SNUH
분당서울대학교병원

개인정보 가명·익명 기술 경진대회 후기

분당서울대학교병원 정보보호팀 김기수

가명정보!

가명정보	익명정보
추가정보의 사용 없이는 **특정 개인을 알아볼 수 없게** 조치한 정보	더 이상 **개인을 알아볼 수 없게** (복원 불가능할 정도로) 조치한 정보

'보건의료 데이터 활용 가이드라인' 공개..
가명정보 활용방안 제시

데이터 심의위원회 구성

1. **근거**
   - 「개인정보 보호법」에 따른 가명정보 활용을 위한 위원회 설치
   - 「보건의료 데이터 활용 가이드라인」에서 데이터 심의위원회 설치운영을 권고

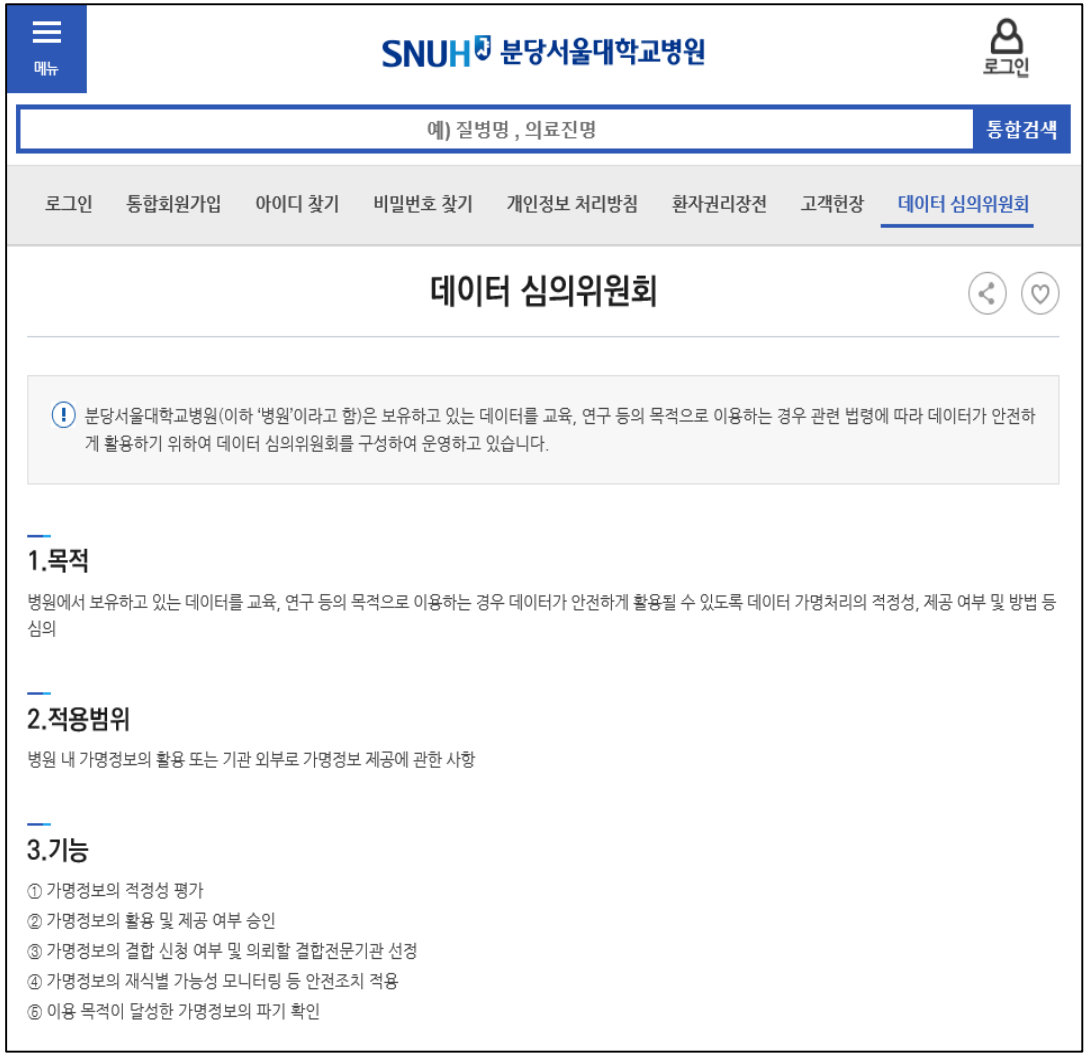
2. **목적 및 범위**
   - 병원에서 보유하고 있는 데이터를 교육, 연구 등의 목적으로 이용하는 경우 데이터가 안전하게 활용될 수 있도록 데이터 가명처리의 적정성, 제공 여부 및 방법 등 심의
   - **병원 내 가명정보의 활용 또는 기관 외부로 가명정보 제공에 관한 사항**

3. **기능**
   - 가명정보의 적정성 평가
   - 가명정보의 활용 및 제공 여부 승인
   - 가명정보의 재식별 가능성 모니터링 등 안전조치 적용
   - 이용 목적을 달성한 가명정보의 파기 확인

