

NISTIR 8053

개인정보의 비식별처리

Simson L. Garfinkel

본 간행물은 아래 주소에서 무료로 열람하실 수 있습니다.
<http://dx.doi.org/10.6028/NIST.IR.8053>

NIST

국립표준기술연구소

미국 상무부

NISTIR 8053

개인 정보의 비식별처리

Simson L. Garfinkel

정보접근 부문(Information Access Division)

정보기술연구소(Information Technology Laboratory)

본 간행물은 아래 주소에서 무료로 열람하실 수 있습니다.
<http://dx.doi.org/10.6028/NIST.IR.8053>

2015 년 10 월



미국 상무부

장관, 페니 프리츠커(Penny Pritzker)

미국 국립표준기술연구소

상무부 장관 산하 표준기술연구소장 윌리 메이(Willie May)

국립표준기술연구소 내부보고서(NIST IR) 8053

vi + 46 페이지(2015 년 10 월)

본 간행물은 다음 주소에서 무료로 열람하실 수 있습니다.

<http://dx.doi.org/10.6028/NIST.IR.8053>

본 보고서에서 실험 절차 또는 개념을 적절하게 묘사하기 위해 특정 기업, 장비, 재료를 언급할 수 있다. 그러한 언급은 NIST 가 그것들을 추천하거나 보증하기 위한 것이 아니며, 또한 필연적으로 그러한 것들이 최적의 업체, 재료 또는 장비임을 암시하기 위한 것이 아니다.

본 보고서에는 NIST 정관에 제시된 NIST 의 책임에 따라 NIST 가 현재 작성하고 있는 다른 간행물에 대한 참조가 있을 수 있다. 본 보고서에 있는 개념, 방법론 등의 정보는 그러한 다른 간행물이 완성되기 전이라도 연방기관이 사용할 수 있다. 그러므로 각각의 간행물이 완성될 때까지는 현행 요구조건, 가이드라인, 절차가 계속 효력을 유지하게 된다. 연방기관은 NIST 가 작성하고 있는 그러한 신규 간행물을 기획과 전환을 목적으로 긴밀히 확인하기를 희망할 수 있다.

의견수렴 기간 중에 모든 간행물 초안을 검토하고 NIST 에 피드백을 제공할 것을 기관과 단체에게 권장하고 있다. 위에 언급된 간행물을 제외한 NIST 컴퓨터 보안 부문의 모든 간행물은 다음 주소에서 열람할 수 있다.

국립표준기술연구소

수신: 컴퓨터 보안 부문(Computer Security Division), 정보기술연구소(Information Technology Laboratory)
100 Bureau Drive (Mail Stop 8930) Gaithersburg, MD 20899-8930

컴퓨터 시스템 기술에 관한 보고서

국립표준기술연구소(이하 'NIST'로 약칭)의 정보기술연구실(Information Technology Laboratory, 이하 'ITL'로 약칭)은 측정과 표준 인프라에 관한 기술적 리더십을 제공함으로써 미국 경제와 공공 복지에 기여하고 있다. ITL 은 정보기술의 발전과 정보기술의 효과적인 이용을 위해, 시험, 시험방법, 참고자료, 개념증명의 이행, 기술적 분석법을 개발하고 있다. ITL 은 연방정부의 정보시스템에 있는 국가안보 관련 정보를 제외하고, 비용효과적인 안전과 프라이버시를 위한 경영 표준, 관리 표준, 기술 표준, 물리적 표준, 가이드라인을 개발하는 등의 업무를 맡고 있다.

초록

비식별처리(de-identification)는 정보 집합에서 식별정보(identifying data)를 제거함으로써 개인정보를 특정한 인물과 연결할 수 없도록 한다. 비식별처리를 통해 개인정보의 수집, 처리, 기록, 배포, 간행에 관련된 프라이버시 리스크를 줄일 수 있다. 그러므로 비식별처리는 프라이버시를 보호하면서 개인정보를 이용하고 공유한다는 상반되는 목표 간에 절충점을 찾고자 한다. 미국의 몇몇 법규와 정책은 개인정보를 공유하기 전에 반드시 비식별처리할 것을 규정하고 있다. 최근 연구에 따르면, 일부 비식별처리된 개인정보를 재식별(re-identify)할 수도 있다. 구조화된 정보, 자유로운 형식의 문서, 멀티미디어, 의학 영상 등 다양한 종류의 정보를 비식별처리할 수 있다. 본 보고서는 약 20년 동안 이루어진 비식별처리 관련 연구를 요약하고, 현재의 실태에 관하여 논의하며 향후의 연구과제를 제시하고 있다.

키워드

비식별처리(de-identification), HIPAA 프라이버시 규칙(Privacy Rule), k-익명성(k-anonymity), 차분 프라이버시(differential privacy), 재식별(re-identification), 프라이버시(privacy)

감사의 글

John Garofolo 와 Barbara Guttman 이 본 프로젝트에 많은 가이드와 지원을 제공하였다. 질문에 답변하고 이 문서의 초기 판본을 검토해 주신 Daniel Barth-Jones, David Clunie, Pam Dixon, Khaled El Emam, Orit Levin, Bradley Malin, Latanya Sweeney, Christine M. Task 에게도 감사한다. 이 문서의 최초 초안에 대해 Anonos, 민주주의와 기술 센터(Center for Democracy and Technology), 프라이버시의 미래 포럼(Future of Privacy Forum), 글락소스미스클라인(GlaxoSmithKline), IMS 헬스(IMS Health), 마이크로소프트(Microsoft),

옵툼 프라이버시(Optum Privacy), 환자 프라이버시권(Patient Privacy Rights), 프라이버시 어널리틱스(Privacy Analytics), 세계 프라이버시 포럼(World Privacy Forum) 등 많은 단체로부터 30 편 이상의 서면 논평을 받았다. 또한 COST Action IC1206 '멀티미디어 콘텐츠의 프라이버시 보호를 위한 비식별처리'의 회원, Marija Krlic 박사, Samir Omarovic 교수, Slobodan Ribarić 교수, Alexin Zoltan 교수 등이 제시한 논평도 받았다.

독자

본 보고서는 개인정보를 포함하고 있는 정보의 생성, 이용, 공유에 관한 정책적 사안에 관심이 있는 관리, 시민단체, 연구자, 기타 공동체의 구성원을 위한 자료이며 개인정보의 비식별처리에 포함된 기술적 사안에 대한 대략적인 내용을 과학기술 전문가와 학자들에게 제공하기 위한 문서이다. 본 보고서에 있는 정보는 정부의 개인정보보호 관련 공무원, 산업체, 학계에도 유익할 것이다.

메모 [한1]: 일반적인 정보를 의미합니다.

본 보고서는 정보시스템 보안 기술에 관한 높은 수준의 지식을 가진 독자를 대상으로 하고 있지만, 일반 대중도 쉽게 접근할 수 있도록 하였다. 따라서 수식의 사용을 최대한 자제하였다.

목차

1 서론.....1

1.1 본 보고서의 목적과 범위 1

1.2 예상 독자 1

1.3 구성 1

1.4 용어 안내 2

1.4.1 ‘비식별처리,’ ‘재편집,’ ‘가명처리,’ ‘익명처리’ 2

1.4.2.. ‘개인적으로 식별 가능한 정보(PII)’ 및 ‘개인 정보’ 3

2 비식별처리, 재식별, 정보 공유 모형3

2.1 비식별처리를 하는 이유 3

2.2 프라이버시를 보호하며 개인정보를 사용하기 위한 모형 6

2.2.1 프라이버시 보호 정보 마이닝(PPDM) 7

2.2.2 프라이버시 보호 정보 간행(PPDP) 8

2.3 비식별처리 정보 흐름 모형 9

2.4 재식별 공격 및 정보 공격자 9

2.5 공개 모형과 정보 통제 14

3 구조화된 정보의 비식별처리, 재식별 접근법..... 15

3.1 직접 식별자 제거 15

3.2 가명처리 16

3.3 연결 공격을 통한 재식별 17

3.4 준식별자의 비식별처리 19

3.5 HIPAA 에 따른 개인보건정보(PHI)의 비식별처리 22

3.5.1 HIPAA 전문가 활용법 22

3.5.2 HIPAA 세이프하버 방식 23

3.5.3 HIPAA 세이프하버 방식의 효과 평가 25

3.5.4 HIPAA 제한된 정보 집합 26

3.6 현장 비식별처리에 대한 평가 26

3.7 재식별 리스크의 추정 29

4 구조화되지 않은 개인정보의 비식별처리 30

4.1 의료 문서의 비식별처리 30

4.2 사진과 비디오의 비식별처리 32

4.3 의료 영상의 비식별처리 35

4.4 유전 정보와 생물학 자료의 비식별처리 36

4.5 지리 정보와 지도 정보의 비식별처리 37

5 결론 38

부록 목차

부록 A 용어집 39

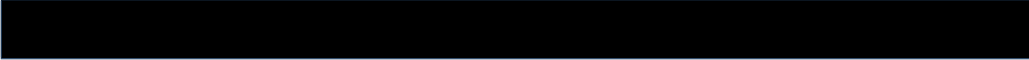
부록 B 참고문헌 44

 B.1 공식 간행물 44

 B.2 법률 검토 논문 및 백서 45

 B.3 보고서 및 서적 46

 B.4 설문조사 논문 46



비식별처리(de-identification)는 기관들이 수집, 사용, 기록하고 다른 기관과 공유하는 정보에서 개인정보를 제거하는데 사용할 수 있는 도구이다.

메모 [한2]: 일반적인 정보를 의미합니다.

비식별처리는 단일한 기법이 아니며, 다양한 종류의 개인정보에 적용할 수 있고 다양한 효과를 내는 접근법, 알고리즘, 도구의 집합체이다. 일반적으로 말하자면, 더 공격적인 비식별처리 기법을 적용하면 프라이버시가 더욱 보호되지만, 그러한 정보 집합(data set)의 유용성은 줄어들게 된다.

외부에 개인정보를 공개하려는 정부기관, 사업체, 기타 기관들에게 있어 비식별처리는 특히 중요하다. 예를 들면, 중요한 의학 연구의 경우, 의료 개인기록부의 프라이버시를 보장하기 위한 미국의 주된 규정인 「의료보험의 이동과 책임에 관한 법률(Health Insurance Portability and Accountability Act, 이하 'HIPAA'로 약칭)」에 설정된 제도적 틀에 따라 비식별처리된 환자 개인정보를 공유함으로써 사회적 편익을 얻을 수 있다.

비록 가능성이 낮지만, 개인정보에서 도출한 정보에 효용성이 남아있는 한, 일부 정보를 해당 정보의 원래 주인에게로 다시 연결할 수 있는 가능성도 있다. 비식별처리된 개인정보를 재식별할 수 있다면, 비식별처리에 따른 프라이버시 보호 효과가 사라질 것이다. 그러므로 재식별에 따른 리스크를 예측하기 어려울 수 있기 때문에, 개인정보를 비식별처리하는 방법 또는 여부에 대한 결정은 그러한 비식별처리된 개인정보의 사용, 공유, 공개 방법에 대한 결정과 연계하여 함께 내려야 한다.

메모 [한3]: 개인정보에서 나온 정보

메모 [한4]: 개인정보에서 나온 정보

비식별처리된 개인정보에도 개인에 대한 리스크가 남아있을 수 있다. 이러한 리스크에는 재식별을 하지 않고도 개인정보 안의 개인에 대한 추론이 가능해지거나 개인정보에 제시된 집단에 영향을 미칠 가능성 등이 있다.

1.1 본 보고서의 목적과 범위

본 보고서는 비식별처리에 관련된 이슈와 용어를 개관하고 있다. 본 보고서는 비식별처리와 재식별에 관련된 지금까지의 주요 간행물을 요약하고 있으나, 비식별처리의 적절성이나 특정한 비식별처리 알고리즘에 관한 권고를 제시하지는 않는다.

1.2 예상 독자

본 보고서는 개인정보를 포함하고 있는 정보의 생성, 사용, 기록, 공유에 관련된 실무적 사안이나 정책적 사안에 관심을 가진 관리, 시민단체, 연구자, 학계, 기타 커뮤니티의

구성원을 대상으로 하였다. 또한 연구자와 학계를 대상으로 개인정보의 비식별처리에 관한 기술적 이슈를 제한적으로 제시한다는 목적도 있다. 비록 이 문서는 전 세계의 연구를 제시하고 있고 전 세계의 독자를 대상으로 작성되었지만, 주로 미국의 정책과 법률적 이슈를 다루고 있다. 본 보고서는 정보시스템 보안 기술에 대한 높은 수준의 지식을 가진 독자를 염두에 두고 있지만, 더 폭넓은 대중이 접근할 수 있도록 하고 있다. 따라서, 이 문서는 수식의 사용을 최대한 자제하고 있다.

1.3 구성

본 보고서의 나머지 부분은 다음과 같이 구성되어 있다. 제 2 절에서는 비식별처리, 재식별, 개인정보 공유 모형의 개념을 소개하고 있다. 제 3 절에서는 통상적으로 성명이나 전화번호와 같은 특정 범주를 제거, 마스킹(masking), 변경함으로써 이루어지는 구조화된 개인정보의 비식별처리 방법에 관해 논의하고 있다. 제 4 절에서는 수형(樹型) 문서, 이미지, 유전자 정보 등 표(table) 형태가 아닌 개인정보의 비식별처리에 관련된 문제를 다루고 있다. 제 5 절은 본 보고서의 결론으로서, 비록 비식별처리가 완벽하지는 않지만 정보주체의 프라이버시를 보호할 수 있는 중요한 기술적 통제수단임을 주장하고 있다. 부록 A 는 용어집이며 부록 B 에는 참고문헌 목록이 제시되어 있다.

1.4 용어 안내

1.4.1 ‘비식별처리(de-identification)’, ‘재편집(redaction)’, ‘가명처리(pseudonymization)’, ‘익명처리(anonymization)’

일부 저자와 간행물은 ‘비식별처리’와 ‘익명처리’를 혼동하여 사용하고 있다. 일부는 ‘비식별처리’를 어떤 과정을 표현하는데 사용하고, ‘익명처리’는 되돌릴 수 없는 특정한 종류의 비식별처리를 지칭하는데 사용하고 있다. 의료에 관련된 특정한 상황에서는 ‘비식별처리’와 ‘가명처리’를 동등하게 취급하고 있으며, 정보주체의 신원에 대응하는 가명이 제거되었음을 표현하기 위해 ‘익명처리’라는 용어를 사용하고 있다. ‘재편집’이라는 용어는 종종 정부기관에서 식별정보(identifying information)나 달리 민감한 정보(sensitive information)를 직접적으로 제거하는 것을 가리키는데 사용한다.

본 보고서는 ISO/TS 25237:2008(E), ‘보건 정보공학 - 가명처리(Health Informatics – Pseudonymization)’에 있는 용어를 사용하고 있다. 그러나 이 표준은 분명한 지침이 되지 못하고 있다. 특히 이 표준에는 다음과 같은 정의가 있다.

비식별처리(de-identification): “식별정보 집합과 정보주체 간의 연계성을 제거하는 과정을 가리키는 일반적인 용어” [3 페이지]

익명처리(anonymization): “식별정보 집합과 정보주체 간의 연계성을 제거하는 과정”
[2 페이지]

가명처리(pseudonymization): “정보주체와의 연계성을 제거하며, 또한 정보주체에 관련된 특정한 특징들과 하나 또는 하나 이상의 가명 간에 연계성을 추가하는 익명처리의 일종”¹ [5 페이지]

ISO/TS 25237:2008(E)의 뒤에는 다음과 같은 설명이 있다.

“주석 – 익명처리는 비식별처리의 하위 범주에 속한다. 익명처리는 가명처리와는 다르게, 복수의 개인정보 기록부 또는 정보시스템에 걸쳐, 정보를 동일한 인물에게 연결할 수 있는 수단을 제공하지 않는다. 따라서 익명처리된 개인정보를 재식별하는 것은 불가능하다.” [6 페이지]

이러한 정의의 문제점은 일부 경우에는 익명처리를 이용하여 산출한 정보를 재식별할 수 있다는 점이다. 이러한 경우, 실제로 개인정보가 익명처리되지 않았지만 익명처리된 것으로 간주된다.

의학 영상의 경우, ‘비식별처리’라는 용어는 ‘실제 환자 식별자(real patient identifier)를 제거하는 과정, 또는 익명처리를 위해 영상 개인정보에서 모든 정보주체 집단을 제거하는 과정’을 가리킨다. 한편, ‘비인격화(de-personalization)’라는 용어는 ‘임상시험 식별자를 포함하여, 영상에서 정보주체와 관련된 모든 정보를 완전히 제거하는 과정’을 의미한다.²

‘익명처리’라는 용어의 용법과 정의가 일관되지 않기 때문에 이 문서에는 본 절과 일부 참조의 제목을 제외한 다른 곳에서는 되도록이면 ‘익명처리’라는 용어를 사용하지 않는다. 대신에 이 문서는 비식별처리된 정보를 재식별할 수 있거나 그렇지 않을 수도 있음을 감안하여 ‘비식별처리’라는 용어를 사용한다.

1.4.2 ‘개인적으로 식별 가능한 정보(Personally Identifiable Information, 이하 ‘PII’) 및 개인 정보(Personal Information)’

비록 다양한 법규와 기관의 지침에 개인적으로 식별 가능한 정보(PII)라는 용어에 관한 다양한 정의가 있지만, 이 용어는 보통 개인에 대한 특정한 식별자를 포함하고 있는 정보를 지칭한다. 다양한 정의가 있기 때문에, 특정인을 식별하지만 PII에 관한 정의에 부합되지 않는 정보가

메모 [한5]: 일반적인 정보를 의미합니다

¹ ‘코드화(coded)’라는 용어를 사용하여 식별자(identifier)가 가명으로 대체된 개인정보 또는 생물학적 표본을 설명하는 경우가 종종 있다. OHRP-Guidance on Research Involving Private Information or Biological Specimens, 미국 보건복지부(Department of Health & Human Services), 인간연구보호청(Office of Human Research Protections, OHRP), 2008년 8월 16일, <http://www.hhs.gov/ohrp/policy/cdebiol.html> 참조.

² Colin Miller, Joe Krasnow, Lawrence H. Schwartz, Medical Imaging in Clinical Trials, Springer Science & Business Media, Jan 30, 2014.

있을 수 있다. 어떤 문서에서는 PII 라는 용어를 개인에게 귀속되는 정보 또는 특정한 개인에게만 귀속시킬 수 있는 정보를 지칭하는 반면, 다른 문서에서는 이 용어를 사실 상, 개인을 식별하고 있는 정보에만 사용하고 있다.

이렇게 일관성이 없기 때문에, 이 문서에서는 ‘개인적으로 식별 가능한 정보’라는 용어의 사용을 자제한다. 대신에 ‘개인정보(personal information)’라는 용어를 이용하여 개인에게서 나온 정보를 지칭하며, 개인을 식별하는 정보는 ‘식별정보(identifying information)’라는 용어를 사용한다. 따라서, 식별정보는 개인정보이지만 개인정보는 식별정보가 아닐 수도 있다.

이 절에서는 비식별처리를 하는 이유를 설명하고 비식별처리의 효과를 알아보기 위해 재식별 공격에 관해 논의하며, 식별한 개인정보를 공유하기 위한 모형에 관하여 설명하고 있다. 본 보고서에 사용한 용어도 소개하고 있다.

2.1 비식별처리를 하는 이유

개인정보를 수집하고 유지하는 기관들은 한편으로는 개인정보를 최대한 광범위하게 사용하고 공유하면서, 다른 한편으로는 개인정보를 보호해야 하는 어려움이 점점 가중되고 있다. 정부 정보는 공유를 통해 투명성을 높이고 민간 기업에 새로운 자원을 제공하며, 전반적으로 효율적인 정부가 될 수 있다. 민간 기업들은 공개성과 시민 참여의 확대라는 형태로 공유되는 정보에서 이익을 실현하고, 심지어는 개인정보 판매나 분석 결과에서 이익을 얻을 수도 있다.

메모 [한6]: 일반적인 정보를 의미합니다

개인정보에 예를 들면 성명, 이메일 주소, 지리적 위치정보 또는 사진 등의 식별정보가 들어 있는 경우, 개인정보의 사용과 프라이버시 보호라는 목표 간에 갈등이 있을 수 있다. 이러한 갈등을 해결하기 위한 비식별처리를 통해, 다른 유용한 정보는 남겨두고 개인을 식별하는 일부 프라이버시에 관련된 민감한 개인정보를 제거할 수 있게 된다.

그러므로 비식별처리는 기관들이 개인정보를 포함하고 있는 정보를 생성, 사용, 기록, 공유, 심지어 간행하는 것에 관련된 프라이버시 리스크를 최소화할 수 있는 중요한 도구이다. 수집 당시 또는 최소한의 처리를 한 이후에 개인정보를 비식별처리하면, 의도하지 않은 공개(즉, 개인정보 침해)에 관련된 프라이버시 리스크를 줄임으로써 개인정보를 사용하고 기록으로 보존하는 비용을 줄일 수 있다. 비식별처리된 정보를 공유하면 기술적 통제와 정책적 통제 수요가 줄어들 수 있다. 그러므로 기관들은 비식별처리를 통해 개인정보를 더욱 잘 활용할 수 있다.

