www.khidi.or.kr, www.khiss.go.kr

KHIDI Brief

발행일_ 2018. 8. 13 발행처_ 한국보건산업진흥원 발행인_ 이영찬

보건산업브리프 Vol. 268

통계분석 | 정책제도 | 동향전망

Khidi Issue Paper

의료데이터 활용을 위한 개인정보 비식별화 기술 및 프로그램 동향

4차보건산업추진단 빅데이터팀
김재한

Contents

Ⅰ 의료데이터 활용을 위한 개인정보 비식별화 조치 기술의 필요성

■ 의료데이터 양의 폭발적인 증가는 의료 서비스의 비용절감 등 사회 시스템의 혁신을 불러일으킬 수 있는 영역으로 평가

⊙ Dell EMC(2014)는 의료데이터가 2013년 153EB에서 2020년 2,314EB로 증가할 것으로 전망
※ 1EB(엑사바이트)는 1,000,000TB(테라바이트)에 해당하는 데이터 단위

⊙ McKinsey(2013)는 미국 보건 의료부문의 빅데이터 활용의 용이성이 높고 경제에서 차지하는 비중이 커질 것으로 기대
- 미국 보건의료 부문에서만 연간 최대 1900억 달러의 비용절감을 실현시킬 수 있다고 전망

⊙ 국내 의료데이터의 활용으로 관련 산업의 성장을 활성화 시키기 위해 정부 산하기관은 보유하고 있는 의료데이터를 활용할 수 있도록 정책 지원 업무를 하고 있음
- 건강보험공단에서는 NHISS 포털을 통해 건강보험통계, 관심질병통계 등 다양한 통계데이터를 제공
- 건강보험심사평가원은 요양기관, 의약품, 진료정보 등의 데이터 셋과, 주요 의료통계, 질병/행위별 의료통계 등을 제공

⊙ 의료 데이터의 분석과 활용은 의료서비스의 품질개선, 개인별 질병진단 및 치료서비스 향상 등 의료 시스템 전반에 걸쳐 혁신을 유도할 수 있을 것으로 예측

■ 의료 데이터 활용에 대한 전망과 기대가 높은 반면, 개인의 민감한 정보를 많이 담고 있어, 개인정보 보호의 기술적 대책이 필요

⊙ 의료데이터의 활용으로 개인정보의 경제적 가치가 부각되면서 안전한 개인정보 보호의 수단으로 비식별화 기술이 주목 받고 있음
- 호주 개인정보 보호위원회(The Australian Privacy Comissioner)는 비식별화 기술을 로켓 사이언스(Rocket science)에 비유하며, 개인정보 활용과 보호의 균형적 조화를 해결할 수단이 될 수 있다고 언급

⊙ 미국 ITRC(Identify Theft Resource Center)는 정보유출이 보건의료 분야에서 매년 높은 비율로 증가하고 있다고 발표하고 데이터 보호에 대한 문제해결이 시급하다고 언급
- 보건의료 분야의 의료 데이터 유출과 개인정보에 대한 보안을 위해 HCIC(Healthcare Industry Cyber-security) 태스크포스 팀*을 구성
  * HCIC(Healthcare Industry Cyber-security) 태스크포스 팀 : 2016년 창설된 TF팀은 사이버보안법(The Cybersecutiry Act of 2015)의 하위로 활동하며 2017년 사이버보안을 발전시키기 위한 전략방안을 보고

■ 의료데이터의 활용은 정보유출의 위험성뿐만 아니라 재식별의 위험성이 부각되며, 이와 관련하여 비식별 기술 개발이 주목받고 있음

〈표 1〉 비식별 조치 데이터의 재식별 사례, 한국정보화진흥원(2014) 재구성

	비식별 조치 데이터	재식별 시도결과
메사추세츠 주 사례(1997, 미국)	메사추세츠주의 보험위원회가 주정부 소속 공무원의 병원 출입 기록을 비식별 처리하여 공개	투표자 명부 데이터와 결합하여 해당 공무원의 신원을 식별
아메리카 온라인 사례(2006, 미국)	65만명의 사용자가 3개월간 검색한 검색기록 리스트 2천만건을 비식별 처리하여 공개	검색 이력으로부터 특정 이용자에 대한 이름과 지역의 재식별에 성공
넷플릭스 사례(2006, 미국)	넷플릭스는 50만명의 이용자들이 영화에 대한 평점을 내린 1억건의 시청이력 데이터를 비식별 처리하여 공개	넷플릭스 데이터와 영화정보 사이트 IMDb 데이터를 결합하여 일부개인 식별에 성공
SNS 재식별 사례(2013, ETRI)	페이스북 667만개, 트위터 227만개의 한국인 계정에 업로드한 데이터 대해 재식별 가능성 분석	비식별 정보로 생각된 정보로 개인식별의 가능성 3% 이상, 제3의 데이터 결합을 통한 개인식별 가능성이 최대 45%로 분석

■ 의료데이터 기반의 보건산업 발전과 성장을 위해 데이터를 보호하고 이를 안전하게 활용할 수 있도록, 비식별화 조치에 대한 기술 개발이 요구되는 상황

