

보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(Ⅱ)

- 가명정보 결합을 중심으로

이혜정·이기호

김기태·안수인·박성준·이지은·김민석

사람을
생각하는
사람들



KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



한국 보건사회연구원
KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



■ 연구진

연구책임자	이혜정	한국보건사회연구원 부연구위원
	이기호	한국보건사회연구원 부연구위원
공동연구진	김기태	(주)유플러스데이터 대표
	안수인	한국보건사회연구원 연구원
	박성준	한국보건사회연구원 부연구위원
	이지은	한국노동연구원 노동패널팀장
	김민석	충남대학교 연구교수

연구보고서 2023-09

보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

- 가명정보 결합을 중심으로

발 행 일 2023년 12월
발 행 인 이 태 수
발 행 처 한국보건사회연구원
주 소 [30147] 세종특별자치시 시청대로 370
세종국책연구단지 사회정책동(1~5층)
전 화 대표전화: 044)287-8000
홈페이지 <http://www.kihasa.re.kr>
등 록 1999년 4월 27일(제2015-000007호)
인 쇄 처 (사)아름다운사람들

© 한국보건사회연구원 2023
ISBN 978-89-6827-966-9 [93510]
<https://doi.org/10.23060/kihasa.a.2023.09>

발|간|사

데이터를 개방하는 추세에 따라 데이터가 다량으로 개방되고 있을 뿐만 아니라 여러 가지 법, 제도 등의 도입·개정으로, 여러 개의 데이터를 서로 연계하고 통합한 데이터를 활용할 수 있는 환경이 조성되고 있는 상황이다. 그러나 여전히 데이터가 개별적으로 분산된 채 제공되고 있고, 여러 개의 데이터를 서로 연계하고 통합한 데이터를 활용하는 것은 저조하다는 평가가 이어지고 있다.

이 연구는 다양한 주제별로 분산된 데이터 자원을 연계·통합하는 적극적인 시도를 통해 데이터의 활용 가치를 높이는 연구로, 2년 동안 이루어졌다. 1차 연도(2022년)에는 데이터 연계·통합을 위한 활용 가능한 데이터 현황을 파악하였고, 통계적 매칭 방법을 활용하여 데이터를 연계·통합하는 실증 연구를 하였다. 2차 연도(2023년)에는 가명정보 결합의 개념과 활용 방법을 살펴보았다. 그런 다음 가명정보 결합을 활용한 통합데이터 구축 사례와 관련한 국내외 선행연구를 검토하였다. 가명정보 결합 활성화를 위해 현행 법률 및 가이드라인을 검토하였고, 이를 통해 우리 연구원의 익명·가명 데이터 제공 절차 개선 방안을 마련하였다. 마지막으로 데이터 통합 실증 연구로, 가명정보 결합을 활용하여 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터를 통합한 다음, 이 ‘통합데이터’에 대한 분석을 실시하였다. 이러한 연구 결과에 따라 데이터 통합의 중요성을 확인하였으며, 가명정보 결합 제도를 활성화하기 위한 방안에 대해 제안하였다.

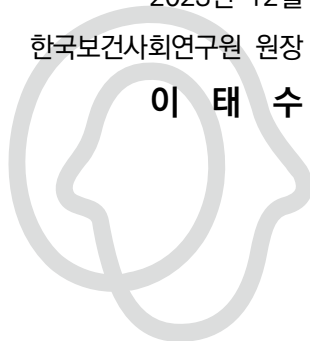
이 보고서는 이해정, 이기호 부연구위원의 공동 책임하에 박성준 부연구위원, 안수인 연구원이 원내 연구진으로 참여하였다. 외부 연구진으로 유피에스데이터 김기태 대표, 한국노동연구원 이지은 팀장, 충남대학교

김민석 연구교수가 참여하였다. 모든 연구진의 노고에 감사드린다. 보고서 작성과 관련하여 유익한 의견을 주신 원내 오미애 연구위원, 원외 고려대학교 송주원 교수, 그리고 익명의 검독위원들에게도 감사의 마음을 전한다. 마지막으로, 본 보고서의 내용은 본원의 공식적인 의견이 아님을 밝힌다.

2023년 12월

한국보건사회연구원 원장

이 태 수



목 차

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



Abstract	1
요 약	5
제1장 서론	15
제1절 연구의 배경 및 목적	17
제2절 연구 내용 및 방법	23
제2장 가명정보 결합 개념 및 활용 방법	27
제1절 가명정보 결합의 개념	29
제2절 가명정보 결합 및 반출 절차	34
제3절 소결	75
제3장 가명정보 결합을 활용한 통합데이터 구축 및 연구 사례	79
제1절 국내 사례	81
제2절 해외 사례	94
제3절 소결	122
제4장 가명정보 결합 활성화를 위한 제반 사항 검토	127
제1절 현행 법률 및 가이드라인 검토	130
제2절 데이터 제공 관련 현황 분석	142
제3절 데이터 제공 절차 수립	152
제4절 데이터 제공 절차 개선 방안	171
제5절 소결	178



제5장 가명정보 결합을 활용한 데이터 통합 실증 연구 181

제1절 가명정보 결합을 활용한 통합데이터 구축 183

제2절 통합데이터 실증 분석 191

제3절 소결 239

제6장 결론 243

제1절 연구 결과 요약 및 함의 245

제2절 가명정보 결합 제도 활성화를 위한 방안 258

제3절 향후 과제 263

참고문헌 265

부록 271

표 목차

KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



〈표 1-1〉 연도별 가명정보 결합 선도 사례(2021~2023년)	18
〈표 1-2〉 분야별 데이터 결합 분야	22
〈표 1-3〉 분야별 데이터 결합 목적	22
〈표 2-1〉 가명정보 처리 및 결합 관련 「개인정보 보호법」 내용	31
〈표 2-2〉 가명정보 처리 안전성 확보 관련 「개인정보 보호법」과 「개인정보 보호법 시행령」 내용	32
〈표 2-3〉 가명정보 관련 제도 현황에 대한 참고	33
〈표 2-4〉 가명정보 처리 가능 목적별 예시	39
〈표 2-5〉 위원회 구성 시 참고사항	45
〈표 2-6〉 결합전문기관 리스트	49
〈표 2-7〉 결합키 생성을 위한 대표적인 권고 해시 알고리즘	55
〈표 2-8〉 결합률 예시	56
〈표 2-9〉 처리 환경 위험도 분류 측정 예시	60
〈표 3-1〉 행정자료 활용 현황(2019년 기준)	84
〈표 3-2〉 자료출처 및 이용 항목	87
〈표 3-3〉 임상 및 상황 정보 보고서(BIC) 작성 항목	110
〈표 3-4〉 NDB에 저장·제공되는 일본의 의료보험제도 종류별 전자화된 익명 진료(조제) 보수 명세서 정보	117
〈표 3-5〉 NDB 활용 사례	121
〈표 4-1〉 데이터 제공 절차 관련 주요 컨설팅 과업 내용	129
〈표 4-2〉 정보공개법에 따른 데이터 제공 단계별 설명	154
〈표 4-3〉 통계법에 따른 데이터 제공 단계별 설명	156
〈표 4-4〉 공공데이터법에 따른 데이터 제공 단계별 설명	158
〈표 4-5〉 데이터기반행정법에 따른 데이터 제공 단계별 설명	160
〈표 4-6〉 개인정보보호법에 따른 데이터 제공 단계별 설명	162
〈표 4-7〉 생명윤리법에 따른 데이터 제공 단계별 설명	164
〈표 4-8〉 결합정보 제공과 이용 형태에 따른 해당 절차 - 결합전문기관 결합 절차	166

〈표 4-9〉 결합정보 제공과 이용 형태에 따른 해당 절차 - 데이터 전문기관 결합 절차 ...	169
〈표 4-10〉 데이터 제공 체계를 위한 역할 및 책임에 따른 담당 부서 배치	172
〈표 4-11〉 가명처리 총괄의 담당 부서 조직 구성 방안	173
〈표 4-12〉 3개 안의 장·단점	173
〈표 4-13〉 제2조 이용자의 의무에서 수정·보완 사항	174
〈표 4-14〉 자료 활용 시 주의사항에 대한 수정·보완 사항	175
〈표 4-15〉 단계별 개선 내용 및 적용 시기	178
〈표 4-16〉 우선순위에 따른 적용 내용	179
〈표 5-1〉 한국의료패널조사와 사회보장정보시스템에서의 활용 가능 변수	192
〈표 5-2〉 행정데이터 기준 사회보장 급여 관련 분포	195
〈표 5-3〉 행정데이터 기준 사회보장 급여액 현황	196
〈표 5-4〉 거주 지역에 대한 조사데이터와 행정데이터의 일치 여부	196
〈표 5-5〉 조사데이터 기준 가구원 수 분포	197
〈표 5-6〉 통합데이터의 기초연금 수급 여부에 대한 분포	198
〈표 5-7〉 조사데이터와 행정데이터의 기초연금 수급 여부 일치 여부	198
〈표 5-8〉 기초연금 수급에 대한 일치 여부에 따른 거주 지역 분포	200
〈표 5-9〉 기초연금 수급에 대한 일치 여부에 따른 가구원 수 분포	201
〈표 5-10〉 기초연금 수급에 대한 일치 여부에 따른 주거형태 분포	202
〈표 5-11〉 기초연금 수급에 대한 일치 여부에 따른 연간 가구 총소득 분포	203
〈표 5-12〉 기초연금 수급에 대한 일치 여부에 따른 가구의 성별 분포	204
〈표 5-13〉 기초연금 수급에 대한 일치 여부에 따른 가구의 연령대 분포	204
〈표 5-14〉 기초연금 수급에 대한 일치 여부에 따른 가구의 혼인상태 분포	205
〈표 5-15〉 기초연금 수급에 대한 일치 여부에 따른 가구의 최종 학력 분포	206
〈표 5-16〉 기초연금 수급에 대한 일치 여부에 따른 가구의 경제활동 상태 분포	207
〈표 5-17〉 통합데이터의 생계급여 수급 여부에 대한 분포	208
〈표 5-18〉 조사데이터와 행정데이터의 생계급여 수급 여부 일치 여부	209
〈표 5-19〉 통합데이터의 의료급여 수급 여부에 대한 분포	209



〈표 5-20〉 조사데이터와 행정데이터의 의료급여 수급 여부 일치 여부	210
〈표 5-21〉 통합데이터의 주거급여 수급 여부에 대한 분포	210
〈표 5-22〉 조사데이터와 행정데이터의 주거급여 수급 여부 일치 여부	211
〈표 5-23〉 통합데이터의 교육급여 수급 여부에 대한 분포	212
〈표 5-24〉 조사데이터와 행정데이터의 교육급여 수급 여부 일치 여부	212
〈표 5-25〉 통합데이터의 맞춤형 급여 수급 여부에 대한 분포	213
〈표 5-26〉 통합데이터의 맞춤형 급여 수급 여부 일치 여부	213
〈표 5-27〉 맞춤형 급여 수급에 대한 일치 여부에 따른 거주 지역 분포	214
〈표 5-28〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구원 수 분포	216
〈표 5-29〉 맞춤형 급여 수급에 대한 일치 여부에 따른 주거형태 분포	216
〈표 5-30〉 맞춤형 급여 수급에 대한 일치 여부에 따른 연간 가구 총소득 분포	217
〈표 5-31〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구의 성별 분포	218
〈표 5-32〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구의 연령대 분포	219
〈표 5-33〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구의 혼인상태 분포	220
〈표 5-34〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구의 최종 학력 분포	221
〈표 5-35〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구의 경제활동 상태 분포	222
〈표 5-36〉 급여별·자격별 의료비 부담 수준(5점 척도)	225
〈표 5-37〉 의료비 및 급여액 분위 구분	227
〈표 5-38〉 의료비 분위별·급여액 의료비 평균 비교	228
〈표 5-39〉 급여 분위별 의료비 평균 비교	232
〈표 5-40〉 연간 의료비 및 급여액의 분위별 가구 수 분포-1	234
〈표 5-41〉 연간 의료비 및 급여액의 분위별 가구 수 분포-2	235
〈표 6-1〉 가명정보 결합을 사용하여 통합데이터를 구축한 사례 정리	247
〈표 6-2〉 우선순위에 따른 적용 내용	251
〈표 6-3〉 한국의료패널조사와 사회보장정보시스템의 가명정보 결합 절차	253
〈부표 1〉 보건복지 분야 전문가 대상 가명정보 결합을 활용한 연구 주제 관련 설문조사 결과	283

그림 목차

[그림 1-1] 데이터 결합 건수(누적)와 데이터 결합 참여 횟수	21
[그림 1-2] 연차별 연구 목적	23
[그림 2-1] 가명정보 결합 절차(개요)	34
[그림 2-2] 가명정보 결합·반출 단계 및 관련 기관의 역할	36
[그림 2-3] 사전준비 단계	38
[그림 2-4] 가명정보 이용·제공 신청서(예시)	42
[그림 2-5] 결합신청 단계	47
[그림 2-6] 가명정보결합 종합지원시스템	48
[그림 2-7] 가명정보 결합신청서	51
[그림 2-8] 결합 목적을 증명할 수 있는 서류	52
[그림 2-9] 가명 대상 가명정보에 관한 서류	53
[그림 2-10] 결합키 생성 예시	55
[그림 2-11] 결합 유형	57
[그림 2-12] 가명처리 단계별 절차도	58
[그림 2-13] 가명정보 활용 종합지원플랫폼	59
[그림 2-14] 항목별 가명처리계획 작성 예시	62
[그림 2-15] 가명정보 제공 기관의 가명처리 현황 공개(예시) 일반기관으로 예시 변경 ..	69
[그림 2-16] 결합 단계	70
[그림 2-17] 결합전문기관의 가명정보 결합 과정	71
[그림 2-18] 반출 및 사후관리	71
[그림 2-19] 국세통계센터 이용 절차	72
[그림 2-20] 반출 신청서	74
[그림 3-1] 사회보장 행정데이터 구축 과정	91
[그림 3-2] SINAVE(국립 역학 감시 정보 시스템)와 SICO(국립 전자 사망 증명서 정보 시스템) 데이터베이스의 데이터 연계에 관한 흐름	112
[그림 3-3] NDB에 포함된 정보	116
[그림 3-4] NDB 수집 경로	116



[그림 3-5] 익명화 처리 과정	118
[그림 3-6] 해시 함수에 대한 유의사항	120
[그림 4-1] 연구 참여 동의서-익명화 관련 예시	137
[그림 4-2] 연구 참여 동의서-제3자 제공 관련 예시	137
[그림 4-3] 가명정보 처리 가이드라인의 가명처리 절차	139
[그림 4-4] 보건의료데이터 활용 가이드라인 데이터 심의위원회 세부 구성 요건	140
[그림 4-5] 정보 주체의 인권 및 사생활 보호에 중대한 피해를 야기할 수 있어 본인의 동의를 받아 사용하는 것을 원칙으로 하는 정보	140
[그림 4-6] 한국복지패널조사 데이터 제공 웹 페이지	146
[그림 4-7] 한국의료패널조사 데이터 제공 절차 안내 페이지	148
[그림 4-8] 한국의료패널 자료활용동의서	150
[그림 4-9] 연구 참여 동의서 포함 항목	151
[그림 4-10] 익명·가명 데이터 제공 절차도	153
[그림 4-11] 정보공개법에 따른 데이터 제공 절차도	155
[그림 4-12] 통계법에 따른 데이터 제공 절차도	157
[그림 4-13] 공공데이터법에 따른 데이터 제공 절차도	159
[그림 4-14] 데이터기반행정법에 따른 데이터 제공 절차도	161
[그림 4-15] 개인정보보호법에 따른 데이터 제공 절차도	163
[그림 4-16] 생명윤리법에 따른 데이터 제공 절차도	165
[그림 4-17] 개인정보보호법에 따른 데이터 제공 절차도 - 내부 결합	167
[그림 4-18] 개인정보보호법에 따른 데이터 제공 절차도 - 결합전문기관 결합	168
[그림 4-19] 개인정보보호법에 따른 데이터 제공 절차도 - 데이터 전문기관 결합	170
[그림 4-20] 데이터 제공 양식 보완-연구 참여 동의서	176
[그림 4-21] 데이터 제공 양식 보완-개인정보 수집 및 이용 동의서	177
[그림 5-1] 통합데이터를 활용한 분석 주제	194
[그림 5-2] 기초연금 자격별 의료비 부담 수준(5점 척도)	226
[그림 5-3] 의료급여 자격별 의료비 부담 수준(5점 척도)	226



[그림 5-4] 의료비 분위별 연간 의료비 평균	230
[그림 5-5] 급여액 분위별 의료비 평균	233
[그림 5-6] 연간 의료비 및 생계급여 분위별 가구 분포	236
[그림 5-7] 연간 의료비 및 기초연금(전체) 분위별 가구 분포	236
[그림 5-8] 연간 의료비 및 기초연금(단독) 분위별 가구 분포	237
[그림 5-9] 연간 의료비 및 기초연금(부부2인) 분위별 가구 분포	237
[부도 1] 적정성 검토 위원회에 제출할 필요 서류	271



Abstract

A Study on How to Link, Intergrate, Utilize Multi-Source Data to Vitalize the Data Economy in the Health and Welfare Sectors(II):

With a Focus on the Linkage of Pseudonymized Data

Project Head: Lee, Hyejung · Lee, Kiho

In line with the trends of open data, a growing amount of data is becoming accessible for users. However, such open data is primarily provided in unlinked, separate sets by different sources.

In this study, our aim was to examine how data from multiple sources, when linked and integrated, can increase in utility and explore steps needed to be taken to overcome challenges in implementing data integration and linkage.

We constructed a dataset by integrating Korea Health Panel Survey data and social security administrative data using pseudonymous data linkage. The survey data and administrative data showed similar proportions of individuals receiving social security benefits. There were instances where survey responses regarding benefit receipt did not align with the corresponding administrative data. For basic pensions, the proportion of households identified in the survey as recipients but confirmed to be otherwise by administrative data was higher than the pro-

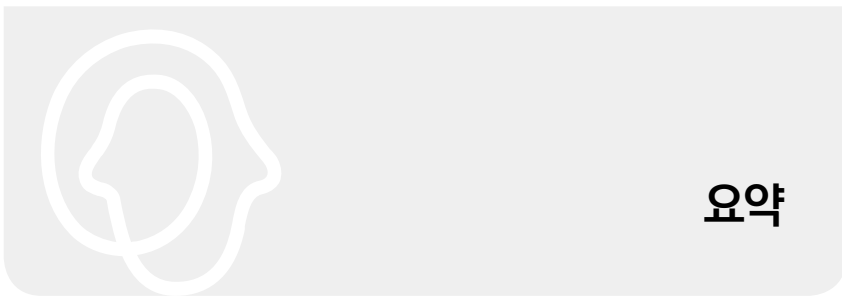
portion of households identified as non-recipients in the survey data but confirmed as recipients according to the administrative data. For customized benefits, we found that the proportion of households identified in survey data as non-recipients but confirmed as recipients in administrative data exceeded the proportion of survey-identified recipients that administrative data classified as non-recipients. Meanwhile, of the various customized benefits, housing benefit was where the proportion of households identified as non-recipients by survey data but as recipients by administrative data was the highest.

The discrepancy between survey data and administrative data can be due to several reasons. First, it could stem from the time difference between the survey responses and administrative data extraction. Second, despite efforts to ensure accuracy, there may be cases where information provided by the representative household member is not precise. Third, administrative data, generated through administrative processing, may lack the refinement of survey data. Fourth, as name, date of birth, and sex were used as matching keys, there may be instances where there may arise inconsistencies in name and sex for the same person due to differences between the resident registration date of birth and the actual date of birth (lunar or solar calendar).

The usability of administrative data was confirmed through a

review of the consistency of integrated data constructed by combining pseudonymous information. The accuracy of survey responses as data can be determined through comparison with administrative data. The results of such comparisons can be used as useful information to collect more accurate responses. Additionally, linking administrative data with survey data can reduce the burden of response on survey participants.

Key words : Korea Health Panel Survey, Social Security Administrative Data, pseudonymization, pseudonymized data



1. 연구의 배경 및 목적

행정안전부는 올해 연말까지 공공데이터포털을 통해 30개의 국가중점 데이터를 개방할 계획이라고 밝혔다. 또한 개인정보보호위원회, 기획재정부 등 관계부처 합동으로 ‘가명정보 활용 확대 방안’을 발표하였다(개인정보보호위원회 보도자료, 2023.7.20.). 이 방안은 가명정보 활용을 위한 데이터 확보의 어려움, 가명처리·결합 절차의 이행부담 및 제약요인, 가명정보 활용 인프라·지원 부족, 가명정보 재식별·유출에 대한 안전성 우려 등을 해소하기 위한 방안을 마련하는 데 목적이 있다. ‘가명정보 활용 확대 방안’의 주요 내용은 양질의 데이터 제공·공유 확대, 가명정보 활용 절차 합리화, 가명정보 활용지원 확대, 가명정보 활용 절차의 안전관리 강화에 대한 것이다. 이렇듯 정부의 적극적이고 지속적인 공공데이터 개방, 가명정보 활용 확대 방안 마련 등의 조치로, 데이터를 활용할 수 있는 분야는 넓어지고 있으며 환경도 좋아지고 있다.

종합해 보면, 데이터를 개방하는 추세에 따라 데이터가 다량으로 개방되고 있을 뿐만 아니라 여러 가지 법, 제도 등의 도입·개정으로, 여러 개의 데이터를 서로 연계하고 통합한 데이터를 활용할 수 있는 환경이 조성되고 있는 상황이다. 그러나 여전히 데이터가 개별적으로 분산된 채 제공되고 있고, 여러 개의 데이터를 서로 연계하고 통합한 데이터를 활용하는 것은 저조하다는 평가가 이어지고 있다.

이 연구의 목적은 곳곳에 분산된 데이터의 활용 가치를 높이기 위해서 다출처 데이터를 연계·통합하는 시도를 해 보고, 이에 따른 어려움을 해결하여 통합 데이터의 활용이 활성화될 수 있도록 하는 방안을 마련하는 것이다.

2. 주요 연구 결과

제2장에서는 가명정보 결합 및 반출 절차에 대해 구체적으로 살펴보았다. 활용하고자 하는 가명정보에 따라 절차가 조금 상이한 경우도 있기 때문에 각각의 가이드라인을 아울러서 정리하였다. 가명정보 결합은 개인 식별이 불가능하도록 가명처리된 가명정보에 대하여 결합전문기관이 두 개 이상의 데이터를 결합키를 이용하여 통계작성, 과학적 연구, 공익적 기록보존 등의 목적에 한해 결합하는 것이다. 가명정보 결합 및 반출 절차는 5단계로, 사전준비, 결합신청, 가명처리, 결합수행, 반출 및 관리의 순서로 진행된다. 각 단계에서 주의해야 할 사항들을 숙지해야만 가명정보 결합에 소요되는 시간을 줄이면서, 가명정보 결합 신청 시 제출했던 연구계획서의 연구 목적에 맞는 정확한 분석을 수행할 수 있다.

제3장에서는 가명정보 결합을 활용하여 ‘통합데이터’를 구축하고 연구한 국내외 사례에 대해 살펴보았다. 국내의 경우 가명정보 결합을 통해 구축된 통합데이터를 활용한 연구가 많지 않은 편으로 나타났다. 통계청에서 ‘조사데이터’와 ‘행정데이터’의 결합을 적극적으로 실시하고 있는 것으로 보였다. 최근에는 사회보장위원회에서도 사회보장 행정데이터를 구축하고 있는 것으로 나타나 앞으로 연구 사례가 점차 많아질 것으로 생각한다. 영국 사례를 보면, 2018년 공공기관의 기능을 수행하기 위해 데이터 결합이 필요하다는 법적 근거가 마련되어 모든 응답자를 연계할 수 있게 되었다. 법의 적용과 연계 방법론의 개선으로 가족자원조사와 행정데이터의 결합률은 95% 이상으로 나타났다. 포르투갈 사례를 보면, 정확매칭을 통해 성별, 연령, 거주지 및 사망일을 기반으로 두 개의 정보시스템 데이터를 결합하였다. 코로나19 감염 및 관련 사망 정보 모니터링 중에 기록된 질환 상태에 관한 두 개의 데이터에서 이용 가능한 정보의 잠

제적인 상보성을 비교·식별하였다. 그 결과 기록된 질환 상태는 일치하는 경우도 있었지만, 서로 다른 정보를 가지는 경우도 확인할 수 있었다. 일본에서는 건강보험 청구데이터와 건강검진 데이터를 결합한 데이터를 구축하고 있다. 환자의 이름, 생년월일, 보험 의료 기관의 소재지와 명칭 등 특정 개인이 식별되는 정보는 삭제하고, 이를 함수로 처리한 두 개의 해시 ID를 사용하여 데이터를 매칭하였다. 이렇게 결합된 데이터는 정보 통계 추출, 시계열 자료 구축 및 평가, 건강검진 정보와 진료 정보 연계 등에 활용되고 있다.

국내에서도 가명정보 결합에 대한 수요가 점차 높아지는 가운데, 해외 사례와 같이 다양한 분야의 행정데이터의 활용 가능성, 응답값의 정확성과 측정오차 등의 검토를 통해 조사데이터의 품질 향상 등의 목적으로 가명정보의 결합이 활성화될 수 있다고 생각한다.

제4장에서는 「정보공개법」, 「통계법」 등 7개의 법령³⁾과 가명정보 관련 3개의 가이드라인⁴⁾을 살펴보고, 한국보건사회연구원의 데이터 제공 절차에 대한 현행 규정⁵⁾과 절차를 분석하였다. 「정보공개지침」에는 정보에 대한 가명처리 등의 가명정보 제도에 대한 내용이 없는 것으로 나타났다. 이러한 종합적인 분석 결과를 토대로, 한국보건사회연구원은 익명·가명 데이터 제공 절차 및 지침을 수정·보완하였고 새롭게 수립하여, 일원화된 데이터 제공 체계를 갖출 수 있도록 방안을 제안하였다. 그 결과 한국보건사회연구원의 다양한 데이터를 더 빠르게 제공할 수 있을 뿐

3) 7개의 법령은 「공공기관의 정보공개에 관한 법률」(이하 '정보공개법'), 「통계법」, 「공공데이터의 제공 및 이용 활성화에 관한 법률」(이하 '공공데이터법'), 「데이터기반 행정 활성화에 관한 법률」(이하 '데이터기반행정법'), 「개인정보 보호법, 생명윤리 및 안전에 관한 법률」(이하 '생명윤리법'), 「신용정보의 이용 및 보호에 관한 법률」(이하 '신용정보법')이 해당함.

4) 가명정보에 대한 가이드라인은 가명정보 처리 가이드라인, 보건의료데이터 활용 가이드라인, 공공분야 가명정보 제공 실무 안내서가 해당함.

5) 정보공개지침, 조사자료관리지침

만 아니라, 안전하게 데이터를 제공할 수 있는 기반을 마련하였다고 생각한다. 특히 데이터제공심의위원회라는 일원화된 의사결정기구를 통해 다양한 데이터의 제공에 대해 통합적인 관리를 할 수 있게 되었으며, 데이터 제공으로 인한 개인 식별 등의 다양한 위험에 대해 효과적으로 고민하고 대응할 수 있게 되었다. 또한, 데이터 제공 체계에서 각 역할과 책임에 따라 담당 부서를 배치하는 방안도 마련하였다. 한국보건사회연구원의 현재 상황과 여건을 고려했을 때, 우선 ‘임시 전담 부서’로 조직 및 운영하도록 제안하였다. 향후 데이터 제공 건수가 많아지면 ‘내부 전담 총괄 부서’로 조직 및 운영하는 것이 효율적이라고 생각한다.

제5장에서는 가명정보 결합 절차에 따라 실제로 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터를 통합하여 구축하였다. 그리고 구축된 ‘통합데이터’를 활용하여 두 개의 분석 주제에 대해 분석하였다.

먼저, 한국의료패널조사 데이터(이를 ‘조사데이터’라고 함)와 사회보장정보시스템의 행정데이터(이를 ‘행정데이터’라고 함)의 가명정보를 결합하기 위해 사전준비, 결합신청, 가명처리, 결합수행, 반출 및 관리의 절차에 따라 결합을 실시하였다.

다음은 가명정보 결합을 통해 구축된 통합데이터를 사용하여 실증 분석한 분석 결과를 요약하였다. 분석 주제는 (1) 조사데이터와 행정데이터의 정합성 검토와 (2) 사회보장 급여 수급자의 의료비 지출 특성 분석이었다. 한국의료패널조사는 의료비 지출을 포함한 각종 보건의료 정보를 포함하고 있는 조사데이터이다. 그런데 사회보장 급여는 세부적인 정보가 부족하거나, 응답자의 기억에 의존하여 응답을 받기 때문에 정확도가 낮아질 수 있다. 그러므로 가명정보 결합을 통해 구축된 통합데이터를 사용하면 조사데이터가 가지는 부족한 부분을 보완할 수 있다는 점이 가장

큰 강점이라고 볼 수 있다.

첫 번째 분석 결과를 살펴보면, 조사데이터와 행정데이터에서 사회보장 급여를 수급 받는다고 응답한 비율이 유사하게 나타났다. 조사데이터와 행정데이터 간 수급 여부가 일치하지 않는 경우, 기초연금은 조사데이터에서 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 가구의 비율이 높았던 반면에, 맞춤형 급여는 반대의 경우가 높았다. 한편, 맞춤형 급여 중에서 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서 수급 받는다고 한 비율이 가장 높은 급여는 주거급여로 나타났다.

조사데이터와 행정데이터의 값이 일치하지 않는 요인은 다음의 것들을 생각해 볼 수 있다. (1) 조사데이터의 응답 시점과 행정데이터의 추출 시점 간의 차이 때문으로 볼 수 있다. 맞춤형 급여는 조사데이터와 행정데이터의 시점이 2019년 12월 31일로 동일하였다. 그러나 기초연금은 조사데이터의 경우 응답 시점이 2019년 1월 1일~12월 31일까지인 반면에, 행정데이터는 2019년 12월 31일이 추출 시점으로, 동일하지 않다고 볼 수 있다. (2) 조사데이터의 경우 사회보장 급여 수급 여부를 가구 조사표에서 조사하고 있는데, 이는 대표 가구원이 모두 응답하는 구조이므로 정확히 파악하지 못하는 경우가 있을 수 있다. (3) 행정데이터는 행정 처리 과정에서 생성되는 데이터이므로, 조사데이터와 같이 정제된 데이터가 아니다. 이러한 특성이 기인했을 가능성도 있다고 생각한다. (4) 두 데이터를 결합할 때 결합키로 이름, 생년월일, 성별을 사용하였다. 이때 동일인이라도 주민등록 생년월일과 실제 생년월일(음력, 양력)의 차이, 또는 이름과 성별의 차이가 생길 수 있다.

조사데이터와 행정데이터의 사회보장 급여 수급 여부가 일치하지 않는 집단의 비율은 낮은 편이었으나, 이러한 차이를 가구 특성과 가구주의 인

구사회학적 특성 분포를 통해 살펴보았다. 조사데이터에서 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 집단은 서울 및 경기도, 광역시에 거주하고, 아파트에 거주하며, 연간 가구 총소득이 높다고 볼 수 있다. 가구주의 특성은 남성이며, 연령대가 젊은 편이고, 혼인 중이고, 최종 학력도 고졸 이상으로 높은 편이고, 경제활동을 하고 있는 것으로 나타났다. 이러한 결과는 한국의료패널조사의 표본이탈 경향 분석에서 나타난 패널가구 및 패널가구원의 특성과 유사하게 나타났다(이혜정 외, 2023). 즉, 조사에 참여하더라도 소극적으로 응답하는 경향이 있다고 볼 수 있다.

가명정보 결합을 통해 구축된 통합데이터의 정합성 검토를 통해 행정데이터의 활용 가능성을 확인하였다. 행정데이터와의 비교를 통해 조사데이터의 응답의 정확성을 파악할 수 있으며, 조사 수행 및 실사 관리에서 정확한 응답을 수집할 수 있도록 유용한 정보로 활용될 수 있다. 또한 향후 행정데이터의 정보를 연계하여 응답자의 부담을 낮추는 데 기여할 수 있다고 생각한다.

두 번째 분석 주제에 대한 결과를 살펴보면, 두 가지로 요약할 수 있다. 첫째, 맞춤형 급여 수급 규모가 작은 가구일수록 의료비 지출 수준이 낮은 것으로 나타났다. 한편 본인부담금 수준이 높은 의료급여 2종 급여의 대상은 상대적으로 의료비 지출에 대한 부담 수준이 높은 것으로 나타났다. 이를 통하여 맞춤형 급여 수급액이 높은 저소득층과 근로무능력층의 의료비 지출이 의료서비스의 필요에 따라 충분히 이루어지고 있는지를 살펴볼 필요가 있음을 알 수 있다. 또한 각 수급 자격에 따른 본인부담금 수준과 의료비에 대한 부담 등을 다각적으로 살펴볼 필요가 있음을 알 수 있다.

둘째, 기초연금 수급 가구는 수급 자격에 따라 의료비 지출 특성이 상이하게 나타났다. 기초연금 ‘단독’ 수급 가구는 기초연금 ‘부부2인’ 가구

에 비해 의료비 지출 규모가 작았으나 의료비에 대한 부담이 큰 것으로 나타났다. 또한 기초연금 ‘단독’ 수급 가구는 급여 수준이 낮을수록 의료비 지출 규모가 큰 것으로 나타났으며, 기초연금 ‘부부2인’ 가구는 급여 수준이 높을수록 의료비 지출 규모가 큰 경향을 보였다. 이를 통하여 급여 자격별로 의료비 지출 특성을 구분하여 분석하여야 함을 알 수 있다.

특히 단독 수급 가구의 경우, 기준연금액 100%를 수급하는 가구의 의료비 지출 수준이 상대적으로 낮다는 점에 주목할 필요가 있다. 이 집단은 연금 수급을 통한 소득이 높지만 소득·재산의 수준이 상대적으로 낮을 가능성이 높다. 향후 이들의 건강상태나 의료비 지출에 대한 부담 등이 실제 의료비 지출에 대한 위축으로 작용하게 되는지 여부를 살펴볼 필요가 있다. 한편, 부부2인 가구의 경우 기준연금액 100%를 수급하는 가구의 의료비 지출이 상대적으로 높다는 점을 확인하였다. 이들은 부부가 모두 65세 이상이면서 기초연금 외의 소득이 없을 가능성이 높기 때문에 과도한 의료비 지출로 인한 가계부담 증대의 위험을 가지고 있다. 향후 부부2인 가구의 기초연금 소득의 대부분을 의료비로 지출하는 가구를 확인하여 이에 대한 보충적 지원 정책을 뒷받침할 수 있는 자료를 생성할 필요가 있다.

3. 결론 및 시사점

가명정보 결합 제도가 활성화되기 위한 방안을 제안하고자 한다.⁶⁾ 가명정보 이용자, 데이터 ‘보유기관’(결합전문기관 등), 가명정보 결합에 사용되는 결합키, 가명정보 결합 절차에 대한 측면으로 구분하여 살펴보고자 한다.

6) 전문가 자문 의견 결과를 통합적으로 정리하였음.

가명정보 이용자에 대한 지원 강화가 필요하다. 이용자는 데이터 보유 기관이 제공할 수 있는 데이터에 대한 정보에 쉽고 편리하게 접근할 수 있어야 한다. 이러한 정보를 한 곳에서 볼 수 있는 플랫폼이 있다면 유용할 것이라고 생각한다. 이용자에게 데이터 보유기관의 가명정보를 이용하는 절차를 자세하게 안내하는 것이 필요하다. 가명정보를 활용하거나 결합하기 위한 절차를 처음 접하는 데이터 이용자를 지원하는 제도가 유지 및 확대되어야 한다. 현재 개인정보위원회, 한국인터넷진흥원, 한국지능정보사회진흥원 등을 중심으로 선도 사례 발굴, 경진대회, 가명정보 활용 지원 서비스⁷⁾ 등이 이루어지고 있는 상황이나, 더 지속적으로 널리 확대될 필요가 있다고 생각한다. 특히 가명정보 이용자가 가명정보 활용 지원 서비스를 받을 수 있도록 적극적인 안내 및 홍보를 하는 것도 중요하다고 생각한다.

데이터 보유기관에 대한 지원 강화가 필요하다. 데이터 보유기관을 대상으로 한 컨설팅 서비스 제공, 가명정보 관련 지식이 없는 일반 업무 담당자에 대한 교육 실시, 가명정보 전문가 인력 확보, 관련 업무를 수행할 수 있도록 예산 지원, 가명정보 제공에 대한 인센티브제도 마련 등을 고려해 볼 수 있다.

한편, 가명정보 결합률을 높일 수 있도록 데이터 보유기관과 결합키 연계기관의 적극적인 역할이 필요하다. 보통 결합키로 이름, 생년월일, 성별을 조합하여 사용하고 있는데, 다양한 결합키 조합에 대한 연구가 필요하다고 생각한다. 또한, 데이터 보유기관은 가명정보 결합 사례별로 결합키 생성 시 투입되는 개인정보의 속성을 확인해야 한다. 결합키 연계기관은 결합률 향상을 위해 적극적으로 대응하고 사례 연구 결과를 공유해야 한다.

7) 가명정보 활용 지원 서비스에는 가명·익명 처리 관련 법·제도 등 교육, 데이터 활용에 대한 맞춤형 컨설팅, 가명정보 결합 지원 등 안전한 가명정보 활용을 위한 기술지원 서비스 등이 해당함.

가명정보 결합 절차의 측면에서 살펴보면 다음과 같다. 가명정보 결합 절차 간소화, 가명정보 결합 신청 양식의 일원화, 가명처리 수준 완화 등이 필요하다. 이러한 문제들은 ‘개인정보 안심구역’을 통해 많은 부분이 해결될 수 있다고 생각한다. 개인정보 안심구역은 제로 트러스트 보안모델 기반 안전조치, 사전·사후적 데이터 처리과정 통제 등 환경적 안전성을 갖추면, 기존에 사실상 제한했던 다양한 데이터 처리가 가능하도록 지원하는 제도다(전자신문, 2023.10.31.). 개인정보 안심구역 내에서는 데이터 분석 시 환경의 안정성 강화 수준에 비례하여 가명처리 수준을 적정 수준으로 완화하거나, 연계정보 일부 등 다양한 결합키 활용을 허용할 수 있다. 또한, 시계열 분석 등 지속적·반복적 연구를 위한 가명정보의 장기간 보관과 제3자 재사용 등의 활용성을 강화할 수 있을 것이다. 현재는 개인정보 안심구역을 시범으로 운영할 계획이나, 가명정보를 더 유연하게 활용할 수 있도록 개인정보 안심구역 제도가 정착되고 활성화될 필요가 있다.

주요 용어 : 가명정보, 가명정보 결합, 한국의료패널조사, 한국사회보장정보원, 행정데이터

사람을
생각하는
사람들



KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



제 1 장

서론

제1절 연구의 배경 및 목적

제2절 연구 내용 및 방법

제 1 장 서론

제1절 연구의 배경 및 목적

행정안전부는 2023년 국가중점데이터 개방사업 통합 착수보고회를 7월 21일에 개최하고 ‘2023년 국가중점데이터 개방사업’을 본격 추진한다고 밝혔다(행정안전부 보도자료, 2023.7.20.).¹⁾ 올해 연말까지 공공데이터포털(data.go.kr)을 통해 30개의 국가중점데이터가 개방될 계획이다. 이 국가중점데이터 개방사업은 민간에서 서비스를 완성하는 데 필요한 공공데이터를 묶음(패키지) 형태로 제공하여 활용도를 높이고, 개인정보 등 민감한 정보가 포함된 데이터는 진위 확인 서비스 형식으로 데이터를 개방하는 것이 특징이다.

같은 날자에 개인정보보호위원회, 기획재정부 등 관계부처 합동으로 ‘가명정보 활용 확대 방안’을 발표하였다(개인정보보호위원회 보도자료, 2023.7.20.). 이는 가명정보 활용을 위한 데이터 확보의 어려움, 가명처리·결합 절차의 이행부담 및 제약요인, 가명정보 활용 인프라·지원 부족, 가명정보 재식별·유출에 대한 안전성 우려 등을 해소하는 방안을 마련하는 데 목적이 있다. 가명정보 활용 확대 방안의 주요 내용은 양질의 데이터 제공·공유 확대, 가명정보 활용 절차 합리화, 가명정보 활용지원 확대, 가명정보 활용 절차의 안전관리 강화에 대한 것이다.

이렇듯 정부의 적극적이고 지속적인 공공데이터 개방, 가명정보 활용

1) 행정안전부 보도자료. (2023. 7. 20.) 국민생활 밀접한 데이터, 묶음으로 개방 국민 편의 높이고 데이터 경제 활성화한다.

확대 방안 마련 등의 조치로, 데이터를 활용할 수 있는 분야는 넓어지고 있으며 환경도 좋아지고 있는 상황이다.

한편, 2020년 8월 5일 데이터 3법²⁾이 개정·시행됨에 따라 개인정보보호위원회는 가명정보 결합제도를 도입하여 이종 산업 간 데이터의 안전한 연계를 통해 새로운 부가가치를 창출하고자 하였다(과학기술정보통신부 보도자료, 2022.2.18.). 이에 따라 개인정보보호위원회와 관계부처는 가명정보 제도를 확산시키기 위하여 2021년에 5대 분야 7건의 시범사례를 발굴하여 결합 성과를 발표하였다(〈표 1-1〉 참조). 2022년에는 가명정보 결합 4건의 중점 선도 사례를, 올해(2023년)는 사례 건수가 더 늘어난 9건으로 주제도 더 다양해졌다.

〈표 1-1〉 연도별 가명정보 결합 선도 사례(2021~2023년)

연도	분야	선도 사례	연구기관	결합 대상 정보
2021 (1기)	의료+인구	폐암 환자의 사망동향 및 사망예측 연구	국립암센터	(암센터) 암 환자정보 (국민건강보험공단) 암 환자 진료정보 (통계청) 사망정보
		암 환자의 장기합병증 및 만성질환 예측 연구	국립암센터	(암센터, 병원) 약 20만 명의 암 환자 정보 (국민건강보험공단) 암 환자 진료정보
	통신+유통	불법스팸 실태연구	한국인터넷진흥원	(한국인터넷진흥원) 스팸신고정보 (SK텔레콤) 고객정보
		지역별, 상권별, 상품별 소비패턴 분석	케이티, 롯데멤버스	(KT) 이동통신 고객정보 (롯데멤버스) 유통소비정보
	레저+건강	맞춤형 산림치유프로그램 분석	국립산림과학원	(국민건강보험공단) 의료 건강정보 (주)비글) 운동 활동정보
	소득+복지	노후소득보장 종합연구	보건복지부 사회보장위원회	(국세청) 약 340만 명 대상 소득정보 (행안부) 세대·인적정보 (은행사) 개인연금정보

2) 데이터 3법은 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」인.

연도	분야	선도 사례	연구기관	결합 대상 정보
	금융+보훈	국가보훈대상자 신용 실태연구	국가보훈처	(국가보훈처) 보훈대상자 정보 (신용정보원) 보훈대상자 신용정보
2022 (2기)	소득 불평등	건강보험 급여에 따른 소득분배 영향 분석	가천대학교	(국민건강보험공단) 건강보험 가입현황 및 보험료·급여 정보 등 (통계청) 가계금융복지조사 정보
	의료 형평성	생애주기에 따른 의료이용 실태분석 및 형평성 비교	심사평가원 서울대학교	(한국노동연구원)노동패널정보 (건강보험심사평가원) 진료자료 (국민건강보험공단) 검진·장기요양자료 등
	장애인 복지	장애인 복지정책 효과 분석	서울대학교	(국민건강보험공단) 국민건강보험 정보 (사회보장원) 장애인 연금정보 등
	저탄소 경제	친환경차 충전 인프라 수요예측 및 최적입지 선정	성남시 티맵모빌리티	(성남시) 성남시 차량등록정보 (티맵모빌리티) Tmap 운행이력정보
2023 (3기)	복지(노인)+경제	고위험 취약노인 예측 모형개발·고령자 지원 정책 효과 분석	KDI 국제정책대학원	(복지부) 기초연금, 노인돌봄 등 (고용부) 고용보험 취득·상실정보 등 (건보공단) 보험자격 등
	인구+사회	인구감소지역 지원을 위한 생활인구 산정 및 활성화 방안	행안부, 통계청	(행안부) 주민등록 (법무부) 외국인등록 (통신사) 이동통신
	복지(여성)+경제	일·가정 양립 정책이 여성 근로자의 경제활동·출산에 미치는 영향 분석	서울대	(고용부) 육아휴직 관련 등 (건보공단) 보험자격 등
	의료+건강	실업이 의료서비스 이용과 국민 건강에 끼치는 영향 분석	고려대	(고용부) 고용보험 취득·상실정보 등 (건보공단) 보험자격 등
	사회+복지	공공부조 제도가 확대가 대상자의 고용·건강·의료 이용에 미치는 영향 분석	강남대	(고용부) 고용보험 취득·상실정보 등 (복지부) 기초생활보장 수급 (건보공단) 보험자격 등
	복지(아동)+건강	복지서비스 이용 아동과 청년의 서비스 이용실태 및 효과분석	아동권리보장원	(권리보장원) 아동 신체 마음·건강 등 (사회보장원) 급여 서비스 내역 등
	교육+취업	졸업 후 경력경로에 기반한 맞춤형 학생역량 시스템 개발	군산대	(군산대) 졸업생 학적 등 (건보공단) 건강보험 정보

20 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

연도	분야	선도 사례	연구기관	결합 대상 정보
	지역 소비 +경제	부산 지역화폐 운영 정책개선을 위한 이용행태 분석	부산시, 부경대	(부산시) 지역화폐 승인내역 (롯데멤버스) 구매이력
	지역 의료 +건강	코로나19 환자치료 지역격차 분석	강원도	(질병청) 코로나19 역학조사 (통계청) 사망원인 정보 (강원도) 코로나19 역학조사 (심사평가원) 진료·처방정보 등

자료: 과학기술정보통신부 보도자료(2022. 2. 18.) 및 개인정보보호위원회 보도자료(2023. 5. 4.)
의 내용을 재구성함.

2023년 현재 가명정보 결합전문기관은 23개 기관³⁾이며, 데이터 전문 기관은 올해 7월 비씨카드 등의 8개 민간기업이 새롭게 추가 지정되어 총 12개⁴⁾로 늘어났다. 이러한 민간기업의 참여로 인하여 다분야·이종데이터 간 결합은 더욱더 활성화될 것으로 기대하고 있다(금융위원회 보도자료, 2023.7.19.).

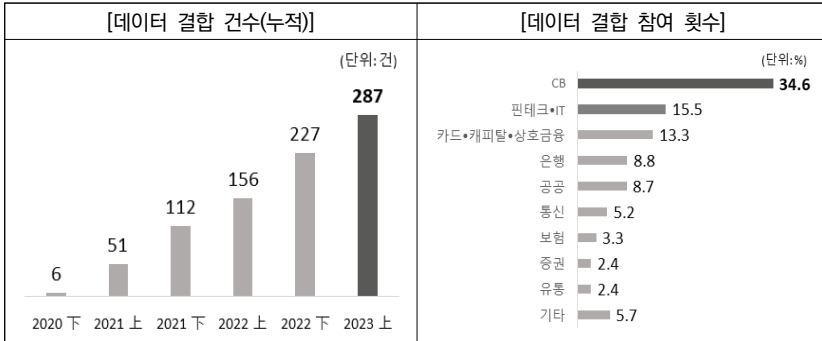
이러한 환경에서 데이터 결합 현황을 살펴보면 다음과 같다. 금융위원회에서 발표한 결과로(금융위원회 보도자료, 2023.7.19.),⁵⁾ 2020년 6건으로 시작된 데이터 결합은 점점 증가하여 2023년 6월 말 기준 총 287건(누적)으로 나타났다. 데이터 결합 참여 횟수는 CB(신용정보회사)가 34.6%로 가장 많았고, 테크·IT 기업(15.5%), 카드·캐피탈·상호금융 기업(13.3%) 순으로 나타났다([그림 1-1] 참조).

3) 개인정보위원회 11개(통계청, 삼성SDS, 케이씨에이, 한국지역정보개발원, 롯데정보통신, 한국정보인증, 신세계아이앤씨, 한국사회보장정보원, 국세청, LG CNS, CJ 주식회사), 보건복지부 3개(국민건강보험공단, 건강보험심사평가원, 국립암센터), 과기정통부 5개(한국지능정보사회진흥원, SK 주식회사, 더존비즈온, 한국데이터산업진흥원, BC카드), 국토교통부 1개(한국도로공사), 교육부 1개(한국교육학술정보원), 산업통상자원부 1개(한전 KDN), 행정안전부 1개(국가정보자원관리원)

4) 금융위원회 12개(한국신용정보원, 금융보안원, 국세청, 금융결제원, 비씨카드, 삼성SDS, 삼성카드, 신한은행, 신한카드, LG CNS, 쿠팡, 통계청)

5) 참고로 데이터 결합 현황은 금융위원회에서 추진된 결과만을 정리한 것으로, 다른 정부 관계부처에서의 현황은 포함되지 않은 결과임.

[그림 1-1] 데이터 결합 건수(누적)와 데이터 결합 참여 횟수



주: 2023년 6월 말 기준임.

자료: 금융위원회 보도자료(2023. 7. 19.) 내용을 재인용함.

〈표 1-2〉는 데이터 결합 분야에 대한 것으로, 비금융 간 데이터 결합이 54%로 금융 간 데이터 결합(46%)보다 8%p 높게 나타났다. 비금융 간 데이터 결합에서는 공공분야(13.2%)가 가장 높았고, IT(12.5%), 통신(9.1%)이 그 뒤를 이었다. 둘 이상의 비금융 데이터가 결합된 다중결합도 12.2%를 차지했다.

다음으로 분야별 데이터 결합 목적을 살펴보면, CSS(신용평가시스템) 개발이 42.2%로 가장 높았고, 통계작성·연구 등(21.6%), 기타 모형 개발(19.5%) 순으로 나타났다. 통계작성·연구 등의 목적으로 한 데이터 결합은 62건인데, 학술적 연구를 진행한 논문이나 연구보고서가 많지 않은 편으로 나타나 주로 통계 작성을 위해 활용했다고 보인다.

종합해 보면, 데이터 개방 추세에 따라 데이터가 다량으로 개방되고 있을 뿐만 아니라 여러 가지 법, 제도 등의 도입·개정으로, 여러 개의 데이터를 서로 연계하고 통합한 데이터로 활용할 수 있는 환경이 조성되고 있는 상황이다. 그러나 데이터가 개별적으로 분산된 채 제공되고 있고, 여러 개의 데이터를 서로 연계하고 통합한 데이터로 활용하는 것은 저조하다는 평가가 이어지고 있다.

22 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

〈표 1-2〉 분야별 데이터 결합 분야

(단위: 건, %)

금융&금융	금융&비금융						
		공공	IT	통신	유통	기타	다중결합
132 (46.0)	155 (54.0)	38 (13.2)	36 (12.5)	26 (9.1)	11 (3.8)	9 (3.1)	35 (12.2)

주: 2023년 6월 말 기준임.
자료: 금융위원회 보도자료(2023. 7. 19.) 내용을 재인용함.

〈표 1-3〉 분야별 데이터 결합 목적

(단위: 건, %)

CSS 개발	통계작성·연구 등	기타 모형 개발	전략 수립·신상품 개발	합계
121 (42.2)	62 (21.6)	56 (19.5)	48 (16.7)	287 (100.0)

주: 2023년 6월 말 기준임.
자료: 금융위원회 보도자료(2023. 7. 19.) 내용을 재인용함.

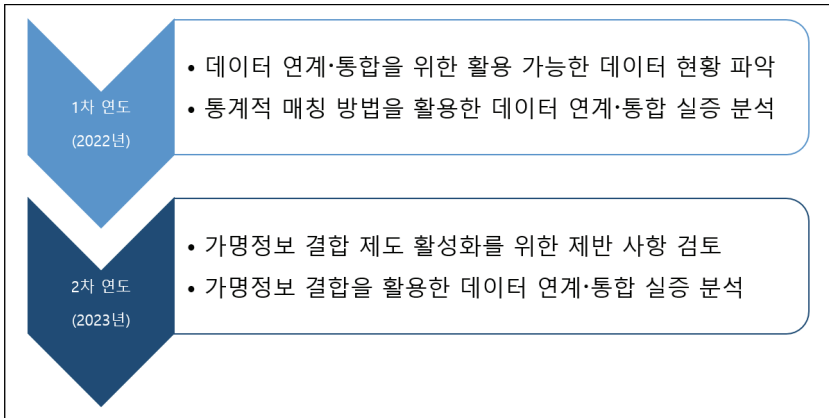
이 연구의 목적은 곳곳에 분산된 데이터의 활용 가치를 높이기 위해서 다출처 데이터를 연계·통합하는 시도를 해 보고, 이에 따른 어려움을 해결하여 통합 데이터 활용이 활성화될 수 있도록 하는 방안을 마련하는 것이다. 이 연구는 여러 분야의 데이터를 연계·통합하여 풍부한 정보를 포함하는 강력한(powerful) 데이터를 구현하여 사회 현상을 이해하고 미래를 예측하여 정책을 제안하고 평가하는 데 기반이 되고자 한다. 즉 데이터의 활용 가치를 높이고 확산하며 정책 수립 및 효과 분석 시 증거 기반의 의사결정을 가능하게 하여 정확한 정책과 맞춤형 서비스를 제공하는 데 기여하고자 한다.

제2절 연구 내용 및 방법

1. 연구 내용

이 연구는 다양한 주제별로 분산된 데이터 자원을 연계·통합하는 적극적인 시도를 통해 데이터의 활용 가치를 높이는 연구로, 2년 동안 이루어졌다. 올해가 마지막 연도 과제이고, 연차별 연구 목적은 [그림 1-2]와 같다. 1차 연도(2022년)에는 데이터 연계·통합을 위한 활용 가능한 데이터 현황을 파악하였고, 통계적 매칭 방법을 활용하여 데이터를 연계·통합하는 실증 연구를 하였다. 2차 연도(2023년)에는 가명정보 결합을 활용한 데이터 연계·통합 실증 연구를 실시하고, 데이터 연계·통합을 활성화하기 위해서 가명정보 결합과 관련된 제반 사항을 검토하고자 한다.

[그림 1-2] 연차별 연구 목적



자료: 필자 작성

먼저, 1차 연도의 주요 연구 내용에 대해 정리한 후에 2차 연도 연구 내용을 살펴보고자 한다.

