creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

법학석사 학위논문

개인정보 위험기반 비식별 조치와 가명처리

2021 년 2 월

서울대학교 대학원

법학과

김병필

개인정보 위험기반 비식별 조치와 가명처리

지도교수 고학수

이 논문을 법학석사 학위논문으로 제출함
2020 년 11 월

서울대학교 대학원
법학과
김병필

김병필의 법학석사 학위논문을 인준함
2021 년 2 월

위 원 장 _____

부위원장 _____

위 원 _____

국문초록

비식별 조치는 개인정보의 활용과 정보주체 보호를 위해 유용하게 활용될 수 있다. 개인정보처리자는 과학적 연구 활동을 위해서 또는 개인정보를 통계분석하여 그 결과를 통해 의사결정을 보조하기 위한 목적으로 개인정보를 비식별화하여 활용할 수 있고, 나아가 정보주체에 대한 보호를 더욱 강화하기 위해서도 비식별 조치를 활용할 수 있다. 다만, 적정한 개인정보 비식별화를 위해서는 체계적이고 정식화된 절차에 따라 재식별 위험성을 평가하고, 위험성에 비례한 조치를 취할 필요가 있다. 위험기반 비식별 방법론은 위 과정을 통해 데이터 유용성과 개인정보 보호 수준 간의 형량을 하고, 이를 통해 최적의 '골디락스(Goldilocks)' 지점을 찾는 것을 목적으로 한다.

한편, GDPR의 개인정보 개념에 관한 법적 논쟁을 살펴보면, 어떤 정보가 개인정보인지 여부를 판단하는 것은 역동적인 작업임을 알 수 있다. GDPR은 개인정보의 개념 판단에 있어 상대적·주관적 접근을 취한 것으로 해석될 수 있으며, 이는 GDPR 전반에 걸친 위험기반 접근법에 부합하는 것이다. 한편 GDPR에 도입된 '가명처리'의 개념은 약한 가명처리와 강한 가명처리로 구분할 수 있으며, 양자에 부여되는 인센티브는 차이가 있다. 특히 강한 가명처리가 이루어진 경우 정보주체의 권리에 대한 예외가 인정될 수 있음에 비해, 약한 가명처리만이 이루어진 경우에는 이러한 예외가 적용되기 어렵다.

위와 같은 GDPR의 규정은 우리 개인정보 보호법이 도입한 가명처리와 익명처리의 해석에 있어 시사점을 준다. 이에 따라 본 연구에서는 비식별조치를 익명처리, 강한 가명처리, 약한 가명처리로 구분하였다. 이를 간략히 요약하면, ① 익명처리는 개인식별 가능성이 희박(remote)해지도록 개인정보를 변형하는 것으로, 주로 데이터를 공중에 공개하거나 재식

별 공격에 노출될 우려가 있는 상황을 염두에 둔 조치이다. ② 그보다 완화된 조치는 강한 가명처리이다. 강한 가명처리는 해당 정보를 처리하는 자가 체계적이고 확실한 수단에 의해 정보주체를 재식별해 낼 수 없도록 만드는 것으로서, 이는 주로 개인정보를 동의받은 목적 이외의 범위에서 이용하거나, 연구자 등 제3자에게 제공하는 상황을 염두에 둔 조치이다. ③ 약한 가명처리는 개인식별정보를 삭제하거나 암호화하는 방식으로 이루어지는 것으로서, 개인정보처리자가 개인정보 보호 수준을 높이기 위해 채택할 수 있는 조치이다.

본 연구는 비식별 조치의 실무 적용상 어려움을 보여주는 예시로서 종단 데이터의 문제를 검토한다. 종단 데이터에는 정보주체에 관한 많은 정보가 포함되어 있어, k-익명성과 같이 전통적 비식별화 기법을 적용하기 어렵다는 문제가 있다. 더욱이 아직 국제적으로도 체계화된 비식별 조치 방법론이 충분히 개발되어 있지 못한 실정이다. 이러한 점은 국내 비식별 조치 가이드라인에서도 마찬가지로 드러난다. 따라서 종단 데이터에 대해서는 기존 비식별 조치 이외의 대안적 방법, 예컨대 합성 데이터 기법 등이 고려될 필요가 있다.

주요어 : 비식별조치, 가명처리, 익명처리, 재식별화, 종단 데이터, 합성 데이터

학 번 : 2018-29594

목 차

제 1 장 서론 ····· 1

제 2 장 위험기반 비식별 조치 ····· 4

제 1 절 비식별(de-identification) 조치의 목적 ····· 4

1. 오픈 이노베이션과 비식별 조치 ····· 4

2. 데이터 애널리틱스(analytics)와 비식별 조치 ····· 7

3. 정보주체 보호를 위한 비식별 조치 활용 ····· 8

제 2 절 비식별화된 정보의 재식별 위험 ····· 9

제 3 절 위험기반 비식별 조치의 개념 ····· 12

1. 목록기반 접근법 ····· 12

2. 휴리스틱 접근법 ····· 14

3. 위험기반 접근법 ····· 15

제 4 절 위험기반 비식별 조치의 절차 ····· 21

1. 직접식별자, 간접식별자의 선정 ····· 21

2. 재식별 위험 수용 한도의 산정 ····· 24

3. 가능한 재식별 공격 방법의 고려 ····· 27

제 5 절 가명처리 ····· 34

1. 가명을 활용할 필요성 ····· 35

2. 가명정보에 대한 공격 ····· 37

3. 가명처리 개념의 확장 ·····38
제 6 절 소결 ·····39

제 3 장 GDPR과 개인정보 비식별 조치 ·····40
제 1 절 개인정보 판단의 역동성 ·····40
제 2 절 GDPR상의 개인정보 개념과 해석상 쟁점 ·····41
1. GDPR 상의 개인정보의 개념과 개인식별 가능성 ·····41
2. 규범적 판단의 쟁점 ·····43
3. GDPR과 위험기반 접근법 ·····48
제 3 절 GDPR에서의 가명처리 ·····49
1. GDPR 상의 가명처리의 개념: 약한 가명처리와 강한 가명처리 ·····49
2. GDPR상 가명처리 인센티브 ·····53
3. 가명처리와 무관한 의무 ·····57
제 4 절 소결 ·····59

제 4 장 개인정보 보호법과 개인정보 비식별 조치 ··60
제 1 절 개정 개인정보 보호법상의 개인정보의 범위 ·····60
제 2 절 익명처리와 익명정보 ·····62
1. 개정 전 개인정보 보호법과 개인식별 가능성 판단 ···62
2. 개정 개인정보 보호법 상의 해석 ·····66

제 3 절 가명처리와 가명정보 ·····68

1. 가명처리의 정의와 '추가 정보'의 개념 ·····68

2. '가명정보'의 해석 ·····70

3. 가상 사례를 통한 검토 ·····72

제 4 절 소결 ·····79

제 5 장 종단 데이터(longitudinal data)의 문제 ·····82

제 1 절 종단 데이터의 개념 ·····82

제 2 절 종단데이터의 비식별 조치 ·····83

1. 종단 데이터 비식별 조치의 어려움 ·····83

2. k-익명성 기법 적용의 문제점 ·····84

3. 날짜 정보 변경의 필요성 ·····87

제 3 절 국내 가이드라인 적용상 문제점 ·····88

1. 종단 데이터에 대한 익명처리 ·····89

2. 종단 데이터에 대한 가명처리 ·····90

제 4 절 종단 데이터 비식별화의 대안과 합성 데이터 ·····91

제 6 장 결론 ·····93

참고문헌 ·····95

Abstract ·····98

표 목 차

[표 1] 이용자 기본정보 테이블 ·····35
[표 2] 구매이력 테이블 ·····36
[표 3] 학생 기본정보 테이블 ·····72
[표 4] 학점정보 테이블 ·····73
[표 5] 가명학번 대응표(가명처리 비밀) ·····74
[표 6] 가명처리 결과물 ·····74

그 림 목 차

[그림 1] 데이터 유용성과 개인정보 보호수준의 타협 ·····17
[그림 2] 파레토 개선이 가능한 상황 ·····18
[그림 3] 가능한 타협점 간의 비교 ·····19
[그림 4] 맥락에 따른 수용 가능한 재식별 위험 수준의 한도 차이 ·····20
[그림 5] 수용 가능한 위험 한도의 결정 ·····26
[그림 6] k-익명성 설명을 위한 예제 병원 입원 환자 기록 ·····33
[그림 7] 개인정보와 비(非)개인정보의 구분 ·····40
[그림 8] 비식별 조치와 개인정보 해당 여부 ·····79
[그림 9] 환자 내원 이력 데이터 예제 ·····85
[그림 10] k-익명성 적용 예시 ·····85
[그림 11] k-익명성 적용을 위해 변환된 데이터 ·····86

제 1 장 서론

최근 비식별 조치를 통한 개인정보 활용에 대한 관심이 커지고 있다. 이는 빅데이터 활용 가능성이 증대되고, 머신러닝·인공지능 기술이 급속하게 발전하면서 데이터 활용의 중요성이 점차 증가하고 있기 때문이다. 특히 최근 인공지능 기술 발전에 있어 공개된 데이터가 중요한 역할을 한 것으로 이해된다. 전세계의 인공지능 연구자들은 공개된 데이터를 이용하여 공통된 과제 해결을 위해 상호 선의의 경쟁을 펼친다. 다수의 연구자들이 동일한 데이터를 활용하고 같은 과제를 해결하고자 하므로 다른 연구자들과의 성과를 비교할 수 있게 된다. 이를 공통 과제 플랫폼(common task framework)이라 부르기도 하는데, 인공지능 발전에 중요한 역할을 한 것으로 평가된다.[1] 이처럼 공중에 공개된 데이터를 오픈소스(open source) 데이터라고도 하는데, 현재 캐글(kaggle) 등의 인공지능 경진대회 사이트에는 다양한 유형의 오픈소스 데이터가 업로드되어 널리 활용되고 있다. 그러나, 인공지능 학습 데이터를 공개하거나 연구자에게 제공할 때는 이를 통한 개인정보 침해 위험성도 함께 고려될 필요가 있다. 본 연구에서 소개하는 바와 같이 연구 목적으로 데이터를 공개하거나 연구자에게 제공하였으나, 추후 개인정보가 재식별되어 프라이버시 침해 위험성이 밝혀진 사례가 다수 존재한다.

이러한 상황에서 데이터의 유용성을 유지하면서도 정보주체에 대한 프라이버시 보호를 제공할 수 있는 선택지로 비식별 조치가 주목받고 있다. 개인정보의 비식별 조치란 “개인정보에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 활용하여 개인을 알아볼 수 없도록 하는 조치”를 말한다.[2] 즉, 개인정보처리자가 보유한 개인정보를 변형하여 더 이상 개인을 알아볼 수 없도록 만드는 것이다.

개인정보처리자는 정보주체로부터 동의를 얻어 그 목적 범위 내에서 개

1) Donoho, David. "50 years of data science." Journal of Computational and Graphical Statistics 26.4 (2017): 745-766.

2) 개인정보 비식별 조치 가이드라인(2016), 제3면.

인정보를 이용하거나 정보주체의 동의를 얻어 제3자에게 제공하면 되는 것이지, 굳이 비식별 조치를 취할 필요가 있는지 의문이 들 수 있다. 그래서 비식별 조치를 개인정보 보호법을 우회하기 위한 수단이라고 볼 여지도 있다.[3] 다른 한편 개인정보처리자의 입장에서도 개인정보에서 개인을 식별할 수 있는 요소를 삭제한다면 그 데이터가 갖는 가치가 줄어들 수밖에 없다.

이러한 한계에도 불구하고, 개인정보에 대해 적정한 비식별 조치를 취하여 활용함으로써 얻을 수 있는 중대한 장점은 개인정보 침해 위험을 상당히 낮출 수 있다는 것이다. 데이터 유출 사고가 발생하더라도 적정한 비식별 조치가 이루어졌다면 정보주체에 대한 직접적 위험은 줄어든다. 이러한 점에서 유럽연합의 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (이하 "**GDPR**")은 비식별 조치를 통한 데이터 처리에 여러 인센티브를 부여하고 있다. 우리 개인정보 보호법도 2020. 2. 개정을 통해 가명처리와 가명정보의 개념을 도입하였고, 개인정보의 범위를 보다 합리화하는 규정을 신설하는 등, 개인정보를 비식별화하여 활용하는 것을 장려하고 있다. 이에 본 연구에서는 비식별 조치의 방법론과 그 법적 쟁점을 검토하고자 한다.

본 연구의 구성은 다음과 같다. 우선 제2장에서는 비식별 조치의 목적과 비식별화된 정보의 재식별 위험성을 살핀다. 이를 통해 적정한 개인정보 비식별화를 위해서는 체계적이고 정식화된 절차에 따라 재식별 위험성을 평가하고, 위험성에 비례한 조치를 취할 필요가 있다는 점을 보일 것이다. 이러한 목적을 달성하기 위한 방안으로 '위험기반 비식별 조치 방법론'을 구체적으로 소개한다. 위험기반 비식별 조치는 데이터 유용성과 개인정보 보호 수준 간의 형량을 하고, 이를 통해 최적의 지점을 찾아 내는 것을 목적으로 한다. 제3장에서는 위험기반 비식별 조치가 GDPR 하

3) 함께하는 시민행동, "고객 몰래 개인정보 결합한 기업과 공공기관에 대한 고발", 기자회견문, http://www.privacy.or.kr/archives/5642 (2020. 11. 20. 접속)

에서 어떻게 이해되고 수용될 수 있는지는 검토한다. 특히 GDPR 상의 개인정보 범위에 관한 법적 쟁점과 가명처리의 개념을 살펴봄으로써 우리 개인정보 보호법 해석을 위한 시사점을 찾는다. 제4장에서는 앞선 GDPR에 관한 분석을 토대로 국내 개인정보 보호법상의 비식별 조치에 대한 체계적 해석을 시도한다. 특히 본 연구에서는 비식별조치를 익명처리, 강한 가명처리, 약한 가명처리로 구분하고, 각각의 조치들이 어떠한 법적 함의를 갖는지 제시하고자 한다. 마지막으로 제5장에서는 비식별 조치의 구체적 적용 사례로서 종단 데이터의 문제를 검토한다. 실무상 활용되는 데이터의 상당수는 종단 데이터임에도 아직 체계화된 비식별 조치 방법론이 충분히 개발되어 있지 못한 실정이다. 국내 비식별 조치 가이드라인 상의 종단 데이터에 대한 비식별 조치에 관한 규율 내용을 살핀 후, 그 대안을 모색한다.

참고로, 본 연구에서는 '비식별 조치(또는 비식별화, de-identification)'를 '익명처리(또는 익명화, anonymization)'와 '가명처리(또는 익명화, pseudonymization)'를 포함하는 상위 개념으로 사용하다. 용례에 따라서는 비식별 조치와 익명처리를 동의어로 보는 경우 또는 상호 구분되는 개념으로 파악하는 경우 등도 있다.[4] 하지만 개정법상 '가명처리'와 '가명정보'에 관한 정의 조항이 도입되었고, 나아가 개정법 제58조의2 상의 "시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아 볼 수 없"도록 하는 조치를 강학상 '익명처리'로 칭하는 것이 일반화되어가고 있는 것으로 보인다. 따라서 양자를 포괄하여 '비식별 조치'라고 부르는 것이 적절하다고 본다.

4) 각 용어의 비교에 관해서는 관하여는 고학수 외, "개인정보 비식별화 방법론: 보건의료정보를 중심으로", 박영사, (2017) 제195면 이하 참조.

제 2 장 위험기반 비식별 조치

제 1 절 비식별(de-identification) 조치의 목적

개인정보처리자는 비식별 조치를 취하는 이유는 크게 나누어 보면 ① 개인정보 보호법의 적용을 받지 않으면서 해당 정보를 활용하기 위한 경우와 ② 정보주체에 대한 보호를 더욱 강화하기 위한 경우로 나누어 볼 수 있다. 전자의 경우는 기술 개발 등의 연구 활동을 위해 개인정보를 비식별화하여 활용하는 경우와 개인정보를 통계 분석하는 등 그 결과를 이용하여 의사결정을 보조하기 위한 경우로 세분화해 볼 수 있다.

1. 오픈 이노베이션과 비식별 조치

비식별 조치가 이루어진 정보를 활용하고자 하는 대표적인 사례로는 새로운 제품이나 서비스의 개발을 위해 활용하는 경우가 있다. 이는 기업의 '오픈 이노베이션(open innovation)' 전략과 밀접하게 관련되어 있다. 전통적으로 기업의 기술 개발 전략은 자신의 연구개발부서를 통해서 이루어져 왔고 이를 통해 개발된 기술은 영업비밀로 유지되거나 특허권을 취득하여 독점적 지위를 보장받는 방식으로 보호가 이루어져 왔다. 그러나 기업들은 점차 외부 기관이나 개인들과의 협력 관계를 통해 새로운 아이디어를 수용해야 필요성을 느끼게 되었다. 이처럼 "기업이 자신의 기술을 발전시키기 위하여 내부 아이디어뿐만 아니라 외부의 아이디어를 활용하는 패러다임"을 흔히 '오픈 이노베이션'이라 한다. 그런데, 제품이나 서비스 개발을 위해서 기업이 보유한 개인정보를 활용해야 하는 경

우, 기업은 개인정보 비식별 조치를 취함으로써 해당 정보를 기업 외부자에게 제공할 수 있게 된다.

오픈 이노베이션의 장점을 잘 보여 준 사례로는 넥플릭스사 의 'Netflix Prize' 경진대회가 있다.[5] 넥플릭스 사는 2006년 10월 자사의 영화 추천 알고리즘을 10% 개선할 경우 100만 달러의 상금을 지급하는 경진대회를 개최했다. 'Netflix Prize' 경진대회가 시작되었을 당시에는 성능 10% 개선이라는 목표가 달성되기란 어렵다는 전망이 많았다.[6] 넥플릭스의 연구자들은 영화 추천 알고리즘의 개선을 위해 오랫동안 상당한 자원을 투여 왔다. 그런데 외부 연구자들이 참여한다고 하여 손쉽게 알고리즘의 성능 개선이 이루어질 것이라고 기대하기는 쉽지 않았을 것이다.

넥플릭스 사가 활용한 영화 추천 알고리즘은 여러 이용자가 영화에 대해 매겨 놓은 영화 평점을 기반으로 하여 이용자의 영화에 대한 선호도를 추정하는 방식이다.[7] 따라서 이러한 알고리즘을 개발하기 위해서는 넥플릭스 사가 보유한 영화평 데이터를 공개할 필요가 있었다. 이에 넷플릭스 사는 약 48만명의 이용자의 17,000여 건의 영화에 대한 평가 데이터 약 1억 건을 "Netflix Prize" 웹사이트를 통해 공개했다. 전세계의 연구자는 누구라도 위 웹사이트로부터 데이터를 다운로드 받아 새로운 영화 추천 알고리즘을 개발할 수 있다. 알고리즘의 성능은 공개되지 않은 테스트 데이터셋 140여만 건을 통해 이루어졌다. 따라서 경진대회의 목적은 테스트 데이터셋에 포함된 실제 이용자 평가 결과와의 오차를 최소화하는 알고리즘을 개발하는 것이었다.

위 대회를 위해 데이터를 공개하기에 앞서 넥플릭스는 자신이 보유한 이용자 영화평 데이터에 대해 비식별조치를 취하였다. 넥플릭스가 공개한 데이터는 이용자 ID, 영화 ID, 평가일, 평점(1~5점) 1억여건으로 구성되어 있다. 이때 이용자 ID는 넥플릭스가 보유한 실제 이용자 ID가 아니라 임의의 일련번호로 변형되었다. 즉, 개인정보의 일부를 대체함으로써 개

---

5) https://en.wikipedia.org/wiki/Netflix_Prize (2020. 11. 20. 접속)

6) Kate Greene, "The $1 Million Netflix Challenge", MIT Technology Review (2006. 10. 6.)

7) 이러한 추천 알고리즘을 '협업 필터링(collaborative filtering)'이라 부른다.

인을 알아볼 수 없도록 만든 것이다.

넥플릭스의 경진대회는 우선 홍보 목적에 있어 대단한 성공을 거두었다. 전세계 186개국에서 41,305팀의 51,051명이 참여하였다. 그보다 더 놀라운 것은 원래 예정했던 기간보다 훨씬 빨리 목표를 달성할 수 있었다는 점이다. 넥플릭스는 당초 대회 기간을 5년으로 정해 두었는데, 3년도 되지 않아 목표한 10%의 성능 향상을 이루었다. 이처럼 넷플릭스의 경진대회는 기업 외부자 다수의 참여를 이끌어냄으로써 놀라운 기술 혁신을 달성할 수 있다는 점을 보여 주었다.

넥플리스의 성공에 자극받아 다른 경진대회도 이어졌다. 예를 들면 캘리포니아의 의료보험 및 의료기관 네트워크인 Heritage Provider Network (HPN)는 2011년 4월 의료정보 분석 경진대회를 개최하였다.[8] 그 목적은 의료보험 청구 데이터를 이용하여, 다음 연도에 병원에 입원할 가능성이 높은 환자를 예측하는 것이었다. 이러한 예측이 정확히 이루어진다면 의료비를 많이 지출할 가능성이 높은 환자들을 미리 선별하여 사전적인 조치를 취함으로써 전체 의료보험 지출을 줄일 수 있게 된다. 이 대회를 위해서 HPN은 환자 113,000명의 보험청구 2,668,990건에 대해 비식별 조치를 취하여 이를 공개하였다.

비식별 조치는 개인정보를 일반 공중에 공개하는 경우뿐만 아니라 기업 외부의 연구자에게 제공할 때에도 활용된다. 의료기관이 보유한 환자에 대한 의료정보를 제약사 또는 의료 연구기관 등에 제공하는 경우가 대표적이다. 그런데 개인정보 보호법상 개인정보를 제3자에게 제공하기 위해서는 법률에 특별한 규정이 있는 경우 등 예외적 사유를 제외하고는 정보주체로부터 별도의 동의를 얻어야 한다.[9] 하지만 개인정보처리자가 제3자에게 개인정보를 제공하기 위해서는 새로운 동의를 얻기 어려운 경우가 적지 않다. 개인정보처리자는 이러한 어려움을 덜기 위하여 개인정보

---

8) El Emam, Khaled, et al. "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset." Journal of medical Internet research 14.1 (2012); 고학수·최경진, "개인정보의 비식별화 처리가 개인정보 보호에 미치는 영향에 관한 연구", 개인정보보호위원회 연구보고서, (2015), 29-35면은 위 사례를 상세하게 소개하고 있다.

9) 개인정보 보호법 제18조 제2항 제1호.

에 대해 비식별 조치를 취하여 더 이상 개인정보가 아닌 형태로 변형함으로써, 정보주체의 동의 없이 이를 제3자에게 제공하고자 할 수 있게 된다.

2. 데이터 애널리틱스(analytics)와 비식별 조치

개인정보처리자가 개인정보 보호법의 적용을 받지 않으면서 해당 정보를 활용하기 위해 비식별 조치를 활용하는 다른 사례로는 데이터 애널리틱스를 들 수 있다. 데이터 애널리틱스는 원시 데이터를 분석하여 새로운 유용한 정보를 발견하거나, 의사결정을 보조하는데 활용하는 것을 말한다. 최근 들어 빅데이터가 광범위하게 축적되고, 그 활용 가능성이 증가하면서 그 유용성이 주목받고 있다. 보다 구체적으로는 데이터 활용의 단계를 세분화한다면, 데이터를 이용하여 현재 상황을 파악하는 설명(descriptive) 단계, 데이터를 이용하여 발생한 사건의 원인을 파악하는 진단(diagnotic) 단계, 데이터를 이용하여 장래 발생한 사건을 예측하는 예측(predictive) 단계, 데이터 분석 결과를 이용하여 종국적으로 어떠한 행동을 해야 할 것인지를 알려주는 처방(prescriptive) 단계로 구분할 수 있다.[10]

예를 들어, 법률 서비스 분야에서 이루어지는 데이터 애널리틱스를 리걸 애널리틱스(Legal Analytics)라 한다.[11] 주로 변호사가 사건 전략 수립을 위하여 판결문 등으로부터 통계 정보를 추출하여 활용하는 형태로 이루어진다. 예컨대 미국의 Lex Machina 사는 재판지별 승소율, 법관별 사건 처리 기간, 신청 사건에 대한 인용률 등의 통계 정보를 제공함으로써 변호사의 소송 활동을 돕는다. 국내에서도 2020년 11월 형량 예측 서비스가 개시되었는데, 이 서비스는 이용자가 사건의 유형(범죄, 행위 태양 및 양형 요소)를 입력하면 그 사건에서 선고된 형량을 1순위부터 3순위까지 보여준다.[12]

---

10) Gartner, “A Framework for Applying AI in the Enterprise”, (2019. 1. 29.)
11) 고학수 외, “사법부에서의 인공지능(AI) 활용방안”, 법원행정처, (2020), 143면.
12) https://www.lawtalk.co.kr/sentence-estimation

이처럼 데이터 애널리틱스는 특정 개인에 관한 추론을 하는 것이 아니라 총량, 평균, 분산 등의 통계치를 이용하는 경우가 많다. 이때에는 개인정보를 비식별화함으로써 해당 정보를 활용할 수 있게 된다. 가령 형량 예측 서비스에 있어서는 피고인이 자신에 대한 판결문을 통계 처리하여 활용하는 것에 대해 동의하였는지와 무관하게 특정 유형의 형사사건에 대한 형량 분포를 파악하여 제시할 수 있게 된다.[13)]

데이터 애널리틱스는 정부에 의해서도 유용하게 활용될 수 있다. 최근 들어 증거에 기반한 정책 수립(Evidence-based Policymaking)의 필요성이 강조되면서, 특히 빅데이터 활용의 유용성이 크다고 주장된다.[14)] 이는 빅데이터를 통해 정책 수립 방향을 정하고, 정책의 효과를 예측하거나, 구체적 해결책을 제시할 수 있을 것이라는 주장이다. 나아가 조직, 인사, 재정 등 정부 내부 작용에 있어서도 데이터 애널리틱스가 유용하게 활용될 수 있을 것으로 전망된다.

3. 정보주체 보호를 위한 비식별 조치 활용

개인정보처리자가 정보주체를 보호하기 위하여 비식별 조치를 활용하는 것이 바람직한 경우도 있다. 대표적인 사례가 개인정보처리자가 내부적인 이용자 프라이버시 보호를 강화할 목적으로 비식별 조치를 사용하는 것이다. 예를 들어 IT 개발부서가 개인정보 처리 소프트웨어를 개선하는 업무를 하는 경우를 상정해 보자. 이때 개발부서가 반드시 정보주체에 대해 진실된 정보를 보유하고 있을 필요까지는 없고, 실제와 유사한 데이터가 있으면 되는 경우도 많다. 하지만 개발자는 실제 상황과 동떨어진 가상 데이터를 활용하기도 어렵다. 실제 소프트웨어의 구동 상황에서 적합하게 작동하는지 확인할 필요가 있기 때문이다. 이러한 경우 좋은 관행은 비식별 조치를 수행하여 실제 이용자의 개인정보와 유사한 형식

13) 참고로, 형사소송법 제59조의3 제2항은 확정된 사건의 판결서 등을 열람 및 복사하기에 앞서 판결서 등에 기재된 개인정보가 공개되지 아니하도록 대법원규칙으로 정하는 보호조치를 하도록 정하고 있다.

14) 오철호, “정책결정, 증거 그리고 활용 - 연구경향과 제언 - ”, 한국정책학회보, 제24권 1호, (2015), 54-55면.

를 취하지만 개인을 식별할 수 없는 형태의 데이터베이스를 활용하는 것이다. 기업이 이러한 조치를 취한다면, 개발부서 직원이 임의로 유명인의 개인정보를 검색해 보는 시도를 막을 수 있다. 만에 하나 개발부서에서 개인정보가 유출되더라도 그 위험의 범위를 줄일 수 있다는 장점도 있다.

현재 논의되고 있는 더욱 발전된 형태의 비식별 조치 활용 가능성은 개인정보에 접근하는 직원의 권한에 따라 상이한 수준의 비식별조치를 적용하는 방법이다.[15] 예컨대 의료기관에 있어 환자를 담당하는 의사, 간호사, 행정직원이 그 환자의 모든 개인정보에 대해 접근할 수 있도록 하는 것이 아니라, 해당 업무 처리에 필요한 수준의 범위에서만 접근할 수 있도록 하는 것이다. 이를 통해 정보주체의 프라이버시 보호를 강화할 수 있음은 물론이다. 다만, 이러한 기능을 구현하기 위해서는 개인정보를 저장한 서버 데이터베이스에 자료 조회 요청이 이루어졌을 때 실시간으로(on-the-fly) 비식별 조치를 수행해야 하는데 아직까지는 기술적으로 용이하지 않다. 그러나 앞으로 정보주체의 프라이버시 보호를 위해 다양한 형태의 비식별 조치 기술이 도입될 여지가 있다.