■ 본고에서는 의료데이터의 활용을 위한 최근 국내외 비식별화 기술 동향을 조사하고, 국내의 비식별화 기술에 대한 정책적인 시사점을 도출하고자 함

Ⅱ 국내의 개인정보 비식별화 기술 및 솔루션 프로그램 동향

■ 범부처 합동으로 '개인정보 비식별 조치 가이드라인(2016)'을 발간하고 이를 통해 비식별화 조치 기법과 적정성 평가 모델을 제시

⊙ 개인 식별정보*는 원칙적으로 삭제 하되, 데이터 목적상 필요한 부분에 대해서만 비식별화 조치를 권고

* 개인 또는 개인과 관련한 사물에 고유하게 부여된 값 또는 이름

- 일반적 비식별 조치 기법으로 가명처리, 총계처리, 데이터 삭제 등 5가지 기법과 17가지 세부기술을 제시

〈표 2〉일반적 비식별 조치 기법, 범부처 합동(2016)

기법	세부기술
가명처리 (Pseudonymization)	① 휴리스틱 가명화(Heuristic Pseudonymization) ② 암호화(Encryption) ③ 교환방법(Swapping)
총계처리 (Aggregation)	④ 총계처리(Aggregation) ⑤ 부분총계(Micro Aggregation) ⑥ 라운딩(Rounding) ⑦ 재배열(Rearrangement)
데이터 삭제 (Data Reduction)	⑧ 식별자 삭제 ⑨ 식별자 부분삭제 ⑩ 레코드 삭제(Reducing Records) ⑪ 식별요소 전부 삭제
데이터 범주화 (Data Suppression)	⑫ 감추기 ⑬ 랜덤 라운딩(Random Rounding) ⑭ 범위 방법(Data Range) ⑮ 제어 라운딩(Controlled Rounding)
데이터 마스킹 (Data Masking)	⑯ 임의 잡음 추가(Adding Random Noise) ⑰ 공백(blank)과 대체(impute)

⊙ 비식별 조치된 데이터의 재식별 가능성을 낮추기 위해 비식별화 조치가 적정하게 이루어졌는지 파악하는 적정성 평가에 관한 모델을 제시

- 개인정보 비식별 조치 가이드라인에서는 적정성 평가 모델로 k-익명성 모델을 최소한의 평가수단*으로 제시

  * k-익명성 모델을 최소한의 평가수단으로 k-익명성 '3'이 되도록 권고함

- 필요시 추가적인 평가모델로 l-다양성 모델과 t-근접성 모델 활용을 제안

■ 비식별 조치 적정성 평가 모델인 k-익명성, l-다양성, t-근접성 모델 제안

⊙ k-익명성(k-anonymity) 모델

- 전체 데이터에서 동일한 속성 값을 갖는 레코드를 'k'개 이상으로 유지하여 식별 확률을 1/k로 낮추는 모델
- 'k'값이 증가한다는 것은 동일한 속성을 갖는 레코드의 개수가 증가한다는 의미
- 동질성 공격과 배경지식에 의한 공격에 취약하며, 이를 보완하기 위해 l-다양성 모델이 등장

⊙ l-다양성(l-diversity) 모델

- 데이터의 민감한 속성에 대해 각 레코드별로 'l'개 이상의 서로 다른 값을 가질 수 있도록 하는 모델
- 'l'값이 증가한다는 것은 전체 집합에서 민감 속성의 속성 값이 다양해진다는 의미
- 쏠림 공격과 유사성 공격에 취약하며, 이를 보완하기 위해 t-근접성 모델이 제시

⊙ t-근접성(t-closeness) 모델

- 특정 데이터 집합의 분포와 전체 데이터 집합의 분포가 't'이하의 차이를 보일 수 있도록 하는 모델
- 't'값은 0~1의 범위를 갖으며, 0에 가까울수록 특정 데이터의 분포와 전체 데이터 분포의 유사성이 강해진다는 의미

※ 가이드라인에서는 k-익명성 모델을 최소한의 적정성 평가모델로 권고하였으며, 필요시 l-다양성, t-근접성 모델을 활용하도록 제시

◉ k-익명성, l-다양성, t-근접성 모델을 활용한 비식별 조치 적용예시

k-익명성 모델

·데이터의 연결공격 취약점 개선
·같은 값이 적어도 k개 이상 존재

구분	지역	연령	성별	질병
1	13053	28	M	☆
2	13068	21	M	☆
3	13068	29	F	○
4	13053	23	M	○
5	14853	50	F	◇
6	14853	47	M	☆
…	…	…	…	…

↓

구분	지역	연령	성별	질병
1	130**	〈30	*	☆
2	130**	〈30	*	☆
3	130**	〈30	*	○
4	130**	〈30	*	○
5	1485*	〉40	*	◇
6	1485*	〉40	*	☆
…	…	…	…	…

l-다양성 모델

·k-익명성 모델의 취약점 개선
·동질성 및 배경지식 공격 방어

구분	지역	연령	성별	질병
1	130**	〈30	*	☆
2	130**	〈30	*	☆
3	130**	〈30	*	○
4	130**	〈30	*	○
5	1485*	〉40	*	◇
6	1485*	〉40	*	☆
…	…	…	…	…