1차 연도 연구에서는(이혜정 외, 2022) 데이터 연계·통합의 개념과 데이터 통합 방법론에 대해 주로 살펴보았다. 이때 데이터 통합 방법론은 자료연계와 통계적 매칭으로 구분하여 초기부터 최근 연구 동향을 파악하였다. 그런 다음 국내외 데이터 통합 현황과 통합데이터를 활용한 연구 사례를 소개하였다. 활용 가능한 데이터 현황 파악을 위해 우리 연구원에서 제공 중인 조사데이터와 국내의 개방된 데이터 포털 및 플랫폼을 중심으로 살펴보았다. 우리 연구원에서 제공 중인 조사데이터 중에서 주요 조사데이터의 공통변수와 유일변수에 대해 정리하였다. 국내의 개방된 데이터 포털 및 플랫폼은 과학기술 및 공공 분야와 보건 의료 분야를 중심으로 살펴보았다. 이를 통해 국내의 개방된 데이터 자원을 활용하여 우리 연구원 조사데이터 등과 연계, 통합, 활용하는 방안을 모색하였다. 마지막으로 실제 데이터를 활용하여 모의실험과 데이터 통합 및 분석을 실시하였다. 조사데이터 간의 데이터 통합으로, 2기 한국의료패널조사 2019년 연간데이터는 기준데이터이고, 생활시간조사 데이터는 제공데이터로 선정하였다. 모의실험은 한국의료패널조사 데이터를 사용하여 통계적 매칭 방법을 활용하여 데이터를 통합하는 것이었다. 그 방법으로는 다양한 거리 함수에 따른 최근접 이웃 핫덱과 랜덤 핫덱이 해당하였고, 모의실험을 통해 최적의 통계적 매칭 방법을 선정하였다. 모의실험 결과를 기반으로 하여 실제 한국의료패널조사와 생활시간조사에 대한 데이터를 통합하였다. 그 결과물인 통합데이터에 대한 품질을 평가한 다음에 데이터 통합 목적에 따라 데이터 분석을 실시하였다. 통합데이터 분석 결과, 다출처 조사데이터에 대한 통계적 매칭 방법을 이용한 통합데이터의 구축 가능성을 확인하였다. 또한, 주관적 건강상태, 스트레스 인지 정도와 생활시간 사용과의 관계에 대해 통합데이터를 분석한 연구 결과는 선행연구와 유사한 결과를 보였다.

2차 연도 연구 내용은 다음과 같다.

제2장은 가명정보 결합 개념 및 활용 방법으로, 가명정보 결합과 관련된 개념을 정리한다. 그리고 가명정보 결합과 관련된 법적 근거, 가명정보 결합 및 반출 절차에 대해 살펴보려고 한다.

제3장에서는 국내외에서 가명정보 결합을 활용하여 통합데이터를 구축하고 연구한 사례를 고찰하고자 한다. 국내 사례의 경우 통계청의 가계금융복지조사 데이터와 행정데이터, 통계기업등록부와 행정데이터 간의 결합, 그리고 행정데이터 간의 결합에 대해 살펴보려고 한다. 조사데이터와 행정데이터 간의 결합을 살펴본 뒤, 해외 사례의 경우 영국은 조사데이터와 행정데이터 간의 결합을, 포르투갈과 일본은 행정데이터 간의 결합을 중심으로 살펴보려고 한다.

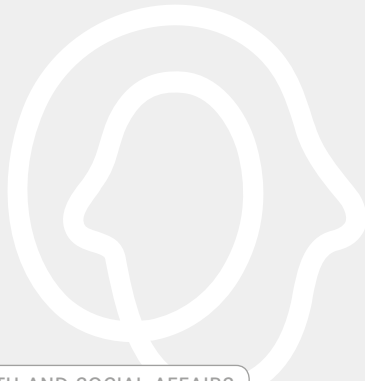
제4장에서는 가명정보 결합 제도의 활성화를 위하여 한국보건사회연구원의 데이터 제공에 대한 맞춤형 방안을 마련하고자 한다. 이를 위해 다양한 법령과 가명정보 관련 가이드라인을 종합적으로 분석하고자 한다. 그리고 현행 한국보건사회연구원의 데이터 제공 절차에 대한 규정과 절차를 검토하고자 한다. 이를 기반으로 하여 한국보건사회연구원의 데이터 제공 절차를 수립하고자 한다.

제5장은 가명정보 결합으로 구축된 통합데이터에 대한 실증 연구로, 가명정보 결합 절차에 따라 실제로 2기 한국의료패널조사 데이터와 한국사회보장정보원 사회보장행정시스템의 행정데이터의 통합을 실시하고자 한다. 가명정보 결합을 활용한 데이터 통합 과정에 대해 정리하고, 가명정보 결합으로 구축된 통합데이터에 대한 실증 분석을 한다. 데이터 통합 목적은 한국의료패널조사 데이터의 품질 제고를 위해 사회보장정보시스템의 행정데이터와의 정합성을 검토하는 것이며, 이를 위해 사회보장 급여 지급자의 의료비 지출 특성을 살펴보려고 한다.

제6장은 결론으로 주요 연구 결과를 요약하고, 가명정보 결합 제도 활성화를 위한 방안을 제시한다. 마지막으로 향후 과제에 대해 논의하고자 한다.

2. 연구 방법

국내외 문헌 연구, 가명정보를 결합한 통합데이터 분석, 전문가 자문회의 등을 다양하게 활용하여 연구한다. 제5장에서 사용한 데이터는 2기 한국의료패널조사의 2019년 연간데이터와 한국사회보장정보원 사회보장행정시스템의 행정데이터를 가명정보 결합을 통해 구축한 통합데이터이다. 통계패키지는 SAS 및 Stata를 사용한다.



제2장

가명정보 결합 개념 및 활용 방법

제1절 가명정보 결합의 개념

제2절 가명정보 결합 및 반출 절차

제3절 소결

제2장 가명정보 결합 개념 및 활용 방법

제1절 가명정보 결합의 개념

1. 가명정보 결합의 개념

데이터 이용 수요가 급증하는 가운데, 데이터 3법⁶⁾이 시행(2020.8.5.) 되면서, 정보 주체⁷⁾인 개인의 동의 없이, 통계 작성, 과학적 연구, 공익적 기록보존 등을 위한 목적으로 개인정보를 가명처리하여 사용할 수 있게 되었다(개인정보보호법 제28조의 2 제1항, 제28조의 3 제1항). 가명처리란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말한다(개인정보보호법 제2조 제1의 2호).

가명정보의 결합이란 개인 식별이 불가능하도록 가명처리된 가명정보에 대하여 결합전문기관⁸⁾이 두 개 이상의 데이터를 결합키⁹⁾를 이용하여 통계 작성, 과학적 연구, 공익적 기록보존 등의 목적으로 결합하는 것이다.

6) 데이터 3법은 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」임.

7) '정보 주체'란 처리되는 정보에 의하여 알아볼 수 있는 사람으로서 그 정보의 주체가 되는 사람을 말함(개인정보보호법 제2조 제3호).

8) '결합전문기관'이란 서로 다른 개인정보처리자 간의 가명정보의 결합을 위하여 개인정보보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관을 말함(개인정보보호법 제28조의 3).

9) '결합키'란 결합 대상 가명정보의 일부로서, 서로 다른 가명정보를 결합할 때 매개체로 이용되는 값을 말함. 예를 들어, 이름, 성별, 생년월일 등의 정보를 하나의 변수로 만들어 사용할 수 있음.

우리나라에서 행정데이터와 설문조사의 가명정보의 결합 사례는 아직 초기단계이다. 제한적인 목적에 한해서 가명정보 결합을 위해 개인의 동의를 받지 않아도 된다는 편리함은 있지만, 개인 식별 위험성을 방지하기 위한 엄격한 법적 근거와 복잡한 절차들을 두고 있다. 이에 본 장에서는 가명정보의 법적 근거와 가명정보의 결합 및 반출 절차에 대하여 살펴보고자 한다.

2. 가명정보 결합의 법적 근거 및 운영 가이드라인

가명정보 결합은 개인정보 보호 차원에서 모든 과정들에 대하여 「개인정보보호법」과 「개인정보보호법 시행령」에 명시되어 있다.

〈표 2-1〉은 가명정보 처리 및 결합, 반출에 관한 개인정보 보호법 조문이다. 개인정보 보호법 제28조의 2는 개인정보의 처리에 관한 내용으로 가명처리가 허용되는 목적을 통계 작성, 과학적 연구, 공익적 기록보존으로 제한함을 기술하였다. 제17조 제4항은 정보 주체에게 불이익이 발생하지 않고, 암호화 등의 안전성 확보에 필요한 조치를 한 경우에는 정보 주체의 동의 없이 개인정보를 제공할 수 있다는 내용이다. 이 조문을 통해서 개인의 동의 없이 설문조사나 행정정보의 가명정보를 결합하는 것이 현실적으로 가능하게 되었다. 제3조 제7항은 익명정보 사용으로 목적 달성이 가능하다면 익명정보를 사용하고, 목적 달성이 불가능한 경우에만 가명정보를 사용해야 한다고 명시하였다. 제28조의 3은 가명처리된 데이터들 간의 결합은 결합전문기관에서만 가능하며, 결합 및 반출 과정은 대통령령으로 정함을 고시하였다.

〈표 2-1〉 가명정보 처리 및 결합 관련 「개인정보 보호법」 내용

조문	조문 내용
제28조의 2	① 개인정보처리자는 통계 작성, 과학적 연구, 공익적 기록보존 등을 위하여 정보 주체의 동의 없이 가명정보를 처리할 수 있다. ② 개인정보처리자는 제1항에 따라 가명정보를 제3자에게 제공하는 경우에는 특정 개인을 알아보기 위하여 사용될 수 있는 정보를 포함해서는 아니 된다.
제17조 4항	개인정보처리자는 당초 수집 목적과 합리적으로 관련된 범위에서 정보 주체에게 불이익이 발생하는지 여부, 암호화 등 안전성 확보에 필요한 조치를 하였는지 여부 등을 고려하여 대통령령으로 정하는 바에 따라 정보 주체의 동의 없이 개인정보를 제공할 수 있다.
제3조 7항	개인정보처리자는 개인정보를 익명 또는 가명으로 처리하여도 개인정보 수집 목적을 달성할 수 있는 경우 익명처리가 가능한 경우에는 익명에 의하여, 익명처리로 목적을 달성할 수 없는 경우에는 가명에 의하여 처리될 수 있도록 하여야 한다.
제28조의 3	① 제28조의 2에도 불구하고 통계 작성, 과학적 연구, 공익적 기록보존 등을 위한 서로 다른 개인정보처리자 간의 가명정보의 결합은 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관이 수행한다. ② 결합을 수행한 기관 외부로 결합된 정보를 반출하려는 개인정보처리자는 가명정보 또는 제58조의 2에 해당하는 정보로 처리한 뒤 전문기관의 장의 승인을 받아야 한다. ③ 제1항에 따른 결합 절차와 방법, 전문기관의 지정과 지정 취소 기준·절차, 관리·감독, 제2항에 따른 반출 및 승인 기준·절차 등 필요한 사항은 대통령령으로 정한다.

자료: 개인정보 보호법, 법률 제19234호 (2023).

〈표 2-2〉는 가명정보를 처리할 때 안전성 확보와 관련된 개인정보 보호법과 개인정보 보호법 시행령의 내용이다. 개인정보 보호법 제28조의 4는 가명정보를 처리하는 과정에서 별도 분리 보관과 정보가 훼손되지 않도록 기술적·관리적·물리적 조치를 의무화하였다. 개인정보 보호법 시행령 제29조의 5는 가명정보와 추가정보를 분리 보관하는 것과 접근권한 관리·통제에 관한 내용이다.

〈표 2-2〉 가명정보 처리 안전성 확보 관련 「개인정보 보호법」과 「개인정보 보호법 시행령」

내용	
조문	조문 내용
개인정보 보호법 제28조의 4	① 개인정보처리자는 제28조의 2 또는 제28조의 3에 따라 가명정보를 처리하는 경우에는 원래의 상태로 복원하기 위한 추가정보를 별도로 분리하여 보관·관리하는 등 해당 정보가 분실·도난·유출·위조·변조 또는 훼손되지 않도록 대통령령으로 정하는 바에 따라 안전성 확보에 필요한 기술적·관리적 및 물리적 조치를 하여야 한다. ② 개인정보처리자는 제28조의 2 또는 제28조의 3에 따라 가명정보를 처리하는 경우 처리목적 등을 고려하여 가명정보의 처리 기간을 별도로 정할 수 있다. ③ 개인정보처리자는 제28조의 2 또는 제28조의 3에 따라 가명정보를 처리하고자 하는 경우에는 가명정보의 처리 목적, 제3자 제공 시 제공받는 자, 가명정보의 처리 기간(제2항에 따라 처리 기간을 별도로 정한 경우에 한한다) 등 가명정보의 처리 내용을 관리하기 위하여 대통령령으로 정하는 사항에 대한 관련 기록을 작성하여 보관하여야 하며, 가명정보를 파기한 경우에는 파기한 날부터 3년 이상 보관하여야 한다.
개인정보 보호법 시행령 제29조의 5	① 개인정보처리자는 법 제28조의 4 제1항에 따라 가명정보 및 가명정보를 원래의 상태로 복원하기 위한 추가정보(이하 이 조에서 “추가정보”라 한다)에 대하여 다음 각 호의 안전성 확보 조치를 해야 한다. 1. 제30조 또는 제48조의 2에 따른 안전성 확보 조치 2. 가명정보와 추가정보의 분리 보관. 다만, 추가정보가 불필요한 경우에는 추가정보를 파기해야 한다. 3. 가명정보와 추가정보에 대한 접근 권한의 분리. 다만, 「소상공인기본법」 제2조에 따른 소상공인으로서 가명정보를 취급할 자를 추가로 둘 여력이 없는 경우 등 접근 권한의 분리가 어려운 정당한 사유가 있는 경우에는 업무 수행에 필요한 최소한의 접근 권한만 부여하고 접근 권한의 보유 현황을 기록으로 보관하는 등 접근 권한을 관리·통제해야 한다. ② 법 제28조의 4 제3항에서 “대통령령으로 정하는 사항”이란 다음 각 호의 사항을 말한다. 1. 가명정보 처리의 목적 2. 가명처리한 개인정보의 항목 3. 가명정보의 이용내역 4. 제3자 제공 시 제공받는 자 5. 그 밖에 가명정보의 처리 내용을 관리하기 위하여 보호위원회가 필요하다고 인정하여 고시하는 사항

자료: 개인정보 보호법, 법률 제19234호 (2023).
개인정보 보호법 시행령, 대통령령 제33723호 (2023).

〈표 2-3〉은 가명정보 관련 제도 현황으로, 가명정보 처리는 모든 부처에서 하나의 통일된 방법을 사용하는 것이 아니라, 구체적인 사항은 개별

부처의 상황에 따라 적용하도록 함을 알 수 있다. 개인정보보호위원회에서 발표한 가이드라인에 따라 보건복지부, 교육부, 행정안전부, 금융위원회는 각각의 가이드라인을 가지고, 정보의 안전성과 처리자 및 연구자의 법적 안정성을 도모하고 있다.

〈표 2-3〉 가명정보 관련 제도 현황에 대한 참고

연번	구분	내용	소관부처
1	법률	개인정보 보호법	개인정보위
2	법률	신용정보의 이용 및 보호에 관한 법률	금융위
3	시행령	개인정보 보호법 시행령	개인정보위
4	시행령	신용정보의 이용 및 보호에 관한 법률 시행령	금융위
5	고시	가명정보의 결합 및 반출 등에 관한 고시	개인정보위
6	고시	신용정보업감독규정	금융위
7	가이드라인	가명정보 처리 가이드라인	개인정보위
8	가이드라인	보건의료 데이터 활용 가이드라인	보건복지부
9	가이드라인	교육분야 가명·익명정보 처리 가이드라인	교육부
10	가이드라인	공공분야 가명정보 제공 실무안내서	행정안전부
11	가이드라인	금융분야 가명·익명처리 안내서	금융위

자료: 장인성 외(2022)와 개인정보보호위원회(2022)의 내용을 참고하여 필자 작성

다음 제2절에서는 각각의 가이드라인을 아울러서 가명정보 결합 및 반출 절차를 설명하였다.

제2절 가명정보 결합 및 반출 절차

가명정보 결합 및 반출 절차는 사전준비, 결합신청, 가명처리, 결합수행, 반출 및 관리로 구성된다([그림 2-1] 참조).

[그림 2-1] 가명정보 결합 절차(개요)

1단계	2단계	3단계	4단계	5단계
사전준비	결합신청	가명처리	결합수행	반출 및 관리
데이터 이용자 (연구자)	데이터 제공자 (=결합신청자)	데이터 제공자 (=결합신청자)	결합전문기관	데이터 이용자 결합전문기관
• 목적 설정 • 데이터 협의 • 데이터 이용신청 • 적합성 검토	• 결합 신청 • 결합키 생성 • 결합률 • 결합 유형	• 대상 선정 • 위험도 측정 • 가명처리 수준 정의 • 가명처리 • 적정성 검토 • 추가 가명처리 • 데이터 전송 • 사후관리	• 결합 • 적정성 검토 • 추가 가명처리	• 데이터 분석 • 반출 신청 • 반출심사위원회 • 사후관리

자료: 필자 작성

해당 절차를 단계별로 개략적으로 살펴보면, 사전준비 단계에서는 데이터를 이용하고자 희망하는 연구자의 역할이 가장 중요하다. 연구 목적을 설정하고, 데이터 보유기관과 데이터 내용과 제공 가능 범위를 협의하면서, 가명처리한 정보가 연구 목적을 달성할 수 있는지 판단해야 한다. 데이터의 형태가 개략적으로 정해지면, 데이터 보유기관에 가명정보 데이터 이용 제공 신청서를 제출한다. 결합신청을 받은 데이터 보유기관은 데이터 심의위원회에서 연구 목적이 가명정보 결합을 통해 달성될 수 있는지, 결합 시 정보의 안전성이 있는지, 데이터 활용 시 안전성이 있는지 등을 검토하여 제공 승인 여부를 결정한다.

2단계에서는 데이터 보유기관 중 대표 기관이 한국인터넷진흥원에 결합전문기관을 지정하고, 결합을 신청한다. 데이터 보유기관이 협의하여 결합키를 생성하고, 한국인터넷진흥원은 결합키를 이용하여 데이터를 결합한다. 연구자는 결합키에 의해 결합된 데이터의 결합률¹⁰⁾을 확인하고, 결합 방법을 결정한다.

3단계는 데이터 보유기관에서 진행되는 가명처리 과정이다. 가명처리는 특정 개인을 알아 볼 수 없도록 처리하는 것으로, 연구에 사용될 각각의 변수에 모두 시행한다. 가명처리에 앞서, 가명처리할 대상자와 변수를 선정하고, 가명처리 환경과 기관의 개인정보 보호 수준 등을 고려하여 위험도를 측정하며, 이에 준하여 가명처리 수준을 정의한다. 삭제, 평균 사용, 범주화 등의 다양한 방법을 이용하여 개인이 식별되지 않도록 가명처리를 실시한다. 데이터 심의위원회에서는 가명처리가 적절하게 수행되었는지, 처리 이후에 연구 목적이 달성될 수 있는지를 심의한 후 결합전문기관에 데이터를 발송한다. 개별 데이터 보유기관은 결합전문기관에 데이터를 보낸 후 데이터의 관리 및 보완에 관하여 안전한 사후관리를 수행한다.

4단계는 결합전문기관에서 수행하는 가명처리된 데이터의 결합이다. 결합전문기관은 개별 정보를 보유하고 있는 기관으로부터 데이터를 제공받고, 이를 2단계에서 생성한 결합키 연계정보인 일련번호를 이용하여 각각의 데이터를 결합한다. 결합전문기관의 데이터 심의위원회에서 결합한 데이터의 적정성 평가를 다시 시행한 후, 개인 식별이 우려되는 정보에 대해서는 추가 가명처리를 진행한다.

5단계는 결합전문기관에서 수행하는 반출 과정이다. 결합된 데이터의

10) 결합키 결합에서는 연구에 필요한 모든 정보를 결합하는 것이 아니라, 개인정보(예를 들어, 성명, 성별, 생년월일 등)로 만들어진 결합키만 결합함. 모든 정보의 결합은 3단계 가명처리 후 4단계 결합전문기관에서 수행됨.

적정성 평가가 완료된 이후에 연구자는 결합전문기관이 제공한 장소에서 데이터를 분석하고, 분석 결과표는 결합전문기관에 반출 신청을 한다. 반출 심의위원회는 결합 목적과 개인의 식별 가능성, 결합된 정보를 관리적·기술적·물리적으로 안전한 환경에서 이용하는지 등에 대하여 심의한다. 데이터 보유기관에서는 반출 기간 이후 데이터를 폐기 처분하는 등의 사후관리를 수행한다.

가명정보 결합에 앞서 가명정보 결합을 시행할 때 각자의 역할을 숙지하는 것이 중요하다([그림 2-2] 참조).

[그림 2-2] 가명정보 결합·반출 단계 및 관련 기관의 역할

연구자	A, B, C 기관	결합 전문기관	결합기 관리기관 (한국인터넷진흥원)
① 연구 목적 설정 및 데이터 협의 ① 가명정보 이용·제공 신청서 제출 - A, B, C 기관	② 적합성 검토 후 데이터제공 결정		
	③ 결합 신청(대표) - 한국인터넷진흥원		④ 결합신청번호발급
	⑤ 결합기 신청 - 한국인터넷진흥원		⑥ 결합률 확인 - 결합 방법 결정
	⑦ 결합신청서 제출 - 결합전문기관	⑧ 적합성 검토	
	⑨ 가명처리 - 적정성 검토 - 결합전문기관에 데이터 제출	⑩ 결합수행	
		⑪ 적정성 검토	
⑫ 데이터 이용	⑬ 반출 신청(대표) - 결합전문기관	⑭ 반출 심사	
	⑮ 사후관리	⑮ 사후관리	

자료: 필자 작성

예를 들어, 연구자가 A, B, C 기관에서 보유하고 있는 데이터들을 결합하여 연구를 수행하고자 한다. ① 연구자는 연구 목적을 설정하고, 정보를 보유한 기관과 데이터 구성과 제공 범위에 대하여 사전 협의한다. ② 연구자는 A, B, C 기관의 데이터 담당 부서에 데이터 활용 신청서를 제출한다. ③ 신청서를 수령한 3개 기관은 연구의 목적이 가명정보 결합으로 달성 가능한지, 그리고 가명정보 처리 후 안전성 등을 각각 검토하여, 데이터 제공을 결정한다. ④ 데이터 제공이 승인되면, 3개 기관 중 대표 기관을 선정하여, 가명정보결합종합지원시스템¹¹⁾에 결합을 신청한다. ⑤ 한국인터넷진흥원은 결합신청번호를 발급하여 3개 기관에 전달하고, ⑥ 3개 기관은 시스템에 접속하여 결합키 생성을 위한 결합신청서를 각각 작성한다. ⑦ 한국인터넷진흥원에서는 3개 기관과 협의하여 결합키를 생성하고, 생성된 결합키를 이용하여 결합을 실시한다. 이 과정은 데이터의 전체를 결합하는 것이 아니라, 결합키만 결합하는 것으로, 대상자의 결합 정도만을 확인할 수 있다. 연구자는 결합률을 확인하고, 결합 방법을 결정한다. ⑧ A, B, C 기관은 데이터의 실제 결합을 위해 결합전문기관에 가명정보 결합신청서를 제출하고, ⑨ 결합전문기관은 가명처리 대상 데이터의 개인 식별 위험성을 평가(적합성 검토)하여, 가명정보 결합을 승인한다. ⑩ A, B, C 기관은 가명처리된 데이터의 적정성 검토를 통해 개인 식별 위험성이 없다는 것을 확인하고 결합전문기관에 데이터를 제출한다. ⑪ 결합전문기관은 결합 연계키를 이용하여 3개 기관의 정보를 결합한다. ⑫ 결합전문기관은 가명정보 결합으로 생성된 새로운 데이터에 대하여 연구 목적과 결합 시 안전성, 식별 위험성 등의 적정성 검토를 실시한다. 이 과정에서 추가적으로 가명처리가 발생하기도 한다. ⑬ 적정성

11) 가명정보결합종합지원시스템은 한국인터넷진흥원에서 위탁 운영하는 사이트임.
<https://link.privacy.go.kr/nadac/index.do>

검토에서 데이터 사용이 ‘승인’되면 연구자는 결합전문기관이 지정한 방법(가명정보 활용 형태, 처리 장소, 처리 방법)으로 신청 기간 동안 데이터를 분석한다. ⑬ 데이터 이용자는 분석 결과표를 결합전문기관에 반출 신청한다. ⑭ 결합전문기관은 반출심사위원회에서 결합된 데이터의 안전한 사용 가능성에 대해 심사한다. ⑮ A, B, C 기관과 결합전문기관은 가명처리하여 제공한 데이터의 관리적, 기술적, 물리적 보호 조치를 하고, 개인정보 처리방침에 따라 가명정보 처리를 공개해야 한다.

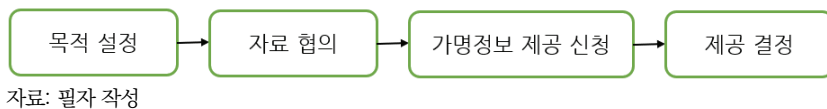
연구자는 ① 가명정보 이용·제공 신청서를 제출한 이후부터 ⑭ 반출 승인된 데이터를 사용하기까지 3개의 데이터 보유기관과 결합전문기관, 결합키 관리기관과 유기적으로 관계를 맺으며 협업해야 한다.

이상 간략하게 살펴본 가명정보 결합 및 반출 단계에 대하여 각 단계의 세부 과정을 구체적으로 살펴보고자 한다.

1. 사전준비

사전준비 단계는 연구자가 연구 목적을 달성하는 데 필요한 데이터를 보유하고 있는 기관과 협의하고, 가명정보 이용·제공 신청서를 제출, 가명정보 이용·제공 신청서의 적합성 검토 후 제공 결정하는 과정이다.

[그림 2-3] 사전준비 단계



가. 목적 설정

연구자는 개인정보보호법 제28조의 2에 따라 통계 작성, 과학적 연구, 공익적 기록 보존 등의 목적일 때에 한해, 정보 주체의 동의 없이 가명정보 처리를 할 수 있다는 것을 명심해야 한다.

‘통계 작성’을 위한 가명정보 처리는 통계 작성을 위해 가명정보를 이용, 분석, 제공하는 것 등을 말하며, 시장조사를 위한 통계 작성 시에도 가능하다. ‘과학적 연구’를 위한 가명정보 처리는 체계적이고 객관적인 방법으로 검증 가능한 질문에 대해 연구하는 것으로, 자연과학, 사회과학 분야에서 기초연구, 응용연구뿐만 아니라 새로운 기술·제품·서비스 개발 및 실증을 위한 산업적 연구도 해당된다. ‘공익적 기록보존’을 위한 가명정보 처리는 공공의 이익을 위하여 지속적으로 열람할 가치가 있는 정보를 기록하여 보존하는 것이다(개인정보보호위원회, 2022).

〈표 2-4〉 가명정보 처리 가능 목적별 예시

목적	예시
통계 작성	A공사가 도로 구조 개선 및 휴게공간 추가 설치 등의 고객센터 개선을 위하여 월별 시간대별 차량 평균속도, 상습 정체구간, 사고 구간 및 원인 등에 대한 통계를 작성하는 경우
과학적 연구	코로나19 위험 경고를 위해 생활패턴과 코로나19 감염률의 상관성에 대한 가설을 세우고, 건강관리용 모바일앱을 통해 수집한 생활습관, 위치정보, 감염증상, 성별, 나이, 감염원 등을 가명처리하고 감염자의 데이터와 비교·분석하여 가설을 검증하려는 경우
공익적 기록보존	연구소가 현대사 연구과정에서 수집한 정보 중 사료 가치가 있는 생존 인물에 관한 정보를 가명처리하여 기록·보존하고자 하는 경우

자료: 개인정보보호위원회(2022), 가명정보처리 가이드라인

나. 데이터 협의

연구자는 가명정보 처리에 합당한 연구 목적을 설정하고, 데이터를 보유한 기관과 협의한다. 데이터를 협의할 때는 정보의 종류, 범위를 명확히 하여 대상을 선정해야 한다. 이 과정에서 연구자가 주의할 점은 다음과 같다. 첫째, 사용하고자 하는 데이터가 대부분 행정데이터이기 때문에, 통계 분석을 목적으로 가공된 데이터와 같이 사용자 편의적이지 않다. 코드북이 존재하지 않거나, 대분류 코드북만 공개되고, 세부 내용은 직접 방문하여야만 열람할 수 있는 경우가 많다. 세부 변수를 파악했다라고, 개별 변수의 응답값을 미리 확인할 수 없기 때문에, 결합 이후 주요 변수에 결측값이 많거나, 사례 수가 적은 경우도 발생한다. 즉 연구자는 가명정보 결합을 원하는 모든 기관의 변수들을 정확히 파악하기가 어렵기 때문에 추후 결합된 데이터의 분석 가능성을 예측하기가 어렵다.

둘째, 가명정보로 제공될 수 있는 데이터의 시점이 기관마다 다르기 때문에 대상자 선정에 어려움이 있다. 예를 들어, 국세청의 데이터는 2015년 이후의 데이터만 제공받을 수 있고, 건강보험공단의 데이터는 2011년 이후, 건강보험 심사평가원의 데이터는 2007년 이후 것만 제공받을 수 있다. 또한 사회보장정보원의 경우에는 보유한 데이터의 성격에 따라 각각 제공 연도가 다르다. 따라서 데이터를 협의하는 과정에서 초기의 연구 계획을 수정하는 경우가 발생한다.

셋째, 데이터의 성격상 데이터를 보유하고 있는 기관의 승인뿐만 아니라, 해당 정부 관계부처의 승인도 각각 받아야 한다. 예를 들면, 사회보장정보원이 보유한 기초연금 및 장애인 연금 관련 정보는 보건복지부에서, 국민기초생활보장 주거급여 정보는 국토교통부에서, 국민기초생활보장 교육급여 정보는 교육부에서 각각 승인을 받아야 하기 때문에, 협의 과정

이 길어질 수밖에 없다.

넷째, 연구자가 다수의 기관 데이터를 결합하고자 한다면, 각각의 기관으로부터 데이터 제공 승인을 받아야 한다. 만약 일부 기관에서 제공을 거부할 경우, 기존에 설정했던 연구 목적을 달성하기 어렵기 때문에, 연구 목적 설정부터 다시 시작해야 한다. 따라서 데이터 보유기관과의 사전 협의 과정이 매우 중요하다. 행정 절차인 신청서 제출에 앞서, 이미 사전 협의 과정에서 데이터 제공 여부와 데이터 제공 범위까지도 결정되어야 한다.

다섯째, 다양한 정보를 확보하기 위해 다수 기관의 정보를 결합하고자 시도할 수 있다. 이는 정보의 다양성 측면에서 보면 연구의 분야를 확장할 수 있어서 바람직하지만, 여러 기관의 데이터를 결합할수록 결합률이 낮아질 수 있음에 유의해야 한다. 따라서 결합하고자 하는 모든 기관과 협의하는 과정에서, 연구에 필요한 정보를 보유한 기관과의 결합만을 권고받기도 한다.

이상의 유의점들에 유념하여 데이터 보유기관과의 가명정보 결합을 협의하면서 연구계획서를 수정, 보완해야 한다. 이 과정을 충실히 할수록 이후 결합 과정에 발생할 수 있는 문제들을 예측하고, 결합된 데이터의 분석 범위를 명확히 할 수 있다.

다. 가명정보 이용·제공 신청

연구자는 데이터를 보유한 기관에 데이터를 요청하기 위해 가명정보 제공 신청서를 작성하여 제출한다. 가명정보 제공 신청서는 표준신청서가 있는 것이 아니라, 기관의 특성에 따라 자체적으로 제작되었다.

[그림 2-4]를 보면 기재 사항은 신청인의 인적 사항, 가명처리 대상 데

42 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

이터 명칭 및 데이터 내용, 이용 기간, 처리 목적(통계 작성, 과학적 연구, 공익적 기록보존), 활용 형태(내부 이용(동일부서, 타 부서), 외부 이용(제3자 제공, 결합전문기관 제공)), 처리 장소(폐쇄 환경 여부) 등이다.

[그림 2-4] 가명정보 이용·제공 신청서(예시)

가명정보 이용·제공 신청서					
접수번호				접수일	
신청자	조직/부서명				
	담당자 직위		담당자 성명		
	전화번호		이메일 주소		
처리 목적		☐ 통계작성 ☐ 과학적 연구 ☐ 공익적 기록보존			
		세부내용 :			
활용 형태	☐ 내부이용 ☐ 제3자 제공 ☐ 결합전문기관을 통한 결합				
이용 주체	☐ 동일 개인정보처리자 ☐ 제3자(제공받는 자) 제공				
처리 환경	☐ 내부 ☐ 외부				
처리 장소	☐ 폐쇄 환경 ☐ 제한 없음				
반복 제공 여부	☐ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)				
제공 방법	☐ 온라인 ☐ 오프라인				
제공 받는 자	(내부) ☐ 동일부서 ☐ 타부서 (외부) ☐ 제3자, 결합전문기관 제공				
신청 명세	가명처리 대상 명칭				
	데이터 내역				
	가명정보 이용 기간	년 월 일 ~ 년 월 일(개월)			
위와 같이 가명정보 이용·제공을 신청합니다.					
년 월 일					
신청인(부서장명) (서명 또는 인)					
첨부서류	1. 개인정보 유형 분류표 2. 활용데이터 요구 수준표(선택) 3. 데이터 처리 위탁 계약서(위탁 계약의 경우) / 추후 제출 4. 데이터 제공 계약서(제3자 제공의 경우) / 추후 제출				

자료: 개인정보보호위원회. (2022). 가명정보처리 가이드라인. p.103.

함께 첨부해야 하는 서류는 활용 목적 증빙 서류(연구계획서), 관리적·기술적·물리적 보호조치를 포함한 활용자의 안전조치 계획 증빙 서류, 개인정보 유형 분류표, 활용데이터 요구 수준표, 데이터 처리 위탁 계약서, 데이터 제공 계약서 등이다. 보건의료 관련 데이터는 IRB(Institutional Review Board, 기관생명윤리위원회)를 추가로 요청한다.

가명정보 이용·제공 신청서 제출 시 주의할 점은 다음과 같다. 첫째, 연구자가 여러 기관의 정보를 결합하고자 한다면 각 기관의 신청서 양식에 따라 작성하고 제출해야 한다. 둘째, 연구자가 본인이 속한 기관의 정보를 이용하고자 하는 경우는 내부 이용에 해당하므로, 가명정보를 처리하는 담당부서에 신청서를 제출해야 한다. 셋째, 신청서가 승인된 이후에는 수정이 불가능하기 때문에 연구 목적에 해당하지 않는 변수 신청은 불가하며, 분석과정에서 연구자를 추가하는 것도 불가능하다.

라. 가명정보 이용·제공 신청서 적합성 검토

데이터 보유기관은 데이터 이용을 희망하는 연구자가 제출한 신청서를 바탕으로 연구 목적의 성격, 가명정보 결합을 통한 연구 목적의 달성 여부, 개인정보의 안정성 확보 여부 등에 대하여 법률에 근거하여 데이터 제공을 결정한다. 만약 연구자가 3개 기관의 데이터를 결합하고자 한다면, 3개 기관 각각에 신청서를 제출하고, 3개 기관은 각각 내부 규정에 맞춰 적합성 검토를 진행한다. 적합성 검토 결과 ‘제공 승인’이 결정되면 연구자에게 통보하고, 결합키 신청 단계로 넘어간다.

데이터 보유기관이 가명정보를 제공하기 위해서는 개인정보보호 규정과 가명정보 처리 내부관리계획이 기관에 수립되어 있어야 한다. 기관은 내부 규정에 맞춰 적합성 검토를 실시한다.

1) 내부 관리 계획

개인정보보호법 제28조의 4와 개인정보보호법 시행령 제29조의 5, 제30조의 1항 1호에 따라 개인정보처리자는 가명정보를 관리하기 위해 내부관리계획을 수립, 시행하여야 한다. 기관에 개인정보 내부관리계획이 수립되어 있다면, 가명정보 내부관리계획을 따로 수립하지 않고, 개인정보 내부관리계획 내에 가명정보 처리에 관한 내용만 추가하여 수립하는 것도 가능하다.

개인정보보호위원회(2022. 4)의 「가명정보 처리 가이드라인」에 따르면 가명정보 처리 내부관리계획에 포함될 사항은 다음과 같다.

- 가명정보 및 추가정보의 분리 보관에 관한 사항
- 가명정보 및 추가정보에 대한 접근 권한 분리에 관한 사항
- 가명정보 또는 추가정보의 안전성 확보 조치에 관한 사항
- 가명정보를 처리하는 자의 교육에 관한 사항
- 가명정보 처리 기록 작성 및 보관에 관한 사항
- 개인정보 처리 방침 공개에 관한 사항
- 가명정보의 재식별 금지에 관한 사항
- 가명정보 및 추가정보에 대한 파기에 관한 사항

2) 적합성 검토

적합성 검토위원회는 개인정보보호법 제28조의 2 제1항에 대해 가명처리 전 적합성 검토를 수행하는 위원회이다. 적합성 검토위원회의 구성은 개별 기관마다 필수는 아니고, 기관의 성격에 맞게 다른 위원회로 대

체하는 것이 가능하다. 공공분야는 기존의 별도 공공데이터 제공 심의위원회에서 본 기능을 수행할 수 있고, 보건의료 분야는 데이터 심의위원회, 금융 분야는 데이터 전문기관을 통해 위탁 평가할 수도 있다.

적합성 검토위원회는 개인정보보호 법·제도 및 가명처리 기법 등에 관한 전문가¹²⁾로 3인 이상 7인 이내의 위원으로 구성하도록 되어 있으나, 3인 이상 구성이 필수 요건은 아니다.¹³⁾ 내부 위원 및 외부 위원의 자격은 다음과 같다.

〈표 2-5〉 위원회 구성 시 참고사항

구분	기관
내부위원	• 해당 기관의 개인정보보호 및 정보 보안 등과 관련된 부서장 • 해당 기관의 데이터 제공 및 분석 업무 관련 부서장이 추천하는 사람 중 위원장이 지명하는 사람 • 가명처리, 익명처리 기술 이해도가 높은 사람 • 해당 기관에서 법무, 리스크관리 업무 등을 수행하는 법률 관련 내부 전문가(기관에 해당 인력이 없으면 기관 임원 중 1인으로 대체 가능)
외부위원	• 데이터 활용 전문가 • 해당 기관에 법률 전문가가 없는 경우 외부 법률전문가 • 개인정보보호 관련 전문가 • 개인정보 가명처리, 익명처리 등 비식별 처리 전문가 • 기타 위원장이 필요하다고 인정하는 전문가

자료: 행정안전부·개인정보보호위원회. (2021). 공공분야 가명정보 제공 실무 안내서.

적합성 검토는 가명정보의 제공 여부를 검토하기 위해 기관 내부 여건 및 활용 위험성, 법·제도 부합 여부, 기술적 가능 여부, 데이터 이용 기관의 가명정보 활용을 위한 법적 안전성 조치 등을 검토한다.

12) 가명정보 전문가는 한국인터넷진흥원에서 2021년부터 ‘가명정보 전문가 풀’을 구성하여 운영 중에 있다. 2023년 8월 2일 출력 <https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS061&mCode=C010010000&nttId=8839#LINK>

13) 김기태(2022.4.)에 따르면, 가명처리 전 과정을 검토하는 적합성 검토위원회는 최소 3인 이상의 위원 구성이 필수 요건은 아니지만, 가명처리 후 과정을 검토하는 적정성 검토는 최소 3인 이상으로 구성하고, 반드시 외부 위원을 최소 1인 이상 포함해야 함을 유념해야 함.

내부 여건 및 활용 위험성 검토는 첫째, 가명처리·제공에 소요되는 인적·물적 비용 충당 여부, 공공기관 본래 업무수행에 상당한 지장을 초래할 우려가 있는지 등의 내부 여건을 검토한다. 둘째, 제공된 가명정보가 목적에 맞게 활용되고 창출되어, 사회·경제적 가치 및 공공성의 제공 가치가 있는지 검토한다. 셋째, 제공받는 가명정보를 악용하거나 제3자의 권리를 침해할 가능성, 해당 가명정보 활용이 공공기관이 추진 중이거나 계획 중인 정책에 배치될 가능성 여부를 검토한다.

법·제도 부합여부 검토는 활용 목적이 통계 작성, 과학적 연구, 공익적 기록보존에 해당하는지를 신청자가 기재한 사항을 기초로 검토하되 기관이 주체적으로 판단한다.

기술적 가능 여부 검토는 데이터의 가명처리 가능 여부와 비공개 정보 등의 기술적 분리 여부, 목적 달성에 필요한 가명·익명 처리 수준을 검토하는 것이다.

법적 안전성 조치 검토는 가명정보를 활용하고자 하는 연구자가 속한 기관이 「개인정보의 안전성 확보 조치 기준」에 따른 관리적·물리적 보호 조치를 시행하는지 판단하는 것이다. ‘가명정보 안전조치 의무이행 협약서’가 있다면 안전성 확보 조치 검토 항목을 ‘만족’하는 것으로 간주한다 ([부도 1] 참조).

적합성 검토를 위해 필요한 서류는 데이터 활용 신청서, 상세 데이터 목록, 데이터 처리 위탁 계약서(위탁 계약의 경우), 데이터 제공 계약서(제3자 제공의 경우), 가명정보 안전조치 의무이행 협약서 등이다.

적합성 검토위원회에서는 세부 항목들에 대해 종합적으로 검토한 후, 가명정보 제공에 대하여 ‘승인’ 또는 ‘거부¹⁴⁾’를 결정한다.

14) 행정안전부·개인정보보호위원회(2021.1)에 따르면, 가명정보 제공 결정은 의무사항이 아니라 기관 재량사항임. 따라서 제공 거부에 대한 불복절차를 기관에 마련하는 것도 기관이 자율적으로 결정할 수 있음.

2. 결합신청

결합신청 단계에서는 결합신청, 결합키 생성, 결합률 확인, 결합 유형 결정 등으로 구성된다.

[그림 2-5] 결합신청 단계



자료: 필자 작성

가. 결합신청

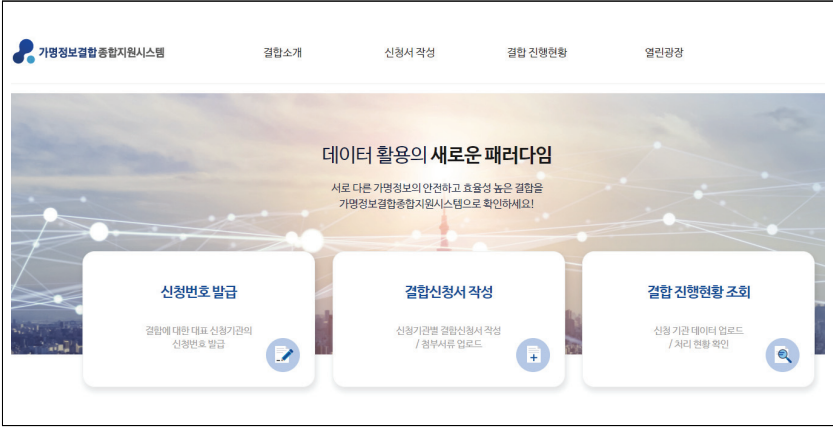
연구자는 가명정보결합종합지원시스템에 결합을 신청한다([그림 1-7] 참조). 결합신청은 데이터 이용기관 또는 데이터 보유기관 중 대표 기관이 신청할 수 있다. 연구자가 서두른다고 결합신청을 빨리 할 수 있는 구조가 아니기 때문에, 데이터 보유기관의 협조를 거듭 독려하는 수밖에 없다. 연구자는 데이터 보유기관과 수시로 연락하면서, 신청 진행 현황을 확인해야 한다.

1) 신청번호 발급

[그림 2-6]은 가명정보결합종합지원시스템 홈페이지의 첫 화면이다. 신청번호 발급을 클릭하면, 본인 인증 후 신청서 작성 화면이 나온다. 결합전문기관을 선택하고, 결합신청기관 정보(기관명, 담당자 연락처)를 입력한다. 여기서 결합신청기관은 데이터 보유기관 또는 데이터 이용기관을 의미하고, 결합전문기관에 기관이 보유한 데이터를 제공하면서 데이터들간의 결합을 신청하는 역할을 수행한다. 앞의 [그림 2-2]의 A, B, C 기관이다.

48 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

[그림 2-6] 가명정보결합종합지원시스템



자료: 개인정보보호위원회·한국인터넷진흥원, 가명정보결합 종합지원시스템.
<https://link.privacy.go.kr/>에서 2023.6.22. 인출

결합전문기관은 가명정보 결합을 수행하기 위해 개인정보보호위원회 또는 중앙행정기관의 장이 지정하는 전문기관¹⁵⁾이다. 2023년 6월 현재 개인정보보호위원회 지정 10기관, 보건복지부 지정 3기관, 국토교통부 지정 1기관, 과학기술정보통신부 지정 5기관, 교육부 1기관, 산업통상자원부 1기관, 행정안전부 1기관 등 총 22개 기관이 있다(〈표 2-6〉 참조).

15) 개인정보보호법 제28조의 3 제1항에 따라 지정되었다.

“① 제28조의 2에도 불구하고 통계 작성, 과학적 연구, 공익적 기록보존 등을 위한 서로 다른 개인정보처리자 간의 가명정보의 결합은 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관이 수행한다.”

〈표 2-6〉 결합전문기관 리스트

구분	기관
개인정보보호위원회	(주)케이씨에이, 삼성SDS, 통계청, 한국지역정보개발원, 롯데정보통신, 한국정보인증, 국세청, 한국사회보장정보원, LG CNC, CJ 주식회사
보건복지부	국민건강보험, 건강보험심사평가원, 국립암센터
국토교통부	한국도로공사
과학기술정보통신부	한국지능정보사회진흥원, (주)더존비즈온, SK주식회사 C&C, 비씨카드, 데이터산업진흥원
교육부	한국교육학술정보원
산업통상자원부	한전KDN
행정안전부	국가정보자원관리원

자료: 개인정보보호위원회·한국인터넷진흥원. 가명정보결합종합지원시스템, 2023.6.22. 인출.
<https://link.privacy.go.kr/>

결합전문기관은 연구자가 임의로 선택할 수 있지만, 보통 데이터를 보유한 기관 중에서 선택한다. 예를 들면 건강보험공단 데이터와 통신사 데이터를 결합하는 경우, 결합전문기관으로 통계청을 지정할 수도 있지만 건강보험공단을 지정하는 것이 업무 진행상 원활할 것이다. 건강보험공단 데이터와 건강보험심사평가원 데이터, 국립암센터 데이터를 결합하는 경우는 세 기관 모두 결합전문기관이므로, 세 곳 중 임의로 지정할 수 있다. 다만, 연구자가 결합 데이터를 빨리 사용하고 싶다면 세 기관에 각각 예상 반출 시기를 사전에 문의하여 가장 빠른 곳을 선택할 수 있다.

2) 결합신청서 작성

발급받은 결합신청번호를 이용하여 결합신청서를 작성한다. 결합신청기관의 사업자등록번호, 대표자명, 주소, 담당자 연락처 및 가명정보 제공자의 연락처 등을 기재한다. 결합 개요는 반복결합 여부, 결합키 제출 방법(온라인, 오프라인)을 선택한다. 가명정보 제공자와 결합정보 이용자

는 동일한 사람이면 안 된다. 가명정보 제공자는 데이터 보유부서의 책임자이고, 결합정보 이용자는 정보를 분석하고자 하는 연구책임자이다.

첨부 서류는 ① 서면 결합신청서, ② 사업자등록증, ③ 결합 목적을 증명할 수 있는 서류, ④ 결합 대상 가명정보에 관한 서류(전체 항목명, 가명처리 대상 항목명, 가명처리 내역 등), ⑤ 기타(추가 서류)이다. 이 중 서면 결합신청서와 결합 목적을 증명할 수 있는 서류는 가명정보결합종합지원시스템에서 다운로드할 수 있다([그림 2-7] 및 [그림 2-8] 참조).

제출 서류 중 가명 대상 가명정보에 관한 서류(전체 항목명, 가명처리 대상 항목명, 가명처리 내역 등)는 결합전문기관으로부터 수령할 수 있다. 결합신청 단계에서는 전체 항목명과 가명처리 대상 항목명을 제출하고, 가명처리 내역은 추후 결합 대상 정보가 확정된 이후에 제출한다([그림 2-9] 참조).

3) 신청 완료

관련 서류를 모두 제출하면 지정된 결합 전문기관과 한국인터넷진흥원은 신청서를 검토하고, 필요에 따라 서류의 수정 보완을 요청한다. 이후 승인되면 [그림 2-6]의 가명정보결합종합지원시스템에서 결합 진행현황 조회가 가능하다.

[그림 2-7] 가명정보 결합신청서

■ 「가명정보의 결합 및 반출 등에 관한 고시」 [별지 제3호서식]

결합 신청서

신청번호		
접수번호		

결합신청자		
기관명	사업자등록번호 또는 법인등록번호	
주소	대표자명	
담당자	담당자 연락처 (전화, e-mail)	
유형	☐ 개인 ☐ 공공기관 ☐ 비영리법인 ☐ 민간기관	

결합 개요		
전체 결합신청기관명	총 기관 수	☐ 개 기관
반복결합	☐ 해당없음 ☐ 최초 ☐ 추가(결합접수번호 :)	
추가결자 신청	결합을 확인 ☐ 가명정보 추출 ☐ 모의결합 ☐	

가명정보 제공자		해당없음 ☐
제출 방법	☐ 온라인 ☐ 오프라인	
지원 요청 사항	☐ 결합 신청에 필요한 가명처리	
가명정보 제공 담당자	이름	연락처 (전화, e-mail)

결합정보 이용자		해당없음 ☐
결합 목적	☐ 통계작성 ☐ 과학적 연구 ☐ 공익적 기록보존 등	
세부 결합 목적		
분석공간 이용	☐ 추가 가명처리 ☐ 결합정보 분석 ☐ 이용안함	
지원 요청 사항	☐ 반출 전 처리 ☐ 분석	
결합정보 이용 담당자	이름	연락처 (전화, e-mail)

「개인정보 보호법」 제28조의3제1항 및 같은 법 시행령 제29조의3제1항에 따른 결합을 위하여 결합전문기관에 결합신청서를 위와 같이 제출합니다.

년 월 일

결합신청기관 (서명 또는 인)

(결합전문기관)의 장 귀하

첨부 서류	1. 사업자등록증, 법인등기부등본 등 결합신청자 관련 서류 1부 2. 결합 목적을 증명할 수 있는 서류 1부(결합정보 이용자에 한함) 3. 결합 대상 가명정보에 관한 서류(전체 항목명, 가명처리 대상 항목명, 가명처리 내역 등) 1부(가명정보 제공자에 한함) * 결합키 생성에 사용된 항목 제외 ** 결합 대상 정보가 확정된 이후에 제출
----------	--

자료: 개인정보보호위원회·한국인터넷진흥원. 가명정보결합종합지원시스템, 2023.6.22. 인출.
<https://link.privacy.go.kr/>

[그림 2-8] 결합 목적을 증명할 수 있는 서류

[서식2] 가명처리 및 결합 목적 증명 자료(예시)
※ 아래 서식은 가명처리 및 결합 목적 증빙에 대한 이해를 돕기 위해 작성한 자료로써 표준이 아니며
자체적으로 제작하여 다른 자료로 활용할 수 있음

통계작성 계획서		
통계명		
대표 참여진	소속	
	담당자명	
통계작성 배경 및 목적		
통계작성 대상자 수		
통계작성 계획 및 방법		
기대효과 및 활용방안		
붙임. 상세 통계작성 계획서 등		

과학적 연구 계획서		
연구명		
연구진	소속	
	연구책임자	
연구 배경 및 목적		
예상 연구 기간		
연구 대상자 수		
연구 방법		
연구내용		
기대효과 및 활용방안		
붙임. 상세 연구계획서 등		

공익적 기록보존 계획서		
공익적 기록보존명		
대표 참여진 (기록보관 기관)	보관기관명	
	담당자명	
공익적 기록보존 목적		
보존기간		
공익적 기록보존 방법		
내용		
기대효과 및 활용방안		
붙임. 상세 계획서 등		

자료: 개인정보보호위원회·한국인터넷진흥원. 가명정보결합중합지원시스템, 2023.6.22. 인출.
<https://link.privacy.go.kr/>

[그림 2-9] 가명 대상 가명정보에 관한 서류

[작성 예시]

1. 데이터 기본정보

가. 데이터 크기 및 생성방법

데이터 명	데이터 설명	크기 및 형식	생성방법	가명처리 목적
대상자	대상자의 개인일반내역	- (데이터 용량) 00GB - (레코드 수) 000,000건 - (컬럼 수) 0개 - (파일명) 000 - (파일 형식) CSV - (인코딩 방식) UTF-8	0000	0000

나. 데이터 세부 항목별 명세서

연번	항목	개인정보 유형	데이터 타입	최대 길이	내용
1	대상자 생년월일	QI	숫자형	15	대상자 생년월일 8자리
2	대상자 성별구분	QI	문자형	2	성별 구분코드 (코드값: 10(남)/20(여))
3	시도	QI	문자형	8	관리할정지역코드 기준 시도명
4	시군구	QI	문자형	12	관리할정지역코드 기준 시군구명
5	내외국인	QI	문자형	3	내외국인 구분명 (코드값: 내국인/외국인)
6					
7					
8					
9					
10					

※ ID: 개인정보, Q: 개인정보(준식별자), SA: 개인정보(민감정보), NSA: 개인정보(비민감정보)

자료: 한국사회보장정보원. (2023). 결합신청서 및 첨부서류 양식. 2023. 6. 22. 제공

나. 결합키 생성

결합신청이 완료되면 한국인터넷진흥원과 결합신청기관 담당자들이 협의하여 결합키를 생성한다. 결합키란 서로 다른 데이터의 연계를 위해 사용되는 정보로, 특정 대상을 알아보기나 구별할 수 없도록 가명처리한 정보이다. 일반적으로 이름, 성별, 연령, 생년월일, 휴대전화번호, 거주지 주소 등을 조합하여 특정한 값을 갖는다. 정확한 결합을 위해서는 다양한

변수들을 조합하는 것이 마땅하다. 그러나 모든 정보가 결측 없이 존재하는 경우에만 결합되기 때문에, 다양한 정보를 사용할수록 결합률은 낮아진다. 예를 들어, 동일한 개인이 A기관에는 2010년도 당시 소유했던 휴대전화 번호가, B기관에는 2015년도 당시 휴대전화 번호가 다르게 기록되어 있을 수 있다. 결합키를 이름, 성별, 생년월일만으로 조합할 때는 동일한 개인으로 인식되어 A 기관의 정보와 B 기관의 정보가 결합된다. 하지만, 휴대전화번호를 포함하여 결합키를 생성한다면 다른 개인으로 인식되어 두 기관의 정보는 결합되지 않게 된다.

이와 다른 예를 들면, 건강보험이나 국민연금과 같이 거의 전국민 정보를 보유한 기관의 데이터를 결합하고자 한다. 이름, 성별, 생년월일만을 키 변수로 사용했을 때, 3개의 정보가 정확하게 일치하는 동일인이 많아서 높은 중복결합의 확률 때문에 결합된 데이터의 신뢰성 문제가 발생할 수 있다. 따라서 결합키를 생성할 때는 개별 기관들이 보유한 정보의 성격을 세부적으로 파악하고 협의를 통해 결정해야 한다.