데이터 심의위원회 운영

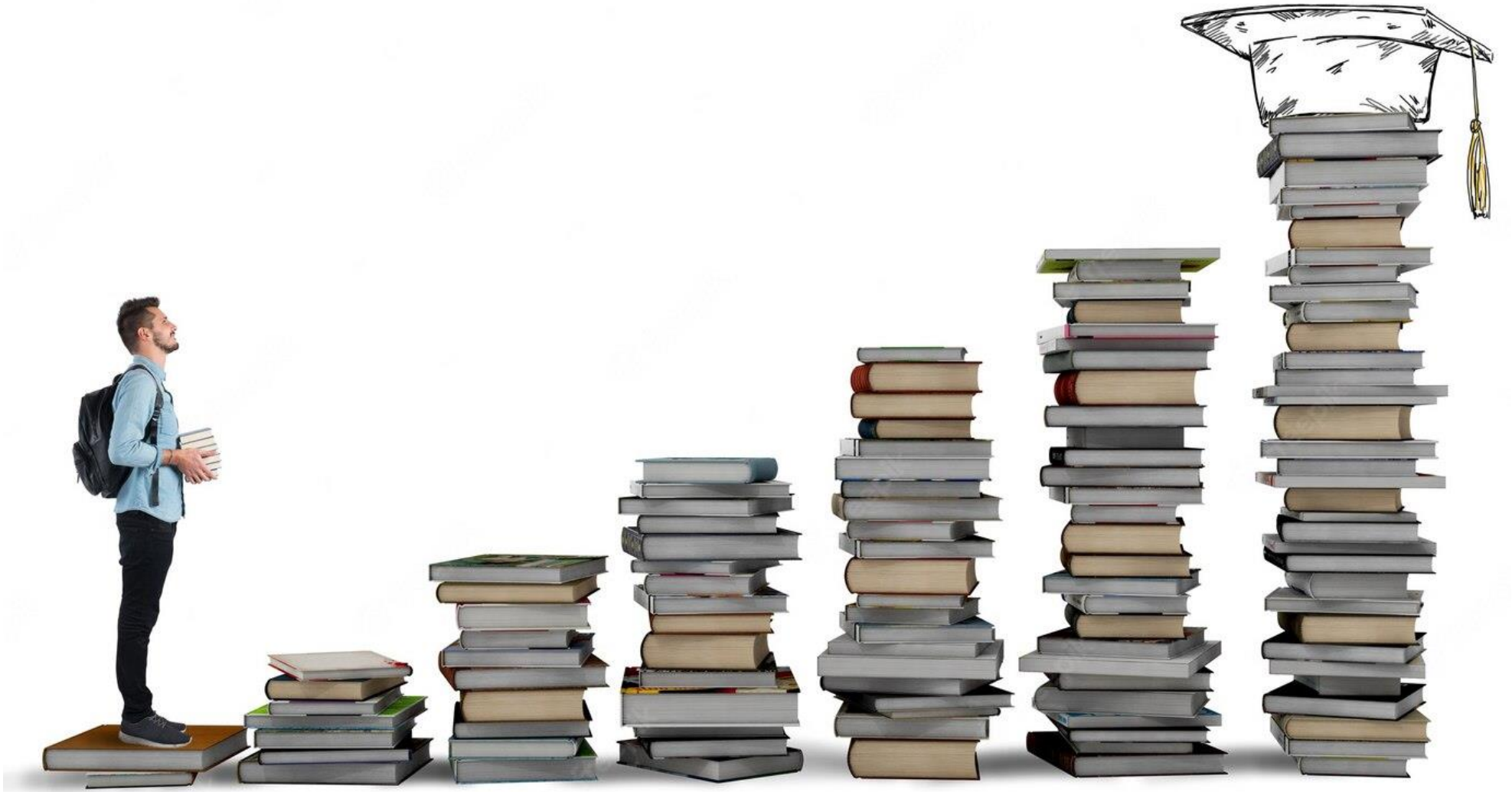
일자	경과
2021.05.01	분당서울대학교병원 데이터 심의위원회 규정 제정
2021.08.01	데이터 심의위원회 구성 (위원장 **1**명, 내부위원 **5**명, 외부위원 **7**명, 간사 **1**명) - 외부위원 과반 이상, 정보주체(환자) 대표 **3**명, 변호사 **2**명, 데이터 활용 전문가 **3**명
2021.09.27	**2021-1**차 데이터 심의위원회 (**4**건 심의)
2021.10.28	**2021-2**차 데이터 심의위원회 (**2**건 심의)
2021.11.19	**2021-3**차 데이터 심의위원회 (**3**건 심의)
2021.12.28	**2021-4**차 데이터 심의위원회 (**11**건 심의)
2022.03.18	**2022-1**차 데이터 심의위원회 (**7**건 심의)
2022.05.13	**2022-2**차 데이터 심의위원회 (**17**건 심의)

데이터 심의위원회 운영


메뉴
SNUH 분당서울대학교병원
로그인
예) 질병명 , 의료진명
통합검색
로그인 통합회원가입 아이디 찾기 비밀번호 찾기 개인정보 처리방침 환자권리장전 고객현장 데이터 심의위원회
데이터 심의위원회
분당서울대학교병원(이하 '병원'이라고 함)은 보유하고 있는 데이터를 교육, 연구 등의 목적으로 이용하는 경우 관련 법령에 따라 데이터가 안전하게 활용하기 위하여 데이터 심의위원회를 구성하여 운영하고 있습니다.
1.목적
병원에서 보유하고 있는 데이터를 교육, 연구 등의 목적으로 이용하는 경우 데이터가 안전하게 활용될 수 있도록 데이터 가명처리의 적정성, 제공 여부 및 방법 등 심의
2.적용범위
병원 내 가명정보의 활용 또는 기관 외부로 가명정보 제공에 관한 사항
3.기능
① 가명정보의 적정성 평가
② 가명정보의 활용 및 제공 여부 승인
③ 가명정보의 결합 신청 여부 및 의뢰할 결합전문기관 선정
④ 가명정보의 재식별 가능성 모니터링 등 안전조치 적용
⑤ 이용 목적이 달성한 가명정보의 파기 확인


데이터 심의위원회
통합 VIEW 이미지 지식iN 인플루언서 동영상 쇼핑 뉴스 어학사전 지도
www.snubh.org › member
데이터 심의위원회 - 분당서울대학교병원
병원에서 보유하고 있는 데이터를 교육, 연구 등의 목적으로 이용하는 경우 데이터가 안전하게 활용될 수 있도록 데이터 가명처리의 적정성, 제공 여부 및 방법 등 심의 병원 내... 분당서울대학교병원 디지털헬스케어연구사업부 교수 보건복지부 의료질평가심의위원회 위원, 국가건강검진위원회 위원 한국의료분쟁조정중재원 조정위
VIEW
전체 블로그 카페
한국스마트헬스케어협회 2020.10.27.
의료 소프트웨어 접근과 보건의료 데이터 심의위원회 운영방안 심층세미나 개최
소프트웨어와 데이터 심의위원회 운영 - 빅데이터 및 인공지능(AI) 기술이 적용된 의료기기의 허가·심사 가이드라인 소개 - 의료 소프트웨어 기업의 GMP 접근 방법 - 병원에서의 데이터 심의위원회 운영 사례...
#의료소프트웨에 #보건의료 #데이터 #심의위원회 #빅데이터
IoMT 2020.10.27.
의료 소프트웨어 접근과 보건의료 데이터 심의위원회 운영방안 심층세미나 개최
소프트웨어와 데이터 심의위원회 운영 - 빅데이터 및 인공지능(AI) 기술이 적용된 의료기기의 허가·심사 가이드라인 소개 - 의료 소프트웨어 기업의 GMP 접근 방법 - 병원에서의 데이터 심의위원회 운영 사례 o...
#의료소프트웨어 #보건의료 #데이터 #심의위원회 #빅데이터

