 적용하는 방안을 구체화하면 비식별화 데이터의 안전성을 강화할 수 있을 것이다. 클라우드 환경에서의 안전한 데이터 저장 같은 세부적인 보안 프로토콜, VPN 사용 방안 등을 포함하여 데이터 저장 및 활용 환경에 대한 기술적 보호 조치도 가능하다. 데이터를 비식별화한 후 데이터 재식별 가능성을 모니터링하고, 외부 요인에 따라 노출 위험이 증가하는 경우 이를 수정할 수 있는 절차를 체계화하는 방안도 필요하다. 데이터를 주기적으로 모니터링하고 민감 정보에 대한 노출 위험 평가를 주기적으로 실시하여 위험을 사전에 차단하는 시스템을 구축하는 일 등이 필요하다. 데이터 재식별 가능성이 검토되는 시점과 기준, 필요한 경우 데이터를 다시 비식별화하는 절차를 포함하여 유연한 모니

터링 체계를 구축한다면 비식별화 데이터의 활용 가치 제고와 함께 데이터의 품질에 대한 신뢰성을 확보할 수 있을 것이다.

제2절 시사점

1. 비식별화 처리 관련 가이드라인 개발의 의미

데이터를 분석하고자 하는 개인과 기업은 개인정보 보호와 관련된 다양한 국가와 지역의 법적 요구사항을 충족해야 하며, 이러한 요구에 부응하는 국제 표준을 마련하여 규제 환경에 맞게 데이터를 보호할 수 있도록 지원해야 한다. 따라서 비식별화 방법론 개발은 사이버 보안 위협이 증가하는 환경에서 데이터를 안전하게 관리하고 보호하는 능력을 의미하며, 데이터 환경의 신뢰성과 직결된다. 또한, EU의 GDPR과 국내의 관련 법규정에 대응하여 통계작성기관은 통계자료 제공을 위해 비식별화 업무를 지원해야 하는 필요성이 생겼다. 통계청의 가이드라인이 2023년 하반기에 배포됨에 따라 통계작성기관인 한국보건사회연구원도 데이터를 제공할 때 익명처리 절차 및 비식별화 처리에 대한 기준이 필요하다는 것이다. 현재 연구원에서 공개를 전제로 검토하고 있는 조사자료의 경우, 비밀보호 처리 부분에서 명확한 기준 없이 실무자들의 주관적 판단에 의해 데이터를 처리하고 있기 때문에 최소한의 개인식별 정보를 판단하는 기준과 처리에 대한 기준 설정 내용을 가이드라인에 포함하였다. 비식별화 처리 관련 가이드라인을 개발한 것의 의미는 실무자들에게는 일관된 최소한의 업무 지침이며, 연구자들에게는 조사데이터의 개인정보 노출 위험 정도를 알려주는 설명자료라고 할 수 있다.

2. 비식별화 처리 관련 절차 및 적용

다양한 영역에서 개인정보 및 데이터 활용을 위한 비식별화 처리 방법 등이 논의되고 있으며, 이에 따라 보건·복지 영역에서도 관련 데이터의 비식별화 기준 및 지침 등을 마련하고 개선해 나갈 필요가 있다. 본 연구의 결과로 제시한 비식별화 처리 관련 가이드라인 설정은 한국보건사회연구원 내 데이터의 안전성 제고에도 기여할 수 있을 뿐만 아니라 관련 영역에서 생산·관리되는 데이터의 안전성 및 신뢰성 확보에 일정 부분 기여할 수 있을 것으로 보인다.

그러나 비식별화를 위한 기준 마련과 관련 업무의 설정이 실질적인 제도 도입과 유지를 담보하지는 않는다. 다시 말해 비식별화의 원칙과 기준을 설정하고 관련 업무에 대한 지침을 마련하여도 비식별화 처리 대상의 관리와 비식별화된 데이터의 이관 절차, 비식별화 처리 과정 시 지침 범위에 벗어난 쟁점의 대응 방안 등이 뒷받침되지 않는다면 연구 결과를 통하여 제시된 지침이 적용되기 어려울 수 있다. 그러므로 비식별화 데이터의 관리와 데이터 특성에 따른 검토 절차 설정 등 체계 마련이 무엇보다 중요하다.

