

# 정책연구용역사업 최종결과보고서

(뒷면)

(측면)

(앞면)

<div>주 의 ( 주 의 내 용 기 재 )  (훈글 14 point 고딕체)</div>	국 민 건 강 영 양 조 사 — 사 망 원 인 통 계  연 계 자 료  개 인 정 보  영 향 평 가  2 0 1 9  질 병 관 리 본 부	<div><div>발 간 등 록 번 호 00-0000000-000000-00 정책연구용역사업 최종결과보고서</div> <div>국민건강영양조사-사망원인통계 연계자료 개인정보 영향평가</div> <div>Disclosure Risk Assessment for linking KNHANES microdata to Cause of Death microdata</div> <div>주관연구기관 : 고려대학교 산학협력단</div> <div>질병관리본부</div></div>
---	---	---

※ 주의 내용

주 의

1. 이 보고서는 질병관리본부에서 시행한 정책연구용역사업의 최종결과보고서입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 질병관리본부에서 시행한 정책연구용역사업의 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 아니 됩니다.

## 정책연구용역사업 최종결과보고서

연구사업명	국민건강영양조사-사망원인통계 연계자료 개인정보 영향평가		
발주부서	건강영양조사과	과제담당관	오경원
주관연구 기관	기관명	소 재 지	대 표
	고려대학교 산학협력단	서울	허준
책임연구원	성 명	소속 및 부서	직위/전공
	안형진	고려대학교 의과대학 의학통계학교실	교수/의학통계
총 연구기간	2019.10.28 - 2020.01.27	총 연구비	20,000천원
당해연도 연구기간	2019.10.28 - 2020.01.27	당해연도 연구비	20,000천원
보안 여부	보안( ), 일반(✓)	결과 공개 여부	가( ), 부(✓)
연구참여자	총 7명 [책임연구원 1명, 연구원 0명, 연구보조원 6명, 보조원 0명]		
세부사업 여부	해당( ), 해당없음( ) (해당사항 ✓표기)	세부사업 수	총 개

2019년도 정책연구용역사업의 최종결과보고서를 붙임과 같이 제출합니다.

붙임1. 최종결과보고서 제본(발주부서에서 요구한 부수: 붙임1 엑셀파일)

2. CD 2매

2020 년 1 월 23 일

책임연구원 안 형 진 (인 또는 서명)  
주관연구기관장 허 준 (직인)

질병관리본부장 귀하

# 목 차

## I. 연구결과 요약문

(한글) 국민건강영양조사-사망원인통계 연계자료 개인정보 영향평가

(영문) Disclosure Risk Assessment for Linking KNHANES Microdata to Cause of Death Microdata

제 1장 최종 목표 .....	10
제 1절 목표 .....	10
제 2절 목표달성도 및 관련분야에 대한 기여도 .....	11
1. 목표 달성도 .....	11
2. 기대효과 .....	12
제 2장 최종 정책연구용역사업 내용 및 방법 .....	13
제 1절 정보의 노출 .....	13
1. 개인정보의 노출 .....	13
2. 노출위험의 요인 .....	13
제 2절 노출위험의 측정 .....	14
1. 유일성에 근거한 노출위험 사례 .....	14
2. 유일성에 근거한 노출위험에 영향을 주는 요인 .....	14
3. 노출위험 확률 추정 .....	17
제 3절 노출제한 방법 .....	20
1. 표본추출 .....	20
2. 자료 감추기 .....	20
3. 재범주화 .....	21
4. 자료 변조 방법 .....	22
제 4장 연구결과 고찰 및 결론 .....	95
제 5장 연구성과 및 활용계획 .....	97
제 6장 정책연구용역사업 진행과정에서 수집한 해외과학기술정보 .....	99
제 7장 기타 중요변경사항 .....	99
제 8장 연구비 사용 내역 및 연구원 분담표 .....	100
제 9장 참고문헌 .....	102

## 보고서 요약문

연구사업명	국민건강영양조사-사망원인통계 연계자료 개인정보 영향평가		
색인어	국민건강영양조사, 개인정보, 정보노출, 사망		
주관연구기관	고려대학교 산학협력단	책임연구원	안형진
연구기간	2019.10.28 - 2020.01.27		
<p>현재 국민건강영양조사에서 제공하고 있는 원시자료는 다양한 건강결과를 포함하고 있지만 가장 핵심적인 건강결과인 사망 및 그 원인에 대한 자료는 포함하고 있지 않았다. 최근 많은 연구자들의 요청에 의해 국민건강영양조사 자료와 통계청 사망원인 자료와의 연계 및 일반 공개 방안을 고려하고 있다. 하지만, 개인정보 보호의식 강화에 따라 조사참여자의 개인정보 노출에 대한 위험성이 함께 제기되고 있다. 따라서 본 연구는 연계자료의 공개 이전에 자료의 개인의 노출 위험성을 조사하고 평가한다. 또한 이 결과를 근거로 개인정보 노출위험을 제어하는 방안을 적용하여 일반 공개가 가능한 연계자료를 구축한다.</p> <p>공개할 고려하고 있는 연계자료에는 원시자료에 포함된 일부 변수들과 통계청 사망원인 자료인 사망여부, 사망원인 변수들로 구성되어 있다. 먼저 개인노출에 이용될 수 있는 인구학적인 주요변수를 선정하고, 표본 유일성에 대한 분석을 실시하였다. 이 결과 연령 변수의 경우 연속형 변수를 그대로 공개하는 경우 빈도가 낮은 고령자들의 개인노출의 위험이 높으므로 토크코딩 방법이나 재범주화 방법을 이용하여 재식별화의 위험을 제어하는 것이 필요하다. 그리고 인구학적 주요변수들에 사망여부, 사망원인 변수들을 하나 씩 포함시켜 주요변수들의 조합에서 유일성 분석을 시행하고 개인 및 파일수준의 노출위험 확률을 계산하고 비교하였다. 이 결과 사망원인의 범주를 3개 이상 공개할 경우 사망자의 유일성이 상당히 높아져서 개인정보의 노출위험이 커졌다. 기 실시한 「국민건강영양조사 원시자료 개인정보 영향평가」의 결과와 공개자료와의 노출 위험성을 비교하였으며, 유병률이 작은 질병을 가지고 있는 사람들은 대부분이 표본 유일성을 보여 개인의 재식별화 위험이 높았다. 마지막으로 인구지리학적인 변수를 주요변수로 선정하여 조사 기수에 따라 개인정보 노출위험을 평가하여, 지리학적인 정보가 개인의 재식별화에 어떤 영향을 미치는지 확인하였다. 그 결과 거주 시도 정보를 포함하는 경우 사망자에서 표본 유일성이 매우 높았다.</p> <p>해당 연계자료를 현 수준에서 공개하면 정보의 노출로 인한 개인정보 침해의 문제를 야기할 가능성이 높다. 따라서 연계자료를 공개하기 위해서는 이용자의 무분별한 접근을 통제하기 위한 방안을 마련하고, 개인노출의 위험성이 일정수준 이상인 경우에 주요변수를 숨기거나 보정하는 등의 정보보호 조치를 적용하여 개인정보의 노출 위험을 최소화하여야 한다.</p>			