개인정보 가명·익명처리 기술 경진대회 후기



2021 PIDICON

개인정보 가명·익명처리 기술 경진대회

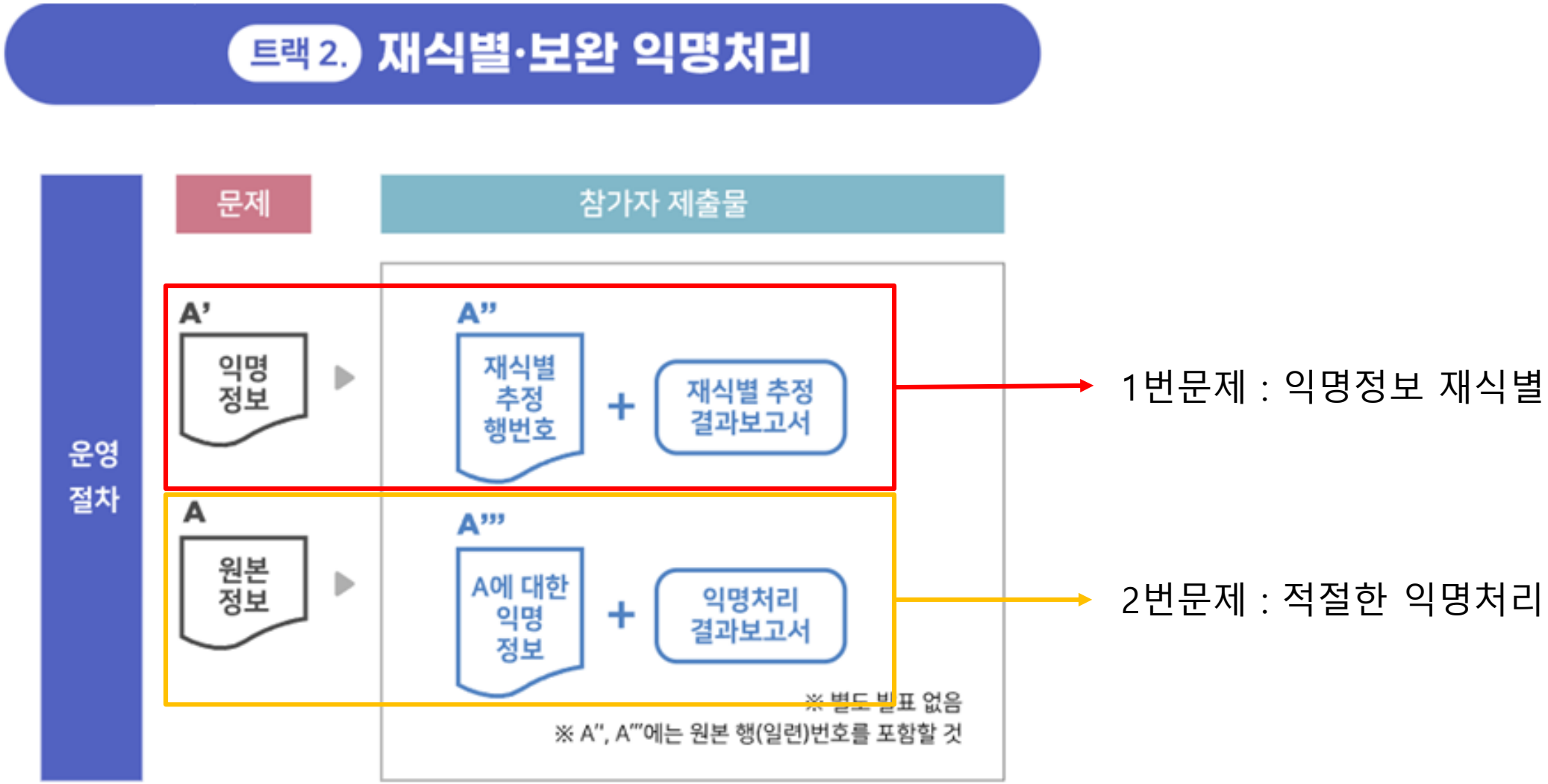
대회접수 2021.09.01.(수) ~ 09.30.(목)

기술경연 2021.11.08.(월) ~ 11.09.(화)

발표평가 2021.11.12.(금) ※ 트랙1만 진행

주최 과학기술정보통신부 주관 KISA 한국인터넷진흥원 운영기관 CM (주)컬처메이커스

운영 및 평가 절차


트랙 2. 재식별·보완 익명처리
운영 절차
문제
참가자 제출물
A'
익명 정보
A''
재식별 추정 행번호
+
재식별 추정 결과보고서
A
원본 정보
A'''
A에 대한 익명 정보
+
익명처리 결과보고서
※ 별도 빌표 없음
※ A'', A'''에는 원본 행(일련)번호를 포함할 것
1번문제 : 익명정보 재식별
2번문제 : 적절한 익명처리

운영 및 평가 절차

트랙 2. 재식별·보완 익명처리

평가 절차

1	재식별 처리의 우수성(정량적 평가)	50%
+		
2	보완 익명처리의 유용성(U1~U9, 정량적 평가)	25%
+		
3	보완 익명처리의 안전성(기술적, 정량적 평가)	25%
=		
4	최종 점수 산출	100%

※ A는 완전재현데이터, A'은 A를 익명처리한 정보로 대회 취지상 보완 익명처리를 위해 주최측이 제작한 다소 결함이 있는 익명정보

※ A"은 원본 행번호 + 재식별추정 행번호(매핑테이블), A""은 원본정보 A를 참가자들이 익명 처리한 정보

※ 제출물(4종) : 재식별 추정 행번호(매핑테이블 A"), 재식별 추정 결과보고서, 원본정보 A에 대한 익명정보 A"", 익명처리 결과보고서

평가기준1(재식별 처리의 우수성)

A : 원본데이터 셋

원본ID	이름	나이	신장	속성...
100001	곽두팔	102세	203cm	...
100002				
100003				
100004				
...				