미국의 몇몇 법규는 개인정보 비식별처리의 중요성과 효용을 특별히 인식하고 있다. 다음과 같은 예가 있으나 이에 한정되지는 않는다.

- 교육부는 「가족 및 교육 개인기록부 프라이버시 법(the Family and Educational Records Privacy Act, FERPA)」에 따른 제한이 비식별처리된 학생 개인기록부에는 적용되지 않는다는 입장을 취하고 있다. ‘교육관련 정부기관과 교육기관은 개인적으로 식별 가능한 모든 정보를 제거함으로써, 비식별처리된 교육 개인기록부 또는 교육 개인기록부에 있는 개인정보를 동의 없이 공개할 수 있다.’³
- 「의료보험의 이동과 책임에 관한 법률(HIPAA)」 프라이버시 규칙⁴ 과 「경제와 임상보건을 위한 의료정보기술 법(the Health Information Technology for Economic and Clinical Health Act, HITECH 법)⁵」은 ‘해당 정보를 이용하여 개인을 식별할 수 있다고 믿을만한 합리적인 사유가 없는’ 경우, 비식별처리된 의료 정보에는 명시적으로 적용되지 않는다.⁶
- 질병 통제 및 예방본부(the Centers for Disease Control and Prevention)가 운영하는 식품유래 질병 감시체계(the Foodborne Illness Surveillance System)⁷에 따르면, ‘종합되고 비식별처리된 감시 개인정보에 대한 적시적인 일반인의 접근(access)’을 허용해야 한다.⁸
- 미국 보건복지부와 계약을 체결하여 약품안전 정보를 제공하는 업체들은 그러한 정보를 비식별처리된 형태로 제공할 능력이 있어야 한다.⁹
- 미국 연방항공국(the Federal Aviation Administration)에 제출되는 자발적 안전 보고서는 만약 보고서에 담긴 개인정보가 비식별처리되었다면 대중에 공개할 수 있다.¹⁰

메모 [한기]: 문맥 상 일반적인 정보를 의미

메모 [한8]: 원문은 the Federal Aviation Submission 이나 각주에 포함된 문헌(49 USC 44735)을 검토한 결과, ‘the Federal Aviation Administration’이므로 미 연방항공국으로 번역하였습니다.

³ Dear Colleague Letter about Family Educational Rights and Privacy Act (FERPA) Final Regulations, U.S. Department of Education, December 17, 2008. <http://www2.ed.gov/policy/gen/guid/fpco/hottopics/ht12-17-08.html>

⁴ 45 CFR 160, 45 CFR 162, and 45 CFR 164. See also ‘Combined Regulation Text of All Rules,’ U.S. Department of Health and Human Services, Office for Civil Rights, Health Information Privacy. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/combined/index.html>

⁵ 42 USC 17935

⁶ 42 CFR 164.514

⁷ <http://www.cdc.gov/foodsafety/fdoss/surveillance/index.html>

⁸ 21 USC 2224

⁹ 21 USC 355

¹⁰ 49 USC 44735

이러한 법규는 개인정보를 제거하고 나머지 개인정보를 정보주체의 신원을 공개하지 않는 식으로 사용할 수 있도록 하여, 비식별처리를 개인정보에 적용할 수 있는 기술적 통제 수단으로 간주한다.

그러나 데이터베이스에 있는 특정 항목을 범주화(suppressing)하거나 일반화(generalizing)하는 비식별처리 접근방식은 프라이버시에 대한 절대적인 보장 수단이 될 수 없다. 왜냐하면 보조 정보 집합을 이용하여 나머지 개인정보를 재식별할 수 있는 가능성이 항상 있기 때문이다. 제 3.6 절에는 적절히 비식별처리된 것으로 간주되었던 개인정보가 공개되었고, 나중에 연구자나 언론이 이를 재식별한 유명한 몇 가지 사례를 논의하고 있다. 이러한 재식별 사례 중 일부에서는 정보주체의 신원을 노출시켰다. 정보 집합을 신원에 연결하였던 특정한 속성을 공개함으로써 발생할 수 있는 프라이버시에 관련된 이슈도 있다. 그럼에도 불구하고 HIPAA 비식별처리 기준(특정 항목의 범주화를 기초로 함)을 충족시킬 수 있는 두 가지 접근법 가운데 하나인 HIPAA '세이프하버(Safe Harbor)' 방식을 통한 시험에 따르면, 남은 식별 정보가 출생연도, 성별, 3 자리 우편번호(ZIP code)뿐인 경우에는 비식별처리된 개인기록부 중에 재식별할 수 있는 부분은 1% 미만이라는 사실이 밝혀졌다(제 3.5 절 참조).

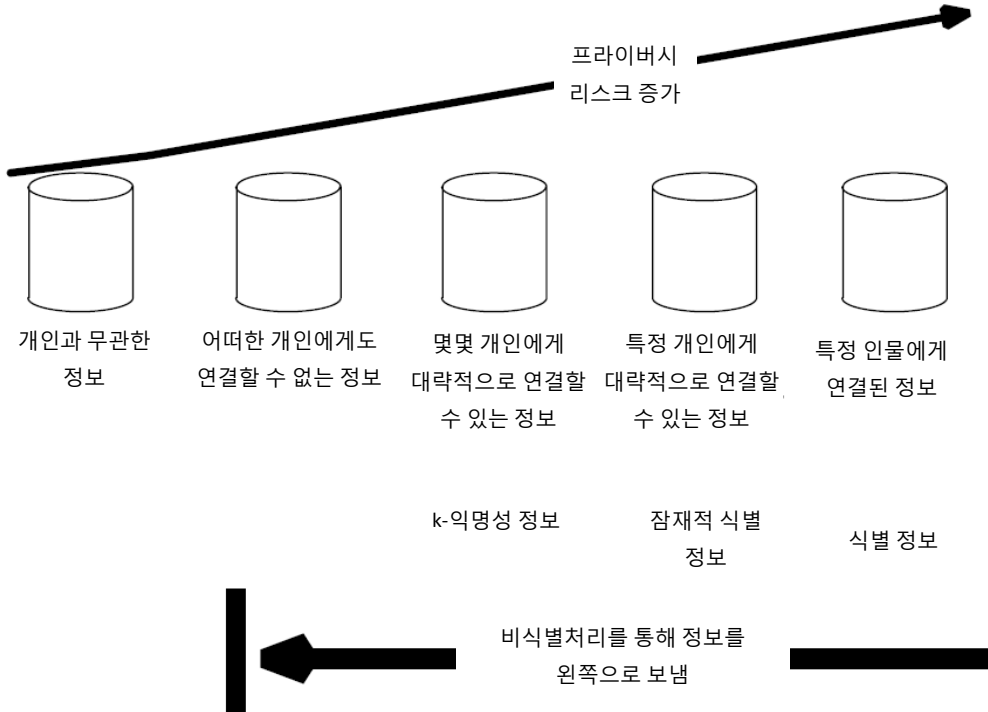
재식별이라는 리스크가 있기 때문에 비식별처리된 개인정보를 공유하는 일부 기관들은 정보 이용자와 정보 이용에 관한 합의서(data use agreement, 이하 'DUA'로 약칭)를 체결하려 할 수 있다. 예를 들면 DUA 를 체결함으로써, 비식별처리된 개인정보의 이용자가 정보주체를 재식별하려 하거나 외부 정보와 연결하거나 무단으로 개인정보를 공유하는 행위를 막을 수 있다.¹¹

그림 1 에 제시된 것처럼, 식별가능성을 기준으로 모든 개인정보를 분류할 수 있다. 왼쪽 끝에 있는 개인정보는 개인에 관련된 정보가 아니기 때문에(예를 들면 과거의 날씨 기록) 프라이버시에 관련된 리스크가 없다. 오른쪽 끝에 있는 개인정보는 특정인에게 직접 연결되어 있다. 이러한 양쪽 극단의 중간에는 노력을 통해 연결할 수 있는 개인정보, 집단에만 연결할 수 있는 개인정보, 개인을 기반으로 한 정보이지만 역으로 연결할 수 없는 정보가 있다. 일반적으로 보면, 비식별처리라는 접근방식은 필요한 효용성을 유지하면서 개인정보를 왼쪽으로 밀어냄으로써, 비식별처리된 개인정보를 광범위한 집단이나 일반 대중에게 배포할 리스크를 낮추도록 하는 방식이라고 할 수 있다.

메모 [한자]: 개인기록부가 아닙니다.

¹¹ HIPAA 프라이버시 규칙에 따라, 더 느슨한 규정에 따라 비식별처리된 제한된 정보 집합을 공유할 때 데이터 이용 합의서가 필요하며, 프라이버시 규칙에 따라 비식별처리된 정보 집합을 공유할 때에는 반드시 필요한 것은 아니다(예, 전문가 활용법에 따라 전문가가 DUA 를 요구할 수 있다).

프라이버시를 옹호하는 사람들은 공정정보원칙(Fair Information Practice Principles, 이하 'FIPP'으로 약칭)¹²에 따라, 설령 공유하는 개인정보를 비식별처리하더라도, 그리고 정보주체에게 통지해야 할 법적 의무는 없지만, 정보주체에게 그들의 개인정보가 공유될 것이라는 사실을 통지해야 한다는 입장이다. 그러나 미국의 현행 정책과 법률은 기관들이 이용하는 비식별처리된 개인정보의 용도에 대해 상당한 자유를 부여하고 있다. 이러한 정책들은 일반적으로 비식별처리된 개인정보를 사용함으로써 유발되는 사회적 편익과 개인정보의 재식별에서 유발될 수 있는 정보주체에 대한 인지된 리스크 사이에 균형을 맞추기



위한 정책이다. 기술이 발전함에 따라 이러한 리스크가 변할 수 있기 때문에, 비식별처리된 개인정보의 사용에 관한 정책을 주기적으로 검토하는 것이 중요하다.

그림 1: 개인정보 식별가능성 스펙트럼

2.2 프라이버시를 보호하며 개인정보를 사용하기 위한 모형

학계에서는 정보주체의 프라이버시를 보호하면서 데이터베이스에 있는 개인정보를 이용하기

메모 [한10]: 원문은 privacy preserving(프라이버시 보존)이지만, (흔치 않은 경우이긴 하지만) 아래에 있는 설명에 'preserving'이 'protection'과 같은 의미로 사용되었으므로 공히 '프라이버시 보호'로 번역하였습니다.

¹² National Strategy for Trusted Identities in Cyberspace, Appendix A—Fair Information Practice Principles. April 15, 2011. <http://www.nist.gov/nstic/NSTIC-FIPPs.pdf>

위한 두 가지 모형을 식별하였다.

- **프라이버시 보호 정보 마이닝(Privacy Preserving Data Mining, 이하 'PPDM')**. 이 모형에서는 개인정보가 공개되지 않으며, 대신에 통계적 처리나 기계 학습(machine learning)¹³에 사용된다. 계산 결과는 기계 학습 알고리즘을 이행하는 요약화(summarization), 총계처리(aggregation), 분류자(classifier)를 기초로 한 통계표의 형태나 및 다른 형태의 결과로 공개할 수 있다.
- **프라이버시 보호 정보 간행(Privacy Preserving Data Publishing, 이하 'PPDP')**. 이 모형에서는 개인정보를 처리하여 사용자에게 배포할 수 있는 새로운 비식별처리되거나 합성한 정보를 생산한다.

이 두 모형은 모두 원본 정보 집합에 있는 특정인에게 귀속시킬 수 있는 정보를 노출하지 않고 일부 개인정보(예, 총계처리 한 정보, 통계적 결과, 분류자 또는 합성한 개인정보)를 노출하도록 하기 위한 것이기 때문에 '프라이버시 보호' 모형이라 부른다.

2.2.1 프라이버시 보호 정보 마이닝(PPDM)

PPDM은 공식 통계를 발표하기 위해 민감한 개인정보를 이용하는 것을 가리키는 일반적인 용어이다. 비밀 설문조사 정보를 요약한 통계 보고서가 그 예라고 할 수 있다.

통계적 공개 한도(Statistical Disclosure Limitation)란 “3 자가 정보를 이용하여 정보에 있는 개인을 인식하는 것을 방지하기 위해 통계적 정보를 변경하는 원칙이다.”¹⁴ 공개 제한을 위해 개발된 기법에는 보고된 정보를 더 큰 범주로 일반화하는 기법, 유사한 개체 간에 정보를 교환하는 기법, 문서에 잡음(noise)을 삽입하는 기법 등이 있다.¹⁵

차분프라이버시(Differential Privacy)란 정보 집합의 계산에서 비롯되는 신원 공개와 정보 누출에 관한 수학적 정의를 이용하는 기법을 말한다. 차분프라이버시는 수학적 계산의 결과를 보고하기 전에 비결정적 잡음(non-deterministic noise)(보통 작은 임의 값)을 삽입함으로써 공개를 방지한다.¹⁶ 차분프라이버시의 수학적 정의에 따르면, 정보 집합의 분석 결과는 한 정보 개인기록부(보통은 단일한 인물의 개인정보로 간주됨)을 추가하거나 제거한 전과 후에

¹³ 기계 학습은 컴퓨터가 명시적으로 그렇게 프로그래밍되어 있지 않고도 정보를 분류하고 패턴을 인식하도록 하는 컴퓨터 알고리즘과 기법이다.

¹⁴ Leon Willenborg and Ton de De Wall, Elements of Statistical Disclosure Control, 2001. Springer

¹⁵ Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005.

¹⁶ Cynthia Dwork, Differential Privacy, in ICALP, Springer, 2006

대략적으로 동일해야 한다. 삽입한 잡음이 개인의 기여를 마스킹(masking)하기 때문에 효과가 있다. 동일한 정도는 변수 ϵ (입실론)으로 정의한다. ϵ 값이 작을수록 더 많은 잡음이 첨가되며, 단일한 개인기록부의 기여도를 분간하기가 더욱 어려워진다. 결과적으로 정보주체 전체의 프라이버시가 높아진다. 가장 기본적인 형태의 차분프라이버시는 온라인 문의 시스템에만 적용되지만, 차분프라이버시를 이용하여 기계 학습 통계 분류자와 합성 정보를 생성할 수도 있다.¹⁷

차분프라이버시에 관한 연구가 활발하지만 지금까지는 다음과 같은 몇몇 계산 시스템에만 적용되어왔다.

- 합리적이고 정확한 블록(block) 단위의 합성 인구조사 정보를 생성하기 위해 차분프라이버시를 이용하고 있는 미국 통계국(the Census Bureau)의 '온더맵(OnTheMap)' 웹사이트¹⁸
- 사용자의 컴퓨터와 사용자의 홈페이지에서 구동되는 윈도우즈(Windows) 프로세스에 관한 총계 통계를 수집하기 위해 임의화 된 반응을 사용하는 구글(Google)의 '크롬(Chrome)' 웹브라우저. 비록 통계는 총계 수준에서 정확하지만 임의화를 하기 때문에, 사용자의 프로세스나 홈페이지를 신뢰성 있게 확인할 수 없다.¹⁹

차분프라이버시를 적용하면 결과의 정확도가 낮아진다. 예를 들면, Fredrikson 등은 임상시험 자료를 이용하여 유전자 정보와 와파린(warfarin) 투여의 상관관계에 관한 통계적 모형을 수립하기 위해 차분프라이버시를 사용하는 경우에 미치는 영향을 확인하였다.²⁰ 이 연구는 비록 실제 임상시험이 아니라 시뮬레이션에서만 시험되었지만, 차분프라이버시를 이용하여 수립한 모형은 차분프라이버시를 이용하지 않고 수립한 모형에 비해 유의미한 숫자의 환자에게서 더 열악한 임상적 결과가 나오는 경향이 있음을 밝혔다.

2.2.2 프라이버시 보호 정보 간행(PPDP)

PPDP 를 이용하여 개인정보에 기반한 정보를 간행할 수 있어, 다른 연구자들이 새로운

¹⁷ Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron, Zhiwei Steven Wu, Dual Query: Practical Private Query Release for High Dimensional Data, Proceedings of the 31st International Conference on Machine Learning, Beijing, China. 2014. JMLR: W&CP volume 32.

¹⁸ Abowd et al., 'Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data,' Joint NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection, Suitland, MD, July 31, 2009.

¹⁹ Úlfar Erlingsson, Vasył Pihur, Aleksandra Korolova, RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14). ACM, November 3-7, 2014, Scottsdale, AZ. <http://dl.acm.org/citation.cfm?doid=2660267.2660348>

²⁰ Fredrikson et al., Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, 23rd Usenix Security Symposium, August 20-22, 2014, San Diego, CA.

분석을 할 수 있게 된다. PPDP 를 이용하는 목적은 정보주체의 신원을 노출시키지 않으면서도 효용성이 높은 정보를 제공하기 위한 것이다.²¹

메모 [한11]: 일반적인 정보를 의미

비식별처리(de-identification)는 “식별 정보와 정보주체 간의 연계성을 제거하는 과정을 나타내는 일반적인 용어이다.”²² 비식별처리는 개인의 신원을 보호하기 위한 것이며, 다른 목적을 위한 정보 집합의 유용성의 일부를 보존하면서, 정보 집합에 있는 개인정보가 만약 특정 개인에 관련되어 있다면 알기 어렵거나 불가능하도록 한다. 비식별처리는 PPDP 를 달성하기 위한 주요한 도구이다.

합성 정보 생성(synthetic data generation)은 원본 정보와 유사한 정보 집합을 생성하기 위해 PPDM 기법을 일부 이용하지만, 이 때 정보 요소의 일부 또는 전부가 생성되고, 실제 개인을 참조(mapping)하지 않는다. 그렇기 때문에 합성 정보 생성은 PPDM 과 PPDP 가 융합된 것으로 간주할 수 있다.

2.3 비식별처리 정보 흐름 모형

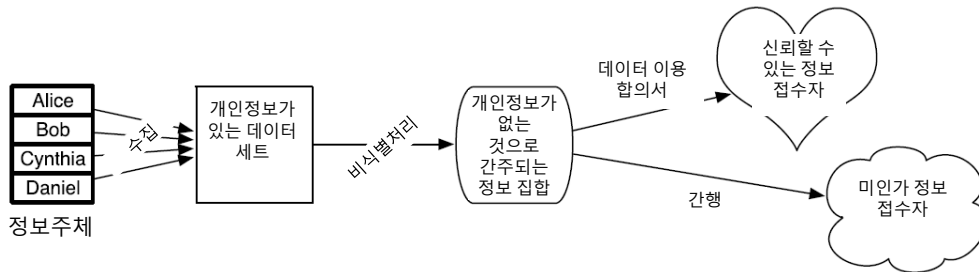


그림 2: 개인정보 수집, 비식별처리 및 이용

그림 2 에는 비식별처리 과정이 개관되어 있다. ‘정보가 참조하는 사람인’ 정보주체(Data Subject)로부터 개인정보를 수집한다(ISO/TS 25237-2008). 이러한 개인정보는 개인정보를 포함하고 있는 정보 집합으로 결합된다. 비식별처리를 통해, 식별정보가 없는 것으로 간주되는 새로운 정보 집합을 생성한다. 기관들은 프라이버시 리스크를 줄이기 위해 이 정보 집합을 원본 정보 집합 대신에 내부적으로 사용할 수 있다. 이 정보 집합은 DUA 등의 추가적인 관리적 통제를 받는 신뢰할 수 있는 정보 접수자에게 제공할 수 있다.²³ 그 대신에,

²¹ Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, Computing Surveys, June 2010.

²² ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008.

²³ 앞에서 언급한 대로, HIPAA 프라이버시 규칙에 따르면 전문가 활용법이나 세이프하버 방식을 이용하여 비식별처리된 정보 집합이 아니라, 제한적인 정보 집합의 경우에 DUA 를 요구한다. 그럼에도 불구하고 비록 HIPAA 기준에 따라 비식별처리한 경우라도 비식별처리된

예를 들면 비식별처리된 개인정보를 인터넷에 게시함으로써 이 정보를 수많은(아마도 무수한) 미지의 미인이 정보 접수자에게 널리 배포할 수 있다.

그림 2 에 제시된 대로, 비식별처리된 개인정보를 대중에게 공개할 필요는 없다. 더욱이, 비식별처리는 정보주체의 신원을 보호하기 위해 적용되는 몇 가지 통제 가운데 하나에 지나지 않을 수 있다.

비식별처리는 자동 프로세스, 수작업 또는 이 둘을 혼합하여 수행할 수 있다. 비식별처리가 이루어진 경우, 개인정보를 수작업으로 검토하거나 다른 방법을 이용하여 확인함으로써 식별 정보가 남아 있는지, 또는 다른 정보 출처와 연계하여 비교함으로써 제거된 신원을 알 수 있는지 확인할 수 있다.

2.4 재식별 공격 및 정보 공격자

재식별은 비식별처리된 개인정보에서 제거된 신원을 분간하려는 과정을 의미한다. 비식별처리를 하는 주된 목적은 승인되지 않은 재식별을 방지하는 것이기 때문에 그러한 시도는 종종 재식별 공격(re-identification attack)이라고 부른다.

‘공격’이라는 용어는 컴퓨터 보안에 관한 문헌에서 따왔는데, 공격에서는 특수한 기법, 지식, 접근권(access)을 갖고 있는 가상적인 ‘공격자(attacker)’를 이용하여 컴퓨터 시스템이나 암호화 알고리즘의 보안을 분석한다. 리스크 분석에는 잠재적인 공격자를 나열하고 각 공격자의 성공 확률을 분석한다.