## Summary

Title of Project	Disclosure Risk Assessment for linking KNHANES microdata to Cause of Death microdata		
Key Words	KNHANES, microdata, disclosure, cause of death, confidentiality, masking		
Institute	Research & Business Foundation of Korea University	Project Leader	Hyonggin An
Project Period	2019.10.28 - 2020.01.27		
<p>The Korea National Health and Nutrition Examination Survey (KNHANES) data including a variety of health outcomes have been provided for public use, but information about mortality and causes of death are not included. Recently, at the request of many researchers, linking KNHANES data to cause of death microdata is being considered. However, there is a concern about the risk of re-identification of participants. In this study, we investigate and evaluate the disclosure risk for linking KNAHNES data to cause of death microdata provided by Statistics Korea. Besides, we apply statistical disclosure limitation methods to the linked data and evaluate the methods.</p> <p>The linked data consist several variables selected from KNHANES raw data and limited causes of death information. First, we select demographic variables as the key variables for risk evaluation and calculate the proportion of sample uniqueness. The disclosure risk for elders increase when age is included in the key variables, so it is recommended to apply top coding or re-categorization method. We include mortality and causes of death in the key variables and calculate the proportion of sample uniqueness and probability of re-identification. As a result, if cause of death is categorized with three levels or more, the proportion of uniqueness and disclosure risk among the deceased significantly increase. We compare the disclosure risk of 「Disclosure Risk Assessment for Microdata of KNHANES」 studied before to that of linked data. Most people diagnosed with specific diseases are unique in the sample. We also include a geographic variable categorized as city/province in the key variables and calculate disclosure risk. We investigate the effect of geographic information on re-identification.</p> <p>In conclusion, the current disclosure risk of the linked data may lead to invasion of privacy. Therefore, it is advised to make proper process for public use of the linked data and to apply masking methods such as suppression and recoding to key variables.</p>			

## 정책연구용역사업 연구결과

### 제 1장 최종 목표

#### 제 1절 목표

국민건강영양조사는 보건 정책과 사업이 효과적으로 전달되고 있는지를 평가하는데 필요한 통계를 산출하기 위하여 『국민건강증진법』 제 16조에 근거하여 실시하고 있다. 국민의 건강 및 영양 상태에 관한 현황 및 추이를 파악하여 정책적 우선순위를 두어야 할 건강취약 집단을 선별하기 위해 국민의 건강수준, 건강관련 의식 및 형태, 식품 및 영양섭취 실태에 대해 조사한다. 국민건강영양조사는 제 1기(1998)부터 제 3기(2005)까지 2~3개월 단기조사체제로 실시하였으나 제 4기(2007~2009)부터 연중조사체제로 개편됨에 따라 3개년도가 전국을 대표하는 확률표본이 될 수 있도록 독립적인 순환표본조사(Rolling Sampling Survey)방식을 도입하여 조사하고 있다.

조사된 원시자료는 『통계법』 제 31조 및 『개인정보보호법』 제 18조 제 2항에 근거하여 개인별 자료를 공공이용자에게 선별적으로 공개하고 있다. 공개된 자료는 정책 및 학술 연구에 중요한 자료원으로 이용되었다. 연구의 결과물은 국민의 건강위험행태를 모니터링하고 국민건강증진의 목표지표 및 평가근거로 사용되었으며 의학 및 보건학 연구의 질을 향상시키는 역할을 하였다.

현재까지 조사된 국민건강영양조사는 다양한 건강결과를 포함하고 있지만 가장 핵심적인 건강결과인 사망 및 그 원인에 대한 자료는 포함하고 있지 않다. 많은 연구자이 국민건강영양조사 자료와 통계청 사망원인 자료와의 연계를 요청하였다. 이에 부응하여 최근 질병관리본부와 통계청은 국민건강영양조사와 사망원인통계의 연계체계를 마련하고 「국민건강영양조사-사망원인통계 연계자료」의 구축 및 일반 공개방안을 고려하고 있다.

「국민건강영양조사-사망원인통계 연계자료」의 공개는 국민건강증진이라는 대의의 목적을 가지고 있다. 그러나 정보공개를 통하여 개인식별이 가능하게 된다면 생존자의 경우 건강관련 정보가 파악이 될 것이므로 사생활 및 인권침해의 위험이 있고, 사망자의 경우 사망자 및 그 가족의 사생활 침해로 이어질 수 있다. 따라서 연계자료의 공개는 신중해야 하며 개인을 식별할 수 있는 정보를 파악하고 보정하는 등의 정보보호(또는 비밀보호)의 조치가 공개 이전에 이행되어야 한다.

현재 국민건강영양조사에서 제공하고 있는 원시자료는 원자료에서 개인식별자(이름, 주소, 생년월일 등) 및 변수간 결함을 통해 노출위험을 높이는 정보 제거 후 공개하고 있다. 또한 공개된 자료 분석 시 조사참여자 개인의 재식별화(re-identification) 가능성에 대한 평가를 실시하여 나이 관련 변수에는 탑코딩(top-coding)을 적용하고 가구총소득 변수에는 탑코딩과 바텀코딩(bottom-coding)을 적용하여 공개하고 있다. 그 외에 가구원수, 주택형태 등의 변수에는 재범주화를 적용하여 비식별조치를 강화하였다. 그러나 사망원인통계 자료를 연계한 후에 정보 노출의 위험정도를 평가한 연구는 없다. 따라서 「국민건강영양조사-사망원인통계 연계자료」의 개인 노출 위험성을 통계적인 방법으로 면밀히 조사하여 개인의 재식별 위험정도를 평가하는 것이 본 연구의 첫 번째 목적이다.

개인의 노출 위험 정도의 평가결과를 바탕으로 노출의 위험성이 있는 경우에는 노출의 위험을 최소화하기 위한 통제방법을 적용하여 자료를 재구축하여야 한다. 만일, 특정변수가 노출의 위험을 높인다면 이를 제거하거나 변경하는 등의 방법을 적용하여야 할 수 있다. 하지만 제공하는 자료의 비밀보호 조치를 너무 강조하여 위험통제 방법을 적용하게 되면 사용자에게 제공하는 자료에 제약이 너무 많아 자료 제공으로부터 기대되는 효과를 얻지 못할 수 있다. 따라서 정보보호 조치는 노출 위험을 최소화하는 동시에 원자료의 정보를 이용하여 얻는 공익적 목적을 함께 고려하여 효과적으로 수행하고 적용해야 할 것이다. 따라서 본 연구의 두 번째 목적은 국민건강영양조사 개인정보 노출위험의 제어방안을 적용하고 평가하여 일반 공개가 가능한 연계자료의 공개 범위를 제시하는 것이다. 이를 위해 질병관리본부 및 전문가와의 자문회의를 개최하여 도출된 방법이나 결과에 수정이 필요한 경우는 의견을 반영하여 수정 및 보완하였다.