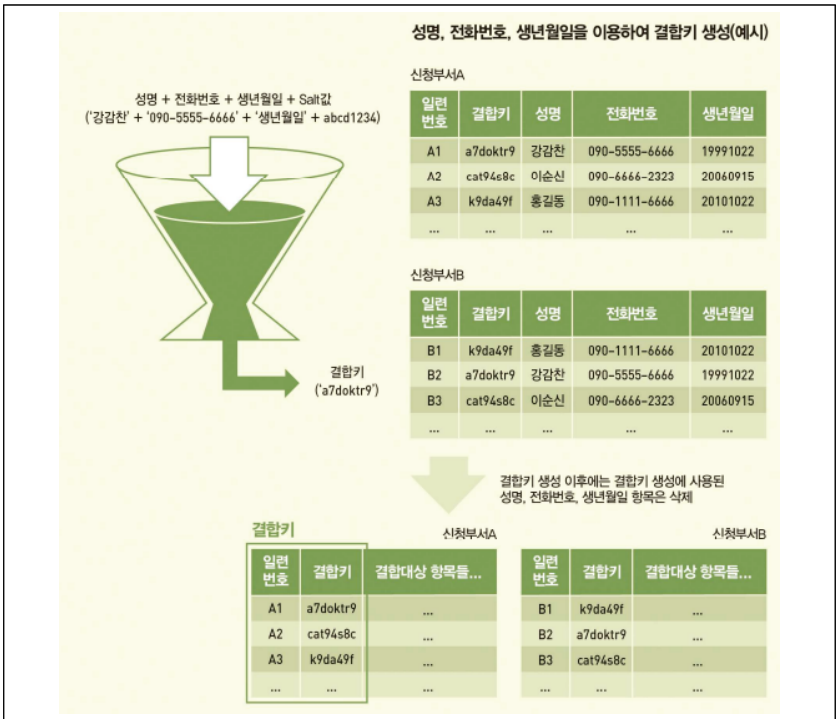
개인정보보호위원회는 결합키를 생성할 때 특정 개인을 식별할 수 없도록 일방향 암호화 기법 중 SHA-2 이상의 알고리즘(salt 값 포함)을 이용할 것을 권고한다. 개인정보보호위원회의 경우 salt 값을 결합키 관리 기관인 한국인터넷진흥원에서 제공한다. 해시 값 생성 시 사용되는 salt 값 또는 키 값은 쉽게 유추할 수 없도록 복잡하게 구성하여야 한다(금융위원회, 금융감독원, 2022). 또한 식별자에 단순히 salt 값 또는 키 값만 추가하여 해시를 생성하기보다는 자체적으로 해시 값 생성 규칙을 마련하여 해시 값의 안전성을 높일 것을 권고한다(금융위원회, 금융감독원, 2022).

〈표 2-7〉 결합키 생성을 위한 대표적인 권고 해시 알고리즘

종류		출력값 길이(비트)	보안강도(비트)	참조 규격
SHA-2	224	224	112	NIST FIPS 180-4
	256	256	128	
	386	384	192	
	512	512	256	
SHA-3	224	224	112	NIST FIPS 202
	256	256	128	
	386	384	192	
	512	512	256	

자료: 금융위원회·금융감독원. (2022). 금융분야 가명·익명처리 안내서.

〔그림 2-10〕 결합키 생성 예시



출처: 교육부, 개인정보보호위원회. (2020). 교육분야 가명·익명정보 처리 가이드라인.

다. 결합률 확인

정보를 보유한 결합신청기관은 기관들 간 협의로 생성된 결합키와 각 기관의 가명정보 결합 대상자의 일련번호를 결합키 관리기관인 한국인터넷진흥원에 전송한다.

결합키 관리기관은 결합키를 이용한 대상자 결합을 진행한 후, 결합률을 해당 결합신청기관에게 통지한다. 이 과정에서는 개별 기관이 보유한 모든 정보를 결합하는 것이 아니라 결합키와 일련번호만 결합하여, 어느 정도 결합할 것인지를 선제적으로 파악한다.

〈표 2-8〉에서 A 기관 데이터를 기준으로 결합률을 측정하면 80%이고, B 기관 데이터를 기준으로 결합률을 측정하면 40%이다. 연구자는 결합률을 확인하고, 가명정보 결합을 진행할 것인지 종료할 것인지 결정한다.

〈표 2-8〉 결합률 예시

(단위: 건, %)

	전체 정보 (A)	결합한 정보 (B)	미결합 정보 (A-B)	결합률 (B/A)
A 기관	1,000,000	800,000	200,000	80%
B 기관	2,000,000	800,000	1,200,000	40%

주: 연구자가 임의로 만든 수치임.

라. 결합 유형 결정

연구자가 가명정보 결합을 결정하였다면, 연구 목적에 따라 다양한 결합 유형을 선택할 수 있다([그림 2-11] 참조).

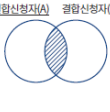
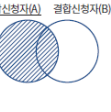
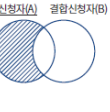
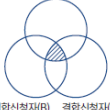
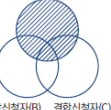
첫째, 세 개 기관의 데이터 중 세 개 기관에 모두 결합된 데이터가 필요하다면, 공통결합 중 단일결합을 선택할 수 있다.

둘째, 세 개 기관이 아니라 한 개의 기관과 결합된 모든 데이터를 분석

하고자 한다면, A 기관의 데이터를 기준으로 공통결합 중 다중결합을 선택해야 한다.

셋째, A 기관의 데이터 중에서 B 기관의 데이터와 결합된 데이터의 성격을 알고 싶다면 결합되지 않은 데이터도 파악되어야 한다. 이때는 확대결합을 선택해야 한다.

[그림 2-11] 결합 유형

공통결합 (INNER JOIN)			확대결합 (OUTER JOIN)	잔여결합 (ANTI-INNER JOIN)
단일 (INNER SINGLE)	다중 (INNER MULTI)	완전 (INNER FULL)	단일 (OUTER SINGLE)	단일 (ANTI-INNER SINGLE)
 결합신청자(A) 결합신청자(B)	 결합신청자(A) 결합신청자(B) 결합신청자(C)	 결합신청자(A) 결합신청자(B) 결합신청자(C)	 결합신청자(A) 결합신청자(B)	 결합신청자(A) 결합신청자(B)
 결합신청자(A) 결합신청자(B) 결합신청자(C)	 결합신청자(A) 결합신청자(B) 결합신청자(C)		 결합신청자(A) 결합신청자(B) 결합신청자(C)	 결합신청자(A) 결합신청자(B) 결합신청자(C)

자료: 개인정보보호위원회. (2022). 가명정보처리 가이드라인.

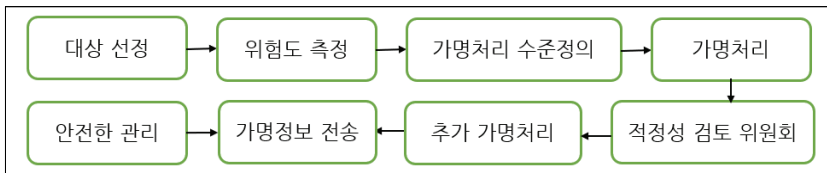
3. 가명처리 단계

가명처리란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 것이다(개인정보보호위원회, 2022). 가명처리의 일차적인 목적은 개인 식별 위험성을 방지하는 것이다. 그러나 가명처리의 궁극적인 목적은 가명처리를 통해 개인이 식별되지 않는 데이터를 이용하여 통계 작성, 과학적 연구, 공익적 기록보존 등의 업무를 수행할 수 있게 하는 것이다. 따라

서 개인이 식별되지 않으면서 가명정보를 사용하고자 하는 목적을 달성할 수 있도록 가명처리를 실시하여야 한다.

가명처리 단계는 데이터를 보유한 개별 기관이 제공할 정보에 대하여 수행하는 과정이다. 3개 기관의 데이터를 이용한다면, 3개 기관 모두 개별적으로 수행하여야 한다. 개별 기관에서 가명처리하는 과정은 가명처리하고자 하는 변수와 대상자 선정, 위험도 측정, 가명처리 수준 정의, 가명처리 수행, 적정성 검토위원회 운영, 추가 가명처리, 처리된 가명정보를 결합전문기관에 전송, 사후 안전하게 관리하는 단계로 구성된다.

[그림 2-12] 가명처리 단계별 절차도



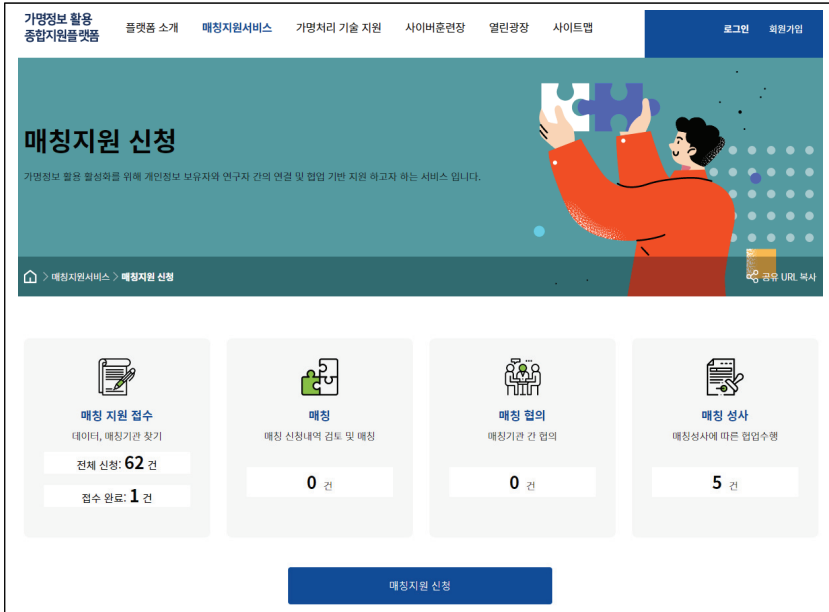
자료: 필자 작성

가. 대상 선정

연구계획서에 적시한 목적을 달성하기 위해 데이터 및 대상의 특성을 확인하고 최소 항목으로 추출한다.

연구자는 어느 기관이 어떠한 정보를 보유하고 있는지 알기 쉽지 않다. 이 경우에는 가명정보 활용 정보지원플랫폼의 매칭지원서비스를 이용할 수 있다([그림 2-13] 참조). 이곳에서 데이터 검색을 해보면, 원하는 데이터를 어느 기관이 보유하고 있는지 알 수 있다. 연구 주제가 있는 경우 홈페이지에 신청하면 심사 과정을 통해 필요한 데이터가 어느 기관에서 보유하고 있는지 알려주고 기관과 연결해 주기도 한다.

[그림 2-13] 가명정보 활용 종합지원플랫폼



자료: 개인정보보호위원회·한국인터넷진흥원, 가명정보 활용 종합지원플랫폼, 2023.8.4. 출력
<https://data.privacy.go.kr/mtcsprtInfo/mtcsprtSvc.html>

연구자가 가명정보 결합을 시도하고자 하는 데이터는 대부분 공공에서 열람되는 데이터가 아니고, 보완이 필요한 행정데이터이다. 따라서 데이터 보유기관은 대분류 수준에서 정보의 성격을 공개하고, 세분류는 보안상 기관에 직접 방문하여 열람하도록 하는 경우가 흔히 발생한다. 변수의 세분류를 파악하여 세분류 단위까지 데이터 제공을 신청했는데, 가명처리 후에도 개인 식별 위험성이 커진다면, 대분류 수준으로 다시 가명처리될 것이다. 이러한 경우 연구 목적을 달성하기 어려울 수도 있다. 따라서 변수를 선정할 때 충분한 시간을 가지고 데이터 보유기관과 협의하여 다양한 가능성에 대하여 고민해야 한다.

나. 위험도 측정

〈표 2-9〉는 교육부에서 제공하는 가명정보 처리 환경 위험도 분류 측정 예시이다. 점수가 높을수록 위험도가 높은 것이다. 내부보다는 외부에서 사용할 경우, 데이터를 소유하고 있는 부서보다는 데이터를 소유하지 않은 타 부서가 사용할 경우, 활용할 때보다는 제공할 때, 제공할 때보다는 결합할 경우, 교육기관보다는 민간에서 데이터를 사용할 경우 위험도가 높다. 기관의 개인정보 보호 수준과 재식별 수준에 따라서도 상, 중, 하로 구분할 수 있다. 개별 항목을 모두 합산하여 점수가 높을수록 위험도가 높은 처리 환경이므로, 이를 고려하여 가명처리를 더욱 엄격하게 할 것을 권고한다.

〈표 2-9〉 처리 환경 위험도 분류 측정 예시

처리 환경				기관의 개인정보 보호수준	재식별 수준 (다른 정보 보유 등)
활용장소	활용부서	처리유형	기관유형		
내부(1)	소유부서(1)	활용(1)	교육기관(1)	상(1)	상(3)
외부(2)	타부서(2)	제공(2)	민간기관(2)	중(2)	중(2)
		결합(3)		하(3)	하(1)

자료: 교육부·개인정보보호위원회. (2020). 교육분야 가명·익명정보 처리 가이드라인.

이 과정에서 연구자는 가급적 많은 정보를 사용하고자 하고, 데이터 보유기관은 개인 식별 위험성을 낮추기 위해 가급적 많은 정보를 가명처리 하고자 한다. 따라서 관계 기관과의 긴밀한 협의를 통해 가명처리의 방법 및 수준을 선택해야 한다. 또한 예시에서 보여준 데이터를 처리하는 환경 뿐만 아니라, 데이터를 제출하는 환경에 따라서도 가명처리 수준이 달라진다. 예를 들면 결합전문기관의 안심구역¹⁶⁾에서 데이터를 분석하는 경

16) 안심구역이란 가명정보를 처리할 수 있는 물리적 폐쇄 환경으로, 가명처리 활용 전 과정의 추적 관리가 가능하고 분석 결과물만 반출할 수 있는 공간임.

우는 거주지의 ‘읍면동’이 제공 가능하고, 데이터를 온라인으로 접속하여 사용하거나 다른 공간에서 사용하는 경우는 ‘시군구’로 제공 가능하다.

다. 가명처리 수준 정의

[그림 2-14]는 항목별 가명처리 계획을 내부 이용과 제3자 제공으로 구분하여 작성한 예시이다. 나이는 내부 이용 시에는 14~19세 사이만 범주화하지만, 제3자 제공 시에는 10살 단위로 범주화하였다. 혹은 상하단 코딩 방식을 사용한다면 내부 이용 시에는 13세 미만 삭제 및 90세 이상은 90세로 경계치를 입력한 반면에, 제3자 제공 시에는 20세 미만은 삭제하고 80세 이상은 80세를 입력하여 처리환경 위험도에 따라 개인 식별이 더욱 어렵도록 엄격하게 가명처리하였다. 개별 구매액에 대해서도 내부 이용 시에는 상단 99.9%를 초과하는 경우 경계치로 변경하고, 모든 금액에 대해 1만 원 단위로 라운딩 처리를 했다. 반면에, 제3자 제공 시에는 상단 99.9%는 동일하게 하고, 모든 금액에 대해서도 10만 원 단위 미만은 1만 원 단위 라운딩, 10만 원 이상 1,000만 원 이하 단위는 10만 원 단위 라운딩, 1,000만 원 이상 단위는 100만 원 단위 라운딩으로 범주화 범위가 넓어진다. 고객 등급은 내부 이용 시에는 VIP와 S만 하나로 처리한 반면에, 제3자 제공 시에는 VIP, S, A는 1등급, B, C는 2등급, D, E, F는 3등급으로 처리했다.

62 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

[그림 2-14] 항목별 가명처리계획 작성 예시

순번	항목명	개인정보유형	내부 이용 또는 제3자 제공			
			처리 방법	처리 수준	처리 방법	처리 수준
1	고객ID	개인식별정보	대체	- 일련번호 대체	대체	- 일련번호 대체
2	나이	개인식별가능정보	변주화	- 14~19세 사이는 14~19세로 변주화	변주화	- 10살 단위 변주화
			상하단 코딩	- 13세 미만 삭제 - 90세 이상은 90세 이상 경계치 입력	상하단 코딩	- 20세 미만 삭제 - 80세 이상은 80세 이상 경계치 입력
3	주소	개인식별가능정보	부분삭제	- 동단위 이하 삭제	부분삭제	- 동단위 이하 삭제
4	성별	개인식별가능정보	처리 없음		처리 없음	
5	2019년 1월 여행용품 구매액	개인식별가능정보	변주화	- 상단 99.9%를 초과하는 경우 경계치로 변경 - 모든 금액에 대해 1만단위 라운딩 처리	변주화	- 상단 99.9%를 초과하는 경우 경계치로 변경 - 금액은 다음과 같이 변주화 적용 - 0원 : 0원 - 10만단위 미만 : 1만단위로 라운드 업 - 1,000만 단위 미만 : 10만 단위로 라운딩 - 1,000만 단위 이상 : 100만 단위로 라운딩
6	2019년 1월 식품류 구매액	개인식별가능정보				
7	2019년 1월 의류 구매액	개인식별가능정보				
8	2019년 1월 취미용품 구매액	개인식별가능정보				
9	2019년 1월 생활용품 구매액	개인식별가능정보				
10	2019년 1월 구매 총금액	개인식별가능정보	변주화	- 각 구매액과 동일한 처리	변주화	- 각 구매액과 동일한 처리
~	~	~	~	~	~	~
217	2021년 12월 식품류 구매액	개인식별가능정보	변주화	- 상단 99.9%를 초과하는 경우 경계치로 변경 - 모든 금액에 대해 1만단위 라운딩 처리	변주화	- 상단 99.9%를 초과하는 경우 경계치로 변경 - 금액은 다음과 같이 변주화 적용 - 0원 : 0원 - 10만단위 미만 : 1만단위로 라운드 업 - 1,000만 단위 미만 : 10만 단위로 라운딩 - 1,000만 단위 이상 : 100만 단위로 라운딩
218	2021년 12월 의류 구매액	개인식별가능정보				
219	2021년 12월 취미용품 구매액	개인식별가능정보				
220	2021년 12월 생활용품 구매액	개인식별가능정보				
221	2021년 12월 구매 총금액	개인식별가능정보	변주화	- 각 구매액과 동일한 처리	변주화	- 각 구매액과 동일한 처리
222	2021년 고객 등급	개인식별가능정보	변주화	- 식별가능성이 높은 VIP와 S를 하나로 묶어 VIP로 처리	변주화	- 다음과 같이 변주화 처리 - VIP, S, A → 1등급 - B, C → 2등급 - D, E, F → 3등급

자료: 개인정보보호위원회. (2022). 가명정보처리 가이드라인. p.32.

라. 가명처리 수행¹⁷⁾

가명처리 방법은 한 가지 표준이 정해져 있는 것이 아니라, 여러 가지 방법을 혼합하여 다양하게 사용할 수 있다.

17) 가명정보 처리 수행 방법은 개인정보보호위원회(2022)의 '가명정보 처리 가이드라인' p.67을 참고하여 정리하였음.

1) 삭제

개인정보를 삭제하는 방법은 전체를 삭제하거나, 일부 혹은 특이 정보에 해당하는 항목만 삭제할 수 있다. 삭제하는 방법은 공백이나 문자('*', '-', 등)로 대체한다. 예를 들어, 성명, 거주지 주소, 전화번호 등은 직접적으로 개인 식별이 가능하므로 삭제하는 것이 원칙이다. 그러나 두 기관의 데이터 결합을 위해서는 해시함수(SHA-256)나 salt 값 등을 적용하여 결합키를 생성하고 이용한다.

2) 통계 처리

평균값이나 최댓값, 최솟값, 최빈값, 중간값 등으로 처리하는 것이다. 예를 들면, 30대 남성 임금근로자 임금은 개별 정보가 아니라 그 평균 값을 사용하는 것이다.

3) 일반화(범주화)

주로 숫자 변수에서 사용하는 방법으로, 올림, 내림, 반올림 등의 기준을 적용하여 처리하는 방법이다. 양극단 값의 경우 사례가 적기 때문에 식별이 가능하여 상하단 코딩을 한다. 예를 들면 월급여액의 경우 상위 0.1% 혹은 상위 1%에 해당하는 사람은 상단코딩하여 경계치로 변경하고 동일한 수치를 개별 데이터에 부여함으로써 식별성을 낮춘다. 혹은 연령을 20세, 21세, 35세, 46세 등으로 제공하는 것이 아니라, 20대, 30대, 40대 순으로 범주화하여 제공한다.

4) 기타

개인정보에 임의의 숫자를 더하거나 곱하여 다른 수치로 잡음을 추가하거나, 주어진 입력값에 대해 예측이 불가능한 값을 생성하는 난수 생성, 식별이 높은 열 항목에 대하여 무작위로 순서를 변경하는 등 다양한 방법들이 있다. 또한 원문에 대한 암호화 작업을 통해 복호화 적용이 불가능하게도 할 수 있다.

그 외 모집단이 아니라 표본을 추출하거나, 원본과 유사한 통계적 성질을 가진 가상의 데이터를 생성하는 방법도 존재한다.

마. 적정성 검토위원회 운영

적정성 검토위원회는 가명처리 후 적정성 검토를 수행하는 위원회이다.¹⁸⁾ 적정성 검토위원회의 구성은 적합성 검토위원회와 동일하게 기관의 성격에 맞게 다른 위원회로 대체 가능하다. 공공분야는 공공데이터 제공 심의위원회에서, 보건의료 분야는 데이터 심의위원회에서, 금융 분야는 데이터 전문기관을 통해 위탁 평가할 수도 있다.

1) 적정성 검토위원회 구성

검토위원회의 위원의 자격은 개인정보보호 법·제도 및 가명처리 기법 등에 관한 전문가¹⁹⁾로 일반적인 환경에서 3인, 그리고 보건의료 환경에서는 5인 이상으로 구성할 수 있다. 가명정보를 외부에 제공하는 경우 외부 검토위원 1명을 반드시 포함해야 하고, 익명정보를 외부에 제공하는

18) 개인정보보호법 제24조 제2항에 근거함.

19) 가명정보 전문가는 한국인터넷진흥원에서 2021년부터 '가명정보 전문가 풀'을 구성하여 운영하고 있음.

경우에는 외부 검토위원을 과반으로 구성해야 한다. 적정성 검토를 위해 검토하게 될 데이터는 개인정보에 해당하므로, 개인정보 보호법에 따라 각 위원들에게 ‘보안서약서’를 제출받아야 한다.

2) 적정성 검토 과정

적정성 검토 과정은 가명처리 전 단계에서 해당 연구가 가명정보 이용의 목적에 부합하게 설계되었는지부터, 실제 가명처리는 개인이 식별되지 않도록 안전하게 이루어졌는지, 가명처리 후 가명정보 이용 목적을 달성할 수 있는지 등 총괄적으로 평가한다. 세부 검토 과정은 다음과 같다.

첫째, 사전준비 단계에서 제출된 필요 서류가 법·제도의 목적에 적합하게 작성되었는지 검토한다. 해당 연구가 처리 목적을 구체적으로 설정하였는지, 설정된 목적이 가명정보 이용의 목적인 통계 작성, 과학적 연구, 공익적 기록보존에 적합한지 검토한다. 검토 서류는 가명정보 이용·제공 신청서, 가명정보에 대한 안전조치 이행 약약서, 통계 작성 계획서 등이다.

둘째, 가명처리 단계에서 데이터의 식별 위험성과 처리 환경의 식별 위험성 등을 반영하였는지 검토한다. 이때 식별 위험성 검토 점검표 및 결과보고서를 기반으로 검토한다. 검토하는 서류는 개인정보 유형 분류표, 활용데이터 요구 수준표, 식별 위험성 검토 결과 보고서 등이다.

셋째, 항목별 가명처리계획이 적정하게 계획되었는지 확인한다. 즉 제3자 제공의 경우 위험도가 높아지므로 가명처리도 더욱 엄격하게 진행되었는지에 대해 항목별 가명처리 계획서를 통해 검토한다.

넷째, 계획한 가명처리 방법 및 수준에 따라 실제 가명처리를 수행하였는지 검토한다. 본 과정에서는 가명정보 처리 기초데이터 명세서, 원본데이터 세부 항목별 명세, 원본데이터, 개인정보 속성, 원본데이터 분포, 평가 대상 데이터 세부 항목별 명세 집계 데이터 등에 대하여 종합적으로

검토한다.

다섯째, 가명처리된 데이터가 법·제도·기술적 측면에서 본 연구의 목적을 달성할 수 있을지 검토한다.

위 과정들 속에서 가명처리된 정보가 식별의 위험이 있거나, 혹은 과도한 식별로 연구의 목적을 달성하기 어렵다고 판단되면 추가 가명처리를 권고 할 수 있다.²⁰⁾

3) 검토 결과 통보

모든 위원은 적정성 검토 결과, 최종결과에 대하여 각각 ‘적정성 검토 결과서’를 적성하고, 위원장은 각 검토위원들의 의견을 종합해 ‘적정성 검토 종합 결과서’를 작성하여 정보제공 요청자에게 통보한다. 검토 결과는 적정(승인), 조건부 승인, 부적정(반려) 중에 선택한다. ‘적정’은 즉시 결합전문기관에 데이터를 제공하고, ‘조건부 승인’은 추가 가명처리나 필요사항 들을 수행한다. ‘부적정’ 판정 시에는 가명정보 제공 계획을 중단한다.

20) 적정성 검토 과정에서 필요한 서류인 가명정보 이용·제공 신청서, 가명정보에 대한 안전 조치 이행 약속서, 통계 작성 계획서, 개인정보 유형 분류표, 활용데이터 요구 수준표, 식별 위험성 검토 결과 보고서, 항목별 가명처리 계획서, 가명정보 처리 기초데이터 명세서, 원본 데이터 세부 항목별 명세, 원본데이터, 개인정보 속성, 원본데이터 분포, 평가 대상 데이터 세부 항목별 명세 집계 데이터 등은 [부도 1] 참고

바. 추가 가명처리

적정성 검토 결과, 식별 위험성에 따라 추가 가명처리가 권고된 경우에는 앞에서 수행한 가명처리 단계인 대상 선정, 위험도 측정, 가명처리 수준 정의, 가명처리 과정을 다시 수행한다.

사. 가명정보 전송

개별 데이터 보유기관은 적정성 검토에서 데이터 제공이 ‘승인’되면 결합전문기관에 가명처리된 정보를 전송한다.

제공 방법은 가명정보의 위험성 정도를 고려하여 온라인 제공, 오프라인 제공, 안심 구역 설정 등으로 결정한다. 온라인 제공 시 암호 알고리즘을 적용하여 암호화 후 전송하여야 한다. 오프라인 제공 시에도 암호 알고리즘을 적용하여 암호화하고 2개 이상의 복사본을 함께 저장하여 보완 USB 등 보조저장 매체에 저장하여 전달한다. 안심구역에서 사용한다면 허가를 받은 사람만 출입 또는 이용 가능하며, 관련 기록을 남겨야 한다(행정안전부·개인정보보호위원회, 2021).

아. 가명정보의 안전한 사후관리

가명처리된 정보가 특정 개인을 식별할 위험성이 증가하지 않도록 지속적으로 모니터링하여 안전하게 처리해야 한다. 가명정보의 관리적·기술적·물리적 보호조치가 필요하다.

1) 가명정보의 관리적 보호조치

가명정보를 제공하기 위해서는 내부관리계획의 수립이 필수적이며, 이를 시행하여야 한다. 또한 가명정보 처리업무를 외부에 위탁하는 경우 재식별 금지, 타인에게 다시 제공 또는 재위탁 제한, 재식별 위험 발생 시 통지 등에 관한 내용을 포함하여 위탁계약을 작성해야 한다.

2) 가명정보의 기술적 보호조치

가명정보와 추가정보를 분리하여 별도로 저장, 관리해야 한다. 또한 가명정보 취급자가 가명처리하는 시스템 외의 특정 개인을 알아 볼 수 있는 다른 개인 정보처리 시스템에 접근할 수 없도록 권한을 제한해야 한다.

가명정보를 제공했던 가명정보 보유기관(금융 기관 해당)은 개별적으로 기관의 홈페이지에 가명정보 처리 및 제3자 제공에 관한 사항을 공개하여야 한다. 공개 내용은 제공받는 자, 제공받는 자의 이용 목적, 제공하는 가명정보의 종류, 제공하는 가명정보의 항목, 담당부서 등이다([그림 2-15] 참조).

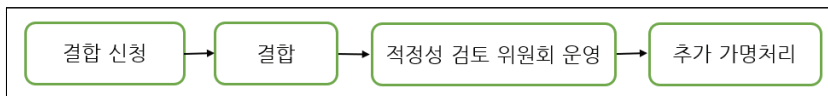
3) 가명정보의 물리적 보호조치

가명정보가 전산실이나 데이터 보관실에는 보관되어 있다면 허가를 받지 않은 사람의 접근으로부터 보호하기 위해서 출입통제 등의 절차를 수립하여야 한다. 가명정보가 보조저장기기에 저장되어 있다면 잠금장치가 있는 안전한 장소에 보관하여야 한다.

4. 결합 단계

결합 단계는 결합전문기관에서 수행한다. 정보를 보유한 개별 기관은 가명처리 후 결합전문기관에 결합을 신청한다. 이미 사전협의 과정을 거쳤기 때문에 특별한 사유가 없는 한 승인이 된다. 결합전문기관은 전송받은 가명처리된 정보들을 모두 결합하고, 모든 정보가 결합된 데이터에 대하여 반출심사위원회를 운영한다. 필요에 따라 추가 가명처리를 진행한다.

[그림 2-16] 결합 단계



자료: 필자 작성

가. 결합신청 및 결합

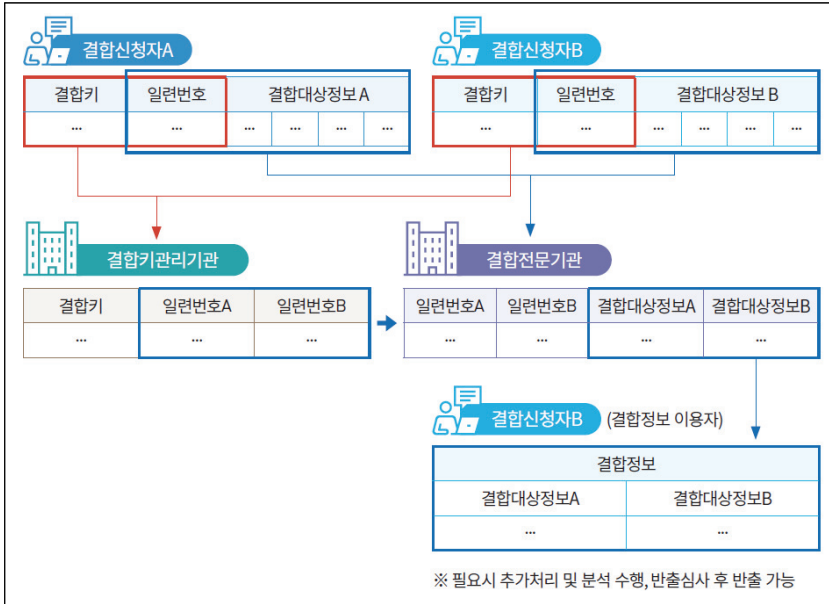
개별 데이터 보유기관은 결합전문기관에서 결합을 신청하고 결합이 ‘승인’되면 결합키 연계정보, 가명처리된 데이터를 결합전문기관에 전송한다. 결합전문기관은 결합키 관련 정보를 이용하여 개별 기관의 결합대상 정보를 모두 결합한다.

나. 추가 가명처리

결합전문기관은 모든 정보가 결합된 데이터에 대하여 데이터의 분포와 내용을 검토하여 특이치가 새로 발견되면 재식별을 낮추기 위해 추가 가명처리를 실시한다. 금융 분야 데이터의 경우 데이터전문기관에서 추가 가명처리를 수행하고, 개인정보보호법이 적용되는 기관은 반출 기관이

추가 가명처리를 수행한다.

[그림 2-17] 결합전문기관의 가명정보 결합 과정

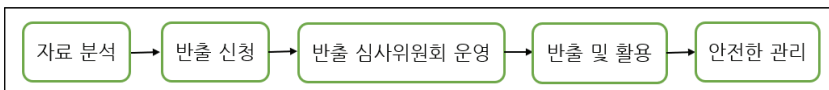


자료: 개인정보보호위원회. (2022). 가명정보처리 가이드라인. p.55.

5. 반출 및 사후관리

결합을 수행한 결합전문기관에 반출 신청을 한다.

[그림 2-18] 반출 및 사후관리



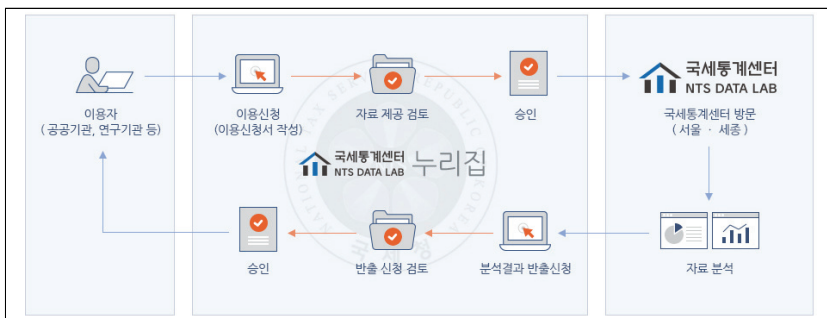
자료: 필자 작성

가. 데이터 분석

연구자는 가명처리가 완료된 이후에 결합전문기관의 안심구역에서 데이터를 분석할 수 있다. 결합전문기관의 안심구역은 정보의 안전한 관리를 위해 기술적·관리적·물리적으로 외부와 단절된 폐쇄 환경으로, 출입 통제시스템과 실내 감시 카메라 등이 설치되어 있다. 또한 휴대전화, 개인 노트북, USB 등의 저장 장치, 기타 전자 기기의 반입이 통제되며 인터넷 사용도 불가능하다. 안심구역 안에서 SAS, STATA, SPSS, R 등의 통계프로그램이 설치된 기관의 컴퓨터를 이용하여 분석하여야 한다. 따라서 방문하기 전에 분석에 필요한 코드나 참고자료를 준비해서 사용 가능 여부를 안심구역 관리자와 상의해야 한다.

안심구역 이용 방법은 결합전문기관마다 거의 유사하다. 해당 사이트에서 이용 신청 후 승인이 되면 사용할 수 있다. 예를 들어, 결합전문기관인 국세통계센터는 세종 8석과 서울 8석, 총 16석 규모의 데이터 분석 공간이 있으며, 운영 시간은 평일 오전 9시~오후 6시(점심시간 12시~13시)이다. 자세한 이용 절차는 [그림 2-19]와 같다.

[그림 2-19] 국세통계센터 이용 절차



자료: 국세청, 국세통계센터, <https://datalab.nts.go.kr/websquare/websquare.html?w2xPath=ui/main/index.xml>, 2022. 8. 4 인출

나. 반출 신청

연구자는 데이터 분석 과정이 완료되면 결합전문기관에 분석 결과표 반출을 요청할 수 있다. 반출 신청 시에는 반출 신청서와 함께 반출 대상 정보에 관한 서류, 반출 목적을 증명할 수 있는 서류, 반출 정보의 안전조치 계획 및 이를 증빙할 수 있는 서류 등을 제출한다.

다. 반출 심사위원회 운영

가명정보 반출 심사위원회 운영은 「가명정보의 결합 및 반출 등에 관한 고시」에 따라 이루어진다.

결합전문기관은 결합신청자가 반출을 요청하면 영업일 기준 5일 이내에 반출 심사위원회를 구성하고, 일정을 정하여 결합신청자에게 통지하여야 한다. 반출 심사위원회 위원²¹⁾은 3명 이상으로 구성되며 개인정보 보호에 관하여 학식과 경험이 풍부한 사람으로 선정한다.

반출 심사위원회는 첫째, 결합 목적과 반출 정보가 관련성이 있는지, 둘째, 특정 개인을 알아볼 가능성이 없는지, 셋째, 반출 정보에 대한 안전조치 계획이 있는지 등을 심사한다.

라. 안전한 관리

결합된 가명정보의 안전한 관리는 앞에서 설명한 가명처리된 정보의 안전한 관리와 동일하다. 결합전문기관도 개별 가명정보 제공 기관과 마찬가지로 가명정보의 관리적·기술적·물리적 보호조치를 수행해야 한다.

21) 위원회 위원 구성과 관련하여서는 〈표 2-5〉 참고

제3절 소결

이 장에서는 가명정보 결합 및 반출 절차에 대해 구체적으로 살펴보았다. 활용하고자 하는 가명정보에 따라 절차가 조금 상이한 경우도 있기 때문에 각각의 가이드라인을 아울러서 정리하였다. 가명정보 결합은 개인 식별이 불가능하도록 가명처리된 가명정보에 대하여 결합전문기관이 두 개 이상의 데이터를 결합키를 이용하여 통계작성, 과학적 연구, 공익적 기록보존 등의 목적에 한해 결합하는 것이다. 가명정보 결합 및 반출 절차는 사전준비 단계, 결합신청, 가명처리, 결합수행, 반출 및 관리 단계로 구성되고, 각각의 단계에서 데이터 이용자와 데이터 제공자, 결합전문기관, 결합키 생성기관이 유기적으로 관계하면서 작업을 수행한다.

가명정보 결합은 개인정보를 정보 주체의 동의 없이 사용할 수 있기 때문에 개인 식별의 위험성을 없애고 개인정보의 안전한 사용을 위해 복잡한 절차들을 수행하도록 법적으로 규제하고 있다. 따라서 각 과정에서 주의해야 할 사항들을 숙지해야만 가명정보 결합에 소요되는 시간을 줄이면서, 결합 후 정확한 분석을 시도할 수 있다. 가명정보 결합 시 유의사항은 다음과 같다.

1. 결합 시 유의 사항

첫째, 가명정보 결합에 앞서 결합하고자 하는 정보의 특성을 파악하여 제공받을 수 있는 변수의 수준을 파악해야 한다. 대부분의 행정데이터는 데이터 이용을 목적으로 생성된 데이터가 아니기 때문에, 코드북과 같이 데이터의 특성을 설명한 문건이 충분하지 않다. 예를 들면, 직종 변수에 대하여 기관 홈페이지에는 대분류만 제공하지만, 직접 기관을 방문하면

세분류를 제공하는 경우가 있다. 또한 코드북이 없기 때문에 변수의 결측치 비율이나 응답 비율을 알 수 없어 막상 결합한 이후에 변수를 사용하지 못하는 경우가 발생하기도 한다. 이외에도 데이터의 제공 가능 시기가 기관마다 상이할 수도 있으므로, 결합하고자 하는 정보의 특성을 기관의 협조를 받아 가급적 자세히 파악하도록 노력해야 한다.

둘째, 행정데이터의 경우는 데이터 보유기관의 승인뿐만 아니라, 해당 정부 부처의 승인을 받아야 한다. 예를 들어, 기초연금은 보건복지부에, 주거급여는 국토교통부에, 교육급여는 교육부에 각각 승인을 받아야 한다. 또한 여러 개 부처 중 한 개 부처에서 승인을 해주지 않을 경우, 가명정보 결합 데이터를 이용하고자 하는 목적을 달성하지 못하게 된다. 이 경우에는 연구계획서를 새로 작성해야 하기 때문에 결합 전 사전준비 기간이 길어질 수 있음에 유념해야 한다.

셋째, 복잡한 과정들을 거쳐 결합이 완료된 이후 데이터를 분석하는 과정에서 실제 결합률이 낮거나, 변수 결측값이 많아 분석에 사용할 응답 값이 적거나, 가명처리 수준이 높아 계획했던 분석이 불가능한 경우가 발생할 수 있다. 특히 가명처리의 경우 연구자는 가급적 자세한 정보를 제공받기를 원하고, 데이터 보유기관과 결합전문기관은 개인 식별 위험도를 낮추는 것을 원하기 때문에 연구자와 기관 간의 협의가 매우 중요하다.

2. 가명정보 결합의 한계

행정데이터와 조사데이터의 가명정보 결합은 연구의 확장성 측면에서 기존의 조사데이터만 사용하거나 행정데이터만 사용했던 많은 연구자들에게 희소식이다. 그럼에도 불구하고, 앞에서 언급한 유의 사항들뿐만 아니라 가명정보의 결합 및 반출 과정 전반에 대한 구조적인 한계가 존재한다.

첫째는 가명처리와 결합 및 반출의 복잡한 절차이다. 개인정보를 정보 주체인 개인의 동의를 받지 않고 사용하기 때문에 개인정보의 오남용을 막기 위해서 복잡한 절차를 도입했을 것이라 생각한다. 공공기관에서 가명정보 결합을 신청했음에도 익숙하지 않은 신청 과정 때문에 여러 번 수정·보완해야 하는 것을 고려해 보면, 개인 연구자가 가명정보 결합을 시도하는 것은 매우 어려울 수 있다.

둘째는 이렇게 복잡한 절차와 과정 속에 연구자의 참여가 제도적으로 보장되어 있지 못하는 점이다. 구인회 외(2021)는 다층노후소득보장체계 분석을 위해 행정안전부, 국세청, 보건복지부, 교육부, 고용노동부, 금융위원회 등 10개 소관부처의 데이터를 결합하는 과정 속에서, 연구자가 변수의 선정 및 표본 추출 과정에 직접 참여하지 못하여 발생한 결합데이터의 한계를 기술하였다. 요청데이터에 비해 제공데이터의 시점이 제한적이며, 일부 변수는 포함되지 못함으로써 연구 설계 당시의 분석 목표를 달성하지 못하였다고 하였다.

셋째는 사전협의 과정과 가명정보 처리, 결합 및 반출에 걸리는 기간이 너무 길다는 것이다. 장인성 외(2022)²²⁾에 따르면 데이터 보유기관과 결합전문기관과의 사전협의 과정이 거의 6개월이 소요되며, 실제 신청부터 반출까지를 서둘러서 진행하여도 6개월 이상 소요된다는 것이다. 즉 데이터를 결합하는 기간이 1년 이상 걸리므로 장기 프로젝트에서만 시도해 볼 수 있을 것이다.

넷째, 연구자가 가명처리에 적합한 목적으로 가명정보 결합을 신청하여도 데이터를 보유하고 있는 기관이 가명정보를 제공해야만 하는 의무

22) 한국노동연구원에서 발간한 ‘장인성 외(2022), 패널자료 품질개선 연구(XII)’ 보고서의 제2장 ‘가명정보 결합 가능성 검증 연구’는 이지은과 장인성이 집필하였으며, 한국노동패널조사 자료와 국민건강보험공단 및 건강보험심사평가원 행정 자료를 가명처리하여 결합을 실시한 연구임.

가 있는 것은 아니다. 각 기관에서 가명정보 이용·제공 신청에 대한 적합성을 검토할 때, 해당 기관에 가명정보와 관련된 내부관리계획이 수립되어 있지 않거나, 기관의 본래 업무수행에 상당한 지장을 초래할 경우 제공을 거부할 수도 있다.

다섯째, 가명정보 결합 데이터는 신청자 외에는 접근할 수 없으며, 데이터 활용 기간이 끝난 후에는 지체 없이 파기해야 하는 일회성 데이터이다. 연구 도중에 연구자를 충원해도 결합신청 당시 신청한 연구자가 아니면 데이터에 접근할 수 없다. 또한 동일한 데이터를 이용하여 다른 연구를 진행하고자 하여도, 이미 데이터가 폐기되었기 때문에 사전협의 단계부터 다시 시작해야 하는 한계가 있다.

가명정보 결합의 한계가 개선된다면 행정데이터와 조사데이터의 가명정보 결합은 연구의 확장성 측면에서 많은 도움이 될 수 있다고 생각한다.

사람을
생각하는
사람들



KOREA INSTITUTE FOR HEALTH AND SOCIAL AFFAIRS



제3장

가명정보 결합을 활용한 통합데이터 구축 및 연구 사례

제1절 국내 사례

제2절 해외 사례

제3절 소결

제3장

가명정보 결합을 활용한 통합데이터 구축 및 연구 사례

제1절 국내 사례

1. 조사데이터와 행정데이터의 결합

가. 통계청(2020). 「가계금융복지조사에서의 조사자료와 행정자료의 통합방법의 이해」

통계청에서는 가계의 경제상황, 재무건전성, 소득분배, 빈곤 현황 같은 가구의 전반적인 경제 수준, 규모의 변화 및 구조적 특성을 파악하기 위해 매년 2만 가구를 대상으로 가계금융복지조사를 실시하고 있다. 가계금융복지조사는 조사원이 가구를 방문하여 가구의 대표자 또는 가구원을 대상으로 직접 면접하여 조사표를 완성하는 대면조사 방식을 사용하는 데, 이러한 현장조사는 소득 등 민감한 조사항목을 정확하게 측정하는 것이 어렵고, 개인정보보호 및 사생활 보호에 대한 인식이 높아지는 인구사회적 구조 변화로 인해 더욱 어려워지고 있는 것이 현실이다.

이러한 현실적 어려움으로 인해 발생하는 응답 거부나 응답 착오로 인한 응답오차는 자료의 신뢰성을 떨어뜨릴 수 있기 때문에 조사자료의 품질을 높이기 위한 방안으로 행정자료와 매칭하는 방법이 활용되고 있다.

통계청(2020)에서는 가계금융복지조사의 각 조사 소득 항목들이 다양한 행정자료들과 매칭되는 과정을 자세하게 설명하고 있다. 우선, 활용 가능한 행정자료의 여부를 확인하게 되는데, 국세청의 과세자료, 보건복지부의 사회수혜금, 각 연금공단의 국민연금, 공무원연금 등과 각 공단의

연금기여금 및 건강보험료 지출 등이 이에 해당한다. 다음으로, 행정자료의 개념이 얼마나 같은지 검토하게 되는데, 행정자료를 통해 조사통계의 품질 제고와 표본조사의 효율성을 높이기 위해 조사자료를 비교 기준으로 삼게 된다. 마지막으로, 개념 비교가 끝난 조사자료와 행정자료를 개인 단위로 결합을 진행한다.

조사자료와 행정자료 매칭의 첫 번째 과정은 두 자료를 연계할 연계명부를 작성하는 것이다. 연계명부에는 개인 단위의 성별, 이름, 생년월일, 주소 같은 텍스트 정보를 포함하고 있다. 두 번째 과정은 연계명부에 대한 연계키를 해당 행정기관에 요청하고 해당 기관으로부터 명부에 붙은 연계키를 제공받는 것이다. 마지막으로 연계키를 포함한 행정자료 입수 명부를 각 행정자료 보유기관에 보내고, 각 해당 기관으로부터 연계키 단위의 행정자료를 제공받게 된다. 이러한 과정에서 두 자료를 연계할 수 있는 매칭키를 찾는 것이 중요한데, 개인의 이름, 성별, 생년월일, 주소 등의 변수들이 매칭변수로 활용되고 있으나 이들 변수들은 면담조사 시기입오류, 기억오류 등의 응답오차가 발생할 수 있어, 등록기반 행정자료의 확인 과정이 필요하다.

통계청(2020)에 따르면, 가계금융복지조사의 매칭키 확보율은 약 92~95%로 높은 편으로 매칭 편향에 대한 우려는 크게 없는 것으로 나타났다. 다만 특정 그룹, 즉 혼자 떨어져 사는 특성이 있는 1인 가구 또는 30대 연령 그룹에서 연계율에 차이가 있을 수 있으므로 통계적 기법 등에 의해 편향을 보정할 필요가 있다. 보정은 조사값이 있는데 매칭키가 없는 가구원에 대해 이루어지며, 아이디어는 행정자료와 조사자료의 상대적 비율만큼 조정하는 데 있다. 조사값이 있는 경우 회귀모형 기반 혼합비(ratio) 대체 방법을 사용한다. 모형에서 행정자료로 보완된 값이 관심변수이며 해당 변수의 조사값은 보완값과 강한 양의 상관을 가지므로 조사

값이 보완값의 핵심 예측변수가 될 수 있다. 연구 결과, 근로소득과 금융 소득 항목에서 편향 보정 효과가 큰 것으로 나타났다.

행정자료를 활용하기 위해서는 조사자료의 측정오차를 최소화하기 위한 자료의 전처리 과정이 필요하다. 기본적으로 조사자료와 행정자료의 개념이 일치하거나 포괄 범위가 같은 경우에 대해서는 행정자료가 더 정확하다고 가정한다. 조사자료를 기준으로 특정 항목에 대해 응답하지 않거나, 응답 착오일 것으로 판단되는 경우에는 행정자료로 조사자료를 대체한다. 개인 단위에서 조사자료를 수정하고 대체하는 것을 원칙으로 하며, 조사 항목의 개념 및 포괄 범위에 부합하도록 행정자료가 포함하고 있는 다양한 신고 서식의 분류별 상세코드를 확인하여 두 자료의 포괄범위가 일치할 수 있도록 여러 단계의 자료처리 과정을 거치게 된다. 조사자료와 비교하기 위해 소득원천별로 다출처 행정자료를 통합하여 행정자료 기반 통합근로소득, 통합사업소득, 통합소득세 등을 작성하고, 이러한 다출처 자료들은 결합 분석을 통해 자료의 중복성 및 일치성 등을 복합적으로 고려하여 최종적으로 행정자료의 활용 기준을 정하게 된다.

통계청(2020)은 2019년의 조사 항목에 대한 행정자료의 활용 현황을 <표 3-1>과 같이 제시하였는데, 표에서 'x'는 행정자료가 미활용되었음을 나타내고, 음영은 무응답에만 이용되었음을 의미한다.

통계청(2020)에서 행정자료를 활용하여 가계금융복지조사 자료를 대체·보완한 결과는 다음과 같다. 보완 전에 비해 전체적인 가구소득이 상향 조정되었는데, 2016년 자료의 경우 보완 후의 가구당 평균소득은 9.2%, 중위소득은 6.1% 각각 증가하였다. 소득구간별로는 보완 전에 비해, 소득구간 2~3천만 원 미만, 8천만 원 이상 구간에 속한 가구 비율이 증가하였고, 이러한 현상은 연도별로 유사한 것으로 보인다. 조사자료의 보완으로 지니계수는 2016년 처분가능소득 기준으로 0.342에서 0.355

로 증가하였고, 상대적 빈곤율은 16.1%에서 17.6%로 증가하였다. 특히 가구주가 60세 이상인 가구의 소득을 보완하는 데 크게 기여한 것으로 나타났다. 이는 조사자료에서 60세 이상의 고연령층 가구에서 실제로 금융자산이 있음에도 불구하고 금융소득이 없다고 응답한 가구가 많았는데, 행정자료를 통해 이러한 무응답들이 크게 보완된 데 기인한다.

〈표 3-1〉 행정자료 활용 현황(2019년 기준)

조사 소득 항목			입수한 행정자료
소득	근로소득		• 근로소득(종합소득세, 원천세), 일용근로소득, 자활급여
	사업소득		• 사업소득(종합소득세, 원천세)
	재산소득	금융소득	• 금융소득(종합소득세, 원천세)
		임대소득	• 사업소득 중 부동산 임대소득(종합소득세)
	공적이전소득	공적연금	• 연금지급액
		기초연금	• 기초연금
		양육수당 (출산장려금 포함)	• 양육수당, 양육보조금, 한부모가족아동양육수당, 아동수당(2018년 정책) 등
		장애수당 (장애아동수당, 장애인연금 등)	• 장애연금, 장애수당, 장애아동수당, 장애인급여
		기초생활보장지원금	• 생계급여, 주거급여, 교육급여, 해산·장제급여
		근로자녀장려금	• 근로·자녀장려금
지출	공적이전지출	소득세	• 소득세(종합소득세, 원천세, 일용근로소득세)
		재산세*	×
		자동차세*	×
	공적연금·사회보험료	국민연금기여금	• 국민연금(연금공단)
		기타 연금기여금	• 공무원연금, 사학연금, 별정우체국연금, 군인연금(공단)
		건강보험료	• 건강보험료, 노인장기요양보험료
		고용보험료*	×

자료: 통계청. (2020). 가계금융복지조사에서의 조사자료와 행정자료의 통합방법의 이해. p.55.

통계청(2020)의 연구를 통해 행정자료를 활용하여 조사 효율성을 높일 수 있음을 확인하였는데, 이렇게 행정자료를 활용할 경우 통계품질 개선 및 응답 부담을 줄이는 효과를 기대할 수 있다. 다만 행정자료 활용을 통한 통계품질 개선은 행정자료의 품질이 얼마나 높은가와 밀접한 관련이 있다는 점과 다출처 자료를 결합하는 데 있어서 각 자료가 갖는 속성과 구조에 대한 충분한 이해를 바탕으로 명확한 자료처리 기준을 마련하는 것이 필요하다는 점을 해당 연구는 강조한다.

나. 임성희, 김진옥(2021). 「개인 창업사례 생존분석: 2010~2018 행정데이터를 중심으로」

임성희·김진옥(2021)은 우리나라 창업생태계의 대부분을 차지하는 개인에 대한 생존율 및 생존요인을 분석하고자 하였다. 현재 우리나라 기업의 생존율은 통계청의 기업생멸행정통계로 확인할 수 있는데, 기업생멸행정통계의 생존율은 기준 연도별로 작성되기 때문에 만약 전체 기간에 대한 생존율을 알고 싶은 경우에는 연도별 평균을 이용할 수밖에 없고, 업종별 교차영역에 대한 생존율에 대해서는 파악할 수 없는 경우 등 제약 사항이 존재했다. 이러한 이유로 임성희·김진옥(2021)은 개인 창업사업체에 대한 종단적인(longitudinal) 자료 분석이 가능한 연간 SBR(statistics business registers, 통계기업등록부)을 사용하여 분석을 진행하였다. SBR은 통계청에서 경제통계 분석을 위해 제공되는 조사자료와 과세자료를 결합한 자료이다.

임성희, 김진옥(2021)은 SBR 자료(2010~2018년 기준 9개년)를 기본 베이스로 하고, 개별 단위의 식별정보를 이용하여 패널자료 형태로 DB화한 후 추가적인 2차 행정자료를 결합하여 분석에 이용하였다. 결합된 2차

행정자료는 가용한 자료 중에서 SBR 분석 단위와 연계 가능한 식별정보가 있는지, SBR에 결합되었을 때 그 숫자가 유용한 일정 수준 이상이 되는지, 분석에 유의한 변수를 포함하는지 등을 검토하였으며, 최종적으로 공동사업자명세자료(2010~2018년)와 서비스업조사의 프랜차이즈 가맹점 모집단(2010~2018년)이 선정되었다.²³⁾

임성희·김진옥(2021)의 자료 추출 과정은 다음과 같다. SBR의 ‘조직구분’ 항목을 이용하여 개인 단위만을 추출하고, 사업자등록번호에서 조직형태가 개인이 아닌 것으로 추정되는 사업체는 삭제하였다. ‘개업일’ 항목으로 각 연도마다 신규로 창업한 사업체를 추출하여 기초 DB를 생성한 후, 식별정보를 이용하여 2차 자료인 공동사업자명세자료와 같은 프랜차이즈 가맹점 모집단 자료를 연계하였다. 연간 데이터를 식별정보를 이용하여 동일한 단위별로 종단적으로 연계하여, ‘사업자등록일’, ‘개업일’, ‘폐업일’에 오류가 있거나 ‘사업자등록일’에서 ‘폐업일’까지가 한 달 이내 이거나, ‘활동구분’ 항목에서 비활동에 해당하는 경우 삭제 처리를 한 결과, 최종 분석 대상 사업체 수는 8,180,131개이고, 분석 항목은 <표 3-2>와 같다.

