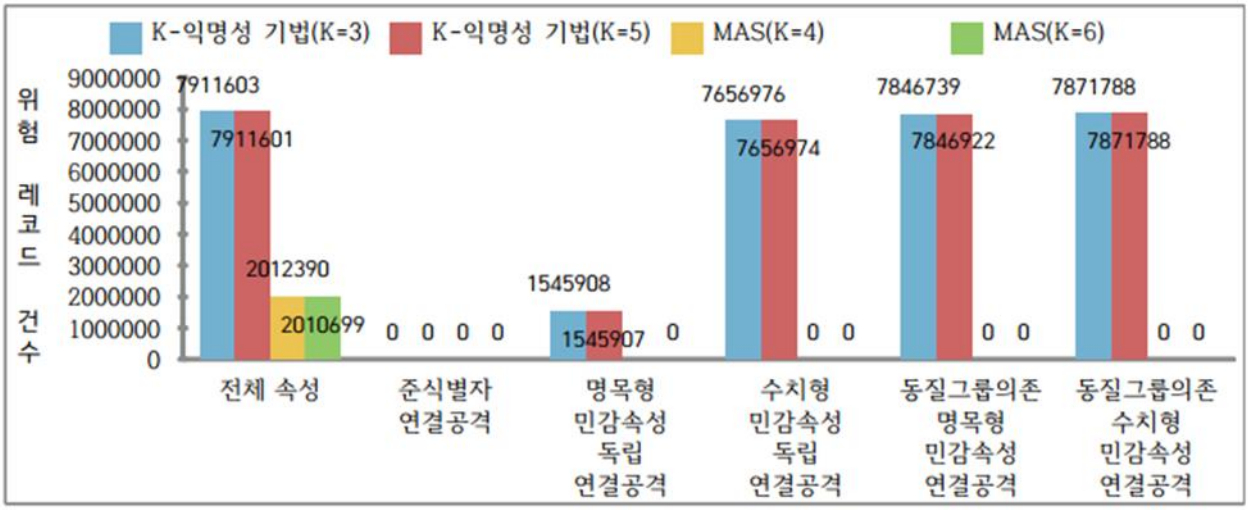
정보화사회실천연합			보 도 자 료
보도일시	2020. 04. 30 (목) 조간(온라인 금일)부터 보도하여 주시기 바랍니다		
배포일시	2020. 04. 29 (수) 09:00		총 7쪽(붙임 5쪽 포함)
담당부서	개인정보보호(070-7797-2583)	작 성 자	대표 손영준

비식별 정보 재식별 위험성에 관한 연구 보고서 3년간 미 공개

-과기부 정책 추진에 부정적 결론, 보고서 누락-

□ 미래부의 산하기관인 한국과학기술정보연구가 2016-2017년도에 수행한 **'개인정보 비식별 자료 생성유통의 현장 적용을 위한 실증 연구'** 에 의하면 **'개인정보를 비식별 조치를 하여도 재식별 위험성이 높다'**는 결론을 내고 있다.



[그림 3-21] 비식별화 기법에 따른 m 유일성 측정 결과: 신용도

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함

□ 외부기관인 TTA(한국정보통신기술협회)를 통한 비식별 데이터에 대한 검증에 따르면 10만건을 표본으로 3회 시험 결과 평균 98.64%가 원본데이터와 유사하다는 결과가 나타났다.

□ 또한 보고서도 비식별한 데이터의 유일성을 측정결과 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결에 공격시 취약하며, 이는 민감속성(일반 속성자) 값을 미리 알고 시도되는 민감속성 기반 공격을 막을 수 없다고 결론을 내고 있다.

- (예:공격자가 신림동에 사는 37세의 여자인 김화복씨가 페암에 걸린 사실을 미리 알고 있고 비식별 데이터에 30대, 여자, 신림동에 사는 5명의 동질 그룹에 페암 걸린 사람이 오직 한사람일 경우 이 레코드가 김화복씨의 레코드임을 재식별함과 동시에 해당 레코드에 있는 김화복씨의 다른 민감속성 값을 알아낼 수 있음)

□ 해당 과제는 국가과학기술지식정보서비스(NTIS)에 등록되어 있으나 연구에 대한 성과물인 연구 보고서는 사업이 완료되면 성과물로 등록하여야 하나 **3년이 지난 현재에도 해당 보고서가 누락**되어 있다.