A' : 익명처리 데이터 셋

익명ID	이름	나이	신장	속성...
200001				
200002				
200003	곽XX	100대	2미터 이상	...
200004				
...				

A" : 매핑테이블

원본ID	익명ID
100001	**200003**
100002	-
100003	-
100004	-
...	...

평가기준2(보완 익명처리의 유용성)

지표명	지표 설명
U1 : MA (Mean Attribute)	◦ 지정된 특정 속성들에 대한 평균값의 차이
U2 : MC (Mean Correlation)	◦ 지정된 2개 이상의 특정 속성쌍들에 대한 피어슨 상관계수에 대한 차이(평균절대오차)
U3 : CS (Cosine Similarity)	◦ 코사인 유사도로 원본과 비식별 동일 속성집합 간 벡터의 스칼라곱과 크기
U4 : NED_SSE (Normalized Euclidian Distance_Sum of Squared Errors)	◦ 정규화된 유클리디안 거리(NED, Normalized Euclidian Distance)를 이용한 제곱합오차(SSE, Sum of Squared Errors)
U5 : SED_SSE (Standardized Euclidian Distance_Sum of Squared Errors)	◦ 표준화된 유클리디안 거리(SED, Standardized Euclidian Distance)를 이용한 제곱합오차(SSE, Sum of Squared Errors)
U6* : MD_ECM (Mean Distribution Equivalence Class Metric)	◦ 동질집합별 속성들에 대한 평균 분포도(분산)
U7* : NA_ECSM (Normalized Average Equivalence Class Size Metric)	◦ 정규화된 동질집합들의 평균 크기
U8* : NUEM (Non-uniform Entropy Metric)	◦ 비균일 엔트로피 방법을 이용한 k-익명성 프라이버시 보호 모델에서의 정보손실 측도
U9 : AR (Anonymisation Ratio)	◦ 원본 데이터셋 대비 익명처리된 데이터셋의 정보량

평가기준3(보완 익명처리의 안전성)

F.1 안전성 평가기준

가명처리 | 식별가능성 | 복원가능성

익명처리 | 특정가능성 | 연결가능성 | 추론가능성

- 식별가능성 : 가명처리 된 데이터로부터 주어진 데이터셋만으로 정보 주체를 알아볼(식별할) 가능성
- 복원가능성 : 가명처리 과정에서 생성된 추가적인 정보(암호키, 매핑테이블 등)가 없는 상황에서 원본 데이터 주체를 복원해 낼(가명처리 이전으로 되돌릴) 가능성
- 특정가능성 : 데이터 주체를 고유 식별하기 위해 데이터 세트의 특성 집합을 관찰하여 해당 데이터 주체에 속한 레코드를 격리(isolation)해 낼 가능성
- 연결가능성 : 동일한 데이터 주체 혹은 데이터 주체 그룹과 관련된 레코드를 별도의 데이터 세트에 연결하여 정보 주체를 알아볼(식별할) 가능성
- 추론가능성 : 무시할 수 없는 확률로 다른 속성 집합의 값에서 속성의 값을 추론하여 정보 주체를 알아볼(식별할) 가능성

* 출처 : ISO/IEC 20889

Technique Name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of		
				Singling out	Linking	Inference
Statistical tools						
Sampling						
Aggregation	N.A.	Continuous, discrete	All attributes	Yes	Yes	Yes
Cryptographic tools	Yes					
Deterministic encryption	Yes	All	All attributes	No	Partially	No
Order-preserving encryption	Yes	All	All attributes	No	Partially	No
Homomorphic encryption	Yes	All	All attributes	No	No	No
Homomorphic secret sharing	Yes	All	All attributes	No	No	No
Suppression	Yes					
Masking	Yes	Categorical	Local identifiers	Yes	Partially	No
Local suppression	Yes	Categorical	Identifying attributes	Partially	Partially	Partially
Record suppression	Yes	N.A.	N.A.	Partially	Partially	Partially
Sampling	Yes	N.A.	N.A.	Partially[a]	Partially	Partially
Pseudonymization	Yes	Categorical	Direct identifiers	No	Partially	No
Generalization	Yes	All, subject to meaning	Identifying attributes			
Rounding	Yes	Continuous	Identifying attributes	No	Partially	Partially
Top/bottom coding	Yes	Continuous, ordinal	Identifying attributes	No	Partially	Partially
Randomization	No		Identifying attributes			
Noise Addition	No	Continuous	Identifying attributes	Partially	Partially	Partially
Permutation	No	All	Identifying attributes	Partially	Partially	Partially
Micro aggregation	No	Continuous	Indirect identifiers, and all attributes	No	Partially	Partially
Differential privacy	No	All	Identifying attributes	Yes	Yes	Partially
K-anonymity	Yes[b]	All	Quasi identifiers	Yes	Partially	No

[a] If the data principal record is not included in the sample.

[b] Unless *K*-anonymity is implemented using microaggregation.

문제1(익명정보 재식별 추정)

A : 원본데이터 셋

원본ID	복지주체ID	생년월일	관리행정지역코드	핸드폰번호	성별	이름	장애 유형코드	종합장애등급코드	자격개시일	장애인 보장구분코드	소득금액	결혼상태코드	결혼 여부	종교코드	종교	학력코드	학력	주거유형	개인근로 능력 유무	재산가액	소득산정코드
1000001	DG	[redacted]	4313012800	[redacted]	M	이[redacted]	10	20	19[redacted]	4	3[redacted]	0	0	2	개신교	1	고졸 미만	3	3	[redacted]	[redacted]
1000002	DG	[redacted]	1120012200	[redacted]	M	정[redacted]	50	10	19[redacted]	4	3[redacted]	0	0	2	개신교	2	고졸	6	9	[redacted]	[redacted]
1000003	DG	[redacted]	4148010700	[redacted]	F	안[redacted]	10	10	19[redacted]	4	2[redacted]	0	0	0	무교	3	대졸	2	1	[redacted]	[redacted]
1000004	DG	[redacted]	5011010900	[redacted]	F	유[redacted]	60	10	19[redacted]	4	2[redacted]	0	0	0	무교	2	고졸	6	9	[redacted]	[redacted]
1000005	DG	[redacted]	2717010300	[redacted]	F	김[redacted]	10	20	19[redacted]	4	5[redacted]	0	0	0	무교	2	고졸	3	3	[redacted]	[redacted]
1000006	DG	[redacted]	1150010900	[redacted]	M	한[redacted]	10	10	19[redacted]	4	4[redacted]	0	0	0	무교	1	고졸 미만	4	9	[redacted]	[redacted]
1000007	DG	[redacted]	1171010100	[redacted]	F	박[redacted]	10	20	19[redacted]	4	1[redacted]	5	0	0	무교	3	대졸	3	3	[redacted]	[redacted]
1000008	DG	[redacted]	4128510600	[redacted]	M	김[redacted]	30	10	19[redacted]	2	2[redacted]	0	0	0	무교	2	고졸	8	9	[redacted]	[redacted]
1000009	DG	[redacted]	4128710200	[redacted]	F	정[redacted]	80	10	19[redacted]	4	5[redacted]	0	0	1	불교	1	고졸 미만	3	9	[redacted]	[redacted]
1000010	DG	[redacted]	4117110100	[redacted]	M	김[redacted]	10	20	19[redacted]	4	2[redacted]	0	0	0	무교	2	고졸	1	3	[redacted]	[redacted]