임성희·김진옥(2021)의 분석자료는 2010~2018년간 창업한 개인사업체들의 패널 데이터이며 연구 대상의 종료 시점은 2018년 12월 31일이 된다. 연구종료 시점까지 존속한 사업체는 폐업(death)이 아닌 중도 절단(censoring)된 경우로서, 자료 분석 시 폐업과 중도 절단은 구분되어야 한다. 생존 분석(survival analysis)은 중도 절단 자료를 포함할 수 있으며, 생존시간 추정, 집단 간의 생존율 차이 분석, 생존위험 요인 분석 등을 분석할 수 있는 방법론이다. 임성희·김진옥(2021)의 연구 대상이 되는 개인사업체의 생존 자료는 진입 시기가 사업체의 창업 일자마다 다르고,

23) 공동사업자명세자료와 프랜차이즈 가맹점 모집단은 통계청 내부 자료임.

연구 종료 시점을 제외하고는 신고 누락 등으로 인한 중도 절단 시기도 사업체마다 다르기 때문에 임의중도절단²⁴⁾에 해당된다고 볼 수 있다.

〈표 3-2〉 자료출처 및 이용 항목

자료출처	분석 항목	비고
SBR	사업자등록번호/종원사업장일련번호	종단 DB 연계 시 이용
	개인대체식별번호	2차 자료 결합 시 이용
	대표자 생년월일	창업 당시 대표자 연령 산출
	대표자 성별	
	행정구역코드(7자리)	도시(동)/비도시(읍면) 구분 수도권·광역시/기타 지역 구분
	등록일자/폐업일자	사업체별 생존시간 산출
	업종코드(6자리)	
	SBR 연간 매출액	
	SBR 연간 종업원 수	
공동사업자 명세서	공동 창업 여부	
서비스업 조사	프랜차이즈 가맹점 여부	

자료: 임성희·김진옥. (2021). 개인 창업사업체 생존분석: 2010~2018 행정데이터를 중심으로. p.7.

임성희·김진옥(2021)은 캐플란-마이어 방법을 적용하여 2010~2018 년간 창업한 개인 사업체들의 생존시간을 추정하였다. 추정 결과, 중위수 생존시간은 2.6년(평균 3.8년)이었으며, 1년 생존율은 78.9%, 3년 생존율은 45.6%, 5년 생존율은 31.4%였다. 창업 이후 1.5년까지가 폐업 위험이 가장 높고, 이후부터는 점차 감소하는 것을 확인하였다. 개인 창업 사업체의 중위수 생존시간은 여성 대표자가 남성 대표자보다 6개월 짧게 나타났으며, 창업 당시 연령층별 생존율은 청년층의 생존율이 다른 연령

24) 임의중도절단(random censoring): 중도 절단의 3가지 유형 중 하나로 일정한 기간을 두고 진행되는 추적조사에서 연구집입 시기가 다르고, 중도 절단된 시기도 다른 경우의 중도 절단을 가리킴.

층보다 현격하게 낮으며 폐업 발생 위험도는 가장 높은 특징을 보였다. 공동사업자 유무에 따른 중위수 생존시간은 공동사업자가 있는 경우에 3.3년이 더 길었으며, 프랜차이즈 가맹점 여부에 따른 중위수 생존시간은 가맹점에 해당하는 경우 1.8년이 더 길게 나타났다. 마지막으로 종업원 유무에 따른 중위수 생존시간은 종업원이 있는 경우에 1.3년이 더 길었고, '수도권+6대 광역시'와 '기타 지역'으로 지역을 구분했을 때 '기타 지역'의 생존시간이 1.2개월 더 긴 것으로 나타났다. 분석 결과를 정리하면 '공동사업자 있음'과 '프랜차이즈 가맹점 해당'은 통계적으로 유의한 대부분의 업종에서 생존위험에 감소 효과가 있는 것으로 나타났다. 연간 종업원 수와 연간 매출액 1단위 증가는 생존위험에 감소 효과를 나타냈으며, 성별과 도시지역은 업종마다 다른 결과를 나타내었다.

2. 행정데이터와 행정데이터의 결합

가. 정찬미(2021). 「1인 가구의 사회보장 수급현황 및 정책방향: 행정데이터를 활용하여」

정찬미(2021)는 1인 가구의 인구·경제학적 특성과 사회보장 정책 수급 현황을 지역, 성별, 소득분위 등으로 구분하여 분석하고, 1인 가구 내의 이질적 특성 및 사회보장 욕구에 대해 파악한 것을 토대로 적절한 정책적 대안을 제시하고자 하였다. 이를 위해 사회보장위원회가 사회보장 정책 연구를 위해 수집한 행정데이터를 이용하였는데, 이 데이터는 사회보장정책을 수행하는 11개 부처, 18개 자료 보유기관의 협조로 생성된 자료이다. 주민등록전산정보에서 전 국민의 7%를 표본추출(층화추출)하여 개인 단위로 사회보장 수급 관련 행정데이터를 결합하였다. 이 자료의

특징은 주민등록상의 동일 세대 구분이 가능하여 세대 단위의 분석도 가능한 점인데, 이로 인해 정찬미(2021)의 연구에서의 ‘가구’는 사실상 주민등록상의 ‘세대’를 의미하고 있다. 데이터 수집 시, 정보 보유기관의 상황 및 정책 시행 간격(월, 분기, 연 등)에 따라 처리 시점이 다양하기에 전체 기관의 개인의 사회보장 수급 상황을 동일 시점으로 수집 가능한 2017년을 기준으로 데이터를 수집하였다. 이러한 데이터의 활용은 사회보장 수급 현황 분석에 있어서는 정확한 분석이 가능하나, 소득·자산 등에서는 자료의 종합적 결합을 이루지 못하여 하위 소득·자산을 파악하지 못하는 한계를 보인다.

정찬미(2021)는 성별, 연령, 지역, 소득분위 등 다양한 상황에 따라 이질적 특성이 발생하는 1인 가구의 사회보장 수급 현황을 분석하기 위해, 우선 1인 가구의 분표를 성별, 연령, 지역 등의 인구학적 및 경제학적 특성에 따라 살펴보고, 사회보장 수급 현황에 대해서는 주로 연령에 따른 인구집단과 소득분위로 구분하여 분석을 진행하였다. 인구집단은 18세 이하의 아동, 19세 이상 39세 이하의 청년, 40세 이상 64세 이하 중장년, 65세 이상을 노인으로 구분하여 연령 기준을 정의하였다. 아동복지법상 15세 미만의 아동은 1인 가구를 구성할 수 없으나 이 연구에 사용된 자료는 주민등록 전산자료의 세대를 기준으로 형성된 것으로, 아동세대는 주민등록 서류상으로만 분리되어 있는 것으로 추측된다. 지역은 특·광역시 및 특별자치도는 대도시, 시 단위는 중소도시, 군·읍·면을 농어촌으로 구분하였다. 가구균등화지수를 적용한 시장소득을 10분위로 나누어 소득분위 분석을 진행하였는데, 소득신고 대상자의 자료만 확보할 수 있어 과소신고, 무신고 등의 사유로 소득이 없는 비율이 10% 이상이었다. 이로 인해 소득 1, 2분위가 구분되지 않는 한계가 있었다.

정찬미(2021)는 이상의 분석을 통해 1인 가구 내 인구집단별 이질성이

두드러짐을 확인하였다. 청년은 소득수준 및 사회보험 가입률은 높으나 자산 수준이 낮으며, 근로하는 노인은 많으나 양질의 근로를 하지 못하는 것으로 나타났다. 이러한 특성들을 기반으로 청년의 자산을 형성할 수 있는 프로그램 지원과 중장년층의 근로연계복지 확대 및 사회보험료 지원, 그리고 노인을 위한 양질의 일자리 제공 및 공공부조의 급여 충분성 제고 등의 정책 필요성을 강조하였다.

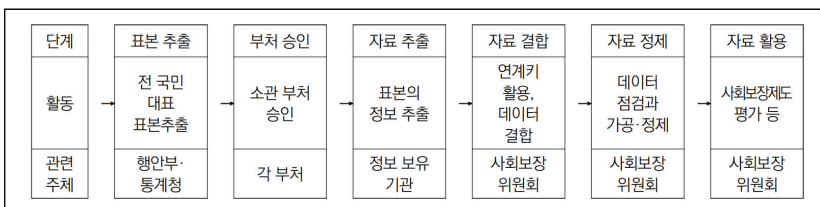
나. 이현주(2022). 「사회보장제도 근거 강화와 사회보장 행정데이터 구축: 사회보장 행정데이터 연계·활용을 위한 과제」

이현주(2022)는 최근 사회보장제도의 정책 근거를 강화하는 중요한 기반으로 기대되고 있는 사회보장 행정데이터의 발전을 위해, 이의 연계와 활용을 위한 과제를 도출하고자 하였다. 이를 위해 2010~2019년 이전까지 수행된 주요 정책연구를 정리한 뒤 정책에 대한 분석을 주로 설문조사 자료에 의존하였던 과거 연구들의 한계를 지적하였다. 설문조사 자료의 경우, 사회보장제도 중 극히 일부분의 정보만을 포괄하고 있다는 가장 큰 문제점과 여러 정책의 수급 여부와 급여 수준을 응답자가 정확하게 답하기 쉽지 않고 이로 인해 정확성이 낮다는 문제점을 안고 있다. 이러한 설문조사 자료의 한계를 극복하기 위해 행정데이터의 사용, 특히 여러 행정데이터를 결합한 통합 행정데이터는 정보의 포괄성과 정확성이 높아지고, 다양한 자료 형태로의 변형 가능성 등 많은 장점을 가지고 있어 이에 대한 활용이 필요하다. 통합 행정데이터를 활용하여 사회보장제도를 분석하면 개별 지원의 효과 분석을 넘어 전체 현금 지원의 소득계층별, 재산 분위별, 지역별 효과를 정치하게 분석할 수 있을 뿐 아니라 서비스 지원의 효과도 유사 목적 서비스의 전반적 효과를 상세한 수준으로 파악할

수 있게 된다.

이현주(2022)는 사회보장 행정데이터 구축 과정을 다음과 같이 소개하고 있다. 표본을 추출하고 필요한 정보를 담당하는 소관 부처의 승인을 거쳐, 정보 보유 기관으로부터 데이터를 추출한 뒤, 추출한 데이터를 가명처리된 결합키를 활용하여 연계하는 과정을 거친다. 연계 과정으로 마련된 통합 행정데이터에 대하여 다시 비식별 조치를 한 후 안정성을 점검하고, 추가로 필요한 가공·정제를 한 후 활용이 가능하도록 준비한다. [그림 3-1]의 각 단계에서 사회보장위원회와 관련 기관의 협력이 필요하다. 자료를 추출하여 제공하는 기관에서는 정보 협조에 대한 심의와 결정, 정보 추출과 비식별 조치 등을 수행하는 역할을 맡는다. 이 외의 정보 보유 기관은 지속적으로 정보의 질 향상을 위해 노력하는 역할을 수행한다. 한편 행안부 또는 통계청과 같이 전 국민의 정보를 보유한 기관은 표본 추출과 표본의 비식별 연계기 마련 등과 관련하여 중요한 역할을 수행한다. 사회보장위원회는 협력을 위한 협상, 통합 행정데이터에 대한 추가적 비식별 조치, 데이터의 점검과 정제 등의 과정을 수행한다. 또한 해당 데이터의 활용을 위한 공간 구성 등 활용을 지원하는 역할을 포함하여 전 과정에서 핵심 역할을 수행하게 된다.

[그림 3-1] 사회보장 행정데이터 구축 과정



자료: 이현주, (2022). 사회보장제도 근거 강화와 사회보장 행정데이터 구축: 사회보장 행정데이터 연계·활용을 위한 과제. p.14.

이현주(2022)는 이러한 사회보장 행정데이터의 범위를 다음과 같이 설명하고 있다. 사회보장 행정데이터가 포괄하는 정보의 범위는 법에 명시되어 있는데, 사회보장 행정데이터의 정보 범위를 정하는 가장 기본적인 기준은 사회보장제도에 대한 법적 정의라 할 수 있다. 「사회보장기본법」 제3조 정의에는 ①항에 “사회보장”에 대한 정의와 ④항에 “사회서비스”에 대한 정의가 명시되어 있다.²⁵⁾ 관련된 법적 개념 정의에 따르면, 사회보장 행정데이터는 사회보험·공공부조·사회서비스 제도와 관련된 정보를 포괄하며 사회적 위험인 출산, 양육, 실업, 노령, 장애, 질병, 빈곤 및 사망과 관련된 정보를 포괄하게 된다. 「사회보장기본법」 제42조, 「사회보장기본법 시행령」 제20조, 「사회보장기본법 시행령」 제20조 제2항 관련 별표에는 사회보장 행정데이터로 요청할 수 있는 정보의 목록이 적시되어 있다. 정보 목록을 법에 명시하는 것은 법 조항의 수시 변경 등으로 인한 소모적인 측면이 있음에도 불구하고 정보 추출의 안정성을 도모하려는 것으로 이해된다.

이제 데이터 연계와 정제에 대해 살펴보면, 추출한 정보들은 결합기를 활용하여 연계되고 이러한 과정을 거쳐 통합 행정데이터를 구축하게 된다(이현주, 2022), 결합키는 생년월일, 성, 이름을 가명처리하여 생성하는데, 법적 규정 등으로 주민번호를 가명처리하여 활용할 수 없기 때문에 대안으로 활용하는 결합키라 할 수 있다(이현주, 2022). 생년월일, 성, 이름이 같은 중복키가 적지 않은데, 이현주 외(2020)의 연구에 따르면 표본

25) 사회보장기본법 제3조(정의) ① “사회보장”이란 출산, 양육, 실업, 노령, 장애, 질병, 빈곤 및 사망 등의 사회적 위험으로부터 모든 국민을 보호하고 국민 삶의 질을 향상시키는데 필요한 소득·서비스를 보장하는 사회보험, 공공부조, 사회서비스를 말한다. ④ “사회서비스”란 국가·지방자치단체 및 민간부문의 도움이 필요한 모든 국민에게 복지, 보건 의료, 교육, 고용, 주거, 문화, 환경 등의 분야에서 인간다운 생활을 보장하고 상담, 재활, 돌봄, 정보의 제공, 관련 시설의 이용, 역량 개발, 사회참여 지원 등을 통하여 국민의 삶의 질이 향상되도록 지원하는 제도를 말한다.

중 중복 개인이 개인 중 12.6%, 중복 개인이 속한 가구는 가구 중 25.7%, 중복 개인이 속한 가구에 속한 가구원 전체는 개인의 33.1%에 이르는 것으로 나타났다. 이는 결과적으로 개인의 30% 이상의 표본이 손실되는 것을 말한다. 결합이 완료된 통합 행정데이터는 점검, 가공과 정제 과정을 거쳐 자료의 정확성을 높이고 비식별화한다. 통합 행정데이터의 활용 전 안전성에 대한 검토를 거쳐 완료되면 사회보장위원회 분석센터에 데이터를 활용할 수 있도록 구비한다.

이현주(2022)는 지금까지 논의한 사회보장 행정데이터의 효율적인 연계·활용을 위해, 요청 정보의 범위에 대한 합리적 판단 기제 마련, 데이터의 안정적 확보를 위한 협치 강화, 여러 정보를 연계하기 위한 연계기의 정확성 제고, 통합 데이터를 활용한 새로운 연구 주제 개발과 분석 결과를 반영할 수 있는 기제 강화 등을 과제로 제시하였다. 협치 강화 과제를 자세히 살펴보면, 우리나라는 공급자 중심의 행정데이터 플랫폼 특성으로 인해 정책 분석에 많은 한계를 보이고 있는데, 이러한 현황을 해결하기 위해서는 통합 행정데이터 구축이 필요하고, 이러한 구축이 가능하도록 협치 강화가 전제되어야 된다는 내용이다. 연계기의 경우, 현재 활용 중인 생년월일, 성, 이름을 가명처리한 결합키로는 정확한 매칭이 가능하지 않은데, 이와 관련하여 주민번호 중 여덟 번째 자리의 가명처리와 이를 활용한 결합키의 보완 등의 대안이 고려될 수 있다. 그러나 이러한 방식으로 중복키가 완전하게 해소되는 것은 아니어서, 이에 다른 대안으로 정부가 주체인 경우에 한하여 주민번호를 가명처리하여 데이터를 연계하고 그 통합데이터를 정책 평가에 활용하는 방안 등이 필요함을 강조하였다.

제2절 해외 사례

1. 영국 가족자원조사(Family Resource Survey, FRS)²⁶⁾

가. 개요

1) 조사 배경 및 대상

가족자원조사(Family Resources Survey, FRS)는 현재 영국 국민들의 생활 수준과 상황을 파악하기 위하여, 영국 내 개인 가구 대표 표본을 통해 개인들의 소득과 그 상황에 관한 정보를 매년 수집하는 조사이다. FRS는 1인 가구와 자녀가 있는 가구, 그리고 노인 및 돌봄 책임이 있는 사람들 등 전 국민을 대상으로 생활 수준과 현황을 추적하고, 이를 통해 영국 국민들을 위한 정책을 개발하는 데 중요한 역할을 담당하고 있다. FRS는 1992년 10월 영국 본토(잉글랜드, 웨일스, 스코틀랜드)에서 시작되었으며, 2002~2003년 조사에 북아일랜드로 확대되었다.

2) 조사 목표

FRS의 주요 목표는 영국 노동연금부(Department for Work and Pension, DWP)에 사회복지 정책의 개발과 모니터링 및 평가를 위한 정보를 제공하는 것이다. 자료는 수급자의 소득(수급, 세액공제, 연금 등) 및 생활 현황, 주택 소유 형태, 돌봄의 필요성과 책임, 장애 여부, 주택 비용,

26) Department for Work & Pensions. Family Resources Survey: background information and methodology(updated 21 July 2023), UK의 내용을 참고하여 작성함
<https://www.gov.uk/government/statistics/family-resources-survey-financial-year-2021-to-2022/family-resources-survey-background-information-and-methodology#linking-frs-data-to-administrative-data>, retrieved from 2023.10.02.

교육, 육아, 가족 사항, 자녀 양육비, 가구의 식품 보장 및 푸드뱅크 이용 등에 관한 세부적인 정보를 수집한다. DWP의 데이터 활용에 있어서 마이크로시뮬레이션은 중요한 역할을 한다. 따라서 정확한 정보의 수집 이후에는 데이터 처리, 편집 및 품질 보증에 세심한 주의가 기울여진다.

FRS 데이터는 통계규제국(Office for Statistics Regulation, OSR)에 의해 국가 통계로 지정되며, DWP의 여러 국가 통계 발표에 데이터를 제공한다. 제공되는 데이터에는 ‘평균소득 이하 가구’, ‘노인 소득 시리즈’ 등이 포함된다. 또한 ‘소득 관련 혜택에 대한 공식 통계: 수급 추정치’ 및 ‘저소득 가구의 자녀: 지역 통계’에 기초를 제공한다. 이러한 FRS는 특별히 재무성과 국세청, 그리고 관세청의 세금 및 혜택에 관한 정책 분석에 활용되고 있으며, 학계 및 연구 기관에서도 사회 연구를 위해 광범위하게 활용되고 있다.

3) FRS 데이터와 행정데이터의 결합

FRS 데이터와 행정데이터의 결합에 대해 살펴보면, DWP 디지털 데이터 전략에 따라 해당 부서는 행정데이터를 연결 가능한 다양한 소스(예: 정부의 다른 부분)에 연결하여 FRS를 변형하는 일을 시행하고 있다. 또한 국가 통계로서 통계 관행 코드(Value V4.1)와 일치하며 DWP는 매년 FRS 조사 개선에 노력을 기울이고 있다.

2018년 EU의 General Data Protection Regulation(GDPR)이 도입되어 데이터 결합을 법적으로 동의하는 대안이 제공되었다. 이로써 공공기관의 기능 수행을 위해 데이터 결합이 필요하다는 법적 근거를 바탕으로 모든 응답자를 연결할 수 있게 되었다(GDPR 제6(1)(e)조). 이러한 법의 적용과 함께 연결 방법론의 개선으로 최소한 95% 이상의 FRS 응답자를 그들의 행정 기록에 연결할 수 있게 되었다. 이러한 행정데이터와

FRS 데이터의 통합으로, 다양한 금융에의 활용, 데이터 품질 제고, 응답자 및 사용자 이점을 실현할 수 있는 가능성이 열렸다.

FRS 데이터와 행정데이터 연결 프로젝트는 모든 행정데이터 소스의 잠재력을 조사하고 이점을 실현하기 위해 출발하였다. 이는 DWP 디지털 데이터 전략과 관련이 있으며 행정데이터의 최적 활용과 재사용을 위한 것이다. 또한 영국 통계 권한의 데이터 연결 전략과 관련이 있으며, 국가 통계 규제청의 데이터 수집 변환 전략과 관련이 있다.

나. 가족자원조사(Family Resource Survey, FRS) 활용 사례

1) Serena Pattaro, Nick Bailey, and Chris Dibben(2020). “Using Linked Longitudinal Administrative Data to Identify Social Disadvantage”

Pattaro et al.(2020)은 행정데이터(Administrative Data)를 사용하여 사회 불평등을 식별하고자 하였다. 행정데이터는 정책적인 목적으로 무료 학교 급식(Free School Meals) 자격이나 다중 결핍 지수(Indice of Multiple Deprivation) 등 사회적 불이익 지표(Indicators of Social Disadvantage)를 구성하기 위해 널리 사용된다. 이러한 지표들은 횡단면 자료에 의존하기 때문에, 연구자들은 이러한 지표들을 정확성과 간편성 사이의 절충안으로 사용한다. 장기적인 행정데이터(longitudinal administrative data)들의 증가는 연구에 필요한 더 정확한 지표들을 구성하는 데 도움을 줄 수 있다. 이러한 잠재적 가능성을 설명하기 위해, 노동연금부(DWP)의 National Benefit Database의 복지 혜택(Welfare Benefit) 및 영국 국세청(Her Majesty's Revenue and Customs, HMRC)의 P14/P60 데이터에서 연간 수입에 관한 행정데이터를 이용하

여, 5년 동안의 개인 노동 시장 이력을 재구성하였다. 이러한 행정데이터들은 Poverty and Social Science UK 2021의 설문조사 데이터와 연결되어 있다. 기술통계(descriptive data)와 로지스틱 회귀분석(logistic regression) 결과, 장기적인 측정값은 동일한 주제에 관한 설문조사 응답과 높은 상관관계가 있으며, 횡단면 데이터를 이용한 측정값보다 빈곤 위험을 예측하는 강력한 지표임을 보여주었다. Pattaro et al.(2020)의 연구 결과는 장기적인 행정데이터 측정값이 정책 및 빈곤 연구뿐만 아니라 다양한 응용 분야에서 잠재적으로 폭넓게 활용될 수 있음을 보였다.

영국에서는 복지 혜택 및 세금 체제에 관한 행정데이터를 다양한 정책 분야에서 사용하기 위해 사회 불이익(social disadvantage) 지표를 구성에 사용한다. 이 지표들의 쉬운 이용성과 정책상의 중요성은 사회 불평등에 관심 있는 연구자들에게 매우 매력적이다. 그러나 이 지표들은 정책과 실무적인 측면에서 설계되었으므로, 종종 정확성보다는 간편성과 실행의 용이성이 우선시된다. 이 지표들은 주로 단일 자료에서 비롯된 횡단면 자료 지표에 의존한다. 대표적인 예로는 무료 학교급식 자격이나 소득 및 고용을 포함한 다양한 자료에서 구성된 다중 결핍 지수 등이 있다.

연구 목적으로 사용 가능한 행정데이터의 가용성이 증가함에 따라 연구자들은 다년간의 여러 소스의 데이터를 연결하고, 정책 결정자가 직면한 운영 제약에서 벗어나 더 정교한 사회 불이익 측정값을 설계하고 적용할 수 있게 되었으며, 단순한 횡단면 데이터에만 의존하는 대신 개인의 장기적인 정보를 활용하고, 다른 여러 소스의 데이터를 결합하여 좀 더 정확하게 사회 불이익을 식별할 수 있는 기회가 생겼다. Pattaro et al.(2020)의 연구 목표는 연결된 장기적인 행정데이터를 사용하여 사회 불이익을 식별하는 잠재력을 탐구하는 것이었는데, 복지 혜택 및 수입에 관한 장기적인 수급 및 수입의 측정값이 현재의 빈곤 위험을 어느 정도

예측하는지 비교하고, 동일한 주제에 대한 설문조사 측정값 및 행정데이터의 횡단면 측정값과 비교하였다.

Pattaro et al.(2020)의 연구에서 데이터는 소득과 빈곤에 관련된 두 가지의 국가적인 설문조사에 국가 행정 시스템의 데이터를 연결한 독특한 조합에 기반하고 있다. 영국의 빈곤과 사회 배제(Poverty and Social Exclusion in the UK, PSE-UK) 조사는 2012년에 실시된 횡단면 가구 조사이다. 이 조사는 가족자원조사(Family Resource Survey, FRS) 2010~2011년 조사에 참가한 응답자 샘플로 구성되어 있으며, 재연락에 동의한 응답자들을 포함하고 있다. Department for Work and Pensions(2018)에 따르면, FRS는 소득 및 빈곤에 관한 공식 통계를 도출하기 위해 사용되는 연례적인 횡단면 조사 자료이다(Pattaro et al., 2020에서 재인용). Bailey and Bramley(2018)에 따르면, PSE-UK 조사는 FRS에서 얻은 자료를 보완하여 특히 빈곤과 박탈 지표의 더 다양한 측정값을 포함하여 진행되었다(Pattaro et al., 2020에서 재인용). 영국의 FRS 응답자들은 재연락에 대한 허가 외에도 조사 응답과 관련하여 다양한 행정데이터를 연결할 것에 대한 동의를 요청받았다. 연결 허가가 북아일랜드에서는 요청되지 않기 때문에 이 지역은 여기서 제외되었다.

영국 내에서 FRS는 우편 번호 섹터를 주요 추출 단위로 삼은 계층적 클러스터 분포(clustered distribution) 표본 설계를 기반으로 하였으며, 추출 단위 내에서 무작위로 선택된 가구로 구성되어 있다. 조사는 16세 이상의 모든 거주자와의 면담식 인터뷰를 포함하여, 컴퓨터 지원 개인 면담 방식(computer-assisted personal interviewing, CAPI)을 사용한다. PSE-UK는 재연락 동의를 한 FRS 응답자들을 무작위로 샘플링하여 이전 조사의 복잡한 설계를 유지하였다. FRS는 영국 내 약 25,000개의 가구를 대상으로 하였으며, 응답률은 59%였다. PSE-UK 인터뷰는 영국

내 4,205개의 가구에서 실시되었으며(응답률 63%), 이러한 가구 내에서는 9,786명의 개별 응답자와 전체 또는 부분적인 면담을 완료하였다.

FRS와 PSE-UK 데이터는 DWP의 노동 및 연금 패널 연구(Work and Pension Longitudinal Study, WPLS)로부터의 행정데이터와 연결되어 있다. 이는 DWP에서 보유한 복지 혜택 기록과 HMRC의 세금 기록에서 나온 수입 및 고용 정보를 결합한 것이다. DWP와 HMRC의 데이터는 일반적으로 국민보험번호(National Insurance Number) 같은 고유 인식자를 이용하여 연결될 수 있다. DWP는 이름, 성별, 출생일 및 거주 주소 같은 식별 변수 집합을 기반으로 한 확률적인 연결 방법을 사용하여 행정 데이터를 설문조사 데이터에 연결하였다. 연구팀에게는 익명화된 연결 데이터만을 제공하며, 현재 작업의 목적상 연결 오차가 작고 모델 추정치에 중요한 영향을 미치지 않는다고 가정을 한다.

Pattaro et al.(2020) 연구에서는 두 개의 결합된 WPLS 데이터를 사용하였다. 첫 번째로는 DWP의 National Benefits Database(NBD)가 있으며, 이는 무직 수당인 Jobseeker's Allowance(JSA)와 소득 지원인 Income Support(IS)의 수령 데이터를 제공한다. JSA는 실직 상태이며 적극적으로 일자리를 찾고 있는 다양한 연령대의 성인들에게 지급된다. 일부는 최근의 고용 기록을 기반으로 자격을 얻지만, 다른 사람들은 소득이 낮다는 기준으로 자격을 얻는다. 추가 자격 기준에는 저금액 저축 및 배우자의 고용상태가 포함된다. IS는 주로 혼자 사는 부모를 포함한 소득이 낮은 다양한 연령대의 성인들에게 지급된다. 두 번째 WPLS 소스는 연간 소득에 관한 HMRC의 데이터로, P14/P60 세금 신고 양식에서 파생되었다. 이 양식들은 모든 고용주가 각 세금 연도의 끝에 발급하여 그해 동안 해당 고용에서 과세 소득을 상세히 설명하고 있다. HMRC의 소득 및 고용 기록은 직원만을 대상으로 하며, 자영업자는 포함되어 있지 않다.

Pattaro et al.(2020)은 2012~2013년 세금 연도를 포함하여 약 10년 동안의 행정 기록을 연결하였으며, 이 기간 동안 PSE-UK 조사가 실시된 연도까지를 다룬다. 이 연구에서는 조사 검증을 목적으로, 회고적 질문이 포함된 참고 기간인 조사 직전 5년에 중점을 두었다. 이는 현재 빈곤 위험에 가장 영향을 미칠 가능성이 높은 소득과 실직이 발생할 가능성이 있는 기간이다.

Pattaro et al.(2020)의 연구에서는 사회적 불이익 범주로 빈곤 상태를 살펴보고, 행정 및 설문 조사 데이터에서 유도된 노동시장 참여의 장기적인 측정값이 현재의 빈곤 위험을 얼마나 잘 예측하는지 분석하였다. Pattaro et al.(2020)의 연구에서는 일반 선형 모델(분산 분석 및 로지스틱 회귀 모형)을 사용하여 모형 적합을 실시하였으며 성과 변수로는 PSE-UK로부터 도출된 저소득 및 빈곤에 대한 정보를 결합한 빈곤 척도를 사용하였다. 또한, Pattaro et al.(2020)의 연구에서는 ‘가난한’ 성인을 다음과 같이 정의하였다. 주택 비용을 공제한 가구 소득이 주당 £295 미만이고 Bailey and Bramley(2018)와 Gordon(2017)의 합의된 접근 방식에 따라 식별된 박탈 척도에서 3개 이상의 항목이 부족한 경우이다.

PSE-UK 설문조사에서는 설문조사 일자와 자가 보고(self-reported)된 수개월 기간을 포착할 수 있는 회상 질문을 통해 지난 5년간의 고용 상태 및 실직 여부에 대한 정보를 가지고 있다. 행정데이터에는 조사 날짜의 JSA 및/또는 IS 수령과 PSE-UK 조사 날짜 이전 5년간의 연도 합계에 따른 실직 기간 정보가 있다. 또한 조사 이전 5년간의 고용에서 유사한 장기적 측정값은 P14/P60 양식의 HMRC 세금 기록에서 재구성되었다. 특정 세금 연도에 대한 총 연간 순수입은 모든 고용에서의 수입을 합산하여 계산되었다. 고용기간 일자 중 빈도가 높은 관측치가 누락된 경우, 수입은 각 세금 연도별로 동일한 분포를 가정하여 달력 연도에 할당되었다.

제어 변수는 대부분 PSE-UK에서 가져왔으며, 나이, 성별, 가구 구성, 부모만 사는 가구 여부, 인종, 거주 국가, 최고 교육 수준 및 응답자가 장기 간 질병 또는 장애가 있는지 여부를 포함한다.

Pattaro et al.(2020)은 지난 5년 동안의 실직 기간을 나타내는 PSE-UK의 회상 조사에서 얻은 측정값과 행정데이터에서 구성된 동등한 측정값을 비교하였는데, 이 두 가지 측정값 간의 일치율은 해당 기간 동안 어떤 실직도 경험하지 않은 사람들이 89.3%로 가장 높다. 두 번째로 높은 일치율은 12개월 이상 실직한 사람들에게 관찰되지만 일치율은 46.8%에 불과하다. 중간 범주에 대한 일치율은 여전히 낮은 것으로 나타났다. 일치율이 낮아지는 주요 이유는 조사에서 물어볼 때 어떤 실직에 대한 보고를 하지 않았는데도 기록된 실직의 비율이 높기 때문이다.

두 측정값 간의 불일치의 주된 이유는 고용되지 않은 많은 사람들이 자신을 실직자로 보지 않고 경제적으로 비활동적인 것으로 여기기 때문인 것으로 보인다. 조사 당시, 그들은 JSA 및/또는 IS를 12개월 이상 수령하고 있음에도 불구하고, 자신을 병/장애가 있다거나(44.2%) 가족 돌봄(house care)을 하는 사람(18.2%)으로 보고하고 있다. 만약 이러한 범주들을 성별로 나누어 본다면, 여성들은 주로 가정에서 가족을 돌보는 경향이 있는 것으로, 남성들은 주로 병/장애가 있음으로 나타날 것이다. 이는 영국의 두 종단 자료원을 비교하는 Dex and McCulloch(1998)의 연구결과와 일치한다(Pattaro et al., 2020에서 재인용). JSA 및/또는 IS를 12개월 이상 수령한 사람들 중 매우 적은 수의 사람들이 자신을 고용자(11.5%)나 자영업자(1.2%)라고 보고했다. 이는 행정데이터 측정값이 고용되지 않은 사람들을 여전히 포착하고 있으며, 따라서 자신을 ‘실직자’로 보지 않은 사람들도 실직 위험이 큰 사람들을 포함하고 있다는 것을 보여준다. 이는 복지 혜택 데이터에서 비고용 상태를 식별하는데 어려움

이 존재한다는 것을 보여주고 있다.

HMRC 데이터에 기록된 연간 총수입과 FRS 설문조사에서 보고된 총 수입을 이용하여 비슷한 비교가 가능하다. FRS 설문조사 자료는 월별로 수집되기 때문에 연간 측정값으로 변환하였다. HMRX 데이터의 측정값은 연간이고, FRS 설문조사에서 보고된 측정값은 월별 데이터를 연간화한 데이터임에도 불구하고 이 두 측정값은 강력한 상관관계가 있다($r=0.84$). 그러나 이상치 몇 개가 관찰되는데, 응답자가 일부 수입을 보고한 경우 10건 중 4건에서 행정데이터가 기록되지 않았다. 또한 FRS에서 수입을 보고하지 않은 응답자의 경우 기록된 수입이 £100,000까지의 범위가 관측되었다. FRS에서 보고된 수입이 0인 경우, 누락된 값이 10건 중 5건 이상을 차지하며, 이는 FRS 값에 대한 대체 또는 편집이 수행된 것으로 추정된다. 이는 McKay(2012)에서 보고한 결과와 일치하며, 이는 행정 세무 데이터 대비 FRS의 과소 보고를 나타낸다고 볼 수 있다.

2) Stephen P. Jenkins, Fernando Rios-Avila(2021), “Reconciling Reports: Modelling Employment Earnings and Measurement Errors Using Linked Survey and Administrative Data”

노동 소득에 관한 가구 조사 데이터의 측정 오차에 대한 연구는 연계된 행정데이터가 오차가 없는 기준 수입 측정값을 제공한다고 가정하는 경우가 많다. 그러나 최근의 연구에서는 행정데이터에서의 측정 오차도 고려하는 연구들이 발표되고 있다. Jenkins and Rios-Avila(2021)의 논문은 이러한 두 번째 세대의 연구에 속한다.

Jenkins and Rios-Avila(2021)의 연구는 미국의 연구 결과가 주를 이루는 분야에서 영국의 측정 오차(measurement errors) 및 고용소득(employment earnings)에 대한 새로운 증거를 제시하였다. 저자들은

설문조사 데이터와 행정데이터(administrative data)가 결합된 데이터에서 다양한 유형의 측정 오차를 고려하고 관련된 특성에 따라 오차 분차를 다르게 한 계량모형을 개발·적용하였다. 일반적으로 적용되는 Kapteyn and Ypma(2007)의 계량경제학 모델을 발전시킨 새로운 계량경제학 모델이다. 이 새로운 계량경제학 모델은 데이터 품질을 개선하고 수입 불평등 수준을 평가하는 데 관련이 있으며, 행정데이터의 평균회귀(mean-reverting) 측정 오차를 포함할 뿐만 아니라 연결 불일치(linkage mismatch)의 평균회귀 측정 오차 및 ‘참조 기간(reference period)’ 오차를 포함하며, 오차의 분포가 개인마다 변화할 수 있도록 허용하였다.

연간화된 소득 조사는 평균적으로 실제 연간 소득을 과소 평가한다. 또한 설문조사 측정 오차에서 평균 회귀는 나타나지 않는다. 따라서 이 두 가지 소득 원천은 실제 연간 소득을 과소 평가한다. 설문조사를 통한 소득 측정은 행정데이터를 통한 소득보다 신뢰성이 높지만, 설문조사 데이터와 행정데이터를 결합한 데이터를 기반으로 혼합 소득 예측은 각각의 소득별 측정보다 훨씬 더 신뢰성이 있다. 이러한 측정 오차의 분포가 다른 모델은 어떻게 데이터의 품질을 개선할 수 있는지를 보여준다. 예를 들어, Jenkins and Rios-Avila(2021)의 분석 결과, 설문조사 품질의 경우 설문 응답자가 설문 면접관에게 급여명세표를 제시함으로써, 조사의 오차 분산이 더 작아진다는 것을 강조하였다. 행정데이터의 경우 비표준(non-standard) 직업, 민간 부문의 직업 및 좋은 급여체계가 없는 고용주와 관련하여 더 큰 오차 분산과 관계가 있다는 것을 시사한다.

Jenkins and Rios-Avila(2021)의 연구는 FRS-P14 결합데이터를 이용하였다. DWP 통계 분석가들은 P14 기록을 2011~2012년 FRS 데이터와 결합에 동의한 응답자들을 추출하였다. 통계 분석가들은 성명, 우편

번호, 성별 및 생년월일 정보를 이용하여 작성된 키를 사용하여 P14 데이터에 연결하였다. DWP는 저자들에게 두 개의 데이터를 제공하였다, 하나는 FRS 개인 식별자와 P14 수입 데이터를 포함하고 있었고, 두 번째 데이터에는 FRS의 안전화된 데이터(secured-data) 버전이다. 저자들은 FRS 개인 식별자를 일치키로 사용하여 P14 및 안전화된 데이터 병합하였다. 2011~2012년의 안전화된 데이터에는 적어도 하나의 직장을 가진 13,851명의 고용 인원이 있었으며, 그중 9,014명(65%)이 데이터를 연결하는 데 동의하였다. 데이터 연결에 동의한 응답자 중 6,432명(71%)의 FRS와 P14 기록을 성공적으로 결합하였다. 데이터 결합 후, FRS에서 '자영업자'로 자신을 신고한 소수의 관측치(23개)와 FRS 또는 P14 수입이 0인 관측치를 제외하여 6,391명(남성 2,794, 여성 3,599)의 응답자를 얻었다. 마지막으로, 총수입이나 작업의 참조 기간에 대한 추정 또는 편집된 값이 있는 관측치 420개를 삭제하여 추정 샘플로 총 5,971명(남성 2,595, 여성 3,376)의 개인 데이터를 최종적으로 얻게 되었다. 이 샘플의 응답자 연령 범위는 16세에서 84세이며, 대부분은 25세에서 59세 사이(84%)이다. 16세에서 24세는 6%, 60세 이상은 10%로 나타났다.

Jenkins and Rios-Avila(2021)의 연구 결과, FRS의 수입의 90% 이상이 측정 오차의 영향을 받는 것으로 추정되며, 측정 오차의 확률은 약 1/3 정도로 나타났다. 그러나 P14 오차의 분산이 FRS의 분산보다 더 큰 것으로 나타났다. 결합된 P14 데이터의 신뢰성은 불일치된 기록의 작은 부분 때문에 더욱 저하되는 것으로 나타났다. FRS의 경우에서 측정 오차가 유일한 문제가 아닌 것으로 분석되었다. 약 1/10 정도의 참조 기간 오차 확률이 존재하며, 이때 FRS의 연간화된 현재 수입의 측정치는 평균적으로 연간 수입을 과소평가하며 잡음(noise)이 많은 것으로 나타났다. 전반적으로 FRS 수입은 P14의 수입보다 더 신뢰성이 있다. 오차의 조합으

로 인해 FRS 데이터든 P14 데이터든 관측된 수입은 참수입의 불평등을 상당히 과대 평가한다. 측정 오차에서 발생하는 더 큰 잡음을 상쇄하기에는 평균 회귀가 충분하지 않다.

공변량을 고려한 모델에서는 데이터 품질이 떨어지는 데 영향을 미치는 요인들에 대하여 강조하였다. 예를 들어, 조사의 측정 오차 분산이 급료명세서를 제시하지 않는 근로자와 더 나이 많은 근로자에서 큰 것으로 나타났다. 한편, 노동 경제학자들을 위한 긍정적인 결과가 있다. ‘표준’ 연간 근무 상근직(full-time job)에 종사하는 수입자로 표본을 제한하면, 측정 오차의 분산이 감소하고, 연간 수입 측정값이 연간 수입의 편향이 적은 측정값이 된다는 것을 밝혀냈다. 반면, 이러한 표본 선택은 불평등과 빈곤에 관한 국가 통계의 기반이 되는 가계 소득 측정을 위해 적절하지는 않다. 이 연구 결과는 데이터 품질을 개선하기 위한 노력의 대상 영역을 가리킨다. 예를 들어, 조사 인터뷰에서 급료명세서 및 기타 기록의 사용을 증가시키고, 특히 비상근직(part-time job) 및 덜 안정적인 직업에 종사하는 근로자들을 위해 고용주의 급료 체계를 개선하는 것이 좋은 방법일 수 있다. 다른 긍정적인 점은 조사 및 연결된 행정데이터를 기반으로 하는 혼합 수입 측정값을 도출하는 것이 가능하며, 개별로 사용되는 것보다 훨씬 더 신뢰성이 있다는 것이다. 또한 이러한 혼합 측정값은 데이터 제공자가 연구자들에게 기밀성 있는 행정데이터를 공개하지 않고도 높은 품질의 데이터를 제공할 수 있는 방법을 제공한다.

2. 포르투갈 사례: 국립 역학 감시 정보 시스템(SINAVE)과 국립 전자 사망 증명서 정보 시스템(SICO)의 결합

가. 시스템 개요²⁷⁾

1) 국립 역학 감시 정보 시스템(SINAVE)

포르투갈에서는 1949년 8월 9일 전염병에 대한 예방, 치료 및 교육 조치에 관한 내용을 포함하는 법률 제2036호가 발표되면서 전염병 모니터링을 위한 기초가 마련되었다. 해당 법률에서는 “전염병 사례를 진단하거나 의심하는 모든 의사는 48시간 이내에 보건 담당자에게 신고해야 한다”고 규정하고 있다. 이후 2009년 8월 21일 제정된 법률 No. 81/2009에 의해 포르투갈의 의무 신고 관련 전염병 감시 시스템이 재편되었으며, 이후 국립 역학 감시 정보 시스템(Sistema Nacional de Vigilância Epidemiologica, 이하 SINAVE)을 구축하였다. 해당 법률에서는 전산으로 연결되어 역학 감시 질을 향상시킬 수 있는 기회를 제공하는 의사, 의료 기관, 연구소 및 보건 당국을 통합하는 감시 네트워크를 규제하였다.

SINAVE는 질병의 확산과 추가 사례 발생을 억제하면서 예방 및 통제 조치 시행을 위해 의사가 전염병 발생을 지역 보건 당국에 실시간으로 알릴 수 있게끔 하는 전염병 및 기타 공중 보건 위험에 대한 의무 신고를 전산화하였다. 전국 신고자 분포를 살펴보면, 92.3%가 소프트웨어를 통한 의사의 직접 신고인 것으로 나타났으며, 대부분의 신고 사례에서는 1차 의료 기관이 임상 프로세스 지원 소프트웨어와의 자동 상호 운용성 체계

27) 두 시스템에 관한 내용은 단국대·청운대 컨소시엄 특수외국어사업단의 무료 번역서비스를 지원받아 포르투갈 원문 자료(국립 역학 감시 정보시스템(SINAVE) 모니터링 보고서, SICO 사용을 위한 실용 가이드 - 의사/의료 기관, 국립 전자 사망 증명서 정보 시스템(SICO)을 번역하여 정리함.

를 통해 신고하였다.

의무신고대상 전염병을 목록화하여 관리하고 있으며, 역학적 상황이 변화하면 감시 시스템이 실행되어야 할 필요성이 있으므로 그 목록에 신종 전염병을 포함시키고 있다. 2014년 에볼라 바이러스 질병, 2015년 중동 호흡기 증후군 코로나바이러스(MERS-CoV) 감염, 2020년부터는 코로나바이러스 질병(COVID-19)도 의무신고대상 전염병 목록에 포함되어 SINAVE에 관련 정보가 등록되었다.

지역 보건 당국은 감염 가능성 사례, 의심 사례 또는 확진 사례에 대한 신고를 받은 후 역학조사를 시작한다. 이 역학조사는 사례 분류를 확인하고, 통제조치를 위해, 질병 기록에 대한 정보를 수집하는 것을 목적으로 한다.

역학조사는 사례 연구를 포함하며 접촉자/동거인/거주자 파악, 임상병력의 상세한 기록, 사회적 결정요인 및 역학 데이터, 연구소 검사 수행을 위한 생물학적 샘플 또는 필요한 경우 환경 샘플 수집, 추가 사례 조사, 감염원 또는 전염 경로 파악, 잠재적 감염원 또는 매개체 검색, 사례 발생에 기여했을 수 있는 기타 요인 파악을 포함한다. 반복해서 새로운 정보가 확보되면 다음 단계를 결정하며, 조사가 완료되고 통제조치가 효과적이었음이 입증될 때까지 부분적인 위험 평가가 수행된다.

역학조사의 다음 단계를 제시하고 사건을 포괄적으로 특징짓기 위해 다른 기관(의료 기관, 연구소, 학교, 기업 등)과 정보 전달과 교환이 필요하며, 관련 보건 당국에 정보를 제공해야 할 필요가 있을 수도 있다. 특히 공중보건 측면에서 긴급한 상황일 경우 연구소 검사 결과를 기다릴 필요는 없다. 확진 사례가 확인되지 않으면 이후 통제조치는 중단된다. 사용 가능한 모든 정보를 평가하여 사건의 규모와 범위를 결정하고, 필요할 때마다 통제조치를 확대하며, 전염 사슬을 차단하기 위한 더 구체적인 조치

를 시행할 수 있도록 해야 한다.

SINAVE에는 신생물, 당뇨병, 인체 면역 결핍 바이러스(HIV), 신경학적 또는 신경근 질환, 천식, 만성 폐 질환, 간 질환, 만성 혈액 질환, 만성 신장 등 13가지 동반 질환의 발생이 기록되어 있다. 질병, 만성 신경학적 결핍, 급성신부전, 울혈성 심부전 및 응고병증 등의 해당 사례를 통보한 경우 동반질환이 있는지 여부를 나타내는 변수도 제공된다.

2) 국립 전자 사망 증명서 정보 시스템(SICO)

포르투갈에서는 법률에 따라 2014년 1월 1일 0시 이후에 발생한 사망자의 경우 국립 전자 사망 증명서 정보 시스템(Sistema de Informação do Certificados de Obito, 이하 SICO)을 이용한 사망 증명서 발급이 의무화되었고, 이에 따라 시스템이 구축되었다. SICO는 국가 보건 정보 기술 기관인 보건부 공유 서비스(Serviços Partilhados do Ministério da Saúde, 이하 SPMS)와 이에 대한 데이터를 제공하는 포르투갈 보건국(Direcao-Geral da Saude, 이하 DGS)이 관리한다.

SICO를 통해 의사가 증명서를 등록하는 순간부터 국가 통계 업데이트, 그리고 사망이 담당 기관에 통보될 때까지 검찰청과 경찰 당국의 개입(해당되는 경우)을 시작으로 모든 사망을 전자적으로만 인증하도록 하고 있다. SICO에서는 사망증명서 발급 절차에 참여하는 모든 기관 간의 연계가 가능하다. 민사 등록 사무소에 사망증명서(전자 사망증명서, Emissao de Certificado de Obito eletronico, 이하 CO)를 전산으로 통보하고 사망 원인 코드화, 정보 통계 처리 및 사망률 역학 감시의 질적 향상을 도모할 수 있다. 또한, 국민보건서비스(Servico Nacional de Saude, 이하 SNS) 이용자 데이터베이스 및 국가이용자등록(RNU) 범위 내에서 부여된 해당 식별번호 업데이트가 가능하다. 그리고 SICO를 통해

시신운송 안내서(Guia de Transporte)를 출력할 수 있으며 시신운송 안내서는 사망 기록을 작성하고 시신 안치 및 운송의 목적으로 사용된다.

의사가 CO를 발급할 때, 필수 입력 정보, 사망원인 기입 규칙, 일반적인 입력 사항들은 다음과 같이 지침으로 제시되어 있다. CO를 작성할 때 먼저 SICO 지원 신청서를 작성해야 하며, 이 과정에서 자연사와 부자연사(외인사) 또는 원인불명사를 구분하고, 동시에 사망이 발생한 장소도 고려하여 작성해야 한다. 공공 또는 민간 의료 기관 내부에서 사망이 발생한 경우와 의료 기관 외부에서 발생한 경우를 구분하여 작성한다.

외인사 또는 원인 불명의 사망이 공공 또는 민간 의료 기관에서 발생한 경우 의사는 DGS의 웹사이트(www.dgs.pt)를 통해 SICO에 접속하고, ‘임상 및 상황 정보’ 탭에서 임상 및 상황 정보 보고서(Boletim de Informacao Clinica e/ou Circunstancial, 이하 BIC) 양식을 작성해야 한다. 외인사 또는 원인 불명 사망이 발생하면, 해당 사실은 발행된 BIC 번호와 함께 전화 또는 팩스 등을 통해 관할 검찰청에 통보된다. 외인사 혹은 원인 불명의 사망이 의료기관 외부에서 확인되는 경우, 현장에 출동한 의사는 사망을 확인한 후, DGS의 웹사이트를 통해 SICO에 접속하여 소프트웨어가 제공하는 양식을 작성하여 CO를 등록해야 한다. 태아 또는 신생아 사망의 경우, 임신 기간 22주 이상의 태아 또는 생후 28일 이전에 사망한 아이에 대해 사용되는 전자 양식 ‘태아 및 신생아 사망 증명서’를 작성해야 한다. 자발적인 임신 중절의 경우와 임신 중단이 자연적으로 발생한 경우(임신 24주 이내)에는 태아 사망 진단서를 제출하지 않아도 된다.

〈표 3-3〉 임상 및 상황 정보 보고서(BIC) 작성 항목

구분	항목
개인정보	이름, 아버지 이름, 어머니 이름, 성별, 생일, 결혼 여부, 직업, 연락처 유형, 가족 연락처, 출생지, 국가명, 지역명, 지자체명, 지자체의 행정 구역
상황 정보	의료기관명, 의료기관 입원 상황, 진료 정보, 이송 정보, 이송된 의료기관 입소 날짜, 허가날짜, 시신 도착 유무, 사망 장소, 원인, 사망 날짜/시간, 들어올 때의 진료 상황, 입원할 때의 병실, 병실에서 진행한 진료, 관련 추가 검사, 관련 병력, 최종 또는 가능성 있는 진단, 가장 중요한 임상 및/또는 수술 요법 혹은 소생 요법, 의료기관에서 수집한 생물학적 제제, 입원 또는 입소 시 수집된 흔적 및 기타 요소, 중요한 기타 요소
추가 사항	정보 제공 의사 신원, 이름, 자격증 번호, 연락처 유형, 연락처, 서명

자료: 포르투갈 보건국. (2012). Sistema de Informação dos Certificados de Óbito(국립 전자 사망 증명서 정보 시스템(SICO) 시행절차). p. 12. <https://docplayer.com.br/19840316-Sistema-de-informacao-dos-certificados-de-obito.html>, 2023.04.13. 인출

나. Feteira-Santos et al.(2022), “Improving morbidity information in Portugal: Evidence from data linkage of COVID-19 cases surveillance and mortality systems”

Feteira-Santos et al.(2022)에서는 포르투갈에서 2020년 코로나19 관련 이환율을 분석하기 위해 등록 자료 기반 관찰 연구(registry-based observational study)를 수행하였다.

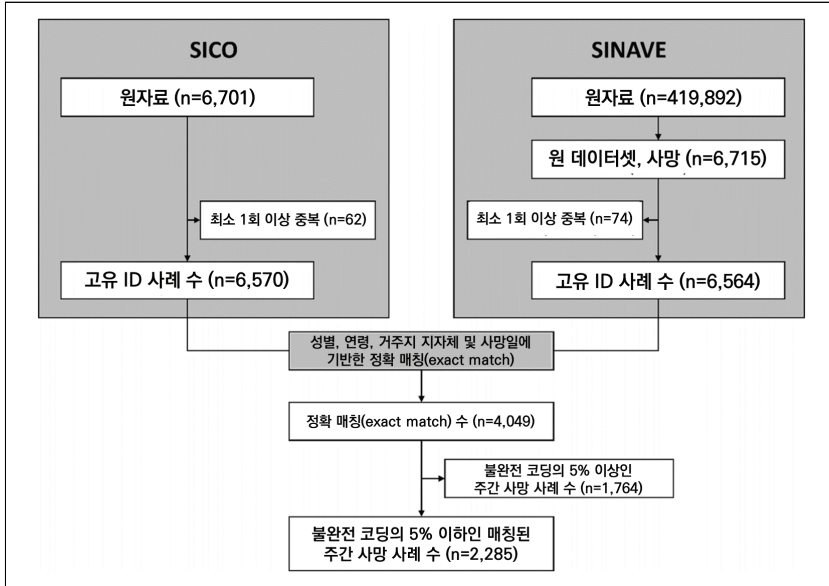
코로나19 모니터링을 위한 가장 간단한 방식으로는 통상의 건강 정보 시스템(Health Information System, 이하 HIS)과 코로나19 관련 보고 플랫폼 또는 기존 감시 시스템을 활용하는 것이다. 그러나 이번 팬데믹으로 인해 코로나19 모니터링을 위한 보건 감시 시스템의 미흡한 점이 드러나게 되었다. 이는 전체 의료 시스템에서 시의적절하고 명확한 대응을 하기 위해서는 조정(adjustments)이 필요했음을 의미한다. 그 예로 포르투갈에서는 더 많은 의료 전문가와 접촉 추적 훈련을 받은 군 인력을 배치하여 인적 자원을 강화하였으며 접촉 추적을 수행하기 위한 구체적인 HIS(예: Trace COVID-19 플랫폼)가 구축되었다.

유사한 건강정보를 독립적으로 수집하기 위해 서로 다른 플랫폼이 구축되면 분리된 데이터가 생성되어 효율적인 활용에 장애가 될 수 있으며, 부분적으로 신속·강력·정확한 대응이 저하될 수 있다. 데이터 수집 및 분석 중에 사용되는 신기술 및 자동화의 역량을 고려하여 HIS를 개선시키고자 여러 전략들을 연구하였다. 그 전략들 중 하나는 서로 다른 HIS의 정보 처리 상호 운용(interoperability)으로, 이를 통해 특히 팬데믹 기간 동안 조직적인 방식으로 소통하고 이용 가능한 건강 정보를 증대시킬 수 있다.