## 제 2절 목표달성도 및 관련분야에 대한 기여도

### 1. 목표 달성도

<표 1> 연구사업내용 및 추진 일정

연구내용	연구기간			비고
	1	2	3	
연계자료의 개인정보 노출위험 평가				
연계자료 개인정보 노출위험 제어방법 선정				
일반 공개가 가능한 연계자료 구축				
최종 결과보고서 작성				

－ 연계자료의 개인정보 노출위험 평가 (100% 완료)



- 개인정보 노출위험 제어방법 선정 (100% 완료)
- 일반 공개가 가능한 연계자료 구축 (100% 완료)
- 최종 결과보고서 작성 (100% 완료)

## 2. 기대효과

국민건강영양조사 및 통계청의 사망원인 자료를 연계하였을 때 개인정보 노출위험 제어방법을 적용함으로써 개인정보를 보호할 수 있으며, 연계자료 제공에서 발생할 수 있는 법적인 문제들을 미연에 방지할 수 있다. 또한 자료제공자의 신뢰성을 쌓아 조사의 참여율을 높이고 탈락률을 줄일 수 있어 질 높은 조사결과를 확보할 수 있으므로 자료의 활용성 및 연구가치를 높일 수 있다.

## 제 2장 최종 정책연구용역사업 내용 및 방법

### 제 1절 정보의 노출

#### 1. 개인정보의 노출

표본조사 또는 전수조사에서 얻어진 마이크로자료를 대중에게 공개하는 것은 학술 연구와 정부의 정책개발 등에 활용 가능하다는 점에서 매우 중요한 가치를 가지고 있다. 그러나 조사한 내용을 가감없이 그대로 공개하는 것은 조사에 응한 응답자들의 개인정보가 유출되는 문제가 발생할 수 있다.

자료노출 시도자가 개인정보를 노출시키고자 할 때 가장 간단하게 적용하는 방법은 조사된 주요변수(key variable)의 조합을 이용하여 외부 자료와의 연계를 시도하는 것이다. 보통 주요변수는 조사단위의 인구지리학적 정보를 가지고 있는 변수들로 이루어진다. 조사단위가 개인인 경우 주요변수는 성별, 나이, 거주지역, 거주형태를 나타내는 변수들로 이루어지며 조사단위가 가구인 경우에는 해당지역, 가구원수, 주거형태 등 가구의 특성을 나타내는 변수로 구성된다.

최근 개인의 사생활이 중시되고 있는 환경에서 공개자료에 적절한 비밀보호 조치를 취하는 것은 개인의 사생활 보호를 위해서 반드시 필요한 과정이다. 개인의 사생활 보호가 보장된다면 응답자들의 조사 참여율 또한 높일 수 있을 것이다.

#### 2. 노출위험의 요인

개인의 정보가 노출 되는 것은 대부분 공개자료와 자료노출 시도자가 가지고 있는 외부 자료를 연결하여 공개된 자료에 포함된 대상의 민감한 정보를 알아내려고 할 때 발생한다. 자료가 공공에게 공개되었을 때 개인 노출의 위험성의 정도는 여러 가지 요인에 의하여 결정된다. 다음은 개인 노출 위험을 결정하는 중요한 요소들이다.

- 자료노출 시도자의 종류와 의도
- 자료노출 시도자가 가지고 있는 자료의 종류와 품질(quality)
- 자료노출 시도자의 자료와 공개된 자료의 유사성
- 공개된 자료의 정확성(또는 불변성) 및 측정오차(measurement error)

## 제 2절 노출위험의 측정

마이크로자료에서 개인 식별의 위험은 자료에 포함된 주요변수의 특성이나 주요변수의 조합을 이용해 측정할 수 있다. 특히 주요변수 중 어떤 특성을 갖는 개인이 유일하거나 주요변수의 조합 중 어떤 조합을 갖는 개인이 유일해지는 경우 '유일성(uniqueness)'이 있다고 한다. 유일성은 마이크로자료에서 노출위험을 이해하는 중요한 개념이다. 만약 모집단이 주어진다면 단순히 모집단에서 주요변수의 조합 중 유일한 개인을 포함하는 조합의 비율을 계산하여 노출위험을 평가할 수도 있으며, 이는 매우 직관적이고 단순하다는 장점이 있다.

공개자료에서의 개인의 노출은 공공에게 제공되는 자료와 자료노출 시도자가 가지고 있는 자료를 연계하는 과정에서도 발생할 수 있다. 공공에게 제공되는 공개자료에 있는 특정 변수의 값이 드물거나 유일한 경우, 외부의 자료에 동일한 조사단위가 포함되어 있다면 두 자료에 있는 동일한 조사단위를 정확하게 연결시킬 수 있다. 공개자료와 외부의 자료가 정확하게 연계되는 경우, 개인정보가 유출되고 사생활 침해의 문제를 야기할 수 있다. 만약 개인이 어떤 변수들의 특성에 의해 유일해지는지를 파악할 수 있다면 해당 변수를 제외하고 자료를 공개하거나 제 3절에서 설명할 노출제한 방법 등을 적용하여 노출을 제어할 수도 있다.

### 1. 유일성에 근거한 노출위험 사례

가상의 사례로 나이가 100세인 사람이 전국에서 유일하고, 이 사람의 개인 특성 및 민감한 정보를 모두 조사했다고 가정하자. 이 자료를 아무런 조치 없이 공개한다면, 외부의 연구자는 이 사람의 민감한 정보를 아무런 제약 없이 얻을 수 있다. 이렇게 자료에서 개인이 유일한 경우 개인이 식별될 가능성이 매우 높으며 노출위험이 높다고 한다. 주어진 자료에서 유일성은 하나의 변수 뿐만 아니라 여러 개의 변수를 조합하여 파악할 수도 있다. 앞선 가상의 사례에서 100명 중 나이가 60세인 사람이 3명이 있다면 이 사람들은 유일하지는 않다. 그러나 나이와 직업 변수를 함께 고려하여 유일성을 평가한다면 나이가 60세이면서 자영업업을 하는 사람은 전국에서 유일한 사람일 수 있다. 유일성을 평가할 때 개인을 식별할 수 있는 특성을 갖는 변수가 많아질수록 자료에서 유일성을 갖는 사람들의 비율은 높아지게 된다.

### 2. 유일성에 근거한 노출위험에 영향을 주는 요인

어떤 조사단위가 모집단에서 유일한 경우인 모집단 유일성에 근거한 노출위험은 다음 세 가지 조건을 모두 만족하는 경우에 발생한다.

첫째, 어떤 조사단위가 주요변수들의 조합에 대해 모집단에서 유일하다.

둘째, 그 조사단위는 공공에게 공개되는 마이크로자료에 포함되어 있다.

셋째, 그 조사단위는 자료노출 시도자가 가지고 있는 다른 자료에도 포함되어 있다.

어떤 조사단위가 위의 조건에서 언급한 자료들 중 어느 하나에도 나타나지 않는다면 정보의 노출은 일어나지 않는다. 한편 위의 노출의 발생 조건 하에서 노출위험은 확률적 모형으로 표현될 수 있으며, 이를 위해 다음과 같이 기호를 정의하기로 한다.