□ 개인정보를 비식별 처리하여도 재 식별 위험성이 존재한다는 해외 연구는 있었으나, 국내 연구 사례로는 처음으로 정부가 연구 과제로 수행한 연구 결과에 대하여 미공개한 사유가 **"정부가 개인정보 활용 정책을 추진하는데 부정적인 결과가 도출"**된 내용으로 인하여 누락한 것이 아닌가 의구심이 든다.

□ 만약 정부가 정책을 추진하는데 걸림돌이 된다는 판단으로 해당 보고서를 누락하였다면 이는 개인정보 활용이란 사회적 관심을 고려해 볼 때 단순한 보고서의 누락이 아니라, 국민의 알 권리를 심각하게 침해하는 행위로 해당 기관에 대한 감사 및 관련자에 대한 징계를 통하여 이와 같은 행위를 근절하여야 한다.

□ 또한 데이터3법 통과에 따른 입법 예고된 시행령은 가명정보를 전문기관의 안전한 공간에 저장하도록 하고 있으나 **예외적으로 심의를 통하여 반출을 허용**하는 조항에 의하여 이는 **타 기관이 수집한 개인정보도 같이 반출**되는 것으로 **재식별을 통한 악용의 소지가 높다.**

□ 따라서 부득이한 사유로 반출을 할 때는 전체 데이터가 아닌 **표본 데이터만 반출**하도록 하여 **재식별을 통한 가명정보의 오남용을 방지**하여야 한다.

-끝-

□ 붙임

1. 연구 과제 내용
2. 연구보고서 결론 부분

붙임 1 연구 과제 내용

□ 해당 연구 보고서는 과기정통부의 산하기관인 한국과학기술정보연구가 2016-2017년도에 수행한 '**개인정보 비식별 자료 생성유통의 현장 적용을 위한 실증 연구**' 과제의 성과물로 과기정통부에 정보공개 청구를 통하여 입수하였다.

사업 2016 / 미래창조과학부 / 일반사업
미래성장동력플래그십프로젝트 (조사분석사업명 : 미래성장동력플래그십프로젝트)

▾ 본 과제에 참여한 연구자
연구책임자 김정선
참여연구원 강미나

과제

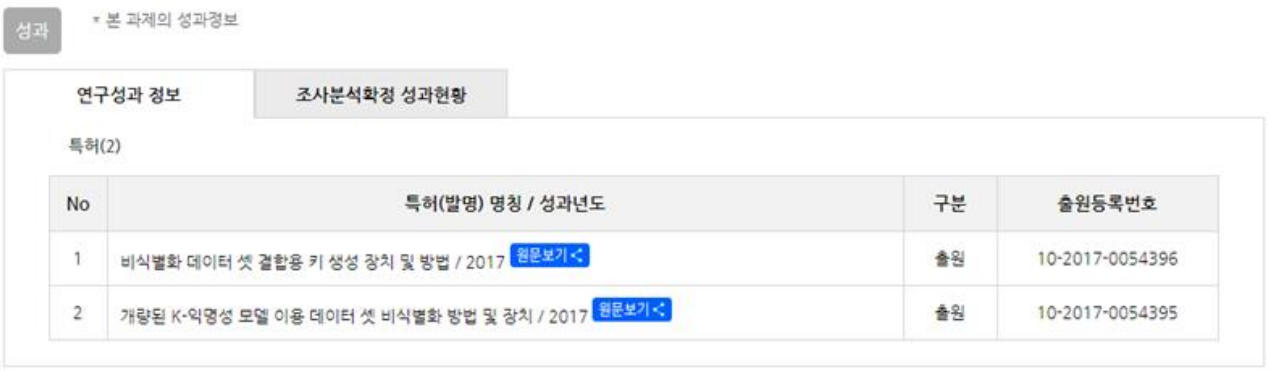
개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증
1711047795 / 에스케이텔레콤㈜ / 주관과제 / 총 연구비 1,722.00 백만원
과학기술표준 분류 1 : 정보/통신 / 기타 정보/통신 / 달리 분류되지 않는 정보/통신 / 100%