A' : 익명처리 데이터 셋

가명ID	나이	관리행정지역코드	성별	이름	장애 유형코드	종합장애등급코드	자격개시일	장애인 보장구분코드	소득금액	현재 결혼 여부	종교코드	학력코드	주거유형	개인근로 능력 유무	재산가액	소득산정코드
5000001	30	41	F	ooo	10	20	1988	4	3~5백만원	0	0	고졸 이하	기타임대	근로능력 있음	1억~3억원	5
5000002	50	11	M	ooo	기타	10	1985	4	5백~천만원	0	0	고졸 이하	기타임대	근로능력 없음	3억~7억원	8
5000003	20	41	F	ooo	10	10	2007	4	1~2백만원	0	1	대졸 이상	자가	근로능력 없음	2천만원 미만	1
5000004	50	11	F	ooo	기타	10	1962	4	1~2백만원	0	0	고졸 이하	기타임대	근로능력 없음	3억~7억원	3
5000005	20	41	F	ooo	10	20	2011	4	1~2백만원	0	3	고졸 이하	기타임대	근로능력 있음	5천~1억원	2
5000006	20	44	F	ooo	기타	10	2005	1-3	2~3백만원	0	1	대졸 이상	기타임대	근로능력 없음	2천~5천만원	5
5000007	20	11	M	ooo	10	20	2009	4	3~5백만원	0	0	고졸 이하	기타임대	근로능력 있음	5천~1억원	3
5000008	50	26	M	ooo	기타	10	1973	1-3	2~3백만원	0	0	대졸 이상	기타임대	근로능력 없음	3억~7억원	3
5000009	60	41	M	ooo	기타	10	1975	4	1~2백만원	0	0	고졸 이하	기타임대	근로능력 없음	3억~7억원	3
5000010	30	11	M	ooo	30	20	2002	4	1~2백만원	0	0	고졸 이하	기타임대	근로능력 있음	1억~3억원	5

- **Question :** 익명처리된 데이터 셋의 재식별 방법?
- **Answer :**

1) 원본데이터를 어떤 방식으로 익명처리 했는지 확인 필요!
2) 수작업으로 식별 가능한 케이스 확인 후 속성 분류 확인!
3) 분류 작업 후 원본데이터 재구성!

문제1(익명정보 재식별 추정)

※ 속성 값 분류 추정 예시

A : 원본데이터 셋

복지주체ID	생년월일	관리행정지역코드	핸드폰번호	성별	이름	길이	장애 유형코드	종합장애등급코드	자격개시일	장애인 보장구분코드	소득금액	결혼상태코드	결혼 여부	종교코드	종교	학력코드	학력	주거유형	개인근로 능력 유무	재산가액	소득산정코드
[illegible]	1972-[illegible]	11[illegible]	[illegible]	M	[illegible]	2	10	10	1987-[illegible]	4	[illegible]	0	0	2	개신교	2	고졸	5	9	[illegible]	[illegible]
[illegible]	1956-[illegible]	44[illegible]	[illegible]	F	[illegible]	2	90	10	1975-[illegible]	4	[illegible]	0	0	0	무교	2	고졸	4	1	[illegible]	[illegible]
[illegible]	1969-[illegible]	41[illegible]	[illegible]	F	[illegible]	2	140	20	1971-[illegible]	4	[illegible]	0	0	0	무교	1	고졸 미만	6	3	[illegible]	[illegible]
[illegible]	1983-[illegible]	41[illegible]	[illegible]	M	[illegible]	2	10	10	1986-[illegible]	4	[illegible]	1	1	0	무교	2	고졸	6	9	[illegible]	[illegible]
[illegible]	1989-[illegible]	45[illegible]	[illegible]	F	[illegible]	2	10	20	1997-[illegible]	4	[illegible]	0	0	2	개신교	1	고졸 미만	4	3	[illegible]	[illegible]
[illegible]	1999-[illegible]	11[illegible]	[illegible]	M	[illegible]	2	30	20	2014-[illegible]	4	[illegible]	0	0	0	무교	1	고졸 미만	2	3	[illegible]	[illegible]
[illegible]	1985-[illegible]	41[illegible]	[illegible]	M	[illegible]	2	30	20	1991-[illegible]	4	[illegible]	0	0	3	천주교	2	고졸	1	3	[illegible]	[illegible]
[illegible]	1975-[illegible]	11[illegible]	[illegible]	F	[illegible]	2	10	20	1977-[illegible]	4	[illegible]	0	0	1	불교	3	대졸	6	3	[illegible]	[illegible]
[illegible]	1982-[illegible]	41[illegible]	[illegible]	F	[illegible]	2	10	20	1992-[illegible]	2	[illegible]	0	0	3	천주교	2	고졸	3	3	[illegible]	[illegible]
[illegible]	1989-[illegible]	11[illegible]	[illegible]	F	[illegible]	2	10	20	1998-[illegible]	3	[illegible]	0	0	0	무교	2	고졸	5	3	[illegible]	[illegible]