$A$  : 관심의 대상인 조사단위

$S_1$  : 마이크로자료(공개자료)

$S_2$  : 자료노출 시도자가 가지고 있는 다른 자료

$U_p$  : 모집단에서 유일한 응답 단위들의 모임

$U_s$  : 표본에서 유일한 응답 단위들의 모임

조사과정에서 측정오차가 없다고 가정하고, 모집단에서 유일한 조사단위  $A$ 가 표본(공개자료)에도 있다면 당연히 표본에서도 유일한 조사단위일 것이다. 즉, 조사단위  $A$ 가 모집단에서도 유일하고 표본에도 있을 확률은 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\Pr[(A \in U_s) \cap (A \in U_p)] &= \Pr(A \in U_p) \Pr[(A \in U_s) | (A \in U_p)] \\ &= \Pr(A \in U_p)\end{aligned}$$

노출위험은 자료노출 시도자가 외부자료  $S_2$ 에 특정 조사단위  $A$ 가 포함된 것을 모르고 있는 경우와 알고 있는 경우로 구분하여 계산할 수 있다.

– 자료노출 시도자가 외부자료  $S_2$ 에 특정 조사단위  $A$ 가 포함된 것을 모르는 경우

만약 의료 서비스 기관 또는 마케팅 회사와 같은 자료노출 시도자가 그들이 가지고 있는 자료  $S_2$ 에 특정 조사단위  $A$ 가 포함되어 있다는 것을 모르고 있다면, 위의 관계식으로부터  $A$ 의 노출위험  $DR(A)$ 는 다음과 같이 정의할 수 있다.

$$DR(A) = \Pr[(A \in S_1) \cap (A \in S_2) \cap (A \in U_p)]$$

그리고 조건부 확률공식과 모집단에서 유일한 조사단위  $A$ 가 표본으로 추출되거나 추출되지 않는 사건이 독립이라는 사실을 이용하여 위의 식은 다시 다음과 같이 표현할 수 있다.

$$\begin{aligned} DR(A) &= \Pr[(A \in S_1) \cap (A \in S_2) | (A \in U_p)] \Pr(A \in U_p) \\ &= \Pr[(A \in S_1) \cap (A \in S_2)] \Pr(A \in U_p) \end{aligned}$$

만약 조사단위  $A$ 가 공개자료  $S_1$ 에 존재하는 사건과 자료노출 시도자가 가지고 있는 자료  $S_2$ 에 존재하는 사건이 서로 독립이면, 조사단위  $A$ 가 공개자료와 외부의 자료에 모두 있을 확률은  $\Pr[(A \in S_1) \cap (A \in S_2)] = \Pr(A \in S_1) \times \Pr(A \in S_2)$ 이 된다.

– 자료노출 시도자가 외부자료  $S_2$ 에 특정 조사단위  $A$ 가 포함된 것을 아는 경우

만약 자료노출 시도자가 가지고 있는 다른 자료  $S_2$ 에 관심 대상인 조사단위  $A$ 가 포함되어 있다는 것을 알고 있다고 한다면  $\Pr(A \in S_2) = 1$ 이 될 것이다. 따라서 노출위험은 다음과 같이 된다.

$$DR(A) = \Pr(A \in S_1) \Pr(A \in U_p)$$

조사단위  $A$ 가 표본으로 추출되는 확률  $\Pr(A \in S_1)$ 는 조사 단계에서 정해져 있으므로, 노출위험  $DR(A)$ 을 측정하는 것은  $\Pr(A \in U_p)$ 를 어떻게 결정하는지에 따라 달라진다. 즉, 외부자료  $S_2$ 에 조사단위  $A$ 가 포함된 것을 알고 있었을 때  $A$ 가 노출될 위험은 모집단에서  $A$ 가 유일할 확률을 추정하는 문제라고 할 수 있다.

자료의 유일성을 평가하는 주요변수의 개수가 늘어나거나 범주의 개수가 늘어나면 주요변수들의 조합의 개수가 급격히 증가하고, 특정 조합에 해당하는 조사단위가 하나만 존재하는 일이 더욱 쉽게 발생한다. 다시 말해 변수의 개수나 범주가 늘어나면 자료에서 유일성을 가지는 개체들이 증가할 확률이 높아진다. 따라서 국민건강영양조사에서 유일성에 근거한 노출위험을 낮추는 방법은 여러 가지 주요변수에 의한 조합에서 유일하게 나오는 조사단위의 수를 낮추는 것이다. 유일성에 의한 노출위험을 축소하는 방법은 일반적으로 변수의 가능한 범주를 축소하는 재범주화(recoding 또는 grouping) 방법이 주로 사용된다.

### 3. 노출위험 확률 추정

노출위험을 측정하는 방법은 크게 두 가지로 나누어 설명할 수 있다. 첫 번째는 조사단위 수준의 노출위험을 측정하는 것이고, 두 번째는 조사단위를 모두 포함한 파일 수준의 노출위험을 측정하는 것이다.

#### 가. 조사단위(개체)의 노출위험 측정

자료를 공공에게 공개하기에 앞서 노출위험이 높은 개체가 많은 자료는 자료노출 시도자가 외부의 자료와 연결하여 민감한 개인정보를 찾아낼 수 있는 가능성이 있다. 따라서 자료 공개 이전에 조사단위의 노출위험을 추정하고 평가하여 노출위험이 높은 개체가 많은 자료에 매스킹 방법들을 적용할 필요가 있다.

개인단위의 노출위험은 자료노출 시도자가 가진 모집단의 정보  $P$ 와 공개자료가 가진 정보  $s$ 가 주어진 경우 조사단위  $i$ 를 모집단에 속한 단위  $i^*$ 와 정확하게 연결하는 확률로 추정할 수 있다.

$$\rho_i = \Pr(\text{공개자료의 조사단위 } i \text{와 모집단의 단위 } i^* \text{를 정확하게 연계} | s, P)$$

위와 같이 정의되는 개인의 노출위험은 실제로 정확하게 측정하기 어렵다. 자료노출 시도자가 가진 정보의 정확성이 낮을 수 있으며 공개자료의 값이 측정오차(measurement error)에 의해 오염되었을 수 있기 때문이다. 따라서 개인정보의 노출위험을 측정하는 방법은 모집단에 대한 외부정보가 정확하고 조사자료에 측정오차가 없는 경우에 자료노출 시도자가 모든 조사단위를 연계하려고 의도한 최악의 경우를 가정하고 노출위험 확률  $\rho_i$ 의 최대한계인  $r_i$ 를 추정한다.

$$\rho_i \leq r_i = \Pr(\text{공개자료의 조사단위 } i \text{와 모집단의 단위 } i^* \text{를 정확하게 연계} | s, P, \text{최악의 경우 가정})$$

주어진 주요변수들로 가능한 모든 조합의 수를  $K$ 개라고 하고,  $k$ 번째 조합에 포함되는 모집단의 개체수를  $F_k$ , 공개자료의 개체수를  $f_k$ 라고 하자. 예를 들어 성별(남/여)과 거주지의 시도(16개)를 주요변수로 고려한다면 가능한 모든 조합의 수인  $K$ 는 32이다. 주요변수의  $k$ 번째 조합에 속하는 조사단위의 수가 표본에서 1이면( $f_k=1$ ) 표본 유일성이고 모집단에서 1이면( $F_k=1$ ) 모집단 유일성이다.