기본정보 | 과제구성정보 | 요약서 | 참여연구원

과제고유번호	1711047795	당해연도 연구기간	2016-10-18 ~ 2017-04-17
(기관)세부과제번호	CN16040-협동7	총연구기간	2016-10-18 ~ 2017-04-17
내역사업명	미래성장동력플래그십프로젝트		
과제명 국문	개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증		
과제명 영문	Field-appicability substantiation for creation and districation of anonymized personal information		
과제수행기관	에스케이텔레콤㈜		
연구관리전문기관	미래창조과학부	과제관리(전문)기관	한국과학기술기획평가원

과제진행상태	종료	실용화대상여부	실용화대상
연구개발단계	기타	연구수행주체	대기업
세부과제성격	연구관리	연구개발성격	기타개발
기술수명주기	기타	지역	서울특별시

< 출처 : NTIS 의 국가 R&D 과제 정보의 해당 과제 등록 내용 갈무리 >



성과 ▾ 본 과제의 성과정보

연구성과 정보 | 조사분석확정 성과현황

특허(2)

No	특허(발명) 명칭 / 성과년도	구분	출원등록번호
1	비식별화 데이터 셋 결합용 키 생성 장치 및 방법 / 2017 원문보기	출원	10-2017-0054396
2	개량된 K-익명성 모델 이용 데이터 셋 비식별화 방법 및 장치 / 2017 원문보기	출원	10-2017-0054395

< 출처 : NTIS의 국가 R&D과제 정보의 해당 과제 연구성과 등록 내용 갈무리 >

붙임 2 연구보고서 결론 일부

□ 외부기관 검증 수행 내용

마. 외부기관 검증 수행

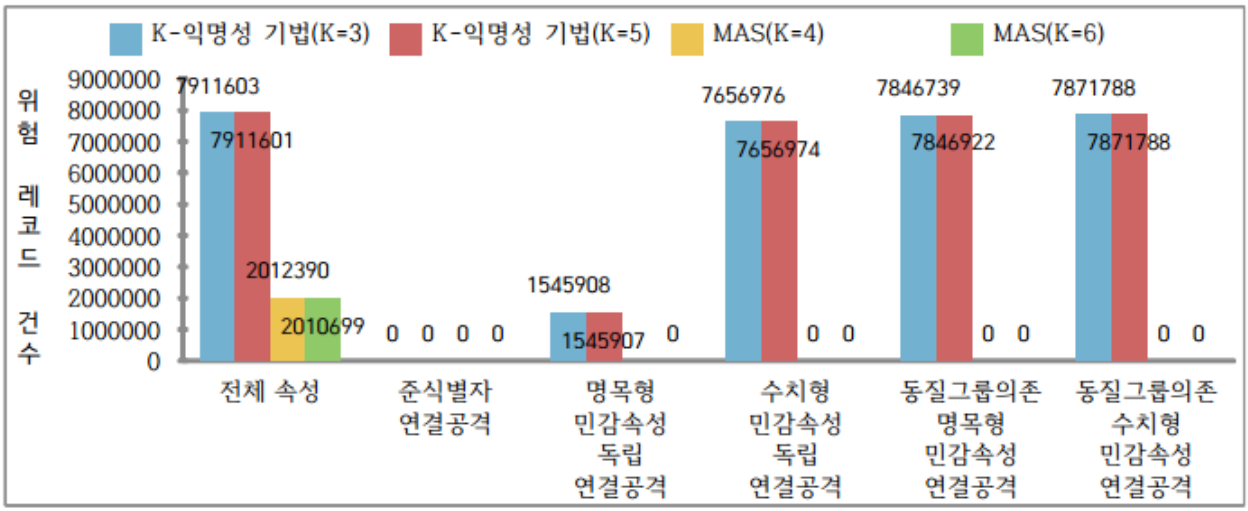
(1) TTA V&V(Verification & Validation) 인증 수행