개인이나 조직이 재식별 공격을 하는 데에는 많은 이유가 있다.

- **비식별처리의 질을 시험하기 위해.** 예를 들면, 연구자는 정보 처리자(data controller)의 요청에 따라 재식별 공격을 수행할 수 있다. 만약 원본 정보에 개인 보건정보 또는 교육 개인기록부 등의 일종의 법적으로 보호받는 정보가 담겨 있다면, 연구자는 사전에 적절한 비밀준수 합의를 체결하고 적절한 보안 절차를 시행해야 한다. 그렇지 않을 경우, 재식별의 성공은 ‘법규 위반(breach)’으로 간주될 것이며, 적용 가능한 법규나 정책에 따라 의무적인 보고 요건이 작동하게 될 것이다.
- **재식별 수행에 관한 대중의 관심 또는 전문가적 관점을 얻기 위해.** 재식별에 성공한 몇몇 사례는 뉴스거리가 되거나 그를 수행한 학자에게 보상이 뒤따랐다. 이러한 공격이

개인정보가 재식별될 리스크가 항상 있다. 그렇기 때문에 재식별 리스크를 최소화시키면서 비식별처리된 개인정보를 교환하기를 희망하는 기관은 추가적인 관리 통제 수단으로 DUA 를 적용하기를 원할 수 있다.

재식별을 금지하는 DUA 를 체결하지 않고 합법적으로 수행되었음에 유의한다.

- **비식별처리를 수행한 기관을 곤경에 처하게 하거나 해할 목적으로.** 비식별처리를 수행하는 기관은 일반적으로 원본 정보에 담긴 개인정보를 보호할 의무를 진다. 그렇기 때문에 부적절한 프라이버시 보호 조치가 적용되었음을 보여줌으로써 이러한 기관을 곤경에 처하게 하거나 해할 수 있다. 특히 비식별처리된 개인정보가 공개적으로 유출된 경우에 그러하다.
- **재식별된 개인정보에서 직접적인 이익을 얻기 위해.** 예를 들면, 마케팅 회사는 비식별처리된 건강정보를 구입하여 그러한 정보를 신원과 매칭시켜, 재식별된 사람들에게 특정한 처방약 쿠폰을 발송할 수 있다.
- **재식별을 통해 민감한 개인정보를 알 수 있는 사람들에게 곤경이나 해와 같은 문제를 야기하기 위해.** 직접 개인정보를 공개하거나 개인을 위협 또는 비방하기 위해 개인정보를 사용하거나 사임을 강요하거나 다른 부정적인 결과를 강제하기 위해 개인정보를 사용함으로써 문제가 발생할 수 있다.

문헌에 따르면, 재식별 공격은 종종 비식별처리된 정보 집합과 약간의 추가적인 *배경 정보*를 갖고 있는 가상의 *정보 공격자(data intruder)*가 수행하는 것으로 기술되어 있다(대중에게 공개된 정보 집합의 경우, 정보 공격자는 컴퓨터 시스템이나 정보 집합에 대한 비승인 접근권(unauthorized access)을 획득할 필요가 없다는 점에 주의한다).

*재식별 리스크(re-identification risk)*²⁴란 정보 집합에 있는 식별자와 기타 개인에 관련된 정보를 비식별처리된 개인정보로부터 알게 되는 리스크를 나타내는 척도이다. 재식별 능력은 원본 정보 집합, 비식별처리 기법, 공격자의 기술적 능력, 공격자의 가용한 자원, 비식별처리된 개인정보에 연결할 수 있는 추가 정보의 가용성에 달라 달라지기 때문에 이러한 리스크를 정량화하기는 어렵다.

기법이 개선되고 상황적 정보가 많아짐에 따라(예, 공개 또는 구매를 통해), 재식별 리스크가 증가하게 될 것이다. 그렇기 때문에 미래에 재식별에 어떤 종류의 상황적 정보가 활용될 수 있는지를 알고리즘을 이용하여 정하는 것은 불가능하다.

연구자들은 재식별 리스크를 계산하고 보고하기 위해 다양한 접근법을 취해 왔다. 그러한

²⁴ Elliot M, Dale A. Scenarios of attack: the data intruders perspective on statistical disclosure risk Netherlands Official Statistics 1999;14(Spring):6-10.

접근법에는 통상적으로 성공 정도를 기술하는 시나리오와 공격자가 갖고 있는 자원과 능력을 측정하기 위한 도구 등이 있다. 재식별 시나리오에는 다음과 같은 예가 있다.

- 만약 특정인이 정보 집합 안에 있음을 공격자가 알고 있는 경우, 그러한 특정인을 재식별할 리스크(‘검사(prosecutor) 시나리오’)
- 정보 집합 안에 재식별할 수 있는 사람이 최소한 1 명이 있는 리스크. 누군가를 재식별할 수 있음을 증명하는 것이 중요하다. 이 경우, 비식별처리를 수행한 기관을 곤경에 처하도록 하거나 신뢰를 떨어뜨리는 것이 재식별의 목표인 경우가 많다(‘기자(journalist) 시나리오’²⁵).
- 정보 집합 안의 신원의 비율을 정확하게 재식별할 수 있는 경우(‘마케터(marketer) 시나리오’)
- 어떤 개인이 포함된 정보 집합에 대해 수행한 분석과 해당 개인이 포함되지 않은 정보 집합에 대해 수행한 같은 분석 간의 구분 가능성(‘차분 식별 가능성(differential identifiability)’ 시나리오’²⁶).

정보 공격자의 능력을 기술하기 위해 다음과 같은 기준을 사용한다.

- 공개된 정보에 대한 접근권(access)이 있는 일반 대중(‘일반 대중’)
- 재식별에 숙련된 컴퓨터 과학자(‘전문가’)
- 정보 집합을 생산한 기관의 구성원(‘내부자’)
- 비식별처리된 개인정보를 접수하는 기관의 구성원이나, 일반 대중에 비해 더 많은 배경 정보에 대한 접근권(access)이 있을 수 있는 사람(‘내부 접수자’)
- 정보를 조합하여 풍부한 정보 상품을 생산하고, 이를 내부적으로 활용하거나 다시 판매할 목적으로 식별된 정보와 비식별처리된 정보를 모두 체계적으로 획득하는 정보 브로커(‘정보 브로커’)
- 특수한 상황에 있는 정보주체의 친구 또는 가족(‘참견하는 이웃’)

비식별처리를 하는 목적은 정보주체의 신원을 가림으로써 약간의 프라이버시를 보호하면서도 비식별처리된 개인정보를 활용할 수 있도록 하는데 있다.

비식별처리 정도와 생성된 정보의 유용성 간에 상충 관계가 있다는 점에서 볼 때, 이 두 가지

메모 [한12]: 일반적인 정보를 의미

²⁵ Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. J. Bus. Econ. Statist., 6, 487-500.

²⁶ 25 Jaewoo Lee and Chris Clifton. 2012. Differential identifiability. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 1041-1049. DOI=10.1145/2339530.2339695 <http://doi.acm.org/10.1145/2339530.2339695>

목표는 상반된 목표이다. 그러나 비식별처리를 하면, 과거에는 프라이버시에 대한 우려로 인해 사용이 금지되었던 개인정보를 새롭게 활용할 수 있게 된다. 그러므로 적절한 보안 수준, 따라서 비식별처리와 유용성 간에 수용할 수 있는 절충점을 찾는 역할은 정보 관리자, 표준기관, 규제기관, 입법자, 법원이 맡는다.

어떤 경우에는 비식별처리된 개인정보를 사용함에 따라 정보 집합에 있는 사람에 해가 되거나 프라이버시에 부정적인 행위가 있을 수 있다. 일반적인 분류학에 따르면, 그러한 리스크는 신원 공개, 속성 공개, 추론적 공개(inferential disclosure) 등 세 가지로 구분된다.²⁷

신원 공개(identity disclosure)는 공격자가 특정한 정보 항목을 특정인에게 연결시킬 수 있을 때 일어난다. 몇 가지 시나리오를 통해 신원 공개가 발생할 수 있다.

- 비식별처리가 불충분한 경우.** 통제가 불충분할 때, 비식별처리된 정보 집합 안에 의도치 않게 식별정보가 남아있을 수 있다. 2006년에 AOL 이 누출한 검색 문의(query) 정보의 경우가 그러하였다. AOL 은 약간의 식별 정보를 제거했지만 사용자가 입력한 검색어를 보존하였다. 아라비아 숫자 코드 하나를 통해 정보에서 사용자를 식별하였다. 해당 코드는 임의적으로 생성된 가명(pseudonym)이었기 때문에 자체적으로는 사용자의 신원에 다시 연계할 수 없었다. 검색 문의 자체에 식별정보가 나타났고(예를 들어 자신의 자산에 관한 정보를 검색한 사람들), 가명이 있었기 때문에 동일한 사용자의 다른 복수의 문의에 매칭시킬 수 있었다. 기자들은 그러한 검색어로부터 몇몇 사용자를 식별할 수 있었고, 사용자와 연락하여 논평을 요청하였다.²⁸ 다른 사례에서 Sweeney 는 워싱턴 주립 병원의 비식별처리된 퇴원 개인기록부에 있는 일부 환자의 경우, 퇴원 개인기록부에 있는 정보와 입원을 유발하게 된 사건에 관한 신문 기사를 수작업으로 연관시킴으로써, 환자 정보를 식별할 수 있음을 보여주었다.²⁹
- 연결을 통한 재식별.** 남아있는 일부 정보를 다른 식별 정보 집합에 있는 유사한 속성과 연결시킴으로써 특정한 개인기록부를 재식별할 수 있다. 예를 들어, 검색 기록을 비식별처리하면 검색자의 성명을 제거할 수 있지만, IP 주소가 남게 되어 해당 정보를 IP 주소를 성명과 연계하는 데이터베이스에 연결할 수 있다.
- 가명 역추적.** 어떤 개인정보를 가명처리하였고, 만약 가명이 신원 정보에서 도출된

메모 [한13]: 일반적인 기록물입니다

²⁷ Li Xiong, James Gardner, Pawel Jurczyk, and James J. Lu, 'Privacy-Preserving Information Discovery on EHRs,' in Information Discovery on Electronic Health Records, edited by Vagelis Hristidis, CRC Press, 2009.

²⁸ Barbaro M, Zeller Jr. T. A Face Is Exposed for AOL Searcher No. 4417749 New York Times. 9 August, 2006.

²⁹ Sweeney L. Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1. June 2013.

것이라면, 가명처리 과정을 역으로 되돌릴 수 있다. 제 3.2 절에 논의된 뉴욕 시 택시 리무진 위원회가 공개한 택시 승차 정보 집합의 경우가 이에 해당한다.

속성 공개(attribute disclosure)는 약간의 비밀 정보를 정보주체에 귀속시킬 수 있을 때 발생한다.

정보 집합에 비밀정보가 조금이라도 포함되어있는 경우, 신원이 공개되면 반드시 속성이 공개된다. 그러나 한 정보 집합이 어떤 특징을 공유하고 있는 개인이 특정한 속성을 모두 갖고 있음을 드러내고, 만약 적이 샘플에서 그러한 특성을 갖고 있는 개인을 알고 있는 경우, 신원이 공개되지 않고 속성이 공개될 수 있다. 예를 들면, 한 병원이 치료한 20 세 여성 모두가 특정한 진단을 받았음을 나타내는 정보를 공개하고, Alice Smith 가 20 세 여성이며, 해당 병원에서 치료를 받은 사실이 알려져 있다면, 비록 그녀의 비식별처리된 치료 개인기록부를 다른 사람들의 것과 분간할 수 없더라도, Alice Smith 의 진단 결과를 추론할 수 있다.³⁰ I-다양성 기법을 활용하면, 특정한 기준에 일치되는 개인기록부 집단 각각이 일정 수준의 다양성을 갖고 있도록 함으로써 추론적 공개로부터 보호하는데 도움이 될 수 있다.³¹ 이러한 접근방식은 심지어 그렇게 할 필요가 없을 경우, 예를 들면 정보주체에 낙인화(stigmatization) 또는 해를 입힐 리스크가 없는 경우라도 정보의 유용성을 낮출 수 있기 때문에 적용할 때 신중해야 한다.

추론적 공개는 “공개된 정보의 통계적 특징으로부터 높은 신뢰도로 개인정보를 추론할 수 있는 경우에 발생한다. 예를 들면, 어떤 정보는 소득과 주택 구매 가격 간에 높은 상관관계를 나타내고 있을 수 있다. 주택의 구매 가격은 통상적으로 공개되는 정보이기 때문에 제 3 자는 이 정보를 이용하여 정보주체의 소득을 추론할 수 있다.”³²

추론적 공개가 발생하는 경우에는 자신의 개인정보가 정보 집합에 포함되어 있지 않은 사람을 포함하여 개인들로 이루어진 전체 계층이 집단적인 해(group harm)를 입을 수 있다. 예를 들면, 어떤 정보 집합에 특정 인구학적 집단이 잘 대표되어 있고, 해당 집단이 정보 집합에서 낙인화되는 비율이 높다면, 비록 해당 인구 집단에 포함된 모든 사람들이 낙인화되는 것이 적절하지는 않지만 그렇게 될 수 있다.

³⁰ El Emam, Methods for the de-identification of electronic health records for genomic research. Genome Medicine 2011, 3:25 <http://genomemedicine.com/content/3/4/25>

³¹ I-다양성 기법은 비식별처리된 정보 집합의 개인기록부들이 범위 값을 대표하도록 함으로써 추론적 공개로부터 보호하는데 도움이 될 수 있다. 이 기법의 단점은 비식별처리된 개인정보의 원본 정보 집합에 대한 충실성(fidelity)이 줄어든다는 점이다. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramaniam. I-diversity: Privacy beyond k- anonymity. In Proc. 22nd Intl. Conf. Data Engg. (ICDE), page 24, 2006.

³² Glossary of statistical terms, OECD, November 30, 2005. <https://stats.oecd.org/glossary/detail.asp?ID=6932>

법률과 규정에 반영되어 있는 미국의 프라이버시 정책은 일반적으로 신원 공개를 우려하고 있으며, 식별된 개인정보를 이용하거나 배포하여 발생할 수 있는 기타 해악이나 문제는 우려하지 않고 있다. 이러한 리스크를 해결하려는 기관은 명시적 윤리 검토(explicit ethics review)를 통해 그러한 문제를 해결하려 할 수 있다. 그러한 검토는 다음과 같은 내용을 적절히 감안하여 조절해야 한다.

- 해당 기관이 비식별처리 절차를 수행하고 시험하는데 투입할 수 있는 노력
- 비식별처리된 개인정보를 통한 효용(즉, 해당 정보에 대표된 개인 및/또는 집단에 대한 이익)
- 비식별처리된 개인정보를 이용하여 발생할 수 있는 개인 또는 집단의 문제
- 리스크를 최소화할 수 있는 다른 통제를 활용할 능력
- 공격자가 개인정보 재식별을 시도할 확률과 공격자가 할 수 있는 노력의 양

2.5 공개 모형과 개인정보 통제

재식별이 발생할 가능성을 줄일 방법에는 개인정보를 획득하고 사용하는 방식을 통제하는 방법이 있다. 이러한 통제는 공개 모형에 따라 분류할 수 있다. 문헌에는 무제한(no restriction)에서 엄격한 제한(tight restriction)에 이르는 몇 가지 모형이 제안되어 있다.

- **일반 공개 모형(the Release and Forget model)**³³ : 통상적으로 인터넷에 게시함으로써 비식별처리된 개인정보를 대중에게 공개할 수 있다. 이러한 방식으로 일단 개인정보가 공개되면 기관이 개인정보를 회수하기는 거의 불가능하다.
- **데이터 이용 합의서(DUA) 모형**: 개인정보를 어떻게 이용할 수 있는가를 세부적으로 규정한 법적 구속력이 있는 데이터 이용 합의서에 따라 비식별처리된 개인정보를 공개할 수 있다. 통상적으로 데이터 이용 합의서는 재식별 시도, 다른 정보와의 연결, 정보 재배포를 금지한다. DUA 는 개인정보 보유자와 자격을 갖춘 연구자(‘유자격 조사자 모형’³⁴) 간에 협상하는 것이 보통이지만 정보를 다운로드하기 전에 합의해야

³³ Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, Vol. 57, p. 1701, 2010

³⁴ K El Emam and B Malin, ‘Appendix B: Concepts and Methods for De-identifying Clinical Trial Data,’ in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC.

하는 사용자 클릭(click-through) 라이선스 합의서로 인터넷에 쉽게 게시할 수 있다('사용자 클릭 모형'³⁵).

- **밀실 모형(the Enclave model)**^{36, 37} : 비식별처리된 개인정보를 원본 정보의 수출(export)을 제한하는 일종의 밀실(enclave)에 유지하고, 대신에 유자격 연구자의 문의를 수락하고 비식별처리된 개인정보에 대한 문의를 운영하며 결과를 응답할 수 있다.

프라이버시와 정보 정책 컨설턴트인 Robert Gellman은 데이터 이용 합의를 보장할 입법안을 제시한 바 있다.³⁸ Gellman은 그가 잠재적으로 식별 가능한 개인 정보(potentially identifiable personal information, 이하 'PI²')라고 부르는 새로운 정보 범주를 제안한다. 이에 동의하는 당사자는 자신들의 데이터 이용 합의서에서, 해당 정보에서 개인 식별자는 제거되었지만 여전히 재식별될 수 있다는 정보 제공자의 약속을 추가할 수 있다. 이후에 접속자는 재식별을 시도할 경우에 민형사상 처벌을 받을 수 있다. 따라서 제안된 입법안에 따르면, 비식별처리된 개인정보가 계속 그러한 상태로 남아있게 되리라는 신뢰가 더해질 것이다. Gellman의 지적에 따르면, '왜냐하면 언제라도 정보를 재식별할 수 있는지를 알 수 없으며, 사실 상 명시적으로 식별 가능한 모든 개인정보를 제외하고는 모두 PI²가 되기 때문이다.'

구조화된 정보를 비식별처리하기 위한 대부분의 접근법은 정보 집합에서 특정한 식별 정보 요소를 지정하고 제거하려 한다. 이 절에서는 그러한 접근법에 사용하는 용어를 소개하고 HIPAA 프라이버시 규칙에 있는 두 가지 비식별처리 기준을 논의하며, 이러한 기법에 대한 비판과 학계의 문헌에 나타난 노력에 관하여 논의한다.

3.1 직접 식별자 제거

많은 비식별처리 접근법은 각각의 열에 다른 개인의 개인정보가 담긴 단일한 정보 표가 있는 정보 집합에 적용했을 때 가장 쉽게 이해할 수 있다. 표 1에는 가상적인 정보 집합이 제시되어

2015

³⁵ Ibid.

³⁶ Ibid.

³⁷ O'Keefe, C. M. and Chipperfield, J. O. (2013), A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. *International Statistical Review*, 81: 426-455. doi: 10.1111/insr.12021

³⁸ Gellman, Robert; The Deidentification Dilemma: A Legislative and Contractual Proposal, 21 *Fordham Intellectual Property, Media & Entertainment Law Journal* 33 (2010), <http://ipj.net/blog/wp-content/uploads/2013/09/Deidentificatin-Dilemma.pdf>.

있다.

직접 식별 변수(*directly identifying variable*) 및 직접 식별 정보(*direct identifying data*)라고도 부르는 **직접 식별자(*direct identifier*)**는 “개인을 직접적으로 식별하는 정보이다.”³⁹ 직접 식별자의 예에는 성명, 사회보장번호, 이메일 주소 등이 있다.

메모 [한14]: 일반적인 정보를 의미

ISO/TS 25237:2008(E)에서는 직접 식별자를 “추가적인 정보 없이, 또는 공개 영역에 있는 다른 정보를 통해 교차 연결함으로써 사람을 식별하는데 사용할 수 있는 정보”라고 정의하고 있다.⁴⁰ 그러나 정보를 신원과 연결하기 위해서는 비록 추가적인 정보가 필요하기는 하지만, 의료보험번호와 전화번호와 같은 개인적인 다른 정보를 직접 식별자로 간주하는 것이 좋다. 왜냐하면 이러한 형태의 식별은 광범위하게 사용되고 있으며, 따라서 신원과 연결하는데 사용할 수 있기 때문이다.

HIPAA 프라이버시 규칙은 직접 식별자의 정의를 확대하여, 성명, 전화번호, 이메일 주소, 기타 고유 식별번호, 특징 또는 코드 등의 특정한 개인정보를 포괄하고 있다. 전체 목록은 제 3.5 절에 제시되어 있다.

비식별처리 중에 직접 식별자는 제거하거나 달리 변환해야 한다. 다음과 같은 접근법이 있다.