이제  $1/F_k$ 을 주요변수들의  $k$ 번째 조합에 속하는 하나의 조사단위의 신원이 노출되는 확률이라고 하면, 개인노출 위험의 확률은  $1/F_k$ 의 평균이고 이것은 다시  $F_k|f_k$ 의 분포에 대해 나타낼 수 있다.

$$r_i = E\left(\frac{1}{F_k} | f_k\right) = \sum_{h \geq f_k} \frac{1}{h} \Pr(F_k = h | f_k)$$

이 때 모집단의 개체수인  $F_k$ 는 대부분 알 수 없기 때문에, 모집단이 어떤 초모집단(superpopulation)에서 추출되었음을 가정하고  $F_k|f_k$ 의 확률질량함수(probability mass function)를 찾는 방법을 사용한다. 이를 위해 다음과 같이 포아송분포와 이항분포 모형을 가정한다.

$$\begin{aligned} \pi_k &\sim 1/\pi_k, \text{ 독립적으로 } k=1,2,\dots,K \\ F_k | \pi_k &\sim \text{Poisson}(N\pi_k), \text{ 독립적으로 } k=1,2,\dots,K \\ f_k | F_k &\sim \text{bin}(F_k, p_k), \text{ 독립적으로 } k=1,2,\dots,K \end{aligned}$$

위의 가정 하에서  $F_k|f_k$ 의 사후분포(posterior distribution)는 다음과 같은 음이항 분포(negative binomial)를 갖는다.

$$\Pr(F_k = h | f_k = j) = \binom{h-1}{j-1} p_k^j (1-p_k)^{h-j}, h \geq j$$

음이항 분포 하에서 조사단위의 개인 노출에 대한 확률은 다음과 같이 표현된다. 이 때  $q_k = 1 - p_k$  이고  $F_{21}(f_k, f_k, f_k + 1; q_k)$  는 Gauss hypergeometric series이다.

$$r_k = \frac{p_k^{f_k}}{f_k} F_{21}(f_k, f_k, f_k + 1; q_k)$$

개인의 노출위험 확률의 최대한계  $r_k$ 를 계산하려면 위의 식을 적분으로 나타내거나 수치적인 방법으로 구해야 한다. 그러나 계산이 매우 까다롭기 때문에, 이를 간단하게 나타낼 수 있는 근사식이 다음과 같이 전개되었다. 이는  $\mu$ -argus 또는 R package sdcMicro에서 사용되는 계산식과도 동일하다.

$$- f_k = 1 \text{ 인 경우 : } r_k = -\log(p_k) \frac{p_k}{1-p_k}$$

- $f_k = 2$  인 경우 :  $r_k = \frac{p_k}{q_k} (p_k \log p_k + q_k)$
- $f_k = 3$  인 경우 :  $r_k = \frac{p_k}{2q_k^3} (q_k(3q_k - 2) - 2p_k^2 \log p_k)$

위의 근사식을 살펴보면  $r_i$ 를 추정하기 위해 필요한 값은  $p_k$ 가 된다. 앞에서 가정한 분포에 의해  $p_k$ 는  $F_k$ 가 주어졌을 때의 이항분포 하에서 최대우도추정치(maximum likelihood estimator)로 추정할 수 있다. 결과적으로  $p_k$ 는  $f_k/F_k$ 로 추정한다. 그러나  $F_k$ 가 관측될 수 없는 경우에는 표본조사에서 사용되는 조사단위에 대한 가중치  $w_i$ 를 이용할 수 있다.

$$\hat{p}_k = \frac{f_k}{\sum_{i: k(i)=k} w_i}$$

일반적으로  $\hat{p}_k$ 가 0 또는 1에 가까운 값이면  $r_k$ 가 불안정(unstable)한 값으로 추정되므로,  $\hat{p}_k$ 의 값이 0 또는 1과 같은 극단적인 값은 잘 다루지 않는다. 만약  $\hat{p}_k = 0$ 이면  $f_k = 0$ 이 될 것이고,  $\hat{p}_k = 1$ 이면  $F_{21}(f_k, f_k, f_k + 1; 1 - \hat{p}_k) = 1$ 이 되어 개인의 노출위험 확률은  $1/f_k$ 이 된다.

#### 나. 파일단위(자료)의 노출위험 측정

매스킹 기법을 적용한 여러 개의 파일 중 노출위험이 가장 낮은 파일이 무엇인지 파악하기 위해서 파일단위의 노출위험을 측정할 수 있다. 대부분의 파일 수준의 노출위험은 모집단 유일성의 개념을 기초로 하고 있다. 즉, 모집단에서 유일한 조합을 갖는( $F_k = 1$ ) 조사단위의 수를 직접 추정하여 노출위험을 측정하는 방법을 사용할 수 있다.

간단하게는 개인 수준의 노출위험을 구한 뒤 그 값들의 합이나 평균을 파일단위의 노출위험으로 제시하는 경우도 있다. 예를 들어 개인 수준의 노출위험의 평균으로 제시하는 경우, 재식별화 비율(re-identification rate)  $\xi$ 는 다음과 같이 계산한다.

$$\xi = \frac{1}{n} \sum_{k=1}^K f_k r_k$$



## 제 3절 노출제한 방법

조사된 자료에 수집된 변수가 사람의 특성을 나타내는 나이, 성별, 거주지, 직업 등을 포함한다면 자료를 이용하는 연구자에게는 매우 유용한 정보일 수 있다. 그러나 동시에 자료 제공자는 개인의 자료가 유출되어 사생활이 침해당할 가능성도 존재한다. 따라서 외부에 공개하는 자료는 민감한 정보가 노출될 수 있는 가능성을 줄이는 노력이 필요하다. 원자료에 적절한 변환 방법을 적용하여 여러 변수를 조합하였을 때 개인의 식별이 불가능하게 하는 방법들을 매스킹(masking)이라고 한다. 매스킹 기법들은 크게 네 가지 범주로 구분할 수 있다. 첫 번째는 부분적으로 자료값을 제공하는 표본추출(sampling) 방법이고, 두 번째는 자료값을 가리는 자료 감추기(suppression) 방법이고, 세 번째는 구분할 수 없는 그룹으로 값을 섞어 제공하는 통합(aggregation) 방법이고, 마지막은 자료 값을 변조(perturbation)하는 방법이다.

자료를 매스킹 하는 방법은 매우 다양하기 때문에 매스킹 된 여러 개의 자료 중 어떤 자료를 공개하는 것이 타당한가를 결정하는 것이 더 어려운 문제일 수 있다. 정보 노출을 막기 위한 목적으로 매스킹을 적용하여 제공한 자료는 노출 위험은 작아질 수 있으나 자료의 유용성은 낮아져 잘못된 분석결과를 만들어낼 수 있다. 즉, 노출위험과 정보손실 사이에는 상충관계(trade-off)가 있다. 자료를 공공에게 제공하기 이전에 개인정보 노출의 위험을 최소화하면서 자료의 유용성을 유지할 수 있는 방법을 모색하는 것이 필요하다.