제 2 절 비식별화된 정보의 재식별 위험

이처럼 개인정보 비식별화 조치는 개인정보 데이터를 활용할 수 있는 범위를 확장하고, 정보주체 보호를 위해 활용될 수 있다는 순기능이 있다. 그러나 개인정보에 대하여 비식별 조치를 취하더라도 개인을 다시 식별해 낼 위험성을 완전히 제거하는 것은 매우 어렵다. 즉 비식별 조치는 항상 재식별 위험을 수반하는 것으로 이해될 수 있다.

비식별 조치된 정보의 재식별 위험성을 국제적으로 알린 사례로는 매사

15) El Emam, Khaled, and Luk Arbuckle, “Anonymizing health data: case studies and methods to get you started”, O′Reilly Media, Inc., (2013), p.222.

추세츠 주지사의 의료정보 재식별 사건이 있다.[16] 매사추세츠 주정부는 단체보험위원회(Group Insurance Commission, GIC)를 통해 주 공무원들에게 건강보험을 제공하고 있었다. 매사추세츠 주 정부는 1990년대 중반 약 135,000명의 주 공무원들과 그 가족의 의료정보를 연구 목적으로 공개하였다. 이 때 GIC는 이름, 주소, 사회보장번호 등 개인을 식별할 수 있을 것으로 보이는 정보를 삭제하였다. 따라서 주 정부는 GIC의 의료정보를 통해 개인을 식별할 수 없을 것으로 기대하고, 위 정보를 연구자들에게 제공하거나 산업계에 유상으로 판매하였던 것이다.

그런데 당시 대학원생이던 Sweeney는 매사추세츠 캠브리지 지역의 유권자 명부를 구입하여 GIC 의료정보와 대조하는 작업을 수행하였다. GIC 의료정보에 포함된 우편번호, 생일 및 성별 정보는 유권자 명부에도 포함되어 있었으므로 양자를 서로 비교하여 연결(link)이 가능하였다. 그 결과 당시 매사추세츠 주지사인 William Weld의 의료기록을 재식별해 낼 수 있었다. 즉, 매사추세츠 주 캠브리지 지역의 유권자 명부에는 6명이 주지사와 생일이 일치하였고, 그 중 3명만이 남자였고, 해당 우편번호에 거주하는 사람은 주지사가 유일하였다. 이처럼 간단한 대조만으로 GIC 의료정보로부터 개인을 재식별낼 수 있었다는 점 및 그 재식별 대상이 주지사 본인이었다는 점에서 이는 놀라운 사건이었다.

Weld 주지사의 의료기록 재식별 사례는 적절한 비식별 조치를 취해지지 못할 경우의 위험성을 알리는 중요한 국제적 선례가 되었다. 국내 개인정보 비식별조치 가이드라인(2016)의 36-37면에서 공개 의료데이터와 선거인명부을 예시로 들어 개인정보 재식별 사례를 설명하고 있는 점도 이 사례의 영향력을 반증한다.

그 이후에도 공개 데이터에서 개인을 성공적으로 재식별해 낼 수 있다는 점을 보여 주는 다른 사건들이 이어졌다. 그 대표적인 사례가 넥플릭스 사건이다. 앞서 소개한 바와 같이 'Netflix Prize' 경진대회에서 넥플릭스는 이용자 ID를 임의의 값으로 된 일련번호로 대체하였다. 더욱이 넥플

16) Sweeney, Latanya, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002), pp. 557-570; 고학수·최경진(2015) (각주 8) 36-38면은 위 사례를 소개하고 있다.

릭스가 공개한 데이터에는 단지 어떤 이용자가 언제 어떤 영화에 대해 몇 점의 평점을 주었는지에 관한 정보만이 포함되어 있을 뿐이었다. 따라서 넥플릭스 입장에서는 자신이 공개한 데이터를 통해 이용자를 재식별해 내고, 그 결과 이용자에 대한 프라이버시 침해 위험이 발생할 것이라고는 생각하기란 쉽지 않았을 것이다.

하지만, Netflix Prize 데이터가 공개된 지 얼마 지나지 않아, 재식별 공격에 성공했다는 결과가 발표되었다.[17] 이러한 재식별 공격은 이용자가 Internet Movie Database(IMDb)에 영화평을 공개한 데이터를 배경지식(background knowledge)로 하여 이루어진 것이었다. 즉, 이용자가 IMDb에 올려놓은 영화평과 해당 영화에 대한 넥플릭스 평점을 상호 연결함으로써 해당 이용자를 재식별해 낸 것이다(IMDb에는 평화 평을 올린 이용자 ID가 공개되어 있으므로, 이를 이용하여 해당 이용자에 관한 더 많은 정보를 확인할 가능성이 열리게 된다).

정보주체의 영화 평점에 관한 정보가 얼마나 보호할 가치가 있는 것인지 의문이 들 수도 있다. 그러나 영화에 대한 선호를 통해서 그에 관한 민감정보를 추론해 낼 수 있다. 예컨대 특정 종교적 이야기를 다룬 영화, 정치적 다큐멘터리, 동성애를 다룬 영화에 대한 평점을 통해 그의 종교, 정치적 성향, 성적 지향을 추론해 내는데 활용될 수 있다.[18] 그 결과 미국의 한 여성 동성애자는 2009년 12월 넥플릭스 사를 상대로 캘리포니아 북부지구 연방지방 법원에 프라이버시 침해를 원인으로 한 집단소송을 제기하기도 하였다.[19] 이 소송에서 원고는 자신이 동성애를 주제로 한 영화에 관심이 있다는 사실이 공개되었다고 주장하였다(이 소송은 합의를 통해 종결되었다). 또한 미국 연방거래위원회(Fair Trade Commission, FTC)가 넥플릭스 사를 접촉하여 경진대회에 관해 조사하기도 하였다고 알려져 있다. 당초 넥플릭스 사는 2009년 8월 후속 경진

17) Narayanan, Arvind, and Vitaly Shmatikov, "Robust de-anonymization of large sparse datasets", 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, (2008); 고학수·최경진(2015) (각주 8) 39-41면은 위 사례를 소개하고 있다.

18) 위의 글., p. 123.

19) https://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf (2020. 11. 20. 접속).

대회를 준비하겠다고 발표하였으나, 위와 같은 사정을 고려하여 후속 대회를 취소하였다.[20]

이러한 일련의 사건들은 개인정보 비식별 조치에 관한 관심을 가져왔다. 그 결과 국제적으로 재식별 위험성 평가 방법 및 위험성을 고려한 비식별 조치 방법에 관한 연구 및 실무 관행이 발전하게 되었다.

제 3 절 위험기반 비식별 조치의 개념

위험기반 비식별 조치 방법론은 재식별 위험 수준을 계량적으로 평가하여 그에 비례한 비식별 조치를 취하는 것을 의미한다.[21] 그 이해를 위해서는 다른 비식별 조치와 비교하는 것이 유용하다. 이하에서는 목록(list) 기반 접근법과 휴리스틱(heuristic) 접근법과의 비교를 통해 위험기반 접근법을 설명한다.

1. 목록기반 접근법

목록기반 접근법은 제거을 해야 할 식별자 항목의 목록을 사전에 정의해 놓고, 이를 삭제한 경우 더 이상 개인정보에 해당하지 않는다고 간주하는 것이다.[22] 목록기반 접근법의 대표적 사례로는 미국 Health Insurance Portability and Accountability Act(HIPAA)의 Safe Harbor 기준이 있다.[23] HIPAA Safe Harbor 기준은 총 18개의 식별자, 즉 이름, 주(state) 단위보다 작은 지리적 위치, 개인과 연관된 날짜들, 전화번호,

---

20) Ryan Singel, “NetFlix Cancels Recommendation Contest After Privacy Lawsuit”, Wired, 2018. 12. 3. https://www.wired.com/2010/03/netflix-cancels-contest/ (2020. 11. 20. 접속)

21) El Emam, Khaled and Luk Arbuckle (2013) (각주 15), p. 34.

22) 위의 글., p. 14

23) 김은수, “비식별화 방식을 적용한 개인정보보호에 대한 연구”, 서울대학교 박사학위 논문, (2018), 130-132면.

팩스번호, 이메일 주소, 사회보장번호, 의료기록번호, 건강보험 수혜자 번호, 계좌 번호, 운전면허번호, 자동차식별자와 일련번호, 장치 식별자와 일련번호, URL, IP주소, 지문 및 성문(聲文) 등 생체인식(biometric) 식별자, 전체 얼굴 사진과 유사 이미지, 기타 유일한 식별 번호·문자·코드를 나열하고 있다. 특히 지리적 위치나 개인과 연관된 날짜에 관해서는 제거해야 하는 정보의 범위를 구체적으로 제시하고 있다. 예컨대, 주소의 경우 거주자가 20,000명을 초과하는 지역에 대해서는 우편번호 앞 3자리까지 기재하는 것은 허용하되, 20,000명 이하의 지역에 대해서는 '000'으로 변경해야 한다. 또한 생일, 입원일자, 퇴원일자, 사망일자 등 개인과 직접 관련된 날짜는 연도만을 남겨둘 수 있고, 89세를 초과하는 나이를 암시할 수 있는 경우에는 90세 이상이라는 단일 범주로 일반화해야 한다.

목록기반 접근법은 수범자의 입장에서 비용이 낮은 비식별 조치 방법이다. HIPAA Safe Harbor의 경우 수범자는 위 18가지 식별자를 정해진 규칙대로 제거하면 개인정보에 해당하지 않게 된다.[24] 이와 유사한 목록기반 접근법은 의료 분야에 있어 전세계적으로 널리 활용되고 있다.[25] 하지만 목록기반 접근법에 대해서는 2가지 비판이 제기된다. 우선 목록에 정의된 제거 대상 식별자의 범위가 지나치게 넓은 경우 데이터 분석 시의 유용성이 낮아질 수 있다는 점이다.[26] 또한, 목록에 정의된 식별자를 정의하더라도 재식별 위험성이 여전히 큰 경우가 존재하여 정보주체에 대한 충분한 보호가 이루어지지 못할 수 있다는 상반된 문제도 존재한다.[27]

한편, HIPAA Safety Harbor 비식별화 기준에 대해서는 재식별 위험에 대한 '실제 지식(actual knowledge)'이 없어야 한다는 요건이 존재한다는

---

24) 다만 "기타 유일 식별 번호·문자·코드(Any other unique identifying number, characteristic, or code)"의 경우 해석의 여지가 열려 있다.

25) El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p. 14에 따르면, 캐나다에서는 연구, 정부, 민간기업 가이드라인에 이러한 방식이 규정되어 있고, European Medicines Agency는 임상 실험에 있어 최소한의 기준으로 이러한 접근법을 고려하고 있다.

26) 김은수(2018) (각주 23), 135면

27) El Emam, Khaled and Luk Arbuckle (2013) (각주 15), p. 15.

점도 유의할 필요가 있다. 즉, HIPAA Safe Harbor에 따르면, 규정된 식별자를 제거하는 것에 추가하여 다른 정보를 이용하여 개인정보 주체를 식별할 수 있다는 점에 대해 실제 지식이 없어야 한다.[28] 여기서 실제 지식의 범위는 개인정보가 처리되는 맥락에 따라 달라진다. 예를 들면 Safe Harbor 규정에 규정된 18개 항목 중에는 '직업'은 포함되어 있지 않다. 하지만 만약 어떤 환자의 직업이 'OO시 시장'이라면, 해당 데이터는 당연히 삭제되어야 할 것이다. 나아가 데이터를 제공받는 연구자의 가족에 관한 의료정보가 해당 데이터에 포함되어 있다는 사실을 알고 있는 경우에는 해당 가족에 대한 의료기록은 삭제되거나, 식별할 수 없을 정도로 변형될 필요가 있다. 또한 언론을 통해 자녀가 많은 것으로 잘 알려진 경우에는 자녀 수에 대한 정보를 삭제해야 한다. 이와 같이 목록 기반 접근법에 의하더라도 '실제 지식'을 고려하여 재식별 위험에 대한 평가가 제한적으로나마 이루어질 필요가 있다.

2. 휴리스틱 접근법

휴리스틱(heuristics)[29] 접근법은 자신의 경험이나 지식에 기반하여 적용하는 대강의 규칙(rules of thumb)을 적용하는 것을 의미한다.[30] 예컨대 의료정보에서 환자의 생일은 삭제하고, 진료 일자나 방문 일자는 포함하되, 희귀 질병의 경우는 진료 일자나 방문 일자도 포함시키는 것과 같은 것이다. 예컨대 앞서 예로든 매사추세츠 주지사 사건에서 GIC가 이름, 주소, 사회보장번호 등 개인을 식별할 수 있을 것으로 보이는 정보를 선별하여 삭제한 사례, Netflix Prize 데이터셋에서 이용자의 ID는 임의의 일련 번호로 대체하되, 영화명은 그대로 유지한 사례는 휴리스틱 접근법에 따른 것이라고 평가할 여지가 있다. 비식별 조치를 수행하는 자는 자

28) 45 CFR. § 164.514(b)(2)(ii), "수범기관은 나머지 정보가 그 자체로 또는 다른 정보와 결합하여 정보의 주체가 되는 개인을 식별할 수 있다는 실제 지식(actual knowledge)이 없어야 한다."

29) 휴리스틱(heuristics)은 경험에 기반하여 문제를 해결, 학습 또는 발견하는 방법을 의미하는데, 어림짐작법, 어림법, 주먹구구법, 발견법 등으로 다양하게 번역된다.

30) El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p. 15.

신의 경험이나 지식에 기반하여, 자신이 보기에 적절하다고 판단되는 일응의 규칙을 적용한 것이기 때문이다.

하지만, 휴리스틱 접근법은 왜 그러한 방법을 취했는지, 왜 다른 방법은 취하지 않았는지 사후적으로 정당화하는 것이 불가능하다는 문제가 있다. 예를 들면 Netflix Prize 데이터셋은 영화명 정보를 포함하고 있었다. 그러나 데이터셋에서 영화에 대해서까지도 모두 임의의 일련번호로 대체하여 공개하는 것도 가능했을 것이다.[31] 그렇다면 IMDb에 공개된 영화 평과의 연결 공격은 훨씬 더 어려워졌을 것이다.

나아가 휴리스틱을 활용하면 개인정보처리자가 재식별 위험의 수준을 적절하게 관리하기 어렵다. 즉 비식별화된 정보가 누구에게 제공되는지(일반 공중에 공개되는지, 아니면 특정 연구자에게만 공개되는지, 특정 연구자에게만 공개되는 경우 그 수령자의 개인정보 보호수준 및 다른정보 보유여부 등)을 체계적으로 고려하지 못하므로, '주먹구구'식으로 비식별 조치가 이루어지기 쉬운 것이다. 이러한 점에서 휴리스틱 접근법은 권장할 만한 방법으로 평가하기 어렵다.[32]

3. 위험기반 접근법

위험기반 접근법은 목록 기반 접근법과 휴리스틱 접근법을 극복하고 재식별 위험을 고려한 체계적인 개인정보 비식별 조치 방법론을 제시하는 것을 목표로 한다. HIPAA 프라이버시 규칙의 전문가 판단 방법(expert determination method)이나 여러 통계적 방법들은 이러한 위험기반 접근법에 부합하는 것으로 볼 수 있다.[33]

El Emam & Arbuckle (2014)는 위험기반 접근법이 네 가지의 전제에 기반한 것이라고 설명한다.[34] 첫째 전제는 재식별 위험을 계량할 수 있

---

31) 넥플릭스 경진대회는 협업 필터링 알고리즘을 적용하므로, 굳이 영화명을 공개하지 않더라도 연구자들이 알고리즘을 개선하는 것이 가능한 상황이었다. 다만, Narayanan and Shmatikov (2008) (각주 17)은 영화명 정보를 삭제하였더라도 재식별이 불가능한 것은 아닐 것이라 평가한다.

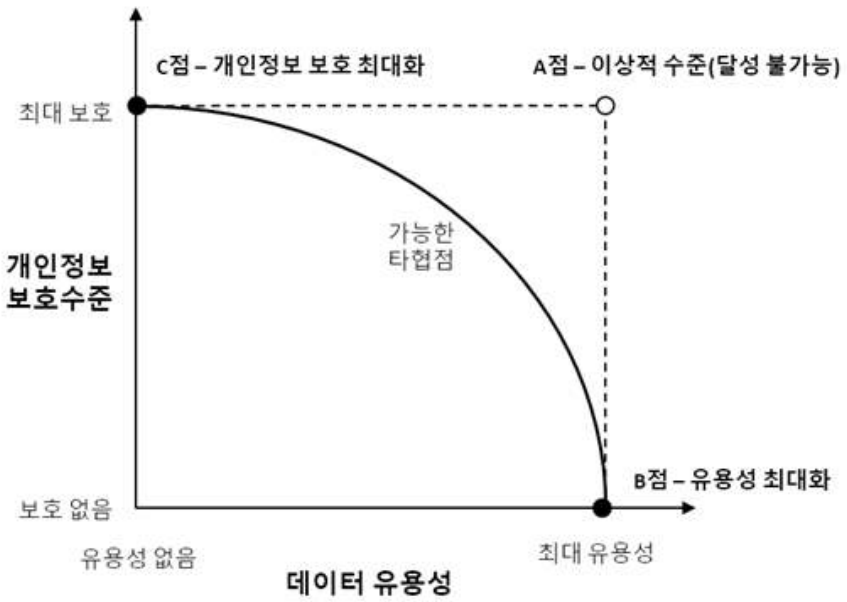
32) El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p. 15.

33) 위의 글. p. 16

다는 것이다. 즉, 재식별 위험을 계량적으로 산정함으로써 기술적으로 비식별 조치가 어느 수준에서 이루어져야 할지 결정할 수 있다고 본다. 재식별 위험을 산정하기 위해서는 데이터셋에 포함된 정보가 정확한 것이고, 재식별 공격자가 정보주체에 관해 상당한 배경지식을 보유하고 있는 이론적 상황을 가정한다. 이러한 이론적 상황 하에서 재식별 공격자가 정보주체를 식별해 낼 확률을 계산함으로써 재식별 위험을 계량적으로 평가할 수 있게 된다.

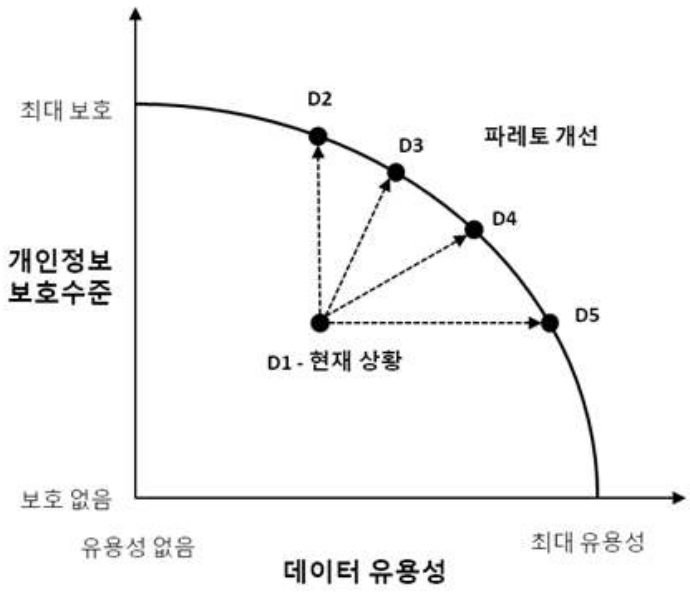
둘째 전제는 데이터 유용성과 개인정보 보호 간의 형량을 통해 중간점을 찾아낼 수 있다는 것이다. 아래 <그림 1>의 X축은 데이터 유용성이고 Y축은 개인정보 보호수준이다. 이상적으로는 A점과 같이 데이터 유용성과 개인정보 보호 수준을 모두 최대화할 수 있으면 좋겠으나, 이를 달성하는 것은 불가능하다. 따라서 프라이버시 보호와 데이터 유용성 사이의 중간점을 찾아야 한다. 만약 개인정보 보호를 포기하고, 아무런 제약없이 데이터를 활용할 수 있게 한다면 그림 우하단의 B점에 위치할 것이다. 반대로 개인정보 보호를 최대화하고 데이터를 활용하지 않는다면 좌상단의 C점에 위치할 것이다. 그 사이에 존재하는 연결선은 수용 가능한 타협점이 된다.

34) 위의 글. pp. 34-36.



<그림 3> 데이터 유용성과 개인정보 보호수준의 타협

그런데, 현실적으로는 현재 위치한 상황이 가능한 최선의 데이터 유용성과 개인정보 보호 수준을 달성하고 있지 못한 경우인 경우가 많을 것이다. 이러한 경우에는 다른 가치를 낮추지 않으면서도 데이터 유용성이나 개인정보 보호 수준을 개선할 수 있는 여지가 발생하게 된다. 예컨대 아래 <그림 2>와 같이 현재 상황이 D1점에 위치해 있다고 가정해 보자. 이 경우에는 데이터 유용성을 그대로 유지하면서 개인정보 보호수준을 높일 가능성(D2점으로의 개선), 개인정보 보호수준을 그대로 유지하면서 데이터 유용성을 높일 가능성(D5점으로의 개선), 다른 최선의 타협을 이룰 가능성(D3점, D4점 기타 경계선 상에 위치하는 무수히 많은 타협점으로의 개선)이 있다. 이러한 변화는 다른 가치를 감소시키지 않고 특정 가치를 증가시키는 것이므로 파레토 개선이라 평가할 수 있다. 위험기반 접근법은 기술적 방법을 통해 데이터 유용성과 프라이버시 간의 최적의 지점을 찾는 최적화 과정을 수행하는 것을 목표로 한다.



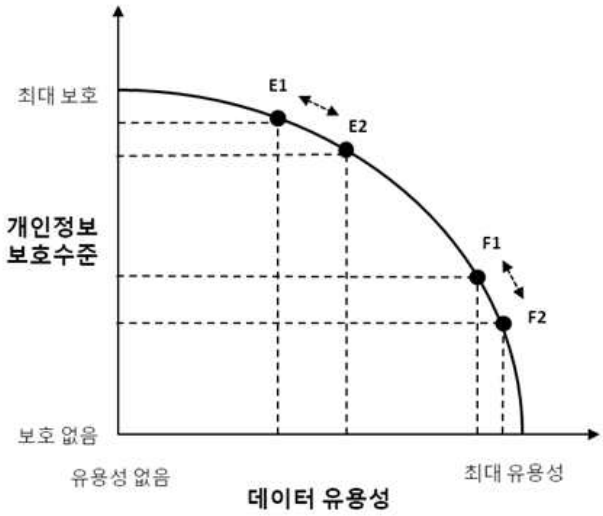
<그림 4> 파레토 개선이 가능한 상황

그렇다면, 경계선에 위치하는 여러 타협점 사이에는 어떠한 점을 선택해야 할 것인가? 이는 일응 규범적 판단의 문제에 해당한다. 데이터 유용성을 증가시키기 위해 개인정보 보호수준을 어느 정도 희생할 수 있는가 혹은 반대로 개인정보 보호 수준을 증가시키기 위해 데이터 유용성을 어느 정도 희생할 수 있는가에 대한 판단은 단순히 수치를 비교하는 것만으로는 답하기 어렵기 때문이다. 또한 이러한 판단은 트레이드-오프 곡선의 형태에 따라서도 달라진다. 곡선이 직선인지, 위로 볼록한 형태인지, 아래로 볼록한 형태인지에 따라 하나의 가치를 변경했을 때 다른 가치에 미치는 영향의 정도가 달라진다.

만약 (일반적으로 가정되는 바와 같이) 트레이드-오프 곡선이 <그림 1>과 같이 위로 볼록한 형태를 취하고 있다고 가정한다면, 위 곡선의 각 지점마다 트레이드오프가 달라진다.[35] 예컨대 아래 <그림 3>에서 E1 점

35) 실제로 트레이드-오프 곡선이 어떠한 형태를 취하고 있을 것인지는 실증적으로 검증할 필요가 있다. El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p.34는 트레이드-오프 곡선이 위로 볼록한 형태를 취한다고 전제한다.

과 E2 점을 비교하는 상황을 고려해 보자. E1점에서 E2점으로 이동하는 경로를 보면, 개인정보 보호 수준을 약간 낮추더라도 데이터 유용성이 상대적으로 크게 증가한다.[36] 따라서 E1점에서 E2점으로 이동은 규범적 승인을 얻기 비교적 용이할 것이다.



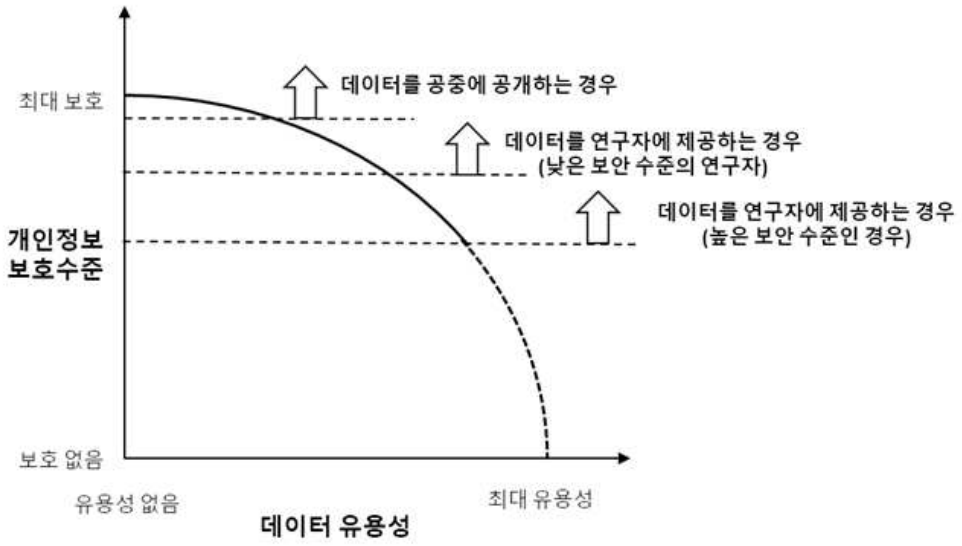
<그림 5> 가능한 타협점 간의 비교

<그림 3>에서 F1점과 F2점을 비교하는 상황은 이와 다르다. F1점에서 F2점으로 이동하는 경로를 보면, 개인정보 보호 수준은 상대적으로 크게 감소하는데 비해 그로 인해 증가하는 데이터 유용성은 크지 않다. 이는 위 그래프의 형태에서 이미 수확 체감(diminishing returns)의 법칙이 적용된다고 가정했기 때문이다. 즉, 이미 데이터 유용성이 상대적으로 큰 상황에서는 데이터 유용성을 추가로 증가시키기 위해서는 개인정보 보호 수준을 더 많이 희생해야 할 수 있다. 따라서 F1점에서 F2점으로의 이동에 대해서는 규범적 승인을 얻기가 상대적으로 어려울 것이다. 반대로 생각하면 F2점에서 F1점으로의 이동은 상대적으로 바람직하다. 이는 데이터 유용성을 크게 감소시키지 않으면서도 개인정보 보호 수준은 상대적으로 크게 증가시킬 수 있게 되기 때문이다.

36) 이러한 설명은 양자가 동일한 척도(scale)로 정규화(normalize)된 상황을 전제한다.

요컨대 위험기반 접근법의 둘째 전제는 데이터 유용성과 개인정보 보호 수준 간의 형량을 통해 '골디락스(Goldilocks)' 지점을 찾아낼 수 있다는 것이다.

위험기반 접근법의 셋째 전제는 매우 낮은 수준의 재식별 위험은 수용 가능하다는 것이다.[37] 여기서 수용 가능한 "매우 작은" 재식별 위험의 수준은 맥락(context)에 따라 달라진다. 가령 아래 그림과 같이 데이터를 웹사이트를 통해 공중에 공개하는 경우, 낮은 보안 수준의 연구자에게 제공하는 경우, 높은 보안 수준의 연구자에게 제공하는 경우는 비식별 조치의 수준이 달라진다. 하지만 어떠한 경우에도 재식별 위험을 0이 되도록 보장하는 것은 불가능하다. 따라서, 이러한 상황 하에서 수용 가능한 재식별 위험 수준의 한도(threshold)를 정의할 수 있는 반복 수행 가능한 절차 수립이 필요하게 된다.



<그림 6> 맥락에 따른 수용 가능한 재식별 위험 수준의 한도 차이

위험기반 접근법의 마지막 전제는 비식별화가 그저 기술적 조치에 그치는 것이 아니라 계약상 조치나 기타 관리적, 물리적 조치를 포함하는 것이라는 것이다.[38] 일반적으로 비식별 조치는 데이터의 총계처리, 삭제,

37) El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p.34

범주화, 마스킹 등의 기술적 조치만을 고려하는 것으로 이해되기 쉽다. 그러나 재식별화의 위험을 수용 가능한 한도 이내로 유지하기 위한 다양한 조치가 가능하다. 가령 연구자에게 데이터를 제공하는 경우 적정한 개인정보 보호 거버넌스 및 감독 제도를 마련할 것을 계약상 요구할 수도 있다. 기타 개인정보 보호를 위한 여러 관리상 조치가 가능할 것이다. 비식별화를 달성하기 위해서는 실무상 여러 조치들을 결합하여 활용할 수 있으며, 오로지 기술적 조치에만 의존하는 것이 아니다. 수용가능한 재식별 위험의 수준 한도는 이러한 기타 조치를 함께 고려하여 판단되어야 한다.