↓

구분	지역	연령	성별	질병
1	1305*	≤40	*	☆
4	1305*	≤40	*	○
5	1485*	〉40	*	◇
6	1485*	〉40	*	☆
2	1306*	≤40	*	☆
3	1306*	≤40	*	○
…	…	…	…	…

1) k-익명성 모델의 취약점

- 동질성 공격(homogeneity attack)에 취약
  ☞ 레코드가 범주화 되었다 하더라도 일부 정보들이 같은 값을 가질 수 있기 때문에 동일한 정보를 이용하여 공격 대상의 정보를 추론
- 배경지식에 의한 공격(background knowledge attack) 취약
  ☞ 주어진 데이터 이외의 배경지식을 통해 공격 대상의 민감한 정보를 추론

2) l-다양성 모델의 취약점

- 쏠림공격(skewness attack)에 취약
  ☞ 정보가 특정한 값에 쏠려 있을 경우 l-다양성 모델이 프라이버시를 보호하지 못함
- 유사성 공격(similarity attack)에 취약
  ☞ 비식별 조치된 레코드의 정보가 서로 비슷하다면 l-다양성 모델을 통해 비식별 된다 할지라도 프라이버시가 노출될 수 있음

〈표 3〉 K,L,T 모델을 활용한 비식별 조치 방법 예시, 개인정보 비식별 조치 가이드라인(2016) 재가공

t-근접성 모델

·l-다양성 모델의 취약점(쏠림공격, 유사성 공격) 개선

· 유사성 공격에 취약한 사례

구분	속성		민감한 정보	
	지역	연령	급여	질병
1	476**	2*	30	위궤양
2	476**	2*	40	급성위염
3	476**	2*	50	만성위염
4	4790*	≥40	60	급성위염
5	4790*	≥40	110	감기
6	4790*	≥40	80	기관지염
…	…	…	…	…

·t-근접성 모델에 의해 비식별 조치된 사례

구분	속성		민감한 정보	
	지역	연령	급여	질병
1	4767*	≤40	30	위궤양
3	4767*	≤40	50	만성위염
4	4790*	≥40	60	급성위염
5	4790*	≥40	110	감기
6	4790*	≥40	80	기관지염
2	4760*	3*	40	급성위염
…	…	…	…	…

3) t-근접성 모델

- t수치가 0에 가까울수록 전체 데이터의 분포와 특정 데이터 구간의 분포 유사성이 강해지기 때문에 그 익명성의 방어가 더 강해지는 경향

☞ 익명성 강화를 위해 특정 데이터들을 재배치해도 전체 속성자들의 값 자체에는 변화가 없기 때문에 일반적인 경우에 정보 손실의 문제는 크지 않음

■ 민간에서는 데이터의 안전한 활용을 위하여 다양한 기법의 비식별 기술 솔루션 프로그램을 개발

⊙ (IDentity SHIELD) 비식별 조치부터 데이터 결합까지 업무 연계 프로세서를 제공하는 국내 비식별화 솔루션 프로그램

- 사전검토를 위한 정형·비정형·반정형 빅데이터에 포함된 개인정보의 실시간 탐지 기능이 있으며, k-익명성을 만족하기 위한 k·l·t 방식의 비식별화 기능을 가짐
- 데이터 전처리, 개인정보 자동탐지, 안전한 익명성 비식별화 및 연계를 위한 기술을 가지고, Hadoop시스템*에서도 비식별 처리 가능한 기술 개발

* Hadoop : 데이터를 분산된 환경에서도 저장하고 처리할 수 있는 자바 기반의 오픈 소스 프레임 워크

⊙ (Analytic DID) 의사 결정에 도움을 줄 수 있도록 빅데이터 분석의 효용성과 위험에 대한 다양한 지표를 가시화하여 제공함

- 개인정보보호법, HIPPA 등 국내외 관련 법률에 맞게 지원하여 안전하고 신뢰할 수 있는 사용 환경을 구축
- 정책에 따른 사용 관리 및 익명화 위임 등 효율적인 업무 프로세스 지원

◉ (DataEye PIDI) 개인정보 비식별 조치 가이드라인을 준수하여 식별자 암호화 등 17가지 비식별 조치 기술을 지원하며, k·l·t 비식별 조치 적정성 평가를 수행할 수 있는 개인정보 비식별 조치 솔루션

- Incognito 알고리즘*을 적용한 k·l·t의 적정성 평가 모델을 지원하는 국내 비식별 소프트웨어
- 데이터 변환 및 이관처리를 위한 ETL 솔루션인 PDI(Penta Data Integrator) 내장

* Incognito 알고리즘 : k-익명성 모델을 만족시키면서 정보 손실을 최소로 하는 최적의 해를 찾아내는 알고리즘

◉ 국내 A대학은 익명화 데이터의 유용성 향상 기술을 위한 'h-ceiling(h-상한)' 기법 연구

- 레코드 삽입을 통한 유용성 향상 기법으로, 위조 레코드를 삽입하여 개인정보를 보호
- 삽입된 위조 레코드 정보를 별도로 관리하면서, 사용자가 원하는 프라이버시 수준과 데이터 유용성 수준을 설정할 수 있는 것이 특징