### 1. 표본추출

전체 자료를 표본추출하여 자료의 일부분을 공개하는 노출제한 방법이다. 표본으로 추출된 자료는 전체 자료의 일부분이기 때문에, 표본추출된 자료에서 유일한 조사 단위들이 생기더라도 전체 자료에서의 유일한 단위인지를 확신할 수 없으므로 노출위험은 감소한다. 그러나 표본으로 추출된 자료가 노출위험이 낮아져서 다른 노출제한 방법을 적용할 필요가 없는 것은 아니다. 표본추출된 자료에서도 노출에 대한 위험성이 있기 때문에 필요한 경우 적절한 노출제한 방법을 적용해야 한다.

### 2. 자료 감추기

자료 감추기는 자료가 공개되었을 때 노출위험이 상당히 높다고 판단되는 자료의 일부를 공개하지 않고 숨기는 방법이다. 다른 노출제한기법과 달리 자료의 일부를 공개하지 않는 것이며, 자료 감추기에는 변수 감추기(variable or attribute suppression)와 레코드 감추기(record suppression)가 있다. 이 방법은 자료를 행과 열로 구성된 행렬처럼 생각했을 때, 한

행 또는 한 열을 통째로 감추는 것이다. 예를 들어 출생지나 특정 질병의 유무 등 개인의 식별을 간접적으로 도울 수 있는 변수를 자료에서 숨기거나, 특정 금액 이상의 고소득자의 자료를 모두 숨기는 것을 생각할 수 있다. 이러한 자료 감추기 방법은 개인 정보의 보호는 확실하지만 너무 많은 자료가 손실된다.

이에 대한 대안으로 국소 감추기(local suppression) 방법을 사용할 수 있다. 이 방법은 노출 위험을 높이는 주요변수 조합의 몇 개의 변수 값을 감추는 것이다. 공개하는 주요변수 조합들에 대하여 유일한 개체가 발견되었을 때 국소 감추기 방법을 적용하여 특정 변수 조합에 대해 같은 값을 갖는 개체의 수를 2 이상으로 만들 수 있다. 국소 감추기를 사용하기 위해서는 개인 별 노출위험을 측정하고 노출 위험이 높은 개인의 변수 중 어떤 변수의 값을 감출지를 결정해야 한다.

### 3. 재범주화

자료에서 현재의 범주 수준이 개인의 노출위험을 크게 만드는 변수인 경우 재범주화 방법으로 노출을 제한시킬 수 있다. 재범주화 방법은 이미 범주화되어 있는 변수 내에서 여러 개의 범주를 합치는 방법이 있다. 예를 들어 국민건강영양조사 자료에서 기초생활수급 여부 변수는 ‘그렇다’, ‘지금은 아니다’, ‘아니다’, ‘모름, 무응답’ 네 가지 범주로 구분되어 있다. 기초생활수급을 받다가 지금은 받지 않는 사람들이 많지 않기 때문에 개인 노출 위험이 높을 경우 기초생활수급 여부 변수를 재범주화 하여 ‘현재 기초생활수급 여부’에 대한 변수를 만들 수 있다. 즉, ‘지금은 아니다’ 범주에 속해 있는 사람들을 ‘아니다’ 범주와 합치면 노출 위험을 감소시킬 수 있다. 각 변수에 대해 빈도수가 너무 적은 그룹이 생기지 않도록 재범주화를 하는 것이 일반적이다.

연속적인 값을 가지는 변수를 범주화하는 방법도 재범주화라고 할 수 있다. 예를 들어 만 나이 변수를 그대로 공개하면 노출의 위험이 커진다고 판단되었을 때, 이를 범주화 하여 ‘10대’, ‘20대’, ‘30대’ 등으로 제시하면 개인의 노출위험을 감소시킬 수 있다. 또한 연속형 변수에 매우 작은 값과 매우 큰 값을 가지고 있는 단위들의 수가 적어서 그대로 공개되면 신분의 노출뿐 아니라 민감한 정보가 유출될 가능성이 높은 경우 극단값 코딩을 할 수 있다. 변수의 값이 정해진 기준 값을 초과하거나 기준 값에 미달하면 값 자체를 공개하지 않고 하나의 대표 값으로 바꾸어 공개하는 방법이다. 예를 들어 만나이가 19세 이하인 경우 ‘19세 이하’의 범주로 코딩하는 것을 바텀코딩(bottom-coding)이라고 하고, 80세 이상인 경우 ‘80세 이상’의 범주로 코딩하는 것을 탑코딩(top-coding)이라고 한다.

#### 4. 자료 변조 방법

##### 가. 잡음첨가방법(Noise addition)

잡음첨가방법은 자료의 값에 잡음을 더하거나 곱하여 원래 자료에 약간의 변형을 가하는 방법이다. 잡음첨가방법은 변수가 연속적인 값을 가지는 경우에 주로 적용하는 방법이지만, 잡음의 형태를 조정하면 범주형 변수에도 적용이 가능하다. 이 방법은 중요한 개인 정보를 포함한 변수에 잡음을 첨가하여 원자료와 자료값이 달라지게 하므로 외부 이용자가 외부 정보를 사용하여 정확 매칭(exact matching)하여 개인을 식별할 수 있는 가능성이 적어진다. 더 나아가서 개인의 신분이 노출되어도 민감한 정보에 잡음이 첨가되었기 때문에 외부 이용자가 알아낸 개인의 정보가 실제 값과는 차이가 있어 직접적인 정보 노출의 피해를 줄일 수 있다.

잡음첨가방법을 적용하는 경우 편향(bias)의 발생을 피하기 위해 자료의 값에 더해주는 잡음의 평균을 0으로 하고 곱해주는 잡음의 평균은 1이 되도록 한다. 따라서 변형된 자료의 평균은 크게 변하지 않지만 분산이 증가하는 단점이 있으며, 하나 이상의 변수에 잡음을 첨가하는 경우에는 변수들 간의 상관관계를 왜곡시킬 위험이 있다. 각 변수에 독립적인 잡음을 첨가하는 것을 무상관 잡음첨가(uncorrelated noise addition)라고 하며, 이 경우 원자료의 상관관계보다 변형된 자료의 상관관계가 더 작아진다. 이러한 단점을 보완하기 위해 모든 변수들의 공분산 행렬에 비례하는 공분산을 가지는 분포에서 잡음을 발생시켜 원자료에 첨가하는 상관 잡음첨가(correlated noise addition) 방법이 있다. 무상관 잡음첨가 방법보다 상관 잡음첨가 방법이 원자료의 구조를 잘 보존하지만, 상관 잡음첨가 방법을 사용하여도 원자료에서 정규성 가정이 성립하지 않으면 변조된 자료의 구조가 심각하게 왜곡될 수 있다. 또한 두 방법 모두 분산이 증가하는 것은 방지하지 못한다.

잡음첨가방법은 변수가 가질 수 있는 값의 범위가 양수인 경우에도 변형된 자료가 음수인 문제가 발생할 수 있다. 이것을 방지하기 위해 평균이 1인 양의 값만 가능한 분포에서 잡음을 생성하여 원자료에 곱하는 승법잡음(noise multiplication) 방법을 적용할 수 있다.