제 4 절 위험기반 비식별 조치의 절차

위험기반 접근법의 목표는 실무상 적용 가능한 체계적 방법론을 개발하는 것이므로, 이를 일정한 단계로 구분하여 정식화하는 것이 바람직하다. 이 때 얼마나 세밀하게 절차를 정의하느냐에 따라 여러 유형의 단계화가 가능할 것이다. 본 절에서는 El Emam & Arbuckle (2013)에서 정리한 비식별 조치 단계를 기준으로 설명한다.[39]

1. 직접식별자, 간접식별자의 선정

첫 단계는 비식별 조치의 대상이 되는 직접식별자(direct identifier)와 간접식별자(indirect identifier)를 선정하는 것이다.

우선, 직접식별자란 개인이나 그의 가족을 유일하게 식별할 수 있는 항목을 말한다. 대표적으로 주민등록번호 등 고유식별정보, 성명, 전화번호, 상세 주소, 이메일 주소, 계좌번호, 신용카드 번호 등이 있다, 국내 가이

38) 위의 글, p. 36.
39) 위의 글, pp. 36-43.

드라인에서는 직접식별자를 그저 '식별자'라고 칭하는 경우도 있고,[40] '식별정보' 또는 '개인식별정보'라 칭하는 경우도 있다.[41] 직접식별자에 해당하는 항목은 비식별 조치를 통해 삭제하거나, 원래 값으로 복원할 수 없는 형태로 변형한다.

다음으로, 간접식별자는 다른 식별자와 결합하여 개인을 식별할 수 있는 항목을 말한다. 국내 가이드라인에서는 속성자(attribute value)라 칭하는 경우,[42] '식별가능정보' 또는 '개인식별가능정보'라 칭하는 경우도 있다.[43] 간접식별자는 개인에게 고유하게 부여된 값이 아니라고 하더라도 다른 정보와 결합하여 개인을 식별할 수 있는 경우를 의미한다. 앞서 설명한 매사추세츠 주지사 사건을 예로 들면, 매사추세츠 주 GIC 의료정보에 포함된 우편번호, 생일 및 성별 정보는 간접식별자에 해당한다. 이러한 정보는 그 자체로서는 개인을 식별할 수 없다고 하더라도 매사추세츠 주지사가 식별될 수 있었기 때문이다.

간접식별자에 해당하기 위해서는 재현 가능성(replicable), 구별 가능성

---

40) 개인정보 비식별 조치 가이드라인(2016) 5면은 식별자를 "개인 또는 개인과 관련한 사물에 고유하게 부여된 값 또는 이름"이라 정하고 있다. 구체적으로, 고유식별정보(주민등록번호, 여권번호, 외국인등록번호, 운전면허번호), 성명(한자·영문 성명, 필명 등 포함), 상세 주소(구 단위 미만까지 포함된 주소), 날짜정보 : 생일(양/음력), 기념일(결혼, 돌, 환갑 등), 자격증 취득일 등, 전화번호(휴대전화번호, 집전화, 회사전화, 팩스번호), 의료기록번호, 건강보험번호, 복지 수급자 번호, 통장계좌번호, 신용카드번호, 각종 자격증 및 면허 번호, 자동차 번호, 각종 기기의 등록번호 & 일련번호, 사진(정지사진, 동영상, CCTV 영상 등), 신체 식별정보(지문, 음성, 홍채 등), 이메일 주소, IP 주소, Mac 주소, 홈페이지 URL 등, 식별코드(아이디, 사원번호, 고객번호 등), 기타 유일 식별번호 : 군번, 개인사업자의 사업자 등록번호 등을 예시하고 있다.

41) 가명정보 처리 가이드라인(2020) 17면은 "고유식별정보(여권번호, 외국인등록번호, 운전면허번호), 성명, 전화번호, 전자우편주소, 의료기록번호, 건강보험번호, 자동차 등록번호 등 외부 연계(식별)를 목적으로 생성된 정보 등"을 "식별정보"라고 표현한다.

42) 개인정보 비식별 조치 가이드라인(2016) 5면은 "'속성자'란 개인과 관련된 정보로서 다른 정보와 쉽게 결합하는 경우 특정 개인을 알아볼 수도 있는 정보"라 정하고 있다. '속성자'는 국어상 표현으로도 어색할 뿐만 아니라, 그 의미를 명확하게 전달하고 있지 못하므로 적절하지 못한 용어 선택이라 본다.

43) 가명정보 처리 가이드라인(2020) 17면은 "성별, 연령(나이), 국적, 혈액형, 신장, 몸무게, 직업, 위치정보 등 가명정보처리자의 입장에서 개인을 알아볼 수 있는 정보" 및 특이정보("국내 최고령, 최장신, 고액체납금액, 고액급여수급자 등 전체적인 패턴에서 벗어나 극단값이 발생 할 수 있는 정보 및 희귀 성씨, 희귀 혈액형, 희귀 눈동자 색깔, 희귀 병명, 희귀 직업 등 정보 자체로 특이한 값을 가지고 있는 정보")를 "식별가능정보"라고 부른다.

(distinguishable), 인지 가능성(knowable)이 인정되어야 한다. ① 재현 가능성은 해당 속성이 시간에 따라 잘 변화하지 않아야 한다는 뜻이다. 예컨대 혈당치는 측정 시점마다 달라지므로 간접식별자에 해당한다고 보기 어렵다. ② 구별 가능성은 데이터셋 내 기록마다 차이가 있어야 한다는 뜻이다. 유방암 환자 데이터베이스에서 유방암 진단 여부는 상호 구별 가능한 정보에 해당하지 않을 것이다. 다만 유방암 이외의 추가적인 진단 병명 항목이 존재한다면 이는 간접식별자에 해당할 수 있다. 마지막으로 ③ 인지 가능성은 비식별 조치가 취해진 정보를 재식별하고자 하는 공격자가 대상에 관해서 알 수 있는 경우에 해당해야 한다는 의미이다. 만약 합리적으로 판단했을 때 공격자가 알 수 없는 정보라면 비식별 조치를 통해 삭제될 필요가 없을 것이다. 하지만 공격자는 다양한 경로를 통해 정보주체에 관한 정보를 취득할 수 있다는 점을 유의해야 한다. 가령 가족이나 친분 관계를 통해 정보주체에 관한 배경 지식을 미리 알고 있을 수도 있고, 다른 공개된 정보를 통해 알 수도 있다.

위 세 가지 기준을 매사추세츠 주 GIC 의료정보의 사례에 적용해 보자. 위 사례에서 간접식별자에 해당하는 우편번호, 생일 및 성별 정보는 재현 가능성, 구별 가능성, 인지 가능성이 모두 인정된다. 생일이나 성별은 변경하기 매우 어렵고, 거주지 우편번호 역시 쉽사리 변하지 않으므로 재현 가능성이 있다. 위 정보를 이용하여 GIC 의료정보 내의 다른 기록와 주지사의 기록을 구별해 낼 수 있으므로 구별 가능성도 인정된다. 마지막으로 유권자 명부라는 다른 공개정보를 통해서 내에서는 공격자가 확인할 수 있으므로 인지 가능성이 인정된다. 따라서 위 항목들은 간접식별자에 해당한다.

간접식별자가 많아질수록 공격자가 개인을 식별해 낼 가능성이 커진다. 그에 적용되어야 할 비식별 조치의 수준 역시 그에 비례하여 엄격해지게 된다. 달리 표현하면 비식별 조치를 취한 결과 데이터에 더 많은 변형이 이루어지게 된다. 그래서 실무상으로는 비식별 조치를 취함에 있어 어느 항목까지 간접식별자로 볼 것인지가 논란이 된다. 이러한 판단이 어려운 이유는 데이터를 활용하는 맥락에 따라 판단 결과가 달라지기 때문이다.

가령 직업 정보는 사안에 따라 재현 가능성, 구별 가능성, 인지 가능성이 인정될 수도, 그렇지 않을 수도 있다. 단순히 공무원, 회사원, 자영업 등으로 구분된 경우라면 다른 정보와 결합하더라도 개인을 식별하기 쉽지 않을 수 있다. 하지만 '국회의원' 등 특수한 직업이 포함된 경우라면 간접식별자로 인정될 가능성이 높을 것이다.

2. 재식별 위험 수용 한도의 산정

위험기반 비식별 조치의 다음 단계는 수용 가능한 위험의 한도(threshold)를 정하는 것이다. 우선 위험 한도는 계량 가능해야 한다. 위험 한도를 구체적 수치로 계량화해야만 그에 비례해서 필요한 비식별 조치의 수준을 결정할 수 있기 때문이다. 다음으로 하고, 정당화 가능(defensible)해야 한다. 앞서 위험 기반 비식별 조치의 전제에서 설명한 바와 같이 비식별 조치는 재식별 위험을 0으로 만드는 작업이 아니다. 따라서 재식별이 이루어질 가능성을 배제할 수 없다. 이 경우 해당 비식별 조치가 합리적인 것이었다고 정당화할 수 있는 근거가 존재해야 한다.

수용 한도를 결정하는 중요한 요소는 크게 두 가지가 있다. 우선 ① 데이터가 제공되는 범위이다. 데이터가 일반 공중에 공개될 것인지, 제한된 연구자에게만 제공될 것인지에 따라 수용 한도가 달라진다. 기업이 자신의 데이터를 (가령, 인공지능 개발 등의 목적으로) 오픈 데이터로 공개하는 경우 또는 정부의 정보 공개 정책 등에 따라 데이터가 공중에 공개되는 경우 데이터 활용에 어떠한 제약도 없게 된다. 이에 비해 만약 수령인이 정해져 있는 경우라면 특정한 제약을 두거나 감독 절차를 수행할 수 있다. 만약 이 경우에는 이러한 제약 조건이나 감독 절차를 고려하여 상대적으로 낮은 수준의 비식별조치를 취하더라도 수용가능하게 될 가능성이 커진다.

② 다음으로 고려할 사항은 데이터가 제공됨으로써 발생할 수 있는 프라이버시 침해의 수준이다. 이는 데이터가 민감정보에 해당하는지에 따라

달라진다. 만약 의료정보라면 환자 정보가 원치 않게 공개될 경우 발생할 위험이 크다. 이는 프라이버시 보호 행동(privacy protective behavior)을 불러 일으킬 수 있다는 점에서 사회적으로 중대한 부작용을 발생시킬 수 있다. 만약 언론이 공개된 의료정보를 통해 어떤 유명인이 성행위로 전파되는 질병에 걸렸다는 사실을 재식별화하여 공개하는 경우를 상정해 보자. 사람들은 프라이버시 침해를 우려하여 더 이상 해당 질병 검사를 받지 않으려고 하게 될 수 있고, 그 결과 공중보건의 측면에서 심각한 문제 발생할 수 있다. 특히 의료 분야에 있어 프라이버시 보호 행동은 다양한 맥락에서 관찰된다. 환자는 개인 병력, 가족 병력, 자가 복용 약물 등에 대해 의사에게 거짓으로 고지하거나, 아예 따로 돈을 주고 의료기록을 생성하지 않기도 한다. 하나의 질병에 대해 여러 병원을 방문해서 의료기록이 부분적으로만 존재하도록 만들기도 한다. 특히 청소년들은 부모에게 의료 기록을 보여주지 않고 싶거나, 장래 취업에의 영향을 우려한다. 비식별 조치의 수용 한도를 결정함에 있어서는 이러한 가능성이 고려되어야 하는 것이다.

아래 <그림 5>는 수용 가능한 위험 수준에 관해 일반적으로 활용되어 온 기준을 소개한 것이다.[44] 여기서는 데이터가 공개될 경우에는 5~9% 사이의 위험 한도 범위를 설정한다. 5~9% 사이의 구체적인 값은 프라이버시 침해 정도를 고려하여 결정된다. 이는 공격자가 정보주체에 관하여 모든 배경 지식을 보유하고 있다고 가정하더라도 공개된 데이터셋으로부터 해당 정보주체를 식별해 낼 가능성이 10%도 되지 않는다는 의미이다. 데이터를 인터넷에 공개하는 등과 같은 상황에서 5% 수준은 매우 안전한 수준이 된다. 앞서 본 매사추세츠 주지사 사건을 생각해 보자. 공격자가 주지사의 우편번호, 생일 및 성별 정보를 모두 알고 있다고 가정한 상태에서 GIC 의료정보로부터 주지사를 식별해 낼 가능성이 5%에 해당하려면, GIC 의료정보에는 동일한 기록을 가진 사람이 20명이 있어야 한다(20명 중에 1명일 경우, 1/20 = 5%가 된다).[45] 공격자는 20명 중

44) El Emam, Khaled and Luk Arbuckle (2013) p. 57.

45) 엄밀히 말하자면 이는 k-익명성을 기준으로 재식별 위험을 산정한 것이다. 개인정보 비식별 조치 가이드라인(2016) 38면에서 설명하고 있는 바와 같이 이러한 기준에 의

누가 주지사에 해당하는지 알기 어렵다. 하지만 이 정도 수준으로 데이터를 변형할 경우 데이터의 유용성은 낮아지게 된다는 점에 유의할 필요가 있다.

만약 데이터가 특정인에게만 제공되는 경우라면 수용 가능한 한도를 더 높게 설정할 수 있다. 데이터 수령자가 강력한 보안 수준과 프라이버시 보호 정책을 가진 경우, 민감정보가 포함되어 있지 않은 경우, 정보주체가 해당 정보를 비식별 조치를 취하여 공개하는 것에 동의한 경우, 공격자가 해당 정보에 대한 재식별을 시도할 동기가 크지 않은 경우라면 수용 가능한 위험 한도가 10~33% 수준으로 더 높아질 수 있을 것이다. 여기서 재식별 가능성이 33%라고 하면, 공격자가 정보주체에 관하여 모든 배경 지식을 보유하고 가정할 경우 해당 데이터셋에서 해당 정보주체를 식별해 낼 가능성이 33%이 된다는 뜻이다.

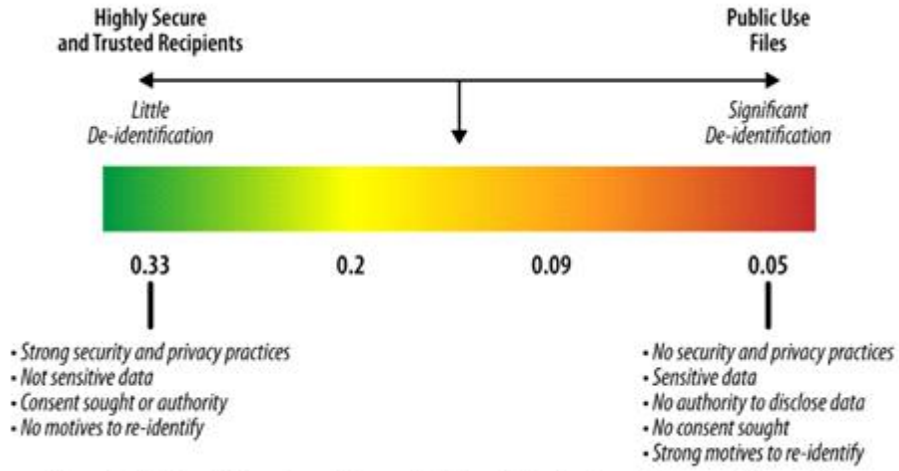


Figure 2-3. The different maximum risk thresholds that have been used in practice

<그림 7> 수용 가능한 위험 한도의 결정

하더라도 개인에 관한 유의미한 정보가 추론될 위험은 상존한다. 예컨대 주지사와 동일한 우편번호, 생일 및 성별을 가진 사람이 20명이 있더라도, 만약 그 20명 중 15명이 '전립선염'이라고 하자. 그러면 주지사가 전립선염을 가질 가능성은 75%나 된다. 만약 주지사의 평소 행동을 관찰한 공격자라면 그가 전립선염이 있다고 확신할 수도 있다. 이러한 공격의 위험을 막기 위해서는 단지 간접식별자가 동일한 기록이 일정 수 이상일 것뿐만 아니라, 그 속성이 다양하게 분포되어 있을 것을 요구할 필요도 있다(l-다양성의 고려).

수용 가능한 위험 한도를 결정하기 위해서는 비식별 조치가 수행되는 맥락을 종합적으로 고려해야 한다. 이를 체계적으로 판단하기 위해 체크리스트를 마련하여 활용하게 된다. 예컨대 국내 개인정보 비식별 조치 가이드라인(2016) 42-49면은 비식별 조치 적정성 평가단의 세부 평가 수행에 있어 평가를 위한 세부 지표를 개발하여 제공하고 있다.

3. 가능한 재식별 공격 방법의 고려

가. 가능한 재식별 공격의 유형

비식별 조치의 다음 단계는 어떠한 형태의 재식별 공격이 가능한 것인지 고려하는 것이다. 가능한 공격의 유형에 따라서 구체적으로 이루어져야 할 비식별 조치의 수준이 달라지기 때문이다. 일반적으로 다음 네 가지 유형의 공격이 가능하다.

① 데이터 수령자가 고의적으로 재식별을 시도하는 경우
② 데이터 수령자가 의도치 않게(또는 즉흥적으로) 정보주체를 식별하게 되는 경우. 예컨대 데이터셋에 지인(智人)이 포함된 경우 등
③ 데이터 유출 사고가 발생하는 경우
④ 데이터에 대해 공격이 가능하다는 점을 과시하기 위한 공격을 하는 경우(과시 공격, demonstration attack)

만약 특정한 수령자에게만 제공되는 경우에는 위 ①, ②, ③번 공격이 현실적으로 가능하다고 생각할 수 있다. 이러한 공격 발생의 가능성에 영향을 미치는 요소로는 재식별 의도 및 역량, 수령자의 개인정보 보호 수준을 들 수 있다. 우선 재식별 의도 및 역량을 판단함에 있어서는 데이터 수령자가 재식별할 수 있는 동기가 있는지, 재식별 시도를 위한 자원 및 기술적 역량이 있는지를 판단하게 된다. 재식별화를 금지하는 계약 등을 체결하고, 주기적인 감독 조치를 실시함으로써 이러한 동기를 줄일 수 있다. 또한 데이터 수령자가 우수한 개인정보 보안 및 프라이버시 관

행을 보유하고 있는지에 따라 공격 발생의 가능성이 낮아지게 된다. 특히 데이터 수령자 내부의 악의적인 직원이 존재할 가능성을 막는 것이 중요하다. 만약 조직 내 위험 감소를 위한 통제가 잘 이루어진다면 데이터에 대한 접근 가능한 직원과 별도 비밀유지계약을 체결할 필요까지는 없을 것이나, 만약 그렇지 않다면 해당 직원들과 별도로 비밀유지계약을 체결하는 것이 바람직하다.

다른 한편, 과시 공격은 공격자가 특정인을 대상으로 재식별을 시도하는 것이 아니라 가장 쉬운 대상(low hanging fruit)을 골라 재식별을 시도하는 것이다. 단일 기록만이라도 재식별화가 이루어질 경우 공격자가 이 사실을 과시할 수 있게 된다. 이러한 점에서 이를 언론 공격(journalist attack)이라고도 부른다. 이 경우, 공격자는 공격 가능성을 보여주기 위해 불법 데이터를 활용하지는 않는 경향이 있다. 데이터셋이 공중에 공개되는 경우 과시 공격도 염두에 두어야 하며, 매우 높은 수준의 비식별 조치가 이루어질 필요가 있을 것이다. 하지만 특정 제3자와 사이에 비밀유지계약을 체결하고 데이터를 제공하는 경우에는 과시 공격까지 고려할 필요는 없을 수 있다.

나. 위험의 유형: 최대 위험, 평균 위험, 엄격한 평균 위험

가능한 공격의 유형에 따라 위험기반 비식별 조치의 판단 기준이 되는 위험의 산정 기준이 달라진다. 가장 보수적인 산출 방법은 공격의 최대 위험(maximum risk)을 기준으로 하는 것이다. 이는 모든 개별 기록의 재식별 확률을 산정하여 가장 높은 재식별 확률을 가진 기록의 위험을 기준으로 전체 데이터셋의 재식별 위험으로 보는 것이다. 특히 앞서 본 과시 공격을 고려해야 할 경우 이를 기준으로 하는 것이 타당하다.

그보다 완화된 위험 산정 방법은 평균 위험(average risk)을 기준으로 하는 것이다. 이는 데이터셋 전체 기록의 재식별화 확률의 평균을 구하는 것이다. 제3자에게 제공되는 데이터셋의 양이 적을수록 평균 위험은

낮아지게 된다. 가령 개인정보처리자가 보유한 전체 데이터 중 일부(가령 30%)만을 임의로 추출하여 제공하는 경우를 생각해 보자(이를 '표본 추출(sub-sampling)'이라 한다). 표본 추출을 할 경우 데이터의 간접식별자에 대해 어떠한 변형을 가하지 않더라도 연구자가 지인을 식별할 위험을 낮출 수 있게 된다. 전체 데이터셋에 재식별 공격 대상이 포함되어 있더라도 그 대상을 식별할 가능성은 서브 샘플링 비율만큼 낮아지게 될 것이기 때문이다. 하지만 표본 추출을 하더라도 최대 위험에는 차이가 없는 경우가 많을 것이다. 이처럼 표본 추출은 평균 위험 수준을 고려하여 데이터 유용성을 높이는 기법에 해당한다.

하지만 평균위험을 기준으로 할 경우 데이터셋에서 특정인을 곧바로 식별해 낼 수 있는 경우가 충분한 조치가 취해지지 못하는 문제가 발생할 수 있다. 예를 들어 특정 우편번호에 거주하는 나이가 99세인 사람이 데이터셋에 포함되어 있다고 생각하자. 아마도 해당 지역 거주자라면 이 사람이 누구인지 곧바로 알 수 있을 가능성이 매우 크다. 만약 비식별 조치를 통해 "90세 이상"이라고 변형하더라도 하더라도 여전히 특정 가능한 경우가 많을 것이다. 이처럼 데이터셋의 전체 평균적으로는 재식별 위험이 낮더라도 특정인에 대해서는 위험이 높을 수 있다.

이러한 상황을 고려하여 엄격한 평균 위험(strict average risk)을 요구할 수도 있다. 엄격한 평균 위험은 최대 위험과 평균 위험을 결합한 것이다. 즉, 1단계로는 최대 위험이 특정 한도 이하일 것을 요구한다. 예를 들어 최대 위험의 한도를 50%로 정하면 2개 이상의 기록이 동일하게 된다. 혹은 보다 안전하게는 33%를 한도로 하여 3개 이상의 기록이 동일할 것을 요구할 수도 있다. 이러한 조치에 따라서 앞서 본 99세 노인의 경우 재식별화 위험을 낮춘다. 2단계로 평균 위험을 측정하고, 이를 데이터셋의 위험 수준으로 파악한다.

이처럼 위험 평가 기준은 다양할 수 있으나, 국내의 개인정보 비식별 조치 가이드라인(2016)이나 가명정보 처리 가이드라인(2020)은 이러한 위험 산정 기준을 명확하게 제시하고 있지 못하고 있다.[46]

46) 비식별 조치 가이드라인(2016)은 k-익명성을 기준으로 하므로 일응 최대 위험을 위험 판단의 기준으로 삼고 있다고 평가할 수 있다.

다. 재식별 공격자의 배경 지식에 관한 가정

다른 한편으로, 가능한 공격의 유형에 따라 위험기반 비식별 조치의 판단 기준이 되는 재식별 공격자의 배경 지식에 관한 가정의 수준도 달라진다.

가장 엄격하고 보수적인 기준은 정확한 완전 지식(exact complete knowledge)의 가정이다. 이러한 가정에 따르면 재식별 공격자는 정보주체에 대한 모든 간접식별자 정보를 보유하고 있는 경우를 산정한다. 가령 공격자가 정보주체를 계속 스토킹해 와서 생일, 성별, 방문일 및 거주지를 모두 알고 있다고 가정하는 것이다. 특히 데이터셋에 유명인이 포함된 경우라면 이러한 위험은 현실적일 수 있다. 이 경우 데이터셋에서 해당 정보주체를 식별해 낼 수 있는지를 기준으로 위험을 평가한다. 데이터를 공중에 공개하는 경우라면 공격자가 정확한 완전 지식을 가정하고 비식별화 처리를 할 필요가 있다.

정확한 완전 지식의 가정은 지나치게 엄격하다는 문제가 있으므로, 비식별 조치의 맥락에 따라서는 보다 완화된 가정을 도입할 수 있다. 정확한 완전 지식을 일부 약화시킨 방법으로는 근사적 완전 지식(approximate complete knowledge)의 가정이 있다. 이는 간접식별자 변수들이 상호 독립이라고 가정하는 것이다. 하지만 근사적 완전 지식 가정도 여전히 상당히 엄격한 수준이다.

보다 현실적인 가정은 공격자가 정보주체에 관하여 일부 정보만을 보유하고 있다고 가정하는 것이다. 가령 의료정보의 비식별화에 있어서 재식별 공격자가 환자의 모든 진단 과정이나 결과를 알고 있기는 가정하기 어렵다(심지어 환자의 가족이라고 하더라도 이를 모두 알고 있지 못할 수 있다). 환자의 지인 중에서는 그가 특정일에 병원에 입원했다는 정보만을 알고 있는 경우도 있을 것이다. 이처럼 공격자가 보유하고 있는 정보의 양을 “power of the adversary”라 한다. 이러한 가정에 따르면, 공격자가 보유하고 있는 정보량의 합리적 추정치(power)를 산정한 다음,

시뮬레이션을 통해 해당 정보를 보유한 공격자가 개인을 식별해 낼 확률을 계산하는 방식으로 재식별 위험을 산출한다.

4. 데이터 비식별화 기법의 적용

이상과 같이 비식별 조치의 위험 수용 한도 및 가능한 공격 기법을 고려한 위험 산출 기준이 정해진 다음에는, 구체적으로 비식별 기법을 적용하여 데이터로부터 개인이 식별될 위험을 낮추는 작업이 진행된다. 데이터를 비식별화하기 위해 적용할 수 있는 기법은 매우 다양하다.[47] 이러한 기법들은 공통적으로 데이터를 삭제, 변형하거나 일반화하는 것이다. 이하에서는 대표적인 기법을 간략히 소개한다.

가. 비식별화의 기술적 기법

우선 직접식별자에 대해서는 이를 삭제하거나 원래값으로 복원할 수 없는 값으로 변형하는 방법이 사용된다. 한편, 간접식별자에 대해서는 이를 일반화하거나 제거하는 방법이 주로 활용된다. 데이터의 일반화(generalization)는 해당 데이터의 정밀도를 낮추는 작업이다. 예컨대 생년월일 정보는 연도와 월 정보까지만 남기고 일자를 삭제하거나, 연도나 남기거나, 더 일반화하여 5년 구간으로 구분할 수 있다. 이러한 변형을 가하면 공격자가 특정 정보주체에 관해 배경지식을 갖고 있더라도 그와 동일한 정보가 포함된 다른 개인이 많아지므로 개인을 식별하기 어렵게 된다. 다만 일반화 기법에 따르면 데이터의 정확도 자체는 유지된다.
만약 일반화를 적용하더라도 특정 가능한 개인이 남게 되는 경우에는 해

47) 개인정보 비식별 조치 가이드라인(2016) 31-35면은 일반적 기법으로 ① 가명처리(휴리스틱 가명화, 암호화, 교환 방법), ② 총계처리(총계처리, 부분총계, 라운딩, 재배열), ③ 데이터 삭제(식별자 삭제, 식별자 부분삭제, 레코드 삭제, 식별요소 전부삭제), ③ 데이터범주화(감추기, 랜덤 라운딩, 범위 방법, 제어 라운딩), ④ 데이터 마스킹(임의잡음 추가, 공백과 대체)를 들고 있다. 한편 가명정보 처리 가이드라인(2020) 42-44면은 가명처리의 기법을 ① 개인정보의 삭제, ② 개인정보 일부 또는 전부 대체, ③ 기타 기술로 구분하고 다양한 기법을 소개하고 있다.

당 개인에 관한 정보는 삭제한다. 앞서 예로 든 바와 같이 특정 우편번호에 거주하는 나이가 99세인 사람에 대해서는 “90세 이상”이라고 일반화하더라도 여전히 특정 가능할 수 있다. 이 경우에는 해당 나이 정보를 삭제하거나, 개인에 관한 정보 전체를 삭제한다. 혹은 의료 데이터에 55세의 나이에 출산한 산모가 있다고 하자. 이는 유일한 정보에 해당할 가능성이 매우 높으므로 나이 정보를 제거한다.

그 이외에도 데이터를 정확하지 않은 다른 값으로 변형할 수도 있다. 예컨대 노이즈 값을 추가하는 것이 대표적이다. 특히 날짜 정보(의료 정보의 경우 입원일자, 퇴원일자, 시술일자 등)의 경우에는 임의의 값을 더하거나 빼는 방식으로 이를 변형하는 것을 고려할 수 있다. 이렇게 하면 공격자가 특정 정보주체에 관한 정확하고 구체적인 배경 정보를 가지고 있더라도 그를 특정하기 어렵게 만들 수 있다. 다만 이에 대해서는 여러 필터링 기법을 이용하여 노이즈를 제거할 수 있을 가능성을 고려해야 한다.