- 직접 식별자를 제거할 수 있다.
- 직접 식별자를 명백히 일반적인 범주 명칭이나 정보로 대체할 수 있다. 예를 들면, 성명은 ‘성명’, 주소는 ‘거리 명 123 호, 타운 명, 미국’ 등으로 대체할 수 있다.
- 직접 식별자를 ‘*****’ 또는 ‘XXXXX’와 같은 기호로 대체할 수 있다.
- 직접 식별자는 임의 값으로 대체할 수 있다. 만약 같은 신원이 두 번 등장하면, 두 개의 다른 값을 받는다. 이렇게 하여 원본 개인정보의 형태를 보존하고 일종의 시험을 할 수 있게 되지만, 개인정보와 개인을 다시 연계시키기는 더 어려워진다.
- 직접 식별자를 체계적으로 가명과 대체하여 동일한 개인을 참조하는 개인기록부들이 매칭되도록 한다. 가명처리(*pseudonymization*)에 관해서는 다음 절에서 논의한다.

| | |
|--------|--|
| 직접 식별자 | |
|--------|--|

³⁹ ISO/TS 25237:2008(E), p.3

⁴⁰ ISO/TS 25237:2008(E), p.3

| 성명 | 주소 | 생일 | ZIP | 성별 | 몸무게 | 진단결과 | ... | ... |
|----|----|----|-----|----|-----|------|-----|-----|
| | | | | | | | | |
| | | | | | | | | |

표 1: 직접 식별자를 나타내는 개인정보 표의 예

초기에는 개인정보를 비식별처리할 때 직접 식별자를 제거하는 것에 그쳤다. 그러한 노력의 결과로 나타나는 정보는 연결 공격(linkage attack)을 통해 재식별할 수 있다. 이에 대해서는 아래의 제 3.3 절에서 논의한다.

3.2 가명처리(pseudonymization)

가명처리는 성명과 개인을 직접 식별하는 다른 정보를 가명으로 대체하는 특정한 변환의 일종이다.⁴¹ 직접 식별자들을 모두 체계적으로 가명처리하면, 가명처리를 통해 복수의 개인정보 기록부나 정보 시스템에 걸쳐, 개인에 속하는 정보를 연결할 수 있게 된다.

만약 가명처리를 수행한 기관이 원래의 신원을 가명과 연결하는 표를 보유하고 있거나, 만약 변수를 알거나 발견할 수 있는 알고리즘을 이용하여 대체한 경우에는 가명처리를 손쉽게 되돌릴 수 있다.

가명처리는 미래의 한 시점에서 가명을 되돌려 정보주체를 재식별할 수 있도록 하는 경우가 많다. 만약 직접 식별자와 가명 간의 참조(mapping)가 보존되었거나 이를 재생할 수 있는 경우에는 가명처리된 정보 집합을 되돌릴 수 있다.⁴² 예를 들어, 비밀 키를 이용하여 식별자를 암호화하여 가명을 생성할 수 있다. 키의 암호를 풀면 가명처리 과정이 역으로 되돌려지며, 원본 식별자가 생성된다. HIPAA 에 따라 직접 식별자를 정보주체의 신원으로 되돌리는 작업은 포괄 기관(covered entity)라고 부르는 HIPAA 의 규정에서 다루는 기관(대체로 의료 제공자)만 수행할 수 있다.⁴³

미 보건부 산하 임상연구안전국(HHS Office for Human Research Protection, 이하 'OHRP')은 만약 가명처리를 쉽게 역으로 되돌릴 수 있다면, 해당 개인정보는 '코드화(coded)'된 것으로 간주되며, 생명윤리규정(Common Rule)⁴⁴에 따라 익명처리된

⁴¹ ISO/TS 25237:2008

⁴² HIPAA 는 가명처리를 위해 적절히 솔팅된(salted) 단방향 암호적 해시를 사용하도록 허용하고 있으나, 전문가 활용법의 일부로서 사용할 수 있을 뿐이며, 암호학적 키가 공개되지 말아야 한다. See Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Office of Civil Rights, Health and Human Services, November 26, 2012, pages 21-22.

⁴³ 45 CFR 164.514 (c)

⁴⁴ U.S. Department of Health and Human Services, 'OHRP—Guidance on Research Involving Coded Private Information or Biological Specimens,' October 16, 2008. <http://www.hhs.gov/ohrp/policy/cdebiol.html>

것으로 간주되지 않는다. 이러한 지침은 HIPAA의 포괄 기관 이외의 정보 집합을 이용하여 수행할 수 있으며 연방정부가 자금을 지원하는 연구에 적용된다. 그러나 만약 코드 키를 공유하는 것을 금지하는 데이터 이용 합의서가 존재한다면, 해당 정보는 식별 가능한 개인정보로 간주되지 않는다.

만약 재식별을 금지하는 데이터 이용 합의서가 없이, 가명처리된 정보 집합이 공개되면, 접수자는 가명을 되돌리거나 식별 정보를 이용하여 재식별을 시도할 수 있다. 예를 들면, 2014년 뉴욕 시 택시 리무진 위원회는 2013년에 뉴욕 시에서 탄 1억 7천 3백만 회의 택시 탑승 정보 집합을 공개하였다. 정보 집합을 비식별처리하기 위해 이 위원회는 택시 번호와 운전사 면허번호를 단방향 암호적 해시(one-way cryptographic hash)로 대체하였다. 이 정보 집합의 이용자는 해시 알고리즘을 발견하였고, 모든 가능한 택시 번호와 면허번호를 반복적으로 적용하고 각각의 암호적 해시를 결정하며, 이 해시를 원래의 숫자로 대체함으로써 가명처리를 되돌릴 수 있었다.⁴⁵ 따라서, 비식별처리 시도는 실제로 운전사의 신원을 보호하지 못했다. 이는 폭력적 공격(brute force attack)으로 알려져 있다.

가명을 되돌릴 수 있는 능력은 가명이 임의로 생성되었는가 또는 알고리즘에 의해 생성되었는가(알고리즘에 임의 키가 적용된 경우), 키의 가용성, 가명에 특이성이 있는가(unique) 아니면 재사용되었는가 등의 많은 요소에 따라 달라진다. 임의적으로 생성된 가명은 참조(mapping)가 유지된 경우에만 되돌릴 수 있다. HIPAA 프라이버시 규칙에서는 코드화 식별자인 가명을 공개하는 것을 허용하고 있다. 왜냐하면 가명은 개인에 연계된 정보에서 도출된 것이 아니기 때문이다.

비록 참조가 유지되지 않더라도, 복수의 정보 집합에 걸쳐 일관된 가명을 이용하면 연결 공격을 통해 개인정보를 재식별하기 더 쉬워질 수 있다. 이러한 이유로 인해, 제 29 조 개인정보 보호 작업 그룹(유럽 위원회(the European Commission)의 작업 그룹)에는 '가명처리된 개인정보는 익명처리된 정보와 동일시할 수 없다. 왜냐하면 가명처리된 개인정보는 개별 정보주체를 식별해 내고 다양한 정보 집합에 걸쳐 연결할 수 있도록 하기 때문이다'라고 언급되어 있다.⁴⁶ 그러나 가명처리를 통해 정보와 원래의 신원을 연결할 가능성이 낮아지기 때문에, 의견(opinion)에서는 가명처리를 여전히 '유용한 보안 수단'이라고 언급하고 있다.

⁴⁵ 'Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset,' Anthony Tockar, September 15, 2014, <http://research.neustar.biz/author/atockar/>

⁴⁶ Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working Party, 0829/14/EN WP216, Adopted on 10 April 2014

오랜 기간 동안 사용한 독특한 가명은 점점 더 많은 정보와 연결되기 때문에 프라이버시 리스크가 증가할 수 있다. 이와 유사하게, 오래 사용한 기기 가명(device pseudonym)도 비슷한 프라이버시 리스크를 야기할 수 있다. 특히 다수의 사람들에게 역동적으로 재지정된 가명과 비교할 때 그러하다.

3.3 연결 공격(Linkage Attack)을 통한 재식별

비식별처리된 정보 집합을 재식별할 수 있는 방법에는 연결 공격도 있다. 연결 공격에서는 비식별처리된 정보 집합에 있는 각각의 개인기록부가, 연결 정보와 정보주체의 신원을 모두 담고 있는 두 번째 정보 집합에 있는 유사한 개인기록부와 연결된다.

가장 널리 알려진 연결 공격의 예로는 Latanya Sweeney의 공격을 들 수 있다. Sweeney는 1990 년대에 자신의 MIT 졸업 과제의 일부로, 메사추세츠 주지사인 William Weld의 의료 개인기록부를 재식별하였다. 당시에 메사추세츠 주는 입원한 적이 있는 메사추세츠 주 공무원의 비식별처리된 의료보험 환급 기록을 담고 있는 연구 정보 집합을 배포하고 있었다. 공무원의 프라이버시를 보호하기 위해, 정보 집합에서 성명이 제거되었으나, 공무원의 생일, 우편번호, 성별은 통계 분석을 위해 보존되었다.

메모 [한15]: 일반적인 기록물입니다.

Weld가 최근에 메사추세츠 병원에서 치료받은 적이 있음을 알고 있는 Sweeney는 주지사의 생일, 우편번호, 성별에 매칭되는 '비식별처리된' 개인기록부를 검색하여 주지사의 개인기록부를 재식별할 수 있었다. 그녀는 이 정보를 자신이 20 달러에 구입한 케임브리지 투표자 등록명부에서 얻었다. 더 나아가 Sweeney는 그녀의 발견결과를 일반화하여, 1990년 인구조사를 기초로, 미국인의 최대 87%를 다섯자리 우편번호, 생일, 성별을 이용하여 특이성이 있게(uniqely) 식별할 수 있다고 주장하였다.⁴⁷ 프라이버시 연구자인 Phillip Golle은 후속 연구에서 2000년 인구조사를 이용하여 62%라는 재식별 비율을 계산하였다.⁴⁸ 컬럼비아 대학교의 Daniel C. Barth-Jones 교수는 이 연구를 비판하며, 1996~1997년 사이에 케임브리지 시 인구의 55%만 등록하였으며, 따라서 투표자 명부와 연계시켜 케임브리지 시 인구의 55% 이하만을 재식별할 수 있었다고 지적하였다.⁴⁹

⁴⁷ Sweeney L., Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh, 2000. <http://dataprivacylab.org/projects/identifiability/paper1.pdf>

⁴⁸ Golle, Phillip, Revisiting the Uniqueness of Simple Demographics in the U.S. Population, in proceedings of WEPS 2006.

⁴⁹ Daniel C. Barth-Jones, 'The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now,' SSRN-id2116396, June 4, 2012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

Sweeney의 연결 공격은 다음과 같이 그림으로 나타낼 수 있다.

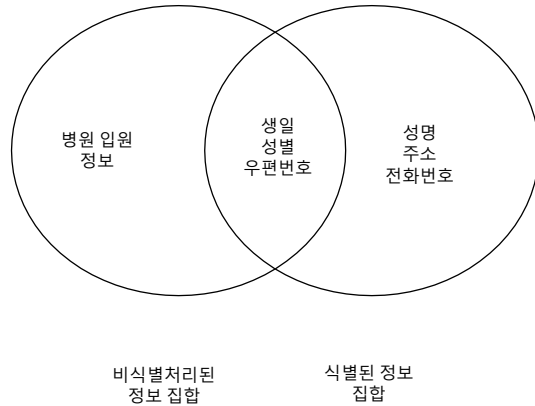


그림 3: 연결 공격은 두 개 이상의 정보 집합의 정보를 조합하여 개인기록부를 재식별

많은 요소들이 그러한 연결 공격을 복잡하게 만든다.

- 연결하기 위해서는 한 사람이 두 정보 집합에 모두 있어야 한다. Sweeney는 Weld가 입원한 적이 있고, 주지사로서 등록된 투표자임을 알고 있었기 때문에 두 정보 집합에 모두 있다는 것을 알았다.
- 두 정보 집합에 모두 있는 변수를 연결함으로써 특이성이 있고 분간되는 개인기록부만 특이성이 있게(unicely) 연결할 수 있다. 그렇지 않을 경우에는 연결에 확률론적 관점을 취해야 한다(예, 만약 두 가지 가능한 매칭이 있다면 정확하게 매칭될 확률은 0.5이다). 특이성이 있게 매칭되는 경우에는 한 개인의 개인기록부는 만약 같은 생일, 성별, 우편번호를 가진 사람이 두 정보 집합에 없어야만 연결할 수 있다. 결과에 나타난 바와 같이, 케임브리지 시의 투표자 명부에 등록된 사람 중에 Weld와 생일이 같은 사람은 없었다. 그러나 만약 Weld의 개인기록부를 현재의 HIPAA 기준에 따라 비식별처리하였다면 개인기록부에는 그의 출생연도와 우편번호의 앞 세 자리만 포함되어 있었을 것이다. 이러한 조건에서는 특이성이 있는(unique) 매칭은 아마도 불가능했을 것이다.
- 두 정보 집합에서 변수가 같지 않다면 연결하기 위해서는 해당 정보를 정규화하거나 달리 일관되도록 해야 한다. 이러한 정규화(normalization) 과정에 오류가 발생할 수 있다. 이는 Weld의 사례에서는 문제가 되지 않았지만, 만약 한 정보 집합이 '연령'을 보고하고 다른 정보 집합이 '생일'을 보고했다면 문제가 될 수 있다.

- 어떤 사람은 정보 집합에 특이성이 있게 나타나지만, 집단에서는 특이성이 없을 수 있다. 즉, Weld 와 생일이 같은 다른 사람이 케임브리지 시에 있었을 수 있지만 그 사람이 투표에 등록하지 않았다면 분석에서 그 사람을 누락했을 것이다.
- 어떤 연결이 정확한지를 검증하기 위해서는 연결 작업의 일부로 사용하지 않은 정보를 이용해야 한다. 이 경우, Weld 의 의료 개인기록부는 Weld 의 입원에 관한 신문기사를 이용하여 검증할 수 있었다.

데이터 이용 합의서와 사용자 클릭 라이선스 합의서를 통해 관리 측면에서 재식별을 금지할 수 있다. 그러나 이러한 합의서는 자유롭게 공개된(예를 들면 인터넷에 공개) 비식별처리 정보 집합에는 적용하기 어려울 수 있다.

3.4 준식별자의 비식별처리

간접 식별자(*indirect identifier*) 또는 간접 식별변수(*indirectly identifying variable*)라고도 부르는 준식별자(*quasi-identifier*)는 자체로서는 특정한 개인을 식별하지 않지만 다른 정보와 조합하거나 '연결'하여 정보주체를 식별할 수 있다.⁵⁰ Sweeney 는 William Weld 의 의료 개인기록부를 재식별할 수 있었는데, 이는 생일, 우편번호, 성별이 준식별자이기 때문이다.

| 직접 식별자 | | 준식별자 | | | | | | |
|--------|----|------|------|----|-----|------|-----|-----|
| 성명 | 주소 | 생일 | 우편번호 | 성별 | 몸무게 | 진단결과 | ... | ... |
| | | | | | | | | |
| | | | | | | | | |

표 2: 직접 식별자와 준식별자를 보여주는 개인정보 표의 예

준식별자는 비식별처리에 중대한 문제를 야기한다. 직접 식별자는 정보 집합에서 제거할 수 있는데 비해, 준식별자는 일반적으로 나중의 분석에서 중요할 수 있는 정보를 담고 있으며, 준식별자를 제거하면 정보 집합의 효용성이 줄어들 수 있다. 그렇기 때문에 재식별 리스크와 준식별자를 포함하여 얻는 유용성 간에 균형을 잡기 위한 세심한 고려가 필요하다.

준식별자를 비식별처리하기 위해 몇 가지 방법을 이용한다.

- 1) 범주화(suppression):** 준식별자를 범주화하거나 제거할 수 있다. 정보를 제거하면 프라이버시 보호가 극대화되지만 정보 집합의 효용이 감소할 수 있다.

⁵⁰ Dalenius, Finding a Needle in a Haystack, or Identifying Anonymous Census Records, *Journal of Official Statistics* 2:3, 329- 336, 1986.

- 2) **일반화(generalization):** 특정한 준식별자 값이 주어진 범위 안에 있다고 보고하거나 어떤 세트의 한 요소로서 보고할 수 있다. 예를 들면, 우편번호 12345 를 12000 과 12999 사이의 우편번호로 일반화할 수 있다. 일반화는 예를 들면 극단치(outlier)를 식별하여, 전체 정보 집합 또는 특정한 개인기록부에 적용할 수 있다.
- 3) **인자변환(perturbation):** 규정된 일반화 수준 이내에서, 각 개인에 대해 일관된 방식으로 특정한 값을 다른 값으로 교체할 수 있다. 예를 들면, 모든 연령을 원래 연령의 (-2 ... 2)년으로 임의로 조정하거나 또는 입원 또는 퇴원 일자를 체계적으로 같은 일자로 (-1000 ... 1000)일을 이동시킬 수 있다.⁵¹
- 4) **교환(swapping):** 규정된 일반화 수준 이내에서 개인기록부 간에 준식별자 값을 교환할 수 있다. 교환은 통계적 특징을 유지할 필요가 있을 경우에는 주의하여 다루어야 한다.
- 5) **하위샘플링(sub-sampling):** 비식별처리를 수행하는 기관은 전체 정보 집합을 공개하는 대신에 샘플을 공개할 수 있다. 만약 하위 샘플만을 공개한다면 재식별 가능성은 낮아진다.⁵²

k-익명성⁵³은 주어진 프라이버시 수준을 달성하기 위해 준식별자에 필요한 조작 횟수를 정량화하기 위해 Sweeney 가 개발한 기법이다. 이 기법은 모든 준식별자 값에 매칭되는 개인기록부 세트인 동질집합(*equivalence class*)이라는 개념을 기반으로 하고 있다. 준식별자로 이루어진 모든 조합에 대해 최소한 k 개의 매칭되는 개인기록부가 있는 경우, 해당 정보 집합은 k-익명성이라고 말한다. 예를 들면, 만약 준식별자인 출생연도와 주(state)가 있는 정보 집합이 k=4 익명성이 있다면, 모든 (출생연도, 주) 조합에 대해 최소한 4 개의 개인기록부가 있다. 이후의 연구를 통해 각각의 동질집합 내의 민감한 속성의 다양성(다양성)⁵⁴에 대한 요건을 추가하고, 결과 정보가 원본 정보에 통계적으로 근접할 것을 요구하여(t-근접성)⁵⁵ k-익명성을 정교화하였다.

⁵¹ Office of Civil Rights, 'Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule', US Department of Health and Human Services, 2010. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>

⁵² El Emam, Methods for the de-identification of electronic health records for genomic research, *Genome Medicine* 2011, 3:25 <http://genomemedicine.com/content/3/4/25>

⁵³ Latanya Sweeney. 2002. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5

⁵⁴ A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian. l-diversity: Privacy beyond k-anonymity. In *Proc. 22nd Intl. Conf. Data Engg. (ICDE)*, page 24, 2006.

⁵⁵ Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). 't-Closeness: Privacy beyond k-anonymity and l-diversity'. *ICDE* (Purdue University).

Khaled El Emam 교수와 Bradley Malin 교수⁵⁶는 식별자와 준식별자를 분류하여 11 단계로 된 개인정보의 비식별처리를 위한 과정을 개발하였다.

- **1 단계: 정보 집합에서 직접 식별자 결정.** 전문가가 정보 집합에서 정보주체를 식별하는 데만 사용하는 요소를 결정한다.
- **2 단계: 직접 식별자 마스킹(변환).** 직접 식별자를 제거하거나 가명으로 대체한다.
- **3 단계: 위협 모형화.** 기관은 '가상의 적', 그들이 재식별을 위해 사용할 수 있는 추가적인 정보, 적들이 재식별을 위해 사용할 수 있는 준식별자를 결정한다.
- **4 단계: 최소 허용 정보 효용을 결정.** 이 단계에서 기관은 각각의 항목에 대해 일어날 수 있는 최대의 비식별처리의 양을 결정하기 위해 비식별처리된 개인정보가 이용될 수 있는 용도를 결정한다.
- **5 단계: 재식별 리스크의 한계치 결정.** 기관은 선례와 기준(예, 작업 문서 22: 통계적 공개 통제에 관한 보고서)을 이용하여 정보에 대한 작업과 통제 완화 가능성에 관련된 허용 가능한 리스크를 결정한다.
- **6 단계: 데이터베이스에서 (샘플) 정보를 수입(import).** (식별된) 소스 데이터베이스에서 정보를 획득하기 위한 노력이 클 수 있기 때문에, 연구자들은 계획에 도움을 주기 위해 시험적인 정보 수입을 권고한다.
- **7 단계: 실제 재식별 리스크를 평가.** 실제의 식별화 리스크를 계산한다.
- **8 단계: 실제 리스크를 한계치와 비교.** 5 단계와 7 단계의 결과를 비교한다.
- **9 단계: 변수를 설정하고 정보 변환을 적용.** 실제 리스크가 허용 가능한 수준이라면, 비식별처리 변수를 적용하고 정보를 변환한다. 리스크가 너무 큰 경우, 새로운 변수 또는 변환을 고려해야 한다.
- **10 단계: 해법에 대한 진단 수행.** 비식별처리된 개인정보가 충분한 효용성을 갖고

⁵⁶ K. El Emam and B. Malin, 'Appendix B: Concepts and Methods for De-identifying Clinical Trial Data,' in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC, 2015

있으며, 허용 가능한 변수 이내에서 재식별이 불가능하도록 하기 위해, 비식별처리된 개인정보를 분석한다.

- **11 단계: 변환된 정보를 외부 정보 집합으로 수출(export).** 마지막으로, 비식별처리된 개인정보를 수출하고, 비식별처리 기법을 서면 보고서에 기록한다.