◉ 국내 B사에서는 '다수준 추상화&동기화(MAS)기법'을 통하여 비식별 조치 기술 개발

- (다수준, Multi-level) 데이터 보안 및 활용 목적에 맞추어 수준을 조정
- (추상화, Abstract) 총계처리(그룹핑)와 차분 프라이버시* 기반의 비식별화 기법

* 차분 프라이버시 : 레코드 자체의 확률적 변형과 정확한 통계적 노이즈를 추가 함으로써 식별가능성을 제한하는 접근법

- (동기화, Synchronization) 추후 연계 분석이 가능하도록 설계

■ 다양한 비식별 프로그램은 의료 데이터의 재식별 위험을 방지하고, 개인정보를 보호하는 수단으로 활용될 것으로 판단

◉ 국내의 비식별 솔루션 프로그램은 '비식별 조치 가이드라인(2016)'에서 언급한 k·l·t 적정성 모델을 기준으로 기술 개발 진행

◉ 개인정보의 철저한 비식별 처리와 추후 재식별 방지를 위한 지속적인 모니터링 등의 사후관리에 필요한 기술적 발전 기대

■ 비식별 기술을 중점적으로 다루고 있는 국내 비식별 처리 기술 전문가들도 표준개발상황을 공유하고 협력할 예정

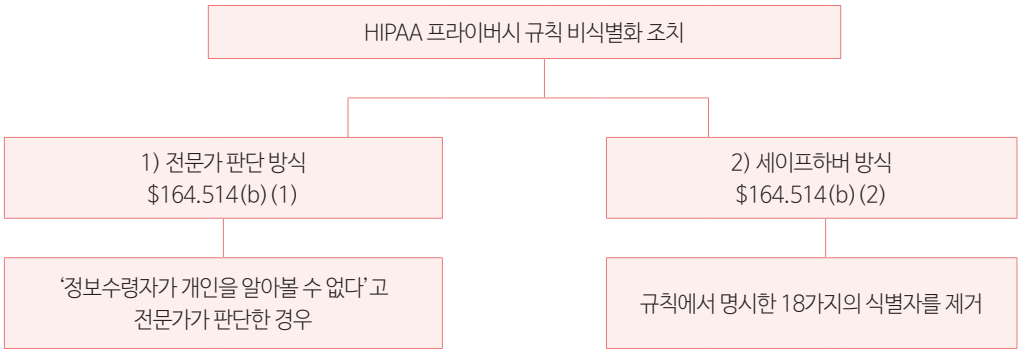
Ⅲ 해외의 개인정보 비식별화 기술 및 솔루션 프로그램 동향

■ 해외에서는 비식별화 기술과 관련한 법제를 마련하고 있으며, 미국에서는 HIPAA 프라이버시 규칙*에 의한 비식별화 기술 조치를 언급

◉ HIPAA 프라이버시 규칙에서 언급하는 두 가지 비식별화 기술 방식에 의해 의료데이터가 비식별화 되었음을 판단

* HIPAA 프라이버시 규칙(Privacy Rule): 보건의료정보의 사용과 공개에 대한 표준을 제공하여 일반인의 이해를 도우며, 개인식별이 가능한 보건의료정보를 보호하고 비식별 처리된 보건의료정보는 자유롭게 공개·사용 가능

〈표 4〉HIPAA 프라이버시 규칙의 두 가지 비식별화 조치 방법



- 전문가 판단 방식(Expert determination method): 적절한 지식과 경험을 갖춘 전문가가 정보주체의 식별 위험성이 매우 낮다고 결정할 경우 비식별화 된 것으로 인정하는 방식
- 세이프하버 방식(Safe harbor method): 18가지 식별속성을 제거된 상태. 단, 의료인 등이 잔존 정보만으로 특정 개인을 인식할 수 없는 경우

〈표 5〉HIPAA 프라이버시 규칙의 비식별화 조치 기법 세이프하버 18가지 항목

세이프하버 방식에서 정하는 18가지 식별자 및 준식별자 항목	
(A) 이름	(I) 건강보험 등록번호
(B) 주(state)보다 작은 지리적 단위(시, 군, 구역), 우편번호 및 이와 상응하는 지역코드	(J) 계좌번호
(C) 개인과 직접적으로 연관되는 모든 날짜로 생일, 입원일, 퇴원일, 사망일, 89세 이상 모든 연령과 그 연령을 나타내는 모든 일자	(K) 수료증 혹은 자격증 번호
(D) 전화번호	(L) 자동차 번호판을 포함한 차량번호
(E) 팩스번호	(M) 기기 아이디 및 시리얼 넘버
(F) 이메일 주소	(N) 웹페이지 주소(URLs)
(G) 사회보장번호	(O) 인터넷 IP주소
(H) 의무기록번호 / 환자번호	(P) 신체특성 지표자, 지문 및 음성포함
	(Q) 얼굴 정면 사진 및 이에 상응하는 사진
	(R) 기타 모든 고유한 숫자, 문자, 코드