A' : 익명처리 데이터 셋

가명ID	나이	관리행정지역코드	성별	이름	장애 유형코드	종합장애등급코드	자격개시일	장애인 보장구분코드	소득금액	현재 결혼 여부	종교코드	학력코드	주거유형	개인근로 능력 유무	재산가액	소득산정코드
5350432	60	44	F	oo	기타	10	1975	4	3~5백만원	0	0	고졸 이하	자가	근로능력 없음	7억~10억원	3
5354450	20	27	F	oo	기타	10	2016	4	2~3백만원	0	0	고졸 이하	기타임대	근로능력 없음	2천만원 미만	8
5372874	20	41	F	oo	30	10	2007	4	3~5백만원	0	2	고졸 이하	자가	근로능력 없음	2천~5천만원	8
5408084	40	26	M	oo	10	10	1988	4	1~2백만원	0	2	고졸 이하	기타임대	근로능력 없음	1억~3억원	2
5408853	30	46	F	oo	10	20	2005	4	3~5백만원	0	1	고졸 이하	기타임대	근로능력 있음	1억~3억원	5
5422767	30	41	M	oo	60	20	2004	4	1~2백만원	0	0	고졸 이하	기타임대	근로능력 있음	1억~3억원	5
5435222	40	11	F	oo	10	20	1977	4	2~3백만원	0	1	대졸 이상	기타임대	근로능력 있음	3억~7억원	3
5453672	40	31	M	oo	60	10	1990	1-3	2~3백만원	1	2	대졸 이상	기타임대	근로능력 없음	1억~3억원	5
5487910	40	11	M	oo	기타	10	1991	4	5백~천만원	0	0	고졸 이하	기타임대	근로능력 없음	1억~3억원	2
5492837	20	11	M	oo	30	20	2014	4	1~2백만원	0	0	고졸 이하	기타임대	근로능력 있음	2천~5천만원	3

※ 확인 가능

① 주거유형 / A : 6 = A' : 기타임대

② 개인근로 능력 유무 / A : 3 = A' : 근로능력 있음

③ 종교코드 / A : 불교 = A' : 1

문제1(익명정보 재식별 추정)

※ 재구성한 데이터셋

☐ 익명처리 데이터셋화 완료한 컬럼 ☐ 분류 필요없는 컬럼

생년월일	만나이	나이(처리)	관리행정지역코드	지역코드(처리)	성별	이름	이름(처리)	장애유형코드	장애유형코드(처리)	중합장애등급코드	자격개시일	자격개시일(처리)
1980-09-10	41	40	4313012800	43	M	이	ooo	10	10	20	1990-08-08	1990
1973-09-21	48	40	1120012200	11	M	정	ooo	50	기타	10	1991-12-05	1991
1958-11-18	62	60	4148010700	41	F	안	ooo	10	10	10	1975-10-27	1975
1962-05-16	59	50	5011010900	50	F	유	ooo	60	60	10	1964-04-06	1964
1979-10-12	42	40	2717010300	27	F	김	ooo	10	10	20	1997-05-26	1997
1964-10-22	57	50	1150010900	11	M	한	ooo	10	10	10	1983-04-12	1983
1967-07-20	54	50	1171010100	11	F	박	ooo	10	10	20	1978-03-04	1978
1964-07-26	57	50	4128510600	41	M	김	ooo	30	30	10	1977-07-25	1977
1973-11-25	47	40	4128710200	41	F	정	ooo	80	기타	10	1990-12-16	1990
1983-03-14	38	30	4117110100	41	M	김	ooo	10	10	20	1999-09-28	1999

장애인보장구분코드	장애인보장구분코드	소득금액	소득금액(처리)	결혼여부	종교코드	종교코드	학력코드	학력	학력(처리)	주거유형	주거유형	개인근로능력유무	개인근로능력유무(처리)	재산가액	재산가액(처리)	소득산정코드
4	4	3	3-5백만원	0	2	2	1	고졸미만	고졸이하	3	기타임대	3	근로능력있음	2	1억-3억원	5
4	4	3	3-5백만원	0	2	2	2	고졸	고졸이하	6	기타임대	9	근로능력있음	3	3억-7억원	1
4	4	2	2-3백만원	0	0	0	3	대졸	대졸이상	2	기타임대	1	근로능력있음	6	3억-7억원	3
4	4	2	2-3백만원	0	0	0	2	고졸	고졸이하	6	기타임대	9	근로능력있음	5	3억-7억원	3
4	4	5	5백-천만원	0	0	0	2	고졸	고졸이하	3	기타임대	3	근로능력있음	2	1억-3억원	2
4	4	4	3-5백만원	0	0	0	1	고졸미만	고졸이하	4	자가	9	근로능력있음	5	3억-7억원	2
4	4	1	1-2백만원	0	0	0	3	대졸	대졸이상	3	기타임대	3	근로능력있음	4	3억-7억원	3
2	1-3	2	2-3백만원	0	0	0	2	고졸	고졸이하	8	기타임대	9	근로능력있음	5	3억-7억원	5
4	4	5	5백-천만원	0	1	1	1	고졸미만	고졸이하	3	기타임대	9	근로능력있음	3	3억-7억원	1
4	4	2	2-3백만원	0	0	0	2	고졸	고졸이하	1	자가	3	근로능력있음	2	1억-3억원	8

➔ 전체 컬럼에 대한 분류 기준 식별 완료

문제1(익명정보 재식별 추정)

A : 익명처리화 한 원본데이터 셋

원본ID	나이(처리)	지역(처리)	성별(처리)	이름(처리)	코드(처리)	자격(처리)	속성...
100001	60	11	M	OO	기타	1990	...
100002							
100003							
100004							
...							

→ **"6011MOO기타1990..."**

A' : 익명처리 데이터 셋

익명ID	나이	지역	성별	이름	코드	자격	속성...
200001							
200002							
200003	60	11	M	OO	기타	1990	...
200004							
...							

A" : 매핑테이블

원본ID	익명ID
100001	**200003**
100002	-
100003	-
100004	-
...	...