나. 비식별화 기법의 적용 수준 - k-익명성의 적용 사례

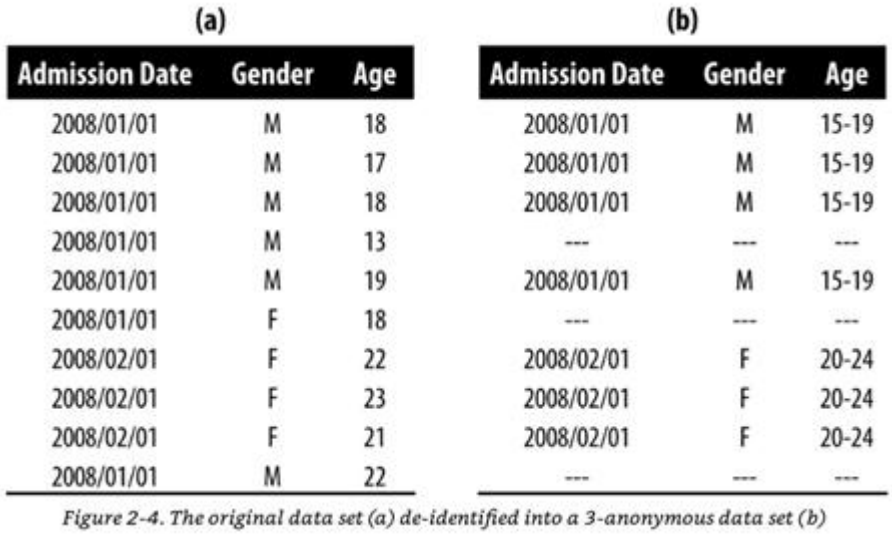
그렇다면 데이터의 비식별 조치의 일반화나 삭제는 얼마나 이루어져야 하는가? 위험기반 비식별 조치 방법론은 앞서 정한 위험 한도를 만족시킬 수준까지 일반화나 삭제를 반복적으로 적용하는 방식으로 진행된다. 이를 평가하는 간명하고도 일반적인 기준은 k-익명성(k-anonymity)이다. 이 기준은 동일한 간접식별자 값을 갖는 기록의 수가 k개 이상이 되도록 하는 것이다. 개인정보 비식별 조치 가이드라인(2016)은 k-익명성 모델을 기본적으로 적용하도록 정하고 있다.[48]

구체적 사례를 들어 살펴보자.[49] 병원 입원 환자 기록을 재식별하는 상

48) k-익명성은 여러 취약점이 있는 것으로 알려져 있으므로 이를 보완하기 위해 l-다양성과 t-근접성 등의 다른 방식을 추가적으로 적용한다. 개인정보 비식별 조치 가이드라인(2016) 28면.

49) 이는 El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p. 61의 사례를 인용한 것이다.

황을 고려한다. 우선 데이터셋에서 환자의 성명이나 연락처 등의 직접식별자 값을 삭제한다. 삭제되지 않고 남은 간접식별자는 아래 <그림 6>의 (a)와 같다고 하자. 여기서는 입원일자와 성별, 나이가 간접식별자로 남아 있다고 한다. 2008. 1. 1. 입원한 환자는 6명이 있다. 그 다음으로 간접식별자 값이 동일한 기록의 수를 계산한다. 아래 그림(a)를 보면 2008/01/01 입원한 18세 남성에 해당하는 기록은 총 2건이 있다. 이에 비해 2018/01/01 입원한 17세 남성의 기록은 오직 1건이 있을 뿐이다. 만약 공격자가 입원일자와 나이, 성별 정보를 알고 있다면 그의 입원 기록을 곧바로 식별해 낼 수 있다.



(a)

Admission Date	Gender	Age
2008/01/01	M	18
2008/01/01	M	17
2008/01/01	M	18
2008/01/01	M	13
2008/01/01	M	19
2008/01/01	F	18
2008/02/01	F	22
2008/02/01	F	23
2008/02/01	F	21
2008/01/01	M	22

(b)

Admission Date	Gender	Age
2008/01/01	M	15-19
2008/01/01	M	15-19
2008/01/01	M	15-19
---	---	---
2008/01/01	M	15-19
---	---	---
2008/02/01	F	20-24
2008/02/01	F	20-24
2008/02/01	F	20-24
---	---	---

Figure 2-4. The original data set (a) de-identified into a 3-anonymous data set (b)

<그림 8> k-익명성 설명을 위한 예제 병원 입원 환자 기록

공격자가 이러한 배경 지식을 보유하고 있다고 하더라도 개인을 식별할 확률이 33% 이하가 되도록 하는 것이 비식별 조치의 목적이라고 해 보자. 그렇게 하려면 동일한 기록이 3건 이상이 되도록 데이터를 변형해야 한다. 데이터 제공자와 수령자가 협의하여 위 데이터에서 나이를 5년 구간으로 일반화하기로 정했다고 하자. 위 그림의 (b)에서 볼 수 있는 바와 같이 2008/01/01 입원한 15-19세 남성 레코드는 총 4건이 된다. 따라서 공격자가 개인을 식별할 확률은 25%(=1/4)로 낮아졌다.

이와 달리 위 그림의 (a)의 4번째 기록인 13세 남성을 보자. 2008/01/01 입원하였으나 10-14세 사이인 다른 환자가 없다. 따라서 그의 기록은 (b)에서는 삭제되어야 한다. 만약 나이를 5세 구간이 아니라 10세 구간으로 일반화하였다면 어떠한가? 2008/01/01에 입원한 10-19세 사이의 다른 남성이 4명이 있으므로, 그의 기록은 삭제되지 않고 포함되었을 것이다. 하지만 그렇게 하면 나이 데이터의 정밀도가 낮아지게 된다는 문제가 있다.

마찬가지로 2008/01/01에 입원한 18세의 여성 환자 기록을 보자. 이 기록과 동일한 간접식별자를 갖는 다른 기록을 찾을 수 없으므로 이 기록은 삭제된다. 이 경우에는 나이를 10-19세 사이로 일반화하더라도 마찬가지이다. 동일한 일자에 입원한 다른 여성 환자가 없기 때문이다. 이 경우 전체 데이터에서 성별 정보를 모두 삭제하는 것도 고려할 수 있다. 하지만 위 데이터에서 성별 정보가 중요하다면 그저 해당 여성 환자 기록을 삭제하는 것으로 만족할 것이다.

이처럼 데이터의 정밀도를 유지하는 것과 더 많은 기록을 유지하는 것은 상충 관계에 있게 된다. 여러 가능한 기술적 조치 중 어떠한 조치를 적용할 것인지는 데이터의 활용 목적에 따라서 선택된다.

제 5 절 가명처리

비식별조치의 또다른 방법은 가명처리이다. 가명처리의 법적 개념에 관하여서는 제3장과 제4장에서 상술하므로, 본 절에서는 가명처리의 기술적 사항에 관해 살핀다. 그 법적 정의와 무관하게 일반적으로 가명처리는 하나 또는 그 이상의 직접식별자를 다른 유일한 값으로 대체하는 것으로 이해된다.[50] 이러한 이해는 개인정보 보호법상 가명처리의 개념이 도입되기 이전에 제정된 개인정보 비식별 조치 가이드라인(2016)의 설명

50) El Emam, Khaled and Luk Arbuckle (2013) (각주 15) p. 218.

에서도 확인된다. 위 가이드라인은 가명처리를 “개인 식별이 가능한 데이터를 직접적으로 식별할 수 없는 다른 값으로 대체하는 기법”이라고 정하고 있다.[51]

1. 가명을 활용할 필요성

앞서 비식별조치에 있어 직접식별자는 삭제하는 것이 일반적이라고 하였다. 하지만 실무상으로 직접식별자를 삭제하는 것이 어려운 경우가 발생한다. 어떤 쇼핑몰 업체가 이용자에 관한 기본정보와 구매이력을 다음 표와 같이 저장하고 있다고 생각해 보자. 기본 정보 테이블과 구매이력 테이블은 모두 고객ID를 포함하고 있고, 이를 통해 양자가 연결된다. 만약 김갑동의 구매이력을 확인하고 싶다면, 이용자 프로파일 테이블에서 김갑동의 고객 아이디 “00000001”를 찾는다. 이로부터 구매이력 테이블에서 그가 2019. 12. 30. 오후 1시 35분에 150만원의 LG TV를 신용카드로 구매하였다는 정보를 확인할 수 있게 된다.

고객ID	성명	생년월일	이메일	우편번호	상세주소	전화번호	최근 접속	...
00000001	김갑동	660102	kkd@hanmail.net	01234	우리APT A동 101호	01012345678	191230 07:01	...
00000002	이을서	770304	les@gmail.com	02345	나라APT B동 202호	01012341234	200102 08:12	...
00000003	박병남	880506	pbn@gmail.com	03456	세종로 1	01045674567	200103 19:32	...
00000004	최정북	990708	cjb@aol.com	04567	대왕로 2	01078907890	200102 20:04	...
...	...		...	...	...	...	..	...

<표 1> 이용자 기본정보 테이블

51) 개인정보 비식별 조치 가이드라인(2016), 31면.

고객ID	구매일시	카테고리	물품명	물품가격	결제방법
00000001	191230 13:35	TV	LG 60AB1234	1,500,000	신용카드
00000003	191003 20:01	휴대폰	K20Plus	800,000	체크카드
00000002	202002 09:14	냉장고	R230L	1,000,000	현금
00000004	200102 21:05	모바일	아이패드 7세대	2,000,000	체크카드
00000001	200101 07:30	휴대폰	K20Ultra	400,000	신용카드
...	...	...			...

<표 2> 구매이력 테이블

이러한 형태로 저장된 데이터를 가명처리한다고 생각해 보자. 쇼핑몰이 보유한 고객ID는 이용자를 식별하기 위해 고유하게 부여한 값이므로 직접식별자에 해당한다. 하지만 고객ID를 삭제하면 위 두 테이블의 연결이 끊어지게 된다. 따라서 테이블간 연결성을 유지하기 위해서는 이 값을 삭제할 수는 없고, 다른 값으로 변경해야 한다. 이처럼 고객ID라는 식별정보를 다른 값으로 변경하는 것이 가명처리이다.

가명처리를 위해 활용하는 데이터 변형 기법에 있어 중요한 특징은 같은 값을 가명처리하면 동일한 값이 나오도록 해야 한다는 것이다. 예컨대 가명처리가 이루어진 후 6개월 후에 해당 정보를 업데이트하려면 어떻게 해야 할까? 이용자가 새로운 물건을 구매한 내역이 가명처리된 데이터베이스 반영되어야 할 것이다. 그러려면 기존 DB에 포함된 이용자와 업데이트된 DB에 포함된 이용자를 서로 매칭시킬 수 있어야 한다. 만약 고객ID를 가명처리할 때마다 서로 다른 값이 나온다면 동일한 이용자 기록을 매칭시킬 수 없게 되는 문제가 발생할 수 있다.

이처럼 같은 값을 입력하면 같은 값이 출력되지만, 역으로 원래 값이 무엇인지 복원할 수 없도록 하기 위한 방법으로는 ① 일방향 암호화와 ② 쌍방향 암호화의 2가지 방법이 일반적으로 활용된다. 일방향 암호화는 수학적 기법을 활용하여 원본 값으로 복원될 수 없는 다른 일련의 숫자 값으로 변형하는 것이다. 일방향 암호화에 있어 널리 활용되는

SHA-256 해시함수를 이용하면 “0000001”이라는 고객 ID는 다음과 같은 16진수 숫자값으로 변형된다.

“0000001” → “E27ADA58F8898BDE36C4BD6BBC544263D7522718603BB9F42917F49932D97C79”

이와 달리 쌍방향 암호화는 암호 키(key)를 이용하여 암호화하는 방법이다. 즉, 암호 키를 이용하면 원본 값을 다시 복호화할 수 있다. 일방향 암호화를 이용할 경우에는 원본 값을 복원할 경우를 대비해서, 원본 값과 암호화된 값을 대응시키는 표(mapping table)을 유지해야 한다는 어려움이 있다. 만약 가명처리의 대상이 되는 데이터가 방대하다면 대응표를 유지하는 것도 상당한 부담이 될 수 있다. 이에 비해 쌍방향 암호화를 이용하면 대응 표를 유지할 필요없이 원본 값을 복원해 낼 수 있다. 그 대신 암호 키를 안전하게 보관하여야 한다는 부담이 발생하게 된다.

한편, 평소에는 가명처리된 정보를 활용하더라도 예외적으로 재식별화가 필요한 경우가 있을 수 있다. 예컨대 공공보건 맥락에서 특정 환자가 유해물질에 유출되었다면 환자를 접촉하여 추적할 필요가 있다. 이러한 경우 원본 값을 복원할 필요가 있을 것이다. 그러나 이러한 가명정보의 재식별은 안전한 환경 하에서, 적절한 통제와 감독이 이루어지는 경우에 이루어져야 한다. 만약 단방향 암호화를 사용한 경우에는 원본 값과 가명 값 사이의 대응표를 유지하여야 하고, 쌍방향 암호화를 활용할 경우에는 암호화 키를 보관하여야 할 것이다. 어느 경우이는 원래 값으로 복원할 수 있는 정보(이를 ‘가명처리 비밀’이라 한다)는 안전하게 유지할 필요가 있다.

2. 가명정보에 대한 공격

가명처리에서는 단방향 암호화를 주로 활용하므로, 암호화된 결과값을

이용하여 원본값을 알아내는 것이 이론적으로는 계산상 불가능(computationally infeasible)하다. 하지만 원본 데이터의 특징을 알고 있다면 가능한 모든 암호값을 미리 계산해 놓는 방식으로 원래 값을 복원할 가능성이 발생한다. 예컨대 “010”으로 시작하는 휴대전화번호를 단방향 암호화하였고, 적용된 암호화 기법이 무엇인지 알고 있다면, 가능한 모든 휴대전화번호 경우의 수에 대해 암호값을 계산해 둘 수 있다. 국내에서는 가능한 휴대전화번호의 총 1억개(010-0000-0000~010-9999-9999)에 불과하다. 이처럼 1억개 번호에 대한 암호값은 손쉽게 계산해 낼 수 있다. 이를 막기 위해서는 단방향 암호화를 하기에 앞서 데이터에 임의의 문자열을 부가하는 조치(이른바 ‘salting’)를 취해야 한다.

또한 가명처리는 동일한 원본 값에 대해 같은 가명값을 생성하므로, 빈도(frequency)가 유지된다는 문제가 있다. 특히 기관의 명칭 등을 가명처리하는 경우는 빈도를 이용한 재식별의 문제가 발생한다. 예컨대 환자가 치료받은 병원의 명칭에 대해 가명처리를 한다고 생각해 보자. 하지만 해당 지역에 매우 큰 병원이 하나 있고, 여러 작은 병원이 있다고 하면, 큰 병원에서 진료를 받은 환자의 빈도가 높을 것이다. 이러한 경우에는 가명처리가 실패하게 된다. 비슷한 사례로 그 지역에 아동 병원이 유일한 경우를 들 수 있다. 아마도 아동이 가장 높은 빈도로 진료받는 곳은 해당 아동 병원일 것이다. 이러한 문제를 해결하기 위해서는 가명처리시마다 확률적으로 서로 다른 값이 나오도록 하는 함수를 활용할 필요가 있게 된다.

3. 가명처리 개념의 확장

앞서 설명한 바와 같이 종래에는 가명처리가 직접식별자를 교체·제거하는 것으로 이해되었으나, 최근 가명처리의 개념이 확장되고 있다. 확장된 의미의 가명처리는 직접식별자를 제거·교체하는 것 뿐만 아니라 정보주체가 공중에 알려진 정보를 이용하여 간접식별자에 의해서 식별될 가능성을 막기 위하여 데이터에 대한 변형을 가하는 것이다. 본 연구에서는

전통적인 가명처리(직접식별자의 교체·제거)를 **“약한 가명처리(basic pseudonymization)”**라 하고, 보다 강화된 방식의 가명처리(간접식별자에 대한 변형을 부가적으로 수행)를 “강한 **가명처리(strong pseudonymization)**”라 한다. 양자의 구분을 위해서는 가명처리의 법적 개념에 관한 고찰이 필요하므로, 이하 제3장 제3절에서 보다 상세히 다룬다.

제 6 절 소결

본 장에서는 비식별 조치의 목적, 비식별화된 정보의 재식별 위험, 위험기반 비식별 조치의 개념과 그 절차, 가명처리의 기술적 사항에 관해 검토하였다. 개인정보처리자는 주로 과학적 연구 활동을 위해서 또는 개인정보를 통계분석하여 그 결과를 통해 의사결정을 보조하기 위한 목적으로 개인정보를 비식별화하여 활용할 수 있다. 나아가 정보주체에 대한 보호를 더욱 강화하기 위해서도 활용될 수 있다. 그러나 적절한 비식별 조치가 이루어지지 못한 경우 개인이 재식별될 수 있다는 점이 여러 사례를 통해 드러났다.

그 결과 재식별 위험성을 평가하고, 위험성을 고려한 비식별 조치를 수행하는 ‘위험기반 비식별 조치’ 방법론이 발전하였다. 최근의 위험기반 비식별 조치 방법론의 요체는 위험성 평가 과정을 통해 데이터 유용성과 개인정보 보호 수준 간의 형량을 하고, 이를 통해 최적의 지점, 즉 ‘골디락스(Goldilocks)’ 지점을 찾아낼 수 있다는 것이다. 이를 위해서는 체계화되고 정식화된 절차를 통해 재식별 위험을 고려하고, 그에 상응하는 기술적·관리적 조치를 취하여야 한다. 그러나, 이러한 위험기반 비식별 조치 방법론이 현행 법제도 하에서 어떻게 수용될 수 있는지는 별도로 검토해야 할 쟁점인바, 이하에서는 GDPR과 국내 개인정보 보호법 상의 비식별 조치에 관해 살핀다.

제 3 장 GDPR과 개인정보 비식별 조치

제 1 절 개인정보 판단의 역동성

비식별 조치는 개인정보를 비(非)개인정보로 변화시키는 작업이다. 개인정보는 개인정보 보호 규제의 적용을 받는 반면, 개인정보가 아닌 정보는 그렇지 않다. 따라서 비식별 조치가 성공적으로 이루어졌는지 여부는 결국 개인정보 보호법상 규정된 개인정보의 범위에 해당하지 않도록 변환이 이루어졌는지 여부에 따라 달라진다.

법적으로는 <그림 7>의 왼쪽 그림과 같이 개인정보와 비개인정보가 이분법적으로 나뉘게 된다. 그러나 현실에서는 <그림 7>의 오른쪽 그림과 같이 가장 극단에는 개인정보, 반대편 극단에는 익명정보임이 분명한 정보가 있고, 그 사이에는 연속적인 스펙트럼의 형태로 존재한다. 제2장에서 예로 든 GIC 의료정보나 Netflix Prize 데이터는 적어도 해당 데이터를 공개하는 자의 입장에서 판단하기에 외견상으로는 개인정보가 아니는 것으로 판단되었을 것이다. 하지만 연구자들이 일정한 재식별 노력을 기울인 결과 일부 항목에 대해 개인을 식별할 수 있음이 밝혀졌다. 이처럼 재식별 가능성의 맥락에 따라 개인정보 판단은 달라질 수 있다.

<그림 7> 개인정보와 비(非)개인정보의 구분

개인정보의 범위에 관한 판단은 시점에 따라서도 달라질 수도 있다. 예

컨대 날씨에 관한 정보는 수집 당시부터 개인정보에 해당하지 않고, 시점이 달라져도 개인정보에 해당하지 않을 것이다. 하지만 개인정보처리자가 수집한 개인정보에 비식별 조치를 취할 경우, 더이상 개인정보에 해당하지 않게 될 수도 있다. 하지만 비식별 정보를 보유하고 있는 자가 추가적으로 다른 정보를 획득하게 된 경우에는 다시금 식별 가능성이 발생할 수도 있다.

이처럼 개인정보인지 여부를 분류하는 것은 역동적인(dynamic) 작업이다. 이는 문맥에 따라 동일한 데이터가 개인정보일 수도, 비개인정보일 수도 있는 것이다. 이러한 개인정보의 역동성은 개인정보 범위에 대한 법적 판단의 어려움을 낳는바, 이하에서는 개인정보 여부에 관한 법적 판단에 관한 구체적인 쟁점을 검토한다. 특히 2020년 개정된 우리 개인정보 보호법은 GDPR 규정을 다수 참조하였는바, 본 연구에서는 GDPR 상의 개인정보 개념을 중심으로 그 핵심적 법적 쟁점을 살펴본다.

제 2 절 GDPR상의 개인정보 개념과 해석상 쟁점

1. GDPR 상의 개인정보의 개념과 개인식별 가능성

GDPR 제4조 (1)항은 개인정보(Personal Data)를 다음과 같이 정의한다. “개인정보란 식별된 또는 식별 가능한 자연인(‘정보주체’)과 관련한 일체의 정보를 가리킨다. 식별가능한 자연인은 직접 또는 간접적으로, 특히 이름, 식별번호, 위치정보, 온라인 식별자를 참조하거나 해당인의 신체적, 심리적, 유전적, 정신적, 경제적, 문화적 또는 사회적 정체성에 특이한 하나 이상의 요인을 참조함으로써 식별될 수 있는 자를 가리킨다.” 이러한 개인정보의 개념 정의에 있어 핵심적 지표는 ‘식별된(identified)’ 또는 ‘식별가능한(identifiable)’이다.

어떤 데이터로부터 개인을 식별할 수 있는지를 판단하는 것이 쉽지 않은

작업인 경우가 많다. GDPR 전문 제26조는 개인정보와 비개인정보를 구별하기 위한 방법을 다음과 같이 규정하고 있다. “어떤 개인이 식별가능한지를 판단하기 위해서는 컨트롤러 또는 다른 사람이 개인을 직접 또는 간접적으로 식별하기 위해 사용할 것으로 합리적으로 가능한(reasonably likely) 모든 수단, 가령 싱글 아웃(singling out) 등을 고려해야 한다. 어떤 수단이 개인을 식별하기 위해 합리적으로 사용 가능한 것인지를 확인하기 위해서는, 식별하기 위해 소요되는 비용과 시간 등 객관적인 요소를 모두 고려하고, 처리 당시 가용한 기술과 기술적 발전을 모두 고려하여야 한다.”

어떤 데이터로부터 개인을 식별할 수 있는지를 판단할 때에는 해당 데이터로부터 특정 개인이 유일하게 식별될 수 있는지만을 기준으로 하는 것은 아니다. 특정 개인이 유일하게 식별되지는 않더라도 개인에 관한 다른 기록과 연결되어 해당 정보주체에 관한 정보가 추가적으로 밝혀질 수도 있고, 해당 데이터에는 포함되어 있지 않은 개인에 관한 다른 속성이 추론될 수도 있기 때문이다.

이러한 점에서 제29조 작업반은 개인식별 가능성 판단 기준으로 유일 식별 가능성(singling out), 연결 가능성(linkability), 추론 가능성(inference) 3가지를 제시하고 있다.[52] ① 유일 식별 가능성(singling out)이란 데이터셋에서 개인을 식별하는 일부 또는 모든 레코드를 분리할 수 있는 가능성을 의미한다. 가장 일반적인 개인식별 가능성 판단 기준이다. ② 연결 가능성(linkability)는 개인과 관련된 기록을 다른 기록과 연결(link)하는 것이 가능한 경우를 뜻한다. 만약 두 개 이상의 데이터셋이 동일한 정보주체에 관한 정보를 포함하는 경우, 재식별 공격자가 어떤 개인을 특정할 수는 없다고 하더라도, (상관관계 분석 등을 통해) 두 개의 레코드가 동일한 개인 집단에 속한다는 점을 알 수 있다면 이는 연결 가능한 것이다. ③ 추론 가능성(linkability)는 특정 속성을 다른 속성값의 집합으로부터 유의미한 확률로 추론할 수 있는 가능성을 의미한다. 제29조 작업반은 위 세 가지 경우가 모두 불가능해야 해당 데이터는 더 이상 개인

52) Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques", (2014), pp. 11-12.

정보에 해당하지 않는다고 설명한다. GDPR 전문 제26조는 명시적으로 유일 식별 가능성만을 언급하고 있으므로, 제29조 작업반이 제시한 바와 같이 연결 가능성과 추론 가능성도 개인 식별 가능 여부 판단의 고려 요소가 될 수 있다.[53]

GDPR 전문 제26조에 규정된 개인식별 가능성 판단 기준에서 도출되는 판단 방법은 '합리적으로 가능한 모든 수단'을 고려해야 한다는 것이다. 하지만 과연 어떠한 수단이 '합리적으로 가능'한 것인지는 규범적 판단의 영역이다. 이러한 규범적 판단에 있어서 고려해야 사항으로는 수용 가능한 재식별 위험의 수준이 어떠한지(절대적 접근과 상대적 접근), 누구의 관점에서 재식별 위험을 판단해야 하는지(객관적 접근과 주관적 접근), 고려 대상이 되는 기술과 시간적 범위는 어떻게 보아야 하는지 등이 있는바, 이하에서는 이에 관한 논의를 살핀다.

2. 규범적 판단의 쟁점

가. 절대적 접근과 상대적 접근

개인정보에 대해 비식별 조치를 취하더라도 개인을 재식별할 가능성이 전혀 없도록 하는 것은 쉽지 않은 일이다. 비식별 조치를 취한 후 남아 있는 재식별 가능성의 수준이 어느 정도 되어야 더 이상 개인정보가 아니라고 판단할 것인지에 대해서는 절대적 접근과 상대적 접근이 있다. 절대적 접근법은 재식별이 비가역적, 영구적, 불가능해야 할 것을 요구하는 것이고, 상대적 접근법은 합리적으로 재식별이 가능하지 않으면 족하다는 보는 것이다.

GDPR 도입 이전에는 유럽 개인정보보호 감독기관의 지침상 절대적 접근법을 취한 것으로 해석될 수 있는 여지가 많았다. 우선 제29조 작업반은 재식별 위험이 없어야 한다(no-risk)고 기준을 제시한 것으로 해석될 수 있다. 제29조 작업반의 개인정보 익명처리 기법 가이드라인은 "익명

53) Data Protection Commission, "Guidance on Anonymisation and Pseudonymisation" (2019. 6.) pp. 6-7.

처리란 비가역적으로(irreversibly) 식별을 방지하기 위해 개인정보를 처리하는 것"이라고 설명한 바 있다.[54] 또한 동 가이드라인은 "익명처리는 비가역적인 비식별화(irreversible de-identification)를 달성하기 위해 개인정보에 적용되는 기술"이라 한다.[55] 또한, 개인정보에 대해 익명처리 기법을 적용한 결과물은 "현재 기술 수준에 비추어 삭제된 경우와 마찬가지로 영구적이어야 하고(as permanent as erasure), 개인정보의 처리가 불가능하도록 만들어야 한다(making it impossible to process personal data)"고 규정한다.[56]

국가별 개인정보보호 감독기관도 유사한 입장을 표명한 바 있다. 예컨대, 프랑스의 Nationale de l'Informatique et des Libertés (CNIL)은 익명처리를 "재식별이 실무적으로 불가능(identification practically impossible)"하게 만드는 방법으로 구성된다고 보면서, 익명처리의 결과 개인정보를 더 이상 처리할 수 없도록 "비가역적으로 만드는 것"을 추구한다고 설명한다.[57] 프랑스 최고 행정법원 Conseil d'État는 개인정보가 개인에 대한 직접 또는 간접적인 식별이 '불가능(impossible)'해 진 경우에만 익명정보로 간주될 수 있다고 본 바 있다.[58]

절대적 접근법에 대해서는 개인정보에 대해 비식별 조치를 취하여 활용하는 것이 사실상 금지하는 것이라는 비판이 제기된다.[59] 절대적 접근법을 극단적으로 취하게 되면 원본 데이터셋이 파괴되지 않고 남아있는 한, 원본 데이터셋을 참조하여 비식별화된 개인정보가 재식별될 가능성을 고려해야 한다. 그렇게 된다면 컨트롤러가 비식별 조치를 취하더라도 여전히 법적으로 해당 정보가 개인정보로 취급되는 결과에 이르게 된다. 이 경우 컨트롤러가 비식별 조치를 취할 유인이 사라지게 되는 부작용을

54) Article 29 Data Protection Working Party (2014) (각주 52), p. 1.
55) 위의 글. p. 6.
56) 위의 글.
57) Finck, Michèle, and Frank Pallas, "They who must not be identified-distinguishing personal from non-personal data under the GDPR", International Data Privacy Law, (2020), p. 15.
58) Conseil d'État, 10ème - 9ème ch. réunies, décision du 8 février 2017, N° 393714, (Finck and Pallas (2020) (각주 56) p. 15에서 재인용).
59) Finck and Pallas (2020) (각주 56), pp. 35-36.

초래할 수도 있다.
이에 비해 GDPR의 문언은 상대적 접근법을 취한 것으로 해석될 여지가 크다. 즉, GDPR 전문 제26조는 재식별 가능성 여부를 판단함에 있어 “합리적으로 가능한(reasonably likely) 모든 수단”을 고려해야 한다고 한다고 규정하고 있다. 따라서 이 규정에 비추어 보면 비식별 조치가 이루어진 데이터에 대해 재식별을 위해 비합리적으로 과다한 시간, 노력, 비용이 소요되는 경우에는 더 이상 개인정보에 해당하지 않는다고 판단될 수 있다.