이러한 관점에서 본 연구의 결과로 제시된 비식별화 처리 관련 가이드라인과 비식별화 데이터의 이관 및 관리 절차는 한국보건사회연구원의 비식별화 관련 처리 업무의 안착뿐만 아니라 데이터 생산 주체의 안전한 데이터 관리 및 제공 체계를 마련하였다는 점에서 의미가 있다. 특히 본 연구에서 검토된 연구데이터 관리 절차에서의 적용, 다양한 주체 및 역할 설정, 데이터 특성별 검토 절차 세분화 등의 결과는 보건·복지 영역 내 여러 데이터 생산 및 제공 주체들이 참고·적용할 수 있는 수준의 결과물로 볼 수 있다.

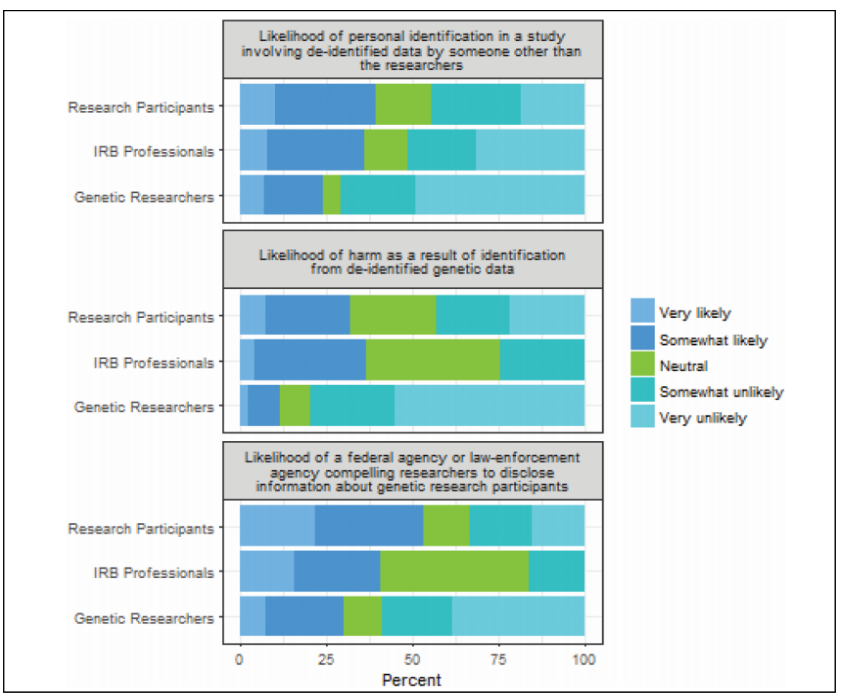
3. 데이터 프라이버시 리터러시 역량 강화

비식별화 데이터를 공개하고 활용하는 데 있어서 이해관계자는 조사에 참여한 연구 참여자, 연구자, 기관생명윤리위원회(IRB) 전문가이다. 이들 간에도 참여자의 개인정보 보호 문제를 인식하는 정도에 차이가 있기 때문에 비식별화 데이터 사용에 대한 견해도 다를 수 있다. Goodman et al.(2018)에서 Northwest Cancer Genetics Registry의 비식별화 처리가 된 데이터의 사용으로 인한 잠재적 피해에 대한 견해를 묻는 온라인 조사를 실시한 결과, 세 그룹 간에 견해 차이가 존재함을 알 수 있었다. 비식별 처리된 데이터에서 제3자에 의한 연구 참여자의 개인식별 가능성에 대한 질문에 유전학 연구자는 연구 참여자 및 IRB 전문가에 비해 연구 참여자가 식별될 것이라고 생각하는 비율이 가장 낮았다. 비식별화 처리가 된 데이터가 포함된 연구에서 연구 참여자가 개인식별이 가능하다고 생각하는 비율은 유전학 연구자에 비해 IRB 전문가가 2배, 연구 참여자가 2.6배 더 높았다. 이 세 그룹 간의 견해 차이는 연구에서의 고유한 역할에 영향을 받았을 가능성이 있다. IRB 전문가라는 인간 피험자 보호 및 준수 여부를 감독하는 IRB 전문가의 역할로 인해 위험 가능성을 최소화할 가능성이 더 높은 유전학 연구자에 비해 더 큰 위험을 인지할 수 있다. 이 연구는 연구자들이 비식별화 처리가 이루어진 데이터 공유에 관한 정책을 수립할 때 조사의 연구 참여자들이 인식하는 위험과 피해에 대한 견해를 더 잘 인식하고 고려해야 한다는 것을 시사한다. 즉, 비식별화 데이터를 생성하는 과정에서는 개인정보 보호, 데이터 유용성뿐만 아니라 다양한 이해관계자의 우려를 관리하는 것 사이에서 균형을 맞추는 것이 필요하다.