따라서 Feteira-Santos et al.(2022)의 연구에서는 포르투갈의 코로나19 사례 및 관련 사망 모니터링 중에 기록된 이환율에 관한 두 데이터베이스 간의 이용 가능한 정보의 잠재적인 상보성을 비교·식별하는 것을 목표로 하였다. 이를 위해 코로나19 감염 및 관련 사망에 대한 정보를 수집하기 위해 두 개의 데이터베이스(SINAVE와 SICO)로부터 성별, 연령, 거주지 지자체 및 사망일을 기반으로 한 정확 매칭(exact match)을 통해 분석 데이터를 구축하였다. 두 데이터베이스에 있는 모든 기록은 2020년 포르투갈의 코로나19의 모든 사망자에 대한 내용이 포함되어 있다.

DGS에서는 해당 연구에 2020년 3월 16일부터 12월 31일 사이에 포르투갈에서 확인된 모든 코로나19 관련 사망증명서 데이터인 SICO 하위 데이터셋을 제공하였다. 해당 데이터셋에서는 ICD-10(질병 및 관련 건강 문제의 국제 통계 분류 10차 개정판)에 따른 COVID-19 코드가 사망 원인인 경우 코로나19 사망자로 식별하였다. SINAVE를 이용하여 코로나19 사례의 임상적 특성에 대한 정보를 추출하였다. 또한, 2020년에 코로나19 관련 사망 원인으로 사망한 사람들의 SINAVE 기록을 검토하였다. SICO의 경우 전문 인력이 현장에서 수집된 개방형 문항에 대한 정보를 코딩하므로 SICO의 이환 데이터는 SINAVE보다 더 정확하고 신뢰할 수 있는 것으로 간주된다.

[그림 3-2] SINAVE(국립 역학 감시 정보 시스템)와 SICO(국립 전자 사망 증명서 정보 시스템) 데이터베이스의 데이터 연계에 관한 흐름



자료: Feteira-Santos et al. (2022). Improving morbidity information in Portugal: Evidence from data linkage of COVID-19 cases surveillance and mortality systems. International Journal of Medical Informatics, 163, 104763. p.3.

SINAVE와 SICO에서 동일한 사망과 관련된 기록을 일치시켜 새로운 데이터베이스를 구축하였다. 각 기록은 SINAVE와 SICO 각 데이터베이스에 비식별화되어 있어, 각 데이터베이스에서 사용 가능한 4개 변수(연령, 성별, 거주지, 사망일)가 동일한 사례는 동일인으로 간주하여 연계하였다. 4개 변수가 중복적으로 일치할 경우에는 데이터베이스에서 제외시켰다. 즉, 공통 변수인 4개 변수를 연계 변수로 하고 연계 변수를 포함한 SINAVE와 SICO의 모든 변수들을 연결하여 새로운 데이터베이스를 구축하였다.

SICO에서는 이환 정보가 먼저 개방형 문항을 통해 보고되고, 이후

ICD-10 코드를 사용하여 코딩된다. 개방형 문항과 ICD-10 코드 간에 일부 불일치되는 경우와 SICO 데이터셋에서 이환 정보에 관한 오픈 문항 데이터는 존재하나 ICD-10 코딩이 되지 않은 경우가 관찰되었다. 연구진은 데이터셋으로부터 기인한 편향(bias)을 최소화하기 위해 개방형 문항에 대한 데이터는 있지만 ICD-10 코딩 정보가 없는 사례가 해당 주에 사망한 사례의 5% 이상일 경우 제외시키는 것으로 기준을 설정하였다.

표본 특성을 설명하고자 연령(4개 연령 그룹으로 분류)과 성별 변수를 사용하였고 국제표준화기구의 주간 날짜 체계(ISO 8601)에 따라 사망일을 해당 연도의 주 번호(1~53)로 코딩하였다. 질환의 중증도를 보정하고자 ICD-10에 따라 Charlson 동반질환 지수(Charlson comorbidity index, CCI)와 Elixhauser의 동반질환 측정치(Elixhauser's comorbidity measure, ECM)를 산출하였다.

ICD-10에 따라 SICO 데이터베이스에 코딩된 질병률은 ICD-10 코드로 전환하여 Elixhauser(ECI) 및 CCI를 산출하였고 이환율에 관해 1) 개인이 SINAVE에 직접 보고한 동반질환, 2) SINAVE에 보고된 질병의 발생 여부, 3) CCI 산출식에 따른 SICO에서의 동반질환, 4) ECI 산출식에 따른 SICO에서의 동반질환을 식별하였다. SINAVE에서의 동반질환 유무 변수와 동반질환 개수 변수를 추가 생성하였으며, 이 경우 만성 신장 질환과 만성 신부전은 동일한 변수로 간주하였다. 동일 사례에 대해 두 데이터베이스에 보고된 동반질환을 비교하였고, 의사 2인, 간호사 및 기타 의료 전문가가 포함된 연구팀의 자문을 받아 SICO에서 확인된 동반질환을 SINAVE에서 보고된 동반질환과 일치시키는 작업을 실시하였다.

범주형 변수에 대한 절대빈도와 상대빈도를 계산하는 등 기술 통계량을 사용한 통계적 분석을 수행하였다. 범주형 변수 간의 연관성을 평가하기 위해 피셔의 정확 검정 및 카이제곱 검정을, 두 데이터베이스 간의 일

치성을 검정하기 위해 코헨의 카파 계수를 사용하였다. 수행된 모든 검정에 대한 통계적 유의 수준은 5%를 기준으로 작성되었다.

두 데이터베이스를 결합한 결합 데이터베이스에서는 전체적으로 2,285건의 일치 사례가 분석되었으며, 그중 53.9%가 남성이고, 평균 연령은 84세로 나타났다. 질병률의 존재 범위는 26.3%에서 62.5% 사이였고, 10개 질환에 대해서는 SINAVE와 SICO에 보고된 정보 간에 비교할 수 있었고 SINAVE와 SICO 모두에서 동시에 보고된 비율은 당뇨병의 경우 5.7%, 인간 면역결핍 바이러스 감염 또는 응고병증의 경우 0.0% 범위였다. SINAVE와 SICO에서 보고된 이환율의 유사성을 평가해 보면, 최소한의 일치 또는 미일치로 나타났다. 결론적으로 데이터가 독립적으로 수집되므로 코로나19 사례 및 관련 사망을 모니터링하는 데 사용되는 두 개의 HIS에서 보고된 질환 상태에 대하여 서로 다른 정보를 찾을 수 있었다.

3. 일본 사례: 건강보험 청구데이터와 건강 검진 데이터의 결합 데이터 베이스 구축

가. 진료(조제) 보수 명세서 정보·특정 건강진단 등 정보 데이터 베이스(NDB)

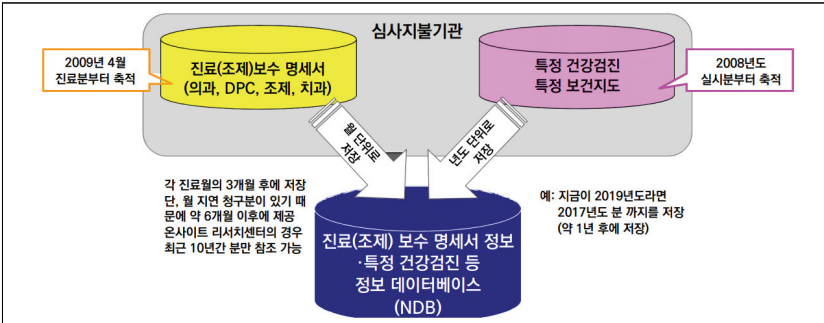
‘진료(조제) 보수 명세서 정보·특정 건강진단 등 정보 데이터베이스’(이하 NDB)는 “고령자의 의료의 확보에 관한 법률”에 근거해 의료비 적정화 계획의 작성, 실시 및 평가를 위한 조사나 분석 등에 이용하는 데이터베이스이다. 명칭에서 알 수 있듯이 진료(조제) 보수 명세서 정보, 즉 의료수가 청구 명세서 정보와 특정 건강진단·특정 보건 지도 정보를 저장하여 구축하고 있다. 의료비 적정화 계획에 관련된 조사나 분석 이외에도 의료 서비스의 질 향상 등을 목표로 증거 기반 정책 추진의 기초 자료로 활용하기 위해 데이터 제공을 시작하였다. 이와 함께 “진료(조제) 보수 명세서 정보·특정 건강 진단 등 정보에 관한 가이드라인”을 정비하여 NDB의 제3자 제공을 2011년부터 시범 실시, 2013년부터는 본격적으로 실시하였다.

국민건강보험공단 단일 보험자 체계의 전 국민 건강보험 제도를 운영하는 우리나라와는 상이하게 일본의 의료보험 제도는 전 국민 건강보험 제도로 다보험자 체계로 보험자 유형별로 가입자격이 상이하며 보험료율 및 보험료액 역시 보험자마다 상이하다. NDB는 보험종별(피용자건강보험, 국민건강보험, 후기고령자 의료제도)을 넘어서 데이터가 수집되고 있어 장기적인 데이터의 추적도 가능하며, 전 국민의 의료와 건강 동향 모두를 포괄적으로 파악할 수 있는 대규모 데이터베이스라는 강점이 있다.

NDB에는 건강보험 청구데이터(의과, 치과 등 진료 및 조제(약제)), DPC 데이터와 건강검진 데이터가 익명화 처리를 거쳐 결합된 후 저장되어 있다. DPC 제도는 우리나라의 DRG 제도와 대응되며 급성기 병원 입

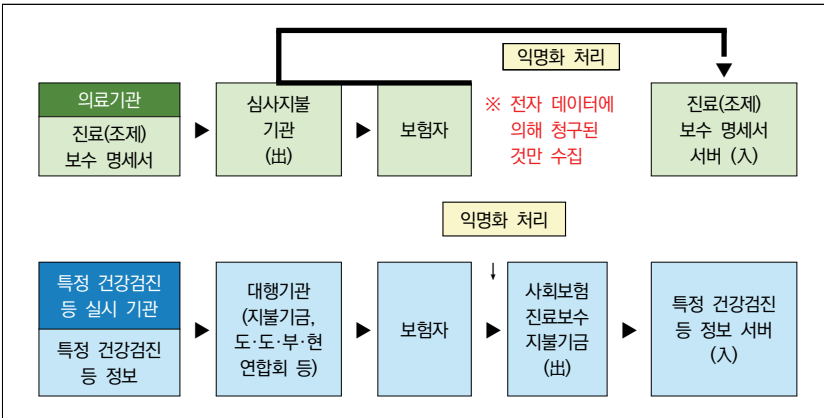
원환자에 대한 지불 제도로, 진단군 분류에 따라 1일당 정액 보수 산정제도이다. 개호 제도의 경우 우리나라의 장기요양보험과 대응된다.

[그림 3-3] NDB에 포함된 정보



자료: 厚生労働省保険局医療介護連携政策課保険データ企画室(후생노동성 보험국 의료 개호 제휴 정책과 보험 데이터 기획실). (2023.10). NDBの利用を検討している方へのマニュアル(NDB의 이용을 검토하고 있는 분을 위한 매뉴얼), p.6, 일본. 2023. 12. 20. 인출.
<https://www.mhlw.go.jp/content/12400000/001162218.pdf>

[그림 3-4] NDB 수집 경로



자료: 厚生労働省 老健局・保険局(후생노동성 노건국·보험국). (2018.9.6.). NDB、介護DB等の役割と解析基盤について(NDB, 개호 DB 등의 역할과 해석 기반에 대해서), p. 6, 일본.
<https://www.mhlw.go.jp/content/12401000/000350567.pdf>, 2023.03.24. 인출

진료(조제) 보수 명세서 정보는 보험 진료를 실시한 의료 기관이 보험자에게 진료 보수(의료 수가)를 청구하기 위한 명세서 정보로 각 보험자로부터 수집한 데이터로 구축된다. 각종 분석을 고려한 적절한 형식을 갖추지 못하고 있으며 이에 NDB 데이터를 이용하기 위해서는 데이터의 특징과 한계를 충분히 숙지할 필요가 있다.

〈표 3-4〉 NDB에 저장·제공되는 일본의 의료보험제도 종류별 전자화된 익명 진료(조제) 보수 명세서 정보

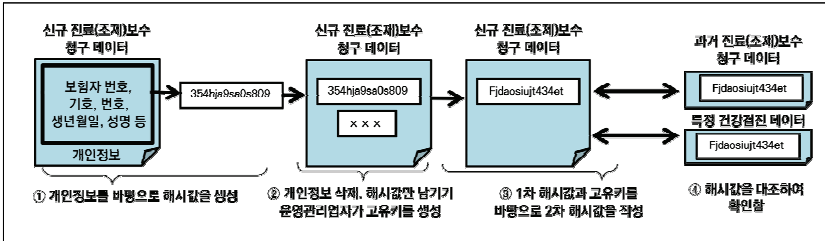
			의과		DPC		치과		조제	
			의료 보험	공비 부담	의료 보험	공비 부담	의료 보험	공비 부담	의료 보험	공비 부담
의료보험 — 전국민보험 —	피용자보험	건강보험	○	○	○	○	○	○	○	○
		선원보험	○	○	○	○	○	○	○	○
		공제보험	○	○	○	○	○	○	○	○
	국민건강보험	시·정·촌 국보	○	○	○	○	○	○	○	○
		국보 조합	○	○	○	○	○	○	○	○
	후기고령자 의료제도		○	○	○	○	○	○	○	○

자료: 厚生労働省保険局医療介護連携政策課保険データ企画室(후생노동성 보험국 의료 개호 제후 정책과 보험 데이터 기획실). (2023.10). NDBの利用を検討している方へのマニュアル(NDB의 이용을 검토하고 있는 분을 위한 매뉴얼), p.8, 일본. 2023. 12. 20. 인출.
<https://www.mhlw.go.jp/content/12400000/001162218.pdf>

NDB에 데이터가 수집될 때 환자의 성명, 생년월일의 일자, 보험 의료기관의 소재지와 명칭, 피보험자증의 기호와 번호 등 특정 개인이 식별되는 정보는 삭제된다. 대신에 이러한 정보를 함수로 처리한 ‘해시 ID’에 의해 동일인에 대한 식별이 가능하다. 예를 들어, 동일 해시 ID를 갖는 진료(조제) 보수 명세서 정보와 특정 건강진단 정보를 함께 추출하여 두 종류

의 정보를 연계하여 분석할 수 있다.

[그림 3-5] 익명화 처리 과정



자료: 厚生労働省 老健局・保険局(후생노동성 노건국·보험국). (2018.9.6.). NDB、介護DB等の役割と解析基盤について(NDB, 개호 DB 등의 역할과 해석 기반에 대해서), p. 6, 일본.
<https://www.mhlw.go.jp/content/12401000/000350567.pdf>, 2023.03.24. 인출

또한, NDB에는 익명화 처리(해시 ID)가 되어 있지만 환자 개인의 연령이나 성별 등의 기본적인 정보나 진료에 대한 자세한 정보가 포함되어 있다. 다른 정보와 대조하면 환자의 식별 가능성이 높아져, 데이터를 이용하기 위해서 데이터 이용 신청자는 일정한 요건을 갖추어야 하며, 연구 내용 및 연구 환경의 보안 등과 관련한 심사를 통과해야 한다.

해시 함수를 사용하여 개인 특징으로 이어지는 정보를 삭제하는 익명화 처리 후, 동일 인물의 정보를 식별·대조하여 확인한 다음 데이터베이스에 보관한다. 주어진 데이터로부터 개인정보(이름, 생년월일 등)를 바탕으로 고정 길이의 의사 난수(해시 값)를 생성하고 그것을 ID로 사용하여 개인 정보를 삭제한 진료(조제) 보수 명세서 정보 등에 대해 동일 인물의 정보로 특정할 수 있다. 서로 다른 데이터에서 동일한 해시 값을 생성하는 것은 매우 어려우며, 생성된 해시 값으로부터는 원래 데이터를 재현할 수 없으며 대응표도 작성하지 않는다.

해시 값은 해시함수에 의해 계산되는데, 해시 함수 자체 그리고 입력되는 개인 정보의 관리 상황 때문에 동일 인물의 정보를 매칭하는 것이 어

럽다. 이에 분석 목적에 따라 불량 데이터의 허용도, 수정 방침 등에 대한 고려가 필요하다.

결혼으로 인해 성씨가 바뀌거나 이직, 전직, 퇴직 등으로 인해 적용되는 건강보험제도 또는 보험자가 변경되는 경우 보험자 번호, 기호 번호, 생년월일, 성별, 성명 등 개인정보에 변화가 있을 수 있다. 개인정보를 바탕으로 해시 값을 생성하기 때문에 이러한 정보에 변화가 있을 때 매칭되지 않을 수 있다. 또한, 진료(조제) 보수 명세서 정보와 건강 검진·보건지도 데이터에서는 성명을 기재하는 규칙이 서로 달라 매칭되지 않을 수 있다. 마지막으로 해시함수의 기술적 특성 때문에 매우 작은 확률이지만 다른 입력 정보에서 동일한 해시 값이 생성될 수 있다.

이러한 해시 함수의 유의점을 보완하기 위해 2개의 해시 값(1차 해시 ID, 2차 해시 ID)을 생성하여 데이터를 매칭한다. 1차 해시 ID는 보험자 번호, 기호 번호, 생년월일, 성별로 생성하고, 2차 해시 ID는 성명, 생년월일, 성별로 생성한다.

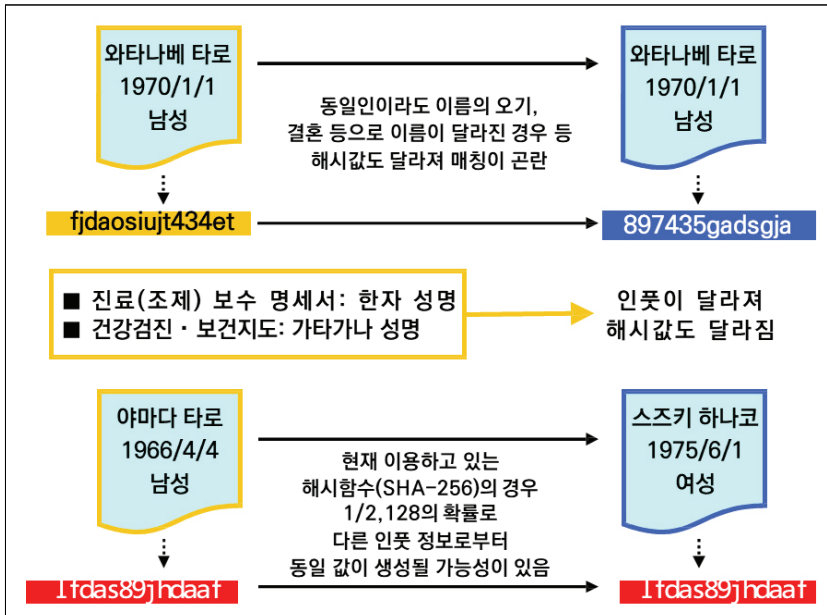
○ 1차 해시 ID = f(보험자 번호, 기호 번호, 생년월일, 성별)

○ 2차 해시 ID = f(성명, 생년월일, 성별)

2개의 해시 값을 생성함으로써 매칭이 가능해지는 경우는 다음 3가지와 같다. 1) 보험자 번호나 기호 번호가 변경되는 경우에는 2차 해시 ID로 매칭 가능하다. 단, 성명, 생년월일, 성별이 동일한 인물이 있을 경우(즉, 2차 해시 ID가 동일할 경우)에는 매칭이 불가능하다. 2) 오기나 결혼 등으로 인하여 성명이 변경된 경우 1차 해시 ID로 매칭이 가능하다. 단, 생년월일과 성별이 동일한 다른 인물이 동일 보험자 번호와 동일 기호 번호 내에 있을 경우에는 매칭이 불가능하다. 3) 진료(조제) 보수 명세서 정보와 특정 건강 검진·특정 보건 지도 데이터를 매칭할 경우 1차 해시 ID로

매칭이 가능하다. 단, 생년월일과 성별이 동일한 다른 인물이 동일 보험자 번호와 동일 기호 번호 내에 있을 경우에는 매칭이 불가하다.

[그림 3-6] 해시 함수에 대한 유의사항



자료: 厚生労働省 老健局・保険局(후생노동성 노건국·보험국). (2018.9.6.). NDB、介護DB等の役割と解析基盤について(NDB, 개호 DB 등의 역할과 해석 기반에 대해서), p. 7, 일본.
<https://www.mhlw.go.jp/content/12401000/000350567.pdf>, 2023.03.24. 인출

보험자 번호, 기호 번호 및 성명 모두 변화가 있을 경우에는 2개의 해시 값을 생성하였음에도 불구하고 매칭이 불가하다. 결혼 등으로 인해 보험자가 변경되고 성명이 변경된 경우 또는 전직 등으로 인해 보험자가 변경되고 성명 기재에 오기가 생겼을 경우가 있다.

제도의 특성에 의해 일본은 전국 단위 데이터가 부재한 실정이며 도·부·현 지역 단위로 데이터가 비교적 잘 구축되어 있는 실정이다. 이러한 지역 데이터를 결합하여 전국 단위 데이터인 NDB로 구축하였고 중앙

부처 및 지자체에서는 의료비 적정화 계획의 작성 및 평가 등에 정책적으로 활용하고 있다. NDB는 보험자, 연구 기관, 대학 및 국가 지원 연구를 수행하는 연구자(민간 기업 포함) 등 제3자에게 제공될 수 있으며, 이들에 의해 의료서비스의 질 향상 등에 기여할 수 있는 분석 및 연구에 활발하게 활용되고 있다. 또한, NDB 외에도 의료와 개호 분야 DB가 완전 별도로 분리되어 있어 의료와 개호 데이터를 결합하는 연구도 활발하게 이루어지고 있다.

〈표 3-5〉 NDB 활용 사례

구분	사례
정책 통계 추출	진료 행위나 투여된 의약품, 진료 보수 등에 대한, 도·도·부·현별, 성별, 연령 구간별의 실적, 상병명이나 보험자 종별 정보 등의 정보를 추가하여 다각적인 평가
시계열 자료 구축 및 평가	특정 사례를 시계열로 추적함으로써 각 질병에 대해 어떤 치료가 행해졌는지 등에 대한 분석 - COPD 환자에서의 후속 재택 산소 요법 도입의 실태 평가
건강검진 정보와 진료 정보 연계	특정 건강 검진 등의 정보를 이용하여 대상자의 건강 상황을 연구하며 진료(조제) 보수 명세서 정보와 연결함으로써 진료와 관련된 다양한 사항을 건강 검진 정보와 관련시켜 분석할 수 있음. - 운동습관이 있는 진찰자 중 대사 증후군을 가진 환자의 비율 - 지역별, 성별, 연령 구간별로 본 식습관이나 음주, 흡연에 대한 정보 - 대사 증후군의 기준을 충족하는 특정 건강 검진 대상자 1인당 평균 의료비

자료: 厚生労働省保険局医療介護連携政策課保険データ企画室(후생노동성 보험국 의료 개호 제휴 정책과 보험 데이터 기획실). (2023.10). NDBの利用を検討している方へのマニュアル(NDB의 이용을 검토하고 있는 분을 위한 매뉴얼), p.7, 일본. 2023. 12. 20. 인출.
<https://www.mhlw.go.jp/content/12400000/001162218.pdf>

제3절 소결

국내 가명정보 연구 결합 사례는 조사데이터와 행정데이터가 결합된 사례(통계청, 2020; 임성희·김진옥, 2021)와 행정데이터 간의 결합이 이루어진 사례(정찬미, 2021; 이현주, 2022)로 구분해 볼 수 있다.

통계청(2020)은 행정자료의 활용을 통해 가계금융복지조사의 품질을 제고하는 데 그 결합 목적을 두고 있다. 국세청의 과세자료, 보건복지부의 사회수혜금, 각 연금공단의 연금과 건강보험료 지출 등 다양한 행정자료들과의 통합을 통해 가계금융복지조사는 정확성, 일관성, 완전성 등의 측면에서 품질을 제고할 수 있게 된다. 이러한 결합은 개인의 성명, 성별, 생년월일, 주소 등의 매칭기를 통해 이루어지며, 결합을 통한 보완 이후 가구소득의 상향과 지니계수 및 상대적 빈곤율의 증가 등의 결과를 보이고 있다. 이와 같이 행정자료의 활용으로 통계품질 개선 및 응답 부담을 줄일 수 있다는 효과를 기대할 수 있으나, 높은 품질의 행정자료 요구와 더불어 명확한 자료처리 기준이 필요하다는 점을 연구는 강조하고 있다.

임성희·김진옥(2021)은 국내 창업생태계에서 개인에 대한 생존율 및 생존요인에 대한 분석을 위해, 통계청의 조사자료와 국세청의 과세자료를 결합한 SBR(통계기업등록부) 자료를 이용하였다. 생존분석을 위해 SBR 자료에서 사업자등록번호를 이용하여 개인사업체만을 추출하였고, 개인대체식별번호를 매칭기로 사용하여 공동사업자명세자료와 프랜차이즈 가맹점 모집단 자료를 연계하였다. 이와 같은 결합을 통해 개인사업체의 생존시간을 추정하였는데, 결합 시 주의점으로 연계 가능한 식별정보의 확인과 응용 가능한 데이터가 일정 수준의 수가 되는지, 그리고 분석에 유의한 변수를 포함하는지 등의 확인이 필요함을 강조하였다.

정찬미(2021)는 사회보장 정책연구를 위해 사회보장위원회에서 주민

등록전산정보와 사회보장 수급 관련 데이터를 결합한 행정데이터를 이용하였다. 이 행정데이터는 사회보장정책을 수행하는 11개 부처, 18개 자료 보유기관의 협조로 생성된 자료이다. 결합을 위한 매칭키로 주민등록 전산정보상의 개인정보를 이용하였으며, 1인 가구의 사회보장 수급 현황 분석을 통해 청년은 소득수준 및 사회보험 가입률은 높으나 자산 수준이 낮음을 확인하고, 노인은 근로하는 노인은 많으나 양질의 근로를 하지 못하는 것을 확인하였다. 다만 이러한 결과 도출을 위한 행정데이터의 결합 시 개인의 수급 현황 파악을 위해 단일 연도 기준으로 자료를 수집한 관계로 소득·자산을 정확히 파악하지 못하는 한계점을 갖고 있다.

이현주(2022)는 사회보장 행정데이터의 효율적인 연계·활용을 위한 과제를 도출하고자 사회보장 행정데이터의 의의와 환경, 그리고 추출과 연계 방안에 대한 논의를 진행하였다. 행정데이터 간의 통합 시 현재 사용되고 있는 매칭키는 생년월일, 성, 이름을 가명처리하여 생성하는데, 법적 규정 등으로 주민번호를 가명처리하여 활용할 수 없기 때문에 대안으로 활용하는 결합키라 할 수 있다. 이현주(2022)가 제시한 과제는 요청 정보의 범위에 대한 합리적 판단 기제 마련, 데이터의 안정적 확보를 위한 협치 강화, 여러 정보를 연계하기 위한 연계키의 정확성 제고, 통합데이터를 활용한 연구 주제 개발과 분석 결과를 반영할 수 있는 기제 강화로 정리된다.

해외 가명정보 연구 결합 사례를 살펴보기 위해 우선 영국의 가족자원조사(Family Resources Survey, FRS)를 소개하고, 가명정보 데이터를 이용한 영국 논문 2편(Pattaro et al. 2020; Jenkins and Rios-Avila, 2021)을 살펴보았다. 이 외에도 포르투갈과 일본의 보건·의료 분야 행정데이터 간 결합 사례를 살펴보았다.

가족자원조사(Family Resources Survey, FRS)는 현재 영국 내 개인 가구 대표 표본을 통해 개인들의 소득과 그 상황에 관한 정보를 매년 수

집하는 조사로, 주요 목표는 국 노동연금부(Department for Work and Pension, DWP)에 사회복지 정책의 개발과 모니터링 및 평가를 위한 정보를 제공하는 것이다. FRS 데이터와 행정데이터의 결합에 대해 살펴보면, DWP 디지털 데이터 전략에 따라 해당 부서는 행정데이터를 연결 가능한 다양한 소스(예: 정부의 다른 부분)에 연결하여 FRS를 변형하는 일을 시행하고 있다. 2018년 GDPR이 도입되어 데이터 결합을 법적으로 동의하는 대안이 제공되었다. 이로써 공공기관의 기능 수행을 위해 데이터 결합이 필요하다는 법적 근거를 바탕으로 모든 응답자를 연결할 수 있게 되었고, 연결 방법론의 개선으로 최소한 95% 이상의 FRS 응답자를 그들의 행정 기록에 연결할 수 있게 되었다. 이러한 행정데이터와 FRS 데이터의 통합으로, 다양한 금융예의 활용, 데이터 품질 제고, 응답자 및 사용자 이점을 실현할 수 있는 가능성이 열렸다.

Pattaro et al.(2020)은 행정데이터(Administrative Data)를 사용하여 사회 불평등을 식별하고자 하였다. 연구에 사용된 데이터는 소득과 빈곤에 관련된 두 가지의 국가적인 설문조사에 국가 행정 시스템의 데이터를 연결한 독특한 조합에 기반하고 있다. 영국의 빈곤과 사회배제(PSE-UK)조사는 2012년에 실시된 횡단면 가구 조사이다. 이 조사는 가족자원조사(FRS) 2010~2011년 조사에 참가한 응답자 샘플로 구성되어 있으며, 재연락에 동의한 응답자들을 포함하고 있다. 이와 함께 무직수당(JAS), 소득지원(IS) 행정데이터를 결합하였는데, 이러한 결합을 위한 매칭키로 국민 보험 번호(National Insurance Number)를 이용하였다. 장기적인 행정데이터의 사용은 정책 및 빈곤 연구뿐만 아니라 다양한 응용 분야에서 폭넓게 활용될 수 있음을 연구 결과가 말해 주고 있다.

Jenkins and Rios-Avila(2021)은 행정데이터에서의 측정오차를 알아보기로, 설문조사 데이터와 행정데이터가 결합된 데이터에서 다양한

유형의 측정오차를 고려하고 관련된 특성에 따라 오차 분차를 다르게 한 계량모형을 개발·적용하였다. 분석을 위해 FRS(가족자원조사)-P14(연도 말 과세 신고자료) 결합데이터를 이용하였는데 두 자료의 결합을 위한 매칭키로 성명, 우편번호, 성별, 생년월일을 사용하였다. 분석 결과, 전반적으로 FRS 데이터 수입이 P14 수입보다 신뢰성이 있으며, 오차의 조합으로 인해 두 데이터 모두 관측된 수입은 참수입의 불평등을 상당히 과대 평가하는 것으로 나타났다.

포르투갈에서 행정데이터들을 결합한 사례로는 SINAVE(국립 역학 감시 정보 시스템)와 SICO(국립 전자 사망 증명서 정보 시스템)의 데이터 결합이 있다. 포르투갈의 모든 의사는 관련 법령에 따라 SINAE 시스템을 통해 전염병 및 기타 공중 보건 위험에 대해 의무적으로 신고해야 한다. 마찬가지로 SICO 시스템을 통해 의무적으로 전자 사망증명서를 발급해야 한다.

Feteira-Santos et al.(2022)의 연구에서는 SINAVE와 SICO를 결합하여 구축한 데이터베이스를 기반으로 2020년 코로나19 관련 이환율 분석을 실시하였다. 코헨의 카파 계수를 통해 두 데이터베이스 간의 일치성을 검정하였고 두 데이터베이스에서 보고된 이환율의 유사성은 최소한의 일치 또는 미일치로 나타났다.

일본의 행정데이터 간의 결합 사례로는 건강보험 청구데이터와 건강검진 데이터의 결합 데이터베이스를 구축한 NDB(진료(조제) 보수 명세서 정보·특정 건강 진단 등 정보 데이터베이스)가 있다. 두 개의 데이터베이스를 해시 함수를 활용해 익명화 처리한 후 결합 데이터베이스를 구축하였다.

NDB 구축을 통해 전국 단위 행정 DB를 구축하여 정책 통계를 추출하고 시계열 자료를 구축하여 정책 평가 등에 활용하고자 한다. 또한, 건강검진 정보와 진료 정보를 연계하여 의료비 적정화 계획을 수립하고 정책의 실시 및 평가를 위해 활용한다.



제4장

가명정보 결합 활성화를 위한 제반 사항 검토

제1절 현행 법률 및 가이드라인 검토

제2절 데이터 제공 관련 현황 분석

제3절 데이터 제공 절차 수립

제4절 데이터 제공 절차 개선 방안

제5절 소결

제4장

가명정보 결합 활성화를 위한 제반 사항 검토

가명정보 결합 활성화를 위해 여러 가지 제반 사항을 검토하여 한국보건사회연구원의 데이터 제공 절차를 마련하였다. 한국보건사회연구원의 데이터 제공 절차에는 가명정보 데이터 제공과 관련된 사항이 포함되어 있지 않다. 그래서 이와 관련하여 2023년 7월 12일부터 10월 11일까지 컨설팅을 수행하였다. 주요 컨설팅 과업 내용은 데이터 제공 절차와 관련된 현행 법률과 제도를 살펴보고, 한국보건사회원의 데이터 제공 절차에 대한 현행 규정과 절차를 분석하였다(〈표 4-1〉 참조). 이를 기반으로 하여 한국보건사회연구원의 데이터 제공 절차를 수립하였다.

〈표 4-1〉 데이터 제공 절차 관련 주요 컨설팅 과업 내용

구분	과업 내용
현행 법률 및 제도 분석	- 7개 법률 분석 1) 「공공기관의 정보공개에 관한 법률」 및 시행령, 시행규칙 2) 「통계법」 및 시행령, 시행규칙 3) 「공공데이터의 제공 및 이용 활성화에 관한 법률」 및 시행령, 시행규칙 4) 「데이터 기반 행정 활성화에 관한 법률」 및 시행령, 시행규칙 5) 「개인정보 보호법」 및 시행령, 시행규칙 6) 「생명윤리 및 안전에 관한 법률」 및 시행령, 시행규칙 7) 「신용정보의 이용 및 보호에 관한 법률」 및 시행령, 시행규칙 - 3개 가이드라인 분석 1) 「가명정보 처리 가이드라인」 2) 「보건의료데이터 활용 가이드라인」 3) 「공공분야 가명정보 제공 실무안내서」
한국보건사회연구원의 현황 분석	- 2개 지침 분석 1) 「정보공개 지침」 2) 「조사자료 관리 지침」 - 「통계법」에 따라 제공하는 국가승인 통계 데이터 제공 절차 검토 - 각종 양식 검토와 개선점 도출

구분	과업 내용
데이터 제공 절차 수립	- 새로운 데이터 제공 절차 수립 - 새로운 데이터 제공 절차에 따른 역할과 책임 수립
데이터 제공 절차 개선 방안	- 현행 절차의 개선점 도출

자료: 필자 작성

제1절 현행 법률 및 가이드라인 검토

한국보건사회연구원이 보유한 데이터를 제공할 때, 이에 대한 근거는 4개 법령²⁸⁾을 따르고 있다. 그리고 개인정보가 포함된 데이터를 제공하면 「개인정보 보호법」을 고려하여야 한다. 또한, 인간 대상 연구를 위한 데이터 제공을 위해서는 「생명윤리 및 안전에 관한 법률」, 마지막으로 금융기관의 데이터와 결합하기 위해서는 「신용정보의 이용 및 보호에 관한 법률」에 대한 검토가 필요하다. 한편, 제공하는 데이터에 개인정보가 포함된 경우 이에 대한 비식별 처리가 필수적으로 필요하다. 비식별 처리 수준에 따라 가명정보와 익명정보의 형태로 데이터가 제공될 수 있다. 이때 제공하는 데이터의 내용에 따라 3개 가이드라인²⁹⁾을 참고하여 제공해야 하므로, 이에 대한 검토도 필요하다.

이 절은 7개 법률과 3개 가이드라인에 대해 정의 및 개요, 특징을 살펴본 후 검토 결과를 정리하였다.

28) 「공공기관의 정보공개에 관한 법률」, 「통계법」, 「공공데이터법」, 「데이터기반행정법」

29) 「가명정보 처리 가이드라인」, 「보건의료데이터 활용 가이드라인」, 「공공분야 가명정보 제공 실무안내서」

1. 공공기관의 정보공개에 관한 법률³⁰⁾

「정보공개법」은 공공기관이 보유·관리하는 정보에 대해 청구인이 정보 공개를 청구할 때 따르게 되는 법률이다. 「정보공개법」은 전체 5장, 29조로 구성되어 있고, ‘정보’와 ‘공개’에 대한 정의와 공개원칙 및 범위, 청구권자, 공공기관의 의무, 정보공개 절차 등을 명시하고 있다. 여기서 ‘공개’는 ‘제공’에 대한 의미를 포함하고 있다.

「정보공개법」만의 특징은 청구, 즉 신청할 때 청구인이 직접 신청서를 작성하지 않고도 해당 공무원이나 임직원에게 말로써 청구할 수도 있다는 것과 정보공개 여부 검토는 정보공개심의회를 거치도록 하는 것이 특징이다. 또한 이의신청 제도를 비롯해 행정심판, 행정소송을 통해 청구인이 결과에 불복해 구제받을 수 있는 절차를 포함하고 있다.

「정보공개법」의 검토 결과, 한국보건사회연구원에서는 「정보공개법」의 데이터 제공과 관련하여 「정보공개지침」을 수립하여 운영하고 있다. 이에 따라 「정보공개법」과 관련한 검토 결과는 「정보공개지침」의 검토 결과 부분에서 함께 서술하도록 한다.

2. 통계법

「통계법」은 통계작성기관이 작성 및 보유한 통계자료에 대해 이용하려는 이용자 또는 통계작성기관의 요청 시 통계자료의 제공을 위해 따르게 되는 법률이다. 한국보건사회연구원은 현행 3가지³¹⁾의 승인통계를 보유하고 있으며 그 외 다양한 승인통계의 조사를 수행하였다. 이러한 승인통

30) 이 장에서 ‘공공기관의 정보공개에 관한 법률’은 약칭으로 「정보공개법」이라고 정의함.

31) 가족과 출산 조사, 한국복지패널조사, 한국의료패널조사

제의 제공 시에는 「통계법」에 따라 제공되어야 한다. 「통계법」은 전체 7장, 42조로 구성되어 있고, 통계의 정의부터 통계의 작성·보급의 이용 기반 구축, 통계작성기관 및 지정통계의 지정, 통계의 작성·보급 및 이용, 통계 응답자의 의무 및 보호 등을 다루고 있다.

「통계법」의 특징은 데이터 제공 여부를 검토할 때 통계작성기관의 장이 하며, 제공결정기간이 요청받은 날부터 30일 이내라는 것이다. 또한 요청에 대한 거부 내용도 없고, 이의신청이나 분쟁조정 내용 또한 없다. 그리고 올해 6월 제정되어 시행된 통계청 예규인 「국가 통계자료 제공 및 이용에 관한 규정」에 따라 공표 외 자료인 통계기초자료 및 미공표 집계자료는 비식별처리 방법에 가명처리를 포함하고 있는 것이 특징이다.

「통계법」의 검토 결과, 「통계법」은 「통계법시행령」 제49조에 통계자료제공심의회를 두도록 하고 있으며, 통계자료의 제공³²⁾ 또는 통계자료의 이용³³⁾ 시 이에 대한 심의를 하도록 하고 있으나, 3건의 국가승인통계를 보유하고 있는 한국보건사회연구원 통계자료제공심의회에 대한 별도의 규정이나 지침이 없어 「조사자료 관리지침」을 적용하여야 한다. 「조사자료 관리지침」에는 조사자료관리위원회를 두고 심의를 할 수 있도록 하고 있으나 데이터 제공 또는 이용 시 적용해야 하는 비식별 처리(「국가통계자료 제공 및 이용에 관한 규정」에 따른 가명처리를 포함)에 대한 세부적인 내용을 다루고 있지 않다. 이에 따라 별도의 데이터 제공을 위한 심의위원회의 구성이나 「조사자료 관리지침」에 이를 반영한 개정이 필요하다.

32) 통계법 제30조에 해당하며 일반적인 용어와 달리 통계법에서는 통계작성을 위해 다른 통계작성기관에 요청하여 제공받는 것을 제공이라 함.

33) 통계법 제31조에 해당하며 통계작성기관이 아닌 모든 제공은 통계자료의 이용이라 함.

3. 공공데이터법

「공공데이터법」은 공공기관이 법령 등에서 정하는 목적을 위해 생성 또는 취득하여 관리하고 있는 모든 종류의 전자자료인 공공데이터를 누구나 쉽게 이용할 수 있게 하기 위한 법률이다. 「공공데이터법」은 전체 6장, 40조로 구성되어 있고, 공공데이터의 정의부터 공공데이터의 정책 수립, 공공데이터 등록 등 제공기반 조성, 공공데이터 제공절차 등을 다루고 있다.

「공공데이터법」의 특징은 데이터 제공 여부 결정을 ‘데이터 보유기관의 장’이 하도록 하고 있으며, 제공결정기간이 요청받은 날부터 10일 이내로 30일 이내인 「통계법」보다 짧다. 또한 제공에 대한 거부를 할 수 있으며, 제공 거부에 따른 분쟁조정을 할 수 있는 분쟁조정위원회 제도를 갖추고 있다.

「공공데이터법」의 검토 결과, 「공공데이터법」은 데이터 제공 시 공공기관의 장이 「정보공개법」 제9조의 비공개 대상 정보와 저작권법 및 그 밖의 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 것으로 해당 법령에 따른 정당한 이용 허락을 받지 아니한 정보에 대해 제공하지 않도록 하고 있어, 앞의 「정보공개법」에서 살펴본 개인정보가 포함된 정보의 제공에 대한 내용이 동일하게 발생한다. 이에 따라 「정보공개지침」의 검토 내용을 동일하게 적용하여야 한다. 또한 「공공데이터법」 제35조, 「공공데이터법시행령」 제28조, 「공공데이터 관리지침」 제16조에 따라 공공데이터의 제공에 비용이 드는 경우 최소한의 필요비용을 받을 수 있으며, 이러한 경우 비용부과 기준을 해당 기관의 규정으로 마련하고 기관 홈페이지 등을 통해 공개하도록 하고 있다. 그러나 현재 데이터 제공에 대한 비용부과 기준이 수립되어 있지 않으며 이에 따라 무상으로 데이터를 제

공하고 있다. 앞으로 데이터 제공이 늘어날 것을 고려할 때 비용부과 기준을 규정으로 수립하고 이를 공개하는 것이 필요하다.

4. 데이터기반행정법

「데이터기반행정법」은 공공기관 간에 데이터를 기반으로 한 행정자료의 ‘공유’와 ‘활용’을 활성화하기 위한 법률이다. 「데이터기반행정법」은 전체 4장, 25조로 구성되어 있고, 데이터기반행정의 정의부터 데이터의 등록 및 제공 절차, 데이터기반행정의 기반 구축 등을 다루고 있다.

「데이터기반행정법」의 특징은 데이터 요청 주체가 공공기관에만 해당되며, 데이터 제공 여부 결정을 ‘데이터 보유기관의 장’이 하도록 하고 있으며, 제공 결정 기간이 요청받은 날부터 20일 이내로 10일 이내인 「공공데이터법」보다 길고, 30일 이내인 「통계법」보다 짧다. 또한 제공에 대한 거부를 할 수 있으며, 제공 거부에 따른 분쟁조정을 할 수 있는 분쟁조정위원회 제도를 「공공데이터법」과 동일하게 갖추고 있다. 앞으로 개정될 사항에는 분쟁조정위원회 제도를 「공공데이터법」을 따르도록 하고 있다.

「데이터기반행정법」의 검토 결과, 「데이터기반행정법」은 데이터 제공 시 개인정보가 포함된 데이터에 대해서는 「개인정보 보호법」을 준용하도록 하고 있으며 이에 대한 내용은 다음 ‘5. 개인정보 보호법’을 참조한다. 또한 「데이터기반행정법」은 데이터 제공 시 발생하는 비용에 대해 법 제 12조 제3항, 시행령 제13조에 따라 요청기관의 장에게 부담하게 할 수 있으나 현재 한국보건사회연구원은 이에 대한 규정이 수립되어 있지 않다. 그러므로 데이터 제공에 대한 비용을 포함한 규정의 수립이 필요하다.

5. 개인정보 보호법

「개인정보 보호법」은 개인정보의 처리 및 보호에 관한 사항을 정하기 위한 법률로 ‘일반법’에 해당한다. 따라서 ‘특별법’과 충돌 시 특별법을 우선 적용하며, 그 외 나머지는 「개인정보 보호법」을 따르게 된다. 「개인정보 보호법」은 전체 10장, 76조로 구성되어 있고, 개인정보의 정의부터 개인정보 보호정책의 수립, 개인정보의 처리, 개인정보의 안전한 관리, 정보 주체의 권리 보장, 개인정보 분쟁조정위원회, 개인정보 단체소송 등을 다루고 있다.

「개인정보 보호법」의 특징은 개인정보처리자가 개인정보를 보호하기 위한 기본 원칙을 정의하고 정보 주체의 개인정보가 침해당했을 경우 별도의 분쟁조정, 집단분쟁조정, 단체소송 등을 할 수 있다. 따라서 최소한으로 수집하고, 수집한 목적으로만 활용하며, 원본이 아닌 익명 또는 가명으로 활용하도록 하고 있다.

「개인정보 보호법」을 검토한 결과, 「개인정보 보호법」은 개인정보에 대해 제28조의 2, 제28조의 3에서 통계작성, 과학적연구, 공익적 기록보존의 목적으로 가명처리하여 제공할 수 있도록 하고 있다. 이에 따라 가명처리하여 데이터를 제공하는 절차를 「가명정보처리 가이드라인」 등 관련 가이드라인을 참조하여 수립할 필요가 있다. 또한 다른 법령에서 개인정보가 포함된 정보에 대해 「개인정보 보호법」을 따르도록 하고 있는 경우나, 비식별 처리 등의 방법에 「개인정보 보호법」의 가명처리가 필요한 경우 역시 「개인정보 보호법」의 관련 내용을 따라야 한다. 이에 따라 가명정보를 제공하는 절차, 적합성 검토, 적정성 검토를 심의하는 절차를 수립하여야 한다.

6. 생명윤리법

「생명윤리법」은 인간과 인체유래물 등을 연구할 때 생명윤리 및 안전을 확보하기 위한 내용을 규정한 법률로 보건의료데이터를 활용하기 위해서는 해당 법률을 따라야 한다. 「생명윤리법」은 전체 9장, 70조로 구성되어 있고, 인간 대상 연구, 인체 유래물 연구, 개인정보 등의 정의부터 생명윤리심의위원회, 인간 대상 연구 및 연구 대상자 보호, 배아 등 생성과 연구, 인체 유래물 연구 및 인체 유래물 은행, 유전자 치료 및 검사 등을 다루고 있다.

「생명윤리법」의 특징은 인간을 대상으로 하는 연구를 하기 위해서는 기본적으로 연구대상자의 동의를 요구(일부 후향적 연구의 경우 생명윤리위원회 심의를 통해 동의 면제 가능)하고 있으며, 생명윤리위원회의 심의를 거쳐야 한다는 것이다. 이후 처리 및 제공에 관한 절차는 「보건의료데이터활용 가이드라인」을 참고할 수 있다.

「생명윤리법」의 검토 결과, 「생명윤리법」에 따라 인간 대상의 연구를 수행할 때 생명윤리위원회의 심의를 받도록 하고 있으며 이에 따라 한국보건사회연구원은 「생명윤리위원회 운영규정」을 제정하고 이에 따라 생명윤리위원회를 운영하고 있다. 생명윤리위원회 운영규정의 연구 참여 동의서의 내용을 보면 3번 항목의 내용과 4번 항목의 내용이 가명정보를 이용한 데이터 제공을 어렵게 하는 부분이 포함되어 있다.

[그림 4-1] 연구 참여 동의서-익명화 관련 예시

연구 참여 동의서 3. 본인의 개인식별정보는 익명화되고 모든 정보는 엄격히 관리되는 것을 알고 있습니다. 4. 익명화된 조사정보는 연구 목적으로 제3자에게 제공될 수 있음을 알고 있습니다.
--

위와 같이 연구 참여 동의서에 개인식별정보를 익명화하도록 하고 있는 경우 생명윤리법의 익명화의 기준(보건복지부 생명윤리정책과, 2020.8.5.)과 개인정보보호법 등 다른 법령의 익명화 기준의 차이점(개인정보보호법의 가명처리는 생명윤리법의 익명화에 포함되는 것으로 해석됨)으로 인해 익명화에 대한 논란이 있을 수 있으며, 이에 대한 명확한 설명을 포함하는 것이 필요하다. 또한 제3자 제공의 경우 개인정보 수집 및 이용 동의서의 가장 하단에 다음과 같이 정의되어 있다.

[그림 4-2] 연구 참여 동의서-제3자 제공 관련 예시

▷ 위와 같이 수집된 귀하의 정보를 연구 목적으로 제3자에게 제공하는 것에 대해 동의 하십니까? ☐ 동의함 ☐ 동의하지 않음
--

이의 경우에도 연구 목적에 대한 해석(현재 동의받은 연구 목적인지 다른 연구 목적도 가능한 것인지)에서 다툼의 여지가 있으며, 실제 다른 기관에서 이런 다툼이 일어나고 있다. 이에 따라 동의서의 내용을 일부 수정하여 개인정보보호법의 가명처리를 통한 데이터의 활용 가능성을 높일 필요가 있다.

7. 신용정보법

「신용정보법」은 신용정보의 이용 및 보호를 위한 법률로 금융위원회가 승인한 신용정보회사 등이 따르게 된다. 「신용정보법」은 전체 9장, 70조로 구성되어 있고, 신용정보의 정의부터 신용정보업의 수집 및 처리, 유통 및 관리, 신용정보 주체의 보호 등을 다루고 있다.

「신용정보법」의 특징은 신용정보업을 하는 신용정보회사들이 따르는 법으로 한국보건사회연구원은 해당이 없다. 다만, 이러한 신용정보회사들의 데이터를 이용하여 가명정보를 결합할 때 해당 법에 따라 결합을 결합전문기관이 아닌 데이터 전문기관을 통해서만 결합할 수 있다.

「신용정보법」의 검토 결과, 「신용정보법」으로 인해 한국보건사회연구원이 별도로 개정해야 할 규정은 없는 것으로 판단된다. 다만 신용정보업을 하는 신용정보회사들과 데이터를 결합하는 경우 「신용정보법」의 절차를 준수하여야 하며 이에 따라 이를 포함한 데이터 제공 절차의 수립이 필요하다.

8. 가명정보 처리 가이드라인

「가명정보 처리 가이드라인」은 「개인정보 보호법」 제3장 제3절 ‘가명정보 처리에 관한 특례’에 관한 설명과 구체적 사례를 제공하기 위해 만들어졌으며, 가명정보의 처리에 대한 이해를 돕고, 처리 과정에서 발생할 수 있는 개인정보 오·남용을 방지하여 안전한 가명정보 활용 방안을 안내하기 위해 작성됐다. 「가명정보 처리 가이드라인」은 목적과 용어 정의부터 가명처리 및 가명정보의 처리 절차, 가명정보 결합 및 반출 절차, 안전성 확보 조치 방안과 각 절차에 필요한 서식까지 담고 있다.

[그림 4-3] 가명정보 처리 가이드라인의 가명처리 절차



자료: 개인정보보호위원회. (2022). 가명정보 처리 가이드라인. p.10-38을 참고하여 필자 작성

「가명정보 처리 가이드라인」의 특징은 다른 가이드라인들에 비해 표준적 절차를 제시한다는 것이다. 다만, 세부적인 방법에 대해서는 「개인정보 보호법」이 적용되는 기관들의 환경에 따라 유연하게 적용할 수 있도록 강제화하고 있지는 않으며, 특정 분야에 속하면 해당 분야의 가이드라인을 따르도록 하고 있다.

이에 따라 가명처리의 기본 절차는 보건의료 데이터가 아닌 경우 「가명정보 처리 가이드라인」의 가명처리 절차에 따라 진행해야 한다.

9. 보건의료데이터 활용 가이드라인

「보건의료데이터 활용 가이드라인」은 「개인정보 보호법」 제3장 제3절 ‘가명정보 처리에 관한 특례’에 관한 사항과 생명윤리법 시행규칙 제13조 제1항(보건복지부령으로 정한 기준에 맞는 연구)을 근거로 하고 있다. 「보건의료데이터 활용 가이드라인」은 목적과 용어 정의부터 가명처리 원칙, 가명정보 활용 원칙, 가명정보 처리 및 활용 절차, 대상정보 및 가명

처리 방법, 결합 및 활용, 「개인정보 보호법」 개정에 따라 「생명윤리법」 유권해석 등을 담고 있다.

「보건의료데이터 활용 가이드라인」의 특징은 데이터 심의위원회에 있다. 「가명정보 처리 가이드라인」의 적정성 검토위원회와 비슷한 성격이지만, 세부 구성 요건이 아래 그림과 같이 명확하다.

[그림 4-4] 보건의료데이터 활용 가이드라인 데이터 심의위원회 세부 구성 요건

- ▶ 보건의료데이터를 이용한 연구 경험이 풍부한 자
- ▶ 개인정보보호 업무 경력과 전문성을 갖춘 자 (외부위원 1명 이상 포함)
- ▶ 정보주체 또는 그 관점에서 이익을 대변할 수 있는 자 (외부위원 1명 이상 포함)

자료: 개인정보보호위원회·보건복지부. (2022). [2022년] 보건의료 데이터 활용 가이드라인. p. 7.

전체적으로 5인 이상으로 구성하되 2명 이상의 외부 위원으로 구성될 수 있도록 하고 명시하고 있고, 그 구성원의 자격요건 또한 명시하고 있다. 또한 정보 주체의 동의 원칙이 필수인 정보와 다양한 보건의료데이터의 처리 방식들이 정의되어 있다.

[그림 4-5] 정보 주체의 인권 및 사생활 보호에 중대한 피해를 야기할 수 있어 본인의 동의를 받아 사용하는 것을 원칙으로 하는 정보

- ▶ 정신질환 정보
- ▶ 감염병예방법 제2조제10호에 따른 성매개감염병 정보
- ▶ 후천성면역결핍증 정보
- ▶ 희귀질환관리법 제2조제1호에 따른 희귀질환 정보
- ▶ 학대 및 낙태 관련 정보 (질병분류코드 기준으로 T74, O04 그 외 의뢰진 판단 활용)

자료: 개인정보보호위원회·보건복지부(2022). [2022년] 보건의료 데이터 활용 가이드라인. p. 8.

또한 정보 주체의 인권 및 사생활 보호에 중대한 피해를 야기할 수 있어 본인의 동의를 받아 사용하는 것을 원칙으로 하는 정보로는 [그림 4-

5]와 같이 5가지가 있다. 이 데이터가 필요한 경우 심의위원회에 그 사유와 인권을 보호할 특별한 보호조치 등의 실시계획을 보고한 뒤 승인을 득한 후 사용할 수 있다.