○ 원본 유사도 검증
- 추상화 기반 비식별화로 변환된 레코드 세트와 원본 레코드 세트 간의 유사도가 90% 이상 보존되는지 확인
- 시료데이터는 국민건강보험에서 제공하는 공개 데이터 중 "건강 검진 정보" 데이터 사용

○ 원본 유사도 검증 결과
- 31가지 속성 값을 가진 100,000건의 데이터가 담겨있는 테이블에 대해 추상화 기반 비식별화를 수행했을 시, 원본 테이블과 변경된 테이블이 90%이상 유사한지 확인함
- 100,000건 데이터가 담긴 테이블 10개를 시료로 사용하여 시험하였고, 10개 테이블에 대하여 총 3회 시험 결과 원본테이블과 평균 98.64% 유사도가 유지되는 것을 확인함

<그림1. 연구 보고서 p56 갈무리>

□ 재식별 가능성 분석 내용

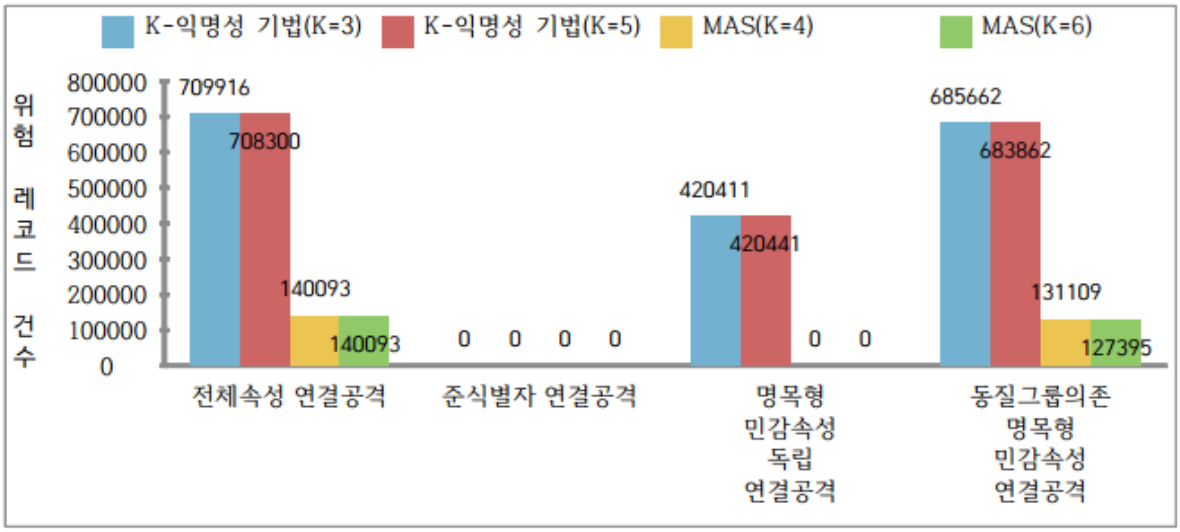