이를 위해 데이터 프라이버시 리터러시 교육은 선택이 아닌 필수이다.

데이터 프라이버시 리터러시는 디지털 사회에서 개인의 데이터 보호와 보안을 유지하기 위한 필수적인 역량이며, 데이터의 가치와 위험성을 인식하고 대응할 수 있는 능력이라고 할 수 있다. 데이터 프라이버시 리터러시 교육을 통해 연구 참여자는 개인의 데이터가 수집되는 과정과 데이터 사용의 이면 및 잠재적 위험을 더 잘 파악할 수 있다. 데이터 프라이버시 리터러시 역량 강화로 개인이 자신의 데이터가 안정하게 보호된다는 확신을 갖는다면 디지털 환경에서의 안전과 신뢰가 높아질 수 있을 것이다.

[그림 6-1] 연구 참여자, 유전 연구자, 기관생명윤리심의위원회(IRB) 전문가 간의 개인식별 가능성 견해 차이 비교



출처: Goodman et al. (2018). A comparison of views regarding the use of de-identified data. p.116.

참고문헌

개인정보보호법, 법률 제19234호 (2024).

개인정보보호위원회. (2024. 2.). 가명정보 처리 가이드라인.

관계부처 합동. (2016. 6. 30.). 개인정보 비식별 조치 가이드라인 -비식별 조치 기준 및 지원·관리체계 안내-.

교육부, 개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인.

금융위원회, 금융감독원. (2022. 1.) 금융분야 가명·익명처리 안내서.

김지우, 권성훈, & 김동하. (2023). 심층 생성 모형을 이용한 재현 데이터 생성 방법론 리뷰 및 향후 연구 제언. 한국데이터정보과학회지, 34(5), 791-810.

김현태, 장가영. (2023). 데이터 가명·익명처리 기법의 현황과 대안: 재현데이터를 중심으로. 보험연구원 연구보고서. p.49.

박민정, 권성훈, 정재호. (2019). 차등정보보호 적용 실험 연구. 통계청 통계개발원.

박민정, 이용희, 권성훈. (2018). 차등정보보호에 관한 연구. 통계청 통계개발원.

박성준, 허혜옥, 심현보, 이기호. (2023). 보건복지 연구데이터의 통합적 관리 및 활용 방안 연구: 한국보건사회연구원 사례를 중심으로. 한국보건사회연구원.

박종서, 임지영, 김은정, 변수정, 이소영, 장인수, ..., 송지은. (2021). 2021년도 가족과 출산 조사-(구)전국 출산력 및 가족보건복지 실태조사. 한국보건사회연구원.

보건복지부, 개인정보보호위원회. (2024. 1.). 보건의료데이터 활용 가이드라인.

안성빈, 트랑, 도안, 이주희, 김지우, 김용재, ... & 박철우. (2023). 유용성과 노출 위험성 지표를 이용한 재현자료 기법 비교 연구. 응용통계연구, 36(2), 141-166.

이혜정, 이기호, 안수인, 임종호, 이상혁, 조용찬. (2022). 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구. 한국보건

사회연구원.

일본 내각부·문부과학성·후생노동성·경제산업성. (2024.4.). 의료분야의 연구개발에 기여하기 위한 익명 가공 의료 정보 및 가명 가공 의료 정보에 관한 법률에 대한 가이드라인(차세대의료기반법 가이드라인).
匿名加工医療情報及び仮名加工医療情報 に関する法律についてのガイドライン(次世代医療基盤法ガイドライン)

전진아, 이수빈, 박주현, 이한나, 최소영, 배은미, ..., 채수미. (2022). 사회정신건강연구센터 운영: 지역사회 거주 정신질환자의 건강증진 및 복지서비스 지원 방안. 한국보건사회연구원.

통계청. (2023. 6. 13.). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인 작성. 통계데이터분과위원회. 제2023-07호.

한국보건사회연구원 정보통계연구실. (2022.4.26). 제공용 조사자료 사전검토 가이드라인(내부자료). 한국보건사회연구원.

한국보건사회연구원. (2021). 가족과 출산조사 마이크로데이터. https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

한국보건사회연구원. (2021). 가족과 출산조사 코딩북. https://data.kihasa.re.kr/kihasa/kor/contents/ContentsList.html

한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 마이크로데이터[내부 자료].