직접 식별자와 준식별자를 기초로 한 비식별처리에 대한 주된 비판은 준식별자를 관리적으로 결정할 때에는, 비식별처리를 수행할 당시에 없었지만 나중에 존재하게 된 정보 등의 외부 정보와 조합되거나 연결되었을 때 특이성이 있게 식별하는 변수를 누락할 수 있다는 점이다.

메모 [한16]: 일반적인 정보를 의미합니다.

현실적인 입장을 취하고 발생 가능성이 높은 공격을 고려하는 것이 중요하다. 재식별, 다른 정보와의 연결, 미인가 공유를 금지하는 데이터 이용 합의서가 있는 경우에는 특히 그러하다. 식별자의 선정에 대한 지침이 되는 기준 및 허용 가능한 리스크 수준에 관한 선례 이외에도, 평가 또는 재식별 리스크는 적이 현실적으로 알 수 있는 정보의 양('공격자의 힘')으로 한정될 수 있다.⁵⁷

3.5 HIPAA 에 따른 개인보건의정보(Protected Health Information, 이하 'PHI')의 비식별처리
1996 년 HIPAA 프라이버시 규칙에는 개인보건의정보(PHI)의 비식별처리에 관하여, 전문가 활용법(Expert Determination Method, §164.514(b)(1))과 세이프하버 방식(the Safe Harbor method, §164.514(b)(2)) 등 두 가지 접근법이 제시되어 있다. 두 방식 모두 재식별 리스크가 전혀 없는 안전한(foolproof) 비식별처리 방식은 아니다. 단, 이 방법들은 재식별 리스크를 낮추면서 비식별처리된 의료 정보를 생산하고 공유할 수 있도록 하기 위한 실용적인 접근법을 취하고 있다.

3.5.1 HIPAA 전문가 활용법

'전문가 활용법(the Expert Determination method)'은 개인정보를 검토하는 전문가를 제공하며, 재식별 리스크를 최소화하는 적절한 비식별처리 수단을 정한다. 프라이버시 규칙에는 다음과 같은 특별한 조항이 있다.

“(1) 개별적으로 식별할 수 없는 정보의 제작에 관해 일반적으로 인정되는 통계학 및 과학적 원리와 방법에 관한 적절한 지식과 경험을 갖춘 사람”

(i) 그러한 원칙과 방법을 적용하여, 예상되는 정보 접수자가 그러한 정보를 이용하거나

⁵⁷ Yabo Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. 2008. Anonymizing transaction databases for publication. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08). ACM, New York, NY, USA, 767-775. DOI=10.1145/1401890.1401982 <http://doi.acm.org/10.1145/1401890.1401982>

또는 합리적으로 이용할 수 있는 다른 정보를 이용하여, 정보주체인 개인을 식별하기 위해 해당 정보를 이용할 수 있는 리스크가 매우 적다고 결정한다.

(ii) 그러한 결정을 뒷받침하는 방법과 분석 결과를 문서로 기록한다.’

제 3.4 절에 제시된 TI Emam 과 Malin 방법론은 전문가 활용법의 예라고 할 수 있다.

본 절과는 별도로, 프라이버시 규칙이나 미국 보건복지부(the Department of Health and Human Services, 이하 ‘HHS’) 시민권실(Office of Civil Rights)⁵⁸은 전문가에 관한 기준이나 자격을 명시하고 있지 않으며, 전문가를 이용하는 기관의 공개 요건이나 심지어 전문가의 결정이 내려졌는지를 인정하기 위한 요건도 명시하고 있지 않다. 또한 재식별 리스크의 계산 또는 정량화 방법도 명시되어 있지 않으며, 허용 가능한 최소한의 재식별 리스크가 ‘매우 적어야’ 한다는 점 외에는 다른 규정이 없다.

전문가 활용법에는 전문가가 “일반적으로 인정되는 통계적, 과학적 원리와 방법”을 알고 적용해야 한다고 명시하고 있으며, 이는 통계적 공개 통제와 비식별처리 방법에 관한 문헌을 알고 있어야 함을 의미할 것이다. 비식별처리와 리스크 수준에 관한 강력한 선례도 존재하고 있으며,⁵⁹ 효과적인 비식별처리 방법론에 적용되어야 한다.

3.5.2 HIPAA 세이프하버 방식

‘세이프하버(Safe Harbor)’ 방식은 포괄 기관이 ‘개인, 친척, 직원 또는 개인의 가족’에 관한 18 가지 특정 유형의 개인정보를 삭제함으로써 개인정보를 비식별처리된 것으로 다루도록 한다. 18 가지 유형은 다음과 같다.

“(A) 성명

(B) 주(state) 이하의 모든 지리적 행정구역으로서, 거리 주소(street address), 시(city), 카운티(county), 지구(precinct), 우편번호(ZIP code) 및 그와 동등한 지오코드(geocode)를 포함하나, 미국 통계국이 현재 공개한 우편번호의 앞 세 자리는 제외한다.

(1) 앞자리 세 개가 동일한 모든 우편번호를 조합하여 형성된 지리적 단위에는 20,000 명 이상이 포함된다.

(2) 20,000 명 내외의 사람이 포함된 그러한 모든 지리적 단위의 우편번호 앞

⁵⁸ Office of Civil Rights, ‘Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule’, US Department of Health and Human Services, 2013. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>

⁵⁹ 예, Working Paper 22: Report on Statistical Disclosure Control.

세 자리를 000 으로 변경한다.

(C) 생일, 입원일, 퇴원일, 사망일, 89 세 이상의 모든 연령 등, 개인과 직접적으로 관련된 모든 날짜의 모든 날짜 요소(년은 제외) 및 그러한 연령을 나타내는 날짜의 모든 요소(년 포함), 단, 단일한 90 세 이상의 범주로 종합할 수 있는 연령 및 요소는 제외한다.

(D) 전화 번호

(E) 팩스 번호

(F) 이메일 주소

(G) 사회보장번호

(H) 의료기록번호

(I) 의료보험(Health Plan) 수혜자 번호

(J) 계좌번호

(K) 자격증/면허 번호

(L) 차량 식별자 및 일련번호, 차량 번호판 번호 포함

(M) 기기 식별자 및 일련번호

(N) 웹 URL(Universal Resource Locator)

(O) 인터넷 프로토콜(IP) 주소

(P) 지문이나 성문(voiceprint)을 포함한 생체인증 식별자

(Q) 얼굴 전체의 사진 및 그와 유사한 이미지

(R) 기타 특이한(unique) 식별 번호, 특징 또는 코드

세이프하버 방식은 전문가 활용법에 비해 규칙을 직접적으로 적용할 수 있고 과정을 반복할 수 있으며, 합법적으로 비식별처리된 정보 집합을 제공한다.

어떠한 방법을 사용하더라도, 비식별처리를 수행하는 포괄 기관은 “정보를 단독으로 사용하거나, 다른 정보와 함께 사용하여 정보주체인 개인을 식별할 수 있다는 실제적인 지식을 갖고 있지” 않아야 한다.⁶⁰ 그럼에도 불구하고 HHS 는 단순히 재식별 기법의 존재를 알고 있는 것은 ‘실제적인 지식’ 기준에 부합되지 않는다는 특별 지침을 내렸다.

Q: ‘3.7: 만약 포괄 기관이 보건 정보를 재식별하는 방법에 관한 특정한 연구를 알고 있거나 비식별처리된 보건 정보를 단독으로 또는 다른 정보와 조합하여 사용함으로써 개인을 식별할 경우, 이는 필연적으로 해당 포괄 기관이 세이프하버 방식에 따른

⁶⁰ §164.514 (c)

실제적인 지식을 갖고 있음을 의미하는가?

A: '아니다. 특정한 분석 능력과 정량화 능력을 가진 연구자들이 정보를 특정한 방식으로 조합하여 건강 정보를 식별하는 능력에 관한 많은 문헌이 있다.^{61 62 63 64} 포괄 기관은 나머지 정보를 식별하기 위한 방법에 관한 연구나 비식별처리된 정보를 단독으로 또는 다른 정보와 조합하여 사용함으로써 개인을 식별하는 연구에 관하여 알고 있을 수 있다.

그러나 그러한 연구와 방법에 대한 포괄 기관의 단순한 지식은 그 자체로서는 그러한 기관이 공개하는 개인정보에 그러한 방법이 이용될 것이라는 '실제적인 지식'을 갖고 있음을 의미하지 않는다. OCR 은 비식별처리된 개인정보를 받는 모든 잠재적인 접수자들이 그러한 능력을 갖고 있다고 포괄 기관이 간주할 것이라고 예상하지 않는다. 이는 어떤 정보가 적절히 비식별처리되었는지 결정하기 위한 단순한 방법을 포괄 기관에게 제공하려는 세이프하버 방식의 의도에 부합되지 않을 것이다.⁶⁵

3.5.3 HIPAA 세이프하버 방식의 효과 평가

HIPAA 세이프하버 방식은 Sweeney 의 연구에 많은 영향을 받았다. 세이프하버 방식은 Sweeney 의 연구를 언급하고 있고, 그녀가 일반화를 위해 식별한 준식별자에 특히 주목하고 있다(우편번호와 생일). 예를 들면, 우편번호 첫 세 자리와 출생연도에 대한 보고를 허용하는 등, 이 방법은 재식별 리스크와 정보 집합의 효용성의 유지 사이에 균형을 이루려는 것으로 보인다.

의료 개인기록부의 비식별처리와 비식별처리된 개인정보의 재식별 리스크의 측면에서 HIPAA 세이프하버 방식의 효과에 관한 논란이 있다. Sweeney 는 미국 전역의 특이성이 있는 재식별 리스크를 0.04%로 예측하였는데, 이는 10,000 개의 개인기록부 가운데 4 개를

⁶¹ K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk. Evaluating the risk of re-identification of patients from hospital prescription records. *Canadian Journal of Hospital Pharmacy*. 2009; 62(4): 307-319.

⁶² G. Loukides, J. Denny, and B. Malin. The disclosure of diagnosis codes can breach research participants privacy. *Journal of the American Medical Informatics Association Annual Symposium*. 2010; 17(3): 322-327

⁶³ B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*. 2004; 37(3): 179-192.

⁶⁴ L. Sweeney. Data sharing under HIPAA: 12 years later. A presentation at the Workshop on the HIPAA Privacy Rule's De- Identification Standard. Washington, DC. March 8-9, 2010.

⁶⁵ Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, US Department of Health and Human Services, Office for Civil Rights, 2010. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/guidance.html#_edn32

특이성이 있게 재식별할 수 있음을 의미한다.⁶⁶ 반더빌트 건강 정보 프라이버시 연구소(Vanderbilt Health Information Privacy Lab)의 Kathleen Benitez 와 Bradley Malin 은 주(state)별로 연구를 수행하여 특이성이 있는 재식별 리스크는 0.01%와 0.25% 사이라고 결론지었다.⁶⁷ 그러나 이는 특이성이 있는 식별정보가 출생연도, 성별, 3 자리 우편번호인 개인정보로 한정된다. Sweeney 는 워싱턴 주가 판매한 건강 개인정보를 공격할 때, 세이프하버 방식에 언급된 준식별자를 제외한 다른 준식별자로 인한 리스크도 있음을 확인하였다.

미국 보건복지부의 국가의료정보기술 조정관실(Office of the National Coordinator for Health Information Technology, ONC HIT)은 2010 년에 HIPAA 세이프하버 방식에 대한 검사를 수행하였다. 이 연구의 일환으로, 의료 시스템에서 나온 2004~2009 년 간 히스패닉계 시민의 병원 입원 개인기록부 15,000 건을 연구자들에게 제공하였다. 연구자들은 비식별처리된 개인기록부와 미국인 2 억 3 천 5 백만 명의 정보를 갖고 있다고 주장하는 회사인 InfoUSA 에서 구입한 30,000 건의 개인기록부의 매칭을 시도하였다.⁶⁸ 미국 인구조사 자료에 근거하여, 연구자들은 판매된 개인기록부 30,000 건이 대략 병원 환자 5,000 명을 포괄하고 있다고 추정하였다. 실험자들이 성별, ZIP3(HIPAA 에 따라 허용된 대로, 우편번호의 앞 세 자리), 연령을 이용하여 매칭했을 때, 병원 자료에서 216 건의 특이성이 있는 개인기록부를 발견하였고, InfoUSA 의 자료에서 84 건의 특이성이 있는 개인기록부를 발견하였으며, 두 자료에서 모두 매칭된 개인기록부는 20 건이었다. 이어서 연구자들은 이 20 건의 매치를 검토하여 20 개 중 2 개만 같은 성(姓), 거리 주소, 전화번호를 갖고 있음을 확인하였다.⁶⁹ 이는 특이성이 있는 재식별 비율이 0.013%임을 나타낸다. 또한 연구자들은 더 보수적인 방법을 이용하여 특이성이 있는 재식별 위험을 0.22%로 계산하였다. 이러한 비율들은 단일한 의료보험 시스템에 속한 단일한 민족 집단을 이용한 것이기 때문에 국가 전체 평균이 아니다.⁷⁰

메모 [한17]: 원문은 data 이지만 문맥 상 '자료'로 번역하였습니다.

⁶⁶ Latanya Sweeney, Testimony before that National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007. As quoted in Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Office of Civil Rights, Health and Human Services, November 26, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

⁶⁷ Kathleen Benitez and Bradley Malin, Evaluating re-identification risks with respect to the HIPAA privacy rule,' J. Am Med Inform Assoc. 2010; 17:169-177.

⁶⁸ <http://www.infousa.com>, accessed June 4, 2015

⁶⁹ 연구자들의 '실제 매칭'은 성만을 고려하였기 때문에 매칭된 개인기록부는 동일인이나 동거하는 가족을 나타낼 수 있다.

⁷⁰ Peter K. Kwok and Deborah Lafky, 'Harder Than You Think: A Case Study of Re-identification Risk of HIPAA-Compliant Records,' Joint Statistical Meeting, August 2, 2011 (초록만).

3.5.4 HIPAA 제한된 정보 집합

개인보건정보를 담고 있는 정보와 비식별처리된 개인정보에 추가하여, HIPAA 프라이버시 규칙은 제한된 정보 집합(limited dataset)이라는 세 번째 유형의 정보를 인정하고 있다. 프라이버시 규칙에 따르면, 제한된 정보 집합은 부분적으로 비식별처리되었지만, 여전히 날짜, 시, 주, 우편번호, 연령을 포함하고 있는 정보 집합이다. 그러한 정보는 개인보건정보로 간주되지만, 해당 개인정보를 공유하는 기관이 데이터 이용 합의를 집행할 경우, 연구, 공중 보건, 의료보험 운영을 위해 공유할 수 있다. 최저 기준으로서, 데이터 이용 합의서는 보안 안전조치(safeguards)를 요구하고 해당 개인정보를 이용하는 모든 사람이 유사한 제한을 받을 것을 요구하며, 재식별 및 다른 정보와의 무단 연결, 정보주체와의 접촉을 금지해야 한다.⁷¹

메모 [한18]: 일반적인 정보를 의미합니다

메모 [한19]: 일반적인 정보를 의미합니다

Benitez와 Malin은 제한된 정보 집합의 재식별 리스크가 10~60%인 것으로 확인함으로써,^{72 73} 제한된 정보 집합은 적절히 비식별처리된 것으로 간주하지 말아야 한다는 점을 다시 한 번 보여주고 있다.

3.6 현장 비식별처리에 대한 평가

비식별처리에서 정보 집합의 일부 개인정보란에 식별될 가능성이 없는 유용한 정보가 들어있다는 것을 가정한다.

최근 몇 년에 발표된 많은 학술 연구에서는 많은 개인정보 항목이 식별정보일 수 있으며, 정보주체의 신원이 있는 적절한 정보에 대한 접근권(access)이 있고, 재식별이나 연결에 관한 금지가 없는 고차원 정보에서 개인을 식별할 수 있는 경우가 많다는 사실을 보여주었다.

- **넷플릭스 상(the Netflix Prize):** 2008년에 Narayanan과 Shmatikov는 사람들이 본 영화를 식별자로 이용할 수 있는 경우가 많다는 점을 보여주었다.⁷⁴ 넷플릭스는 일부 고객이 관람한 영화로 된 정보 집합을 공개하였고, 자사의 넷플릭스 상 경연의 일부로서 순위를 매겼다. 비록 정보 집합에는 직접 식별자가 없었지만, 연구자들은 관람한 영화 세트(특히 인기가 적은 컬트 고전과 외국 영화)를 이용하여, 넷플릭스 정보 집합의 사용자 프로필과 비식별처리가 되지 않았고, 실명이 많이 포함된 사용자 성명이 있는 인터넷영화 데이터베이스(the Internet Movie Data Base, IMDB)에 있는 단일한

⁷¹ http://privacyruleandresearch.nih.gov/pr_08.asp

⁷² Deborah Lafky, The Safe Harbor Method of De-Identification: An Empirical Test, Department of Health and Human Services, Office of the National Coordinator for Health Information Technology, October 8, 2009. http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf

⁷³ Benitez and Malin, *ibid.*

⁷⁴ Narayanan, Arvind and Shmatikov Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125

사용자 프로필을 매칭시킬 수 있는 경우가 많다는 점을 보여주었다. 위협 시나리오에 따르면, 사람들의 IMDB 프로필을 넷플릭스 상 정보와 연결할 수 있기 때문에, IMDB 에 몇 가지 영화의 순위를 매김으로써, 의도치 않게 자신이 관람한 모든 영화를 공개할 수 있다는 것이다.

- 의료 검사:** 2013 년에 Atreya 등은 환자의 실험 결과 5~7 개를 “비식별처리된 생물의학 연구 데이터베이스에 있는 상응하는 개인기록부를 확인할 수 있는 검색 키”로 활용할 수 있다는 점을 보여주었다.⁷⁵ 환자 61,280 명의 비식별처리된 검사 결과 850 만 건으로 된 반더빌트 대학의 정보 집합을 이용하여, 연구자들은 검사에 따라, 정보 집합에 있는 연속적인 실험실 검사 결과 4 개의 34%~100%를 특이성 있게 분간할 수 있다는 점을 발견하였다. 가장 흔한 검사 결과인 CHEM7 과 CBC 는 각각 검사 주체의 98.9%와 98.8%를 분간하였다. 위협 시나리오에 따르면, CHEM7 또는 CBC 검사 결과를 담고 있는 단일한 실험실이 식별한 보고서를 중간에서 가로채는 경우(아마도 쓰레기통 같은 곳에서), 이를 이용하여 해당 개인에 속하는 다른 개인기록부를 비식별처리된 생물의료 연구 데이터베이스에서 검색할 수 있다.
- 신용카드 거래:** Montjoye 등은 무명의 국가에서 나온 샘플 110 만 명의 비식별처리된 신용카드 거래 수집자료에서, 네 개의 구분되는 공간 및 시간적 지점을 이용하면 샘플 안의 개인의 90%를 특이성 있게 식별하는데 충분하다는 점을 보여주었다.⁷⁶ 지리적 정밀도를 낮추고 거래액을 버리면(예, 구매액 14.86 달러를 10.00 달러와 19.99 달러 사이로 보고함), 필요한 지점의 수가 늘었다.
- 이동 궤적:** Montjoye 등은 사람과 차량을 ‘이동 궤적(mobility traces)’(사람이나 차량이 방문한 장소와 시간 기록)을 이용하여 식별할 수 있다는 점을 보여주었다. 그들은 연구에서 150 만 명으로 이루어진 샘플의 궤적 정보를 처리하였고, 시간 값은 시(hour)로 일반화하고, 공간 정보는 핸드폰 시스템이 제공한 해상도로 일반화하였다(통상적으로 10~20 개 도시 블록). 연구자들은 특정한 장소와 시각에 매칭시킨 개인에 대한 임의로 선정한 관찰결과 4 개를 이용하여 정보주체의 95%를 특이성 있게 식별할 수 있다는 점을 발견하였다.⁷⁷ 개인의 위치/시간 지점은 신용카드

메모 [한20]: 일반적인 기록물입니다.

⁷⁵ Atreya, Ravi V, Joshua C Smith, Allison B McCoy, Bradley Malin and Randolph A Miller, 'Reducing patient re-identification risk for laboratory results within research datasets,' J Am Med Inform Assoc 2013;20:95-101. doi:10.1136/amiajnl-2012-001026.

⁷⁶ Yves-Alexandre de Montjoye et al., Unique in the shopping mall: On the reidentifiability of credit card metadata, Science, 30 January 2015, Vol 347, Issue 6221

⁷⁷ Yves-Alexandre de Montjoye et al., Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3 (2013), Article 1376.

구매, 사진, 인터넷 등 다양한 출처에서 수집할 수 있다. Ma 등이 수행한 유사한 연구에 따르면, 10 개의 주변적 정보를 이용하여 개인을 30~50% 식별할 수 있음을 보여주었다.⁷⁸ 위협 시나리오에 따르면, 공격자는 공개된 정보 집합에서 장소/시간 쌍 5 개를 노출한 개인(예를 들면, 1 개월 동안 집과 직장에서 이메일을 4 번 발송)의 전체 이동 궤적을 식별할 수 있을 것이다. 위와 마찬가지로, 공격자는 목표물이 정보 안에 있었음을 알아야 할 것이다.

메모 [한21]: 일반적인 정보를 의미합니다.