[그림 3-21] 비식별화 기법에 따른 m 유일성 측정 결과: 신용도

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함

<그림2. 연구 보고서 p63 갈무리>

② 장애우 거소지 실증데이터에 대한 유일성 분석 결과

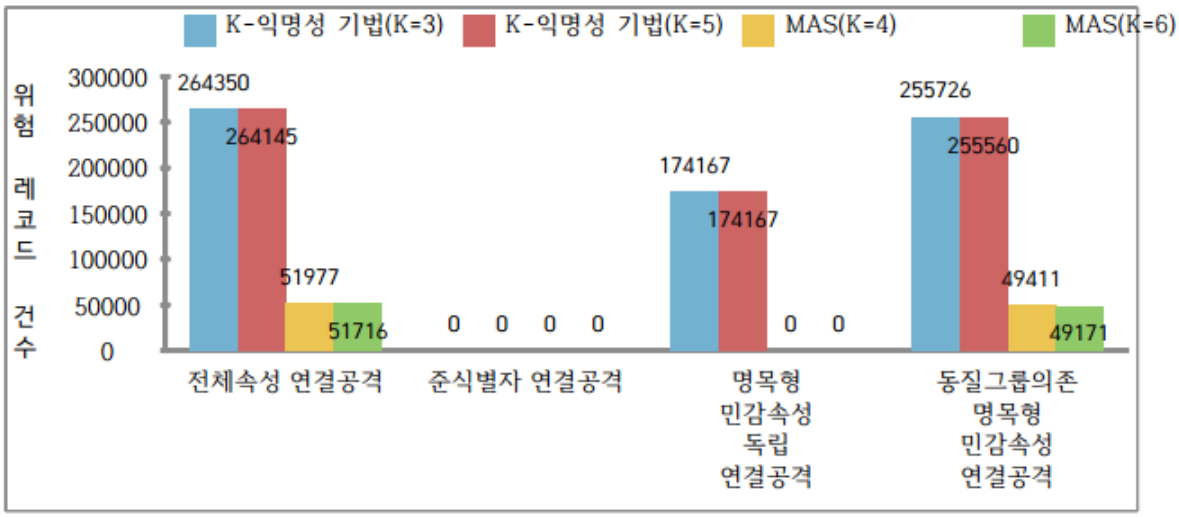


[그림 3-22] 비식별화 기법에 따른 m 유일성 측정 결과: 장애우 거소지

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함

<그림3. 연구 보고서 p63 갈무리>

③ 외국인 체류지 실증데이터에 대한 유일성 분석 결과



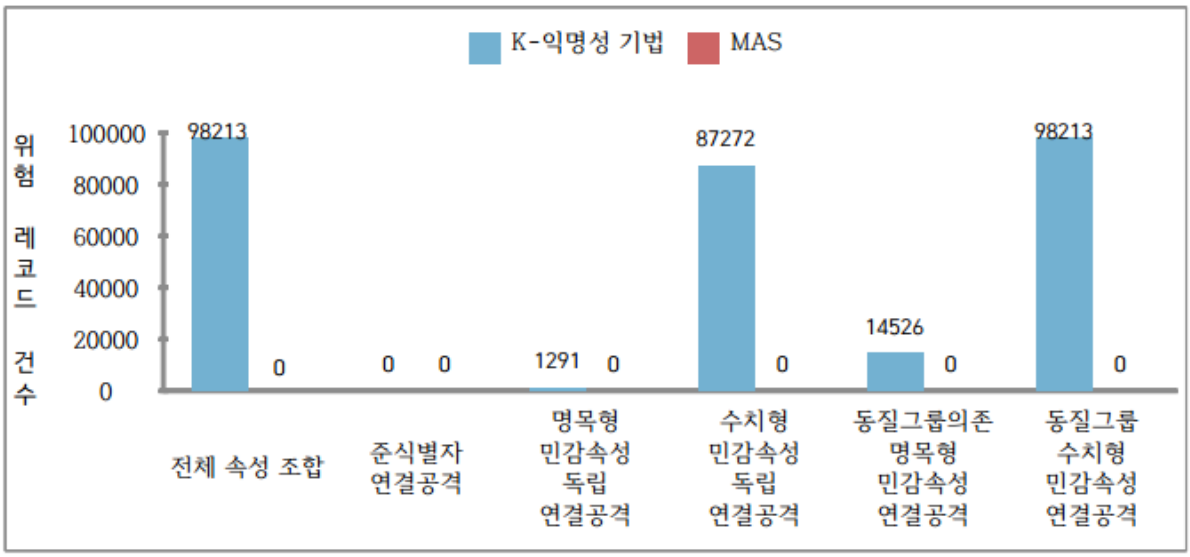
[그림 3-23] 비식별화 기법에 따른 m-유일성 측정 결과: 외국인 체류지