한국보건사회연구원. (2022). 정신질환자의 건강 및 복지서비스 인식 및 이용 경험 조사 코딩북(내부 자료).

한국보건사회연구원. (2023). 한국복지패널조사 데이터. https://www.koweps.re.kr:442/data/data/list.do

한국보건사회연구원. (2023). 한국복지패널조사 코딩북. https://www.koweps.re.kr:442/data/book/list.do

한국보건사회연구원. (2024). 원내 조사자료 검토의견서[내부자료]. 한국보건사회연구원.

unite.ai. (2022.12.09.). 차등 프라이버시란 무엇입니까?, 2024.10.04.인출

https://www.unite.ai/ko/%EC%B0%A8%EB%93%B1-%ED%94%84%EB%9D%BC%EC%9D%B4%EB%B2%84%EC%8B%9C%EB%9E%80-%EB%AC%B4%EC%97%87%EC%9D%B8%EA%B0%80/

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. (2022). How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models, International Conference on Machine Learning, 290–306, PMLR.

Arjovsky Martin, Chintala Soumith, and Bottou Leon.(2017). Wasserstein generative adversarial networks, International conference on machine learning, 214-223, PMLR.

Benedetti, R. and Franconi, L. Statistical and technological solutions for controlled data dissemination. In: Pre-Proceedings of New Techniques and Technologies for Statistics, Sorrento, Italy, 1998, 225-570.

Bishop Christopher M, and Nasrabadi Nasser M.(2006). Pattern recognition and machine learning, Springer, 4(4).

Breiman, L. (2017). Classification and regression trees. Routledge.

Choi Edward, Biswal Siddharth, Malin Bradley, Duke Jon, Stewart Walter F, and Sun Jimeng.(2017). Generating multi-label discrete patient records using generative adversarial networks, Machine learning for healthcare conference, 286-305, PMLR.

Diederik P Kingma and Max Welling. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.

Duchi, J. C. Jordan, M. I. and Wainwright, M. J.(2018), Minimax optimal procedures for locally private estimation, Journal of the American Statistical Association, 113(521), 182-215.

Dwork C., McSherry F., Nissim K., Smith A. (2006). Calibrating Noise

to Sensitivity in Private Data Analysis, In: Halevi S., Rabin T. (eds) Theory of Cryptography, TCC 2006, Lecture Notes in Computer Science, vol 3876. Springer, Berlin, Heidelberg.

Dwork, C. (2006). Differential privacy, In 33rd International Colloquium on Automata, Languages and Programming, Part II (ICALP 2006). Springer, Venice, Italy, 1-12.

Engelmann Justin, and Lessmann Stefan.(2021). Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning, Expert Systems with Applications, 174, 114582, Elsevier.

Goodman, D., Johnson, C. O., Bowen, D., Smith, M., Wenzel, L., & Edwards, K. L. (2018). A comparison of views regarding the use of de-identified data. Translational Behavioral Medicine, 8(1), 113-118.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.(2014). Generative adversarial nets, Advances in Neural Information Processing Systems, 27.

Jared S Murray and Jerome P Reiter. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence, Journal of the American Statistical Association, 111(516): 1466-1479.

Jorg Drechsler and JP Reiter. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey, Journal of Official Statistics, 25(4): 589.

Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. (2018). General and specific utility measures for synthetic data, Journal of the Royal Statistical

Society: Series A (Statistics in Society), 181(3): 663-688.

Karr Alan F, Kohnen Christine N, Oganian Anna, Reiter Jerome P, and Sanil Ashish P.(2006). A framework for evaluating the utility of data altered to protect confidentiality, The American Statistician, 60(3): 224-232, Taylor & Francis.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. (2019). Modeling tabular data using conditional GAN, Advances in Neural Information Processing Systems, 32.

Lin Zinan, Khetan Ashish, Fanti Giulia, and Oh Sewoong.(2018). Pacgan: The power of two samples in generative adversarial networks, Advances in neural information processing systems, 31.

LIN, Jianhua. Divergence measures based on the Shannon entropy. IEEE Transactions on Information theory, 1991.