나. 객관적 접근과 주관적 접근

어떠한 데이터의 개인식별 가능성을 판단할 때에는 누구의 관점에서 이를 판단할 것인지를 결정해야 한다. 만약 해당 데이터를 보유하고 있는 자가 결합 가능한 다량의 정보를 보유하고 있고, 개인식별을 시도할 동기(motivation)가 크다면 개인식별 가능성이 높다고 판단될 여지가 커지게 되고, 이와 반대로 결합 가능한 데이터를 거의 보유하고 있지 않고(즉 오로지 공개 데이터만을 활용할 수 있을 뿐이고), 개인식별 시도의 동기가 적다면 개인식별 가능성이 낮아지게 될 것이기 때문이다. 여기서 개인정보의 컨트롤러만을 고려하면 족한지, 아니면 가능한 모든 제3자를 포함하여 판단해야 할 것인지가 쟁점이 된다. 이렇게 제3자가 정해지면, 그 제3자가 활용 가능한 배경 지식과 그의 능력을 고려해야 재식별 여부를 판단하게 된다.
객관적 접근법은 가능한 모든 제3자의 관점을 포함하여 판단해야 한다는 입장이다. GDPR 전문 제26조는 컨트롤러 이외에도 “다른 사람(another person)”이 재식별을 위한 사용할 수 있는 수단을 고려해야 한다고 규정하고 있으므로, 가능한 제3자를 모두 고려할 필요가 있다는 주장이다. 하지만 객관적 접근법에 따르면 항상 최악의 상황을 가정해야 하므로(예컨대, 고도의 능력을 갖춘 해커가 컨트롤러의 서버에 침입하여 재식별을 시도하는 상황), 재식별 위험 수준을 고려한 관리 방법을 도입하기 어렵

다는 비판이 제기된다.[60] 이에 비해 주관적 접근법은 재식별 가능성은 원칙적으로 컨트롤러의 관점에서 판단하여야 하고, 다만 합리적으로 판단할 때 해당 데이터에 접근 가능한 제3자도 함께 고려해야 한다고 본다. 주관적 접근법에 따르게 되면, 과연 어느 범위까지의 제3자를 고려할 것인지를 정해야 한다.

이 문제에 관하여 영국 개인정보보호 감독기관인 Information Commissioner's Office(이하 "ICO")는 '동기부여된 침입자 테스트(motivated intruder test)'라는 절충적 방법론을 제시하고 있다.[61] 즉 재식별을 시도할 동기가 부여되어 있는 침입자가 존재할 수 있다면, 그가 재식별을 달성할 수 있을지를 기준으로 판단해야 한다는 것이다. 동기부여된 침입자는 웹 검색, 신문 기사 검색, 소셜 네트워크 검색, 공적 기록이나 도서관 자료 등을 활용하여 재식별을 시도할 수 있는 자이다. 하지만 해킹 기술이나 특수 장비를 활용하거나 절도나 협박 등의 범죄적 수단을 통해 보안 조치가 이루어져 있는 데이터에 접근할 수 있는 것으로 가정하지는 않는다.[62]

유럽사법재판소에서 이 문제를 다룬 대표적인 사례로는 Breyer 사건이 있다.[63] 이 사건에 있어서는 독일 연방 정부에 의해 수집된 동적 IP 주소가 개인정보인지 여부가 문제되었다. 독일 연방 정부가 보유한 데이터만으로는 동적 IP 주소를 이용하여 개인을 식별할 수 없었으나, 인터넷 서비스 제공자(Internet Service Provider, 이하 "ISP")는 이러한 데이터를 보유하고 있다. 하지만, ISP의 관점에서 판단한다면 동적 IP 주소는 개인정보에 해당할 수 있으므로, 이를 개인정보로 보아야 할 것인가가 문제된 것이다. 여기서 개인정보에 해당하는지 여부는 컨트롤러의 관점(즉 독일 정부; 상대적 접근법)에서만 평가되어야 하는 것이 아니라 제3자 관점(ISP; 절대적 접근법)에서도 평가되어야 하는지에 따라 결론이 달라질 수 있다.

---

60) 위의 글. p. 17.

61) Information Commissioner's Office (ICO), "Anonymisation: Managing Data Protection Risk Code of Practice", (2012. 11.) pp. 22-25.

62) 위의 글.

63) Case C-582/14 Patrick Breyer [2016] EU:C:2016:779.

그런데, 유럽사법재판소는 독일법상 ISP가 정보주체를 식별할 수 있게 하는 데이터를 웹사이트 운영자에게 전송하는 것이 원칙적으로 금지되어 있지만, 독일 연방 정부는 사이버 공격이 이루어지는 경우 ISP에게 해당 데이터 제공을 강제할 수 있는 권한을 가지고 있다는 점에 주목하였다. 즉, 개인을 식별할 수 있는지 여부를 판단할 때에는 독일 연방 정부의 관점에서 판단해야 하지만, 독일 연방 정부가 개인을 식별할 수 있는 추가적인 데이터를 ISP로부터 취득할 가능성이 있으므로, 이는 개인정보에 해당한다고 본 것이다. 즉, Breyer 사건은 유럽사법재판소가 개인식별의 실제적 위험을 평가하였다는 점에서, GDPR 전문 제26조에 규정된 상대적 접근법을 취한 것으로 해석될 수 있다.

다. 고려 대상 기술과 시간적 범위

GDPR 전문 제26조는 개인을 식별하기 위해 합리적으로 사용 가능한 수단을 고려할 때 현재 이용 가능한 수단뿐만 아니라 '기술 발전'도 고려해야 한다고 규정하고 있다. 따라서 어느 정도의 시간을 고려해야 하는지가 쟁점이 된다. 먼 미래의 기술까지도 고려해야 한다면, 향후 양자 컴퓨팅 기술을 통해 각종 암호화 기술이 무력화되는 상황까지도 고려해야 할 수도 있다.

이에 대해 제29조 작업반은 해당 데이터를 처리하는 시점의 최선의 기술(the state of the art in technology)은 물론 장래 데이터 처리가 이루어질 기간 동안 이루어질 기술 발전 가능성까지도 함께 고려해야 한다는 의견을 제시한 바 있다.[64] 여기서는 해당 데이터의 보관 기간을 함께 고려하여 판단되어야 한다. 만약 어떤 데이터가 1개월 동안만 보관될 것이고, 현재의 기술 수준으로 개인식별이 불가능하다면 개인정보에 해당하지 않을 것이다. 하지만 개인정보가 10년 동안 보관될 예정인데, 9년이 지난 후에 기술 발전을 통해 개인을 식별할 수 있게 된다면, 해당 시점부터는 개인정보로 취급되어야 할 것이다.[65]

---

64) Article 29 Working Party, Opinion 04/2007 on the Concept of Personal Data (WP 136) 01248/07/EN, p. 15.

또 다른 쟁점은 미래의 다른 데이터의 가용성에 대한 고려이다. 현재 비식별조치가 이루어진 데이터를 보유한 기업이 미래에 개인을 식별할 수 있게 하는 추가적인 정보를 취득하거나 이에 대한 접근권을 얻게 된다면, 추가적인 정보에 대한 접근할 수 있게 된 시점부터는 개인정보로 고려될 필요가 있다. 따라서 컨트롤러는 비식별조치가 이루어진 데이터를 둘러싼 환경에 대해 지속적으로 모니터링을 해야 하고, 그 모니터링 결과에 따라 개인정보에 해당할 것인지가 달라지게 된다. 하지만, 미래에 어떤 데이터가 생성되거나 획득될 수 있는지 예측하기 어렵다는 문제가 있다. 따라서 결국 개인정보의 동적 특성을 인정하고 비식별조치된 데이터라고 하더라도 식별이 가능한지 여부를 주기적으로 평가할 필요가 있다.

3. GDPR과 위험기반 접근법

GDPR은 규정 전반에 걸쳐 개인정보의 처리로 인해 발생하는 위험을 평가하고, 그에 비례한 조치를 취하도록 요구하는 경우가 많다. 이처럼 위험은 GDPR에 있어 핵심적인 개념으로서, 반복적으로 언급되고 있다.[66] 예컨대 가명처리는 '관련 정보주체에게 미치는 위험을 줄이기 위해' 장려된다.[67] 민감 정보는 "기본권 및 자유에 대한 중대한 위험"에 비추어 특별한 보호를 받을 필요가 있다.[68] 개인정보 유출 사고가 발생한 경우 컨트롤러는 "개인의 권리와 자유에 대한 위험을 초래할 가능성이 낮다"고 입증할 경우 고지의무가 면제된다.[69] 행동강령(codes of conduct)은 "컨트롤러의 개인정보 처리로부터 발생하는 위험을 고려하여 컨트롤러의 의무를 조율"할 수 있다.[70] 개인정보 보호를 위한 기술 및 관리적 조치에 있어, 컨트롤러는 "개인정보의 처리가 개인의 권리 및 자유에 미치는 위

65) 위의 글.
66) Finck and Pallas (2020) (각주 56), p. 34.
67) GDPR 전문 제28조.
68) GDPR 전문 제51조.
69) GDPR 전문 제85조.
70) GDPR 전문 제98조.

험과 처리의 구체적 성격 등을 고려"해야한다.[71] 설계에 의한 개인정보 보호(Data Protection by Design)이나 기본설정에 의한 개인정보 보호(Data Protection by Default)는 "개인정보의 처리가 개인의 권리와 자유에 미치는 위험을" 고려하여 이루어져야 한다.[72] 개인정보의 처리가 "개인의 권리와 자유에 대한 높은 위험"을 초래할 것으로 예상되는 경우에는 개인정보보호 영향 평가가 요구된다.[73]

이와 같이 GDPR 전반에 걸친 위험을 고려한 접근법은 개인정보의 범위를 해석함에 있어서도 고려되어야 한다. 앞서 검토한 개인식별 가능성에 대한 절대적·객관적 접근법은 구체적 맥락에 맞추어 위험을 고려한 조치를 취하는 것을 어렵게 한다. 절대적·객관적 접근법에 따르면 개인정보에 대한 비식별 조치는 사실상 위험이 0이 하여야 하고, 위험이 잔존하고 있다면 여전히 개인정보에 해당된다고 해석될 가능성이 크기 때문이다. 따라서 GDPR의 체계에 부합하는 해석을 위해서는 GDPR 상의 개인정보의 범위 판단에 있어 상대적·주관적 접근법을 취해야 할 것이다. 이러한 점에서 영국 ICO의 비식별 조치에 관한 가이드라인에서 비식별 조치는 "위험이 전혀 없도록 해야(completely risk free)" 할 필요는 없고, 다만 "재식별의 위험이 희박할 때까지(until it is remote)" 비식별조치가 이루어져야 한다고 설명한 것은 위와 같은 입장에 부합하는 것으로 평가할 수 있을 것이다.[74]

제 3 절 GDPR에서의 가명처리

1. GDPR 상의 가명처리의 개념: 약한 가명처리와 강한 가명처리

71) GDPR 제24조 제1항.
72) GDPR 제25조 제1항.
73) GDPR 제35조 제1항.
74) Information Commissioner's Office (2012. 11.) (각주 61), p. 6.

GDPR에서 고려하는 개인정보의 비식별 조치의 또 다른 중요한 축은 '가명처리(pseudonymisation)'이다. GDPR 제4조 제5항은 "가명처리란 추가 정보의 사용 없이는 더 이상 특정 정보주체에 귀속될 수 없는 방식으로 개인정보를 처리하는 것이다. 다만, 그 같은 추가 정보는 별도로 보관하고, 기술 및 관리적 조치를 적용하여 해당 개인정보가 식별된 또는 식별될 수 있는 자연인에 귀속되지 않도록 해야 한다."고 규정한다(가명처리에 관한 GDPR의 정의는 국내 개인정보 보호법에 거의 그대로 이어졌는바, 국내 개인정보보호법의 해석에 관하여는 제4장에서 검토한다).
가명처리는 일반적으로 직접식별자를 삭제하거나 대체하는 방식으로 이루어지는데, 그 결과 여전히 다른 속성정보에 의해 개인이 식별될 가능성이 남아 있게 된다.[75] 따라서 가명처리된 결과물은 여전히 개인정보에 해당한다. 이에 대해 제29조 작업반은 "가명처리는 익명처리 방법이 아니다. 이는 단지 데이터셋과 정보주체의 원래 식별자 사이의 연결성을 감소시킬 뿐이며, 그 결과 유용한 보안 조치가 된다."고 설명한 바 있다.

가. GDPR 제11조와 가명처리의 관계

GDPR상의 가명처리를 이해하기 위해서는 그 정의조항뿐만 아니라 GDPR 제11조(식별을 요하지 않는 개인정보의 처리)를 함께 고려할 필요가 있다. GDPR 제11조 제1항은 "컨트롤러가 개인정보를 처리하는 목적상 정보주체의 식별을 요구하지 않거나 더 이상 요구하지 않아도 되는 경우, 그 컨트롤러는 GDPR을 준수할 목적에 한하여 정보주체를 식별하기 위한 추가 정보를 유지, 취득, 처리할 의무를 가지지 않는다."고 규정한다. 동조 제2항은 "본 조 제1항의 경우, 컨트롤러가 정보주체를 식별할 수 없음을 입증할 수 있다면, 제15조 내지 제20조까지는 적용되지 않는다. 단, 정보주체가 위 조항에 따라 자신의 권리를 행사하기 위한 목적으로 본인을 식별할 수 있는 추가 정보를 제공하는 경우는 예외로 한다."

75) Article 29 Data Protection Working Party (2014) (각주 52), p. 20.

고 규정한다. 즉, 위 규정은 식별을 요하지 않는 정보처리에 있어 정보주체의 권리를 제한하는 취지인데, 그 적용이 배제되는 GDPR 제15조 내지 제20조는 열람권, 삭제권, 처리 제한권, 전송권 등 정보주체의 권리에 관한 조항들이다.

위 조항의 목적을 이해하는 것은 어렵지 않다. 가령 어떤 웹사이트 운영자가 사이트 이용 통계를 확인하기 위한 목적으로 이용자의 접속 국가와 접속 시간 정보만을 수집하는 경우를 상정해 보자. 나아가 이용자 ID, 이름, 이메일 주소 등 개인을 식별할 수 있는 정보는 전혀 수집하지 않고 있다고 해 보자. 만약 정보주체가 자신이 해당 웹사이트를 접속한 이력을 삭제할 것을 요구하더라도 웹사이트 운영자는 어떠한 접속 이력이 위 요청을 한 정보주체에 해당하는 것인지 알 수 없다. 위 규정의 목적은 이러한 상황에서 열람권, 삭제권 등을 비롯한 정보주체의 권리를 적용하지 않겠다는 것으로 이해된다. 하지만 만약 정보주체가 자신의 컴퓨터에 남아 있는 접속 시간 자료를 제공하면서 삭제를 요청한 경우라면 (즉, '본인을 식별할 수 있는 추가 정보를 제공하는 경우') 정보주체의 권리를 보장할 수 있을 것이다.

그렇다면 GDPR 제4조 제5항에서 정의된 '가명처리'와 제11조의 '식별을 요하지 않는 개인정보의 처리'의 관계는 어떠한가? 가명처리도 식별을 요하지 않는 형태의 개인정보의 처리를 위한 수단이므로 양자는 일응 유사한 것으로 이해될 여지가 있다. 앞선 웹사이트 운영자의 사례에서 본 바와 같이 접속 국가와 접속 시간 정보만으로서는 개인을 식별할 수 없고, 추가 정보가 제공되어야만 개인을 식별할 수 있다. 따라서 웹사이트 운영자가 수집하는 정보는 일응 가명정보라고 볼 여지도 있다.

하지만 특정 국가에서 특정 시간대에 접속하는 이용자의 수가 매우 제한적인 상황을 가정해 보자. 그러면 해당 정보를 이용해서도 정보주체를 식별해 낼 가능성이 존재할 수 있다. 따라서 웹사이트 운영자가 수집한 정보가 곧바로 개인정보가 아니라고 단정할 수는 없다. 개인정보에 해당하는지 여부를 판단하기 위해서는 위 제2장에서 검토한 바와 같이 합리적으로 가능한 모든 수단을 이용하여 개인을 식별할 수 있는지 여부를

판단해야 한다. 판단 결과에 따라서는 위와 같이 수집한 사이트 이용정보 기록(접속 국가와 접속 시간)도 개인정보에 해당한다고 볼 수 있다.
이제 시나리오를 약간 변경해 보자. 웹사이트 운영자가 접속 통계를 작성하기 위한 목적으로 이용자의 IP 주소, 접속 국가 및 접속 시간 정보를 수집하고 있다고 하자. 웹사이트 운영자가 이용자 IP 주소를 암호화하여 보관한다. 쌍방향 암호화 기술을 활용하여 복호화를 위한 암호 키(key)를 사용하지 않고서는 이를 복원할 수 없다고 가정한다면, 위 웹사이트 운영자는 GDPR 제11조의 적용을 받을 수 있는가? 웹사이트 운영자가 암호 키를 삭제하지 않고 보관하고 있는 동안에는 정보주체의 권리행사를 배제할 근거가 없다. 만약 어떤 정보주체가 자신의 접속 이력을 삭제해 달라고 요청한다면, 웹사이트 운영자는 암호 키를 이용하여 원래 데이터를 복원한 다음, 요청받은 정보주체의 기록을 삭제할 수 있기 때문이다.
이처럼 컨트롤러가 정보주체를 재식별하기 위한 추가 정보(위 사례에서는 암호키)를 보유하고 있는지, 이를 이용하여 용이하게(readily) 정보주체를 재식별해 낼 수 있는지 여부에 따라 GDPR 제11조의 적용 여부는 달라질 수 있다. 요컨대, GDPR 제11조는 정보주체에 대한 권리에 대한 적용 예외를 규정하고 있다. 그 에 따라 컨트롤러가 그저 직접식별자를 삭제·대체하는 수준의 가명처리보다 더욱 강력한 조치를 취한 경우에만 적용될 수 있는 것이다.

나. 약한 가명처리와 강한 가명처리

이상과 같이 GDPR 제4조 제5항에서 정한 '가명처리'와 제11조에서 정한 '식별을 요하지 않는 개인정보의 처리'는 구분된다. 후자의 경우에는 정보주체에 대한 권리가 배제되므로, 후자의 적용을 받기 위해서는 컨트롤러가 더욱 강력한 조치를 취할 필요가 있게 된다. 이처럼 전자를 약한 가명처리(basic pseudonymization), 후자를 강한 가명처리(strong pseudonymization)로 구분하는 것은 GDPR 상의 가명처리를 이해하는

데 있어 유용하다.[76]

약한 가명처리는 데이터셋에서 직접식별자를 삭제 또는 대체하는 것을 의미한다. 개인정보를 재식별해 내기 위한 추가 정보는 별도로 보관, 관리되어 있으나,[77] 추가 정보에 용이하게 접근 가능한 상황일 경우 약한 가명처리가 이루어진 것으로 이해된다.[78] 이에 비해 강한 가명처리는 직접식별자를 삭제·대체할 뿐만 아니라, 공개 정보 등을 이용하여 간접식별자에 의해서 식별될 가능성을 막기 위하여 데이터에 대한 변형이 이루어진 경우를 의미한다. 필요하다면 정보주체를 재식별을 위한 추가 정보를 삭제할 필요도 있다.[79] 이러한 구분에 비추어 보면, GDPR 제4조 제5항에서 정한 가명처리는 '약한 가명처리'를, GDPR 제11조는 '강한 가명처리'를 규정한 것으로 해석될 수 있다. 강한 가명처리가 이루어진 경우 컨트롤러가 정보주체를 재식별해 낼 수 없으므로, 정보주체의 권리를 배제하는 것이 정당화될 수 있다. 그럼에도 불구하고 여전히 속성정보를 이용하여 개인을 간접적으로 재식별해 낼 가능성은 상존하므로, 강한 가명처리가 이루어진 경우에도 여전히 개인정보에 해당한다.

이처럼 (강한 익명처리이든 약한 가명처리이든) 개인정보에 대해 (익명처리에 이르지 못한 수준의) 가명처리가 이루어진다면, 해당 정보는 여전히 개인정보이다. 하지만, 가명정보의 형태로 처리함으로써 정보주체에 대한 위험을 줄일 수 있다. 따라서 그 활용을 장려하도록 GDPR은 여러 인센티브가 부여하고 있는바, 이하에서는 가명처리에 따른 어떠한 인센티브가 존재하는지 검토한다.

2. GDPR상 가명처리 인센티브

76) Hintze, Mike, "Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency", International Data Privacy Law 8.1, (2018), pp. 91-93; Arbuckle, Luk and Khaled El Emam, "Building an Anonymization Pipeline: Creating Safe Data", O'Reilly Media, (2020), pp. 140-144.

77) GDPR 제4조 제5항 단서

78) Arbuckle and El Emam (2020) (각주 76), p. 142.

79) 위의 글.

GDPR은 가명처리에 대해 다양한 인센티브를 부여하고 있다. 그 예로는 ① 정보주체의 동의 없는 개인정보의 목적외 처리의 허용 ② 정보주체의 권리 적용 예외(특히 삭제권의 배제) ③ 설계 및 기본설정에 의한 개인정보 보호 ④ 개인정보 보안 조치 의무의 이행 ⑤ 데이터 유출 통지의 예외가 있다.

가. 정보주체의 동의없는 처리의 허용 가능성

GDPR 제6조 제4항에 따르면 컨트롤러는 개인정보를 수집한 당초 목적과 양립가능한(compatible) 범위 내에서 정보주체의 동의 없이 목적외로 처리할 수 있다. 그런데 양립 가능성을 판단하는 요소로 암호화나 가명처리 등의 적절한 안전조치가 존재하지는 여부가 규정되어 있다.[80] 즉, 가명처리가 이루어진 경우에는 양립가능성이 인정되어 정보주체의 별도 동의 없이 당초 수집한 목적과 다른 목적으로 개인정보를 활용할 수 있을 가능성이 커지게 된다. 이는 약한 가명처리나 강한 가명처리 양자 모두에 대해서 부여되는 인센티브이지만, 만약 강한 가명처리가 이루어졌다면 양립가능하다고 인정될 가능성이 더욱 커지게 된다.[81]

나. 정보주체의 권리 적용 예외

GDPR 제15조부터 제20조는 정보주체의 권리에 관한 규정을 두고 있다(제15조 열람권, 제16조 정정권, 제17조 삭제권, 제18조 처리에 대한 제한권, 제19조 정정·삭제·처리제한에 관한 고지, 제20조 이전권). 특히 개인정보를 처리하는 기업 입장에서는 제17조의 삭제권(잊힐 권리)이 실무상 부담스러운 요건에 해당한다. 이를 구현하기 위해서는 보유하고 있는 모든 데이터를 상세히 파악하고, 경우에 따라서는 데이터 시스템을 전체

80) GDPR 제6조 제4항 (e)호

81) Hintze, Mike, and Khaled El Emam. "Comparing the benefits of pseudonymisation and anonymisation under the GDPR." Journal of Data Protection & Privacy 2.2 (2018): 145-158.

적으로 재설계하고 새로운 툴을 적용해야 할 필요가 있기 때문이다.[82]
그런데 앞서 설명한 바와 같이 강한 가명처리가 이루어진 경우, 컨트롤러는 GDPR 제15조 내지 제20조에 규정된 정보주체의 권리를 보장할 필요가 없게 된다.[83] 다만, 이러한 경우 컨트롤러는 그저 정보주체의 요청을 거절할 수 있는 것은 아니고, "정보주체를 식별할 수 있는 지위에 있지 않음을 입증"해야 할 의무를 부담한다.[84] 이와 달리, 약한 가명처리만 이루어진 경우에는 여전히 정보주체의 권리를 보장해야 한다. 즉, 컨트롤러는 추가 정보를 이용하여 여전히 손쉽게 정보주체가 누구인지 식별할 수 있으므로, 정보주체의 열람, 정정, 삭제, 처리제한, 이전 요청을 받으면 그 요청에 응해야 한다.
예를 들어, 데이터 분석회사가 의료데이터에 대한 임상 연구를 위하여 환자의 가명처리된 의료보험번호, 성별 및 나이 정보만 보유하고 있는데, 환자가 자신의 정보를 삭제할 것을 요청한 경우를 상정해 보자. 만약 데이터 분석회사가 가명처리된 의료보험번호로부터 원래 의료보험번호를 복원해 낼 수 있다면, 이는 약한 가명처리에 해당하는 것이고 정보주체의 요청을 받아들여야 할 것이다. 하지만 이와 달리 데이터 분석회사가 원래 의료보험번호를 복원해 낼 수 있는 추가적인 정보를 보유하고 있지 않고, 달리 해당 환자를 식별할 수 없다는 점을 입증할 경우에는 정보주체의 요청을 거절할 수 있게 된다.

다. 설계 및 기본설정에 의한 개인정보 보호

GDPR 제25조는 '설계 및 기본설정에 의한 개인정보 보호(Data Protection by Design and by Default)'을 규정하고 있다. 이는 개인정보를 보호하기 위한 사전적·예방적 조치를 취할 의무를 규정한 것이다. 구체적으로, GDPR 제25조 제1항에 따르면 컨트롤러는 "최신 기술, 실행 비용, 그리고 개인정보 처리의 성격, 범위, 상황 및 목적을 고려하여 기

82) 위의 글. pp. 5-6.
83) GDPR 제11조 제2항
84) GDPR 제12조 제2항

술 및 관리적 조치를 ... 이행"해야 하고, 그러한 기술 및 관리적 조치는 "GDPR과 정보주체의 권리 요건을 충족시키기 위해 필요한 안전조치가 개인정보처리에 통합될 수 있도록 설계(designed)"되어야 한다. 또한 GDPR 제25조 제2항에 따르면 "컨트롤러는 기본설정을 통해(by default) 특정 처리 목적에 필요한 개인정보만 처리되도록 적절한 기술 및 관리적 조치를 이행"해야 한다. 즉, 기업이 서비스를 설계하고, 프로토타입을 만들고, 이를 시험하고 실제로 구현하는 단계에 있어 GDPR 규정의 준수 여부를 사전적으로 고려해야 하고, 개인정보를 보호하는 조치가 기본설정으로 포함해야 하는 것이다.

그런데, 가명처리는 설계에 의한 개인정보 보호를 위해 위 조항 따라 요구되는 기술 및 관리적 조치(appropriate technical and organizsational measures) 유형의 예시로 명시적으로 언급되어 있다. 즉, 컨트롤러가 가명처리를 할 경우 설계 및 기본설정에 의한 개인정보 보호를 위한 조치가 이루졌다고 인정될 가능성이 높아지게 된다. 물론 가명처리는 요구되는 기술 및 관리적 조치의 예시에 해당하므로, 가명처리가 이루어졌다고 하여 곧바로 GDPR 제25조가 준수되었다고 단정지을 수는 없을 것이다. 그러나 가명처리가 이루어진다면 컨트롤러는 부분적으로는 이러한 의무를 준수한 것이라고 판단할 수 있을 것이고, 완전히 개인을 식별 가능한 개인정보와 비교해 볼 때 가명정보에 대해서는 추가적으로 이행해야 할 조치의 부담이 상대적으로 덜할 것이다.[85]

라. 개인정보의 보안 조치

GDPR 제32조 제1항에 따르면 "컨트롤러와 프로세서는 개인정보의 처리가 개인의 권리 및 자유에 미치는 위험의 다양한 가능성 및 정도와 함께 최신 기술, 실행 비용, 그리고 처리의 성격, 범위, 상황 및 목적을 고려하여, 해당 위험에 적정한 보안 수준을 보장하기 위해 ... 적정한 기술 및 관리적 조치를 이행해야 한다". 위 조항은 위험에 적정한 보안 보안 수

85) Hintze and El Emam (2018) (각주 81).

준을 보호하기 위한 적정한 기술 및 관리적 조치의 예시를 들고 있는바 그 중 가장 먼저 나열되어 있는 것이 개인정보의 가명처리 및 암호화이다. 이는 가명처리를 통해 정보주체의 권리 및 자유에 대한 위험을 줄일 수 있다는 점이 명시적으로 반영된 것이다. 만약 약한 가명처리를 수행할 경우에는 부분적으로 위 의무 준수의 근거가 될 것이고, 강한 가명처리의 경우 위 조항의 준수 근거가 증가할 것이다.[86]

마. 데이터 유출 통지의 예외

GDPR 제34조는 컨트롤러에 대해 개인정보의 유출(personal data breach)이 발생할 경우 지체 없이(가급적 이를 알게 된 후 72시간 내에) 감독기관에 해당 개인정보의 유출 사실을 통지할 의무를 규정하고 있다. 다만, 해당 개인정보의 유출로 인하여 "개인의 권리와 자유에 위험을 초래할 것으로 예상되지 않는 경우는 예외"로 하고 있다. 또한, GDPR 제35조는 정보주체에게 개인정보의 유출에 대한 통지의무를 규정하고 있는바, 동조 제3항 (a)호는 이러한 통지의 예외로서 "컨트롤러가 적절한 기술 및 관리적 보호조치를 시행하였고, 그 조치, 특히 암호처리 등과 같이 관련 개인정보를 열람 권한이 없는 개인에게 이해될 수 없도록(unintelligible) 만드는 조치가 해당 유출로 영향을 받은 개인정보에 적용된 경우"를 규정하고 있다.