◉ 미국 NCVHS(National Committee on Vital and Health Statistics)*는 HIPAA 개인정보보호규칙에 따른 비식별화 표준 관행에 대한 개선 방안을 권고

- NCVHS에서는 비식별화된 데이터 사용에 대하여 안전하게 사용할 수 있도록, 비식별화 과정에 중점을 둘 것을 촉구함

* NCVHS (National Committee on Vital and Health Statistics): HSS 장관에게 건강데이터 통계, 국민건강정보 정책 등에 대한 자문을 담당하는 기관

■ 해외 각 국에서는 비식별화 기술 개발을 통해 다양한 비식별화 솔루션 프로그램을 보유하고 공개 소프트웨어로 보급

◉ 국제 프라이버시 전문가 협회(IAPP)는 「2018년 개인정보보호 기술 업체 보고서(2018 Privacy Tech Vendor Report」 비식별 기술 제공 업체 발표

〈표 6〉 2018년 개인정보보호 기술 업체 보고서 재가공, 비식별 기술 제공 업체(2018)

업체명	개요 및 특징
Arcad	·Datachanger 소프트웨어 제품으로 데이터 구성관리를 지원
Immuta	·모든 기업 소스로부터 데이터를 연결하고 마스킹, 익명화, 차등 프라이버시, 목적 기반 제한 등 다양한 정책을 적용
Protigrity	·데이터 중심적인 암호화, 무형식 토큰화 및 마스킹을 활용
SAS Global Data Management	·EU GDPR을 준수하여 데이터에 접근/식별/관리/보호 할 수 있도록 설계된 플랫폼 제공

- 비식별 기술 업체를 공식적으로 발표하여 이슈가 되고 있는 기술에 대해 공유하며, 기술 개발을 위한 자발적인 노력 시도
- 또한, 자체적으로 공개 소프트웨어 프로그램을 개발 및 보급하여 활발하게 사용하며, 기술 개발에 대한 피드백을 공유

〈표 7〉비식별 소프트웨어 현황(공개.상용) 재가공(2016), 한국인터넷진흥원

구분	비식별 소프트웨어	특징
비식별 공개S/W	ARX Data Anonymization Tool	Java 기반의 비식별 툴
	UDT Anonymization Toolbox	앱과 라이브러리 형태로 공개
	Cornell Anonymization Toolkit(CAT)	다양한 공격자 모델을 대응할 수 있게 한 대화형 디자인
	Open Anonymizer	k-익명화 개념을 기반, 데이터 레코드를 일반화
	sdcMicro	R 언어 기반의 비식별 프로그램

◉ 잘 알려진 k·l·t 모델뿐만 아니라 다양한 프라이버시 모델을 기반으로 비식별 작업을 수행할 수 있으며,

◉ 재식별 위험성을 분석하거나, 사용자가 사용하는 프라이버시 모델에 적당한 파라미터를 추천하는 등 다양한 기능을 제공하는 공개 비식별 소프트웨어를 개발하고 보급

■ 보건의료 분야에서 데이터 활용에 대한 자기결정권 및 개인정보보호 요구사항을 충족시키기 위해 암호화 기술(Encryption)강조

⊙ 프랑스의 Thales의 e-security는 보메트릭 데이터 시큐리티(Vormetric Data Security) 라는 DB암호화 기술로 데이터의 비식별 기술 제공

- 데이터 암호화와 키 관리, 접근 제어, 권한 및 역할 관리 등 파일 단위로 암호화를 수행해 비정형 데이터까지 보호할 수 있는 것이 특징
- 기존 시스템이나 API를 변경하지 않고 투명하게 암호화 기술을 적용할 수 있어 시간을 줄이고 비용 감소

■ 국제표준 단체인 ISO에서 'ISO/IEC 20889*'를 통해 재식별 및 관련 기술용어들의 표준 제정을 추진

* ISO/IEC 20889 : Privacy enhancing data de-identification techniques

⊙ ISO/IEC 20889는 11월 중 공개(Publication)를 앞두고 있으며, 본 자료에서는 DIS(Draft International Standard) 버전을 기준으로 작성