##### 나. 자료교환(Data Swapping)

자료교환은 자료의 값을 변화시키지 않고 그 위치를 바꾸는 방법이다. 자료교환 방법을 적용하면 자료교환 후에도 일부 통계량들이 원래 통계량과 같거나 매우 유사하게 얻어질 수 있다. 초기의 연구는 대부분 범주형 자료에서 도수 분포(frequency distribution)가 변하지 않도록 하는 자료교환 방법이었으나 연속형 자료에 대해서도 적용할 수 있도록 확장되었다. 특히 연속형 변수에서 자료를 교환하는 방법 중 하나로 자료를 크기 순서대로 정렬하고 제

한된 범위 내에서 서로 교환하는 순위자료교환(rank swapping) 방법이 있다.

이 방법은 매우 간단하여 자료와 랜덤 난수 생성만 있으면 적용할 수 있고, 자료의 민감한 변수에 대해서만 선택적으로 시도할 수 있다는 장점이 있다. 자료를 교환하면 자료와 응답자 사이의 관계가 제거되어 개인의 식별이 불가능해지므로 외부 이용자들의 비밀 누출 행위에 대한 시도를 차단할 수 있다. 자료가 교환되는 정도가 커지면 민감한 정보의 유출 가능성이 줄어들지만 자료 교환 전과 후의 통계량의 차이가 커지기 때문에 자료의 유용성은 감소한다. 더욱이 변수들의 특성을 고려하지 않고 자료교환을 시행하는 경우 실제로 존재하지 않는 조합을 생성할 수도 있다. 극단적인 사례로 남자가 모유수유를 했다고 응답하는 자료가 생성될 수 있다. 따라서 자료에 대한 전반적인 구성이나 변수의 특성 및 분포를 고려하여 자료교환 방법을 적용해야 한다.

**제3장 최종 정책연구용역사업 결과는  
개인정보 노출위험이 있어 비공개 합니다.**

## 제 4장 연구결과 고찰 및 결론

본 연구는 국민건강영양조사 자료와 통계청의 사망관련 변수를 연계한 「국민건강영양조사-사망원인통계 연계자료」를 공개하기에 앞서 연계자료의 노출 위험성을 평가하고, 개인의 재식별화 가능성의 측면에서 자료의 공개가 적절한지 평가하기 위한 연구이다.

「국민건강영양조사-사망원인통계 연계자료」는 제 4기(2007-2009)부터 제 6기(2013-2015)까지의 조사자료로 구성되어 있다. 모든 조사 참여자의 값을 공개하는 것은 아니며, 19세 이상의 조사 참여자 중 타 기관과의 자료연계에 동의한 일부 대상자들의 자료 공개를 고려하고 있다. 일반적으로 마이크로자료가 표본조사인 경우 전수조사에 비해 개인정보가 노출될 위험은 낮지만, 개인을 식별할 수 있는 인구지리학적인 변수를 많이 공개할수록 여러 변수를 조합하여 어떤 특징을 갖는 단 한 명의 개인이 나타날 가능성이 높아진다. 만약 마이크로자료에서 유일한 대상이 전수조사의 결과에서도 유일하다면, 자료노출 시도자가 자료의 연계를 시도했을 때 개인의 민감정보가 유출될 수 있다. 이에 개인의 정보노출 위험을 낮추기 위해 「국민건강영양조사-사망원인통계 연계자료」는 원시자료에서 인구지리학적인 변수를 일부 제외하고 구축하였다. 하지만 공개하는 인구학적인 변수와 사망관련 정보를 조합하여 정보노출의 시도가 발생한다면 개인정보가 노출될 위험이 어느 정도인지, 이를 줄이기 위한 방안은 무엇인지를 연구하였다.

나이와 같이 연속적인 값을 갖는 인구학적 변수는 범주형 변수와는 다르게 변수들의 조합의 개수가 기하급수적으로 많아진다. 즉, 나이를 연속적인 값 그대로 고려하여 다른 변수들과의 조합을 살펴본다면 표본의 수가 크지 않은 국민건강영양조사의 자료는 대부분이 유일한 특성을 갖는 사람으로 구분된다. 심지어 75세 이상의 나이로 조사에 참여한 사람들은 연속적인 나이 값에서 전체 자료의 1.3% 미만 또는 1명으로 특징지어지기 때문에 나이 변수에는 톱코딩 방법을 적용하여 자료를 공개할 필요가 있다. 그 외의 연속형 변수는 질환의 진단시기와 건강검진 결과로 회상 비뚤림과 측정오차와 같은 불확실성이 있으므로 자료노출 시도자가 다른 자료와의 연계로 개인을 재식별화 하기에는 어려움이 있다. 그럼에도 정보노출이 염려될 경우 연속형 변수 내에서 자료를 교환하거나 잡음을 추가하는 등의 매스킹 방법을 적용할 수 있으며, 이 방법에 대해서 보고서 제3장 3절에 기술하였다.

국민건강영양조사 자료에 연계된 통계청의 사망관련 변수는 사망여부와 사망원인이다. 사망여부는 사망과 생존으로 구분되지만, 사망원인은 다양할 수 있고 사망원인이 특이성을 가질수록 인구학적인 변수와의 조합으로 개인을 식별할 가능성이 높아진다. 따라서 사망원인을 여러개 수준의 범주로 재범주화하여 자료의 유용성은 유지하며 노출의 위험을 낮추는 최선의 공개수준을 결정하고자 하였다. 사망원인은 범주의 수준을 고려하여 5개의 방법으로 재범주화를 하였으며, 변수의 값을 가장 많이 숨긴 두 개의 수준(질병이환 및 사망의 외인,

기타)으로 구분한 사망원인부터 다섯 개의 수준(신생물, 순환계통의 질환, 호흡계통의 질환, 질병이환 및 사망의 외인, 기타)으로 구분한 사망원인을 인구학적 변수와 결합하여 각각의 경우에 개인의 노출위험 및 파일 단위의 노출위험이 얼마나 되는지 평가하였다.

노출 위험도를 평가할 주요변수로 인구학적 변수(조사연도, 동읍면, 구간만나이, 성별, 교육수준, 직업재분류, 결혼여부)를 선정한 경우와 사망여부 및 재범주화한 다섯 개의 사망원인들을 선정하였을 때의 유일성 기반의 노출 위험도 결과를 비교하였다. 자료를 구성하고 있는 전체 대상자와 생존자 및 사망자를 구분하여 분석하고, 위암을 진단받은 적이 있다고 응답한 환자들을 대상으로 분석하였다. 사망자에게서는 총 대상자 중 25% 이상이 표본 유일성을 보이며 사망원인 범주의 수준을 3개 이상으로 구성하여 공개하면 40% 이상이 표본 유일성을 보인다. 자료노출 시도자가 동일한 대상자의 주요변수들로 이루어진 민감정보를 포함한 자료를 가지고 있다면, 자료를 연계하여 40% 이상 대상자의 정보를 연결할 수 있다는 의미이다. 예를 들어, 연구자가 연계자료를 원시자료와 연결한다면 개인의 재식별화 위험을 낮추기 위해 삭제한 인구지리학적 변수를 회복할 수 있는 위험이 있다.