보건의료데이터의 기준은 「보건의료기본법」 제3조 제6호에 따른 보건의료정보(보건의료와 관련한 지식 또는 부호·숫자·문자·음성·음향·영상 등으로 표현된 모든 종류의 자료)로서 광(光) 또는 전자적 방식으로 처리될 수 있는 것에 적용되어야 한다. 한국보건사회연구원의 정보 중 대표적으로 한국의료패널조사의 가명처리 시 「보건의료데이터 활용 가이드라인」에 따라 가명처리를 수행해야 한다. 다만, 「보건의료데이터 활용 가이드라인」에는 가명처리 절차에 대해 「가명정보 처리 가이드라인」을 따르고 있어 이를 적용하면 되며, 적정성 검토의 경우 데이터 심의위원회를 구성해서 평가를 해야 한다.

10. 공공분야 가명정보 제공 실무안내서

「공공분야 가명정보 제공 실무안내서」 또한 「개인정보 보호법」 제3장 제3절 ‘가명정보 처리에 관한 특례’에 관한 사항을 근거로 하고 있으며, 특히 공공기관에서 가명정보 제도를 활용할 때 이 안내서 내용을 사용한다. 「공공분야 가명정보 제공 실무안내서」는 가명정보의 제공 단계별 처리사항과 공공분야에서 사용되는 신청서 및 통지서 등을 담고 있다.

「공공분야 가명정보 제공 실무안내서」의 특징은 최초 제정한 2021년 이후 현재까지 개정이 이루어지고 있지 않다는 점이다. 전체적인 절차는 「가명정보 처리 가이드라인」의 절차와 크게 다르지 않지만, 단계가 한 단계 더 적다. 그 이유는 「가명정보 처리 가이드라인」이 개정되면서 위험성 검토 부분을 강조하게 됐는데, 이 부분이 해당 안내서에는 반영되어 있지

않다. 다만, 절차적으로 다르지 않지만 적합성 검토 부분이 「공공분야 가명정보 제공 실무안내서」에서는 강조되어 있는 것이 특징이며, 이에 따라 한국보건사회연구원은 공공기관으로 적합성 검토를 필수적으로 수행해야 한다.

제2절 데이터 제공 관련 현황 분석

현재 한국보건사회연구원의 데이터 제공 절차를 검토하기 위해서 한국보건사회연구원의 다양한 지침 및 절차를 살펴보았다. 이 중에서 데이터 제공과 관련된 지침은 「정보공개 지침」과 「조사자료 관리 지침」이 해당하였다. 두 개 지침의 정의 및 개요, 특징을 살펴보았고, 이에 대한 검토 결과를 정리하였다.

또한, 데이터 제공 절차에 대해 살펴보았다. 특히 법령상 데이터 제공과 관련하여 정확한 요구사항을 가지고 있는 두 개의 국가승인 통계에 대한 데이터 제공 절차를 검토하였다. 한국복지패널조사와 한국의료패널조사는 데이터를 제공할 때 「통계법」에 따라 제공되어야 한다.

마지막으로 활발한 데이터 제공이 가능할 수 있도록 각종 양식에 대해 살펴보았고, 「생명윤리위원회 운영 규정」의 연구 참여 동의서, 개인정보 수집 및 이용 동의서에 대한 개선점을 도출하였다.

1. 정보공개지침

「정보공개지침」은 「정보공개법」을 그 근거로 하고 있으며, 외부 청구인이 한국보건사회연구원이 보유한 정보에 대해 정보를 공개해달라고 청구

하면 따르게 되는 지침이다. 「정보공개지침」은 목적과 용어 정의부터 적용 범위, 정보공개의 원칙, 정보공개의 의무, 사전정보공표, 정보공개 청구 절차, 정보공개 처리 절차, 정보공개심의위원회, 이의신청 등을 다루고 있다.

「정보공개지침」의 특징은 「정보공개법」을 준수하도록 하고 있어, 절차 및 방식 등이 「정보공개법」과 거의 동일하다. 정보공개 여부 결정을 10일 이내로 하는 것과 정보공개 검토를 정보공개심의위원회를 거치도록 하는 것도 모두 동일하다.

「정보공개지침」의 검토 결과, 「정보공개지침」은 「정보공개법」에 따라 수립된 지침으로 「정보공개법」의 내용을 충실하게 따르고 있다. 그러나 지침으로서 법령에 정의되지 않은 세부적인 절차에 대한 보완이 필요하며 다음과 같은 부분에 대해 현행 지침의 개정이 필요한 것으로 파악되었다.

- 개인정보가 포함된 정보의 제공에 대한 세부 절차
- 가명정보 제공을 위한 심의 절차의 반영

먼저, 개인정보가 포함된 정보의 제공의 경우 「정보공개법」 제9조 제1항 6호에 따라 공개하지 않을 수 있으나, 이를 가명정보로 처리하여 제공하는 것에 대해서는 「정보공개법」 제9조 제1항 제6호 가목에 따라 법령에서 정하는 바에 따라 열람할 수 있는 정보에 대해서는 제외하도록 하고 있어 「개인정보보호법」 제28조의 2에 따라 통계작성, 과학적 연구, 공익적 기록보존의 목적과 제28조의 3에 따라 통계작성, 과학적 연구, 공익적 기록보존 목적의 결합에 대해서는 제공이 가능할 것으로 판단되며, 이를 「정보공개지침」에 반영하는 것이 필요하다.

두 번째로, 이렇게 가명정보로 제공하는 경우, ‘공공분야 가명정보 제공 실무 안내서’의 가명정보 제공 신청 적합성 검토의 내용을 참조하여 ‘가명정보의 제공에 대한 적합성 심의’와 ‘가명정보처리 가이드라인’의 4

단계 적정성 검토의 내용을 참조하여 ‘가명처리된 정보의 적정성 검토를 수행’해야 하며 ‘이를 위한 절차가 수립’되어야 한다. 이러한 절차는 「정보공개법」에 의한 데이터 제공뿐 아니라 국가승인통계에 적용되는 「통계법」을 통한 제공 등 다양한 법률에 따른 ‘제공’에 적용되어야 한다. 이런 점을 고려할 때 다음과 같은 두 가지 방법을 적용할 수 있다.

- 1) 「정보공개법」에 따른 정보공개심의위원회와 다른 법령에 따른 모든 심의위원회의 세부 절차에 이를 포함하여 적용
- 2) 가명정보 제공에 대한 별도의 절차를 수립하여 이를 「정보공개지침」에서 준용

1)의 방법을 통해 적용할 경우 각종 법령에서 요구하는 모든 절차에 적합성 검토와 적정성 검토를 위한 절차를 수립하여야 하며, 이는 각 절차에 따른 위원회에서 모두 가명정보 제도에 대한 이해가 필요하다는 단점이 발생한다. 이에 따라 2)의 방법으로 적용하는 것이 더 효과적인 절차가 될 수 있다고 생각한다.

앞에서 살펴보았듯이, 현행 「정보공개지침」은 「정보공개법」의 개인정보 제공과 관련한 부분에서 「개인정보보호법」의 가명정보를 통한 제공에 대한 내용이 반영되어 있지 않다. 이에 따라 가명정보에 대한 내용을 반영하는 것이 필요하다.

2. 조사자료관리지침

「조사자료관리지침」은 한국보건사회연구원 내에서 수행되는 연구사업 수행을 목적으로 생성되는 조사자료에 대해 관리업무와 자료 이용 활성화를 목적으로 하는 지침이다. 「조사자료관리지침」은 목적과 용어 정의부터 기본

원칙, 적용 범위, 조사자료 관리위원회의 구성과 역할, 조사자료 관리 계획 수립 및 조사 수행, 조사자료의 제출, 보관, 제공, 이용 등을 다루고 있다.

「조사자료관리지침」의 특징은 한국보건사회연구원 내에서 연구사업을 위해 활용되는 자료에 대해서는 모두 해당 지침에 영향을 받게 되며, 이러한 자료의 공개, 제공과 관련한 세부적인 내용을 다루고 있다.

「조사자료관리지침」을 검토한 결과, 「조사자료관리지침」은 「정보공개법」을 제외한 모든 ‘제공’ 관련 법령을 적용하는 지침으로 다음과 같은 문제점을 가지고 있다.

- 「통계법」에 따른 국가승인통계의 제공에 대한 절차 미비
- 데이터 제공 비용 규정 없음
- 보관용 조사자료에 대한 명확한 기준이 없음

「통계법」의 데이터 제공과 관련 절차의 미비는 제1절 ‘2. 통계법의 내용’을 참조하여 보완이 필요하며, 데이터 제공 비용 규정의 경우 「공공데이터법」, 「데이터행정법」에 따른 보완이 필요하다. 마지막으로 보관용 조사자료에 대한 명확한 기준에 대해서는 「개인정보 보호법」의 가명정보를 통한 제공인지, 제58조의 2에 따른 제공인지에 대한 명확한 기준을 수립하는 것이 필요하다.

3. 데이터 제공 절차 검토

한국보건사회연구원의 데이터 제공은 다음과 같은 형태로 구분하여 제공되고 있다. 가족과출산조사와 같이 국가지정통계의 제공 형태, 한국복지패널조사와 한국의료패널조사와 같이 국가승인통계의 제공 형태, 가족과출산조사 외 91종과 같이 공공데이터 개방 형태가 있다. 이 중에서 한

국보건사회연구원이 승인받은 국가승인통계는 다음과 같이 3종(가족과 출산조사, 한국복지패널조사, 한국의료패널조사)으로 확인되었다. 또한 노인실태조사 등의 15종의 통계청 승인 조사와 30종의 미승인 조사를 제공하고 있다. 제공 절차에 대한 명확한 법률적 근거가 있는 국가승인통계에 대해 실제 제공되고 있는 절차와 법률적 요구사항을 검토하였다.

가. 한국복지패널조사의 제공 절차

한국복지패널조사의 제공 절차는 다음과 같다.

[그림 4-6] 한국복지패널조사 데이터 제공 웹 페이지



자료: 한국보건사회연구원. 한국복지패널 사이트(www.koweps.re.kr), 데이터 제공 페이지 2023. 12. 20. 인출

데이터의 다운로드를 위해 로그인을 요구하고 있으며, 로그인을 위한 회원 가입 시 개인정보 수집·이용·제공 동의와 이용약관에 대한 동의를 받고 있다. 이용약관에 대한 동의 후 회원가입을 하도록 하고 있으며, 회

원가입이 되면 위의 정보들을 다운로드할 수 있다. 그러나 회원 가입 시 주소는 수집하지 않고 있는데, 통계법 시행령에서는 통계이용자의 통계자료 신청 시 다음과 같은 정보를 적어 신청하도록 하고 있다.

통계법 시행령 제47조(통계이용자의 통계자료 신청과 제공) ① 법 제31조에 따라 통계자료를 이용하려는 자(이하 “통계이용자”이라 한다)는 다음 각 호의 사항을 문서(전자문서를 포함한다)에 적어 신청하여야 한다. <개정 2016. 7. 26.>

1. 통계이용자의 이름(기관 등인 경우에는 기관 등의 명칭) 및 주소
2. 통계자료의 명칭
3. 통계자료의 사용 목적
4. 통계자료의 내용 및 범위
5. 통계자료의 제공 방법
6. 통계자료의 보호 방법

데이터의 보호에 대한 부분은 이용약관 제4장 제1조(2)의 8에 연구원의 콘텐츠 무단 복제에 대한 제재 내용만을 다루고 있다. 따라서 적절한 보호조치에 대해 명시하고 있다고 하기에는 부족한 부분이 있어서, 이에 대한 보완이 필요하다.

나. 한국의료패널의 제공 절차

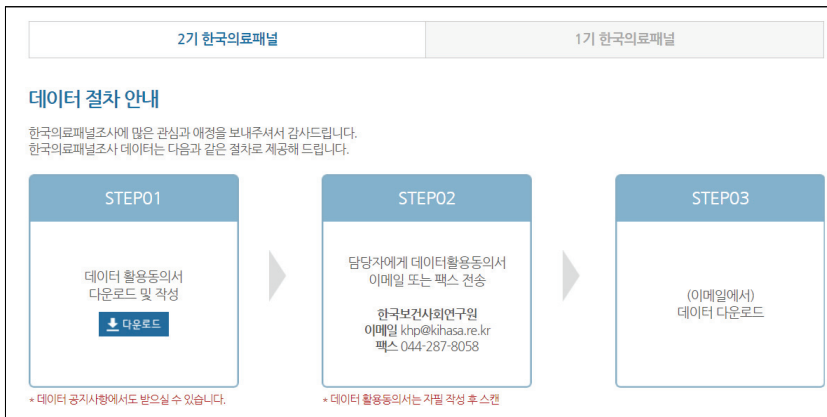
한국의료패널 조사는 <https://www.khp.re.kr:444/web/data/data.do>에서 다운로드할 수 있으며, 데이터 제공 절차는 [그림 4-7]과 같이 게시되어 있다.

데이터 활용 동의서는 [그림 4-8]과 같이 구성되어 있다.

데이터 활용 동의서의 내용을 보면, 데이터 자료 제공 요청 시 소속, 연락처, 이메일 등 6개의 정보를 수집하고 있다. 한국복지패널조사와 마찬가지로

가지로 통계이용자의 주소에 대한 추가적인 수집이 필요해 보인다. 또한, 통계자료의 보호 방법에 대해서는 연구 목적의 사용과 타인에게 양도 금지만을 명시하고 있어 충분한 보호 방법으로 보기에는 부족한 부분이 있다.

[그림 4-7] 한국의료패널조사 데이터 제공 절차 안내 페이지



자료: 한국보건사회연구원. 한국의료패널 사이트(<https://www.khp.re.kr>), 데이터 제공 절차 안내 페이지 2023. 12. 20. 인출

다. 생명윤리위원회 운영 규정

「생명윤리법」은 인간 대상 연구를 수행할 때와 이에 포함된 데이터를 제공하는 경우 원칙적으로 사용자의 동의를 요구하고 있으며, 동의서에 포함해야 하는 내용을 법률로 정의하고 있다.

생명윤리법 제16조(인간대상연구의 동의) ① 인간대상연구자는 인간대상연구를 하기 전에 연구대상자로부터 다음 각 호의 사항이 포함된 서면동의(전자문서를 포함한다. 이하 같다)를 받아야 한다.

1. 인간대상연구의 목적
2. 연구대상자의 참여 기간, 절차 및 방법
3. 연구대상자에게 예상되는 위험 및 이득
4. 개인정보 보호에 관한 사항
5. 연구 참여에 따른 손실에 대한 보상
6. 개인정보 제공에 관한 사항
7. 동의의 철회에 관한 사항
8. 그 밖에 기관위원회가 필요하다고 인정하는 사항

② 제1항에도 불구하고 동의 능력이 없거나 불완전한 사람으로서 보건복지부령으로 정하는 연구대상자가 참여하는 연구의 경우에는 다음 각 호에서 정한 대리인의 서면동의를 받아야 한다. 이 경우 대리인의 동의는 연구대상자의 의사에 어긋나서는 아니 된다.

1. 법정대리인

2. 법정대리인이 없는 경우 배우자, 직계존속, 직계비속의 순으로 하되, 직계존속 또는 직계비속이 여러 사람일 경우 협의하여 정하고, 협의가 되지 아니하면 연장자가 대리인이 된다.

③ 제1항에도 불구하고 다음 각 호의 요건을 모두 갖춘 경우에는 기관위원회의 승인을 받아 연구대상자의 서면동의를 면제할 수 있다. 이 경우 제2항에 따른 대리인의 서면동의를 면제하지 아니한다.

1. 연구대상자의 동의를 받는 것이 연구 진행 과정에서 현실적으로 불가능하거나 연구의 타당성에 심각한 영향을 미친다고 판단되는 경우
2. 연구대상자의 동의 거부를 추정할 만한 사유가 없고, 동의를 면제하여도 연구대상자에게 미치는 위험이 극히 낮은 경우

④ 인간대상연구자는 제1항 및 제2항에 따른 서면동의를 받기 전에 동의권자에게 제1항 각 호의 사항에 대하여 충분히 설명하여야 한다.

제18조(개인정보의 제공) ① 인간대상연구자는 제16조 제1항에 따라 연구대상자로부터 개인정보를 제공하는 것에 대하여 서면동의를 받은 경우에는 기관위원회의 심의를 거쳐 개인정보를 제3자에게 제공할 수 있다.

② 인간대상연구자가 제1항에 따라 개인정보를 제3자에게 제공하는 경우에는 익명화하여야 한다. 다만, 연구대상자가 개인식별정보를 포함하는 것에 동의한 경우에는 그러하지 아니하다.

[그림 4-8] 한국의료패널 자료활용동의서

한국의료패널 자료활용동의서

본 연구자는 한국의료패널 데이터를 사용함에 있어 다음의 사항에 대한 준수를 동의합니다.

☐ **데이터 명 및 버전**
 - 2019년~2020년 연간데이터(Version 2.1)

☐ **자료 활용 시 주의사항**

- 본 데이터를 활용한 연구결과를 게재하거나 출판 시, 연구방법이나 데이터 소개 부분에 “본 연구는 한국보건사회연구원과 국민건강보험공단이 공동으로 주관하는 한국의료패널 2019년~2020년 연간데이터(Version 2.1)를 활용하였다”와 같은 형식으로 데이터 주관기관, 데이터 명, 데이터 버전에 대해 명확하게 명기하겠습니다.
- 본 데이터를 연구목적으로만 사용하며, 개인 및 기관의 영리목적으로 사용하지 않겠습니다.
- 본 데이터를 타인에게 대여 또는 양도하지 않겠습니다.

만일 위의 사항을 위반하였을 시에는 관련 규정에 따라 어떠한 처벌도 감수할 것을 서약합니다.

년 월 일

소 속 : _____

연 락 처 : _____

이 메 일 : _____

활용목적 : _____

성 명 : _____ (서 명)

자료유형 : SAS()/SPSS()/STATA()

※ 본 데이터를 활용한 연구결과를 게재하거나 출판할 경우, 반드시 그 사실을 한국의료패널 담당자에게 통보하여 주시기 바랍니다.

자료: 한국보건사회연구원. 한국의료패널 사이트(<https://www.khp.re.kr>), 데이터 제공 절차 안내 페이지 2023. 12. 20. 인출

생명윤리법 제16조의 내용에 따라 연구 참여 동의서의 내용과 개인정보 수집 및 이용 동의서의 내용을 검토하였다. 연구 참여 동의서의 경우, [그림 4-9]와 같이 작성하도록 하고 있으나 이 부분이 현행 법률의 동의서의 내용과 일부 내용이 달라 수정이 필요하다.

또한 개인정보 이용 수집 동의서의 내용에도 가명정보의 활용에 대한 동의 내용을 추가할 필요가 있다.

[그림 4-9] 연구 참여 동의서 포함 항목

* 설명문은 개별 조사에 따라 적절하게 수정·변경하여 사용할 수 있으며, 아래 내용을 포함함.

- ① 연구 제목, 목적
- ② 연구자 성명, 소속기관 및 연락 담당자 관련 정보, 기관생명윤리위원회의 연락처
- ③ 연구비 지원기관
- ④ 연구대상자의 참여 기간, 절차 및 방법
- ⑤ 연구대상자 모집 방법(해당되는 경우)
- ⑥ 연구 참여로 인한 잠재적 위험과 이익
- ⑦ 연구대상자로부터 얻어지는 정보의 종류와 기밀성에 관한 사항
- ⑧ 연구 참여의 자발성과 참여 거부권, 철회권
- ⑨ 연구 참여에 대한 비용 및 보상
- ⑩ 기타 연구 수행과 관련된 문서

자료: 필자 작성

제3절 데이터 제공 절차 수립

제1절에서 파악된 법령상의 데이터 제공 절차와의 차이점을 해결하기 위해 한국보건사회연구원의 데이터 제공 절차를 수정·보완하였다. 이에 따라 ‘익명·가명 데이터 제공 절차’를 수립하였다. 익명·가명 데이터 제공 기본 절차를 수립한 다음에, 법령별³⁴⁾ 제공 절차가 상이하여 각 법령에 대한 제공 절차도 함께 수립하였다. 한편, 개인정보보호법에 따른 데이터 결합 절차도 수립하였다.

이때 법령별로 데이터 제공 절차도를 작성하였고, 데이터 제공 단계별로 구분하여 해당 내용을 표로 정리하였다. 데이터 제공 단계는 신청, 검토, 공개 결정 시, 종료 등으로 구성되나, 해당 법령의 제공 단계에 맞춰 살펴보았다.

1. 익명·가명 데이터 제공 기본 절차

익명·가명 데이터를 제공하는 기본 절차를 [그림 4-10]과 같이 수립하였다. 데이터 제공을 위한 기본 절차는 각 법령에 따라 데이터 제공에 대한 신청이 접수되면 데이터제공심의위원회를 개최한다. 데이터제공심의 위원회에서는 각 법령의 절차에 따라 제공 여부를 결정하고, 데이터 제공 형태(익명, 가명)에 따라 익명처리, 가명처리를 적용한 다음에 처리된 데이터를 제공하도록 절차를 수립하였다. 이는 각 법령의 내용과 데이터 제공의 형태에 따라 모두 대응할 수 있는 절차라고 볼 수 있다.

34) 「공공기관의 정보공개에 관한 법률」, 「통계법」, 「공공데이터법」, 「데이터기반행정법」, 「개인정보보호법」, 「생명윤리법」

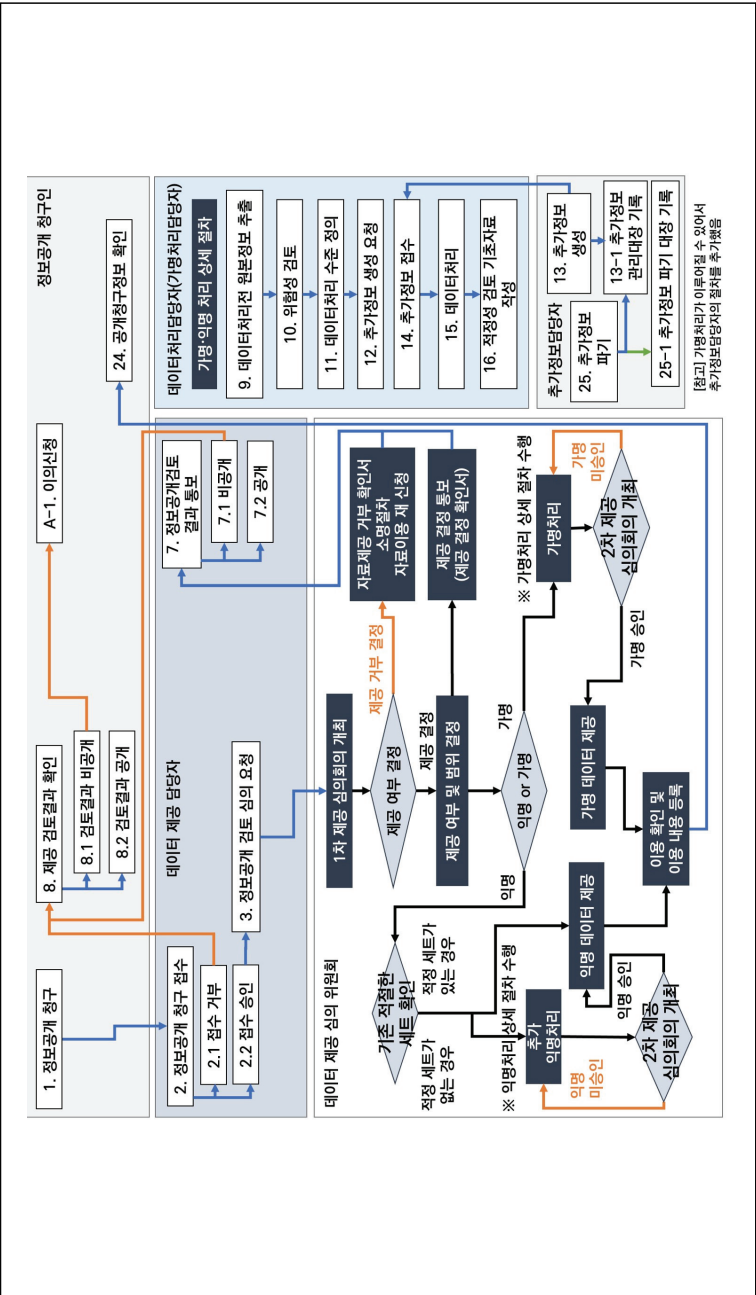
2. 정보공개법에 따른 제공 절차

〈표 4-2〉 정보공개법에 따른 데이터 제공 단계별 설명

단계	설명
신청	- 청구인은 사전공개된 정보는 통합정보공개시스템을 통해 확인할 수 있음 - 청구인은 다음과 같은 내용을 포함한 정보공개 청구서를 제출하거나 말로써 정보의 공개를 청구할 수 있음 1. 청구인의 성명·생년월일·주소 및 연락처(전화번호·전자우편주소 등). 청구인이 법인 또는 단체인 경우 그 명칭, 대표자의 성명, 사업자등록번호 또는 이에 준하는 번호, 주된 사무소의 소재지 및 연락처 2. 청구인의 주민등록번호 3. 공개를 청구하는 정보의 내용 및 공개방법 - 말로 신청하는 경우 담당공무원 등이 청구인의 진술을 바탕으로 정보공개 청구조서를 작성하여 청구함 - 공공기관은 청구인의 정보공개청구를 접수
검토	- 공개기관은 법 제12조에 따라 정보공개심의회를 구성하고 공개 여부를 검토함 - 검토 결과는 요청받은 날부터 10일 이내 결정. 단, 해당 기간 내 결정하지 못하면 기간 만료일 다음 날부터 기산하여 10일 이내의 범위에서 기간 연장할 수 있음. 기간 연장 시 연장된 사실과 연장 사유를 청구인에게 지체 없이 문서로 통지하여야 함
비공개 또는 부분 공개 결정 시	- 청구인에게 비공개 또는 부분 공개 결정에 따른 청구 종결의 내용과 사유, 불복의 방법 및 절차를 지체 없이 문서로 통지 [이의신청] - 청구인은 공개 여부 결정 통지를 받은 날 서면으로 이의신청 가능 - 정보공개 청구 후 20일이 경과하도록 공개 결정이 없는 경우, 정보공개 청구 후 20일이 경과한 날부터 30일 이내 서면으로 이의신청 가능 [행정심판] - 청구인이 정보공개 결정에 대하여 불복이 있거나 정보공개 청구 후 20일이 경과하도록 결정이 없을 경우 이의신청 없이 행정심판법에 따라 청구 가능 [행정소송] - 청구인이 정보공개 결정에 대하여 불복이 있거나 정보공개 청구 후 20일이 경과하도록 결정이 없을 경우 행정소송법에 따라 행정소송 가능
공개 결정 시	1. 공개의 일시 및 장소 등을 청구인에게 통지 2. 타인의 지식재산권, 사생활의 비밀, 그 밖에 타인의 권리 또는 이익이 부당하게 침해되지 않는 정보에 한해 청구인의 요청에 따라 공개 3. (부분공개)공개 청구한 정보가 비공개 정보와 공개 가능한 부분이 혼합된 경우로 공개 청구 취지에 어긋나지 않는 범위에서 분리할 수 있는 경우 분리 공개
종료	- 공개 결정 이후 기본 제공 절차에 따라 처리 후 종료

자료: 법률에 따라 정리하여 필자 작성

[그림 4-11] 정보공개법에 따른 데이터 제공 절차도



자료: 필자 작성

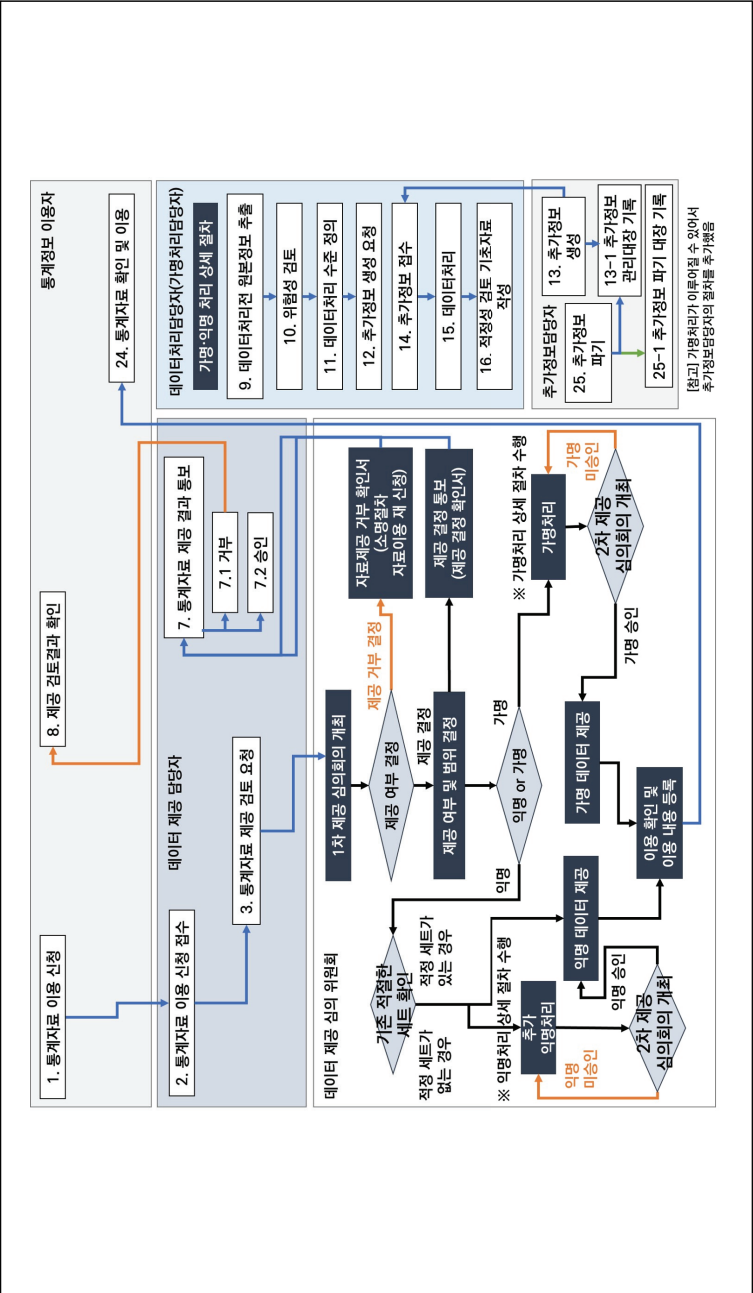
3. 통계법에 따른 제공 절차

〈표 4-3〉 통계법에 따른 데이터 제공 단계별 설명

단계	설명
신청	- 이용자(요청기관을 포함)는 시행령 제46조 및 제47조에 따른 이용 목적, 이용 자료, 이용방법, 이용기간 등 구체적인 사항을 기재하여 문서(전자문서 및 자료 제공시스템을 통한 신청을 포함)로 신청 - 통계작성기관(이하 “제공기관”이라 한다)은 이용자의 신청을 접수
검토	- 제공기관은 시행령 제49조에 따른 통계자료제공심의회를 통해 신청접수된 통계 자료의 이용 목적·내용 및 범위의 타당성을 심사함(통계자료제공심의회의 조직 구성·운영에 관한 사항은 제공기관의 장이 정함) - 검토 결과는 30일 이내 통보. 단, 해당 기간 내 결정하지 못하면 이용자와 협의 해 10일 이내 범위에서 연장할 수 있음
제공 거부 시	- 제공 거부 결과 통보
제공 결정 시	- 제공 승인 결과 통보 - 전자매체, 복사·출력물 등 이용자가 지정한 방법으로 제공 - 특정의 개인이나 법인 또는 단체 등을 식별할 수 없는 형태로 통계자료를 처리한 후 제공. 단, 불가피한 경우, 필요하다고 인정되거나 이용자가 지정한 방법이 적절하지 않을 경우 등 정당한 사유가 있으면 제공기관의 장은 이용자와 협의하여 다른 방법으로 제공할 수 있음
종료	- 제공 결정 이후 기본 제공 절차에 따라 처리 후 종료

자료: 법률에 따라 정리하여 필자 작성

[그림 4-12] 통계법에 따른 데이터 제공 절차도



자료: 필자 작성

4. 공공데이터법에 따른 제공 절차

〈표 4-4〉 공공데이터법에 따른 데이터 제공 단계별 설명

단계	설명
신청	- 이용자(요청기관을 포함)는 공표된 공공데이터는 공공데이터 포털을 통해 제공 받을 수 있음 - 공표 제공 대상 외의 공공데이터 제공 신청을 위해 별도 제공 신청서(전자문서로 된 신청서를 포함)를 작성해 제공기관 또는 활용지원센터에 직접 제출하거나 정보통신망을 통하여 제출 - 활용지원센터가 제공 신청을 받을 경우 이를 즉시 해당 제공기관에 이첩해야 함
검토	- 제공기관은 법 제27조 제3항에 따라 제공기관의 장이 제공 여부를 검토함 - 검토 결과는 요청받은 날부터 10일 이내 결정. 단, 해당 기간 내 결정하지 못하면 기간 만료일 다음 날부터 기산하여 10일 이내의 범위에서 기간 연장할 수 있음. 기간 연장 시 연장된 사실과 연장 사유를 신청인에게 지체 없이 통보하여야 함
제공 거부 시	- 이용자와 행정안전부장관에게 제공 거부 결정의 내용과 사유를 통보 - 이용자는 제공 거부 통보일부터 60일 이내 공공데이터분쟁조정위원회에 분쟁조정을 신청할 수 있음 - 분쟁조정 신청 시 분쟁조정위원회는 신청일부터 30일 이내에 심사하여 조정안을 제공기관에 전달해야 함 - 제공기관은 특별한 사유가 없는 한 분쟁조정안에 따라야 하며, 조정안에 따를 수 없는 특별한 사유가 있는 경우에는 분쟁조정위원회에 소명해야 함 - 조정안에 따라 제공을 결정하면, 지체 없이 이용자에게 제공하고, 공공데이터 포털에 등록해
제공 결정 시	- 제공 승인 결과 통보 - 제공 방법·절차 등을 이용자에게 통지 - 제공공데이터 목록 및 제공 공공데이터를 목록등록관리시스템에 등록 ※ 제공기관의 장은 기관의 웹사이트 등을 통해 이미 공개된 데이터는 특별한 사유가 없으면 공공데이터 포털에도 등록해야 한다.(공공데이터 지침 제10조 제4항)
종료	- 제공 결정 이후 기본 제공 절차에 따라 처리 후 종료

자료: 법률에 따라 정리하여 필자 작성

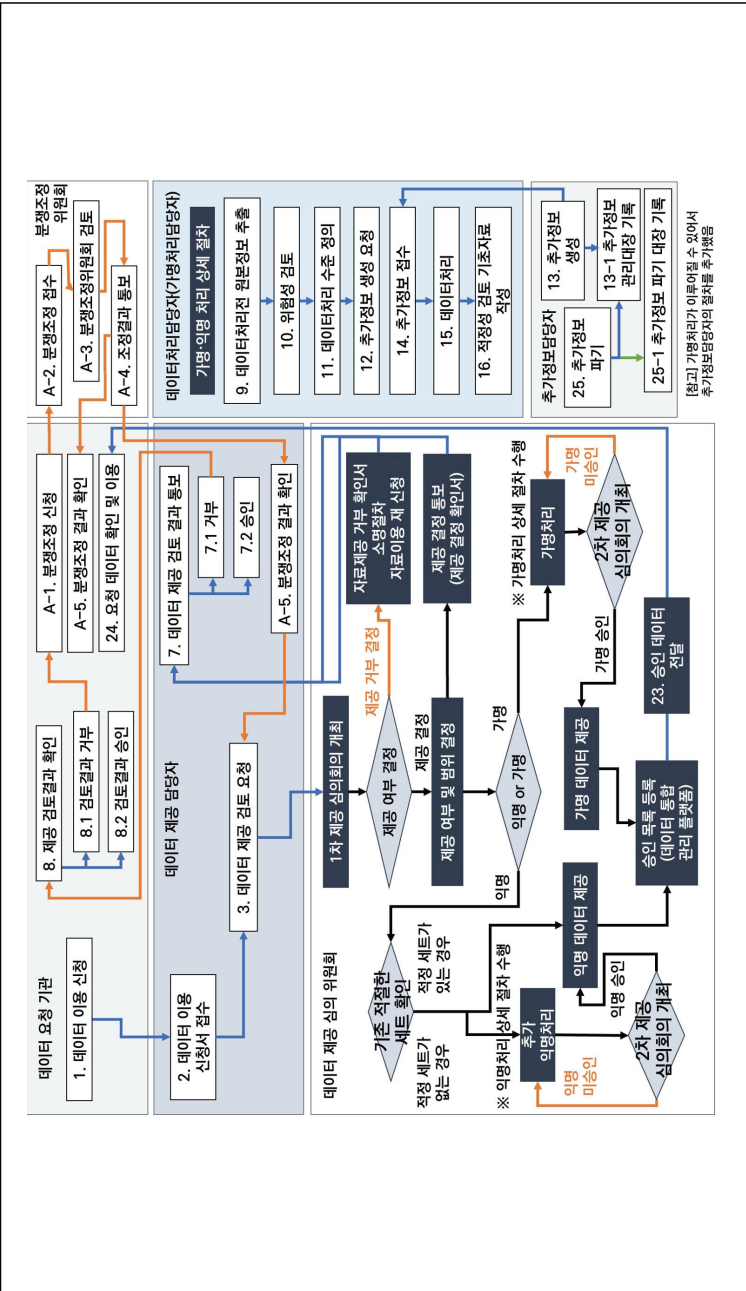
5. 데이터기반행정법에 따른 제공 절차

〈표 4-5〉 데이터기반행정법에 따른 데이터 제공 단계별 설명

단계	설명
신청	- 요청기관의 장은 데이터통합관리 플랫폼을 통해 제공받을 수 있음 - 데이터통합관리 플랫폼에 등록되지 않은 데이터 제공 신청을 위해 별도 제공 요청서(전자문서를 포함)를 작성해 제공기관에 제출
검토	- 제공기관은 법 제12조 제1항에 따라 제공기관의 장이 제공 여부를 검토함 - 검토 결과는 요청받은 날부터 20일 이내 결정. 단, 해당 기간 내 결정하지 못하면 10일 이내의 범위에서 기간을 1회 연장할 수 있음. 기간 연장 시 연장된 사실과 연장 사유를 요청기관의 장에게 통보하여야 함
제공 거부 시	- 요청기관의 장에게 제공 거부 결정(법 제11조 제2항에 따른 기술적 분리 가능 여부 검토 결과 포함)의 내용과 사유를 통보 - 요청기관의 장은 제공 거부 통보일부터 60일 이내에 데이터기반 행정 활성화 위원회(개정안 : 공공데이터분쟁조정위원회, 이하 '분쟁조정위원회'라 한다)에 분쟁조정을 요청할 수 있음 - 분쟁조정 신청 시 분쟁조정위원회는 신청일부터 30일 이내 당사자 간 합의를 권고 - 합의가 이루어지면 조정절차 종결 - 합의가 이루어지지 않으면, 그 사실을 확인한 날부터 30일 이내 위원회의 의결을 거쳐 조정 요청에 대한 결과 확정. 단, 부득이하게 해당 기간 내 결과를 확정할 수 없는 경우 위원회의 의결을 거쳐 30일의 범위에서 그 기간을 1회 연장할 수 있음 - 조정 요청에 대한 결과가 확정되면 조정 결과 통보서를 요청기관의 장에게 전달 - 제공기관은 특별한 사유가 없는 한 분쟁조정안에 따라야 하며, 조정 결과에 따를 수 없는 특별한 사유가 있는 경우에는 분쟁조정위원회에 소명해야 함
제공 결정 시	- 제공 승인 결과를 요청기관의 장에게 통보 ※ 행정안전부장관은 제공기관의 장이 데이터의 제공을 결정한 경우 데이터 통합관리 플랫폼에 해당 데이터를 등록하도록 요청할 수 있음
종료	- 제공 결정 이후 기본 제공 절차에 따라 처리 후 종료

자료: 법률에 따라 정리하여 필자 작성

[그림 4-14] 데이터기반행정부에 따른 데이터 제공 절차도



6. 개인정보보호법에 따른 제공 절차

이 절차는 내부 활용과 제3자 제공에 있어서 모두 동일한 절차를 가진다.

〈표 4-6〉 개인정보보호법에 따른 데이터 제공 단계별 설명

단계	설명
신청	- 가명정보 활용을 위해 요청기관에 대상 정보 신청 - 제공기관은 신청 접수 - 제공기관 내 적합성 검토위원회를 통한 적합성 검토
처리	- 적합성 검토 결과 승인일 때 원본 추출 및 위험성 검토 수행 - 위험성 검토 후 가명처리 대상의 처리 수준 정의 및 위험성 검토 결과보고서 작성 - 처리 수준 정의를 토대로 가명처리 수행
적정성 검토	- 처리 데이터의 적정성 검토를 위해 기초자료 작성 - 적정성 검토위원회 개최 및 검토 수행
부적정 시	- 최초 신청 단계로 회귀
조건부 승인 시	- 문제의 단계로 다시 회귀 후 추가처리
승인 시	- 내부 활용 또는 결합 절차 수행
종료	- 제공 FLOW 종료

자료: 법률에 따라 정리하여 필자 작성

7. 생명윤리법에 따른 제공 절차

〈표 4-7〉 생명윤리법에 따른 데이터 제공 단계별 설명

단계	설명
신청	- 보건의료데이터를 이용하려는 이용자는 사전에 IRB를 통해 연구계획에 대한 심의를 받아야 함 - IRB 승인이 이루어지고 난 후 데이터 제공 신청을 할 수 있음
검토	- 데이터제공심의위원회를 통해 제공 데이터에 대한 적합성 심의를 시행함
종료	- 제공 결정 이후 기본 제공 절차에 따라 처리 후 종료

자료: 법률에 따라 정리하여 필자 작성

8. 개인정보보호법에 따른 결합 절차

결합 절차는 크게 내부 결합, 개인정보보호법에 따른 결합전문기관의 결합 절차, 신용정보법에 따른 데이터 전문기관의 결합 절차로 나누게 되며, 각 절차는 [그림 4-17]과 같다.

가. 내부 결합 절차

나. 결합전문기관 결합 절차

결합전문기관의 결합 절차는 기본 데이터 제공 절차에 결합을 위한 절차(연두색)를 추가하여 절차를 구성하였으며, 데이터의 제공 및 이용 형태에 따라 [그림 4-18]과 같이 적용하여야 한다.

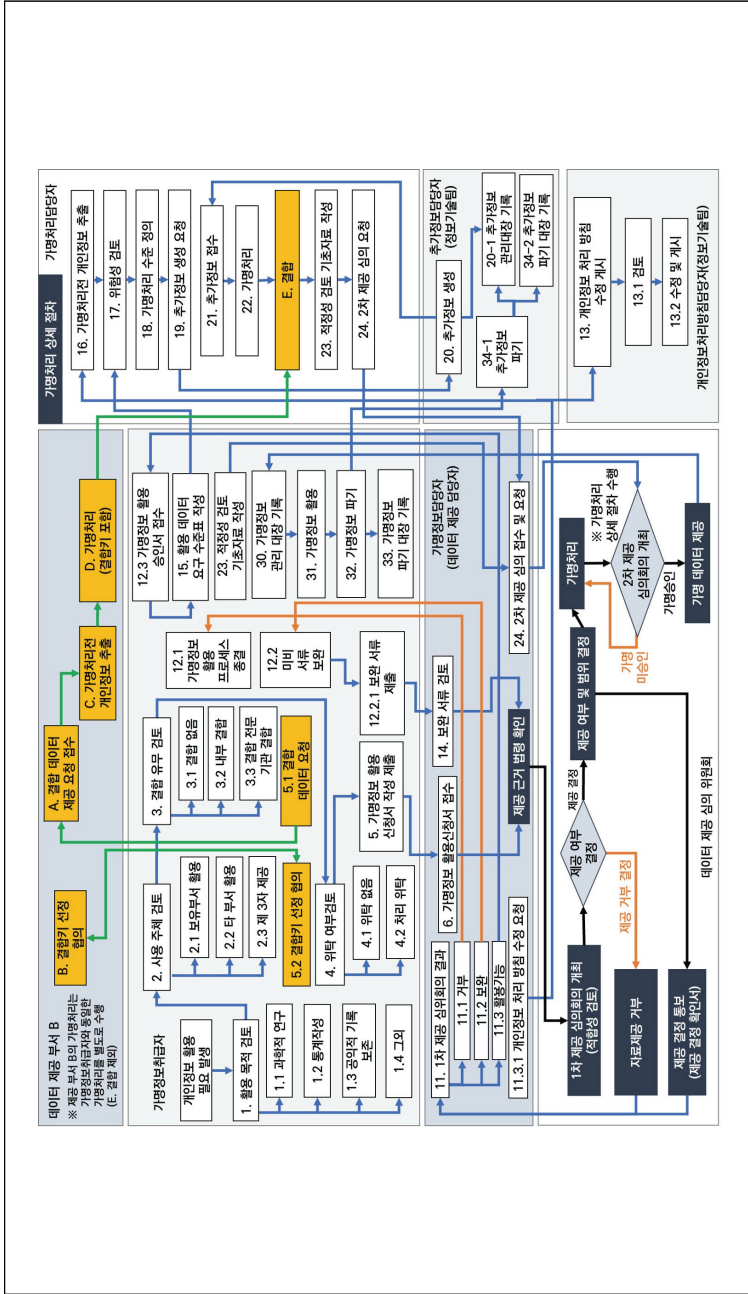
결합 대상 정보의 비대칭으로 인해 추출결합을 하는 경우에는 주황색의 절차(K. 추출, K.1. 추출 정보 접수, K.2. 추출정보 기반 개인정보 추출)를 모두 추가하여 적용하여야 한다.

〈표 4-8〉 결합정보 제공과 이용 형태에 따른 해당 절차 - 결합전문기관 결합 절차

구분	결합정보 제공과 이용 형태	해당 절차
1	결합정보의 제공과 이용을 함께 하는 경우	모든 절차를 진행함
2	결합정보를 제공만 하는 경우	최초 절차부터 가명정보 승인 후 결합 데이터 전송까지만 진행함
3	결합정보를 이용만 하는 경우	결합신청서를 작성하고, 결합신청 이후의 절차만 진행함

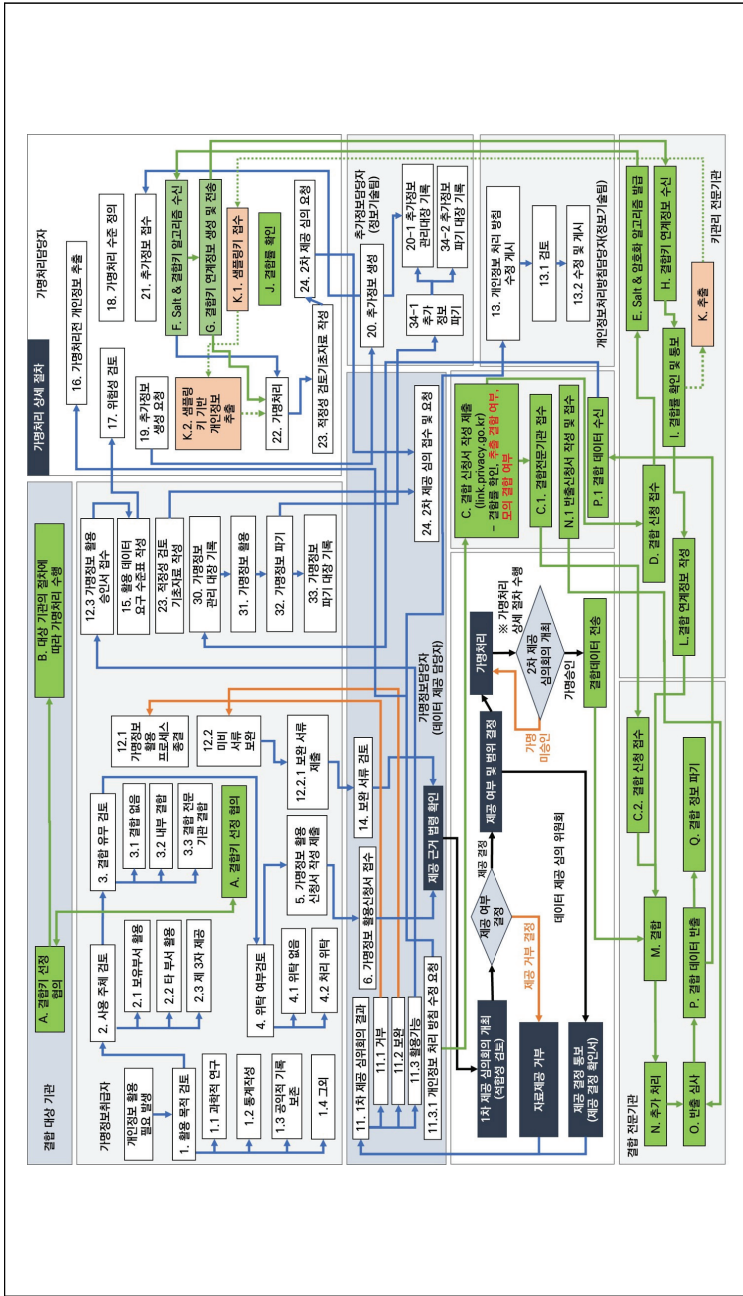
자료: 법률에 따라 정리하여 필자 작성

[그림 4-17] 개인정보보호법에 따른 데이터 제공 절차도 - 내부 결합



자료: 필자 작성

[그림 4-18] 개인정보보호법에 따른 데이터 제공 절차도 - 결합전문기관 결합



자료: 필자 작성

다. 데이터 전문기관 결합 절차

데이터 전문기관의 결합 절차는 기본적인 데이터 제공 절차에 결합을 위한 절차(연두색)를 추가하여 절차를 구성하였으며, 데이터의 제공 및 이용 형태에 따라 [그림 4-19]와 같이 적용하여야 한다.

샘플링 결합을 수행하는 경우에는 주황색의 절차(G.1 샘플링 결합키 추출, G.2. 샘플링 키 전달, L.1. 샘플링키 접수, L.2. 샘플링키 기반 개인 정보 추출)를 모두 추가하여 적용하여야 한다.

〈표 4-9〉 결합정보 제공과 이용 형태에 따른 해당 절차 - 데이터 전문기관 결합 절차

구분	결합정보 제공과 이용 형태	해당 절차
1	결합정보의 제공과 이용을 함께 하는 경우	모든 절차를 진행함
2	결합정보를 제공만 하는 경우	최초 절차부터 가명정보 승인 후 결합 데이터 전송하고 데이터 설명회까지만 진행함
3	결합정보를 이용만 하는 경우	결합신청서 작성 이후 절차만 진행함

자료: 법률에 따라 정리하여 필자 작성

제4절 데이터 제공 절차 개선 방안

익명·가명 데이터 제공 절차를 수립하면서 데이터 제공을 위한 운영 지침, 역할 및 책임(Role & Responsibility, R&R)도 마련하였다. 「통계법」에 근거하여 한국복지패널조사와 한국의료패널조사의 데이터 제공 절차에 대한 보완 사항도 제안하였다.

1. 데이터 제공을 위한 운영 지침 수립

데이터 제공을 위한 운영 지침을 수립하였다. 데이터 제공 운영 지침은 한국보건사회연구원의 내부 지침에 해당하여 자세한 내용은 이 보고서에서 포함하지 않는다. 「데이터 제공 운영 지침」은 총 3장, 27조로 구성되어 있으며, 데이터 제공 체계 운영을 위한 일원화된 데이터 제공 지침이라고 볼 수 있다.

각 장의 구성은 다음과 같다. 제1장은 지침의 목적부터 데이터 처리의 원칙을 명시하고 있으며, 제2장에서는 데이터 제공을 위한 담당 조직 구성과 조직 및 역할, 운영 절차 등에 대해 다루고 있다. 마지막 3장에서는 데이터 제공을 위한 검토위원회의 설치 근거와 구성, 역할, 운영 절차 등을 명시하고 있다.

2. 데이터 제공을 위한 역할과 책임 수립

데이터 제공 체계를 위해 한국보건사회연구원의 현행 조직 구성원과 인터뷰를 실시하였다. 인터뷰 분석 결과를 기반으로 하여 데이터 제공 체계를 위한 역할 및 책임(R&R)에 따라 담당 부서를 <표 4-10>과 같이 배

치하도록 제안하였다.

〈표 4-10〉 데이터 제공 체계를 위한 역할 및 책임에 따른 담당 부서 배치

역할 및 책임	담당 부서
가명처리 총괄	- (1안) 내부 전담 총괄부서 구성 - (2안) 임시 전담 부서 조직(Task Force: TF) 구성 - (3안) 각 데이터 보유 부서에서 개별 구성
심의위원회 운영	- 총괄부서 (*)
가명정보 내부관리계획 수립 및 운영	- 정보기술팀
개인정보처리방침 개정	- 정보기술팀
위험성 검토	- 총괄부서 (*)
가명처리	- 총괄부서 (*) - 가명처리 담당자는 가명정보 담당자와 분리
가명정보 보호 총괄	- 이용부서 - 관리감독은 정보기술팀
추가정보 관리	- 정보기술팀
가명정보 관련 기록 및 보관	- 이용부서 - 관리감독은 정보기술팀
가명정보 파기 총괄	- 이용부서 - 관리감독은 정보기술팀
가명정보 취급자 교육	- 정보기술팀

주: (2안)으로 조직을 운영 시 ‘총괄부서’는 ‘임시 전담 부서’를 의미함.
자료: 필자 작성

이때 ‘가명처리 총괄’의 ‘담당 부서’ 조직 구성은 3개 방안으로 제시하였다(〈표 4-11〉 참조). 1안은 내부 전담 총괄부서이고, 2안은 임시 전담 부서이고, 마지막 3안은 각 데이터 보유 부서이다.

3개 방안에 대한 장·단점은 〈표 4-12〉와 같다. 한국보건사회연구원의 현재 상황과 여건을 고려했을 때, 2안이 적합하다고 판단되어, 최종 2안으로 결정하였다. 우선 2안으로 조직을 운영하되, 향후 데이터 제공 건수가 많아지면 1안으로 조직을 운영하는 것이 효율적이라고 생각한다.