- 택시 승차 정보:** 2014 년 뉴욕 시 택시 리무진 위원회는 2013 년의 모든 뉴욕 시 택시 승차 기록(총 1 억 7 천 3 백만 건)이 담긴 정보 집합을 공개하였다. 이 정보에는 택시 기사나 탑승자의 성명이 포함되어 있지 않았지만, 택시의 번호판 번호로 쉽게 변환할 수 있는 32 자리로 된 알파벳숫자 코드가 포함되어 있었다.⁷⁹ Neustar 사의 한 인턴 정보 과학자는 택시 번호판을 분명히 볼 수 있는 택시에 타고 내리는 유명인사가 담긴 시간이 찍혀 있는 사진을 인터넷에서 찾아낼 수 있음을 알게 되었다.⁸⁰ 그 인턴은 이 정보를 이용하여 1 억 7 천 3 백만 건의 택시 승차 중 2 건의 택시 도착지, 택시요금, 팁 등을 알 수 있었다. Gawker 웹사이트의 한 리포터는 9 개를 더 식별할 수 있었다.⁸¹

메모 [한22]: 일반적인 정보를 의미합니다.

메모 [한23]: 일반적인 정보를 의미합니다.

택시 정보에 관한 경험에 따르면, 개인정보 기록에 있는 다른 정보와 연관될 수 있는 예상치 못한 정보 출처가 많다는 점을 보여주며, 직접 식별자를 변환할 때 매우 조심해야 한다는 점을 보여주고 있다. 택시와 이동 궤적에 대한 연구에 따르면, 지리공간적 정보의 식별 능력이 매우 크다는 점을 알 수 있다. 개인의 위치와 시간에 대한 몇 가지 관찰 정보만 있다면, 비록 일반화되고 잡음이 있는 정보 집합 안의 정보라고 해도, 식별 가능성이 매우 높아질 수 있다. 또한 일부 위치는 독립된 위치이거나 사진정보가 많기 때문에 매우 식별력이 크다.

이러한 유형의 연구에 대한 비판은, 샘플 안에 있는 특이성(uniqueness)과 집단 안에 있는 특이성을 구분하지 못하는 경우가 많다는 점이다. 예를 들면, Montjoye 등의 연구에 따르면, 지리공간적 지점 4 개에서 개인을 분간할 수 있다는 연구결과는 만약 목표로 한 개인이 샘플 안에 있다는 추가적인 정보를 공격자가 갖고 있을 경우에만 적용된다. 만약 샘플 크기가 전체

⁷⁸ Ma, C.Y.T.; Yau, D.K.Y.; Yip, N.K.; Rao, N.S.V., 'Privacy Vulnerability of Published Anonymous Mobility Traces,' Networking, IEEE/ACM Transactions on, vol.21, no.3, pp.720,733, June 2013

⁷⁹ 32 자리 코드는 택시 번호판의 MD5 암호학적 해시였다. MD5 알고리즘은 뒤바뀔(invert) 수 없지만, 가능한 택시 번호판의 수가 매우 적었기 때문에, 직접 폭력적 공격을 이용하고 모두 해시하였다. 만약 택시 리무진 위원회가 키로 된 해시를 이용하고 키를 공개하지 않았거나, 번호판 번호를 해시하지 않고 암호화하였다면(그리고 암호 키를 공개하지 않았다면), 또는 택시 리무진 위원회가 각각의 번호판 번호에 대해 임의로 코드를 생성했다면 문제가 없었을 수도 있다.

⁸⁰ Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset,' Anthony Tockar, September 15, 2014, <http://research.neustar.biz/author/atockar/>

⁸¹ Public NYC Taxicab Database Lets You See How Celebrities Tip,' J. K. Trotter, GAWKER, October 23, 2014. <http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546>

집단까지 커진다면, 특이성을 식별하는데 필요한 지점의 수도 증가할 수 있다.^{82 83}

또한 의료 검사와 택시 연구에서는 모두 정보에 비교적 적은 인자변환(perturbation)이 있을 경우에는 재식별이 어렵거나 불가능하게 될 수 있다는 점도 보여주고 있다. Atreya 등이 검토한 의학 검사 결과에서 저자들은 임상적 중요성을 크게 변화시키지 않으면서도 식별 정보를 제거하는 방식으로 임상적 수치에 적응적 잡음(adaptive noise)을 추가하는 단순한 비식별처리 알고리즘을 제시하고 있다. 택시 연구의 경우, 택시 번호판 가명을 되돌릴 수 있었기 때문에 유명인사들이 식별되었다. 만약 번호판을 적절히 보호하였거나 GPS 위치정보가 100 제곱미터짜리 그리드에 종합되었다면 재식별 리스크는 상당히 줄었을 수 있었으며, 프라이버시에 대한 영향은 무시할 수 있었을 것이다.

El Emam 등⁸⁴은 2001년부터 2010년까지 발간된 재식별 시도 14건을 검토하였다. 저자들은 각각에 대해 의료 정보가 포함되어 있었는지, 적의 직업이 무엇인지, 재식별이 일어난 국가는 어디인지, 재식별된 개인기록부의 비율은 어떠한지, 비식별처리에 적용된 기준은 무엇인지, 재식별이 검증되었는지를 확인하였다. 연구자들은 현행 기준에 따라 비식별처리되지 않은 작은 정보 집합의 경우에 재식별에 성공하는 경우가 많았음을 확인하였다. 많은 경우에 재식별 연구자들은 단지 몇 개의 개인기록부만 재식별하였고, 그러한 재식별이 검증되지 않았을 수도 있었다. 그렇기 때문에, 이러한 사례에서 과학적 결론을 도출하기는 어렵다.

3.7 재식별 리스크의 추정

위의 예에서 알 수 있듯이, 개인정보를 다루는 개인과 기관은 주어진 특정한 비식별처리 절차의 재식별 리스크를 계산하기 위해 쉽고 신뢰성 있는 절차가 필요하다. 샘플링한 정보 집합 안의 다양한 개인의 구분 가능성, 비식별처리 알고리즘, 연결 정보의 가용성, 재식별 공격을 수행할 수 있는 개인의 범위 등의 여러 가지 요소에 따라 달라지기 때문에 이러한 리스크의 계산은 복잡하다.⁸⁵

재식별 리스크의 종류도 다양하고 리스크의 보고 방식도 다양하다. 모형은 식별되는 주체 각각의 평균 리스크를 보고할 수 있고, 어떤 주체가 식별될 수 있는 리스크를 보고하거나,

⁸² Daniel Barth-Jones and Khaled El Emam, Assessing data intrusion threats, Science, 10 April 2015, VOL 348 Issue 6231, p. 194.

⁸³ Jane Yakowitz Bambauer, Is De-Identification Dead Again, INFO/LAW Blog, April 28, 2015.

⁸⁴ K El Emam, E Jonker, L Arbuckle, B Malin (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. doi:10.1371/journal.pone.0028071

⁸⁵ Catherine Marsh, Chris Skinner, Sara Arber, Bruce Penhale, Stan Openshaw, John Hobcraft, Denise Lievesley and Nigel Walford, The Case for Samples of Anonymized Records from the 1991 Census Journal of the Royal Statistical Society. Series A (Statistics in Society) Vol. 154, No. 2 (1991), pp. 305-340

어떤 주체가 다양한 k 명의 개인 가운데 1 인 것으로 식별될 수도 있다.

Dankar 등은 다양한 종류의 정보 출처의 구분가능성(distinctiveness)을 추정하기 위한 통계적 모형과 결정 규칙을 제안하고 있다.⁸⁶ El Eman 등은 단일한 개인기록부를 식별하는 것이 목표인 ‘동기가 약한 적’과 모든 매칭을 식별하고 검증하며, 실제적 또는 금전적 제한만을 받는 ‘동기가 강한 적’이라는 두 가지 공격자 모형을 이용하여 약물 부작용 보고서를 재식별할 리스크를 모형화하는 기법을 개발하였다.⁸⁷ 일반적으로, 기관들은 모든 개인정보 요소(연속적 요소 및 범주적 요소)를 준식별자로 취급하는 결정과 재식별 가능성과 공개하는 실질적으로 변경되지 않은 개인정보의 유용성 간에 균형을 잡아야 한다. 비식별처리를 할 경우에는 정보 안의 통계적 관계에 심각한 손상을 입혀, 비식별처리한 개인정보의 유용성이 제한될 수 있다.

메모 [한24]: 일반적인 정보를 의미합니다.

마지막 절은 구조화된 개인정보의 비식별처리를 주로 다루었지만, 이 절에서는 구조화되지 않은 문서와 멀티미디어를 비식별처리하는 문제를 다루고 있다.

4.1 의료 문서의 비식별처리

의료 문서에는 구조화되지 않은 문안이 상당히 많이 포함되어 있다. 최근 몇 년 동안, 자유로운 양식의 문서(예를 들면, 서술형으로 된 섭취 보고서)에서 평서문 처리 기법을 이용하여 HIPAA 정보 요소 18 개를 제거하기 위한 도구를 개발하고 평가하려는 노력이 있었다. 주로 연구한 두 가지 기법은 규칙을 기반으로 한 시스템과 통계를 기반으로 한 시스템이 있다. 규칙을 기반으로 한 시스템은 특정 유형의 문서에 잘 적용되지만, 새로운 영역에 적용했을 때에는 잘 작동하지 않는 경향이 있다. 통계적 기법은 일반적으로 규칙을 기반으로 한 시스템에 비해 정확도가 낮고 라벨링된(labeled) 훈련 정보가 필요하지만, 새로운 분야에 적용하기 쉽다는 장점이 있다.

메모 [한25]: 일반적인 정보를 의미합니다.

평서문을 비식별처리하는데 있어 다양한 요소들이 장애가 된다.

- 1) 성명과 주소 등의 직접 식별자를 분명하게 표시할 수 없을 수 있다.
- 2) 중요한 의료 정보를 개인정보로 착각하고 삭제할 수 있다. 의학에서 질병을 설명하는데 흔히 사용하는 사람이름이 붙은 병명의 경우에는 특히 문제가 된다(예,

⁸⁶ Dankar et al. Estimating the re-identification risk of clinical data sets, BMC Medical Informatics and Decision Making 2012, 12:66.

⁸⁷ El Eman et al., Evaluating the risk of patient re-identification from adverse drug event reports, BMC Medical Informatics and Decision Making 2013, 13:114 <http://www.biomedcentral.com/1472-6947/13/114>

애디슨병(Addison's Disease), 벨 마비(Bell Palsy), 라이터 증후군(Reiter Syndrome) 등)

- 3) HIPAA 세이브하버 요소 18 개를 제거한 후에도 정보에서 의료 주체를 식별할 수 있을 수 있다.

메모 [한26]: 일반적인 정보를 의미합니다

몇몇 연구자들이 문서 비식별처리 도구에 대한 정식 평가를 수행한 바 있다.

- 2012 년에 신시네티 아동병원 의료 센터(Cincinnati Children's Hospital Medical Center, 이하 'CCHMC')의 Deleger 등은 MALLET 기계학습 패키지를 기반으로 CCHMC 가 자체 개발한 시스템인 MCRF 에 대해 MITRE 식별 스크러버 툴킷(Identification Scrubber Toolkit, MIST)⁸⁸ 을 검사하였다. 참조 말뭉치(corpora)는 2010 년에 CCHMC 에서 작성한 5 백만 건의 개인기록부 에서 선정한 3,503 건의 임상 개인기록부, 2006 i2b2 비식별처리 도전 말뭉치(de-identification challenge corpus)⁸⁹, PhysioNet 말뭉치^{90 91}였다.
- 2013 년에 유타 대학교(the University of Utah) 생의학 정보공학과(Department of Biomedical Informatics)의 Ferrández 등은 자동 비식별처리 시스템 5 개를 2 개의 참조 말뭉치에 대해 평가하였다.⁹² 이 검사는 비식별처리하고 합성 정보를 더한 문서 889 건⁹³으로 이루어진 2006 i2b2 비식별처리 도전 말뭉치와 미 보훈보건청(Veterans Health Administration, 이하 'VHA')이 제공한 2008 년 4 월 1 일과 2009 년 3 월 31 일 사이의 단어 500 개 이상인 문서에서 임의로 추출한 문서 800 건으로 이루어진 말뭉치를 이용하여 수행하였다.
- 2013 년에 국립 의학도서관(the National Library of Medicine, 이하 'NLM')은 도서관 산하 과학 자문관 위원회(Board of Scientific Counselor)에게 '임상 문서 비식별처리

⁸⁸ Aberdeen J, Bayer S, Yeniterzi R, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. *Int. J. Med Inform* 2010;79:849e59.

⁸⁹ Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*. 2007;14:550e63.

⁹⁰ Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008;8:32.

⁹¹ Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. *Circulation* 2000;101:E215e20.

⁹² Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., & Meystre, S. M. (2013). BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *Journal of the American Medical Informatics Association: JAMIA*, 20(1), 77-83. doi:10.1136/amiajnl-2012-001020

⁹³ Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc* 2007;14:550e63.

연구(Clinical Text De-Identification Research)'라는 제목의 보고서를 발간하였는데, 여기에서 NLM 은 자체적으로 개발한 도구인 NLM Scrubber(NLM-S)와 MIT 의 비식별처리 시스템(MITdeid), MIST 를 비교하였다.⁹⁴ 이 검사는 1073 건의 의사관찰보고서(Physician Observation Reports)와 NIH 임상센터(Clinical Center)의 2020 건의 환자 연구 보고서로 이루어진 내부 말뭉치를 이용하여 수행하였다.

CCHMC 와 유타 대학교의 연구는 모두 원 상태의 시스템을 검사하였고, 글뭉치의 일부를 훈련 정보의 일부로 사용하여 조정한 후에 검사하였다. 유타 대학교의 연구에서는 공개하기 위한 VHA 의 개인기록부를 충분히 비식별처리하지 못하였고, 규칙에 기반한 시스템은 특정한 유형의 정보를 잘 찾아냈지만(예, 사회보장번호, 전화번호), 다른 종류의 정보에 대해서는 훈련 가능한 시스템이 더 나은 결과를 보인다는 점을 발견하였다. 시스템에 약간의 변경이 있었지만, 모두 유사한 성능을 나타냈다. NLM 의 연구에서는 NLM-S 가 MIST 와 MITdeid 를 크게 능가하였고, NLM 정보 집합에서 HIPAA 세이프하버 기준에 매칭되는 식별자의 99.2%를 제거하였다. 저자들은 나머지 식별자들이 환자의 프라이버시에 크게 위협이 되지 않을 것이라고 결론지었다.

Meystre, Shen 등의 연구에 의하면, 솔트레이크 시티(Salt Lake City) VHA 의료 센터의 자동적으로 비식별처리된 퇴원 정보에서는 환자의 의사 또는 치료 전문가를 식별할 수 없었다.⁹⁵ Carrell 등의 연구에서는 비식별처리된 문서에서 '환자(PATIENT)' 및 '주소(ADDRESS)'와 같은 일반적인 라벨 대신에 '조지 워싱턴(George Washington)'이나 '1600 Pennsylvania Ave'와 같은 실제와 비슷한 가명을 사용한 경우, '검토자가 가명과 남은(누설된) 식별자를 구분할 수 없었기 때문에' 문서에 남아 있는 일부 성명의 재식별 위험이 낮아지거나 경감되는 것을 확인하였다.⁹⁶

이들 시스템들은 HIPAA 세이프하버 정보 요소 18 개를 제거하는 것 이외에는 비식별처리를 시도하지 않아, 다른 정보를 이용하여 개인을 재식별할 가능성을 열어두었다. 예를 들면, 미국과 캐나다의 법규에 따르면 약물 부작용에 관해 보고해야 한다. 이러한 보고서는 기자와 연구자들이 사망 보고서와 뉴스 보도와 사망 등록 등의 다른 정보 출처를 연관시켜

⁹⁴ Kayaalp M et al., A report to the Board of Scientific Counselors, 2013, The Lister Hill National Center for Biomedical Communications, National Library of Medicine.

⁹⁵ Meystre S et al., Can Physicians Recognize Their Own Patients in De-Identified Notes? In Health – For Continuity of Care C. Lovis et al. (Eds.) © 2014 European Federation for Medical Informatics and IOS Press.

⁹⁶ Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., & Hirschman, L. (2013). Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2), 342-348.

재식별하였다.⁹⁷ 세계의 모든 공개된 정보에 대한 접근권(access)을 갖고 동기를 가진 공격자에 대비하여 자동으로 그러한 보고서를 재편집(redact)할 수 있는 시스템을 개발하기 위한 추가적인 연구가 필요하다.

4.2 사진과 비디오의 비식별처리

정지 화상, 판매용 비디오, 감시용 비디오에는 개인을 식별하는데 사용할 수 있는 많은 정보가 있을 수 있다. 비식별처리를 통해 비디오를 비식별처리함으로써 특정한 용도로 사용하도록 하면서도, 이러한 식별정보를 삭제함으로써 개인이 식별되지 못하도록 한다.

사진이나 비디오를 찍었을 때 카메라가 생성하였거나 사후의 가공 도구를 이용하여 추가한 경우, 식별정보가 이미지 정보와 함께 이미지 파일에 내장된 메타정보 형태로 존재할 수 있다. 예를 들어, 어떤 사람의 주택의 GPS 주소, 카메라의 시리얼번호, 또는 사람의 성명이 헤더에 내장되어 있을 수 있다. 이러한 개인정보를 비식별처리하는 것과 관련된 이슈는 제 3 절에서 다른 종류의 구조화된 정보를 비식별처리하는 것에 관련된 이슈와 유사하다. 그러나 널리 알려지지 않은 바이너리 구조(binary structure)로 이러한 개인정보를 저장할 수 있기 때문에 더욱 문제가 복잡하다. 따라서, 식별정보가 존재할 수 있지만, 알 수 없거나 감춰져 있을 수 있다.

또한 멀티미디어 콘텐츠 자체를 비식별처리하려고 할 때 발생하는 다양한 이슈도 문제를 더욱 복잡하게 한다. 멀티미디어 콘텐츠에는 예를 들면 관찰자가 어떤 사람이 키가 크고 나이 든 여성임을 추론할 수 있도록 하는 등, 특정 인물을 식별할 수 있거나 개인의 신원에 제한을 가할 수 있는 정보가 많다.

ICT COST Action IC1206 “멀티미디어의 프라이버시 보호를 위한 비식별처리”는 멀티미디어 콘텐츠에서 “얼굴, 목소리, 실루엣, 걸음걸이와 같은” 식별정보를 삭제하기 위한 방법을 모색하고 있다.⁹⁸ 그러한 작업의 일환으로, 위원회는 멀티미디어 콘텐츠 안의 식별자 분류체계를 개발해왔다.

- **생체인증 식별자**, 개인을 식별하는데 사용하는 구분되고, 측정할 수 있으며, 일반적으로 특이성이 있고, 영구적인 개인적 특징. 생리학적 생체인증 식별자(얼굴, 홍채, 귀, 지문)와 행동학적 생체인증 식별자(음성, 걸음, 자세, 입술움직임, 타이핑

⁹⁷ K El Emam, E Jonker, L Arbuckle, B MalinB (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. doi:10.1371/journal.pone.0028071

⁹⁸ ICT COST Action IC1206, 'De-identification for privacy protection in multimedia content,' 27 November 2012, http://www.cost.eu/COST_Actions/ict/Actions/IC1206.

스타일) 등이 있다.

- **연성 생체인증 식별자.** 영구적이거나 구분되지는 않지만 모호한 신체적, 행동적이거나 사람에게 부수된 특징(키, 몸무게, 눈 색깔, 실루엣, 연령, 성별, 인종, 사마귀, 문신, 점, 흉터)
- **비생체인증 식별자.** 문체, 어조, 특정 사회-정치적, 환경적 문안, 복장 스타일, 헤어스타일 등.⁹⁹

멀티미디어 콘텐츠의 비식별처리에서는 이미지와 음향 정보(있는 경우)를 변경하여 식별을 어렵게 하거나 불가능하게 한다. 생체인증, 연성 생체인증, 비생체인증 식별자는 함께 있는 경우가 많으며, 개인의 프라이버시를 보호하기 위해서는 이들을 모두 비식별처리해야 한다. 이러한 작업을 다중모드 비식별처리(multimodal de-identification)라고 할 수 있다.

초기 연구에서는 이미지를 변환하여, 자동 얼굴 인식 시스템을 이용하여 개인을 신뢰성 있게 식별할 수 없도록 하는데 목적을 두었다. 예를 들어, 사진 비식별처리 기법을 대규모로 적용한 구글 스트리트 뷰(Google Street View)가 사용한 얼굴 흐리게 하기(face blurring)가 있다.¹⁰⁰ 일부 연구자들은 몸을 식별하고 흐리게 할 수 있는 시스템을 개발하였는데¹⁰¹, 이는 연구 결과에서는 얼굴이 없이도 몸을 식별할 수 있다는 점을 보여주었기 때문이다.¹⁰² 어떤 시험용 시스템은 특정 종류의 정지 이미지에서 식별문신의 위치를 확인하고 제거할 수 있는데,¹⁰³ 품질이 낮은 정보, 복잡한 장면, 일부가 가려져 있거나 높은 곳에서 본 문신 또는 많은 원근이 포착되는 비디오에서 정지 이미지를 어떻게 처리하는지는 분명하지 않다.