노출 위험도 평가 결과로 자료를 공개하였을 때 표본에서 유일한 개인이 모집단(전수)에서도 유일한 개인일 확률을 개인별로 계산하고 평가하였다. 또한, 이를 이용한 전체 표본 수준의 노출위험 확률도 계산하였다. 전체 표본 수준의 노출위험 확률은 공개하는 사망원인 범주의 수준 수가 많아질수록 자료 연계 없이 인구학적 변수만 공개할 때에 비해 증가하였다. 그러나 나이(연속형), 소득분위수, 결혼상태 등을 주요변수로 함께 고려한다면 연계자료의 노출 위험도는 더욱 증가할 것이다.

본 연구결과로 볼 때 사망원인통계 연계자료를 현재의 상태로 불특정 다수가 누구나 이용할 수 있도록 공개 배포하는 경우 개인정보 노출의 위험이 높다고 할 수 있다. 특히, 사망자의 정보노출 위험은 더욱 높아진다. 더불어, 원시자료를 가지고 있는 연구자가 원시자료와 사망원인 연계자료를 연결한다면 사망원인 연계자료에서 삭제된 변수가 복원되어 정보노출의 위험이 발생할 수도 있다. 따라서 공개용 사망원인 연계자료는 사망관련 정보 및 개인을 식별할 수 있는 주요변수를 최소화하여 공개할 필요가 있다. 만일 좀 더 높은 수준의 정보를 원하는 이용자에게는 자료 이용을 위한 승인 절차를 거치거나, 제한된 공간에 방문하여 분석 또는 원격접속을 통해 자료에 접근할 수 있도록 하는 등의 방법을 고려할 필요성이 있다. 연구자의 연구 목적에 맞는 사망원인을 따로 검토하여 제공하는 방법도 사용할 수 있다. 예를 들어, 암과 관련된 사망을 연구하는 연구자에게 암 때문에 사망했는지 여부만을 공개하는 것이다. 그러나 이러한 과정들은 모두 자료를 제공하는 기관의 부담이 있을 것이다. 하지만, 이러한 과정을 통해 사망원인 연계 자료에서 개인정보의 노출의 위험을 최소화할 수 있다.

## 제 5장 연구성과 및 활용계획

### 5.1 연구성과

정책연구용역 사업명	국민건강영양조사-사망원인통계 연계자료 개인정보
책임연구원	안형진 / 고려대학교 / 의학통계

#### 가. 연구논문

번호	논문제목	저자명	저널명	집(권)	페이지	Impact factor	국내/국외	SCI 여부
1								
2								

#### 나. 학술발표

번호	발표제목	발표형태	발표자	학회명	연월일	발표지	국내/국제
1							
2							

#### 다. 지적재산권

번호	출원/등록	특허명	출원(등록)인	출원(등록)국	출원(등록)번호	IPC분류
1						
2						

#### 라. 정책제안 및 활용

「국민건강영양조사-사망원인통계 연계자료」의 개인정보 노출 위험을 평가하고 효율적인 정보공개를 위한 연계자료 공개 관련 관리대책을 제시함.
--

#### 마. 타연구/차기연구에 활용

해당 없음
-------

#### 바. 언론홍보 및 대국민교육

해당 없음
-------

#### 사 기타

해당 없음
-------

## 5.2 활용계획(연구사업 종료 후)

정책연구용역 사업명	국민건강영양조사-사망원인통계 연계자료 개인정보
책임연구원	안형진 / 고려대학교 / 의학통계

### 가. 연구논문

번호	논문제목	저자명	저널명	집(권)	페이지	Impact factor	국내/국외	SCI 여부
1								
2								

### 나. 학술발표

번호	발표제목	발표형태	발표자	학회명	연월일	발표지	국내/국제
1							
2							

### 다. 지적재산권

번호	출원/등록	특허명	출원(등록)인	출원(등록)국	출원(등록)번호	IPC분류
1						
2						

### 라. 정책제안 및 활용

「국민건강영양조사-사망원인통계 연계자료」의 개인정보 노출 위험을 평가하고 효율적인 정보공개를 위한 연계자료 공개 관련 관리대책을 제시함.
--

### 마. 타연구/차기연구에 활용

해당 없음
-------

### 바. 언론홍보 및 대국민교육

해당 없음
-------

### 사. 기타

해당 없음
-------



## **제 6장 정책연구용역사업 진행과정에서 수집한 해외과학기술정보**

- 해당 없음

## **제 7장 기타 중요변경사항**

- 해당 없음

## 제 8장 연구비 사용 내역 및 연구원 분담표

### 8.1 연구비 사용 내역(작성일까지 사용내역 작성)

구분 \ 비목	금액(원)	구성비	비고
○ 인 건 비 소 계	14,495,454	72.48%	
책 임 연 구 원 (총 명)	3,860,236	19.30%	
연 구 원 (총 명)			
연 구 보 조 원 (총 명)	10,635,218	53.18%	
보 조 원 (총 명)			
○ 경 비 소 계	2,119,000	10.60%	
여 유 인 물 비	500,000	2.50%	
전 산 처 리 비	285,000	1.43%	
시 약 및 연 구 용 재 료 비			
회 의 비	1,020,000	5.10%	
임 차 료			
교 통 통 신 비			
위 탁 정 산 수 수 료	314,000	1.57%	
연 구 활 동 비 ( )%	701,564	3.51%	
일 반 관 리 비 ( )%	865,800	4.33%	
부 가 가 치 세	1,818,182	9.09%	
○ 계	20,000,000	100%	

## 제 9장 참고문헌

박민정 (2015) 마이크로데이터 매스킹 기법 실무 활용 안내서, 통계개발원

안형진, 이용희, 송주원, 김규영 (2011) 통계적 정보 보호 방법, 통계교육원

Benschop, T., Machingauta, C., & Welch, M. (2016). Statistical Disclosure Control for Microdata: A Practice Guide. Technical Report July, International Household Survey Network and the World Bank.

Benschop, T., Machingauta, C., & Welch, M. (2019). Statistical Disclosure Control: A Practice Guide.

Duncan, G. T., Elliot, M., and Salazar-Gonzalez, J.-J. (2011). Statistical Confidentiality, Springer: New York

Franconi, L., & Polettini, S. (2004, June). Individual risk estimation in  $\mu$ -Argus: A review. In International Workshop on Privacy in Statistical Databases (pp. 262–272). Springer, Berlin, Heidelberg.

Hundepool, A., Wetering, A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., & Giessing, S. (2008). Mu-argus, version 4.2 user's manual. Statistics Netherlands.

Hundepool, A., De Wolf, P. P., Bakker, J., Reedijs, A., Franconi, L., Polettini, S., ... & Domingo-Ferrer, J. (2014). mu-Argus User's Manual version 5.1. Statistics Netherlands: The Hague, The Netherlands.

Park I, Dohrmann S, Montaquila J, Mohadjer L, Curtin LR. (2006) Reducing the risk of data disclosure through area masking: limiting biases in variance

estimation. Proceedings of the Section on Physical and Engineering Sciences. American Statistical Association: Alexandria, 1761–1767.

Templ, M. (2007). sdcMicro: A package for statistical disclosure control in R. na.

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(1), 1–36.

## 제 10장 첨부서류

- 해당 없음