Multicenter Perioperative Outcomes Group(MPOG). (2022.12.22.). Standardized Data File – User Guide. https://mpog.org/downloads/. 2024. 10. 3. 인출

Multicenter Perioperative Outcomes Group(MPOG). (2024). https://mpog.org/memberhospitals/ 2024. 10. 3. 인출

Multicenter Perioperative Outcomes Group(MPOG). (2024). Research Proposal Process. 2024. 10. 3. 인출 https://mpog.org/write-a-research-proposal/

Multicenter Perioperative Outcomes Group(MPOG). (2024). Security Guidelines for Users. https://mpog.org/securityguidelinesusers/. 2024. 10. 3. 인출

Nowozin Sebastian, Cseke Botond, and Tomioka Ryota.(2016). f-gan: Training generative neural samplers using variational

divergence minimization, Advances in neural information processing systems, 29. on Machine Learning, 214–223, PMLR.

OpenAI. (2024). ChatGPT(October 9 version)[Large language model]. https://chat.openai.com

Park Noseong, Mohammadi Mahmoud, Gorde Kshitij, Jajodia Sushil, Park Hongkyu, and Kim Youngmin. (2018). Data synthesis based on generative adversarial networks, arXiv preprint arXiv: 1806.03384.

Paul R Rosenbaum and Donald B Rubin. (1983). The central role of the propensity score in observational studies for causal effects, Biometrika, 70(1): 41-55.

Portability, I., & Act, A. (2012). Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. Washington DC: Human Health Services.

Gulrajani Ishaan, Ahmed Faruk, Arjovsky Martin, Dumoulin Vincent, and Courville Aaron C.(2017). Improved training of wasserstein gans, Advances in neural information processing systems, 30.

Salimans Tim, Karpathy Andrej, Chen Xi, and Kingma Diederik. (2017). Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications, arXiv preprint arXiv: 1701.05517.

Shokri Reza, Stronati Marco, Song Congzheng, and Shmatikov Vitaly.(2017). Membership inference attacks against machine learning models, 2017 IEEE symposium on security and privacy(SP), 3-18, IEEE.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. Journal of

Official Statistics.

Snoke Joshua, Raab Gillian M, Nowok Beata, Dibben Chris, and Slavkovic Aleksandra. (2018). General and specific utility measures for synthetic data,

Solomon Kullback and Richard A Leibler. (1951). On information and sufficiency, The Annals of Mathematical Statistics, 22(1): 79–86.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely.

Sweeney, L., Yoo, J. S., Perovich, L., Boronow, K. E., Brown, P., & Brody, J. G. (2017). Re-identification Risks in HIPAA Safe Harbor Data: A study of data from one environmental health study. Technology science, 2017.

Sweeney, Latanya. k-anonymity: A model for protecting privacy. (2002). International journal of uncertainty, fuzziness and knowledge-based systems.

Templ, M. (2017). Statistical disclosure control for microdata. Cham: Springer.

Villani, C. (2009). Optimal transport: old and new (Vol. 338, p. 23). Berlin: springer.

Wasserman, L. and Zhou., S. (2010), A statistical framework for the differential privacy, Journal of the American Statistical Association, 105(489), 375-389.

Woo Mi-Ja, Reiter Jerome P, Oganian Anna, and Karr Alan F. (2009). Global measures of data utility for microdata masked for disclosure limitation, Journal of Privacy and Confidentiality, 1(1).

Zhao Zilong, Kunar Aditya, Birke Robert, and Chen Lydia Y. (2021). Ctab-gan: Effective table data synthesizing, Asian Conference on Machine Learning, 97-112, PMLR.

[부록] 익명처리 수준 정의표

〈부표〉 익명처리 수준 정의표

대분류	중분류	항목명	위험성 크기	익명처리 방안	기타
개인식별 정보		이름	8	삭제	-
		주민등록번호	사용 불가	삭제	-
		휴대폰 번호	8	삭제	-
개인식별 가능 정보		성별	6	현행 유지, 삭제	-
		연령	6	범주화, 삭제	5세(10세) 단위 범주화를 원칙으로 처리하고, 연구자의 판단에 따라 처리 방안이 변경될 수 있음
		지역	6	단계별 범주화, 삭제 ① 단계: 시도 ② 단계: 권역	시도 범주를 원칙으로 처리하고, 변수 조합에 의한 식별 가능성을 고려하여 권역으로 단계별 범주화함
		최종학력	6	현행 유지, 범주화, 삭제	-
		혼인상태	4	현행 유지, 범주화, 삭제	-
		가구소득	6	현행 유지, 범주화, 삭제	-
		경제활동상태	6	현행 유지, 범주화, 삭제	-
개인 식별 가능 정보	민감 정보 (연령)	생년	6	현행 유지, 범주화, 삭제	5세(10세) 단위 범주화를 원칙으로 처리하고, 연구자의 판단에 따라 처리