위 조항은 가명처리를 명시적으로 언급하고 있지는 않으나, 개인정보의 유출로 인한 개인의 권리와 자유에 대한 위험 평가에 있어서는 가명처리가 이루어졌는지 여부가 고려될 것이다.[87] 특히 강한 가명처리가 이루어진 경우에는 이와 같이 데이터 유출 통지 의무의 예외가 인정될 가능성이 더욱 커질 것이다.

3. 가명처리와 무관한 의무

86) 위의 글. p. 7.
87) 위의 글.

이상에서 살펴본 바와 같이 GDPR은 가명처리를 취한 경우에 대한 여러 인센티브를 부여하고 있지만, 가명처리를 적용하더라도 여전히 적용되는 조항도 적지 않다. 예컨대, ① 개인정보 보유 기간 제한(제5조 제1항 (e)호), ② 개인정보 수집시 정보주체에 대한 고지의무(제13, 14조) ③ 처리 활동의 기록(제30조) ④ 개인정보의 역외 이전 등이 그 예이다. 이를 구체적으로 살펴보면 다음과 같다.

우선 GDPR 제5조 제1항 (e)호는 개인정보가 "처리목적 달성에 필요한 기간 동안만 정보주체를 식별할 수 있는 형태로 보유"될 것을 요구한다. 따라서 컨트롤러는 개인정보 보유 필요성을 평가하고, 적절한 보유 기간을 수립해야 하는데, 가명처리를 수행하였다고 하여도 이러한 필요성 평가 및 보유 기간 수립의무가 면제되는 것은 아니다. 다만 컨트롤러는 가명정보에 대해서 필요한 보유 기간이 더 길다고 주장할 수 있는 여지는 존재할 것이다.[88]

또한 GDPR 제13조와 제14조는 컨트롤러가 개인정보를 수집할 경우 정보주체에 일정한 사항을 고지할 의무를 부여하고 있는바. 이는 개인정보의 수집·이용·공개에 대해 정보주체에 대한 투명성을 요구하는 GDPR의 기본 원칙을 반영한 것이다. 따라서 컨트롤러는 가명처리를 하더라도 여전히 정보주체에 대한 수집·이용·공개에 관한 고지 의무를 부담하게 된다.[89] 다만, 가명처리가 이루어진 경우에는 고지의 명시성(prominence)의 수준이 낮아질 수 있다. 가령 강한 가명처리가 이루어져 있다면 정보 주체가 해당 고지를 찾아볼 수 있는 수준(discoverable notice)도 적법하다고 인정될 가능성이 높아질 것이다.[90]

나아가 GDPR 제30조는 개인정보 처리 활동에 대한 기록 보존의무를 규정하고 있다. 이때 문서화하여 보존해야 할 기록에는 컨트롤러, 개인정보 담당관의 이름 및 연락처, 처리의 목적, 정보주체의 범주 및 개인정보의 범주, 역외 이전에 관한 사항, 개인정보의 삭제 예상 기한, 기술 및 관리적 안전조치에 대한 설명 등이 포함된다. 이러한 개인정보 처리 활동에

88) 위의 글. p. 8.
89) 위의 글. p. 4.
90) Hintze (2018) (각주 76), pp. 95 - 96.

대한 문서화 및 기록 작성의무는 가명정보에도 동일하게 적용된다.[91]
마지막으로 GDPR 제44조 내지 제49조는 개인정보의 역외 이전을 제한하면서 제한된 조건 하에서만 이를 허용하고 있다. 예를 들면 EU 집행위원회가 개인정보의 이전대상국이나 국제기구의 개인정보 보호 수준이 적합한 수준인지 평가하여 적정성 결정을 받은 경우에는 개인정보의 국외이전이 허용된다.[92] 가명처리된 정보라 하더라도 개인정보의 역외 이전에 대한 제한 조건은 준수하여야 한다.[93]

제 4 절 소결

본 장에서는 GDPR 상으로 개인정보인지 여부를 판단하는 것은 역동적인 작업이라는 점을 보았다. 즉, 누구의 관점에서 개인정보인지 여부를 평가하는 것인지, 개인정보가 어떠한 맥락에서 처리되는 것인지를 고려하지 않은 채, 특정한 정보가 곧바로 개인정보인지 여부를 결정지을 수는 없다. 특히 GDPR은 전체적으로 위험기반 접근법을 취한 것으로 해석되는바, GDPR의 이러한 태도에 비추어 개인정보의 개념 판단에 있어서도 정보주체에 대한 위험성을 고려하여야 한다.
다른 한편, GDPR은 '가명처리'의 개념을 도입하였는바, 이는 약한 가명처리와 강한 가명처리로 구분된다는 점을 살펴 보았다. 이러한 구분은 GDPR상의 다양한 가명처리 인센티브를 이해하는데 도움이 된다. 특히 강한 가명처리가 이루어진 경우 정보주체의 권리에 대한 예외가 인정될 수 있음에 비해, 약한 가명처리만이 이루어진 경우에는 이러한 예외가 적용될 수 없다는 차이가 있다. 이는 다음 장에서 보는 바와 같이 우리 개인정보 보호법의 '가명정보'의 해석에 있어서도 시사점을 준다.

---

91) Hintze and El Emam (2018) (각주 81), p. 8.
92) GDPR 제45조.
93) Hintze and El Emam (2018) (각주 81), pp. 6-7.

제 4 장 개인정보 보호법과 개인정보 비식별 조치

제 1 절 개정 개인정보 보호법상의 개인정보의 범위

개인정보 보호법상의 개인정보의 범위에 있어서는 2020. 2. 4. 개정을 통해 상당한 변화가 이루어졌다. 개인정보 보호법 제2조 제1호는 '개인정보'를 3가지 유형으로 구분한다. 우선, "성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보"가 있다(동호 가.목). 이는 강학상 또는 실무상 '식별정보'라고도 한다.[94] 다음으로, "해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보"가 있다(동호 나.목). 이를 강학상 또는 실무상 '식별가능정보'라고도 한다.[95] 이상의 규정은 개정 이전의 규정과 동일하다. 2020년 개정법상 중요한 변화는 결합용이성 판단 기준을 명확하게 하고, 가명정보의 개념을 도입한 것이다. 우선 '결합용이성'이란 식별가능정보가 다른 정보와 쉽게 결합할 수 있는 정도를 의미하는 것인데, 이는 식별가능정보의 범위를 결정하는데 중요한 역할을 하는 개념 지표가 된다. 개정법은 결합용이성 판단을 위해서 "다른 정보의 입수 가능성 등 개인을 알아보는데 소요되는 시간, 비용, 기술 등을 합리적으로 고려하여야 한다"는 점을 명시하였다(동호 나.목 단서).

또한 개정법은 식별정보나 식별가능정보를 "가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼

94) 개인정보 비식별 조치 가이드라인(2016), 5면. 가명정보 처리 가이드라인(2020), 17면.
95) 개인정보 비식별 조치 가이드라인(2016), 6면. 가명정보 처리 가이드라인(2020), 17면.

수 없는 정보"를 '가명정보'로 정의하였다(동호 다.목). 이 때 가명처리는 "개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것"으로 정의되어 있다(동조 제1의2호).

나아가 법 제58조의2는 "이 법은 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아 볼 수 없는 정보에는 적용하지 아니한다"는 조항을 신설하였다. 이처럼 시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없도록 하는 조치를 강학상 '익명처리'라 하고, 익명처리된 정보를 '익명정보'라 한다.

개인정보 보호법의 개정 이유는 "4차 산업혁명 시대를 맞아 핵심 자원인 데이터의 이용 활성화를 통한 신산업 육성이 범국가적 과제로 대두되고 있고, 특히, 신산업 육성을 위해서는 인공지능, 클라우드, 사물인터넷 등 신기술을 활용한 데이터 이용이 필요한 바, 안전한 데이터 이용을 위한 사회적 규범 정립이 시급한 상황"이기 때문이라 설명된다.[96] 이러한 문제 인식 하에서 개정법은 "정보주체의 동의 없이 과학적 연구, 통계작성, 공익적 기록보존 등의 목적으로 가명정보를 이용할 수 있는 근거를 마련"한 것이다.

위와 같은 개정은 상당 부분 GDPR의 영향을 받은 것으로 평가할 수 있다. 무엇보다도 개정 또는 신설된 조항의 문언 자체가 GDPR의 내용과 상당히 유사하다. 따라서 개정 개인정보 보호법의 해석에 있어 위 제3장에서 검토한 GDPR 상의 개인정보와 가명처리의 개념과 그 해석상 쟁점이 주요한 참고가 될 수 있다. 이하에서는 이를 전제로 개정 개인정보 보호법의 해석상 주요 쟁점을 익명처리/익명정보와 가명처리/가명정보로 구분하여 검토한다.

96) 개인정보보호법(법률 제16930호, 2020. 2. 4. 일부개정) 제정·개정이유

제 2 절 익명처리와 익명정보

1. 개정 전 개인정보 보호법과 개인식별 가능성 판단

개인정보 보호법상 개인정보의 범위에 포함되는지 여부를 판단함에 있어 핵심적인 개념 지표가 되는 것은 “개인을 알아 볼 수 있는”지 여부이다. 제3장에서 검토한 바와 같이 개인식별 가능성 판단에 있어서는 ① 비식별 조치의 수준에 관한 절대적 접근과 상대적 접근, ② 판단의 준거 주체에 관한 객관적 접근과 주관적 접근, ③ 고려 대상 기술과 시간적 범위가 주로 쟁점이 된다. 그런데 개정 전 개인정보 보호법 하의 국내 판결 중에는 절대적·객관적 접근법을 취한 것으로 평가되는 사건(이른바 ‘증권통’ 사건)과, 비식별 조치의 기술적 적절성을 다룬 사건(이른바 ‘약학정보원’ 사건)이 있는바, 이하에서는 각 판결 중 개인정보성 판단 부분을 다룬다.

가. ‘증권통’ 사건과 절대적·객관적 접근

‘증권통’ 사건의 개요는 다음과 같다.[97] 피고인인 증권사는 스마트폰의 증권 관련 앱의 로그인 절차를 간편하게 하기 위하여 스마트폰 이용자의 동의 없이 국제단말기 인증번호(IMEI)와 범용 가입자 식별 모듈(USIM) 일련번호의 조합 정보 또는 IMEI와 개인 이동전화번호의 조합 정보를 수집하였다. 그런데, IMEI나 USIM 일련번호와 결합할 수 있는 개인정보는 이동통신사가 보유, 관리하고 있고, 엄격한 통제를 가하고 있으므로 해당 증권사는 이 정보를 입수할 수 없었다. 이 사건에서는 위와 같은

97) 서울중앙지방법원 2011. 2. 23. 선고 2010고단5343 판결.

경우에도 IMEI나 USIM 일련번호가 “다른 정보와 쉽게 결합하여 개인을 알아볼 수 있는 정보”에 해당하는지 쟁점이 되었다.

이에 대해 서울중앙지방법원은 “‘쉽게 결합하여 알아 볼 수 있다’는 것은 쉽게 다른 정보를 구한다는 의미이기보다는 구하기 쉬운지 어려운지와는 상관없이 해당 정보와 다른 정보가 특별한 어려움 없이 쉽게 결합하여 특정 개인을 알아볼 수 있게 되는 것을 말한다 할 것이다. 오늘날 정보화의 진전에 따라 종전에 쉽게 결합할 수 없었던 정보들이 쉽게 결합되어 개인을 식별할 수 있는 자료로 작용되기에 이르렀고, 따라서 기계적인 정보라 하더라도 특정 개인에게 부여되었음이 객관적으로 명백하고, 이러한 정보를 통하여 개인이 식별될 가능성이 크다면 이를 개인정보로 봄이 상당”하다고 판단하였다.

즉, 이 판결은 해당 증권사가 개인을 식별하기 위한 다른 정보를 용이하게 입수할 수 있는지를 고려하지 않고, 다른 정보를 입수한 상황을 전제로 하여 개인식별 가능성을 판단하였다는 점에서 절대적·객관적 접근법을 취한 것으로 해석된다. 이에 대해서는 개인식별 가능성(특히 결합용이성)의 범위를 지나치게 넓게 해석함으로써 정보의 활용 가능성을 과다하게 제약하는 것이라는 비판이 제기되었다.[98] 이와 달리 상대적·주관적 접근법을 취할 경우 위 증권사의 관점에서는 수집된 정보가 개인정보에 해당하지 않는다고 판단될 여지가 클 것이다.

나. ‘약학정보원’ 사건과 비식별 조치의 기술적 적절성

‘약학정보원’ 사건의 개요는 다음과 같다.[99] 이 사건에서 약학정보원은 처방전 관련 정보를 환자나 의사의 동의 없이 의약 시장조사 업체에 제공하였다. 제공된 정보 중 의사의 성명 및 면허번호, 환자의 주민등록 번

98) 이동진, “개인정보 보호법 제18조 제2항 제4호, 비식별화, 비재산적 손해 - 이른바 약학정보원 사건을 계기로 -”, 정보법학 제21권 제3호, (2017), 267면 및 그 인용 문헌 참조.

99) 서울중앙지방법원 2017. 9. 11. 선고 2014가합508066 판결; 서울고등법원 2019. 5. 3. 선고 2017나2074963 판결; 이하에서는 이 사건에서 개인정보의 범위에 관한 판단만을 살핀다.

호, 성명, 생년월일, 성별 정보 등에 대해서는 시기 별로 구분되는 암호화 등의 비식별 조치가 취해졌다. 즉, 최초 정보 제공시(1기)에는 주민등록번호를 홀짝수 자리에 따라 영문자로 1:1로 대응시키는 규칙에 따라 영어 알파벳으로 치환하고, 임의의 알파벳 2자리의 잡음을 추가하는 방식의 암호화가 이루어졌다. 그 이후 일정 기간 동안(2기)에는 환자의 주민등록번호를 일방향 암호화 기법인 SHA-512 방식으로 암호화하였다. 마지막 기간 동안(3기)에는 주민등록번호가 아니라 성명, 생년월일, 성별로 환자를 특정하여 이를 일방향 암호화하였다. 한편 1기 암호화 방식을 2기와 3기 암호화 방식으로 변환하면서, 1기 암호화의 결과물과 2기 및 3기 암호화의 결과물로 대응시키는 매칭 테이블(matching table)을 제공하였다.

이 사건에서 서울중앙지방법원 민사 1심 판결은 1기 암호화 방식은 주민등록번호 숫자를 단순한 규칙에 따라 알파벳으로 치환한 것에 불과하여 쉽게 복호화가 이루어질 수 있으므로, 1기 암호화 방식으로 암호화된 정보는 재식별 가능성이 현저하여 여전히 개인정보에 해당한다고 판단하였다. 이에 대하여 2기 및 3기 암호화 방식은 기술적 수준이 높은 일방향 암호화 방식을 활용하여 수령자가 이를 복호화하는 것이 원칙적으로 불가능하므로, 적절한 비식별 조치가 이루어졌고, 개인정보에 해당하지 않는다고 판단하였다.

이에 비하여 서울고등법원 민사 2심 판결은 1기 암호화 방식이 적절한 비식별 조치에 해당하지 않는다는 1심 판결의 결론은 유지하였으나, 2기 및 3기 암호화 방식에 의해 비식별 조치가 이루어지더라도, 1심 암호화 방식의 결과물과의 대응 테이블을 제공받은 이상, 여전히 개인정보에 해당한다고 판단하였다. 즉, 2기 및 3기 암호화 방식으로 비식별 조치가 이루어지더라도 여전히 1기 암호화 결과물과의 대응 테이블과 결합하여 개인을 식별할 수 있으므로 여전히 개인정보에 해당한다고 본 것이다.

이 사건은 개인정보의 기술적 비식별 조치의 적절성을 다룬 최초의 사건이라는 점에서 상당한 의의가 있다. 이 사건에서 법원은 비식별 조치의 기술적 수준을 구분하여, (i) 단순한 치환 방식을 적용한 경우에는 손쉽

게 개인을 재식별해 낼 수 있으므로 여전히 개인정보에 해당하나, (ii) 기술적 수준이 높은 일방향 암호화가 이루어짐으로써 복호화가 원칙적으로 불가능한 수준에 이르면 더 이상 개인정보에 해당하지 않는다는 취지로 판단하였다. 이는 비식별 조치의 기술적 적절성 판단에 대한 일응의 기준을 수립한 것으로 평가할 수 있다.

하지만 이 판결의 비식별 조치의 적정성 판단에 있어 다음과 같은 몇 가지 한계가 있다.[100] 우선, 이 사건에서는 SHA-512 기법으로 일방향 암호화를 한 경우, 원칙적으로 복호화가 불가능하다고 판단하였으나, 이는 엄밀히 보자면 기술적으로 정확한 서술이라고 보기 어렵다. SHA-512 알고리즘을 1회 적용하는 것만으로는 무차별 대입 공격(brute force attack)이나 사전 공격(dictionary attack)에 취약하고, 솔팅(salting) 등의 추가적인 보안 조치를 취해야 한다.[101] 그런데 이 사건 판결문에는 2기 암호화 또는 3기 암호화에 있어 단지 SHA-512 기법이 적용되어 있다고만 기재되어 있으므로 이와 같은 추가적인 보안 조치가 이루어졌는지 알 수 없다. 만약 SHA-512 알고리즘에 따른 1회 해시만이 적용되었다면 기술적으로 볼 때 재식별이 불가능한 조치가 이루어졌다고 평가되기는 어려울 것이다.

나아가, 이 사건에서는 개인식별정보에 대한 비식별 조치만을 고려하였을 뿐, 식별가능정보를 포함한 데이터셋 전체를 기준으로 하여 개인식별가능성을 평가하지 않은 한계가 있다. 즉, 이 사건에서 의약 시장정보 업체에 제공된 정보에는 환자의 주민등록번호, 환자의 생년월일, 의료인의 성명 및 면허번호 등 개인식별정보뿐만 아니라, 질병 분류 기호, 처방의 약품의 명칭 및 분량, 처방전 발급 연월일 등 식별가능정보도 포함되어 있었다. 비록 개인식별정보에 대한 삭제나 대체가 이루어진다고 하더라도, 식별가능정보를 다른 정보와 결합하여 개인이 재식별할 가능성은 여전히 남아 있다. 그럼에도 불구하고 식별가능정보 부분에 관한 개인정보성 여부에 대한 판단이 이루어지지 않은 것은 충분한 판단이 이루어졌다

100) 이 사건은 개인정보 침해 행위에 대한 손해배상의 범위가 중요한 쟁점이 되었으나, 본 연구와 직접적으로 관련되어 있지 않으므로 이에 관하여는 다루지 않는다.

101) Finck and Pallas (2020) (각주 57) 참조

고 평가하기 어렵다.

2. 개정 개인정보 보호법 상의 해석

전술한 바와 같이 개정 개인정보 보호법에서 익명처리와 관련하여서는 중요한 두 가지 개정이 이루어졌다. 우선 제2조 제1호 나목에는 결합 용이성 판단에 있어, “다른 정보의 입수 가능성 등 개인을 알아보는 데 소요되는 시간, 비용, 기술 등을 합리적으로 고려”한다는 단서를 추가하였다. 또한 “시간·비용·기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아 볼 수 없는 정보”에는 개인정보 보호법을 적용하지 않는다는 익명정보에 관한 조항이 신설되었다(법 제58조의2).

위와 같이 신설되 조항의 문언은 GDPR 전문 제26조의 영향을 받은 것으로 평가될 수 있다. 즉, 제3장에서 살펴 본 바와 같이 GDPR 전문 제26조는 “어떤 개인이 식별가능한지를 판단하기 위해서는 컨트롤러 또는 다른 사람이 개인을 직접 또는 간접적으로 식별하기 위해 사용할 것으로 합리적으로 가능한(reasonably likely) 모든 수단, 가령 싱글 아웃(singling out) 등을 고려해야 한다. 어떤 수단이 개인을 식별하기 위해 합리적으로 사용 가능한 것인지를 확인하기 위해서는, 식별하기 위해 소요되는 비용과 시간 등 객관적인 요소를 모두 고려하고, 처리 당시 가용한 기술과 기술적 발전을 모두 고려하여야 한다”고 규정하고 있는바, 규정의 전반적인 취지나 내용이 개인정보 보호법과 상당히 유사하다. 즉, GDPR에서는 개인식별 가능성 판단에 있어서는 ‘합리적으로 가능한 모든 수단’을 고려할 것을 요구하고 있고, 우리 개정법에는 개인식별 가능성 판단에 있어 ‘시간·비용·기술 등을 합리적으로 고려’할 것을 요구하고 있다.[102]

따라서 개정 개인정보 보호법의 해석에 있어서는 제3장에서 검토한 GDPR의 해석이 중요한 지침이 될 것으로 판단된다. 무엇보다 GDPR에

102) 다만, 우리 개정법에서는 ‘모든’ 수단을 고려해야 한다는 문구가 포함되어 있지 않으나, 우리 법의 해석에 따르더라도 개인식별 가능성 판단에 있어서는 합리적으로 고려할 수 있는 모든 수단이 검토될 필요는 있을 것이다.

서 규정한 따른 상대적·주관적 접근법을 국내법으로 수용할 가능성이 열린 것으로 평가할 수 있을 것이다. 다만 국내 개인정보 보호법 법문 상으로는 '위험'에 대한 언급이 전혀 없으므로, GDPR 상의 위험기반 접근법을 도입한 것으로 평가할 수 있을지는 명확하지는 않다, 다만, 개인정보 보호법에 관한 가이드라인에는 '위험'에 따른 평가를 명시적으로 언급되어 있다는 점에 주목할 필요가 있다. 즉, 개인정보 비식별 조치 가이드라인(2016)는 비식별 조치의 적정성 판단을 위해서는 "재식별 위험이 없는지 여부 등을 종합적으로 평가"할 것을 요구하고 있다.[103] 나아가 가명정보 처리 가이드라인(2020)은 가명처리에 있어서도 '위험도 측정' 단계를 포함하고 있고, 측정된 위험도에 비례하여 가명처리의 수준을 결정하도록 하고 있다. 이러한 점에서 개정 개인정보 보호법에 따르더라도 개인정보의 비식별 조치에 있어 위험기반 접근법을 채택할 근거는 마련되어 있다고 볼 것이다.

이하에서는 앞서 본 '증권통' 사건 및 '약학정보원' 사건이 개정법 하에서 어떻게 달리 판단될 수 있을 것인지 검토함으로써 개정법의 의의를 구체적으로 살펴 본다.

우선 '증권통' 사건과 같은 판결은 개정법 하에서는 유지되기 어려울 것이다. 이 판결은 절대적·객관적 접근을 취하여, 다른 정보와의 결합 가능성 판단에 있어서는 "[다른 정보를] 구하기 쉬운지 어려운지와는 상관없이 해당 정보와 다른 정보가 특별한 어려움 없이 쉽게 결합하여 특정 개인을 알아볼 수 있게 되는 것"이라고 판시하였다. 그러나, 개정법 제2조 제1호 나목 단서는 결합 용이성 판단에 있어 "다른 정보의 입수 가능성" 등을 "합리적으로 고려"할 것을 명시적으로 요구하고 있다. 따라서 앞으로 이와 같은 사건에서는 해당 증권사가 IMEI나 USIM 일련번호와 결합하여 개인을 식별할 수 있는 정보를 입수할 수 있는지 여부가 고려되어야 한다. 만약 이러한 정보를 입수하기 어려운 경우라면, 증권사의 관점에서는 IMEI나 USIM 일련번호가 개인정보에 해당한다고 판단되지 않을 가능성이 클 것이다.

---

103) 개인정보 비식별 조치 가이드라인(2016), 제47-48면. 67면 등.

다음으로 '약학정보원' 사건의 경우, 개정법 하에서는 익명정보인지 여부가 아니라 가명처리와 관련된 규정이 쟁점이 될 가능성이 크다. 즉 이 사건에서 처방전 정보에 대해 이루어진 비식별 조치는 결국 주민등록번호 등의 개인식별정보를 암호화하여 가명으로 대체하는 것이다. 이러한 조치는 개정법 상의 "개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것"에 해당할 수 있다. 나아가 민사 판결에서 인정된 바와 같이 1기 암호화 방식처럼 단순한 마스킹 기법을 적용한 것은 적정한 가명처리 방법으로 인정되지 못하여 가명정보에 해당하지 않는다고 인정될 여지가 크고(즉, 추가 정보가 없더라도 개인을 재식별해 낼 수 있음), 이와 달리 2기나 3기와 같은 일방향 암호화 방식을 적용한 결과물은 가명정보에 해당한다고 인정될 여지가 클 것이다.

다음으로, 약학정보원이 정보주체의 동의 없이 의약 시장조사 업체에 처방전 관련 정보를 제공한 것은 개정 법의 제28조의2 제1항의 통계작성 또는 과학적 연구 목적을 위한 가명정보의 처리 특례가 적용될 수 있을 것인지 쟁점이 될 수 있다. 이와 관련하여 1심 민사 판결은 약학정보원이 위 정보를 제공한 것이 구 개인정보 보호법(2020. 2. 4. 법률 제16930호로 개정되기 전의 것) 제18조 제2항 제4호의 "통계 작성 또는 학술연구 등의 목적"에 해당한다고 판단하였다. 다만, 1기 암호화 결과물과의 매칭 테이블을 제공한 것은 개정 법의 제28조의2 제2항('특정 개인을 알아보기 위하여 사용될 수 있는 정보'의 제3자 제공 금지)를 위반한 것으로서 볼 여지가 있을 것이다.

제 3 절 가명처리와 가명정보

1. 가명처리의 정의와 '추가 정보'의 개념

전술한 바와 같이 개정 개인정보 보호법 제2조 1의2호는 '가명처리'를 "개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리"하는 것이라 정의하고 있다. 나아가 동조 1호 다.목은 개인식별정보와 식별가능정보를 "가명처리함으로써 원래의 상태로 복원하기 위한 추가 정보의 사용·결합 특정 개인을 알아볼 수 없는 정보"를 가명정보라 정의하고 있다. 또한 가명정보의 처리에 관한 특례를 신설하여 가명정보에 대해서 정보주체의 동의 없이 통계작성, 과학적 연구, 공익적 기록보존 등 목적으로 처리하는 것을 허용하고 있다(제28조의2 제1항). 따라서 개인정보처리자가 위 특례를 적용 받기 위해서는 개인정보에 대해 개정법에서 정한 '가명처리'를 수행하여 '가명정보'에 해당하도록 변형할 필요가 있다. 그렇다면 개인정보에 대해 어떠한 수준의 변형이 이루어진 경우 개정법에서 정한 '가명처리'에 해당하는 것인가?

가명정보의 정의에 부합하는지 여부는 결국 "원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는"지 여부에 따라 달라지게 된다. 그런데 개정 법은 '추가 정보'에 대해 따로 정의하고 있지 않으므로, "원래의 상태로 복원하기 위한 추가 정보"의 범위를 정하는 것이 중요한 과제가 된다.

그런데, 제2장 제5절에서 검토한 바와 같이 일반적으로 활용되는 가명처리 방법은 일방향 암호화와 쌍방향 암호화로 구분된다. 일방향 암호화가 이루어진 경우에는 암호화된 내용(즉 가명)만 보아서는 원칙적으로 원문으로 복호화하는 것이 불가능하다. 따라서 향후 데이터가 추가·수정될 경우를 대비하여 원래 정보와 암호화된 가명 간의 매핑 테이블(mapping table)을 생성하여 별도로 안전하게 보관하게 된다.[104] 다른 한편, 쌍방향 암호화가 이루어진 경우에는 암호키를 이용하여 복호화하여 원문으로 복호화 할 수 있다. 따라서 해당 암호키를 별도로 안전하게 보관하게 된다. 이러한 매핑 테이블(일방향 암호화의 경우)이나 암호키(쌍방향 암호화의 경우)를 통상 '가명처리 비밀'이라 칭한다. 이러한 가명처리 방법론을 고

104) 이러한 매핑 테이블을 두어야 할 필요성에 관하여는 제2장 제5절 참조

려한다면 해당 매핑 테이블과 암호키를 '원래의 상태로 복원하기 위한 추가 정보'로 해석하는 것이 자연스럽다. 가명정보 처리 가이드라인(2020)도 '추가 정보'란 "개인정보의 전부 또는 일부를 대체하는 데 이용된 수단이나 방식(알고리즘 등), 가명정보와의 비교·대조 등을 통해 삭제 또는 대체된 개인정보를 복원할 수 있는 정보(매핑 테이블 정보, 가명처리에 사용된 개인정보 등) 등"이라고 규정하여 '추가 정보'란 '가명처리 비밀'을 의미하는 것이라 해석하고 있다.[105]

그러나 이와는 달리 매핑 테이블이나 암호키 이외의 '다른 정보'를 이용하더라도 개인을 재식별할 수 있는 경우도 존재한다. 즉, 가명처리는 익명처리와 구분되는 것으로 개인식별가능성이 여전히 남아 있는 상태이므로, 이를 공개 데이터, 일반적 배경지식 또는 가명정보처리자가 보유한 다른 데이터와 결합하면 개인을 식별할 수 있다. 따라서 이러한 개인을 재식별해 낼 수 있는 '다른 정보'도 "원래의 상태로 복원하기 위한 추가 정보"에 해당한다고 볼 여지도 있다. 이처럼 '추가 정보'를 '가명처리 비밀'에 국한시키지 않고, 가명정보로부터 개인을 재식별하기 위해 사용될 수 있는 모든 정보가 '추가 정보'에 해당한다는 해석도 가능한 것이다.