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함

<그림4. 연구 보고서 p63 갈무리>

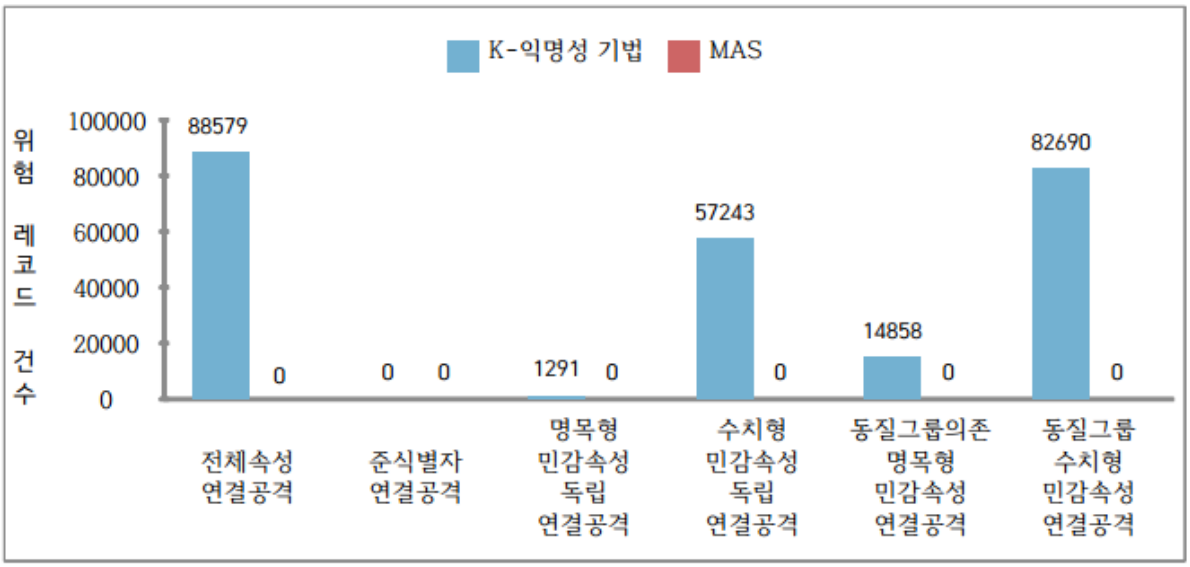
□ 연계 데이터 재식별 가능성 분석에 대한 결과

① k-익명성 기법/MAS 결과 데이터 A에 대한 유일성 결과 검사 비교



[그림 3-27] 연계 기법에 따른 재식별 위험 레코드 건수 비교 : 신용도 A

② k-익명성 기법/MAS 결과 데이터 B에 대한 유일성 결과 검사 비교



[그림 3-28] 연계 기법에 따른 재식별 위험 레코드 건수 비교 : 신용도 B

- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합은 취약함

<그림5. 연구 보고서 p67 갈무리>

□ 현 제도의 제약점

(1) 현재 k-익명성 프라이버시 모델의 재식별 가능성 존재

(가) k-익명성, l-다양성 모델의 한계

- k-익명성은 같은 준식별자 값을 갖는 동질그룹의 수가 최소 k개 이상 되도록 가공하여 준식별자를 미리 알고 시도되는 준식별자 기반 재식별 공격을 방어하는 목적으로 2002년 개발됨
- 단점 : 동질그룹의 민감속성값이 모두 같을 경우 민감정보가 유출될 수 있으므로 동질그룹의 민감 속성값을 l-개 이상으로 가공하는 l-다양성 추가 적용 제안됨
- k-익명성, l-다양성 모델만으로는 민감속성(일반 속성자) 값을 미리 알고 시도되는 민감속성 기반 공격을 막지 못함
  (예:공격자가 신림동에 사는 37세의 여자인 김화복씨가 폐암에 걸린 사실을 미리 알고 있고 비식별 데이터에 30대, 여자, 신림동에 사는 5명의 동질 그룹에 폐암 걸린 사람이 오직 한사람일 경우 이 레코드가 김화복씨의 레코드임을 재식별함과 동시에 해당 레코드에 있는 김화복씨의 다른 민감속성 값을 알아낼 수 있음)

<그림6. 연구 보고서 p82 갈무리>

-끝-