사람의 이미지를 비식별처리하는 것 이외에도, 음향과 문서에서 식별단서(identifying cue)도 처리해야 한다. Cunningham 과 Truta 는 개인정보를 최소한으로 손실하면서 발화자의 신원을 보호하는 말뭉치를 생산하는데 도움이 되는 '통제 음향 왜곡(controlled audio distortion)'을 이용하였다.¹⁰⁴ 물론 단순한 음향 왜곡을 통해서만 단지 발화자의 음성을 이용하여 식별을

⁹⁹ 이러한 분류체계는 본 보고서의 이전 초안에 대한 COST Action IC1206 논평의 일부로 Slobodan Ribaric 이 제시하였다.

¹⁰⁰ Frome, Andrea, et al., 'Large-scale Privacy Protection in Google Street View,' IEEE International Conference on Computer Vision (2009).

¹⁰¹ Prachi Agrawal and P. J. Narayanan. 2009. Person de-identification in videos. In Proceedings of the 9th Asian conference on Computer Vision - Volume Part III (ACCV'09), Hongbin Zha, Rin-ichiro Taniguchi, and Stephen Maybank (Eds.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 266-276. DOI=10.1007/978-3-642-12297-2_26 http://dx.doi.org/10.1007/978-3-642-12297-2_26

¹⁰² Rice, Phillips, et al., Unaware Person Recognition From the Body when Face Identification Fails, Psychological Science, November 2013, vol. 24, no. 11, 2235-2243 <http://pss.sagepub.com/content/24/11/2235>

¹⁰³ Darijan Marc'etić et al., An Experimental Tattoo De-identification System for Privacy Protection in Still Images, MIPRO 2014, 26-30 May 2014, Opatija, Croatia

¹⁰⁴ Scot Cunningham and Traian Marius Truta. 2008. Protecting privacy in recorded conversations. In Proceedings of the 2008 international workshop on Privacy and anonymity in information society (PAIS '08), Farshad Fotouhi, Li Xiong, and Traian Marius Truta

방지할 수 있을 뿐, 발화 패턴이나 이름을 언급하는 등의 오디오 트랙에 있을 수 있는 다른 식별단서를 제거할 수 없다.

멀티미디어 비식별처리의 효과 검증은 다음과 같은 문제가 수반된 다차원적인 문제이다.

- **비식별처리가 필요한 대상의 식별 정밀도와 정확성.** 구글은 자사의 완전 자동 시스템이 얼굴의 89%, 번호판의 94~96%를 흐리게 할 수 있다고 보고하고 있다.¹⁰⁵ 그럼에도 불구하고, 기자들은 많은 얼굴들이 흐리게 처리되지 않은 채로 있다고 구글을 비판해왔다.¹⁰⁶ 기자들은 구글이 종교적 조상(彫像)의 얼굴을 흐리게 처리하였다고도 비판하였다.¹⁰⁷ ¹⁰⁸ 어떤 경우에는 특정 사물, 상징, 또는 인물을 흐리게 처리하거나 달리 훼손하는 것이 용인되지 않을 수 있다.
- **변환의 가역성(reversibility).** 복수의 이미지를 조합하여 모자이크를 제거하고 흐린 이미지를 복원하는 기술이 있기 때문에, 모자이크처리(pixelation)나 흐리게 하기(blurring)를 이용하여 비디오를 모호하게 하는 경우에 주의해야 한다. 일부 연구자들은 모자이크처리나 흐리게 하기 대신에 얼굴을 합성된 얼굴 ¹⁰⁹ ¹¹⁰ 또는 완전히 다른 얼굴로 대체할 수 있는 시스템을 개발해왔다.¹¹¹ ¹¹²
- **생성된 이미지의 시각적 품질.** 흐리게 하기와 모자이크처리는 이상한 그림을 만들어내어 장면에 대한 해석에 영향을 줄 수 있다는 단점이 있다.
- **선택한 신원 모호화 기법의 실제적인 효과.** 일부 연구자들은 얼굴 인식 소프트웨어에

(Eds.). ACM, New York, NY, USA, 26-35. DOI=10.1145/1379287.1379295 <http://doi.acm.org/10.1145/1379287.1379295>

¹⁰⁵ Frome, Andrea, et al., 'Large-scale Privacy Protection in Google Street View,' IEEE International Conference on Computer Vision (2009).

¹⁰⁶ Stephen Chapman, 'Google Maps, Street View, and privacy: Try harder, Google,' ZDNet, January 31, 2013. <http://www.zdnet.com/article/google-maps-street-view-and-privacy-try-harder-google/>

¹⁰⁷ Gonzalez, Robbie. 'The Faceless Gods of Google Street View,' io9, October 4, 2014. <http://io9.com/the-faceless-gods-of-google-street-view-1642462649>

¹⁰⁸ Brownlee, John, 'The Anonymous Gods of Google Street View,' Fast Company, October 7, 2014. <http://www.fastcodesign.com/3036319/the-anonymous-gods-of-google-street-view#3>

¹⁰⁹ Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. In: Protecting Privacy in Video Surveillance, A. Senior, editor. Springer, 2009 Preserving Privacy by De-identifying Facial Images. <http://dataprivacylab.org/projects/facedevid/paper.pdf>

¹¹⁰ E. Newton, L. Sweeney, and B. Malin. Preserving Privacy by De-identifying Facial Images, Carnegie Mellon University, School of Computer Science, Technical Report, CMU-CS-03-119. Pittsburgh: March 2003.

¹¹¹ Saleh Mosaddegh, Lóic Simon, Frederic Jurie. Photorealistic Face de-Identification by Aggregating Donors' Face Components. Asian Conference on Computer Vision, Nov 2014, Singapore. pp.1-16.

¹¹² Umar Mohammed, Simon J. D. Prince, and Jan Kautz. 2009. Visio-lization: generating novel facial images. In ACM SIGGRAPH 2009 papers (SIGGRAPH '09), Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 57, 8 pages. DOI=10.1145/1576246.1531363 <http://doi.acm.org/10.1145/1576246.1531363>

비해 알고리즘에 높은 점수를 주지만, 복장, 자세, 공간-시간적 환경을 조합하여 사람을 식별할 수 있을 수 있다.

본 보고서의 앞 부분에서는 구조화된 개인정보와 문서의 비식별처리의 효과를 정량화하기 위한 몇몇 시도를 논의하였지만, 멀티미디어 비식별처리의 효과를 정량화 하기 위한 의미 있는 노력은 확인된 바가 없다.

다양한 재식별 시나리오가 있기 때문에, 비식별처리에 대한 적절한 검사는 복잡하다.

- 자동 식별화 시스템의 재식별 시도.
- 훈련되어 있지만, 정보주체에 대한 개인적인 지식이 없는 사람에 의한 재식별
- 정보주체의 관련자나 친구에 의한 재식별.

4.3 의료 영상의 비식별처리

과학적 연구 성과의 재현과 검증, 연구 정보 기반의 확대와 품질 평가 등 과학적, 운영적 목표를 달성하기 위해 보다 넓은 범위로 의료 영상을 공유하려는 욕구가 있다. 프라이버시에 대한 우려와 HIPAA 프라이버시 규칙에 따라, 이미지를 공유하기 전에 비식별처리를 해야 한다.

의료 영상은 통상적으로 DICOM¹¹³ (Digital Imaging and Communication in Medicine) 양식으로 된 헤더 정보(메타정보)와 이미징 기기가 생성한 픽셀(pixel)(이미지 정보)로 이루어져 있다. 개인을 식별할 수 있는 정보는 헤더나 픽셀에 존재할 수 있다. 의료 이미지는 비식별처리에 관한 다음과 같은 많은 문제를 야기한다.

- DICOM 헤더에는 환자의 성명 또는 기타 직접 식별자, 준식별자가 포함되어 있을 수 있다.
- 픽셀에는 사진 정보 또는 개인을 분간할 수 있는 기타 생체인증 정보가 있을 수 있다. 예를 들면, 병원이 사람 얼굴이 들어간 부상 사진을 찍을 수 있고, 여기에서 사람을 식별할 수 있다.
- 픽셀에 있는 정보를 수학적으로 처리하여 인식 가능한 이미지나 생체인증 정보를

¹¹³ The DICOM Standard 2015a, Medical Imaging and Technology Alliance, National Electrical Manufacturers Association, 2015. <http://medical.nema.org/>

생산할 수 있다.^{114 115}

- 직접 식별자가 인간이 읽을 수 있는 문서 형태로 픽셀 안에 내장되어 있을 수 있다. 예를 들면, 이미지를 만들 때 디지털화하거나 사진으로 포착한 개인기록부에 수기로 환자의 성명을 써놓을 수 있다. 그 대신에 소프트웨어를 이용하여 이미지 안의 예측가능 하거나 지정된 위치에 문서를 스캔(burn-in)할 수 있다.

다른 파일 양식과는 다르게, DICOM 표준에는 헤더나 이미지의 픽셀 구역에 식별정보가 존재하는지를 지정하고, 어떤 파일이 비식별처리되었는지 지정하기 위한 조항이 있다.¹¹⁶ 예를 들어, DICOM의 픽셀정보 삭제 옵션(Clean Pixel Data Option)을 지정하면, 식별정보가 픽셀 정보로 스캔되었는지 표시할 수 있다. 만약 인식 가능한 시각적 특징 삭제 옵션(Clean Recognizable Visual Features Option)을 지정하면, 픽셀 정보 안의 정보를 이용하여 개인을 인식할 수 있는지를 표시할 수 있다. DICOM PS3.15 부속서 E, 속성 비밀 프로필(Attribute Confidentiality Profiles)에는 비식별처리된 DICOM 이미지를 생산하기 위한 구조화된 절차가 명시되어 있다. 즉, 식별정보를 재편집하거나 암호화하여 ‘암호화 속성 정보 집합(Encrypted Attributes Dataset)’에 저장함으로써 비식별처리된 파일을 나중에 재식별할 수 있다.

그러나 DICOM 표준은 비식별처리를 위한 실제 알고리즘이나 기법을 명시하고 있지 않다. 그러한 기법은 각각의 이미지 작성 방식에 따라 다를 것이며, 다음과 같이 활발하게 연구가 진행되고 있는 분야이다.

- Freymann 등은 이미지를 공개 아카이브에 저장하기 전에 DICOM 헤더 정보를 비식별처리하기 위한 오픈 소스 소프트웨어 세트를 개발하였다.¹¹⁷ 저자들은 상용 소프트웨어는 DICOM 스캔 표시기가 지원되지 않기 때문에, 스캔 방지 의료 정보를 스캔하는 도구가 필요하다는 점에 주목하고 있다.
- 몇몇 연구자들은 컴퓨터 단층촬영(CT) 및 자기공명(MR) 영상으로 만든 3차원 모형을

¹¹⁴ J. Chen, K. Siddiqui, L. Fort, R. Moffitt, K. Juluru, W. Kim, N. Safdar, and E. Siegel, ‘Observer success rates for identification of 3D surface reconstructed facial images and implications for patient privacy and security,’ in Proc. SPIE Int. Soc. Opt. Eng., 2007, pp. 65161B-1-65161B-8

¹¹⁵ F. W. Prior, B. Brunsten, C. Hildebolt, T.S. Nolan, M. Pringle, S.N. Vaishnavi, L.J. Larson-Prior. Facial Recognition From Volume-Rendered Magnetic Resonance Imaging Data. IEEE Transactions On Information Technology In Biomedicine, 13:1, Jan 2009 p.5-9.

¹¹⁶ Digital Imaging and Communications in Medicine (DICOM) Supplement 142: Clinical Trial De-Identification Profiles, DICOM Standards Committee, Working Group 18 Clinical Trials, Jan 25, 2011. ftp://medical.nema.org/medical/dicom/final/sup142_ft.pdf

¹¹⁷ John B. Freymann, Justin S. Kirby, et al., Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-Identification, J. Digit Imaging (2012) 25:14-24.

이용하여 사람의 얼굴을 수학적으로 재구성하기 위한 기법을 시연해왔다.^{118 119} 연구자들은 이러한 위협에 대응하기 위해 얼굴을 만드는데 사용할 수 있지만 진단용으로는 활용 가치가 없는 MR 이미지 부분을 자동으로 제거하기 위한 기법을 개발해왔다.^{120 121}

4.4 유전 정보와 생물학 자료의 비식별처리

유전자 서열과 기타 서열 정보는 HIPAA 프라이버시 규칙에서 개인을 식별하는 정보로 간주되지 않는다. 일부 유전자 서열은 매우 개인적이고, 현재 명시적으로 개인을 식별하기 위하여 유전자 정보를 일상적으로 수집하여 데이터뱅크에 넣기 때문에, 일부 평론가들은 프라이버시 규칙에 관련 내용이 없다는 점을 비판해왔다. 더욱이, 유전자 정보는 유전되기 때문에, 자신은 유전자 서열이 확인되지 않았지만, 대신에 친척의 유전자 서열이 식별 데이터베이스에 포함된 사람의 신원을 확인하는데 유전자 서열을 이용해왔다.

다음과 같은 사례가 있다.

- 2005 년에 나이가 15 세인 한 소년이 DNA 테스트 서비스 업체인 FamilyTreeDNA.com 을 이용하여 정자를 제공한 아버지를 찾았다. 비용이 289 달러인 이 서비스는 이 소년의 아버지를 식별하지 못했지만, Y 염색체가 일치하는 남자 2 명을 식별하였다. 두 남자의 성(姓)은 같았지만 철자가 달랐다. Y 염색체는 아버지에게서 아들로 그대로 전해지기 때문에 유럽의 성(姓)과 같은 방식으로 유전되는 경향이 있다. 정자 제공 일자와 출생지(이 정보는 소년의 어머니에게 제공되었다)와 이 정보를 이용하여 이 소년은 현재 ‘유전관계 삼각측량(genealogical triangulation)’이라고 알려진 온라인 검색 서비스¹²²를 이용하여 아버지를 식별할 수 있었다.
- 2013 년에 MIT 연구 팀은 이 실험을 연장하여, 인간 다형성 연구(the Study of Human Polymorphisms, CEPH)와 1000 게놈 프로젝트(1000 Genomes Project)의 일부로 DNA

¹¹⁸ Kermi, A.; Marniche-Kermi, S.; Laskri, M.T., ‘3D-Computerized facial reconstructions from 3D-MRI of human heads using deformable model approach,’ Machine and Web Intelligence (ICMWI), 2010 International Conference on, vol., no., pp.276,282, 3-5 Oct. 2010 doi: 10.1109/ICMWI.2010.5648144

¹¹⁹ Fred W. Prior, Barry Brunnsden, Charles Hildebolt, Tracy S. Nolan, Michael Pringle, S. Neil Vaishnavi, and Linda J. Larson- Prior. 2009. Facial recognition from volume-rendered magnetic resonance imaging data. Trans. Info. Tech. Biomed. 13, 1 (January 2009), 5-9. DOI=10.1109/TITB.2008.2003335 <http://dx.doi.org/10.1109/TITB.2008.2003335>

¹²⁰ Bischoff-Grethe A, Ozyurt IB, Busa E, et al. A Technique for the Deidentification of Structural Brain MR Images. Human brain mapping. 2007;28(9):892-903. doi:10.1002/hbm.20312.

¹²¹ Milchenko M, Marcus D, Obscuring surface anatomy in volumetric imaging data, Neuroinformatics 2013 Jan; 11(1):65-75.

¹²² Sample, Ian. Teenager Finds Sperm Donor Dad on Internet. The Guardian, November 2, 2005. <http://www.theguardian.com/science/2005/nov/03/genetics.news>

검사 결과가 인터넷에 공개된 50 명 이상의 성(姓)과 완벽한 신원을 식별하였다.¹²³

Erlich 와 Narayanan 은 유전자 서열을 재식별하는 기법과 그러한 공격을 방어할 수 있는 기술적 조치에 관한 자세한 검토결과를 공개하였다.¹²⁴

현재로서는 재식별에 필요한 최소한의 유전자 서열 크기에 관한 과학적 합의는 없다. 또한 재식별을 금지할 데이터 이용 합의를 집행할 필요 없이 연구자들이 비식별처리된 유전자 정보를 이용하도록 하기 위한 적절한 메커니즘에 대한 합의도 없다.

4.5 지리 정보와 지도 정보의 비식별처리

메모 [한27]: 일반적인 정보를 의미합니다

지리 정보의 비식별처리는 연구가 활발한 분야이다. 현재의 방법은 인자변환(perturbation)과 일반화(generalization)에 의존하고 있다. 인자변환은 일부 경우에 문제가 있을 수 있다. 왜냐하면 예를 들면 식당이 물 속으로 움직인다는지, 인자변환된 위치가 상식에서 벗어날 수 있기 때문이다.¹²⁵ 그러나 특히 집단의 분포가 분산되어 있거나 복수의 관찰 결과를 연관시킬 수 있는 경우에는(특히 관찰 결과가 다른 일반화 영역을 이용하는 경우), 일반화로는 신원을 감추기에 불충분할 수 있다.¹²⁶

일종의 일반화나 인자변환이 없다면, 위치를 비식별처리하기가 매우 어려울 수 있는 지리적 정보가 매우 다양하다. 예를 들면, 일정한 기간에 걸쳐 측정된 핸드폰의 가속도계를 이용하여 움직임을 거리 그리드(street grid)에 끼워 넣어 위치를 추론할 수 있다.¹²⁷ 따라서, 일정 기간에 걸쳐 샘플링한 가속 운동은 공개된 정보와 조합할 경우에 식별정보가 될 수 있다.

비식별처리 기법은 정보 집합에서 식별정보를 제거하면서도 나머지 정보의 유용성을 유지하기 위한 것이다. 기관에서는 비식별처리를 이용하여 개인정보의 활용이나 보관에 관련된 프라이버시 리스크를 최소화할 수 있다. 정보 집합을 공유하거나 공개하기 전에도

¹²³ Gymrek et al., Identifying Personal Genomes by Surname Inference, Science 18 Jan 2013, 339:6117.

¹²⁴ Yaniv Erlich and Arvind Narayanan, 'Routes for breaching and protecting genetic privacy,' Nature Reviews Genetics, June 2014, volume 15, p. 409

¹²⁵ Christopher A. Cassa, Shannon C. Wieland, and Kenneth D. Mandi, Re-identification of home addresses from spatial locations anonymized by Gaussian skew, International Journal of Health Geographics, 12 August 2008. <http://www.ij-healthgeographics.com/content/7/1/45>

¹²⁶ Philip Steel and Jon Sperling, 'The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code TrAbulation Geography,' U.S. Census Bureau staff report, 2001. <https://www.census.gov/srd/sdc/steel.sperling.2001.pdf>

¹²⁷ Jun Han; Owusu, E.; Nguyen, L.T.; Perrig, A.; Zhang, J., "ACComplce: Location inference using accelerometers on smartphones," Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on, pp.1,9, 3-7 Jan. 2012

비식별처리를 이용할 수 있다. 비식별처리된 개인정보를 추가적인 통제 없이도 공개할 수 있는 경우도 있고, 어떤 경우에는 개인정보의 접수자가 할 수 있는 것을 선택적으로 제한하는 데이터 이용 합의서와 같은 추가적인 보호 조치가 필요하다.

비식별처리된 개인정보를 배포하거나 이용함으로써 다양한 문제가 발생할 수 있다. 비식별처리된 개인정보를 재식별, 즉 개인정보를 역으로 원래의 정보주체와 일치시킬 수 있는 경우에 이러한 문제 가운데 일부가 발생한다. 발생할 수 있는 중요한 문제에는 사적인 사실의 공개(disclosure of private facts)와 평판의 손상(reputational damage)이 있다. 사적인 사실의 공개는 자신의 개인정보가 재식별된 사람들에게 영향을 미친다. 평판의 손상은 비식별처리를 수행한 기관에 영향을 미친다.

공개된 재식별 사례의 대부분은 연구자나 기자가 수행한 것이며, 많은 경우에 대중인사가 재식별되었다. 대중 인사를 이용하여 재식별 리스크를 측정하는 연구는 리스크가 과도하게 높게 나올 수 있다. 왜냐하면 원래 대중 인사는 공개된 영역에 있는 정보가 더 많기 때문에 일반인에 비해 재식별이 쉽기 때문이다.

재식별 리스크라는 이슈에 대해서는 연구자들 사이에도 상당한 이견이 있다. 일부 연구자는 재식별 리스크라는 용어를 재식별할 수 있는 비식별처리된 개인기록부의 비율을 의미하는 것으로 사용한다. 이 경우, 재식별 리스크는 재식별 공격을 수행함으로써 직접적으로 측정할 수 있다. 어떤 연구자는 재식별 리스크를 미래의 개인기록부 재식별 확률을 의미하는 것으로 간주한다. 이 경우, 리스크를 추정해야 하지만, 결국 정량화할 수 없다. 왜냐하면 지금은 없을 수 있지만 미래에 있는 정보의 가용성에 따라 리스크가 달라지기 때문이다. 또한 어떤 개인기록부가 재식별될 수 있는 확률로서 재식별 리스크를 추정할 수도 있다. 그렇기 때문에 재식별 리스크를 논의할 때에는, 항상 특정한 재식별 시나리오에서의 특정한 리스크를 기술해야 한다.