문제2(적절한 익명처리)

A : 원본데이터 셋(**50**만 건)

원본ID	복지주제ID	생년월일	관리행정지역코드	핸드폰번호	성별	이름	장애 유형코드	중한장애등급코드	자격개시일	장애인 보장구분코드	소득금액	결혼상태코드	결혼 여부	중고코드	중고	학력코드	학력	주거유형	개인근로 능력 유무	자산가액	소득산정코드
1000001		1980-	4313012800		M		10	20		4		0	0	2	개신교	1	고졸 미만	3	3		5
1000002		1973-	1120012200		M		50	10		4		0	0	2	개신교	2	고졸	6	9		1
1000003		1958-	4148010700		F		10	10		4		0	0	0	무교	3	대졸	2	1		3
1000004		1962-	5011010900		F		60	10		4		0	0	0	무교	2	고졸	6	9		3
1000005		1979-	2717010300		F		10	20		4		0	0	0	무교	2	고졸	3	3		2
1000006		1964-	1150010900		M		10	10		4		0	0	0	무교	1	고졸 미만	4	9		2
1000007		1967-	1171010100		F		10	20		4		5	0	0	무교	3	대졸	3	3		3
1000008		1964-	4128510600		M		80	10		2		0	0	0	무교	2	고졸	8	9		5
1000009		1973-	4128710200		F		80	10		4		0	0	1	불교	1	고졸 미만	3	9		1
1000010		1983-	4117110100		M		10	20		4		0	0	0	무교	2	고졸	1	3		8

- **Question :** 원본데이터 셋의 적절한 익명처리?
  - 성별/나이/지역/학력/장애유형에 따른 소득/재산가액/주거유형과의 연관관계 분석을 위함
- **Answer :**

1) 연구 대상과 목적에 따른 컬럼 분류!

2) 재식별 추정을 막기 위해 **1)**에 따라 컬럼 개수 최소화!

3) 라운딩 처리가 필요한 경우 구간 분류 기준 마련!

4) 특정가능성, 연결가능성, 추론가능성을 제거하기 위한 비식별처리 및 검증!

문제2(적절한 익명처리)

- 익명처리 결과보고서 작성

속성명	복지주체ID	생년월일	관리행정지역코드	핸드폰	성별	이름	장애유형코드	종합장애등급코드	자격개시일	장애인보장구분코드
개인정보 유형 코드	식별자	준식별자	준식별자	식별자	준식별자	식별자	준식별자	민감속성	준식별자	민감속성
세부 기술 적용 코드	삭제	상하단코딩	행 항목 삭제	삭제	-	삭제	범위방법	삭제	삭제	삭제
		일반라운딩	범위방법(앞두자리)							

속성명	소득금액	결혼상태코드	결혼여부	종교코드	종교	학력코드	학력	주거유형	개인근로능력유무	재산가액	소득산정코드
개인정보 유형 코드	민감속성	민감속성	비민감속성	민감속성	민감속성	민감속성	준식별자	민감속성	비민감속성	민감속성	민감속성
세부 기술 적용 코드	범위방법	삭제	삭제	삭제	삭제	삭제	범위방법	범위방법	삭제	행 항목 삭제	삭제
										일반라운딩	
비고	k-익명성 적용 1) 준식별자(나이+관리행정지역코드+성별+장애유형코드+학력) 동질집합으로 구성 2) k값 9이하 레코드 삭제(1502개)하여 k값 10확보 l-다양성 적용 - 민감속성(주거유형 l값 2 / 소득 l값 3 / 재산가액 l값 3) 확보										

문제2(적절한 익명처리)

- 익명처리한 데이터 셋의 안전성 검증

속성명	생년월일	관리행정지역코드	성별	장애유형코드	학력	소득금액	주거유형	재산가액
개인정보 유형 코드	준식별자	준식별자	준식별자	준식별자	준식별자	민감속성	민감속성	민감속성
세부 기술 적용 코드	상하단코딩	행항목삭제	-	범위방법	범위방법	범위방법	범위방법	행항목삭제
	라운딩	범위방법						라운딩

데이터셋	특정가능성	연결가능성	추론가능성
k익명성	1	0.5	0
상하단코딩	0	0.5	0.5
일반라운딩	0	0.5	0.5
행항목삭제	0.5	0.5	0.5
범위방법	@	@	@
합계	**1.5**	**2**	**1.5**

➔ 안전성 평가 기준 1이상 만족

Technique Name	Data truthfulness at record level	Applicable to types of values	Applicable to types of attributes	Reduces the risk of		
				Singling out	Linking	Inference
Statistical tools						
Sampling						
Aggregation	N.A.	Continuous, discrete	All attributes	Yes	Yes	Yes
Cryptographic tools	Yes					
Deterministic encryption	Yes	All	All attributes	No	Partially	No
Order-preserving encryption	Yes	All	All attributes	No	Partially	No
Homomorphic encryption	Yes	All	All attributes	No	No	No
Homomorphic secret sharing	Yes	All	All attributes	No	No	No
Suppression	Yes					
Masking	Yes	Categorical	Local identifiers	Yes	Partially	No
Local suppression	Yes	Categorical	Identifying attributes	Partially	Partially	Partially
Record suppression	Yes	N.A.	N.A.	Partially	Partially	Partially
Sampling	Yes	N.A.	N.A.	Partially[a]	Partially	Partially
Pseudonymization	Yes	Categorical	Direct identifiers	No	Partially	No
Generalization	Yes	All, subject to meaning	Identifying attributes			
Rounding	Yes	Continuous	Identifying attributes	No	Partially	Partially
Top/bottom coding	Yes	Continuous, ordinal	Identifying attributes	No	Partially	Partially
Randomization	No		Identifying attributes			
Noise Addition	No	Continuous	Identifying attributes	Partially	Partially	Partially
Permutation	No	All	Identifying attributes	Partially	Partially	Partially
Micro aggregation	No	Continuous	Indirect identifiers, and all attributes	No	Partially	Partially
Differential privacy	No	All	Identifying attributes	Yes	Yes	Partially
K-anonymity	Yes[b]	All	Quasi identifiers	Yes	Partially	No

[a] If the data principal record is not included in the sample.

[b] Unless *K*-anonymity is implemented using microaggregation.