이처럼 '추가 정보'의 범위의 해석에 따라 개정법 상의 '가명정보'의 범위도 달라지게 된다. 본 연구에서는 편의상 '추가 정보'를 '가명처리 비밀'로 한정하는 경우를 '협의의 추가 정보'라고 하고, '추가 정보'를 개인의 재식별하기 위한 모든 정보라 보는 경우를 '광의의 추가 정보'라 부르겠다.

2. '가명정보'의 해석

이상에서는 가명처리의 정의에 있어 핵심적 지표인 '추가 정보'의 범위에 관하여 두 가지 해석이 가능하다는 점을 살펴 보았다. 이러한 해석상의 차이가 '가명정보'의 범위에 있어 어떠한 차이를 가져오는지 살핀다.

개인정보에 대해 변형이 이루어졌을 경우, 해당 변형된 결과물이 가명정

105) 가명정보 처리 가이드라인(2020) 7-8면.

보에 해당하는지 여부를 판단하는 기준은 “원래의 상태로 복원하기 위한 추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는”지 여부이다. 위 규정은 이중 부정 형태로 되어 있어 해석이 쉽지 않다. 간단한 방법은 ‘추가 정보’가 없는 상황을 전제하고, 이러한 상황에서 개인식별 가능성을 판단하는 것이다. 즉, ‘추가 정보’가 없는 상황을 가정했을 때 해당 정보로부터 개인을 재식별해 낼 수 있다면, 해당 정보는 가명정보에 해당하지 않는다(즉, ‘추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 경우’에 해당하지 않는다). 이 경우 해당 정보는 개정법 제2조 제1호 나.목에서 정한 식별가능정보에 해당하고, 개정법의 가명정보 특례 적용의 대상이 되지 못한다. 반대로 ‘추가 정보’가 없는 상황을 가정했을 때 해당 정보로부터 개인을 재식별해 낼 수 없다면, 해당 정보는 가명정보에 해당한다(즉, ‘추가 정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 경우’에 해당한다). 그렇다면 ‘추가 정보’의 개념에 따라 가명정보의 범위는 어떻게 달라지는가?

앞서 본 바와 같이 ‘추가 정보’가 ‘가명처리 비밀’만을 의미한다고 볼 경우(협의의 가명정보), 가명정보처리자가 보유한 다른 데이터, 일반적 배경 지식, 공개 데이터 등은 여전히 보유하고 있는 가정적 상태에서 가명정보처리자가 개인을 재식별 할 수 있는지를 판단하여야 한다. 이러한 가정적 상태에서도 특정 개인을 알아볼 수 없게 하려면 (익명처리에 가까운) 강한 가명처리가 필요하게 된다. 즉, ‘추가 정보’가 ‘가명처리 비밀’만을 의미한다고 보면, 강한 가명처리가 이루어져야 하고, 개정법 상의 가명정보의 개념은 좁아지게 된다(본 연구에서는 논의의 편의상 ‘**강한 가명처리설**’이라 부른다).

이와는 달리 ‘추가 정보’를 가명정보로부터 개인을 재식별하기 위해 사용될 수 있는 모든 정보를 의미한다고 볼 경우(광의의 가명정보), 가명정보처리자가 보유한 다른 데이터, 일반적 배경지식, 공개 데이터 등도 없는 가정적 상태에서 가명정보처리자가 개인을 재식별할 수 있는지를 판단하여야 한다. 이처럼 다른 정보가 전혀 없는 상황을 가정하였으므로, 이러한 상황에서 개인을 알아볼 수 없게 하려면 가명정보 내에 개인식별정보

가 포함되어 있지 않으면 충분하다. 요컨대 '추가 정보'를 가명정보로부터 개인을 재식별하기 위해 사용될 수 있는 모든 정보를 의미한다고 보면, 약한 가명처리로도 충분하게 되고, 개정법 상의 가명정보의 개념은 넓어지게 된다(본 연구에서는 논의의 편의상 '**약한 가명처리설**'이라 부른다). 이하에서는 구체적인 가상의 예시 사례를 통해 위 두 가지 해석 중 어떠한 해석이 개정법에 대한 적절한 해석인지 고찰한다.

3. 가상 사례를 통한 검토

교육학 연구자가 대학생의 학업 성취도에 영향을 미치는 요인을 분석하고자 대학에 데이터를 요청한 가상적 상황을 가정한다. 대학은 개정 개인정보 보호법에 따라 '과학적 연구' 목적을 위하여 보유한 학생 및 학점 데이터를 가명처리하여 연구자에게 제공하고자 한다. 위 대학은 소속 학생에 관한 기본정보를 다음과 같은 형태로 보관하고 있다(이하 "기본정보 테이블").

학번	성명	학과	성별	이메일	우편번호	상세주소	전화번호	수능점수	...
18011001	김갑동	전자	M	kkd@gmail.com	01234	우리아파트 A동 101호	010-1234-5678	400	...
18011002	이을서	전자	F	les@daum.net	02345	나라아파트 B동 202호	010-1234-1234	392	...
18012001	박병남	전산	M	pbn@gmail.com	24088	웰컴로 1	010-4567-4567	384	...
18012002	최정북	전산	M	cjb@naver.com	03456	세종로 2	010-7890-7890	395	...
...	...			...	...	...	...	..	...

<표 3> 학생 기본정보 테이블

각 학생들의 수강과목과 학점정보는 다음 테이블(이하 "학점정보 테이블")에 저장되어 있다고 한다. 이 테이블은 학번 정보를 포함하고 있으므로 기본정보 테이블과 연결(join)시킴으로써 해당 학생의 학점정보를

알 수 있다.

학번	수강연도	수강학기	수강과목	학점
18011001	2018	1	CS101	A0
18011001	2018	1	EE101	A+
18011001	2018	2	PHY101	B-
18011002	2018	1	CS101	A+
18011002	2018	1	EE101	B0
...	...	...	...	...

<표 4> 학점정보 테이블

위 기본정보 테이블과 학점정보 테이블에 어떠한 변형을 가한 경우에 개정 개인정보 보호법상의 '가명정보'에 해당할 것인가? 우선 기본정보 테이블에는 다수의 개인식별정보가 포함되어 있다. 성명, 상세주소, 전화번호, 이메일 주소 등이다. 연구자가 조사대상 학생에게 연락을 취하거나 추적 조사를 수행할 것이 아닌 이상 위 정보는 삭제하면 된다(만약 이러한 방식의 조사가 예정되어 있다면 학생의 동의를 얻어야 할 것이다). 이는 개정법상의 '가명처리' 중 "개인정보의 일부를 삭제"하는 작업에 해당할 것이다.

다음으로, 학생의 학번 역시 개인에게 고유하게 부여된 값이므로 개인식별정보에 해당한다. 그러나 학번을 삭제할 경우 학점정보 테이블과 연결짓지 못하는 문제가 발생한다. 따라서 학번을 다른 값(즉, 가명)으로 대체하여, 기본정보 테이블과 학점정보 테이블 양쪽 모두에서 값을 변경해 주어야 한다. 이는 개정법상의 '가명처리' 중 "개인정보의 일부 또는 전부를 대체"하는 작업에 해당할 것이다. 예컨대, 학번에 대해 단방향 암호화 기법[106]을 적용하면 다음과 같이 원래 학번과 가명 학번 간의 대응표(mapping table)를 생성할 수 있다. 이러한 대응표가 '가명처리 비밀'에 해당한다.

106) 아래 예에서는 SHA-256 해시함수를 적용하였다.

원래 학번	암호화된 값(가명)
18011001	AE5C...(중략)...78D2
18011002	50A9...(중략)...DF07
18012001	5C6B...(중략)...60D9
18012002	CFB5...(중략)...A058

<표 5> 가명학번 대응표(가명처리 비밀)

위와 같이 개인정보의 삭제 또는 대체를 통한 가명처리가 이루어진 결과물은 다음과 같다(이하 '가명처리 결과물'이라 한다).

가명 학번	학과	성별	우편번호	수능점수	...
AE5C...(중략)...78D2	전자	M	01234	400	...
50A9...(중략)...DF07	전자	F	02345	392	...
5C6B...(중략)...60D9	전산	M	24088	384	...
CFB5...(중략)...A058	전산	M	03456	395	...
...			...	..	...

학번	수강연도	수강학기	수강과목	학점
AE5C...(중략)...78D2	2018	1	CS101	A0
AE5C...(중략)...78D2	2018	1	EE101	A+
AE5C...(중략)...78D2	2018	2	PHY101	B-
50A9...(중략)...DF07	2018	1	CS101	A+
50A9...(중략)...DF07	2018	1	EE101	B0
...	...	...	...	...

<표 6> 가명처리 결과물

위 가명처리 결과물은 개인식별정보는 모두 삭제되었으므로, '약한 가명처리'가 수행된 것이다. 그럼에도 가명처리 결과물이 여전히 개인정보임

은 분명하다. 가령 가명정보를 수령할 연구자 중에서 위 대학 졸업생이 있다면, 자신이 다녔던 학과의 다른 학생에 관한 상당한 배경 지식을 가지고 있을 수 있다. 이를 이용하면 위 가명처리 결과물을 이용하여 개인을 식별해 낼 가능성이 크다.

재식별 가능성이 있는 또 다른 문제 상황을 다음과 같이 가정해 볼 수도 있다. 예컨대 기본정보 테이블의 첫 번째 행의 김갑동 학생(가명학번: AE5C...(중략)...78D2)은 수능점수가 만점이다. 이러한 사실은 언론 보도 등을 통해 널리 알려져 있을 수 있다. 만약 해당 대학 전자과의 수능만점자는 김갑동 학생이 유일하다면, 위 가명처리 결과물을 통해 김갑동 학생은 쉽게 재식별이 가능하다. 이처럼 개인식별정보가 삭제 또는 대체되더라도 '특이정보'가 포함된 경우 재식별이 용이하게 가능하다. 이러한 재식별을 막기 위해서는 '특이정보'까지 삭제하는 '강한 가명처리'가 필요하다. 그렇다면 개정 개인정보 보호법상의 '가명정보'에 해당하기 위해서는 약한 가명처리가 족한 것인지, '강한 가명처리'가 이루어져야 할 것인지 쟁점이 된다.

가. 약한 가명처리설에 의할 경우

앞서 설명한 약한 가명처리설에 의할 경우에는 개인식별정보를 삭제·대체하는 가명처리만이 이루어지더라도 개정 개인정보보호법 상의 '가명정보'에 해당한다고 볼 것이다. 이에 따르면, 앞선 학생 정보 사례에서의 가명처리 결과물은 가명정보에 해당하게 된다. 이러한 견해는 다음과 같은 근거로 지지될 수 있다.

우선, 식별가능정보와의 균형이다. 즉, 개정법 제2조 제1호 나목은 '다른 정보'와 쉽게 결합하여 개인을 알아볼 수 있는 정보를 개인정보에 포함하고 있고, 이 경우 결합 용이성은 합리적 고려의 대상이 된다(동조 단서). 이에 비해 가명정보의 정의 조항에는 결합 용이성 판단에 있어 이와 같은 합리적 고려의 제한이 존재하지 않는다. 따라서 강한 가명처리설에 의하면 오히려 가명정보에 대해 더욱 엄격한 개인식별 가능성 판단

이 이루어지게 되어, 나.목의 개인정보의 범위와 균형성이 맞지 않게 되는 문제가 있다.[107] 또한 강한 가명처리와 익명처리의 구분이 어렵다는 문제도 있다. 규범적으로나 실무적으로나 강한 가명처리는 익명처리와 명확하기 구분하기 어려우므로, 결국 가명정보 처리의 특례 적용을 받기 위해서는 익명처리가 이루어질 것이 요구되는 상황이 발생할 수 있다. 그 결과 신설된 특례 규정의 취지가 몰각될 수 있다. 더 나아가 개인정보처리자가 직접식별자 필드를 삭제하거나 대체하는 방법은 손쉽게 적용할 수 있지만, 추가적인 비식별 조치를 취할 것까지 요구할 경우 가명처리 방법에 대한 상당한 혼란이 발생하게 될 우려가 있다. 마지막으로, 개인정보보호법의 개정 취지는 가명처리를 통한 데이터 활용을 장려하는 것이므로 가명정보의 범위를 넓게 해석할 필요도 있다. 이러한 점을 고려하면 약한 가명처리(즉, 개인식별정보만을 삭제하는 것)을 취하더라도 가명정보에 해당한다고 해석하다고 주장될 여지가 있다.

나. 강한 가명처리설에 의할 경우

이에 비하여, 앞서 설명한 강한 가명처리설에 의할 경우에는 개인식별정보를 삭제·대체하는 것에서 더 나아가 가명처리 결과물에서 개인이 식별될 수 있는 정보를 삭제하거나 범주화하는 등의 추가적인 조치를 수행하여야 한다. 그 결과 앞선 사례에서의 가명처리 결과물은 가명정보에 해당하지 않는다. 이러한 견해는 다음과 같은 근거로 지지될 수 있다.

우선, 개정법 제2조 제1호 다목의 '추가 정보'는 "원래의 상태로 복원하기 위한" 것이어야 하므로, 배경 지식, 공개 데이터 또는 다른 보조적 정보까지 '추가 정보'의 범위에 포함시키는 것은 문언의 한계를 넘는 해석이다. 즉, '추가 정보'는 '가명처리 비밀'에 국한되는 것으로 해석되어야 한다. 또한 개인정보보호법의 목적은 개인정보 자기결정권을 실효적으로 보장하는 것에 있는바, 가명정보의 범위를 지나치게 확대할 경우 개인정보 보호법의 취약점으로 기능할 우려가 크다. 또한 가명정보의 범위를

107) 고학수·백대열·구본효·정종구·김은수, "개정 개인정보 보호법상 가명정보의 개념 및 가명처리에 관하여", 서울대학교 인공지능정책 이니셔티브 이슈페이퍼 2020-1, 17면.

넓게 해석할 경우 개인정보처리자가 개인식별정보만을 개인정보를 사실상 마음대로 활용할 수 있게 되는 결과가 초래될 수 있다는 문제도 발생할 수 있다.
강한 가명처리설의 가장 중요한 근거는 개정법상 가명정보에 대해 정보주체의 권리가 제한된다는 점이다. 즉, 개정법 제28조의7에 따르면 가명정보에 대해서는 제20조(정보주체 이외로부터 수집한 개인정보의 수집출처 등 고지), 제21조(개인정보의 파기), 제27조(영업양도 등에 따른 개인정보의 이전 제한), 제34조 제1항(개인정보의 유출 통지), 제35-37조(개인정보의 열람, 정정·삭제, 처리정지) 등이 적용되지 않는다. 따라서 가명정보의 범위를 제한적으로 해석할 필요가 있다는 것이다.

다. 가능한 절충설

한편, 위 양자를 절충한 견해도 고려해 볼 수 있다. 이는 개인정보처리자가 가명정보를 내부적으로 목적외로 이용하거나, 내부의 다른 부서 등에 제공하는 경우에는 약한 가명처리 허용하되, 제3자 제공 시 강한 가명처리가 필요하다고 보는 견해이다. 이는 개정법 제28조의2 제2항을 고려한 것이다. 즉, 위 조항은 가명정보를 제3자 제공하는 경우 “특정 개인을 알아보기 위하여 사용될 수 있는 정보” 포함하는 것이 금지되어 있다. 따라서 내부적 이용과 제3자 제공 시 ‘가명정보’의 범위를 달리 보아야 한다는 해석론도 고려해 볼 수 있다.
이러한 견해는 개정법의 가명정보의 활용 범위를 넓힌다는 목적은 내부적으로 약한 가명처리를 허용함으로써 달성하고, 개인정보 침해 위험을 통제하기 위하여 외부적으로 강한 가명처리를 요구하는 것이 바람직하다는 근거로 뒷받침될 수 있다. 현실적으로는 내부 이용의 경우 상대적으로 위험도가 낮으므로 약한 가명처리를 허용할 필요가 있다고 생각할 수 있다.

라. 가명정보 처리 가이드라인(2020)의 태도

개인정보보호위원회의 가명정보 처리 가이드라인(2020)은 앞서 설명한 견해 중 일응 강한 가명처리설에 따른 해석을 취하고 있는 것으로 보인다. 즉, 위 가이드라인에 따르면 가명처리 절차에 있어 개별 항목별 위험도를 분석하는 단계가 포함되어 있고, 이러한 위험도 분석 단계에 있어 처리(제공)환경과 정보의 규모(레코드나 항목의 크기), 구체성(정보의 정확성 수준) 등을 고려하여야 하는데, 개인식별 가능성이 높은 정보의 예시로서 개인식별정보뿐만 아니라 식별가능정보(이는 특이정보를 포함함 것이다)를 들고 있다.[108] 따라서 개인식별정보를 삭제 또는 대체하는 약한 가명처리만으로는 부족하고 식별가능정보에 대한 추가적인 변형조치를 요구하고 있는 것이다.

위 가이드라인이 제시하는 구체적인 예시 사례에 있어서도 단순히 식별정보를 삭제하는 것만으로는 충분하지 못하는 점을 지적하고 있다. 위 가이드라인은 호텔이 고객의 투숙 정보를 데이터 분석회사에 제공하는 경우, 회원번호와 이름을 가명처리하고 나이, 성별, 등급, 예약방법, 객실정보, 체크인, 체크아웃, 서비스 이용금액 정보를 제공하는 것은 '잘못된 제3자 제공 사례'라 언급하고 있다.[109] 그 이유로 데이터 분석회사의 분석 담당자가 "특정일에 최고급 객실을 이용한 내용을 분석과정에서 인지할 수 있으며, 기존 업무(온라인 SNS정보 수집)를 수행하며 공개된 정보(예 : 개인이 SNS에 올리는 정보, 여행후기 등)를 통해 특정 개인을 식별할 가능성이 있"기 때문이라고 제시하고 있다. 나아가 그 해결방안으로 제공하는 가명정보에 포함된 특이정보(최고급 객실)를 삭제하거나 가명처리 등을 수행해야 한다고 제시한다. 요컨대, 위 가이드라인에 따르면 '강한 가명처리'가 이루어져야 가명정보에 해당한다고 볼 수 있다.

마. 검토

생각건대, 개정법의 해석에 있어서는 GDPR의 구조 및 내용을 참조할 필요가 있다고 본다. 즉, GDPR은 '가명처리'에 대한 상당한 인센티브를

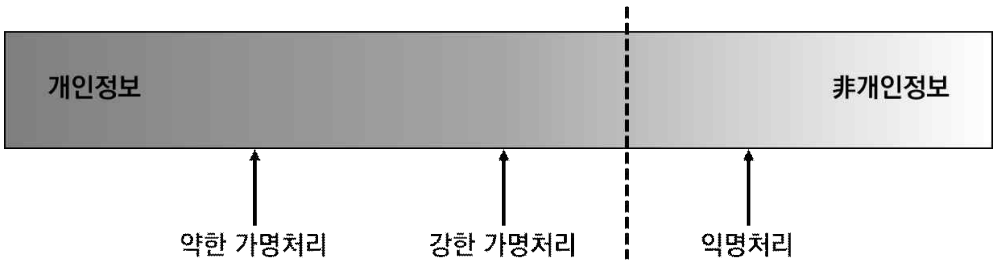
108) 가명정보 처리 가이드라인(2020) 17면.
109) 위의 글. 16면.

부여하고 있으나, 이는 대체로 GDPR 상의 각종 위험 평가시 '가명처리'에 의해 위험을 낮추는 효과가 개별 조항에서 비례적으로 고려되는 형태로 작동한다. 그런데 GDPR 제11조는 '식별을 요하지 않는 개인정보의 처리'에 대해서는 정보주체의 권리를 배제하면서, '가명처리'라는 표현을 사용하지 않고, "컨트롤러가 정보주체를 식별할 수 없음을 입증"해야 한다는 별도의 요건을 부가하고 있다.

이에 비해 개정 개인정보 보호법은 양자를 구분하지 않고, '가명정보'에 대해서는 정보주체의 권리를 배제하는 형태의 입법을 하였으므로, GDPR의 입법례와 비교하여 본다면 우리 법 상의 '가명정보'의 범위를 GDPR 제11조에서 정한 "컨트롤러가 정보주체를 식별할 수 없음을 입증"할 수 있는 수준으로 해석하는 것이 불가피한 것으로 보인다. 이렇게 본다면 가명정보 처리 가이드라인(2020)과 같이 강한 가명처리설을 취하는 것이 자연스러운 해석이 된다.

제 4 절 소결

이상에서는 개정 개인정보 보호법상의 익명처리/익명정보와 가명처리/가명정보의 해석상 쟁점에 관하여 살펴보았다. 이상의 논의를 정리하면 다음 그림과 같다.



<그림 10> 비식별 조치와 개인정보 해당 여부

우선 익명처리는 개인식별 가능성이 희박(remote)해 지도록 개인정보를 변형하는 것이다. 이는 데이터를 공중에 공개하거나 재식별 공격에 노출될 우려가 있는 상황을 염두에 둔 조치이다. Netflix Prize 데이터 재식별 사례에서와 같이 공개 데이터에 대해서는 재식별 공격자에 의해 소수의 정보주체라도 재식별할 위험성이 존재한다. 이를 막기 위해서는 정립된 방법론과 기준(k-익명성 등)을 통해 개인식별 가능성을 평가해야 한다. 이러한 평가 결과 적정한 익명처리가 이루어진 경우에는 더 이상 개인정보 보호법의 적용 대상에 해당하지 않게 된다.

익명처리보다 완화된 조치는 강한 가명처리이다. 이는 주로 개인정보를 동의받은 목적 이외의 범위에서 이용하거나, 연구자 등 제3자에게 제공하는 상황을 염두에 둔 것이다. 강한 가명처리는 해당 정보를 처리하는 자가 개인을 재식별해 낼 수 없도록 만든다. 이로써 강한 가명처리가 이루어진 경우 정보주체의 권리를 배제하는 것이 정당화될 수 있다. 또한 강한 가명처리를 통해 해당 정보의 처리자에 의한 재식별 가능성이 제거된 상황이라는 점이 입증된다면, 익명처리에 이르지 않더라도 해당 정보의 목적외 이용이나 제3자 제공이 허용된다. 하지만 강한 가명처리가 이루어지더라도 여전히 '일부'의 정보주체에 대해서는 재식별이 이루어질 가능성은 상존한다. 따라서 강한 가명처리된 정보도 여전히 개인정보에 해당하고, 각종 기술적 및 관리적 보안 조치 의무가 적용된다.

그렇다면 '익명처리'와 '강한 가명처리'는 실제 적용상 어떠한 차이가 있는 것인가? '익명처리'는 주로 외부적 공격자의 존재를 염두에 두고 데이터셋에 포함된 일부의 정보주체에 대해서도 재식별이 발생하지 않도록 위험의 수준을 낮추는 과정이다. 이에 비해 '강한 가명처리'는 해당 정보의 처리자가 정보주체를 재식별해 낼 수 없도록 하는 것이므로, 체계적(systematic)이고 확실한(reliable) 수단에 의해 정보주체를 식별할 수 없도록 하는 것을 주된 목적으로 한다.[110] 실무상으로는 '익명처리'는 주로 k-익명성 등과 같은 프라이버시 모델에 따른 비식별 조치를 적용하는

110) Hintze (2018) (각주 76) p. 93.

것으로 이해될 수 있고, '강한 가명처리'는 개인식별정보와 특이정보를 삭제·대체하고, 나아가 가명처리 비밀을 삭제하거나 이에 대한 접근이 불가능하도록 하는 방식으로 변형하는 작업이 주가 될 것이다.

강한 가명처리보다 완화된 조치는 약한 가명처리이다. 이는 개인정보처리자의 개인정보 보호 수준을 높이기 위해 채택할 수 있는 조치이다. 이는 개인식별정보를 삭제하거나 암호화하는 방식으로 이루어지므로, 익명처리나 강한 가명처리에 비해 상대적으로 용이하게 구현될 수 있다. 다만, 해당 정보의 처리자는 '추가 정보'(즉 가명처리 비밀)을 이용하여 개인을 손쉽게 재식별해 낼 수 있으므로 여전히 개인정보 보호 규제의 적용을 받는다. 제3장에서 검토한 바와 같이 GDPR은 약한 가명처리에 대해서도 다양한 인센티브를 제공함으로써 컨트롤러가 약한 가명처리를 도입하는 것을 장려하고 있음에 비해, 국내 개인정보 보호법상으로는 약한 가명처리에 대한 충분한 인센티브가 마련되어 있지 못하다는 한계가 있다.

제 5 장 종단 데이터(longitudinal data)의 문제

제 1 절 종단 데이터의 개념

본 장에서는 이상에서 살펴본 비식별 조치의 구체적 적용 사례로서 종단 데이터(longitudinal data)의 문제를 검토한다. 종단 데이터란 여러 시점에서 일정한 조사대상에 관해 수집된 관찰 또는 사건에 대한 정보를 포함하는 데이터를 의미한다.[111] 이는 특정한 시점에서 조사대상에 관한 정보를 수집한 횡단면 데이터(cross-sectional data)와 구분된다. 대표적인 종단 데이터로는 시계열 데이터(time-series data) 또는 패널 데이터(panel data)가 있다.[112]

개인정보처리자는 여러 시점에 걸쳐 정보주체에 관한 정보를 수집할수록 정보주체에 관해 더 가치 있는 사실을 파악할 수 있게 된다. 따라서 개인정보처리자는 정보주체에 관한 정보를 수집하여 종단 데이터 형태로 보관하게 되는 경우가 흔하다. 예컨대 종단 데이터의 예로는 환자의 입원 및 치료 내역, 보험청구 내역, 온라인 쇼핑몰 물품 구매 이력, 신용거래 내역, 학생 성적 데이터 등이 있다.

종단 데이터는 학술적 연구의 목적에서도 횡단면 데이터와 비교했을 때 더 많은 정보를 제공해 준다.[113] 즉, 관찰하고자 하는 변수를 여러 시점

---

111) Ruiz, Nicolas, "A General Framework and Metrics for Longitudinal Data Anonymization", International Conference on Privacy in Statistical Databases, Springer, (2018), p. 216.

112) 이명진, "사회과학도가 알아야 할 자료 II - 시계열 조사 (Times Series Survey)", 한국사회과학자료원 엣세이, <https://kossda.tistory.com/8?category=330415> (2020. 11. 19. 접속)

113) Ruiz (2018) (각주 111), p. 217.

에 걸쳐 반복적으로 수집함으로써 시간에 따른 변화를 파악할 수 있게 되는 것이다. 예를 들어, 실업 문제에 관해 사회과학적 분석을 수행하는 경우를 상정해 보자. 이때 횡단면 데이터는 단지 특정 시점에서 전체 인구 중 실업자가 되는 비율이나 그 실업자의 특성을 파악할 수 있을 뿐이다. 그러나 종단 데이터를 활용하면 특정 실업자가 어느 정도의 비율로 재취업이 되거나 얼마나 오랫동안 실업 상태가 유지되는지 파악할 수 있게 된다. 더욱이 종단 데이터는 인과관계 분석에 있어 중요한 가치가 있다. 가령 어떤 정책의 인과적 효과를 밝히고자 하는 경우에는 해당 정책의 시행 이전과 시행 이후의 상당한 시간에 걸친 종단데이터를 수집하여 분석하게 된다.

제 2 절 종단데이터의 비식별 조치

1. 종단 데이터 비식별 조치의 어려움

종단 데이터는 횡단면 데이터에 비해 비식별화가 어렵다.[114] 종단 데이터에는 횡단면 데이터에 비해 정보주체에 관해 상당히 많은 정보가 포함되어 있을 뿐만 아니라, 개별 레코드 사이에 시간적 상관관계가 있다는 문제가 있기 때문이다. 앞서 예로 든 실업자에 관한 종단 데이터에 있어, 특정인을 재식별하려는 공격자는 누군가가 특정한 시점에 실업 상태에서 채용 상태로 전환되었다는 지식을 보유하고 있을 수 있다. 따라서 시간에 따른 고용상태의 변화 자체가 일종의 준식별자로서 공격자에 의해 추가적인 배경지식으로 이용될 수 있다.

다른 예로 급여 액수를 포함하고 있는 대기업의 직원 인사정보에 관한 횡단 데이터를 고려해 보자. 만약 위 데이터에 비식별 조치를 취하여 급

114) 위의 글, p. 218

여를 1,000만원 단위로 일반화하였다고 하자. 그 결과, 공격자가 특정 직원의 급여 액수를 알고 있더라도 곧바로 해당 직원이 누구인지 특정해 내기 어렵게 된다. 데이터를 이렇게 변형하는 것이 비식별 조치의 주된 목표이다. 하지만 만약 공격자가 특정 기간 사이에 어떤 직원의 급여가 크게 증가했다는 점을 알고 있으면 어떠한가? 이 경우 공격자는 해당 기간 동안 급여가 급격히 증가한 직원을 찾는 방식으로 해당 직원이 누구인지 재식별해 낼 위험이 발생한다.