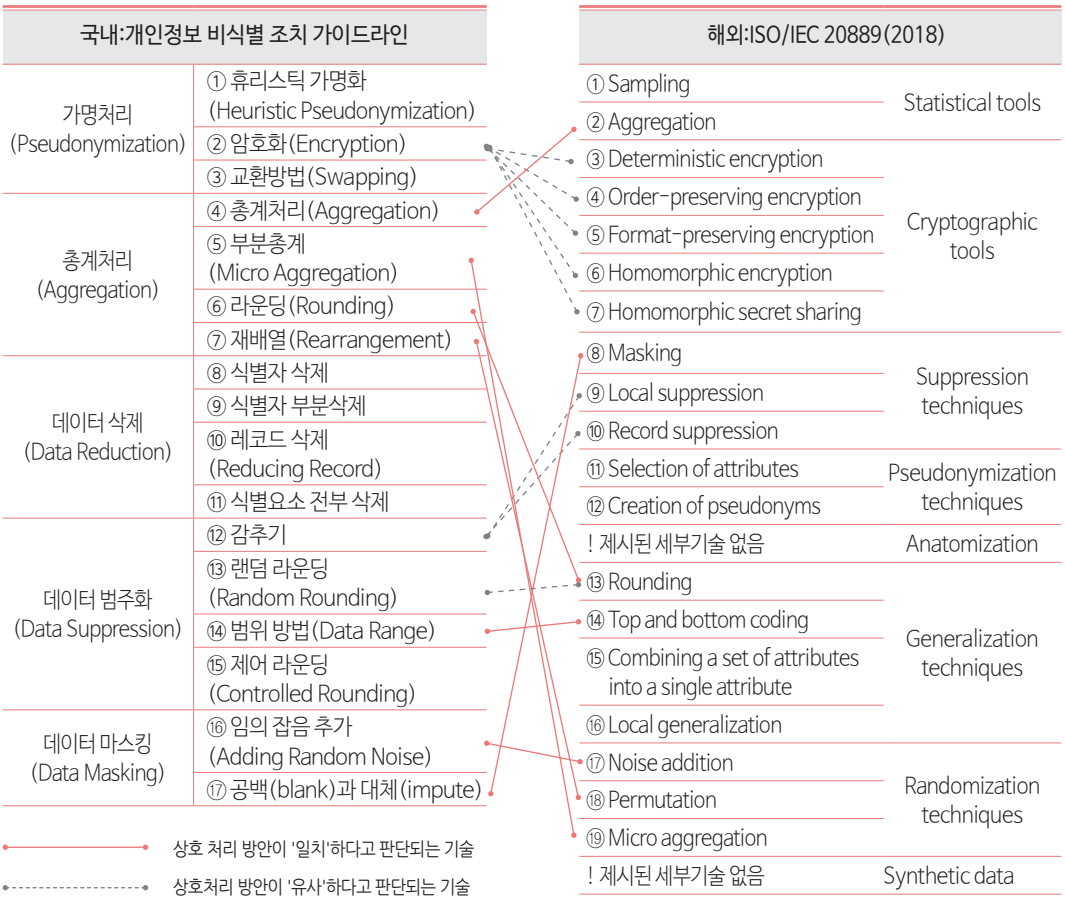
〈표 8〉 ISO/IEC DIS 20889에서 작성한 비식별화 기법(De-identification techniques), ISO(2018)

기법	세부기술
Statistical tools	① Sampling ② Aggregation
Cryptographic tools	③ Deterministic encryption ④ Order-preserving encryption ⑤ Format-preserving encryption ⑥ Homomorphic encryption ⑦ Homomorphic secret sharing
Suppression techniques	⑧ Masking ⑨ Local suppression ⑩ Record suppression
Pseudonymization techniques	⑪ Selection of attributes ⑫ Creation of pseudonyms
Anatomization	! 제시된 세부기술 없음
Generalization techniques	⑬ Rounding ⑭ Top and bottom coding ⑮ Combining a set of attributes into a single attribute ⑯ Local generalization
Randomization techniques	⑰ Noise addition ⑱ Permutation ⑲ Micro aggregation
Synthetic data	! 제시된 세부기술 없음

⊙ 국내의 '비식별 조치 가이드라인(2016)'에서 기술되지 않았던 비식별 조치 기법이 추가

- 특히, ISO/IEC 20889에서 언급된 동형암호는 암호화된 상태 그대로 정보를 사용하며, 제3자에게 개인정보를 노출하지 않고 정보를 활용할 수 있어 최근 비식별 기술로 주목을 받고 있음

〈표 9〉 국내 가이드라인과 해외 ISO/IEC 20889의 세부기술 매칭 테이블



구분	국내 세부기술	해외 세부기술
일치	④ 총계처리(Aggregation)	② Aggregation
	⑤ 부분총계(Micro Aggregation)	⑲ Micro aggregation
	⑥ 라운딩(Rounding)	⑬ Rounding
	⑦ 재배열(Rearrangement)	⑱ Permutation
	⑭ 범위 방법(Data Range)	⑭ Top and bottom coding
	⑯ 임의 잡음 추가 (Adding Random Noise)	⑰ Noise addition
	⑰ 공백(blank)과 대체(impute)	⑧ Masking
유사	② 암호화(Encryption)	③ Deterministic encryption ④ Order-preserving encryption ⑤ Format-preserving encryption ⑥ Homomorphic encryption ➜ 동형암호 ⑦ Homomorphic secret sharing
	⑫ 감추기	⑨ Local suppression ⑩ Record suppression
	⑬ 랜덤 라운딩(Random Rounding)	⑬ Rounding

- 2011년 MIT Tech Review에 '10emerging technologies'중 하나로 소개되었으나, 암호문의 확장과 연산속도 단위의 상승으로 인해 실제 활용이 어렵다고 판단되어 연구가 중단
- 2012년 Coron등의 연구에서 확장연산에 대한 최적화 알고리즘이 제시되고, 최근 2017년 Microsoft에서 'TFHE'공용 오픈 소스 라이브러리를 개발하는 등 동형암호 기술 개발이 활발하게 진행 중