그러한 정보를 공유하려는 기관들은 재식별 리스크를 줄이기 위해 몇 가지 접근법을 조합할 것을 고려할 수 있다. 이러한 접근법에는 준식별자나 정보주체를 재식별하는데 사용할 수 있는 다른 종류의 정보를 삭제하는 등의 기술적 통제, 공유하는 비식별처리된 정보와 연결할 수 있는 정보를 지속적으로 조사하는 방법, 재식별, 다른 정보와의 연결, 타인과의 공유를 금지하는 데이터 이용 합의서와 사용자 클릭 합의서 등의 비식별처리된 개인정보에 대한 통제, 정보 접수자의 활동을 제한하는 기술적 통제 등이 있다.

기관들이 개인정보를 담고 있는 정보를 이용, 공유, 심지어는 금전화하려고 함에 따라

비식별처리 기법은 점점 중요성이 높아지고 있다. 그러나 10 년 이상 연구가 진행되었지만 비식별처리의 기초가 되는 과학에 관해서는 비교적 알려진 내용이 적다. 현재 활용되고 있는 많은 기법과 절차들, 예를 들면 HIPAA 프라이버시 규칙의 세이프하버 비식별처리 기준은 이론적 기반이 견고하지 못하다. 비식별처리 과정의 효과를 검사하거나 비식별처리의 결과 소실되는 효용을 측정하기 위한 널리 수용된 기준은 없다. 비식별처리에 대한 관심이 높아짐에 따라, 본 논문에 기술된 정보와 리스크를 정량적으로 포괄할 수 있는 기준과 평가 기법이 분명히 필요하다.

본 간행물에 사용된 일부 용어에 대한 정의가 아래에 제시되어 있다. 주석이 있는 경우, 해당 정의는 다른 간행물의 정의를 인용한 것이다.

익명성(anonymity): “식별에서, 알려진 신원과 연결하기 위한 충분한 신원 정보가 없이도, 어떤 실체를 별개의 것으로 인식할 수 있는 상태”(ISO/IEC 24760-1:2011)

익명처리(anonymization): “식별 정보 집합과 정보주체 간의 연계성을 제거하는 과정”(ISO/TS 25237:2008)

익명처리된 개인정보(anonymized data): “정보 접수자가 환자를 식별할 수 없는 개인정보”(ISO/TS 25237:2008)

익명 식별자(anonymous identifier): “자연인을 정확히 식별하지 못하도록 하는 식별자”(ISO/TS 25237:2008)

공격자(attacker): 시스템의 잠재적 취약점을 공략하려는 사람

속성(attribute): “상태, 외관, 또는 기타 측면을 기술하는데 사용할 수 있는 어떤 실체의 특징 또는 성질”(ISO/IEC 24760-1:2011)¹²⁸

폭력적 공격(brute force attack): 암호학에서 일치성을 찾아내기 위해 모든 가능한 조합을 시도하는 공격

코드화(coded): “1. 개인의 신원이 개인정보 또는 표본이 숫자, 문자, 기호 또는 그것들의

¹²⁸ ISO/IEC 24760-1:2011, Information technology -- Security techniques -- A framework for identity management -- Part 1: Terminology and concepts

조합(즉, 코드)으로 대체된 사람의 신원을 조사자가 쉽게 확인할 수 있도록 할 식별정보(성명 또는 사회보장번호 등), 2. 존재하는 코드를 풀기 위한 키로서, 식별정보를 개인 정보 또는 표본과 연결할 수 있도록 한다.”¹²⁹

통제(control): “리스크를 변경하는 조치. 주석: 통제에는 과정, 정책, 기기, 방식, 리스크를 변경하는 다른 행위가 포함된다.”(ISO/IEC 27000:2014)

포괄 기관(covered entity): HIPAA 에 따라 개인보건의정보를 전자적으로 전송하는 의료보험(health plan), 의료보험 정산소(health case clearinghouse), 또는 의료보험 제공자(HIPAA 프라이버시 규칙)

정보 연결(data linking): “복수의 데이터베이스에서 나온 정보를 매칭하고 조합하는 것”(ISO/TS 25237: 2008)

정보주체(data subjects): “개인정보가 지시하는 사람”(ISO/TS 25237/2008)

데이터 이용 합의서(data use agreement): 개인정보 제공자와 개인정보 접수자 간에 이행하는 합의로서 개인정보의 활용에 관한 약관을 명시하고 있다.

비식별처리(de-identification): “식별정보들과 정보주체 간의 연계성을 제거하는 과정을 가리키는 일반 용어”(ISO/TS 25237-2008)

비식별처리된 개인정보(de-identified information): “충분한 PII 가 제거되거나 모호하게 되어 나머지 개인정보가 개인을 식별하지 않으며, 해당 개인정보를 개인을 식별하는데 사용할 수 있다고 간주할만한 합리적인 사유가 없는 개인기록부”(SP800-122)

직접 식별자(direct identifier): 직접식별 정보(direct identifying data) 참조

직접식별 정보(direct identifying data): “어떤 개인을 직접적으로 식별하는 정보”(ISO/TS 25237:2008)

직접 식별 변수(directly identifying variable): 직접 식별자를 포함하고 있는 정보 유형

공개(disclosure): “정보를 드러내거나 접근권(access)을 제공함”(ISO/TS 25237:2008)

¹²⁹ OHRP-Guidance on Research Involving Private Information or Biological Specimens, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. <http://www.hhs.gov/ohrp/policy/cdebiol.html>

분간 가능한 정보(distinguishable information): “개인을 식별하는데 사용할 수 있는 정보”(SP800-122)

효과(effectiveness): “계획한 활동을 실현하고 계획된 결과를 달성하는 정도”

실체(entity): “사람, 기관, 기기, 하위시스템, 또는 그러한 항목의 그룹 등 정보 및 커뮤니케이션 기술 시스템 안팎의 항목으로서 식별 가능하게 구분되는 것들”(ISO/IEC 24760-1:2011)

가족 및 교육 개인기록부 프라이버시법(Family and Educational Records Privacy Act, FERPA): 학생의 교육 개인기록부의 프라이버시를 규율하는 미국의 기본 법률

공정정보원칙(Fair Information Practice Principles, FIPP): 1970 년대부터 세계적으로 개발된 원칙들로서, 기관들의 개인정보 취급에 대한 지침이 된다.

유전자 정보(genomic information): DNA 서열이나 유전자 검사 결과 등 개인의 유전자에 관한 정보

해(harm): “PII 의 비밀이 침해된 경우, 개인 또는 기관이 겪을 수 있는 부정적인 영향(즉, 사회적, 신체적, 금전적 피해)”(SP800-122)

1996 년 의료보험의 이동성과 책임에 관한 법(HIPAA): 의료 정보의 프라이버시를 규율하는 미국의 기본 법률

의료보험 식별자(healthcare identifier): “오로지 의료보험 시스템이 이용하기 위한 개인 식별자”(ISO/TS 25237:2008)

HIPAA: 1996 년 의료보험의 이동성과 책임에 관한 법 참조

HIPAA 프라이버시 규칙(Privacy Rule): “의료 개인기록부와 기타 개인 보건 정보를 보호하기 위한 국가적 기준을 설정하며, 의료보험, 의료보험 정산소, 전자적으로 특정 의료보험 거래를 수행하는 의료보험 제공자에게 적용된다.”(HIPAA 프라이버시 규칙, 45 CFR 160, 162, 164)

경제와 임상보건을 위한 의료정보기술 법(Health Information Technology for Economic and Clinical Health Act, HITECH Act): 미국의 전자 보건 개인기록부(HER)의 채택을 장려하기 위해 2009 년에 제정된 법

식별 가능한 사람(identified person): “특히 식별번호 또는 그의 신체적, 생리적, 정신적, 경제적, 문화적, 또는 사회적 신원에 따른 특징적 요소들을 이용하여 직접 또는 간접적으로 식별될 수 있는 사람”(ISO/TS 25237:2008)

식별(identification): “신원에 있는 다른 실체 중에서 특정 실체를 식별하기 위해 어떤 실체에 대해 알려진, 또는 관찰된 속성을 이용하는 과정”(ISO/TS 25237:2008)

식별된 정보(identified information): 명시적으로 개인을 식별하는 정보

식별자(identifier): “상응하는 인증자(authenticator)가 입증하기 전에 신원을 주장하는데 사용하는 정보”(ISO/TS 25237:2008)

간접 식별자(indirect identifier): 다른 정보와 연계하여 개인을 식별하는데 사용할 수 있는 정보

추론(inference): “정보에 포함된 ‘단서(clue)’를 통해 정보 집합에 관련된 개인의 신원을 추론할 수 있는 능력을 지칭한다. 이 분석을 통해, 성명이나 사회보장번호와 같은 특정한 식별자가 제거된 경우라도 해당 개인에 관련된 사실을 조합하여 개인의 신원을 확인할 수 있다.”(ASTM E1869¹³⁰)

k-익명성(k-anonymity): “준식별자를 이용하여 다른 정보에 연결할 수 있는 능력이 제한되는 방식으로 특정 개인정보를 공개하기 위한” 기법.¹³¹ k-익명성은 식별자를 범주화하고 인자변환(perturbation)을 산출하여 이를 달성한다.

I-다양성(I-diversity): 동일한 식별자가 지정한 개인기록부들을 충분히 다양화시킴으로써 추론에 따른 공개를 방지하는 k-익명성 접근법을 정교화.¹³²

가능성(likelihood): “어떤 일이 일어날 수 있는 확률”(ISO/IEC 27000:2014)

제한된 정보 집합(limited dataset): “제한된 정보 집합이란 다음과 같은 개인 또는 친척, 고용자, 개인의 가족의 직접 식별자를 배제하는 개인보건정보를 의미한다. (i) 성명, (ii) 타운

¹³⁰ ASTM E1869-04 (Reapproved 2014), Standard Guide for Confidentiality, Privacy, Access, and Data Security Principles for Health Information Including Electronic Health Records, ASTM International.

¹³¹ L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge- based Systems, 10 (5), 2002; 557-570.

¹³² Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. I-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intl. Conf. Data Engg. (ICDE), page 24, 2006.

또는 시, 주, 우편번호를 제외한 우편주소 정보, (iii) 전화번호, (iv) 팩스 번호, (v) 전자메일 주소, (vi) 사회보장번호, (vii) 의료 개인기록부 번호, (viii) 의료보험 수혜자 번호, (ix) 계좌번호, (x) 자격증/면허 번호, (xi) 면허판 번호를 포함한 차량 식별자 및 일련번호, (xii) 기기 식별자 및 일련번호, (xiii) 웹 URL(Universal Resource Locator), (xiv) 인터넷 프로토콜(IP)주소 번호, (xv) 지문, 성문을 포함한 생체인증 식별자, (xvi) 얼굴 전체 사진 및 그와 동등한 이미지”(HIPAA 프라이버시 규정, 45 CFR 164.514(e)(2)). 제한된 정보 집합에는 완전한 날짜, 연령에서 가장 근접한 시(hour), 시(city), 주, 완전한 우편번호가 포함될 수 있다.

연결 가능한 정보(linkable information): “개인에 대한 다른 정보와 논리적으로 연계할 가능성이 있는 개인에 대한 정보 또는 개인에 관련된 정보”(SP800-122)

연결된 정보(linked information): “개인에 대한 다른 정보와 논리적으로 연계된 개인에 대한 정보 또는 개인에 관련된 정보”(SP800-122)

마스킹(masking): 체계적으로 특정 란을 제거하거나 전화번호를 별표로 대체하거나 임의로 생성된 가명으로 대체하는 등, 값의 분석적 효용성을 보존하지 않는 방식으로 값으로 대체하는 과정¹³³

비결정적 잡음(non-deterministic noise): 예측할 수 없는 임의 값

모호화된 개인정보(obscured data): “정보를 감추기 위해 암호화 또는 기타 수단으로 이용하여 왜곡한 개인정보. 마스킹된 정보 또는 애매화된(obfuscate) 정보라고도 부른다.”(SP800-122)

개인 식별자(personal identifier): “주어진 상황에서 어떤 인물을 특이성 있게 식별할 목적의 정보”(ISO/TS 25237:2008)

개인 정보(personal data): “식별되었거나 식별 가능한 자연인(정보주체)에 관련된 모든 정보”(ISO/TS 25237:2008)

개인적으로 식별 가능한 정보(personally identifiable information, PII): “대리기관이 유지한 개인에 관한 모든 정보로서 (1) 성명, 사회보장번호, 생일이자, 출생지, 모친의 본명, 또는 생체인증 개인기록부 등의 개인의 신원을 분간하거나 추적하는데 사용할 수 있는 모든 정보, (2) 의료, 교육, 재정, 고용 정보 등 개인에 연결된 정보 또는 연결 가능한 모든 기타 정보를

¹³³ El Emam, Khaled and Luk Arbuckle, Anonymizing Health Data, O'Reilly, Cambridge, MA. 2013

포함한다.”¹³⁴(SP800-122)

식별 가능한 개인 정보(potentially identifiable personal information, PI²): Robert Gellman 이 제안한 새로운 유형의 정보로서, 비식별처리되었지만 재식별될 수 있는 정보. “식별 가능한 개인정보란 공개된 식별자가 없는 모든 개인정보를 가리킨다.”¹³⁵ Gellman 의 제안에 따르면, PI2 를 교환하려는 당사자들은 개인정보 접수자가 정보주체의 재식별을 시도하는 경우, 민형사상 처벌을 받도록 하는 규칙에 자발적으로 의탁할 수 있다.

프라이버시(privacy): “개인정보의 부당하거나 불법적인 수집과 이용으로 인한 침입일 경우, 사생활과 개인적 사안에 대한 침입으로부터의 자유”(ISO/IEC 2382-8:1998, 정의 08-01-23)

개인보건정보(protected health information, PHI): “개별적으로 식별할 수 있는 보건 정보: (1) 본 정의의 (2)단락에 제시된 예를 제외한다. 즉 (i) 전자 매체를 통해 전송된 것, (ii) 전자 매체에 유지되는 것, 또는 (iii) 다른 어떠한 형태 또는 매체에 유지되거나 전송되는 것. (2) 개인보건정보는 다음에 포함된 개인적으로 식별 가능한 보건 정보를 제외한다. (i) 가족 교육권 및 프라이버시 수정법, 20 U.S.C. 1232g, (ii) 20 U.S.C. 1232g(a)(4)(B)(iv)에 언급된 개인기록부, (iii) 고용자로서의 역할에 따른 포괄 기관이 유지하는 고용 개인기록부.”(HIPAA 프라이버시 규칙, 45 CFR 160.103)

프라이버시 보호 정보 마이닝(privacy preserving data mining, PPDM): “민감한 정보를 가리기 위해 변경된 개인정보를 다루기 위한 전통적 정보 마이닝 기법의 (연장)”¹³⁶

프라이버시 보호 정보 간행(privacy preserving data publishing, PPDP): “간행된 정보가 실질적인 유용성을 유지하면서 개인의 프라이버시가 보호되도록, 적대적인 환경에서 정보를 간행하기 위한 방법과 도구”¹³⁷

과정(process): “입력물을 출력물로 변환하는 상호 연관되거나 상호작용하는 활동들”(ISO/IEC 27000:2014)

가명처리(pseudonymization): 정보주체와의 연계성을 제거하고, 정보주체에 관련된 특정한

¹³⁴ GAO Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information, May 2008, <http://www.gao.gov/new.items/d08536.pdf>

¹³⁵ Gellman, Robert; The Deidentification Dilemma: A Legislative and Contractual Proposal, 21 Fordham Intellectual Property, Media & Entertainment Law Journal 33 (2010), <http://iplj.net/blog/wp-content/uploads/2013/09/Deidentification-Dilemma.pdf>.

¹³⁶ Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, Computing Surveys, June 2010.

¹³⁷ Ibid.

특징과 하나 이상의 가명 간의 연계성을 추가하는 특정한 유형의 비식별처리.¹³⁸ 통상적으로 직접 식별자를 임의로 생성된 값 등의 가명으로 대체함으로써 가명처리를 이행한다.

가명(pseudonym): “일반적으로 사용하는 개인 식별자와는 다른 개인 식별자”(ISO/TS 25237:2008)

준식별자(quasi-identifier): 간접 식별자(indirect identifier) 참조

접수자(recipient): “정보를 공개하는 자연인 또는 법인, 공공기관, 대리기관 또는 기타 기관”(ISO/TS 25237:2008)

재식별(re-identification): 식별정보와 정보주체 간의 관계를 다시 수립하는 모든 과정을 가리키는 일반 용어

재식별 리스크(re-identification risk): 비식별처리된 개인기록부를 재식별할 수 있는 리스크. 재식별 리스크는 통상적으로 정보 집합 안의 재식별 할 수 있는 개인기록부의 비율로 보고한다.

리스크(risk): “불확실성이 대상에 미치는 영향. 주석: 리스크는 사건의 결과(상황의 변화 포함)와 관련 발생 가능성의 조합으로 표현하는 경우가 많다.”(ISO/IEC 27000:2014)

합성 정보 생성(synthetic data generation): 원료 정보(seed data)를 이용하여 원료 정보와 같은 통계적 특징의 일부를 갖는 인공적 정보를 생성하는 과정

위협(threat): “의도치 않은 사건의 잠재적 원인, 시스템이나 조직에 해를 유발할 수 있다.”(ISO/IEC 27000:2014)

신뢰할 수 있는 정보 접수자(trusted data recipient): 법률, 규정 또는 데이터 이용 합의서와 같은 관리적 통제에 의해 구속된 결과, 자신이 접수하는 정보에 대한 접근성(access)이 제한된 실체.

B.1 표준

- ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and Data

¹³⁸ 주석: 이 정의는 '익명처리'라는 단어가 '비식별처리'로 대체된 것을 제외하고는 ISO/TS 25237:2008 에 있는 정의와 동일하다.

Security Principles for Health Information Including Electronic Health Records

- ISO/IEC 27000:2014 Information technology -- Security techniques -- Information security management systems -- Overview and vocabulary
- ISO/IEC 24760-1:2011 Information technology -- Security techniques -- A framework for identity management -- Part 1: Terminology and concepts
- ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008. 이 ISO 기술규격서는 직접 식별자를 가명으로 대체하는 ‘가명처리 서비스’를 이용하여 프라이버시에 민감한 개인정보를 비식별처리할 수 있는 방법을 기술하고 있다. 본 보고서에 대해 권위 있는 용어와 정의가 포함되어 있다.

B.2 공식 간행물

호주:

- *Privacy business resource 4: De-identification of data and information*, Office of the Australian Information Commissioner, Australian Government, April 2014. http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-resources/privacy_business_resource_4.pdf

EU:

- *Opinion 05/2014 on Anonymisation Techniques*, Article 29 Data Protection Working Party, 0829/14/EN WP216, Adopted on 10 April 2014

영국:

- *Anonymisation: Managing data protection risk, Code of Practice 2012*, Information Commissioner's Office. <https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf>. 108 pages
- UK Anonymisation Network, <http://ukanon.net/>

미국:

- McCallister, Erika, Tim Grance, and Karen Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, Special Publication 800-122, National Institute of Standards and Technology, U.S. Department of Commerce. 2010.
- *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, U.S. Department of Health & Human Services, Office for Civil Rights, November 26, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/hhs_deid_guidance.pdf
- *OHRP-Guidance on Research Involving Private Information or Biological Specimens*, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. <http://www.hhs.gov/ohrp/policy/cdebiol.html>
- *Data De-identification: An Overview of Basic Terms*, Privacy Technical Assistance Center, U.S. Department of Education. May 2013. http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf
- *Statistical Policy Working Paper 22 (Second version, 2005)*, Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology,

December 2005.

B.3 법률 검토 논문 및 백서:

- Barth-Jones, Daniel C., The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now (June 4, 2012). Available at SSRN: <http://ssrn.com/abstract=2076397> or <http://dx.doi.org/10.2139/ssrn.2076397>
- Cavoukian, Ann, and Emam, Khaled, *De-identification Protocols: Essential for Protecting Privacy*, Privacy by Design, June 25, 2014. https://www.privacybydesign.ca/content/uploads/2014/06/pbd-de-identification_essential.pdf
- Lagos, Yianni, and Jules Polonetsky, *Public vs. Nonpublic Data: The Benefits of Administrative Controls*, Stanford Law Review Online, 66:103, Sept. 3, 2013
- *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* Ohm, Paul . UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: <http://ssrn.com/abstract=1450006>
- *Defining Privacy and Utility in Data Sets*, Wu, Felix T. University of Colorado Law Review 84:1117 (2013).

B.4 보고서 및 서적:

- *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015.
- P. Doyle and J. Lane, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, North-Holland Publishing, Dec 31, 2001
- George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice*, Springer, 2011
- Emam, Khaled El and Luk Arbuttle, *Anonymizing Health Data*, O'Reilly, Cambridge, MA. 2013

B.5 설문조사 논문

- Chris Clifton and Tamir Tassa, On Syntactic Anonymity and Differential Privacy, 2013 *Trans. Data Privacy* 6, 2 (August 2013), 161-183.
- Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, *Computing Surveys*, June 2010.
- Ebaa Fayyumi and B. John Oommen, A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. 2010, *Software Practice and Experience*. 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12 <http://dx.doi.org/10.1002/spe.v40:12>

Copyright Notice: Translated with permission, not an official translation of the National Institute of Standards and Technology or the United States Government.