위 사례에서 직원의 급여 정보는 간접식별자에 해당한다. 이처럼 전통적인 비식별조치 방법론을 활용하여 간접식별자 정보를 변형하더라도 횡단 데이터에 있어서 간접식별자의 시간적 상관관계를 배경 지식으로 이용하는 공격을 막기 어렵다.

2. k-익명성 기법 적용의 문제점

종단 데이터에 대한 비식별 조치의 실무상 주요한 어려움은 k-익명성 기법을 적용하기 쉽지 않다는 점이다. El Emam and Arbuckle (2013)에 소개된 간단한 사례를 통해 이를 살핀다.[115]

예를 들어 <그림 9>와 같은 환자 내원 이력에 관한 종단데이터를 고려해 보자. (a) 테이블에는 환자의 기본정보인 환자ID(PID), 출생일(DoB), 성별(Gender)가 포함되어 있다. (b) 테이블에는 환자ID 별로 내원일(Visit Date)와 내원 당시의 거주지 우편번호(Postcal Code)가 포함되어 있다. 예컨대 20000. 8. 7.생 여성인 PID 10번 환자는 2009/01/01, 2009/01/14, 2009/04/18 3회 내원한 것이고, 3회차 내원시에는 거주지의 우편번호가 K7G2C3에서 K7G2C4로 변경되었다는 정보가 포함되어 있다.

115) El Emam and Arbuckle (2013) (각주 15) pp 82-86.

(a)

PID	DoB	Gender
10	2000/08/07	M
11	1975/01/01	F
12	1975/06/24	F
13	1975/08/17	F
14	1975/18/09	F
15	2000/02/12	M

(b)

PID	Visit Date	Postal Code
10	2009/01/01	K7G2C3
10	2009/01/14	K7G2C3
10	2009/04/18	K7G2C4
11	2009/01/01	K1V7E6
11	2009/01/20	K1V7E8
11	2009/02/22	K1V7E8
12	2008/12/15	K1Y4L5
12	2009/01/20	K1V7E8
13	2008/12/22	K1Z5H9
14	2009/01/13	K1Y4L5
15	2009/04/20	K7G2G5

<그림 11> 환자 내원 이력 데이터 예제

이러한 종단 데이터에 대해 k-익명성 기준을 적용해 본다. <그림 10>은 출생일, 성별, 내원일, 우편번호를 준 식별자로 하여 k=2인 k-익명성을 적용한 예이다. k-익명성 요건을 충족시키기 위하여 (i) 출생일은 연단위로, (ii) 내원일은 월단위로, (iii) 우편번호는 앞 3자리까지로 각각 일반화되었다. 또한 (iv) PID 11번 환자의 2009/01 내원시 우편번호, (v) PID 11번 환자의 2009/02 내원이력, (vi) PID 14번 환자의 2009/01 내원시 우편번호는 각각 삭제되었다. 이러한 결과물은 익명정보로 평가할 수 있는가?

(c)

PID	DoB	Gender	Visit Date	Postal Code
10	2000	M	2009/01	K7G
10	2000	M	2009/01	K7G
10	2000	M	2009/04	K7G
11	1975	F	2009/01	K1V
11	1975	F	2009/01	---
11	---	---	---	---
12	1975	F	2008/12	K1Y
12	1975	F	2009/01	K1V
13	1975	F	2008/12	K1Y
14	1975	F	2009/01	---
15	2000	M	2009/04	K7G

<그림 12> k-익명성 적용 예시

만약 재식별을 시도하는 공격자가 어떤 여성 환자가 병원에 2008년 12월 및 2009년 11월에 내원했다는 정보가 있다고 생각해 보자(해당 환자를 치료한 의사나 그 환자의 지인은 이러한 정보를 보유하고 있을 것이다). 그렇다면 해당 환자가 위 데이터셋에서 PID 12로 유일하게 식별될 수 있다(즉, '싱글 아웃'이 이루어진다). 이는 k-익명성 기준을 적용하더라도 공격자에게 여러 시점에 관한 정보가 있다면 이를 막지 못하기 때문에 발생하는 문제이다.

위와 같은 공격에 대한 취약성까지 막고자 한다면, 위 <그림 9>의 데이터를 아래 <그림 11>과 같이 개별 방문 시점과 개별 우편번호 값 자체를 간접식별자로 취급해야 한다. 즉, 특정 방문시점에 특정 우편번호에 거주하는 환자의 방문 기록이 k개 이상일 뿐만 아니라, 동일한 방문 이력까지도 k개 이상이 되도록 하여야 한다. 이렇게 변형하여야 공격자가 2008년 12월 및 2009년 11월에 내원했다는 정보를 보유하고 있더라도 환자를 유일하게 식별해 낼 수 없게 되기 때문이다. 그러나 <그림 11>과 같은 테이블에 k-익명성 기준을 적용한다면 매우 많은 변수에 대해 동일한 기록이 k개 이상이 존재해야 한다. 그 결과 거의 대부분의 값이 삭제되고, 결국은 사실상 무용한 데이터로 변형되고 말 것이다.[116]

(d)

		Visit Date									Postal Code							
DoB	Sex	2009/01/01	2009/14/01	2009/04/18	2009/01/20	2009/02/22	2008/12/15	2008/12/22	2009/01/12	2009/04/20	K7G2C3	K7G2C4	K1V7E6	K1V7E8	K1Y4L5	K1V7E8	K1Z5H9	K7G2G5
2000/08/07	M	X	X	X							X	X						
1975/01/01	F	X			X	X							X	X				
1975/06/24	F				X		X								X	X		
1975/08/17	F							X									X	
1975/09/18	F								X						X			
2000/02/12	M									X								X

<그림 13> k-익명성 적용을 위해 변환된 데이터

116) 위의 글, p. 86.

3. 날짜 정보 변경의 필요성

종단 데이터에 대한 비식별 조치의 다른 어려움은 날짜 정보를 통한 재식별 위험이 크다는 점이다.[117] 의료 데이터를 예로 들면, 해당 환자를 치료한 의사나 환자의 지인(친구, 이웃, 직장 동료 등)은 내원 일자와 해당 내원시의 시술 내역, 그 순서 등을 알고 있을 가능성이 크다. 특히 공적 인물의 경우 이러한 정보가 언론에 보도되는 경우도 적지 않다. 이러한 날짜 정보를 배경지식으로 활용하면 개인을 용이하게 재식별해 낼 가능성이 커지게 된다. 가령 누군가 특정일에 큰 수술을 받았고, 언제 퇴원했다는 사실을 알고 있고, 어떠한 수술을 받았는지는 모른다고 생각해 보자. 데이터셋에 포함된 환자 중에서 해당 일자에 마취에 대한 보험 청구를 했다면, 이는 수술을 받았다는 배경지식과 부합한다. 이러한 조건에 부합하는 환자 중 퇴원 일자가 배경지식에 부합하는 경우로 한정시킨다면 환자가 특정될 가능성은 더욱 커질 것이다. 데이터셋에 많은 준식별자가 있다면 더욱 환자 특정이 용이해질 수 있다.

이러한 공격을 막기 위해서는 날짜 정보에 대한 변형이 필요하다. 가장 간단한 방법은 날짜 정보에 잡음(noise)을 추가하는 것이다(예컨대 +/-15일). 만약 추가되는 노이즈에 패턴이 존재한다면 공격자가 이를 이용할 수도 있으므로, 임의의 잡음을 추가할 필요가 있다. 이때 존재 불가능한 날짜는 고려하지 않아야 한다. 예를 들어 사망자는 사망일 이후로는 변경될 수 없도록 해야 할 것이다.

하지만 임의의 잡음을 추가하면 사건의 선후가 뒤바뀔 수 있다는 문제가 발생한다. 가령 입원일인 2010/01/01이고, 퇴원일이 2010/01/05인 경우, 임의로 잡음을 추가하면 퇴원일이 입원일보다 앞서게 되는 문제가 생길 수 있다. 이를 해결하기 위해 정보주체별 사건의 순서와 사건별 간격은 그대로 유지하는 것도 방안도 고려해 볼 수 있다. 하지만 기간 자체가 간접식별자의 성격을 가져서, 이를 이용하여 개인을 재식별하는데 활용될 수 있는 경우도 있다. 예컨대 입원일과 퇴원일 간격을 이용하면 얼마

117) 위의 글, pp. 101-109.

나 중대한 사유로 입원했는지를 파악할 수 있다. 따라서 날짜의 순서는 그대로 유지하면서 그 기간을 임의화(randomize)시키는 방법도 고려할 수 있다. 이 경우에도 모든 사건 간 간격을 임의화시킬 필요가 있는 것은 아니다. 가령 어떤 병원에서 비교적 간단한 시술을 받고 입원 다음 날 퇴원하는 사례가 많다면, 이를 이용하여 재식별될 위험성이 낮을 것이다. 따라서 단기 입원과 중장기 입원을 구분하여 전자는 그대로 간격을 유지하고, 후자의 경우에 대해서만 간격을 임의화하는 방안도 고려할 수 있다.

한편, 데이터의 유형이나 이용 목적에 따라서는 각 사건 간의 간격이 간접식별자의 성격을 갖지 않거나 사건 간의 간격을 유지해도 무방한 경우도 있을 수 있다. 가령, 온라인 쇼핑몰에서의 구매이력 데이터에서 각각의 구매 건 간의 간격을 고려해 보자. 어떤 전자제품을 구매한 다음 일정 간격 전후에 액세사리 제품을 구매하는 시간적 선후 패턴이 존재할 수도 있다. 만약 이러한 패턴이 흔하다면 굳이 이를 간접식별자로 취급할 필요가 없을 수 있고, 오히려 이러한 패턴을 변형한다면 데이터 유용성이 감소하는 효과가 발생할 수도 있다. 이처럼 모든 종단 데이터에 있어 사건 간 간격을 변형해야 한다고 결론짓기는 어렵다.

제 3 절 국내 가이드라인 적용상 문제점

이처럼 종단 데이터의 비식별화는 고려할 사항이 많고, k-익명성 등 기존 방법을 그대로 적용하기도 어렵다. 그런데 이하에서 검토하는 바와 같이 국내 개인정보 보호법에 대한 각종 가이드라인은 종단 데이터를 주로 염두에 두고 작성된 것으로 보기 어렵고, 그 결과 국내에서는 종단 데이터에 대한 비식별 조치가 쉽지 않은 실정이다.

1. 종단 데이터에 대한 익명처리

우선, 종단 데이터에 대한 익명처리에 관해 살핀다. 비식별 조치 가이드라인(2016)은 여러 비식별 기법을 소개하고 있으나, 종단 데이터에 대해서는 구체적인 예시를 들고 있지 않다. 특히 동 가이드라인은 비식별 조치의 적정성을 판단함에 있어 k-익명성을 기본적인 모델로 삼고 필요시 추가적인 평가모델(l-다양성, t-근접성)을 활용하도록 하고 있다.[118] 그러나 앞서 본 바와 같이 k-익명성 기준은 종단 데이터에 그대로 적용되기 어렵다는 한계가 있다.

종단 데이터에 대한 위험 기반 비식별 조치를 적용하기 위해서는 보다 정교한 기준을 활용할 필요가 있다.[119] 우선, 종단 데이터에 대해서는 공격자 배경지식에 관한 합리적 가정이 필요하다. 국내 비식별 조치 가이드라인(2016)이 기본 모델로 삼고 있는 k-익명성은 공격자가 준식별자에 관한 완전한 지식을 보유하고 있다고 가정하고, 해당 준식별자와 연결 가능한 기록의 수가 k개 이상이 될 것을 요구한다. 그러나 위 제2장 제4절 제3항에서 검토한 바와 같이, 종단 데이터에서는 재식별 공격자가 정보주체에 관하여 일부 정보만을 보유하고 있다고 가정하는 것이 보다 합리적이다. 즉, 공격자가 보유하고 있는 정보량의 합리적 추정치를 산정한 다음, 시뮬레이션을 통해 해당 정보를 보유한 공격자가 개인을 식별해 낼 확률을 계산하는 방법을 적용함으로써 적정한 비식별 수준을 결정할 수 있다.

다음으로 고려해 볼 수 있는 방법은 종단 데이터에 대한 재식별 위험성 평가에 있어 평균 위험(average risk)을 적용하는 것이다. 국내 비식별 조치 가이드라인(2016)에 따르면 전체 데이터셋에서 모든 기록에 대해서 k개 이상의 동일한 기록이 포함될 것을 요구하므로, 이는 최대 위험(maximum risk)을 기준으로 한 것으로 평가될 수 있다. 그러나 앞서 살펴본 바와 같이 평균 위험(average risk)을 기준으로 할 경우 표본 추출을 통해 위험 수준을 낮출 수 있다. 즉, 전체 데이터셋 중 일부(가령

118) 비식별 조치 가이드라인(2016), 9면..
119) El Emam and Arbuckle (2013) (각주 15), p. 87.

30%)만을 임의로 표본 추출하여 제공할 경우, 재식별의 평균 위험을 낮출 수 있게 되는 것이다. 표본 추출 방법은 비식별 조치 가이드라인(2016)에는 언급되어 있지 않지만, 국내에서 이미 활용된 바 있다. 한국신용정보원은 CreDB(금융 빅데이터 개방시스템)을 통해 한국신용정보원에 집중되는 신용정보를 표본추출하고, 비식별 조치를 취한 다음 통계작성 및 학술 연구 목적으로 이용자에게 제공하고 있다.[120]

2. 종단 데이터에 대한 가명처리

다음으로 종단 데이터에 대한 가명처리에 관해 살핀다. 가명정보 처리 가이드라인(2020)은 '가명정보 시계열 분석 절차'를 규정하고 있다.[121] 이는 여러 시점에 걸쳐 추가·수정 되는 횡단면 데이터를 해당 시점마다 가명처리를 하되, 정보주체에 대해서 동일한 가명을 부여함으로써 각기 가명처리된 데이터를 비교하여 시계열 분석을 할 수 있도록 한 것이다. 예컨대 2018. 1. 1. 가명처리한 정보와 2019. 1. 1. 가명처리한 정보가 있는 경우, 정보주체에 대해 동일한 가명을 부여함으로써, 해당 정보주체에 대해 위 기간 동안 어떠한 변화가 있었는지를 분석할 수 있게 된다.

그러나 가이드라인에 언급된 바와 같이 여러 시점에서 가명처리된 횡단면 데이터에 대해 시계열 분석을 수행하는 것은 종단 데이터 자체에 대한 가명처리와는 구별된다. 예컨대 제2장 제5절에서 예시로 든 쇼핑몰 구매이력 데이터(<표 2>)를 가명처리할 때에는 과연 어떠한 조치까지 이루어져야 할 것인지는 명확하지 않다. 물론 약한 가명처리를 수행할 경우에는 직접식별자만 삭제하면 되므로 구매이력에 있어 고객ID를 삭제·대체하는 수준의 조치를 취하면 될 것이다. 그러나 앞서 본 바와 같이 현행 가이드라인은 강한 가명처리를 요구하고 있으므로, 단지 고객ID를 삭제·대체하는 것만으로는 부족하고, 다른 정보를 이용한 식별 가능성을 함께 고려하여야 한다. 이 때에는 ① 개인식별 가능성 판단에 있어 어떠한 다른 정보까지 고려해야 하는 것인지(예컨대 구매이력 테이블의

120) https://credb.kcredit.or.kr/
121) 가명정보 처리 가이드라인(2020), 64-65면.

경우 인터넷에 공개된 다수의 구매후기와 결합될 가능성까지 고려해야 하는지, 아니면 가명정보 처리자가 실제 보유하고 있는 정보만을 고려해야 하는지), ② 종단 데이터에 있어 특이정보의 판단 기준은 어떻게 되는지(가령 고가품의 경우는 특이정보로 보아야 하는지, 그렇다면 얼마나 고가품이어야지 특이정보에 해당하는지, 다수의 구매경력이 있는 경우는 특이정보로 보아야 하는지) 등 불명확한 부분이 적지 않다.

제 4 절 종단 데이터 비식별화의 대안과 합성 데이터

종단 데이터 비식별화의 어려움은 국내의 경우에 한정된 것은 아니다. 종단 데이터에 대해서는 국제적으로도 체계화된 비식별 조치 방법론이 충분히 개발되어 있지 못한 실정이다.[122] 종단 데이터에는 정보주체에 관한 많은 정보가 포함되어 있고, 다양한 유형의 종단 데이터가 존재하며, 각각의 맥락마다 고려해야 하는 배경지식의 정도에도 차이가 있기 때문이다. 그 결과 종단 데이터에 대해서는 비식별 조치 이외의 대안적 방법들이 고려되고 있는 상황인바, 이하에서는 그 대안으로서 최근 주목받고 있는 합성 데이터(synthetic data)를 소개한다.

합성 데이터는 원본 개인정보의 특성을 분석하여 원본과 최대한 유사한 통계적 성질을 보이는 가상의 데이터를 생성한 것이다.[123] 즉, 이는 원본 데이터의 통계적 특성을 추출하는 모델을 만들고, 해당 모델을 이용하여 가상의 데이터를 합성해 냄으로써 개인식별 가능성은 없지만 원본 데이터와 유사한 통계적 특성을 갖도록 한 것이다. 이로써 해당 데이터는 외관상 원본 데이터와 유사해 보이지만, 개인을 식별해 낼 우려가 낮아진다.[124]

---

122) Ruiz (2018) (각주 111), p. 216

123) Arbuckle and El Emam (2020) (각주 76), pp. 150-155.

124) 위의 글, p. 152. 다만 부적절한 합성 방법이 활용될 경우에는 모델 학습에 사용된 원본 개인정보가 그대로 다시 합성될 수 있으므로, 이를 막기 위한 적절한 일반화

합성 데이터는 다음과 같이 3가지로 구분될 수 있다.[125] ① 완전 합성 데이터(fully synthetic data)는 직·간접식별자뿐만 아니라 속성정보 전부를 가상으로 생성한 것이다. 적절한 합성 방법이 활용될 경우에는 더 이상 개인정보에 해당하지 않는다고 인정될 가능성이 크다. ② 부분 합성 데이터(partially synthetic data)는 데이터 중에서 비밀로 여겨지는 일부 데이터만을 가상으로 생성하고, 나머지는 원본 값을 그대로 유지하는 것이다. 다만, 이 경우 원본 값이 유지된 부분에 의해서 정보 주체가 재식별될 가능성이 남아 있는지 확인해야 하고, 해당 부분에 대해서는 추가적 비식별 조치가 필요할 수도 있다. ③ 혼합 합성 데이터(hybrid synthetic data)는 원본 데이터와 합성 데이터를 혼합하되 어느 부분이 가상인지 알 수 없도록 하는 것이다. 하지만, 원본 데이터의 비중이 매우 낮지 않은 이상 이러한 데이터는 여전히 개인정보에 해당할 가능성이 크다.

합성 데이터는 개별 정보주체에 관해 식별된 데이터가 필요하지는 않으나 통계적 속성은 그대로 유지될 필요가 있는 경우, 예컨대 데이터 분석, 각종 경진대회, 해커톤, 교육, 훈련, 오픈소스 데이터, 개념 검증 및 소프트웨어 개발 및 테스팅 등의 분야에서 유용하게 활용될 수 있다. 또한 신용 데이터와 같이 재식별이 이루어질 경우 정보주체의 프라이버시 침해나 영업비밀 침해 우려가 있어, 공개된 데이터가 많지 않은 경우 활용될 여지가 있다.[126] 이러한 경우에는 비록 가상적이기는 하나 온전한 형태로 된 데이터(나이, 성별, 주소, 학력, 소득, 자산 등)를 생성하여 공개할 경우 널리 활용될 여지가 있다.

(generalization) 조치가 필요해 진다.

125) 위의 글. p. 154.

126) Assefa, Samuel, "Generating synthetic data in finance: opportunities, challenges and pitfalls" (2020).

제 6 장 결론

본 연구에서는 위험기반 비식별 조치와 이에 관한 GDPR 및 개인정보 보호법상의 쟁점을 살펴보았다. 본 연구의 주요한 결론은 다음과 같다.

우선, 비식별 조치는 개인정보의 활용과 정보주체 보호를 위해 유용하게 활용될 수 있다. 개인정보처리자는 과학적 연구 활동을 위해서 또는 개인정보를 통계분석하여 그 결과를 통해 의사결정을 보조하기 위한 목적으로 개인정보를 비식별화하여 활용할 수 있고, 나아가 정보주체에 대한 보호를 더욱 강화하기 위해서도 비식별 조치를 활용할 수 있다. 다만, 적정한 개인정보 비식별화를 위해서는 체계적이고 정식화된 절차에 따라 재식별 위험성을 평가하고, 위험성에 비례한 조치를 취할 필요가 있다. 위험기반 비식별 방법론은 위 과정을 통해 데이터 유용성과 개인정보 보호 수준 간의 형량을 하고, 이를 통해 최적의 '골디락스(Goldilocks)' 지점을 찾는 것을 목적으로 한다.

둘째, GDPR의 개인정보 개념에 관한 법적 논쟁을 살펴보면, 어떤 정보가 개인정보인지 여부를 판단하는 것은 역동적인 작업임을 알 수 있다. GDPR은 개인정보의 개념 판단에 있어 상대적·주관적 접근을 취한 것으로 해석될 수 있으며, 이는 GDPR 전반에 걸친 위험기반 접근법에 부합하는 것이다. 한편 GDPR에 도입된 '가명처리'의 개념은 약한 가명처리와 강한 가명처리로 구분할 수 있으며, 양자에 부여되는 인센티브는 차이가 있다. 특히 강한 가명처리가 이루어진 경우 정보주체의 권리에 대한 예외가 인정될 수 있음에 비해, 약한 가명처리만이 이루어진 경우에는 이러한 예외가 적용될 수 없다.

셋째, 우리 개인정보 보호법이 도입한 가명처리/가명정보, 익명처리/익명정보의 해석에 있어 위와 같은 GDPR의 규정이 시사점을 준다. 이에 따라 본 연구에서는 비식별조치를 익명처리, 강한 가명처리, 약한 가명처리로 구분하였다. 이를 간략히 요약하면, ① 익명처리는 개인식별 가능성이 희박(remote)해지도록 개인정보를 변형하는 것으로, 주로 데이터를 공중

에 공개하거나 재식별 공격에 노출될 우려가 있는 상황을 염두에 둔 조치이다. ② 그보다 완화된 조치는 강한 가명처리이다. 강한 가명처리는 해당 정보를 처리하는 자가 체계적이고 확실한 수단에 의해 정보주체를 재식별해 낼 수 없도록 만드는 것으로서, 이는 주로 개인정보를 동의받은 목적 이외의 범위에서 이용하거나, 연구자 등 제3자에게 제공하는 상황을 염두에 둔 조치이다. ③ 약한 가명처리는 개인식별정보를 삭제하거나 암호화하는 방식으로 이루어지는 것으로서, 개인정보처리자가 개인정보 보호 수준을 높이기 위해 채택할 수 있는 조치이다.

마지막으로, 본 연구는 비식별 조치의 구체적 적용 사례로서 종단 데이터의 문제를 검토하였다. 종단 데이터에는 정보주체에 관한 많은 정보가 포함되어 있어, k-익명성과 같이 전통적 비식별화 기법을 적용하기 어렵다는 문제가 있다. 이에 아직 국제적으로도 체계화된 비식별 조치 방법론이 충분히 개발되어 있지 못한 실정이다. 이러한 점은 국내 비식별 조치 가이드라인에서도 마찬가지로 드러난다. 그 결과 종단 데이터에 대해서는 기존 비식별 조치 이외의 대안적 방법, 예컨대 합성 데이터 기법 등이 고려될 필요가 있다.

참 고 문 헌

[국내 문헌]

고학수·백대열·구본효·정종구·김은수, “개정 개인정보 보호법상 가명정보의 개념 및 가명처리에 관하여”, 서울대학교 인공지능정책 이니셔티브 이슈페이퍼 2020-1, (2020)

고학수·최경진, “개인정보의 비식별화 처리가 개인정보 보호에 미치는 영향에 관한 연구”, 개인정보보호위원회 연구보고서, (2015)

고학수 외, “개인정보 비식별화 방법론: 보건의료정보를 중심으로”, 박영사, (2017)

고학수 외, “사법부에서의 인공지능(AI) 활용방안”, 법원행정처, (2020)

관계부처합동, 개인정보 비식별 조치 가이드라인(2016)

개인정보보호위원회, 가명정보 처리 가이드라인(2020)

김은수, “비식별화 방식을 적용한 개인정보보호에 대한 연구”, 서울대학교 박사학위 논문, (2018)

오철호, “정책결정, 증거 그리고 활용 - 연구경향과 제언 - ”, 한국정책학회보, 제24권 1호, (2015)

이동진, “개인정보 보호법 제18조 제2항 제4호, 비식별화, 비재산적 손해 -이른바 약학정보원 사건을 계기로-”, 정보법학 제21권 제3호, (2017)

이명진, “사회과학도가 알아야 할 자료 II - 시계열 조사 (Times Series Survey)”, 한국사회과학자료원 엣세이

함께하는 시민행동, “고객 몰래 개인정보 결합한 기업과 공공기관에 대한 고발”, 기자회견문

[외국 문헌]

Arbuckle, Luk and Khaled El Emam, “Building an Anonymization Pipeline: Creating Safe Data”, O′Reilly Media, (2020)

Article 29 Data Protection Working Party, Opinion 04/2007 on the Concept of Personal Data (WP 136) 01248/07/EN, p. 15.

Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques", (2014)

Assefa, Samuel, "Generating synthetic data in finance: opportunities, challenges and pitfalls" (2020)

Donoho, David. "50 years of data science." Journal of Computational and Graphical Statistics 26.4 (2017): 745-766.

El Emam, Khaled, et al. "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset." Journal of medical Internet research 14.1 (2012)

El Emam, Khaled and Luk Arbuckle, "Anonymizing health data: case studies and methods to get you started", O'Reilly Media, (2013)

Finck, Michèle, and Frank Pallas, "They who must not be identified-distinguishing personal from non-personal data under the GDPR", International Data Privacy Law, (2020)

Gartner, "A Framework for Applying AI in the Enterprise", 2019. 1. 29.

Greene, Kate, "The $1 Million Netflix Challenge", MIT Technology Review (2006. 10. 6.)

Hintze, Mike, "Viewing the GDPR through a de-identification lens: a tool for compliance, clarification, and consistency", International Data Privacy Law 8.1, (2018)

Hintze, Mike, and Khaled El Emam. "Comparing the benefits of pseudonymisation and anonymisation under the GDPR." Journal of Data Protection & Privacy 2.2 (2018).

Information Commissioner's Office (ICO), "Anonymisation: Managing Data Protection Risk Code of Practice", (2012. 11.) pp. 22-25.

Narayanan, Arvind, and Vitaly Shmatikov, "Robust de-anonymization of large sparse datasets", 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, (2008)

Ruiz, Nicolas, "A General Framework and Metrics for Longitudinal Data Anonymization", International Conference on Privacy in Statistical Databases, Springer, (2018)

Singel, Ryan, "NetFlix Cancels Recommendation Contest After Privacy

Lawsuit", Wired, (2018. 12. 3.)
Sweeney, Latanya, "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002)

웹사이트

https://en.wikipedia.org/wiki/Netflix_Prize
https://www.lawtalk.co.kr/sentence-estimation
https://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf
https://credb.kcredit.or.kr/

Abstract

Risk-based De-identification and Pseudonymisation

KIM, Byoung-Pil

School of Law

The Graduate School

Seoul National University

De-identification is useful in both utilizing and protecting personal information. A risk-based approach in de-identification evaluates the risk of re-identification through systematic and formalized process and applies technical measures proportional to that risk. Such methodology aims to balance between data utility and privacy protection.

The legal definition of personal data in the GDPR shows that determining the scope of personal data is a dynamic task. The de-identification methodologies envisaged under the GDPR can be categorized as (i) anonymisation, (ii) strong pseudonymisation and (iii) basic pseudonymisation. The GDPR provides different incentives to each methodology. Such analysis of the GDPR sheds lights on the interpretation of the Korean Personal Information Protection Act of 2020. The anonymisation is a process to ensure that the risk of

re-identification becomes remote and it is mainly concerned about disclosing data to the public. The strong pseudonymisation is more relaxed process, through which the controller cannot readily identify the data subject. Such measure is mainly considered when personal information shared with a third party such as researchers for scientific research purpose. The most relaxed measure is the basic peudonymisation. This is a process to remove or replace only direct identifers, through which the level of privacy protection can be increased.

This paper examines the problem of longitudinal data as an example of the practical difficulties in applying de-identification measures. As longitudinal data contain substantial amount of information about the data subject, it is often very difficult to apply traditional techniques such as k-anonymity. Yet structured methodologies for de-identification have not yet been well developed. This is also the case in the guidelines for de-identification techniques in Korea. Therefore, for longitudinal data, alternative methods such as synthetic data need to be considered.

keywords : de-identification, anonymisation, pseudonymisation, re-identification, longitudinal data, synthetic data

Student Number : 2018-29594