대분류	중분류	항목명	위험성 크기	익명처리 방안	기타
					방안이 변경될 수 있음
		생월	6	삭제	-
	민감 정보 (장애)	장애 유무	6	현행 유지, 삭제	-
		장애 종류	6	단계별 범주화, 삭제 ① 단계: 소분류 ② 단계: 중분류 ③ 단계: 대분류	보건복지부 장애 정도 판정 기준 소분류 범주화를 원칙으로 처리하고, 연구자의 판단에 따라 처리 방안이 변경될 수 있음
	민감 정보 (종교)	종교 유무	4	현행 유지, 삭제	-
		종교 종류	4	현행 유지, 삭제	-
	민감 정보 (가구 소득)	가구근로소득	6	범주화, 일반화(상하단 코딩), 삭제	이상값은 일반화(상하단 코딩)를 원칙으로 처리하고, 연구자의 판단에 따라 다른 방안으로 처리될 수 있음
		근로소득	6	범주화, 일반화(상하단 코딩), 삭제	이상값은 일반화(상하단 코딩)를 원칙으로 처리하고, 연구자의 판단에 따라 다른 방안으로 처리될 수 있음
	민감 정보 (직종)	직업	4	범주화, 삭제	한국표준직업분류(통계청 통계분류포털)를 기준으로 범주화 처리를 원칙으로 하고, 연구자의 판단에 따라 다른 방안으로 처리될 수 있음
		산업분류	4	범주화, 삭제	한국표준직업분류(통계청 통계분류포털)를 기준으로 범주화를 원칙으로 처

대분류	중분류	항목명	위험성 크기	익명처리 방안	기타
					리하고, 연구자의 판에 따라 다른 방안으로 처리될 수 있음
		종사자 수	2	범주화, 일반화(상하단 코딩), 삭제	이상값은 일반화(상하단 코딩)를 원칙으로 처리하고, 연구자의 판단에 따라 다른 방안으로 처리될 수 있음
		직종 경력 기간	2	범주화, 일반화(상하단 코딩), 삭제	이상값은 일반화(상하단 코딩)를 원칙으로 처리하고, 연구자의 판단에 따라 다른 방안으로 처리될 수 있음
		현 소속기관 근무 기간	2	범주화, 일반화(상하단 코딩), 삭제	극단값(이상치)은 일반화(상하단 코딩)를 원칙으로 처리하고, 연구자의 판단에 따라 다른 방안으로 처리될 수 있음
	민감 정보 (가구 구성)				

출처: 교육부·개인정보보호위원회. (2022. 7.). 교육 분야 가명·익명정보 처리 가이드라인을 참고하여 연구진이 원내 상황에 맞게 주로 검토하고 있는 항목에 대해 재구성하였음.

Abstract

A Study on the Development of a System for Creating and Managing De-identified Data in the Health and Welfare Sector

Project Head: Oh, Miae

As the importance of data analysis and use increases, the need for personal information protection is also being emphasized. With the growing risk of personal information exposure, there is a rising demand for developing methods to protect data while effectively utilizing it through de-identification methodologies.

De-identification methodology refers to a process that examines the environment in which personal information within the data is used and assesses the risk of re-identification. It involves taking appropriate measures through various methods to ensure that identification does not occur.

The distinguishing feature of de-identification methods is that their performance is highly sensitive to the data domain and the intended use of the data, making it challenging to adopt a generalized approach. Consequently, no specific technique can yet claim a relative superiority in terms of performance. Despite the inherent challenges in generalizing de-identification methods, generating de-identified data remains crucial to striking a balance between minimizing the risk

Co-Researchers: Park, Seongjun · An, Suin · Kwon, Sunghoon · Song, Jieun · Kim, Hyunkyu

of personal information exposure and maximizing data utility.

This study aims to enhance the utilization of data in the healthcare and welfare sectors by reviewing de-identification methods at the level of anonymization, comparing re-identification risks, and establishing systems for generating and managing de-identified data. The goal is to provide a foundation for the safe utilization and dissemination of data. By ensuring data security and providing a consistent approach to personal information protection, it is possible not only to comply with data protection regulations but also to build societal trust in the processes of data analysis and utilization.

Key words : De-identified data, De-identification techniques, Risk of privacy exposure, Health and welfare