■ **해외 각 국에서 개인정보 보호에 대한 상황을 인식하고 연구 목적 등의 2차 활용을 위한 비식별 기술을 연구하고 그와 관련한 가이드를 제시**

■ **개인정보 보호의 중요성으로 비식별 조치 기술 개발에 대한 공유가 자연스럽게 이루어지고, 관심이 상당히 높아졌으며, 이와 관련한 기술개발도 활발하게 진행 중**

Ⅳ 국내와 해외의 비식별화 기술 및 솔루션 프로그램 동향 비교

■ **법적 제도를 준수하여 비식별화 기술을 개발하고, 이를 바탕으로 국내 및 해외 업체에서 비식별화 솔루션 프로그램을 제공**

⦿ 국내에서는 비식별 솔루션 프로그램들은 국내외 관련 법률에 맞게 프로그램을 지원하여, 안전하고 신뢰할 수 있는 사용 환경을 구축
- '비식별 조치 가이드라인(2016)'의 k·l·t의 적정성 평가 모델을 지원하는 국내 비식별 소프트웨어를 보급하고 있음

⦿ 해외의 'SAS Global Data Management'등은 EU GDPR을 준수하며 데이터의 접근/식별/관리 등을 할 수 있게 설계된 플랫폼을 제공하고 있음

■ **비식별 조치 기술 개발에 대하여 자연스럽게 공유하고, 이에 따른 관심이 상당히 높아졌으며, 국내외에서는 이와 관련한 기술개발도 활발하게 진행 중**

⦿ 국내외 비식별 기술 전문가들은 비식별 처리 기술의 표준개발상황을 공유하고 지속적으로 협력해 나가고 있음

⦿ 특히, 해외에서는 비식별화 솔루션 프로그램 개발뿐만 아니라 데이터 자체를 암호화하고 분석할 수 있는 기술을 개발하고, 강조하며 개인정보유출을 최소화하려고 노력하고 있음

- 프랑스의 Thales의 e-security는 보메트릭 데이터 시큐리티(Vormetric Data Security) 라는 DB암호화 기술로 데이터의 비식별 기술 제공
  - 보건의료 분야에서 데이터 활용에 대한 자기결정권 및 개인정보보호 요구사항을 충족시키기 위해 암호화 기술(Encryption)을 강조하고 있음
- 또한, ISO/IEC 20889에서 언급된 동형암호는 제3자에게 개인정보를 노출하지 않고 정보를 활용할 수 있어 최근 비식별 기술로 주목받고 있음

〈표 10〉 국내 가이드라인과 해외 ISO/IEC 20889의 세부기술 매칭 테이블(재가공)

국내 세부기술(비식별 조치 가이드라인)	해외 세부기술(ISO/IEC20889)
① 총계처리(Aggregation)	① Aggregation
② 부분총계(Micro Aggregation)	② Micro aggregation
③ 라운딩(Rounding)	③ Rounding
④ 재배열(Rearrangement)	④ Permutation
⑤ 범위 방법(Data Range)	⑤ Top and bottom coding
⑥ 임의 잡음 추가 (Adding Random Noise)	⑥ Noise addition
⑦ 공백(blank)과 대체(impute)	⑦ Masking
⑧ 암호화(Encryption)	⑧ Deterministic encryption 결정적 암호화 ⑨ Order-preserving encryption 순위 보존 암호화 ⑩ Format-preserving encryption 형태 보존 암호화 ⑪ Homomorphic encryption ➡ 동형 암호화 ⑫ Homomorphic secret sharing 동형 비밀분산
⑨ 감추기	⑬ Local suppression ⑭ Record suppression
⑩ 랜덤 라운딩(Random Rounding)	⑮ Rounding

V 결론 및 시사점

■ 의료 데이터 활용 가능성은 매우 높지만 개인정보유출 사례와 데이터의 재식별화에 대한 사회적 우려가 의료 데이터 활용에 저해요인으로 작용

◉ 국내의 경우 전 국민 건강보험을 통해 수집된 방대한 의료 데이터베이스를 가지고 있으나 개인정보보호 등의 문제로 데이터 활용 제한

◉ 방대한 의료 데이터의 안전한 활용을 위해 비식별화 기술에 대한 관심이 증가하는 추세이며, 이에 대한 범정부적인 노력 필요

- 의료 데이터를 활용할 수 있는 현행 법·제도의 제정하에 비식별 기술을 개발할 수 있도록 공공과 민간의 노력이 필요

■ 의료 데이터의 안전한 활용을 위해 국내 비식별 기술 개발의 저변 확산 노력과 R&D사업 지원 필요

◉ 정형, 반정형, 비정형 등의 데이터 형태와 데이터 활용 목적 등에 맞춘 사례 중심 기술 세미나를 지속하여 개최

- 일본의 'PWS Cup'*을 벤치마킹 하여 비식별 기술의 개발을 유도하고, 재식별을 시도하여 비식별 기술의 유용성과 성능검사를 시험해 볼 수 있는 기술 콘테스트 개최 등의 방안 제시