〈표 4-11〉 가명처리 총괄의 담당 부서 조직 구성 방안

구분	방안
1안	내부 전담 총괄부서 조직 및 운영
2안	임시 전담 부서 조직 및 운영
3안	각 데이터 보유 부서에서 개별 운영

자료: 필자 작성

〈표 4-12〉 3개 안의 장·단점

구분	장점	단점
1안	- 모든 법령에 따른 데이터 제공 시 제공 수준의 일관성 보장 - 데이터 제공에 필요한 기술적 전문성(가명처리 등) 보장함 - 데이터 제공 건수 증가 시 제공 시간 단축 등 효율성 증가함 - 데이터 제공에 대한 통합관리가 가능하여 동일 기관에서 다수 제공 시 안전성 검토 가능함 - 전문성을 통한 각 데이터 제공 당 소요 비용이 2안과 3안 보다 감소함	- 데이터 제공이 적은 경우 전담 인력의 효율성 저하 - 데이터 제공 관련 다양한 법제에 대한 이해 및 학습 기간 필요함
2안	- 데이터 제공에 대한 전문성(가명처리 등) 습득 가능함 - 데이터 제공 건수가 증가하여 1안으로 변경되는 경우, 해당 인력의 재배치로 전문성을 유지할 수 있음 - 데이터 제공에 대한 통합관리가 가능하여 동일 기관에서 다수 제공 시 안전성 검토 가능함 - 전문성을 통한 각 데이터 제공당 소요 비용이 3안보다 감소함	- 데이터 제공이 증가하는 경우, 1안으로 전환되기 전까지 업무 강도 증가함 - 임시 전담 부서로 결정사항 필요시 데이터 제공의 지연 가능성 존재함
3안	- 개별 법령에 따른 데이터 제공에 대한 전문성 유지함 - 개별 업무 영역으로 데이터 제공 유지 가능함	- 데이터 제공 정책의 일관된 적용이 어려움 - 데이터 제공에 대한 전문성(가명처리 등) 유지 불가함 - 데이터 제공 시 소요비용 증가함

자료: 필자 작성

3. 데이터 제공 절차 보완

「통계법」에 따라 데이터를 제공하는 경우, 다음과 같이 법률에서 요구하는 정보를 수집해야 한다. 이에 따라 한국복지패널조사와 한국의료패널조사는 데이터 제공 절차를 보완해야 할 필요가 있다.

먼저, 한국복지패널조사의 경우 이용 약관을 다음과 같이 변경한다. 제2조 이용자의 의무에서 (5)번을 추가하는 것을 제안한다. 그리고 회원 가입 시 입력받는 정보에 통계이용자의 주소를 추가할 필요가 있다(통계법 시행령 제47조에 근거함).

〈표 4-13〉 제2조 이용자의 의무에서 수정·보완 사항

제2조 이용자의 의무
(1) 아이디와 비밀번호에 관한 모든 관리의 책임은 이용자에게 있습니다. (2) 자신의 아이디가 부정하게 사용된 경우, 이용자는 반드시 연구원에 그 사실을 통보해야 합니다. (3) 이용자는 이 약관 및 관계 법령에 규정한 사항을 준수하여야 합니다. (4) 한국복지패널에서 제공하는 데이터를 내려받기 위해서, 이용자는 사용목적 및 연구원이 요구하는 정보들에 대하여 올바르게 응답해야 합니다. (5) 한국복지패널에서 제공하는 데이터를 사용하는 동안 절대로 개인을 재식별하기 위한 행위를 하지 않으며, 우연하게 재식별되는 경우 즉시 처리를 중지하고 이를 한국보건사회 연구원에 알리고 데이터를 파기하겠습니다. (6) 만약 위의 내용을 어기는 경우 이로 인해 발생하는 모든 형사적/민사적 책임을 질 것을 서약합니다.

자료: 필자 작성

다음으로, 한국의료패널조사 데이터의 자료 활용 동의서 중 ‘자료 활용 시 주의사항’에서 〈표 4-14〉의 내용을 수정·보완하는 것을 제안한다. 그리고 연구자의 소속, 연락처 등을 수집하고 있는데, 추가로 주소도 필요하다.

〈표 4-14〉 자료 활용 시 주의사항에 대한 수정·보완 사항

자료 활용 시 주의사항
- 본 데이터를 연구 목적으로만 사용하며, 개인 및 기관의 영리 목적으로 사용하지 않겠습니다. - 본 데이터를 사용할 때 개인을 재식별하는 행위를 절대 하지 않으며 우연히 개인이 식별되는 경우 즉시 처리를 중지하고 이를 한국보건사회연구원에 알리고 데이터를 파기하겠습니다. - 본 데이터를 타인에게 대여 또는 양도하지 않겠습니다. 만일 위의 사항을 위반하였을 시에는 관련 법령, 규정에 따라 모든 형사상 민사상 처벌을 감수할 것을 서약합니다.
자료: 필자 작성

4. 데이터 제공 양식 보완

「생명윤리법」 따라 데이터를 제공하는 경우, 연구 참여 동의서에 요구하는 양식이 있으며, 현재 일부 맞지 않는 부분이 있다. 다음 [그림 4-20]과 같이 ⑪ 개인정보 보호에 관한 사항과 ⑫ 개인정보 제공에 관한 사항을 추가하여 연구 참여 동의서를 보완하는 것을 제안한다. 또한, 개인정보 수집 및 이용 동의서의 경우 「개인정보보호법」에 따른 가명정보의 활용이 가능하도록 동의 여부 항목을 추가할 필요가 있다.

[그림 4-20] 데이터 제공 양식 보완-연구 참여 동의서

연구 참여 동의서

1. 본인은 연구 목적, 참여기간, 절차 및 방법에 대해 충분한 설명을 들었으며, 추후 조사 내용을 한국보건사회연구원에서 연구 목적으로 이용하는 것에 동의합니다.

2. 본인은 원하지 않을 경우 언제든지 연구 참여를 철회할 수 있으며, 이에 따른 어떠한 불이익도 본인에게 없음을 이해합니다.

3. 본인의 개인식별정보는 개인정보보호법에서 요구하는 안전성 확보조치가 적용되고 모든 정보는 엄격히 관리되는 것을 알고 있습니다.

4. 가명/익명화된 조사정보는 연구 목적으로 제3자에게 제공될 수 있음을 알고 있습니다.

5. 본인은 설명자로부터 설명문 및 동의서 내용에 대한 자세한 설명을 들었고, 이에 대해 잘 이해하였으므로 자발적으로 이 연구에 참여하는 것을 동의합니다.

6. 본인은 설명문과 함께 동의서 사본 또는 확인서를 받을 것을 알고 있습니다.

이 연구에 대하여 궁금한 점이 있으시면 언제든지 연구책임자(000-000-0000)에게 문의하시고, 연구 진행상의 윤리적 측면이나 연구대상자로서 권리에 대한 질문이 있으시면 한국보건사회연구원 생명윤리위원회(000-000-0000)로 문의하시기 바랍니다.

연구대상자 성명 서명 날짜 (년/월/일)

상담자 성명 서명 날짜 (년/월/일)

법정대리인 성명 서명 날짜 (년/월/일)

※ 동의서는 반드시 설명문과 함께 이용하여야 하며, 위원회 승인 후 이용 가능

* 설명문은 개별 조사에 따라 적절하게 수정·변경하여 사용할 수 있으며, 아래 내용을 포함함.

① 연구 제목, 목적

② 연구자 성명, 소속기관 및 연락 담당자 관련 정보, 기관생명윤리위원회의 연락처

③ 연구비 지원기관

④ 연구대상자의 참여 기간, 절차 및 방법

⑤ 연구대상자 모집 방법(해당되는 경우)

⑥ 연구 참여로 인한 잠재적 위험과 이익

⑦ 연구대상자로부터 얻어지는 정보의 종류와 기밀성에 관한 사항

⑧ 연구 참여의 자발성과 참여 거부권, 철회권

⑨ 연구 참여에 대한 비용 및 보상

⑩ 기타 연구 수행과 관련된 문서

⑪ 개인정보 보호에 관한 사항

⑫ 개인정보 제공에 관한 사항

자료: 필자 작성

자료: 필자 작성

5. 데이터 제공 절차 개선 단계

이 절에서 제안한 한국보건사회연구원의 데이터 제공 절차에 대한 개선 사항을 한 번에 모두 적용하는 것은 어려울 수 있다. 이에 따라 <표 4-15>와 같이 적용 시기를 고려하여 단계별로 개선 내용을 적용하는 방식으로 운영하는 것이 적합하다고 생각한다.

<표 4-15> 단계별 개선 내용 및 적용 시기

단계	개선 내용	적용 시기
1	- 데이터 제공 지침 수립함 - 승인통계 제공 절차 변경함 · 한국복지패널조사, 한국의료패널조사 데이터 제공 시 제공자의 주소 수집함 - 데이터 제공 양식 보완함 · 「생명윤리법」의 동의서 양식 보완함	- 즉시
2	- 데이터 제공 TF 조직 구성 및 운영함	- 1단계 완료 후 3개월 이내
3	- 데이터 제공 전담 조직 구성 및 운영함	- 데이터 제공이 월 8회 이상 발생 시

자료: 필자 작성

제5절 소결

이 장에서는 한국보건사회연구원의 데이터 제공에 적용되는 7가지 법령, 3가지 가이드라인에 따라 데이터를 제공하는 통합 절차를 수립하여 일원화된 데이터 제공 체계를 갖출 수 있는 기반을 마련하였다. 이러한 절차는 한국보건사회연구원의 다양한 데이터 제공을 일원화하여 데이터의 빠른 제공과 함께 안전한 제공을 위한 기반이 된다. 특히 데이터제공심의위원회라는 일원화된 의사결정기구를 통해 다양한 데이터의 제공에

대해 통합적인 관리를 할 수 있게 되었으며, 데이터 제공으로 인한 개인에 대한 식별 등의 다양한 위험에 대해 효과적으로 고민하고 대응할 수 있게 되었다.

이러한 개선은 한 번에 모두 이뤄질 수 없으며, 우선순위에 따라 즉시 적용해야 할 내용과 점진적으로 적용해야 할 부분으로 구분하여 운영하는 것이 필요하다(〈표 4-16〉 참조).

〈표 4-16〉 우선순위에 따른 적용 내용

우선순위	내용
1	통합된 데이터 제공 지침 수립, 승인통계 제공 절차의 보완, 데이터 제공 관련 양식의 보완
2	데이터 제공과 관련한 임시 전담 조직을 구성하여 운영
3	데이터 제공 전담 부서를 구성하여 운영

자료: 필자 작성

데이터 제공과 관련한 법제는 계속 변화하고 있고, 이러한 변화에 대해 지속적으로 검토하고 절차를 개선하여야 한다. 이 장의 연구 결과를 통해 현재 「공공데이터법」과 「데이터기반행정법」의 가명정보의 제공에 대한 부분은 우선 반영되었다. 이후에는 데이터 제공 TF 또는 전담 조직 부서에서 개선 내용에 대해 확인하고 반영하는 것이 필요하다. 이러한 활동이 지속되면 한국보건사회연구원의 데이터 제공 체계는 더욱 안전하고, 보유하고 있는 방대한 데이터는 유용하게 활용될 수 있다고 생각한다.



제5장

가명정보 결합을 활용한 데이터 통합 실증 연구

제1절 가명정보 결합을 활용한 통합데이터 구축

제2절 통합데이터 실증 분석

제3절 소결

제5장

가명정보 결합을 활용한 데이터 통합 실증 연구

제2장에서 살펴본 가명정보 결합 절차에 따라 한국의료패널조사와 한국사회보장정보원의 행정데이터를 실제로 통합하였다. 데이터 통합 과정은 사전준비, 결합신청, 가명처리, 결합수행, 반출의 순서로 진행되었다. 제1절은 단계별 진행 과정에 대해 정리하였고, 제2절은 가명정보 결합을 통해 구축된 통합데이터에 대한 실증 분석을 실시하였다.

제1절 가명정보 결합을 활용한 통합데이터 구축

1. 사전준비

현재 가명정보 결합을 처음 실시하는 경우에는 가명정보 결합의 전반적인 과정에 대해 컨설팅 지원을 받을 수 있다.³⁵⁾ 한국인터넷진흥원의 ‘가명처리 컨설팅 서비스’에서 신청하면 되며, 비용은 무료(한국인터넷진흥원에서 전액 지원)이다. 그 목적은 가명정보 결합이 활성화되도록 하는데 있다. 가명처리 컨설팅 서비스를 이용하여 가명정보 결합이 완료되면, 한국인터넷진흥원은 가명정보결합종합지원시스템에 가명정보 결합 우수 사례로 공개한다. 연구진은 컨설팅업체와 계약하였고, 그다음 가명정보 결합을 위한 단계별 절차를 수행하였다(2023.4.20.). 우선, 연구진과 관

35) 통합데이터 구축 과정에서 필요한 한국의료패널 데이터의 개인정보 가명처리와 데이터 심의, 데이터 결합 및 반출 심사 등의 과정에서 개인정보위원회와 한국인터넷진흥원(KISA)이 수행하는 「2023년 가명정보 활용 컨설팅」의 지원을 받았음.

련 담당자는 가명정보 관련 교육을 2회 받았다. 첫 번째 교육은 가명처리, 가명처리 절차 등에 대한 사전교육(2023.3.7.)이고, 두 번째 교육은 적합성, 적정성 검토 방법에 대한 교육(2023.4.24.)이었다.

가명정보에 대한 내부 관리계획을 수정·보완하였고, 한국보건사회연구원 홈페이지에 개인정보처리방침 관련 내용을 공시하였다(2023.6.15.).

가. 목적 설정

한국의료패널조사는 보건의료 분야에 특화된 정보가 많은 편에 속하는데, 이외에도 활용할 수 있는 다양한 정보가 포함되어 있다. 한편, 조사데이터는 응답자의 기억에 의존하여 수집하고 응답값을 입력하는 과정에서 발생하는 오류 등의 비표본오차가 포함되기 마련이다. 이에 따른 조사데이터의 정확도를 파악해 볼 필요가 있는데, 이는 행정데이터를 통해 검증할 수 있다.

가명정보 결합의 연구 목적은 한국의료패널조사에서의 사회보장급여 수급 여부, 소득, 자산, 부채에 대한 정보의 정확도 제고 방안 및 한국의료패널조사에서의 행정데이터 활용 범위 확대 가능성을 검토하는 데 있다. 이러한 연구 목적에 맞는 행정데이터를 검색한 결과, 가명정보 결합 대상 행정데이터는 한국사회보장정보원 정보시스템의 정보를 사용하는 것으로 결정하였다. 한국의료패널조사와 한국사회보장정보원 정보시스템의 데이터에 대한 설명은 다음과 같다.

1) 한국의료패널조사³⁶⁾

한국의료패널조사의 목적은 보건의료서비스의 대응성·접근성 향상과 효율화를 위해 관련 연구 및 정책 수행에 필요한 기초 정보를 수집하는데 있다. 조사 내용은 의료서비스 이용, 의료 관련 지출, 의료 이용에 영향을 미치는 사회경제적 요인, 건강 관련 인식 및 행태 등이 있다. 조사 방식은 조사원이 컴퓨터를 이용하여 면접조사(Computer Assisted Personal Interview, CAPI)한다. 조사 대상은 가구 내에 속한 모든 가구원이고, 조사 주기는 매년 실시한다. 1기 한국의료패널조사는 2008년~2019년까지 실시되었고, 표본 대표성 확보와 조사 효율성 제고 등을 위해 2기 한국의료패널조사가 2019년(사전조사)부터 시작되어 운영 중에 있다. 2008년부터 한국보건사회연구원과 국민건강보험공단이 공동으로 수행하고 있다.

가명정보 결합을 위해 활용한 데이터는 2019년 연간데이터(6,748가구, 16,587명)로 선정하였다.

2) 한국사회보장정보원 정보시스템

한국사회보장정보원은 사회복지 분야의 정보시스템 및 포털을 운영하고 있다. 정보시스템 및 포털은 사회보장정보시스템(행복이음, 범정부), 지역보건의료정보시스템, 사회서비스전자바우처시스템 외 6개³⁷⁾가 있다 (한상필 외, 2023, p.16). 행복이음³⁸⁾은 맞춤형 기초생활보장(생계, 의

36) 2기 한국의료패널 2019년 연간데이터 유저가이드(한국보건사회연구원, 국민건강보험. 2022. 5.)를 참고하여 작성함.

37) 보육통합정보시스템, 사회복지시설정보시스템, 디지털돌봄시스템, 노인맞춤돌봄시스템, 보건복지포털(복지로, e보건소)

38) 행복이음은 사회복지 급여·서비스 지원 대상자의 자격, 이력 정보 통합 관리 및 지자체 복지업무 처리 지원 시스템임.

료, 주거, 교육 급여), 기초연금, 소득, 재산, 부채 등에 대한 정보를 가지고 있어서, 한국의료패널조사와 가명정보를 결합하는 연구 목적에 적합하다고 판단하였다.

가명정보를 결합하기 위해 활용한 데이터는 2019년 12월 31일 맞춤형 기초생활보장 및 기초연금 수급자 기준 6,246,219명으로 선정하였다.

나. 결합 목적에 따른 계획서 작성 및 데이터 협의

가명정보 결합 목적을 증명할 수 있는 서류를 준비하기 위해 양식에 맞춰 과학적 연구계획서를 작성하였다(2023. 3. 14.). 과학적 연구계획서는 다음과 같은 항목이 포함되어 있다. 연구명은 ‘한국의료패널 및 사회보장정보시스템 수급자 정보의 결합을 통한 사회보장급여 수급 실태 및 소득·자산 현황 분석’이다. 연구 배경 및 목적으로, 첫 번째는 한국의료패널조사 데이터에 있는 사회보장급여, 소득, 자산 등의 응답값에 대한 정확도 제고이다. 한국의료패널조사는 대면조사를 통한 응답자의 기억에 의존하여 내용을 수집하기 때문에 정보가 정확하지 않을 가능성이 있다. 그러므로 이와 관련된 행정데이터(사회보장정보시스템)를 활용하여 조사데이터의 정확도를 파악해 볼 수 있다. 더 나아가 한국의료패널조사에서 사회보장급여 수급 여부, 수급 가구의 소득, 자산 등의 정보를 대체할 수 있는지도 검토해 볼 수 있다. 두 번째는 사회보장급여, 소득, 자산 특성에 따른 의료이용 실태 분석이다. 연구대상자 수는 한국의료패널조사의 경우 가구원이고, 사회보장정보시스템의 경우 맞춤형 급여와 기초연금 수급자이다. 연구 방법에는 연구분석 항목, 분석 방법, 결합데이터 이용기관에 대해 작성하였다. 연구분석 항목은, 한국의료패널조사는 가구 및 가구원의 사회보장급여 수급 여부, 소득, 자산, 부채, 의료비 등의 정보가

고, 사회보장정보시스템은 가구 및 가구원의 사회보장급여 수급 여부, 소득, 자산, 부채, 사회보장급여액 등의 정보이다. 분석 방법은 항목별 기술 통계 및 값 비교 분석이고, 결합데이터 이용기관의 경우 가명정보 처리는 한국보건사회연구원과 한국사회보장정보원이며, 데이터 결합은 한국사회보장정보원이고, 데이터 이용자는 한국보건사회연구원 연구진이다. 연구 내용은 두 가지로, 첫 번째는 가명정보 결합을 통해 구축된 통합데이터에서 한국의료패널조사와 사회보장정보시스템의 공통 변수에 대한 검토이다. 공통 변수는 사회보장수급 여부, 소득, 재산, 부채이다. 두 번째는 조사데이터에 행정데이터의 정보를 활용할 수 있는지에 대한 검토이다. 가명정보 결합을 통한 대체 가능한 정보를 파악하고 기존에 없는 정보를 추가할 수 있는지에 대해 파악해 볼 수 있을 것이다. 기대효과 및 활용방안은 조사데이터의 품질 향상 및 활용 범위의 확대 가능성이다. 이 연구 수행을 위한 연구진과 예상 연구기간(2023.7~12.)에 대한 정보를 작성하였다.

한편, 데이터 활용을 위해서는 데이터 보유기관과의 협의가 필요하다. 한국의료패널조사는 한국보건사회연구원과 국민건강보험공단에서 승인을 받았다. 사회보장정보시스템의 행정데이터는 데이터 보유기관의 승인뿐만 아니라 해당 정부 부처의 승인이 필요하다. 그래서 가명정보 결합에 활용되는 생계, 의료, 주거, 교육 급여, 기초연금 관련 행정데이터 제공에 대한 협조를 요청했다. 보건복지부 기초연금과(2023.6.15.), 보건복지부 기초생활보장과(2023.6.20.), 국토교통부 주거복지지원과(2023.6.28.), 교육부 교육복지정책과(2023.7.11.)는 공문을 통해 해당 행정데이터 제공을 승인한다고 통보하였다.

다. 가명정보 이용·제공 신청 및 적합성 검토

한국의료패널조사에 대한 가명정보를 이용하기 위해서 가명정보 이용·제공 신청서를 작성하여 한국보건사회연구원에 제출하였다(2023.5.11.). 이 신청서는 가명정보를 처리하는 담당 부서에 제출하면 되나, 현재 한국보건사회연구원은 담당 부서의 부재로 인하여 한국보건사회연구원에 공문으로 제출하였다.

가명정보 이용·제공 신청서는 다음과 같은 내용이 포함되어 있다. 처리 목적은 과학적 연구로, 한국의료패널조사와 사회보장정보시스템의 수급자 정보 결합을 위한 가명정보 제공(사회보장급여 수급 실태 및 소득·자산 현황 분석)에 있다. 활용 형태는 결합전문기관을 통한 결합이고, 이용 주체는 동일 개인정보처리자이고, 처리 환경은 내부이며, 처리 장소는 폐쇄 환경이다. 1회 제공으로 오프라인으로 제공받으며 제공 받는 사람은 내부는 동일부서이고, 외부는 결합전문기관 제공이다. 가명처리 대상 명칭은 2기 한국의료패널조사 데이터로, 2019년 연간데이터의 가구, 가구원, 의료서비스 이용 등에 대한 정보이다. 가명정보 이용기간은 신청한 날짜로부터 2023년 12월 31까지이다.

다음으로 제출한 가명정보 이용·제공 신청서를 바탕으로 가명정보를 제공하기 위한 적합성 검토위원회를 개최하였다(2023.5.15.). 연구계획서, 한국의료패널조사의 활용데이터 요구 수준표, IRB 심사 결과 통지서 등을 기반으로 하여 3명의 평가위원이 적합성을 검토하였고, 가명정보 제공승인을 통보받았다. 참고로 평가위원은 사회보장재정·데이터연구실 실장, 정보통계연구센터장, 한국의료패널조사 연구책임자로 구성하였다.

한편, 사회보장정보시스템의 행정데이터에 대한 가명정보 제공 승인은 해당 정부 부처의 공문으로 같음하였다.

2. 결합신청

연구진은 가명정보결합종합지원시스템³⁹⁾에서 신청정보를 등록하여 신청번호를 발급받았고(2023.7.5.), 이후 가명정보 결합 신청 관련 서류를 제출하였다(2023.8.11.). 제출한 서류는 결합신청서, 사업자등록증, 개인정보 수집·이용 동의서, 과학적 연구계획서, (상세)연구계획서, 가명정보에 대한 안전조치 이행확약서이다. 참고로 결합 대상 가명정보 관련 서류, 적정성 검토 결과서(위원용), 적정성 검토 종합결과서는 가명정보 제공 때 제출하였다.

데이터 보유기관 간 협의하여 결합키를 생성하였다(2023.8.20.). 결합키는 이름, 성별, 생년월일과 salt 값을 조합하였다. 한국인터넷진흥원은 2개 데이터 보유기관의 결합키만 가지고 결합을 실시한 다음에 결합률을 연구진에게 통보하였다(2023.8.22.). 한국의료패널조사 데이터와 사회보장행정시스템 행정데이터의 결합률은 20%이고, 최종 결합된 데이터는 3,320명이었다. 이는 한국의료패널조사의 응답자 16,587명 기준을 의미한다. 그런데 한국의료패널조사의 응답자 중에서 사회보장급여 수급자(4,673명⁴⁰⁾) 기준으로 보면, 결합률은 71%라고 볼 수 있다. 결합방법은 확대단일결합으로 진행하였다.

3. 가명처리

한국의료패널조사 데이터에 대한 가명처리를 완료하였고(2023.6.1.),

39) 가명정보결합종합지원시스템 홈페이지, <https://link.privacy.go.kr/nadac/index.do>, 2023.7.5. 인출

40) 한국의료패널조사는 가구조사표에서 사회보장급여 수급 여부를 질문하여 가구 단위로 파악이 가능함. 이를 개인 단위 데이터에 연계하여 분석한 결과임.

적정성 검토위원회를 위한 문서를 작성하였다(2023.8.2.). 이후 적정성 검토위원회를 개최하여 5명의 평가위원이 한국의료패널조사 데이터에 대한 적정성 검토를 실시하였다(2023.8.14.). 적정성 검토 기초자료는 총 17개로, (1) 적정성 검토 개요, (2) 연구계획서 및 IRB 승인서, (3) 가명정보 활용 신청서, (4) 적합성 검토 결과서, (5) 활용데이터 요구수준표_가구, 활용데이터 요구수준표_개인, (6) 원본 데이터 명세(종합), (7) 원본 데이터 명세(상세) 및 예시, (8) 식별위험성 검토 체크리스트, (9) 식별위험성 검토 결과보고서, (10) 가명처리 수준 정의표, (11) 가명처리 후 데이터 명세(종합), (12) 가명처리 후 데이터 명세 및 예시, (13) 범주형 데이터 분포, (14) 연속형 데이터 분포, (15) 분포도, (16) 보안서약서, (17) 위탁계약서가 해당하였다. 5명의 평가위원은 가명정보 전문가에 해당하는 외부인으로 구성하였으며, 제출한 기초자료를 검토하였다. 적정성 검토 결과에 따라 한국의료패널조사 데이터에 대한 추가 가명처리를 실시한(2023.8.25.) 다음에, 한국사회보장정보원으로 전송하였다(2023.9.18.). 한국의료패널 데이터의 전송은 한국사회보장정보원의 파일 송수신 프로그램을 통해 진행되었다.

한편, 사회보장정보시스템의 행정데이터도 가명처리 과정을 동일하게 수행하였고, 적정성 검토위원회를 실시하였다(2023.9.4.).

4. 결합수행

한국사회보장정보원은 한국의료패널조사 데이터와 사회보장정보시스템 행정데이터의 가명정보 결합을 완료하여 통합데이터를 구축하였다(2023.9.18.). 앞서 진행했던 적정성 검토 결과를 반영하기 위하여 통합데이터에 대한 추가 가명처리를 실시하였다(2023.9.26.).

5. 반출 및 관리

결합된 가명정보 데이터에 대한 반출을 위해 반출 심사를 개최하였다(2023.10.12.). 반출심사위원회의 구성은 한국인터넷진흥원의 가명정보 전문가 풀(pool)에서 추천받은 3인이었고, 결합 및 반출 관련 신청 서류를 검토하였다. 연구 목적에 따른 데이터 분석을 실시하였는데, 분석 장소는 한국사회보장정보원의 가명정보결합분석센터이었다.

제2절 통합데이터 실증 분석

가명정보 결합으로 구축된 통합데이터에서 활용 가능한 주요 변수 목록은 <표 5-1>과 같다. 사회보장 급여 수급 여부는 2개의 데이터에 모두 포함되어 있다. 이에 반해 각 데이터에만 있는 변수는 한국의료패널 조사의 경우 의료비 변수가 있다. 사회보장정보시스템의 경우는 기초연금에서의 수급 여부, 배우자 수급 구분, 맞춤형 급여에서의 수급 여부, 수급자 구분, 급여액이 있다. 한편, 한국의료패널조사 데이터의 조사 시점은 맞춤형 급여 수급 여부, 부채와 자산의 경우 2019년 12월 31일이고, 이를 제외한 나머지 변수는 2019년 1월 1일~12월 31일이다. 사회보장정보시스템의 행정데이터의 시점은 데이터 추출 시점인 2019년 12월 31일이다.

〈표 5-1〉 한국의료패널조사와 사회보장정보시스템에서의 활용 가능 변수

구분	한국의료패널조사 데이터	사회보장정보시스템 행정데이터
사회보장 급여 수급 정보	- 기초연금 수급 여부 - 맞춤형 급여 수급 여부 · 생계, 의료, 주거, 교육 급여	- 기초연금 수급 여부 - 맞춤형 급여 수급 여부 · 생계, 의료, 주거, 교육 급여
		- 기초연금에서의 배우자 수급 여부 - 맞춤형 급여에서의 수급자 구분 (일반/시설)
사회보장 급여액		- 기초연금 지급액 - 맞춤형 급여별 지급액 · 생계, 의료, 주거, 교육 급여
소득	- 근로·사업소득 유무	- 추출 시점 근로·사업소득 유무
	- 연간 근로소득	- 추출 시점 근로소득
	- 연간 사업소득	- 추출 시점 사업소득
	- 연간 금융소득	- 추출 시점 금융소득
	- 연간 부동산·동산소득	- 추출 시점 부동산·동산소득
	- 연간 기타소득	- 추출 시점 기타소득
	- 연간 공적이전소득	- 추출 시점 공적이전소득
	- 월 평균 가구 총소득	- 추출 시점 가구 총 소득
부채	- 조사 시점 가구 총부채	- 추출 시점 가구 총 부채 · 부채 항목 전체
자산	- 조사 시점 가구 총자산	- 추출 시점 가구 총 자산 · 각종 재산 항목 전체
의료비	- 연간 전체 의료비 지출	

주: 1) 회색 표기는 해당 데이터에는 포함하지 않는 정보임을 나타냄.

2) 한국의료패널조사 데이터의 조사 시점은 맞춤형 급여 수급 여부, 부채, 자산의 경우 2019년 12월 31일이고, 기초연금 수급 여부, 소득, 의료비의 경우 2019년 1월 1일에서 12월 31일임.

3) 사회보장정보시스템의 행정데이터는 추출 시점이 2019년 12월 31일임.

자료: 1) 한국보건사회연구원·국민건강보험공단. (2019). 한국의료패널조사 2019년 연간데이터.

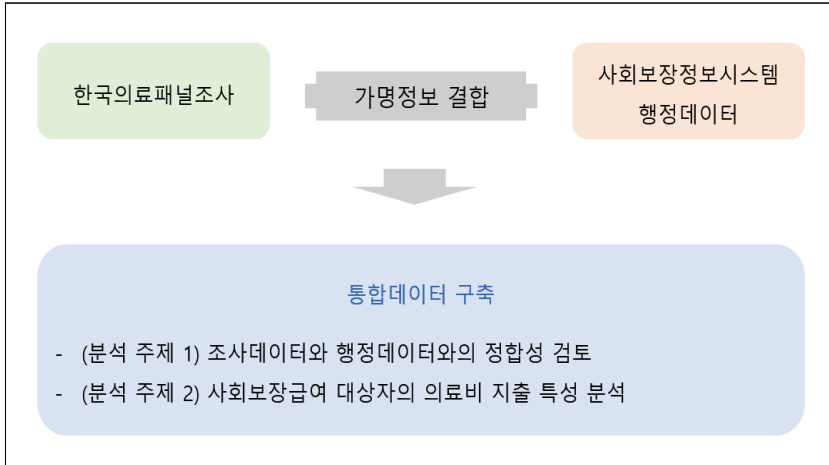
2) 한국사회보장정보원. (2019). 사회보장정보시스템 행정데이터.

통합데이터를 활용한 분석 주제는 2가지로 [그림 5-1]과 같다. 첫 번째는 한국의료패널조사 데이터의 품질 제고를 위해 사회보장정보시스템의 행정데이터와의 정합성을 검토하는 것이다. 즉, 조사데이터와 행정데이터 간의 사회보장 급여 수급자 정보를 비교 분석하는 것이다. 두 번째는 사회보장 급여 수급자의 의료비 지출 특성 분석이다.

분석에 앞서 관련 용어를 정리하면 다음과 같다. 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터가 가명정보 결합된 데이터를 ‘통합데이터’라고 정의하고자 한다. 또한, 한국의료패널조사 데이터는 ‘조사데이터’로, 사회보장정보시스템의 행정데이터는 ‘행정데이터’로 정의하여 표기를 간단하게 하고자 한다.

먼저 통합데이터의 조사데이터와 행정데이터에 대한 기초 분석을 실시하였다. 통합데이터는 개인 단위로 결합되었기 때문에 행정데이터를 가구 단위로 구조화하는 과정을 가졌다. 한국의료패널조사는 가구 조사표에 사회보장 급여와 관련된 설문 문항이 있어서 가구 단위로만 파악이 가능하다. 행정데이터는 개인 단위로 보면 3,320명이고, 가구 단위로 보면 2,473가구가 조사데이터와 최종 결합되었다. 행정데이터의 가구 아이디는 사회보장 급여 수급 단위로 관리하고 있어서, 같은 가구원이라도 가구 아이디가 다를 수 있다.

[그림 5-1] 통합데이터를 활용한 분석 주제



자료: 필자 작성

〈표 5-2〉는 행정데이터에만 있는 주요 변수에 대한 분석 결과이다. 사회보장 급여 수급 가구 현황을 살펴보면, 기초연금 수급 가구는 전체 분석 대상 가운데 33%인 것으로 나타났다. 수급 자격별로 보면, 부부2인의 비율이 16.9%로 가장 높았으며, 다음으로 단독(15.5%)이었고, 부부1인이 0.6%로 낮은 편에 속하였다. 다만, 사회보장정보시스템에서 수급 자격이 ‘부부2인’임에도 불구하고 가구원 중 1인만 분석 데이터에 포함된 사례가 일부 확인되었다. 다음으로 맞춤형 급여의 세부 항목별로 보면 주거급여 수급 가구의 비율이 7.7%였고, 다음으로 의료급여 수급 가구(6.2%)와 생계급여 수급 가구(5.5%) 순이었다. 교육급여 수급 가구 비율이 1.2%로 가장 낮게 나타났다.

〈표 5-2〉 행정데이터 기준 사회보장 급여 관련 분포

(단위: 가구, %)

구분			빈도	비율
기초연금	수급 여부	수급	2,206	33.0
		미수급	4,483	67.0
	수급 자격 구분	단독	1,038	15.5
		부부1인	40	0.6
		부부2인	1,128	16.9
		미수급	4,483	67.0
생계급여	수급 여부	수급	371	5.5
		미수급	6,318	94.5
의료급여	수급 여부	수급	416	6.2
		미수급	6,273	93.8
	종별 구분	1종	433	6.5
		2종	158	2.4
주거급여	수급 여부	수급	514	7.7
		미수급	6,175	92.3
교육급여	수급 여부	수급	77	1.2
		미수급	6,612	98.8
전체			6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

사회보장 급여액 수준을 살펴보면 〈표 5-3〉과 같다. 기초연금 수급자의 연간 평균 수급액은 336.1만 원(표준편차: 109.7만 원)으로 나타났다. 수급 자격별로 살펴보면, 단독 수급 가구가 310.8만 원(표준편차: 67.3만 원), 부부1인 수급 가구가 236.3만 원(표준편차: 112.2만 원), 부부2인 가구가 362.9만 원(표준편차: 131.5만 원)인 것으로 나타났다. 생계급여를 받는 371가구의 연간 평균 생계급여 수급액은 515.3만 원(표준편차: 335.1만 원)으로 나타났다. 이 중에서 의료급여 1종에 해당되는 가구의 연간 생계급여 수급액은 468.2만 원(표준편차: 292.2만 원), 의료급여 2종에 해당되는 가구는 732.8만 원(표준편차: 418만 원)인 것으로 나타났다.

〈표 5-3〉 행정데이터 기준 사회보장 급여액 현황

(단위: 가구, 만 원, 만 원)

구분		빈도	평균	표준편차
기초연금 연간 총액	전체	2,206	336.1	109.7
	단독	1,038	310.8	67.3
	부부1인	40	236.3	112.2
	부부2인	1,128	362.9	131.5
생계급여 연간 총액	전체	371	515.3	335.1
	의료급여 1종	308	468.2	292.2
	의료급여 2종	62	732.8	418.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

17개 시도 기준 거주 지역에 대한 조사데이터와 행정데이터 간의 일치 여부를 살펴보면 〈표 5-4〉와 같다. 거주 지역의 일치는 95.3%로 나타나 높은 편이나, 일부 일치하지 않는 관측치가 있었다.

〈표 5-4〉 거주 지역에 대한 조사데이터와 행정데이터의 일치 여부

(단위: 명, %)

	조사데이터	행정데이터
일치	3,163	95.3
불일치	157	4.7
전체	3,320	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

가구원 수도 조사데이터와 행정데이터에 모두 있는 변수였으나, 가구원 수 정의가 다르므로 인하여 비교해 볼 수 없었다. 조사데이터는 가구조사표가 있어서 가구 단위의 데이터가 있다. 반면에, 행정데이터는 사회보장 급여 수급 단위에 따라 데이터를 관리하고 있어서 같은 가구이더라도 가구원 수가 다를 수 있다. 그래서 가구원 수는 조사데이터에 있는 변수로 활용하였다. 〈표 5-5〉를 보면 2인 가구의 비율이 39.6%로 가장 높

게 나타났으며, 다음으로 1인 가구(21.5%)였다.

〈표 5-5〉 조사데이터 기준 가구원 수 분포

(단위: 가구, %)

	빈도	비율
1인	1,440	21.5
2인	2,650	39.6
3인	1,135	17.0
4인	1,100	16.4
5인 이상	364	5.4
전체	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

이러한 특성으로 인하여 이 절에서는 행정데이터에서 추가된 정보를 제외한 나머지 정보는 조사데이터를 기준으로 분석하였다.

1. 조사데이터와 행정데이터의 정합성 분석

조사데이터와 행정데이터에는 공통으로 가지고 있는 사회보장 급여 수급 정보가 있다. 사회보장 급여인 기초연금과 맞춤형 급여(생계, 의료, 주거, 교육 급여)의 수급 여부에 대한 정합성 분석을 하였다. 정합성 분석 결과에서의 차이를 파악하고자 가구 특성과 가구주의 인구사회학적 특성 분포를 살펴보았다.

가. 기초연금

통합데이터에서 기초연금 수급 여부 현황을 살펴보았다. 〈표 5-6〉을 보면, 기초연금의 경우 조사데이터는 35.8%가 수급 받고 있어 행정데이

터(33%)에 비해 2.8%p 높게 나타났다.

〈표 5-6〉 통합데이터의 기초연금 수급 여부에 대한 분포

	조사데이터		행정데이터	
	가구	%	가구	%
예	2,393	35.8	2,206	33.0
아니오	4,296	64.2	4,483	67.0
전체	6,689	100.0	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음은 가구 응답의 정합성을 살펴보기 위하여 기초연금 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 조사데이터와 행정데이터의 값이 모두 일치하는 가구의 비율은 93.6%로 나타났다(〈표 5-7〉 참조). 반면에 일치하지 않는 경우는 6.4%였는데, 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 가구의 비율이 4.6%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 가구의 비율은 1.8%로 2.8%p 낮게 나타났다.

〈표 5-7〉 조사데이터와 행정데이터의 기초연금 수급 여부 일치 여부

(단위: 가구, %)

	조사데이터	행정데이터	빈도	비율
일치	예	예	2,085	31.2
	아니오	아니오	4,175	62.4
불일치	예	아니오	308	4.6
	아니오	예	121	1.8
전체			6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

앞에서 살펴보았듯이 조사데이터와 행정데이터의 기초연금 수급 여부가 일치하지 않는 집단의 비율이 낮은 편에 속하나, 이러한 차이에 대해 살펴볼 필요가 있다. 그래서 기초연금을 수급 받는 대상 가구의 조사데이터와 행정데이터의 가구 특성 및 가구주의 인구사회학적 특성에 대한 분포를 비교하였다. 조사데이터와 행정데이터에서 모두 기초연금을 수급 받지 않는다고 한 집단을 제외한 3개의 집단에 대해 분석하였다. 3개의 집단은 (1) 조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단, (2) 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단, (3) 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단이다. 이때 가구 및 가구주의 인구사회학적 변수는 조사데이터를 기준으로 분석하였다.

〈표 5-8〉을 보면, 조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 거주 지역 분포는 광역시가 28%로 가장 높았고, 다음으로 전라도(18.8%)와 서울 및 경기(17.9%) 순으로 나타났다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 광역시(35.7%), 서울 및 경기(19.2%)가 높은 편이었다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 서울 및 경기(26.4%), 광역시(24.8%)가 높게 나타났다.

〈표 5-8〉 기초연금 수급에 대한 일치 여부에 따른 거주 지역 분포

(단위: 가구, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
서울 및 경기	374	59	32	17.9	19.2	26.4
광역시	583	110	30	28.0	35.7	24.8
충청도	303	42	18	14.5	13.6	14.9
전라도	393	42	18	18.8	13.6	14.9
경상도	330	37	18	15.8	12.0	14.9
강원 및 제주	102	18	5	4.9	5.8	4.1
전체	2,085	308	121	100.0	100.0	100.0

주: 세종은 충청도에 포함함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 가구원 수 분포를 보면, 2인 가구(51.9%)와 1인 가구(32.2%)의 비율이 높은 편이었다(〈표 5-9〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 2인 가구(39.6%), 1인 가구(27.3%)가 높은 편이었다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 2인 가구(64.5%)가 가장 높게 나타났고, 다음으로 1인 가구(15.7%)와 3인 가구(14.9%)가 차 지하였다.

〈표 5-9〉 기초연금 수급에 대한 일치 여부에 따른 가구원 수 분포

(단위: 가구, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
1인	671	84	19	32.2	27.3	15.7
2인	1083	122	78	51.9	39.6	64.5
3인	214	56	18	10.3	18.2	14.9
4인	51	18	2	2.4	5.8	1.7
5인 이상	66	28	4	3.2	9.1	3.3
전체	2,085	308	121	100.0	100.0	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 주거형태 분포를 보면, 단독주택이 55.7%로 절반 이상을 차지하였고, 다음으로 아파트(29.3%)로 나타났다(〈표 5-10〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 단독주택이 50.6%로 절반을 상회하였고, 다음으로 아파트(32.8%)로 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 단독주택이 53.7%로 절반 이상이었다. 다음으로 아파트가 36.4%로 높았는데, 다른 집단의 아파트 비율과 비교했을 때 가장 큰 값을 가졌다.

〈표 5-10〉 기초연금 수급에 대한 일치 여부에 따른 주거형태 분포

(단위: 가구, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
단독주택	1,162	156	65	55.7	50.6	53.7
아파트	611	101	44	29.3	32.8	36.4
연립주택	82	18	1	3.9	5.8	0.8
다세대주택	202	29	10	9.7	9.4	8.3
기타	28	4	1	1.3	1.3	0.8
전체	2,085	308	121	100.0	100.0	100.0

주: 기타는 영업용 건물 내 주택(상가, 공장 등), 오피스텔, 호텔, 여관, 기숙사 등을 의미함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 연간 가구 총소득 분포를 보면, 2,000만 원 미만이 64.2%로 가장 높고, 다음으로 2,000만 원 이상~4,000만 원 미만이 26.3%로 나타났다(〈표 5-11〉 참조). 4,000만 원 이상의 비율은 다른 집단에 비해 가장 작은 값을 가졌다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 2,000만 원 미만이 57.5%로 가장 높고, 다음으로 2,000만 원 이상~4,000만 원 미만이 20.8%로 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 2,000만 원 이상~4,000만 원이 39.2%로 가장 높고, 다음으로 2,000만 원 미만이 33.3%로 나타났다. 연간 가구 총소득의 평균은 3,254만 원 정도로 다른 집단과 비교했을 때 가장 크게 나타났다.

〈표 5-11〉 기초연금 수급에 대한 일치 여부에 따른 연간 가구 총소득 분포

(단위: 가구, %, 만 원)

	빈도			비율		
조사데이터	예	예	아니오	예	예	아니오
행정데이터	예	아니오	예	예	아니오	예
2,000만 원 미만	1,328	177	40	64.2	57.5	33.3
4,000만 원 미만	544	64	47	26.3	20.8	39.2
6,000만 원 미만	113	36	19	5.5	11.7	15.8
6,000만 원 이상	85	31	14	4.1	10.1	11.7
전체	2,070	308	120	100.0	100.0	100.0
평균	2,063.5	2,742.5	3,254.2	-	-	-

주: 연간 가구 총소득 금액에 대해 무응답인 16가구는 결측 처리한 후 분석함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음으로 가구주의 인구사회학적 특성에 대한 분포를 살펴보았다. 먼저, 조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 가구주의 성별 분포를 보면 남성이 64.2%로 나타났다(〈표 5-12〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 남성이 58.5%로 높게 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 남성이 80.5%였으며, 다른 집단과 비교했을 때 가장 크게 나타났다.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 가구주의 연령대 분포를 보면, 70세~79세가 52.2%로 절반 이상을 차지하였고, 다음으로 60세~69세(23.8%)로 나타났다(〈표 5-13〉 참조). 다른 집단과 비교했을 때 가구주의 연령대가 높은 편에 속한다고 볼 수 있다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고

있다고 한 집단은 70세~79세(35.3%)와 60세~69세(32%)가 높은 편에 속하였다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 60세~69세가 49.2%로 가장 많았고, 다음으로 70세~79세(35.6%)로 나타났다.

〈표 5-12〉 기초연금 수급에 대한 일치 여부에 따른 가구주의 성별 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
남성	1,305	159	95	64.2	58.5	80.5
여성	728	113	23	35.8	41.5	19.5
전체	2,033	272	118	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 91가구임.
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

〈표 5-13〉 기초연금 수급에 대한 일치 여부에 따른 가구주의 연령대 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
30~50세	56	32	4	2.8	11.8	3.4
50~59세	64	32	6	3.1	11.8	5.1
60~69세	483	87	58	23.8	32.0	49.2
70~79세	1,062	96	42	52.2	35.3	35.6
80세 이상	368	25	8	18.1	9.2	6.8
전체	2,033	272	118	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 91가구임.
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 가구주의 혼인상태 분포를 보면, 혼인 중이 55.6%로 절반 이상을 차지하였고, 다음으로 사별/실종(33.3%)으로 나타났다(〈표 5-14〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 혼인 중이 46.7%로 가장 높았고, 다음으로 사별/실종(32%)이 차지하였다. 다른 집단에 비해 미혼과 별거/이혼의 비율이 높은 편에 속하였다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 혼인 중이 75.4%로 가장 높았고, 다른 집단과 비교해도 가장 큰 비율을 가졌다. 다음으로 사별/실종이 16.1%를 차지하였다.

〈표 5-14〉 기초연금 수급에 대한 일치 여부에 따른 가구주의 혼인상태 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
미혼	52	17	3	2.6	6.3	2.5
혼인 중	1,131	127	89	55.6	46.7	75.4
별거/이혼	173	41	7	8.5	15.1	5.9
사별/실종	677	87	19	33.3	32.0	16.1
전체	2,033	272	118	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 91가구임.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 가구주의 최종 학력 분포를 보면, 초졸이 34.4%로 가장 높았으

며, 다른 집단에 비해 초졸 이하의 비율이 52.6%로 가장 높았다(〈표 5-15〉 참조). 다음으로 고졸(20.8%)과 중졸(19.6%)로 나타났다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 초졸(26.1%)과 고졸(24.3%)의 비율이 유사하게 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 고졸이 37.3%로 가장 높았고, 또한 고졸 이상의 비율도 50.9%로 다른 집단에 비해 가장 높게 나타났다. 다음으로 중졸이 26.3%를 차지하였다.

〈표 5-15〉 기초연금 수급에 대한 일치 여부에 따른 가구주의 최종 학력 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
무학	370	51	4	18.2	18.8	3.4
초졸	699	71	23	34.4	26.1	19.5
중졸	399	48	31	19.6	17.6	26.3
고졸	423	66	44	20.8	24.3	37.3
대졸 이상	142	36	16	7.0	13.2	13.6
전체	2,033	272	118	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 91가구임.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 기초연금을 수급 받는다고 일치한 집단의 가구주의 경제활동 상태 분포를 보면, 비경제 활동인구가 48.3%로 가장 높았으며, 다른 집단에 비해 가장 높았다(〈표 5-16〉 참조). 다음

으로 고용주/자영업자(25%)와 임금근로자(14.8%)로 나타났다. 한편, 기타(11.9%)의 경우도 다른 집단과 비교했을 때 높은 편에 속하였다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 비경제 활동인구가 37.5%로 가장 높게 나타났다. 다음으로 고용주/자영업자(27.6%)와 임금근로자(25.7%)로 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 비경제 활동인구가 41.5%로 가장 높았다. 다음으로 고용주/자영업자(33.9%)와 임금근로자(22.9%)로 나타났다. 한편, 기타는 1.7%로 다른 집단과 비교했을 때 가장 낮은 비율에 속하였다.

〈표 5-16〉 기초연금 수급에 대한 일치 여부에 따른 가구주의 경제활동 상태 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	예	아니오	예	예	아니오
행정데이터	예	아니오	예	예	아니오	예
임금근로자	300	70	27	14.8	25.7	22.9
고용주/자영업자	509	75	40	25.0	27.6	33.9
비경제 활동인구	982	102	49	48.3	37.5	41.5
기타	242	25	2	11.9	9.2	1.7
전체	2,033	272	118	100.0	100.0	100.0

주: 1) 가구주의 정보가 결측인 경우가 91가구임.

2) 기타는 무급가족종사자, 자활근로, 공공근로, 희망근로, 노인 일자리가 해당함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

나. 맞춤형 급여

맞춤형 급여는 생계, 의료, 주거, 교육 급여로 구성되어 있어서, 각 급여별로 살펴본 후 맞춤형 급여로 통합하여 살펴보았다.

1) 생계급여

통합데이터에서 생계급여 수급 여부 현황을 살펴보았다. 조사데이터(5.2%)와 행정데이터(5.5%) 간의 차이가 0.3%p로 나타나 유사하다고 볼 수 있다(〈표 5-17〉 참조).

다음은 가구 응답의 정합성을 살펴보기 위하여 생계급여 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 가구 조사데이터와 행정데이터의 값이 모두 일치하는 경우는 98.2%로 나타났다(〈표 5-18〉 참조). 반면에 불일치하는 경우는 1.7%였는데, 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는 경우는 0.7%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 비율은 1%로 0.3%p 높게 나타났다.

〈표 5-17〉 통합데이터의 생계급여 수급 여부에 대한 분포

	조사데이터		행정데이터	
	가구	%	가구	%
예	349	5.2	371	5.5
아니오	6,340	94.8	6,318	94.5
전체	6,689	100.0	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

〈표 5-18〉 조사데이터와 행정데이터의 생계급여 수급 여부 일치 여부

(단위: 가구, %)

	조사데이터	행정데이터	빈도	비율
일치	예	예	301	4.5
	아니오	아니오	6,270	93.7
불일치	예	아니오	48	0.7
	아니오	예	70	1.0
전체			6,689	100

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

2) 의료급여

통합데이터에서 의료급여 수급 여부 현황을 살펴보았다. 조사데이터 (6.1%)와 행정데이터(6.2%) 간의 차이가 0.1%p로 나타나 유사하다고 볼 수 있다(〈표 5-19〉 참조).

〈표 5-19〉 통합데이터의 의료급여 수급 여부에 대한 분포

	조사데이터		행정데이터	
	가구	%	가구	%
예	407	6.1	416	6.2
아니오	6,282	93.9	6,273	93.8
전체	6,689	100.0	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음은 가구 응답의 정합성을 살펴보기 위하여 의료급여 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 가구 조사데이터와 행정데이터의 값이 모두 일치하는 경우는 98.1%로 나타났다(〈표

5-20〉 참조). 반면에 불일치하는 경우는 2%였는데, 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 경우는 0.9%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 비율은 1.1%로 0.2%p 높게 나타났다.

〈표 5-20〉 조사데이터와 행정데이터의 의료급여 수급 여부 일치 여부

(단위: 가구, %)

	조사데이터	행정데이터	빈도	비율
일치	예	예	345	5.2
	아니오	아니오	6,211	92.9
불일치	예	아니오	62	0.9
	아니오	예	71	1.1
전체			6,689	100

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

3) 주거급여

통합데이터에서 주거급여 수급 여부 현황을 살펴보았다. 조사데이터가 5.9%로 행정데이터(7.7%)에 비해 1.8%p 낮게 나타났다(〈표 5-21〉 참조).

〈표 5-21〉 통합데이터의 주거급여 수급 여부에 대한 분포

	조사데이터		행정데이터	
	가구	%	가구	%
예	394	5.9	514	7.7
아니오	6,295	94.1	6,175	92.3
전체	6,689	100.0	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음은 가구 응답의 정합성을 살펴보기 위하여 주거급여 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 가구 조사데이터와 행정데이터의 값이 모두 일치하는 경우는 96.7%로 나타났다(〈표 5-22〉 참조). 반면에 불일치하는 경우는 3.2%였는데, 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 경우는 0.7%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 비율은 2.5%로 1.8%p 높게 나타났다.

〈표 5-22〉 조사데이터와 행정데이터의 주거급여 수급 여부 일치 여부

(단위: 가구, %)

	조사데이터	행정데이터	빈도	비율
일치	예	예	344	5.1
	아니오	아니오	6,125	91.6
불일치	예	아니오	50	0.7
	아니오	예	170	2.5
전체			6,689	100

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

4) 교육급여

통합데이터에서 교육급여 수급 여부 현황을 살펴보았다. 조사데이터가 0.7%로 행정데이터(1.2%)에 비해 0.5%p 낮았다(〈표 5-23〉 참조). 맞춤형 급여 중에서 교육급여가 수급 대상 가구가 가장 작게 나타났다.

다음은 가구 응답의 정합성을 살펴보기 위하여 주거급여 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 가구 조사데이터와 행정데이터의 값이 모두 일치하는 경우는 99.1%로 나타났다(〈표 5-24〉 참조). 반면에 불일치하는 경우는 0.9%였는데, 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 경우는

0.2%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 비율은 0.7%로 0.5%p 높게 나타났다.

〈표 5-23〉 통합데이터의 교육급여 수급 여부에 대한 분포

	조사데이터		행정데이터	
	가구	%	가구	%
예	46	0.7	77	1.2
아니오	6,643	99.3	6,612	98.8
전체	6,689	100.0	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

〈표 5-24〉 조사데이터와 행정데이터의 교육급여 수급 여부 일치 여부

(단위: 가구, %)

	조사데이터	행정데이터	빈도	비율
일치	예	예	32	0.5
	아니오	아니오	6,598	98.6
불일치	예	아니오	14	0.2
	아니오	예	45	0.7
전체			6,689	100

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

한편, 한국의료패널조사에서 맞춤형 급여의 수급 여부에 대한 설문 문항 구조는 다음과 같다. 먼저 맞춤형 급여의 수급 여부를 질문한 다음에 따라, 맞춤형 급여를 수급한다고 응답한 경우로 생계급여, 의료급여, 주거급여와 교육급여의 수급 여부를 각각 질문하는 하위 설문 문항 구조로 구성되어 있다.

이에 따라 통합데이터에서의 맞춤형 급여 수급 여부 현황을 살펴보았다. 조사데이터가 7%로 행정데이터(8.2%)에 비해 1.2%p 낮았다(〈표 5-25〉 참조).

〈표 5-25〉 통합데이터의 맞춤형 급여 수급 여부에 대한 분포

	조사데이터		행정데이터	
	가구	%	가구	%
예	467	7.0	548	8.2
아니오	6,222	93.0	6,141	91.8
전체	6,689	100.0	6,689	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음은 가구 응답의 정합성을 살펴보기 위하여 맞춤형 급여 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 가구 조사데이터와 행정데이터의 값이 모두 일치하는 경우는 96.6%로 나타났다(〈표 5-26〉 참조). 반면에 불일치하는 경우는 3.4%였는데 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 경우는 1.1%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 비율은 2.3%로 1.2%p 높게 나타났다.

〈표 5-26〉 통합데이터의 맞춤형 급여 수급 여부 일치 여부

(단위: 가구, %)

	조사데이터	행정데이터	빈도	비율
일치	예	예	396	5.9
	아니오	아니오	6,070	90.7
불일치	예	아니오	71	1.1
	아니오	예	152	2.3
전체			6,689	100

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음은 맞춤형 급여를 수급 받는 대상 가구의 조사데이터와 행정데이터의 가구 특성 및 가구주의 인구사회학적 특성에 대한 분포를 비교하였다. 조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받지 않는다

고 한 집단을 제외한 3개의 집단에 대해 분석하였다. 3개의 집단은 (1) 조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단, (2) 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단, (3) 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단이다. 이때 가구 및 가구주의 인구사회학적 변수는 조사데이터를 기준으로 분석하였다.