문제2(적절한 익명처리)

- 익명처리 데이터셋(최종)

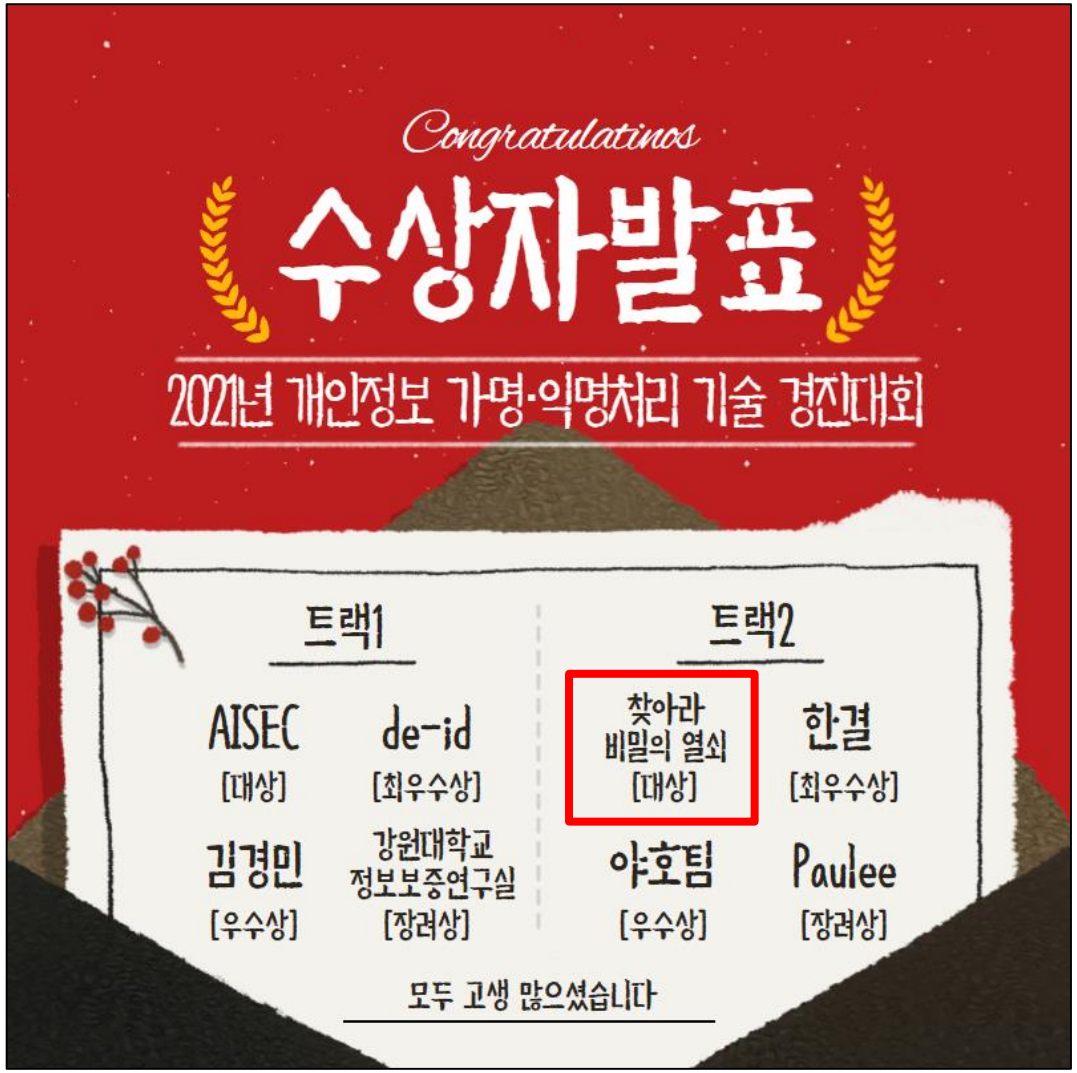
원본ID	나이	관리행정지역코드	성별	장애 유형코드	학력	소득금액	주거유형	재산가액
1000001	40	43	M	신체외부(지체)	0	3백만원5백만원	기타	250000000
1000002	50	11	M	발달장애	0	3백만원5백만원	월세	350000000
1000003	60	41	F	신체외부(지체)	1	2백만원3백만원	기타	650000000
1000004	60	50	F	신체외부(기타)	0	2백만원3백만원	월세	600000000
1000005	40	27	F	신체외부(지체)	0	5백만원천만원	기타	250000000
1000006	60	11	M	신체외부(지체)	0	3백만원5백만원	자가	550000000
1000007	50	11	F	신체외부(지체)	1	백만원2백만원	기타	450000000
1000008	60	41	M	신체외부(청각)	0	2백만원3백만원	기타	550000000
1000009	50	41	F	정신장애	0	5백만원천만원	기타	350000000
1000010	40	41	M	신체외부(지체)	0	2백만원3백만원	자가	200000000
1000011	50	41	F	신체외부(지체)	1	3백만원5백만원	자가	350000000
1000012	60	28	M	발달장애	0	백만원2백만원	자가	550000000
1000013	60	41	M	신체외부(지체)	0	2백만원3백만원	월세	650000000
1000014	50	29	M	신체외부(지체)	1	2백만원3백만원	월세	350000000
1000015	50	11	M	신체외부(기타)	0	2백만원3백만원	기타	300000000
1000016	70	11	F	신체외부(지체)	1	3백만원5백만원	월세	1500000000
1000017	30	41	F	신체외부(지체)	0	3백만원5백만원	전세	100000000
1000018	60	43	F	발달장애	1	2백만원3백만원	기타	600000000
1000019	40	41	F	발달장애	0	5백만원천만원	기타	200000000
1000020	60	41	M	신체외부(청각)	0	3백만원5백만원	기타	700000000

* 출처 : 보건복지부고시 장애등급판정기준

장애 유형코드		
10	지체	신체외부(지체)
30	청각	신체외부(청각)
20	시각	신체외부(기타)
40	언어	
60	뇌병변	
130	안면	
50	지적	발달장애
70	자폐성	
80	정신	정신장애
90	신장	신체내부
100	심장	
110	호흡기	
120	간	
140	장루,요루	
150	간질	

주거유형	
자가	자가
기타자가인정	
전세	전세
월세	월세
미등기,무허가주택소유	기타
영구임대주택	
부분무료임차	
전체무료임차	

수상



의학신문 2021.12.20

분당서울대병원, '개인정보 가명·익명처리 기술 경진대회' 대상

의학전문매거진 MD POST 엠디포스트 2021.12.20

분당서울대병원 정보보호팀, 개인정보 가명 · 익명처리 기술 경진대회 대상 수상

- 재식별, 보완 및 익명처리 기술경연 참가해 최고 점수로 대상 수상 -

헬스미디어 2021.12.20

분당서울대병원 정보보호팀, 경진대회 대상 수상

개인정보 재식별·보완, 익명처리 기술 경연에 참가해 한국인터넷진흥원장상 수상

중부일보 2021.12.20

분당서울대병원, 과기부 주관 개인정보 가명·익명처리 기술 경진대회 대상 수상

마치며..

감사합니다