  * 일본의 PWS Cup: 일본 메이지 대학이 주관하고 개인정보위원회 등의 후원으로 데이터 활용의 기술자·전문가 간 교류와 개인정보보호 기술의 연구개발 활성화를 위해 2015년부터 진행해 온 대회

- 비식별 조치의 다양한 기술발전을 위한 콘테스트 개최 등 민·관의 노력으로 다양한 비식별 기술의 국내 저변을 확산하기 위해 노력하고 있음

◉ 의료 데이터의 활용을 저해하고 있는 데이터 유출 및 재식별화 등 현행 법·제도 하에서 동작할 수 있는 개인정보보호 기술 지원 필요

- 복지부는 '의료 데이터 보호 및 활용에 대한 R&D사업'을 지원할 예정이며, 기술적인 부분을 해결함으로써 공공적 목적의 연구지원 및 의료 빅데이터 기반의 연구 활성화 촉진을 기대
- R&D지원을 통해 공공의 목적을 확보하고 보건의료 분야의 기술성과를 달성함으로써 보건의료의 데이터 활용에 윤활제 역할을 수행할 것으로 예상

■ 의료 데이터를 이용한 보건의료 산업 정책과 개인정보 유출에 대한 우려가 대립되는 상황에서, 의료 데이터를 안전하게 활용할 수 있는 비식별화 기술 개발 지원과 관심은 현재 당면한 상황의 절충점을 찾는 중요한 키로 작용할 것

◉ 의료 데이터를 사용함에 있어 안전성과 활용성 둘 중 하나를 택할 수 없으며, 얼마나 안전하게 데이터를 활용할 수 있는가에 대한 해법은 비식별 기술에 대한 관심과 지원에서 비롯될 것임

◉ 개인정보를 보호하기 위한 법·제도의 개선뿐만 아니라 비식별 기술 개발을 통한 효과성 제고는, 의료 데이터를 안전하게 활용할 수 있는 사회적 인식 전환을 가져올 것으로 예상

Ⅵ 참고문헌

1. 강희정 외(2015), 보건의료빅데이터활용을위한기본계획수립, 보건사회연구원
2. 고학수(2015), 개인정보의 비식별화 처리가 개인정보 보호에 미치는 영향에 관한 연구, 개인정보보호위원회
3. 고학수(2017), 개인정보 비식별화 방법론, 박영사
4. 김동국 외(2015), 빅데이터 기반의 개인정보 비식별화 동향, 한국인터넷정보학회
5. 김동한(2017), 개인정보 비식별화 기술 동향 및 전망, 중앙전파관리소
6. 김원(2017), 개인정보 비식별 조치 주요국가 법제도 및 기술현황, 한국인터넷진흥원
7. 김현일 외(2017), 금융 데이터 상에서의 차분 프라이버시 모델 정립 연구, 공주대학교
8. 박대웅 외(2017), 최신 보건의료 빅데이터 법제 동향 조사분석, 한국보건산업진흥원
9. 박재형(2016), 빅데이터, 개방과 공유의 시대로, 디지에코 보고서
10. 신수용(2018), 보건의료 데이터 비식별화 문제점과 대안, 생명공학정책연구센터
11. 양현철 외(2016), 개인정보 비식별화기술 적용수준이 빅데이터 활성화에 미치는 영향, 한국EA학회
12. 이필우 외(2016), 국내외 비식별화 기술에 관한 검토 분석에 따른 개인건강의료정보 보호를 위한 국내 특화 비식별화 기술 제안에 관한 연구, 인문사회과학기술융합학회
13. 이현승 외1명(2016), 개인정보 비식별화 기술의 쟁점 연구, 소프트웨어정책연구소
14. 정성원(2014), Healthcare에서 빅데이터의 활용, 데이터솔루션
15. 정영철(2015), 의료분야 빅데이터 활용을 위한 개인정보 비식별화 규정 현황과 과제, 한국보건사회연구원
16. 조은지 외(2018), 완전 동형 암호 라이브러리의 성능 분석, 서울과학기술대학교
17. 개인정보보호위원회(2018), www.privacy.go.kr, 개인정보보호위원회
18. 금융보안원(2017), 빅데이터 환경에서의 개인정보 비식별 처리 방법 분석
19. 미래창조과학부(2016), 개인정보 비식별화 기술 활용 안내서 ver 1.0
20. 범부처 활동(2016), 개인정보 비식별 조치 가이드라인
21. 한국인터넷진흥원(2018), 2017년 해외 개인정보보호 동향 분석
22. ISO(2018), ISO/IEC DIS : 20889 - Privacy enhancing data de-identification terminology and classification of techniques, ISO

◉ 집필자 : 빅데이터팀 김재한 ◉ 문의 : 043-713-8020

◉ 본 내용은 연구자의 개인적인 의견이 반영되어 있으며, 한국보건산업진흥원의 공식견해가 아님을 밝혀둡니다.

◉ 본 간행물은 보건산업통계포털(http://www.khiss.go.kr)에 주간단위로 게시되며 PDF 파일로 다운로드 가능합니다.

KHISS 보건산업통계시스템 www.khiss.go.kr

한국보건산업진흥원
Korea Health Industry Development Institute