〈표 5-27〉을 보면 조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단의 거주 지역 분포는 광역시가 42.2%로 가장 높았고, 다음으로 서울 및 경기(18.2%)와 전라도(15.2%) 순으로 나타났다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 광역시(47.9%), 전라도(18.3%)가 높은 편이었다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 광역시(35.5%), 서울 및 경기(21.7%)가 높게 나타났다.

〈표 5-27〉 맞춤형 급여 수급에 대한 일치 여부에 따른 거주 지역 분포

(단위: 가구, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
서울 및 경기	72	9	33	18.2	12.7	21.7
광역시	167	34	54	42.2	47.9	35.5
충청도	37	6	16	9.3	8.5	10.5
전라도	60	13	25	15.2	18.3	16.4
경상도	48	9	13	12.1	12.7	8.6
강원 및 제주	12	0	11	3.0	0.0	7.2
전체	396	71	152	100	100	100

주: 세종은 충청도에 포함함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일정한 집단의 가구원 수 분포를 보면, 1인 가구(47.7%)와 2인 가구(30.3%)의 비율이 높은 편이었다(〈표 5-28〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 2인 가구(36.6%), 1인 가구(32.4%)가 높은 편이었다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 2인 가구(32.2%)가 가장 높게 나타났고, 다음으로 1인 가구(24.3%)와 4인 가구(23.7%)가 차지하였다. 4인 가구 이상의 비율이 다른 집단에 비해 높은 편이었다.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일정한 집단의 주거형태 분포를 보면, 단독주택이 41.9%였고, 다음으로 아파트(36.6%)로 나타났다(〈표 5-29〉 참조). 한편, 다세대주택이 15.2%로 다른 집단에 비해 높은 편에 속하였다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 아파트가 45.1%였고, 다음으로 단독주택(36.6%)으로 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 아파트가 48%이었고, 다음으로 단독주택이 36.2%로 나타났다.

〈표 5-28〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구원 수 분포

(단위: 가구, %)

조사데이터	빈도			비율		
	예	예	아니오	예	예	아니오
행정데이터	예	아니오	예	예	아니오	예
1인	189	23	37	47.7	32.4	24.3
2인	120	26	49	30.3	36.6	32.2
3인	50	13	19	12.6	18.3	12.5
4인	24	7	36	6.1	9.9	23.7
5인 이상	13	2	11	3.3	2.8	7.2
전체	396	71	152	100.0	100.0	100.0

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

〈표 5-29〉 맞춤형 급여 수급에 대한 일치 여부에 따른 주거형태 분포

(단위: 가구, %)

조사데이터	빈도			비율		
	예	예	아니오	예	예	아니오
행정데이터	예	아니오	예	예	아니오	예
단독주택	166	26	55	41.9	36.6	36.2
아파트	145	32	73	36.6	45.1	48.0
연립주택	22	3	8	5.6	4.2	5.3
다세대주택	60	8	15	15.2	11.3	9.9
기타	3	2	1	0.8	2.8	0.7
전체	396	71	152	100.0	100.0	100.0

주: 기타는 영업용 건물 내 주택(상가, 공장 등), 오피스텔, 호텔, 여관, 기숙사 등을 의미함.
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일정한 집단의 연간 가구 총소득 분포를 보면, 2,000만 원 미만이 84.3%로 가장 높고, 다음으로 2,000만 원 이상~4,000만 원 미만이 12.4%로 나타났다(〈표 5-30〉 참조). 4,000만 원 이상의 비율은 다른 집단에 비해 가장 작은 값을 가졌다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 2,000만 원 미만이 67.6%로 가장 높고, 다음으로 2,000만 원 이상~4,000만 원 미만이 23.9%로 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 2,000만 원 미만이 49.3%로 가장 높았으나, 다른 집단에 비해 비율은 가장 낮았다. 6,000만 원 이상이 16%로 다른 집단에 비해 비율이 가장 높았으며, 연간 가구 총소득의 평균도 3,096만 원 정도로 다른 집단과 비교했을 때 가장 크게 나타났다.

〈표 5-30〉 맞춤형 급여 수급에 대한 일치 여부에 따른 연간 가구 총소득 분포

(단위: 가구, %, 만 원)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
2,000만 원 미만	334	48	74	84.3	67.6	49.3
4,000만 원 미만	49	17	37	12.4	23.9	24.7
6,000만 원 미만	8	3	15	2.0	4.2	10.0
6,000만 원 이상	5	3	24	1.3	4.2	16.0
전체	396	71	150	100.0	100.0	100.0
평균	1,447.2	1,985.7	3,095.5	-	-	-

주: 연간 가구 총소득 금액에 대해 무응답인 2가구는 결측 처리한 후 분석함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

다음으로 가구주의 인구사회학적 특성에 대한 분포를 살펴보았다. 먼저, 조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단의 가구주의 성별 분포를 보면, 남성이 50.6%로 절반을 상회하였다(〈표 5-31〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 남성이 52.5%로 절반 이상으로 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 60.1%이었으며, 다른 집단과 비교했을 때 가장 크게 나타났다.

〈표 5-31〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구주의 성별 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
남성	196	36	86	50.6	52.2	60.1
여성	191	33	57	49.4	47.8	39.9
전체	387	69	143	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 20가구임.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단의 가구주의 연령대 분포를 보면, 70~79세가 29.2%, 60~69세가 28.7%로 높게 나타났다(〈표 5-32〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 70~79세(39.1%)와 50~59세(26.1%)가 높은 편에 속하였다. 조사데이터에서는 수급을 받지

않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 60~69세가 27.3%, 50세 미만인 26.6%로 높은 편에 속하였다. 다른 집단에 비해 가구주의 연령대가 젊다고 볼 수 있다.

〈표 5-32〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구주의 연령대 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
50세 미만	54	6	38	14.0	8.7	26.6
50~59세	65	18	33	16.8	26.1	23.1
60~69세	111	13	39	28.7	18.8	27.3
70~79세	113	27	23	29.2	39.1	16.1
80세 이상	44	5	10	11.4	7.2	7.0
전체	387	69	143	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 20가구임.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단의 가구주의 혼인상태 분포를 보면, 사별/실종이 33.1%, 별거/이혼이 31.3%로 나타났다(〈표 5-33〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 혼인 중이 43.5%로 가장 높았고, 다음으로 사별/실종(27.5%)이 차지하였다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단도 혼인 중이 50.3%로 가장 높았고, 다른 집단과 비교해도 가장 큰 비율을 가졌다. 다음으로 사별/실종이 23.8%를 차지하였다.

〈표 5-33〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구주의 혼인상태 분포

(단위: 명, %)

조사데이터	빈도			비율		
	예	예	아니오	예	예	아니오
행정데이터	예	아니오	예	예	아니오	예
미혼	50	5	5	12.9	7.2	3.5
혼인 중	88	30	72	22.7	43.5	50.3
별거/이혼	121	15	32	31.3	21.7	22.4
사별/실종	128	19	34	33.1	27.5	23.8
전체	387	69	143	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 20가구임.
 자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단의 가구주의 최종 학력 분포를 보면, 초졸이 30.2%로 가장 높았으며, 다른 집단에 비해 초졸 이하의 비율이 46.5%로 가장 높았다(〈표 5-34〉 참조). 다음으로 고졸(28.4%)과 중졸(17.8%)로 나타났다. 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단은 중졸(26.1%)이 가장 높았다. 다음으로 고졸(21.7%)과 초졸(20.3%)의 비율이 유사하게 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 고졸이 32.2%로 가장 높았고, 다른 집단과 비교해도 가장 큰 비율을 가졌다. 다음으로 대졸 이상이 23.1%를 차지하였으며 다른 집단에 비해 학력이 높은 편이라고 볼 수 있다.

〈표 5-34〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구주의 최종 학력 분포

(단위: 명, %)

	빈도			비율		
	예	예	아니오	예	예	아니오
조사데이터	예	아니오	예	예	아니오	예
행정데이터	예	아니오	예	예	아니오	예
무학	63	12	16	16.3	17.4	11.2
초졸	117	14	29	30.2	20.3	20.3
중졸	69	18	19	17.8	26.1	13.3
고졸	110	15	46	28.4	21.7	32.2
대졸 이상	28	10	33	7.2	14.5	23.1
전체	387	69	143	100.0	100.0	100.0

주: 가구주의 정보가 결측인 경우가 20가구임.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터에서 모두 맞춤형 급여를 수급 받는다고 일치한 집단의 가구주의 경제활동 상태 분포를 보면, 비경제 활동인구가 76%로 가장 높았으며, 다른 집단에 비해 가장 높았다(〈표 5-35〉 참조). 조사데이터와 행정데이터가 일치하지 않는 경우를 보면, 조사데이터에서는 수급을 받는다고 했으나 행정데이터에서는 수급 받지 않고 있다고 한 집단도 비경제 활동인구가 50.7%로 절반을 상회하였다. 다음으로 임금근로자가 30.4%로 높게 나타났다. 조사데이터에서는 수급을 받지 않는다고 했으나 행정데이터에서는 수급 받고 있다고 한 집단은 임금근로자가 42.7%로 가장 높았는데, 다른 집단과 비교했을 때 가장 높은 비율을 가졌다. 다음으로 비경제 활동인구가 26.6%로 나타났다. 한편, 기타(14.7%)도 다른 집단과 비교했을 때 높은 편에 속하였다.

〈표 5-35〉 맞춤형 급여 수급에 대한 일치 여부에 따른 가구주의 경제활동 상태 분포

(단위: 명, %)

조사데이터	빈도			비율		
	예	예	아니오	예	예	아니오
행정데이터	예	아니오	예	예	아니오	예
임금근로자	61	21	61	15.8	30.4	42.7
고용주/자영업자	16	7	23	4.1	10.1	16.1
비경제 활동인구	294	35	38	76.0	50.7	26.6
기타	16	6	21	4.1	8.7	14.7
전체	387	69	143	100.0	100.0	100.0

주: 1) 가구주의 정보가 결측인 경우가 20가구임.

2) 기타는 무급가족종사자, 자활근로, 공공근로, 희망근로, 노인 일자리가 해당함.

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

조사데이터와 행정데이터와의 정합성 검토 결과를 요약하면 다음과 같다. 먼저 통합데이터에서의 기초연금과 맞춤형 급여의 수급 여부 현황의 경우, 조사데이터와 행정데이터 간의 차이는 각각 2.8%p, 1.2%p로 작게 나타났다. 조사데이터와 행정데이터에서의 사회보장 급여를 수급 받는 비율이 유사하다고 볼 수 있다.

가구 응답의 정합성을 살펴보기 위하여 사회보장 급여 수급 여부에 대해 조사데이터와 행정데이터가 일치하는지를 비교하였다. 조사데이터와 행정데이터의 값이 모두 일치하는 가구의 비율은 기초연금의 경우 93.6%이고, 맞춤형 급여의 경우 96.6%로 나타났다. 반면에 조사데이터에서는 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 가구의 비율은 기초연금의 경우 4.6%이고, 맞춤형 급여의 경우 1.1%였다. 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 가구의 비율은 기초연금의 경우 1.8%이고 맞춤형 급여의 경우 2.3%로 나타났다. 조사데이터와 행정데이터 간 수급 여부가 일치하

지 않는 경우, 기초연금은 조사데이터에서 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 가구의 비율이 높았던 반면에, 맞춤형 급여는 반대의 경우가 높았다.

조사데이터와 행정데이터의 값이 일치하지 않는 요인은 다음과 같은 것들을 생각해 볼 수 있다. (1) 조사데이터의 응답 시점과 행정데이터의 추출 시점 간의 차이 때문으로 볼 수 있다. 맞춤형 급여는 조사데이터와 행정데이터의 시점이 2019년 12월 31일로 동일하였다. 그러나 기초연금은 조사데이터의 경우 응답 시점이 2019년 1월 1일~12월 31일까지인 반면에, 행정데이터는 2019년 12월 31일이 추출 시점으로 동일하지 않다고 볼 수 있다. (2) 조사데이터의 경우 사회보장 급여 수급 여부를 가구 조사표에서 조사하고 있는데, 이는 대표가구원이 모두 응답하는 구조이므로 정확히 파악하지 못하는 경우가 있을 수 있다. (3) 행정데이터는 행정 처리 과정에서 생성되는 데이터이므로, 조사데이터와 같이 정제된 데이터가 아니다. 이러한 특성이 기인했을 가능성도 있다고 생각한다. (4) 두 데이터를 결합할 때, 결합키로 이름, 생년월일, 성별을 사용하였다. 이때 동일인이더라도 주민등록 생년월일과 실제 생년월일(음력, 양력)의 차이, 또는 이름과 성별의 차이가 생길 수 있다.

한편, 조사데이터와 행정데이터의 사회보장 급여 수급 여부가 일치하지 않는 집단의 비율은 낮은 편이었으나, 이러한 차이를 가구 특성과 가구주의 인구사회학적 특성 분포를 통해 살펴보았다. 기초연금의 경우 조사데이터에서 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 집단은 다른 집단에 비해 서울 및 경기에서 거주하고, 2인 가구로 구성되고, 주거형태는 단독주택과 아파트에 거주하고, 연간 가구 총소득이 높은 편으로 나타났다. 가구주의 인구사회학적 특성도 다른 집단에 비해 남성의 비율이 높았고, 연령대도 젊은 편에 속하였고, 혼인 중의 비율

이 높았고, 최종 학력도 고졸 이상으로 높은 편이었고, 경제활동 상태도 근로(임금근로자, 고용주/자영업자)하고 있는 비율이 높은 편이었다.

맞춤형 급여의 경우 조사데이터에서 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 집단은 다른 집단에 비해 광역시, 서울 및 경기에서 거주하고, 주로 1, 2인 가구 구성이 높았는데 4인 이상 가구의 비율도 다른 집단에 비해 높게 나타났다. 주거형태는 아파트에 거주하고, 연간 가구 총소득이 높은 편으로 나타났다. 가구주의 인구사회학적 특성도 다른 집단에 비해 남성의 비율이 높았고, 연령대도 50세 미만의 비율이 높아서 젊은 편에 속하였다. 혼인 중의 비율이 높았고, 최종 학력도 고졸 이상으로 높은 편이었고, 경제활동 상태도 근로(임금근로자, 고용주/자영업자)하고 있는 비율이 높은 편이었다.

종합해 보면, 조사데이터에서 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 집단은 서울 및 경기도, 광역시에 거주하고 아파트에 거주하며 연간 가구 총소득이 높다고 볼 수 있다. 가구주의 특성은 남성이며 연령대가 젊은 편이고 혼인 중이고, 최종 학력도 고졸 이상으로 높은 편이고 경제활동을 하고 있는 것으로 나타났다. 이러한 결과는 한국의료패널조사의 표본이탈 경향 분석에서 나타난 패널가구 및 패널가구의 특성과 유사하게 나타났다(이혜정 외, 2023). 즉, 조사에 참여하더라도 소극적으로 응답하는 경향을 가진다고 볼 수 있다.

통합데이터의 정합성 검토를 통해 행정데이터의 활용 가능성을 확인하였다. 행정데이터와의 비교를 통해 조사데이터의 응답의 정확성을 파악할 수 있으며, 조사 수행 및 실사 관리에서 더 정확한 응답을 수집할 수 있도록 유용한 정보로 활용될 수 있다. 또한 향후 행정데이터의 정보를 연계하여 응답자의 부담을 낮추는 데 기여할 수 있다고 생각한다.

2. 사회보장급여 수급자의 의료비 지출 특성 분석

가. 급여 자격별 의료비 부담 수준 비교

급여 자격별 의료비 부담 수준(5점 척도)을 살펴보면, 의료급여 1종 수급 가구(평균 3.1점)가 의료급여 2종 수급 가구(평균 3.3점)에 비하여 의료비 부담 수준이 낮은 것으로 나타났다. 의료급여 2종 수급 가구는 1종 수급가구에 비하여 본인부담금 수준이 높기 때문에 상대적으로 의료비 지출에 대한 부담이 크게 나타난 것이라 이해할 수 있다.

기초연금 수급 자격별 의료비 부담 수준을 살펴보면, 단독 수급 가구가 약 2.5점, 부부1인 수급 가구가 약 2.6점, 부부2인 가구가 약 2.3점으로 나타났다. 특히 단독 수급 가구가 부부2인 가구에 비하여 의료비 지출에 대한 부담이 큰 것으로 나타났다. 이러한 결과에는 노인 1인 가구의 의료비 부담과 단독 수급 가구의 상대적으로 낮은 급여액 등이 반영되어 있다고 볼 수 있다.

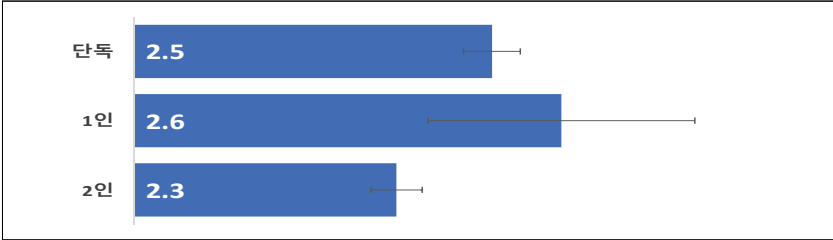
〈표 5-36〉 급여별·자격별 의료비 부담 수준(5점 척도)

급여 구분		의료급여 1종	의료급여 2종	-	전체
의료급여 수급가구	평균(표준편차)	3.1(1.3)	3.3(1.2)	-	3.1(1.3)
	가구 수	304	60	-	364
급여 구분		단독	부부1인	부부2인	전체
기초연금 수급가구	평균(표준편차)	2.5(1.2)	2.6(1.1)	2.3(1.1)	2.4(1.1)
	가구 수	1038	40	1128	2206

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-2] 기초연금 자격별 의료비 부담 수준(5점 척도)

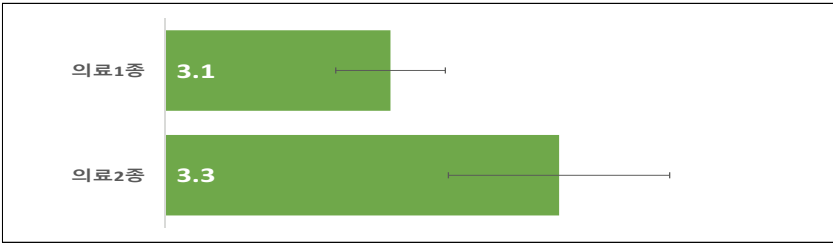
(단위: 점)



주: 각 막대그래프 내 표준오차⁴¹⁾ 표기
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-3] 의료급여 자격별 의료비 부담 수준(5점 척도)

(단위: 점)



주: 각 막대그래프 내 표준오차 표기
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

나. 급여 특성별 의료비 지출 특성 비교

급여의 특성별 의료비 지출을 살펴보기 위하여 연간 사회보장급여 수급액과 연간 의료비 지출금액을 각각 5분위, 10분위로 구분하였으며, 구체적인 사항은 아래 표와 같다(〈표 5-37〉 참조).

41) 표준오차 = $\frac{\text{표준편차}}{\sqrt{\text{관측치}}}$

연간 의료비의 경우, 지출금액이 0원인 사례를 포함하여 분위를 설정하였다. 반면에 생계급여 및 기초연금 수급액에 대한 분위 설정 시에는 수급 자격을 가지면서 급여액이 0원을 초과하는 경우에 한하여 분석 대상에 포함시켰다.

한편 가구원 수가 많을수록 의료비 지출액의 규모가 커질 수 있다. 또한 급여액의 수준은 가구원 수에 따라 상대적으로 인식될 수 있기 때문에 분위별 가구 분포를 제시할 때에는 가구균등화지수를 적용하여 분위 구간을 설정·분석하였다.

〈표 5-37〉 의료비 및 급여액 분위 구분

(단위: 만 원)

구분	가구 수	분위 구분	최솟값/최댓값		비고
			가구원 전체	가구균등화 적용	
연간 의료비	6,343	1분위	0/13.8	0/9.6	의료비가 0만 원인 가구 포함
		2분위	13.9/29.9	9.6/20.6	
		3분위	30.0/47.7	20.7/32.4	
		4분위	47.8/68.7	32.4/46.6	
		5분위	68.8/95.7	46.6/63.8	
		6분위	95.8/129.8	63.8/85.6	
		7분위	129.9/176.3	85.6/117.2	
		8분위	176.4/246.4	117.2/163.8	
		9분위	246.5/390.5	163.8/260.7	
		10분위	390.6/865.0	261/865	
생계급여 연간 총액	365	1분위	3/269	2.3/227.5	미수급자 및 수급액이 0만 원인 가구 제외
		2분위	272/407	227.7/275	
		3분위	409/591	277/414.4	
		4분위	592/702	417.9/603	
		5분위	703/1,658	603.2/1,349	
기초연금 연간 총액	2,206	1분위	15/243	11.5/171.8	
		2분위	244/303	172.5/217	
		3분위	304/345	217.1/303	
		4분위	349/486	303.3/343.7	
		5분위	496/831	344/487.9	

주: 가구균등화를 적용한 값은 연간 의료비 및 급여액을 가구원 수의 제곱근으로 나누어 산출함.
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

수급 가구별 의료비 지출 수준을 살펴보면 <표 5-38> 및 [그림 5-4]와 같다. 먼저, 생계급여 수급 가구와 기초연금 수급 가구의 의료비 지출 특성을 살펴보면, 의료비 지출 분위가 낮은 구간에 가구가 집중되어 있었다. 또한 의료비 지출 분위가 증가함에 따라 평균 의료비 증가 폭이 점차 커지는 특성을 보였다. 특히 기초연금 수급 가구의 경우, 단독 수급 가구와 부부2인 가구의 의료비 지출 차이는 의료비 지출 분위가 높은 구간에서 두드러진 것으로 확인되었다.

<표 5-38> 의료비 분위별·급여별 의료비 평균 비교

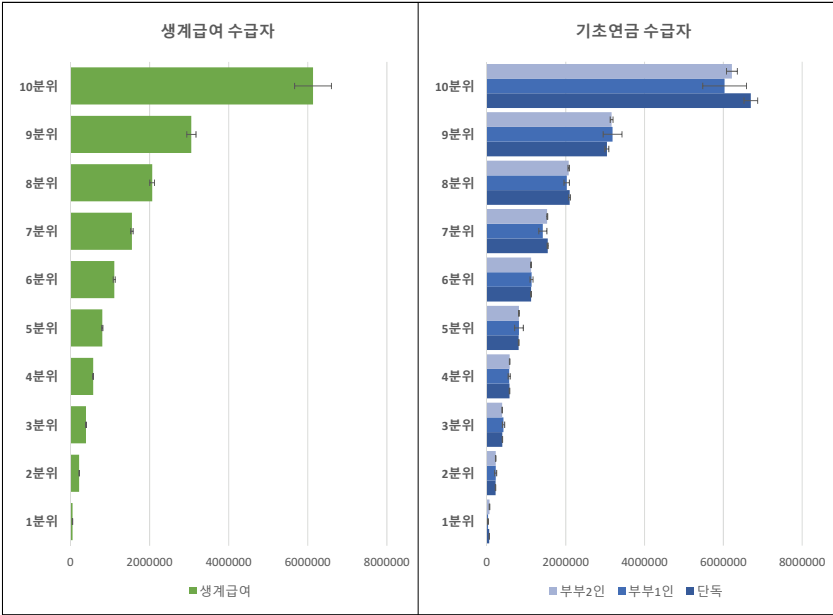
급여		구분	연간 의료비(단위: 원)		
			평균	표준편차	가구 수
맞춤형 급여	생계 급여	1분위	47,300.89	38,828.81	113
		2분위	215,550	48,706.81	60
		3분위	389,333.3	55,378.76	39
		4분위	569,206.9	53,538.62	29
		5분위	803,062.5	65,700.81	16
		6분위	1,104,053	106,447.4	19
		7분위	1,551,158	136,300.8	19
		8분위	2,063,067	231,556.7	15
		9분위	3,056,200	444,940.5	15
		10분위	6,133,133	1,802,357	15
기초 연금	단독	1분위	60,720.93	42,610.7	129
		2분위	218,991.5	47,066.51	118
		3분위	393,471.5	50,648.31	123
		4분위	573,000	62,580.43	104
		5분위	805,113.2	77,076.25	106
		6분위	1,120,247	91,843.99	93
		7분위	1,542,798	133,504.9	84
		8분위	2,097,649	200,872.8	77
		9분위	3,043,819	419,301.1	72
		10분위	6,695,667	1,670,509	93

급여		구분	연간 의료비(단위: 원)		
			평균	표준편차	가구 수
	부부 1인	1분위	29,000	7,071.068	2
		2분위	221,800	52,083.59	5
		3분위	420,200	64,060.13	5
		4분위	565,750	50,598.91	4
		5분위	813,000	156,977.7	2
		6분위	1,132,000	79,132.8	4
		7분위	1,419,667	178,374.1	3
		8분위	2,026,333	120,608.2	3
		9분위	3,188,250	473,815.3	4
		10분위	6,031,571	1,471,197	7
	부부 2인	1분위	66,891.3	40,773.75	46
		2분위	222,898.7	46,469.07	79
		3분위	388,437.5	49,406.82	80
		4분위	577,367.9	57,145.48	106
		5분위	812,296.3	74,803.95	108
		6분위	1,122,055	99,052.52	127
		7분위	1,530,448	135,651	145
		8분위	2,075,145	191,375.4	131
		9분위	3,165,145	412,771.1	138
		10분위	6,216,336	1,697,660	152

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-4] 의료비 분위별 연간 의료비 평균

(단위: 원)



주: 각 막대그래프 내 표준오차 표기

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

연간 사회보장급여 수급액 분위별 연간 의료비 지출액 평균을 살펴보면 <표 5-39> 및 [그림 5-5]와 같다. 먼저, 생계급여 5분위에 해당되는 가구의 평균 연간 의료비 지출액은 약 1,101천 원으로, 가장 높은 수준인 것으로 나타났다. 다음으로 급여액 1분위에 해당되는 집단의 연간 의료비 평균이 약 907천 원으로 높게 나타났다. 반면 급여액 2분위에 해당되는 집단은 다른 집단에 비해 의료비 지출 수준이 낮은 것으로 나타났다. 이를 통하여 생계급여 수급액과 의료비 지출 간 관련성보다 생계급여 수급액 수준이 높은 집단과 낮은 집단 간의 이질적 특성을 확인하는 것이 중요함을 알 수 있다.

기초연금 5분위에 해당되는 가구의 평균 연간 의료비 지출액 추이를 살펴보면, 수급액 1분위에서 3분위까지는 연간 의료비 지출 수준이 낮아지다가 4분위에서 급격하게 증가(연간 평균 2,212천 원)하는 양상을 보였다. 급여 자격별로 살펴보면, 단독 수급 가구는 수급액 1분위부터 3분위까지 급여 대상 가구가 집중되어 있으며, 부부2인 가구는 수급액 4분위에 집중되어 있음을 알 수 있다. 이는 기초연금의 급여액 산정 기준과 관련이 있다고 볼 수 있다. 「2020년 기초연금 사업안내(보건복지부, 2020)」에 따르면, 단독 가구의 최대 수급액은 약 3,057천 원⁴²⁾며, 기초연금 연간 급여액 3분위에 해당되는 값이다(pp.5-7). 한편, 부부2인 가구의 최대 수급액은 약 4,891천 원⁴³⁾이며, 기초연금 연간 급여액 4분위에 해당되는 값이다. 따라서 단독 수급 가구는 3분위 가구까지의 특성을 분석하는 것이 적절하며, 이에 따라 단독 수급 가구는 수급액 규모가 클수록 의료비 지출 수준이 낮은 것으로 볼 수 있다. 한편, 부부2인 가구는 4분위에 해당되는 가구 특성을 살펴보는 것이 중요하며, 상대적으로 단독 수급 가구에 비하여 의료비 지출 수준이 높다고 볼 수 있다.

연간 의료비 지출 규모와 연간 사회보장급여 수급액의 분위별 가구 수 분포를 살펴보면 <표 5-40>과 같다. 먼저, 생계급여 수급 가구의 경우, 생계급여 수급액이 높을수록 연간 의료비 지출 규모가 작은 특성을 보였다([그림 5-6] 참조). 한편, 기초연금 수급 가구의 경우, 수급액이 높을수록 연간 의료비 지출 규모가 큰 것으로 보이거나 이러한 경향은 생계급여 사례에 비하여 완만한 것으로 나타났다([그림 5-7] 참조).

42) 기준연금액 254,760원 × 12개월

43) (기준연금액 254,760원 × 2인 × 부부감액비율 0.8) × 12개월

〈표 5-39〉 급여 분위별 의료비 평균 비교

(단위: 원, 가구)

급여		구분	급여별 분위				
			1분위	2분위	3분위	4분위	5분위
맞춤형 급여	생계 급여	평균	906,925.37	441,000	793,971.4	890,060.6	1,101,448
		표준편차	1,776,942.4	726,660.3	1,332,125	1,270,241	1,719,702
		가구 수	67	65	70	66	67
기초 연금	전체	평균	1,860,401	1,621,992	1,287,497	2,212,489	1,544,918
		표준편차	2,094,352	1,927,674	1,927,577	1,994,008	1,833,348
		가구 수	431	759	368	507	85
	단독	평균	1,951,288	1,477,963	1,308,263	1,661,000	2,049,286
		표준편차	2,421,299	1,829,362	2,003,282	1,167,373	2,064,047
		가구 수	66	570	335	7	21
	부부 1인	평균	2,131,400	1,968,067	1,192,000	568,000	1,896,000
		표준편차	2,399,710	2,227,567	1,332,189	-	-
		가구 수	20	15	2	1	1
	부부 2인	평균	1,827,304	2,063,977	1,069,258	2,223,521	1,371,222
		표준편차	2,012,485	2,144,219	812,503.6	2,003,421	1,749,497
		가구 수	345	174	31	499	63

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

기초연금 수급 가구를 단독 가구와 부부2인 가구로 구분하여 살펴보면, 단독 수급 가구는 연금수급액이 작을수록 의료비 지출이 적은 것으로 나타났으며, 이러한 특성은 급여액 3분위에 해당되는 집단을 중심으로 나타났다([그림 5-8] 참조). 부부2인 가구는 연금수급액이 클수록 의료비 지출이 큰 것으로 나타났으며, 이러한 특성은 급여액 4분위 집단을 중심으로 나타났다([그림 5-9] 참조).

[그림 5-5] 급여액 분위별 의료비 평균

(단위: 원)



주: 1) 기초연금 부부1인 결과는 상대적으로 사례 수가 적어 그림을 통한 비교 생략

2) 각 막대그래프 내 표준오차 표기

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

〈표 5-40〉 연간 의료비 및 급여액의 분위별 가구 수 분포-1

(단위: 가구)

급여	구분		급여액 분위				
			1분위	2분위	3분위	4분위	5분위
맞춤형 급여 (생계급여)	연간 의료비	1분위	19	35	15	19	18
		2분위	12	14	14	11	6
		3분위	9	7	10	8	9
		4분위	6	4	5	9	9
		5분위	2	1	4	2	5
		6분위	2	2	4	9	4
		7분위	2	1	3	4	3
		8분위	4	4	5	2	2
		9분위	4	2	2	6	3
		10분위	4	1	4	1	4
기초 연금 (전체)	연간 의료비	1분위	32	25	39	10	48
		2분위	39	28	52	23	30
		3분위	40	36	67	24	20
		4분위	50	32	68	39	24
		5분위	47	48	57	40	24
		6분위	47	31	66	49	23
		7분위	52	39	61	55	24
		8분위	60	37	62	55	23
		9분위	51	40	73	68	15
		10분위	68	55	72	71	11

주: 의료비 및 급여액 가구균등화 후 분위 산출
자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

〈표 5-41〉 연간 의료비 및 급여액의 분위별 가구 수 분포-2

(단위: 가구)

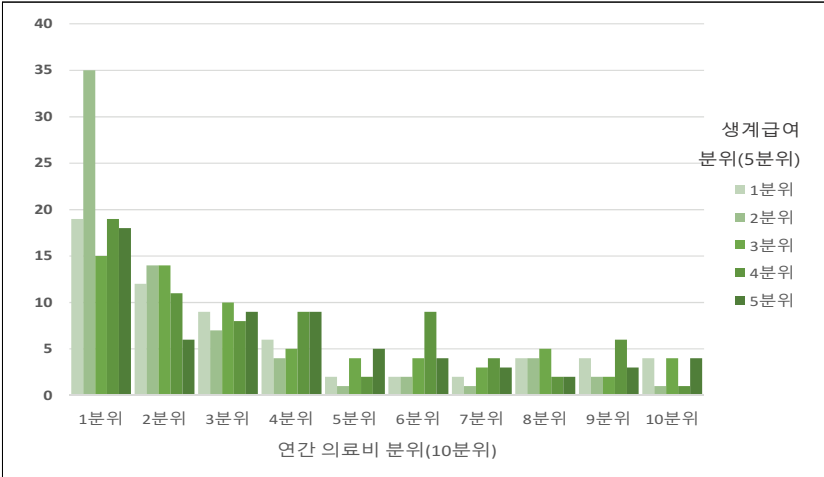
급여	구분		급여액 분위				
			1분위	2분위	3분위	4분위	5분위
기초 연금 (단독)	연간 의료비	1분위	9	15	44	4	57
		2분위	9	14	72	5	18
		3분위	8	23	65	8	19
		4분위	7	17	50	6	24
		5분위	5	20	51	8	22
		6분위	15	24	38	2	14
		7분위	6	25	40	4	9
		8분위	11	14	36	2	14
		9분위	9	23	33	2	5
		10분위	27	25	33	2	6
기초 연금 (부부1인)	연간 의료비	1분위	0	2	0	0	0
		2분위	2	2	1	0	0
		3분위	4	1	0	0	0
		4분위	2	1	0	1	0
		5분위	1	1	0	0	0
		6분위	3	1	0	0	0
		7분위	1	2	0	0	0
		8분위	0	1	1	1	0
		9분위	3	1	0	0	0
		10분위	4	3	0	0	0
기초 연금 (부부2인)	연간 의료비	1분위	23	6	4	9	4
		2분위	28	13	12	20	6
		3분위	24	14	9	23	10
		4분위	37	18	15	28	8
		5분위	33	14	17	40	4
		6분위	40	18	21	44	4
		7분위	47	16	16	57	9
		8분위	46	15	22	45	3
		9분위	39	20	18	59	2
		10분위	43	22	19	64	4

주: 의료비 및 급여액 가구균등화 후 분위 산출

자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-6] 연간 의료비 및 생계급여 분위별 가구 분포

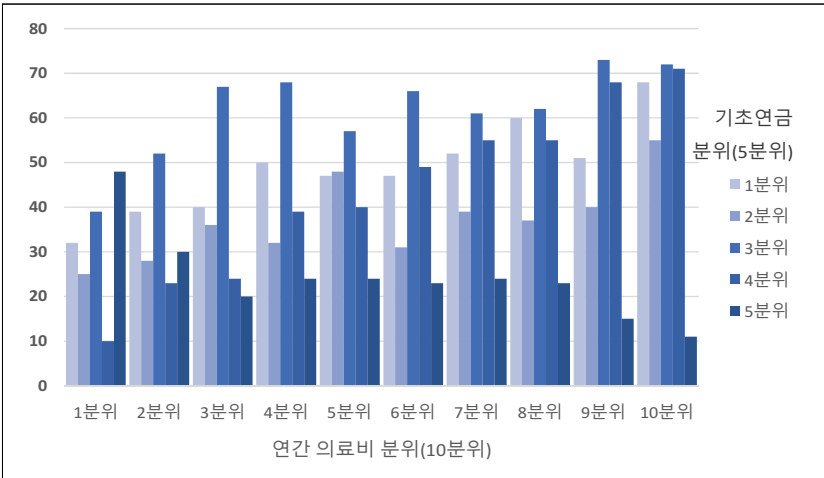
(단위: 가구)



자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-7] 연간 의료비 및 기초연금(전체) 분위별 가구 분포

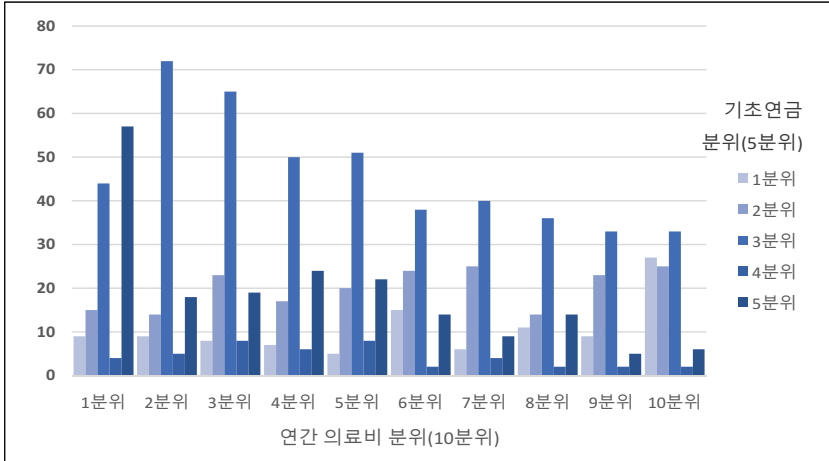
(단위: 가구)



자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-8] 연간 의료비 및 기초연금(단독) 분위별 가구 분포

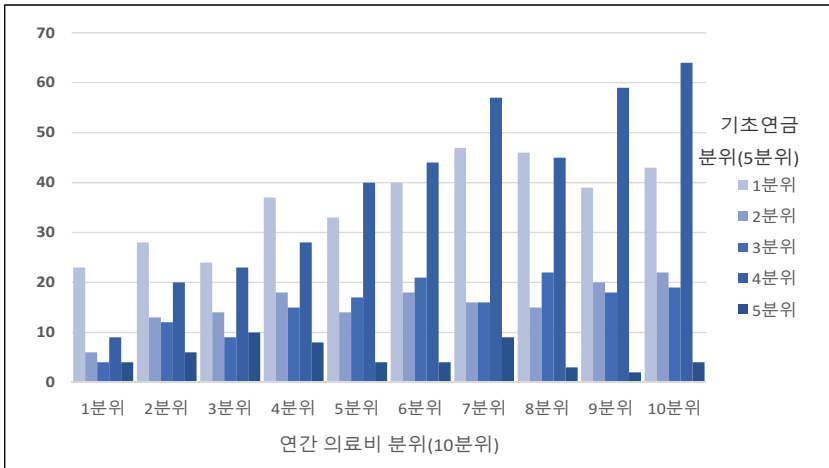
(단위: 가구)



자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

[그림 5-9] 연간 의료비 및 기초연금(부부2인) 분위별 가구 분포

(단위: 가구)



자료: 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템 행정데이터의 통합데이터를 분석함.

사회보장 급여 수급자의 의료비 지출 특성 분석 결과를 요약하면 다음과 같다. 첫째, 맞춤형 급여의 수급 규모가 작은 가구일수록 의료비 지출 수준이 낮은 것으로 나타났다. 한편 본인부담금 수준이 높은 의료급여 2종 급여의 대상은 상대적으로 의료비 지출에 대한 부담 수준이 높은 것으로 나타났다. 이를 통하여 맞춤형 급여 수급액이 높은 저소득층과 근로무능력층의 의료비 지출이 의료서비스의 필요에 따라 충분히 이루어지고 있는지를 살펴볼 필요가 있음을 알 수 있다. 또한 각 수급 자격에 따른 본인부담금 수준과 의료비에 대한 부담 등을 다각적으로 살펴볼 필요가 있음을 알 수 있다.

둘째, 기초연금 수급 가구는 수급 자격에 따라 의료비 지출 특성이 상이하게 나타났다. 기초연금 ‘단독’ 수급 가구는 기초연금 ‘부부2인’ 가구에 비해 의료비 지출 규모가 작았으나 의료비에 대한 부담이 큰 것으로 나타났다. 또한 기초연금 ‘단독’ 수급 가구는 급여 수준이 낮을수록 의료비 지출 규모가 큰 것으로 나타났으며, 기초연금 ‘부부2인’ 가구는 급여 수준이 높을수록 의료비 지출 규모가 큰 경향을 보였다. 이를 통하여 급여 자격별로 의료비 지출 특성을 구분하여 분석하여야 함을 알 수 있다.

특히 단독 수급 가구의 경우, 기준연금액 100%를 수급하는 가구의 의료비 지출 수준이 상대적으로 낮다는 점에 주목할 필요가 있다. 이 집단은 상대적으로 연금 수급을 통한 소득이 높지만, 소득·재산의 수준이 상대적으로 낮을 가능성이 높다. 향후 이들의 건강상태나 의료비 지출에 대한 부담 등이 실제 의료비 지출에 대한 위축으로 작용하게 되는지 여부를 살펴볼 필요가 있다. 한편, 부부2인 가구의 경우 기준연금액 100%를 수급하는 가구의 의료비 지출이 상대적으로 높다는 점을 확인하였다. 이들은 부부가 모두 65세 이상이면서 기초연금 외의 소득이 없을 가능성이 높기 때문에 과도한 의료비 지출로 인한 가계부담 증대의 위험을 가지고 있다. 향후에 부부2인 가구의 기초연금 소득의 대부분을 의료비로 지출하

는 가구를 확인하여 이에 대한 보충적 지원 정책을 뒷받침할 수 있는 자료를 생성할 필요가 있다.

제3절 소결

가명정보 결합 절차에 따라 실제 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터를 통합하여 통합데이터를 구축하였다. 그리고 구축된 통합데이터에 대한 실증 분석을 실시하였다.

한국의료패널조사는 의료비 지출을 포함한 각종 보건의료 정보를 포함하고 있는 조사데이터이다. 그런데 사회보장 급여에 대한 세부적인 정보가 부족하거나, 응답자의 기억에 의존하여 응답을 받기 때문에 정확도가 낮아질 수 있다. 그러므로 가명정보 결합을 통해 구축된 통합데이터를 사용한다면 조사데이터가 가지는 부족한 부분을 보완할 수 있다는 점이 가장 큰 강점이라고 볼 수 있다.

먼저, 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터의 가명정보 결합을 위해 사전준비, 결합신청, 가명처리, 결합수행, 반출 및 관리 절차에 따라 실시하였다. 가명정보 결합 절차⁴⁴⁾는 올해 3월부터 시작하여 10월 중순에 완료하였고, 그 이후부터 데이터 분석을 할 수 있었다. 가명정보를 결합한 통합데이터를 제공받는 것에만 8개월 남짓 소요되었다. 1년 단위의 과제 수행에서는 매우 빠듯할 수 있는 기간이라고 볼 수 있다. 보통 결합전문기관에 따라서는 1년 이상 소요되는 것이 다반사이기 때문에 현재로서는 연구 기간이 충분하게 확보되었을 때 안정적

44) 가명정보 결합 절차를 본격적으로 진행하기 전인 1~2월에는 관련 제도, 선행연구 등을 검토하는 과정을 가졌음.

으로 진행할 수 있다고 생각한다. 한편, 5계 단계 중에서 사전준비 단계가 5개월로 가장 오랜 시간이 걸렸다. 사전준비 단계에서는 한국의료패널조사 데이터와 사회보장행정시스템의 행정데이터를 사용할 수 있도록 승인 받는 과정을 거쳤다. 다음으로 가명처리 단계에 시간을 많이 할애한 것을 알 수 있다. 한국의료패널조사 데이터에 대한 가명처리를 하였고, 적정성 검토를 통해 추가 가명처리를 하였다. 그리고 가명정보를 결합한 통합데이터에 대한 적정성 검토 결과에 따라 다시 추가 가명처리를 실시하였다. 이렇듯 가명처리가 여러 번에 걸쳐 진행되었다.

다음은 가명정보 결합을 통해 구축된 통합데이터를 사용하여 실증 분석한 분석 결과를 요약하였다. 분석 주제는 (1) 조사데이터와 행정데이터의 정합성 검토와 (2) 사회보장 급여 수급자의 의료비 지출 특성 분석이었다.

첫 번째 분석 결과를 살펴보면, 조사데이터와 행정데이터에서의 사회보장 급여를 수급 받는 비율이 유사하게 나타났다. 조사데이터와 행정데이터 간 수급 여부가 일치하지 않는 경우, 기초연금은 조사데이터에서 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 가구의 비율이 높았던 반면에, 맞춤형 급여는 반대의 경우가 높았다. 한편, 맞춤형 급여 중에서 주거급여가 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서 수급 받는다고 한 비율이 가장 높게 나타났다.

조사데이터와 행정데이터의 값이 일치하지 않는 요인은 다음의 것들을 생각해 볼 수 있다. (1) 조사데이터의 응답 시점과 행정데이터의 추출 시점 간의 차이 때문으로 볼 수 있다. 맞춤형 급여는 조사데이터와 행정데이터의 시점이 2019년 12월 31일로 동일하였다. 그러나 기초연금은 조사데이터의 경우 응답 시점이 2019년 1월 1일~12월 31일까지인 반면에, 행정데이터는 2019년 12월 31일이 추출 시점으로 동일하지 않다고 볼 수 있다. (2) 조사데이터의 경우 사회보장 급여 수급 여부를 가구 조사

표에서 조사하고 있는데, 이는 대표 가구원이 모두 응답하는 구조이므로 정확히 파악하지 못하는 경우가 있을 수 있다. (3) 행정데이터는 행정 처리 과정에서 생성되는 데이터이므로, 조사데이터와 같이 정제된 데이터가 아니다. 이러한 특성이 기인했을 가능성도 있다고 생각한다. (4) 두 데이터를 결합할 때, 결합기로 이름, 생년월일, 성별을 사용하였다. 이때 동일인이더라도 주민등록 생년월일과 실제 생년월일(음력, 양력)의 차이, 또는 이름과 성별의 차이가 생길 수 있다.

조사데이터와 행정데이터의 사회보장 급여 수급 여부가 일치하지 않는 집단의 비율은 낮은 편이었으나, 이러한 차이를 가구 특성과 가구주의 인구사회학적 특성 분포를 통해 살펴보았다. 조사데이터에서 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 집단은 서울 및 경기도, 광역시에 거주하고, 아파트에 거주하며 연간 가구 총소득이 높다고 볼 수 있다. 가구주의 특성은 남성이며, 연령대가 젊은 편이고, 혼인 중이고, 최종 학력도 고졸 이상으로 높은 편이고, 경제활동을 하고 있는 것으로 나타났다. 이러한 결과는 한국의료패널조사의 표본이탈 경향 분석에서 나타난 패널가구 및 패널가구원의 특성과 유사하게 나타났다(이혜정 외, 2023). 즉, 조사를 참여하더라도 소극적으로 응답하는 경향을 가진다고 볼 수 있다.

통합데이터의 정합성 검토를 통해 행정데이터의 활용 가능성을 확인하였다. 행정데이터와의 비교를 통해 조사데이터의 응답의 정확성을 파악할 수 있으며, 조사 수행 및 실사 관리에서 좀 더 정확한 응답을 수집할 수 있도록 유용한 정보로 활용될 수 있다. 또한 향후 행정데이터의 정보를 연계하여 응답자의 부담을 낮추는 데 기여할 수 있다고 생각한다.

두 번째 분석 결과를 살펴보면, 두 가지로 요약할 수 있다. 첫째, 맞춤형 급여의 수급 규모가 작은 가구일수록 의료비 지출 수준이 낮은 것으로

나타났다. 한편 본인부담금 수준이 높은 의료급여 2종 급여의 대상은 상대적으로 의료비 지출에 대한 부담 수준이 높은 것으로 나타났다. 이를 통하여 맞춤형 급여 수급액이 높은 저소득층과 근로무능력층의 의료비 지출이 의료서비스의 필요에 따라 충분히 이루어지고 있는지를 살펴볼 필요가 있음을 알 수 있다. 또한 각 수급 자격에 따른 본인부담금 수준과 의료비에 대한 부담 등을 다각적으로 살펴볼 필요가 있음을 알 수 있다.

둘째, 기초연금 수급 가구는 수급 자격에 따라 의료비 지출 특성이 상이하게 나타났다. 기초연금 ‘단독’ 수급 가구는 기초연금 ‘부부2인’ 가구에 비해 의료비 지출 규모가 작았으나 의료비에 대한 부담이 큰 것으로 나타났다. 또한 기초연금 ‘단독’ 수급 가구는 급여 수준이 낮을수록 의료비 지출 규모가 큰 것으로 나타났으며, 기초연금 ‘부부2인’ 가구는 급여 수준이 높을수록 의료비 지출 규모가 큰 경향을 보였다. 이를 통하여 급여 자격별로 의료비 지출 특성을 구분하여 분석하여야 함을 알 수 있다.

특히 단독 수급 가구의 경우, 기준연금액 100%를 수급하는 가구의 의료비 지출 수준이 상대적으로 낮다는 점에 주목할 필요가 있다. 이 집단은 상대적으로 연금 수급을 통한 소득이 높지만, 소득·재산의 수준이 상대적으로 낮을 가능성이 높다. 향후 이들의 건강상태나 의료비 지출에 대한 부담 등이 실제 의료비 지출에 대한 위축으로 작용하게 되는지 여부를 살펴볼 필요가 있다. 한편, 부부2인 가구의 경우 기준연금액 100%를 수급하는 가구의 의료비 지출이 상대적으로 높다는 점을 확인하였다. 이들은 부부가 모두 65세 이상이면서 기초연금 외의 소득이 없을 가능성이 높기 때문에 과도한 의료비 지출로 인한 가계부담 증대의 위험을 가지고 있다. 향후에 부부2인 가구의 기초연금 소득의 대부분을 의료비로 지출하는 가구를 확인하여 이에 대한 보충적 지원 정책을 뒷받침할 수 있는 자료를 생성할 필요가 있다.



제6장

결론

제1절 연구 결과 요약 및 함의

제2절 가명정보 결합 제도 활성화를 위한 방안

제3절 향후 과제

제6장 결론

보건복지 분야 데이터 경제 활성화를 위한 다출처 연계, 통합, 활용 방안 연구는 2년간 실시하였다. 제1절에서는 주요 연구 결과를 요약하였고, 제2절에서는 가명정보 결합 제도 활성화를 위한 방안을 제안하였다. 마지막 제3절에서는 향후 과제를 제시하였다.

제1절 연구 결과 요약 및 함의

제2장에서는 가명정보 결합 및 반출 절차에 대해 구체적으로 살펴보았다. 활용하고자 하는 가명정보에 따라 절차가 조금 상이한 경우도 있기 때문에 각각의 가이드라인을 아울러서 정리하였다. 가명정보 결합은 개인 식별이 불가능하도록 가명처리된 가명정보에 대하여 결합전문기관이 두 개 이상의 데이터를 결합키를 이용하여 통계작성, 과학적 연구, 공익적 기록보존 등의 목적에 한해 결합하는 것이다. 가명정보 결합 및 반출 절차는 5단계로, 사전준비, 결합신청, 가명처리, 결합수행, 반출 및 관리의 순서로 진행된다. 각각의 단계에서 데이터 이용자와 데이터 제공자, 결합전문기관, 결합키 생성기관은 유기적으로 관계하면서 가명정보 결합 및 반출 절차를 수행한다. 가명정보 결합은 개인정보를 정보 주체의 동의 없이 사용할 수 있기 때문에, 개인 식별 위험성을 없애고 개인정보의 안전한 사용을 위해 복잡한 절차를 밟도록 법으로 규제하고 있다. 따라서 각 단계에서 주의해야 할 사항들을 숙지해야만 가명정보 결합에 소요되는 시간을 줄이면서, 가명정보 결합 신청 시 제출했던 연구계획서의 연구 목적에 맞는 정확한 분석을 수행할 수 있다.

제3장에서는 가명정보 결합을 활용하여 통합데이터를 구축하고 연구한 국내외 사례에 대해 살펴보았다(〈표 6-1〉 참조). 국내의 경우 가명정보 결합을 통해 구축된 통합데이터를 활용한 연구가 많지 않은 편으로 나타났다. 통계청에서 ‘조사데이터’와 ‘행정데이터’의 결합을 적극적으로 실시하고 있는 것으로 보였다. 최근에는 사회보장위원회에서도 사회보장행정데이터를 구축하고 있는 것으로 나타나 앞으로 연구 사례가 점차 많아질 것으로 생각한다. 영국 사례를 보면 2018년 공공기관의 기능 수행을 위해 데이터 결합이 필요하다는 법적 근거가 마련되어 모든 응답자를 연계할 수 있게 되었다. 법의 적용과 연계 방법론의 개선으로 가족자원조사와 행정데이터의 결합률은 95% 이상으로 나타났다. 포르투갈 사례를 보면 정확 매칭을 통해 성별, 연령, 거주지 및 사망일을 기반으로 두 개의 정보시스템 데이터를 결합하였다. 코로나19 감염 및 관련 사망 정보 모니터링 중에 기록된 질환 상태에 관한 두 개의 데이터에서 이용 가능한 정보의 잠재적인 상보성을 비교·식별하였다. 그 결과 기록된 질환 상태는 일치하는 경우도 있었지만, 서로 다른 정보를 가지는 경우도 확인할 수 있었다. 일본에서는 건강보험 청구데이터와 건강 검진 데이터를 결합한 데이터를 구축하고 있다. 환자의 이름, 생년월일, 보험 의료 기관의 소재지와 명칭 등 특정 개인이 식별되는 정보는 삭제하고, 이를 함수로 처리한 두 개의 해시 ID를 사용하여 데이터를 매칭하였다. 이렇게 결합된 데이터는 정보 통계 추출, 시계열 자료 구축 및 평가, 건강검진 정보와 진료 정보 연계 등에 활용되고 있다.

국내에서도 가명정보 결합에 대한 수요가 점차 높아지는 가운데, 해외 사례와 같이 다양한 분야의 행정데이터의 활용 가능성, 응답값의 정확성과 측정오차 등의 검토를 통해 조사데이터의 품질 향상 등의 목적으로 가명정보 결합이 활성화될 수 있다고 생각한다.

〈표 6-1〉 가명정보 결합을 사용하여 통합데이터를 구축한 사례 정리

연구목록	가명정보 결합 목적	결합한 데이터 목록	매칭키	연구 결과	고려 및 주의사항
통계청 (2020)	행정자료의 활용으로, 조사자료의 품질 제고	가계금융복지조사 + 행정자료 (근로소득, 기초연금, 국민연금, 소득세, 건강보험료 등)	개인의 성명, 성별, 생년월일, 주소 등	[행정자료 활용 결과] • 보완 전에 비해 전체적인 가구소득 상향조정 • 보완으로 지니계수 및 상대적 빈곤 율 증가 • 특히 가구주 60세 이상인 가구의 소득 보완에 크게 기여	• 품질 높은 행정자료의 필요 • 명확한 자료처리 기준 필요
임성희, 김진옥 (2021)	경제통계 분석	통계기업등록부(통계청 조사자료 + 국세청 과세자료)+공동사업자명세자료 +프랜차이즈 가맹점 모집단 자료	사업자 등록번호, 개인대체 식별번호	[개인사업체 생존시간 추정] • '공동사업자 있음'과 '프랜차이즈 가맹점 해당'은 대부분의 업종에서 생존위험에 감소 효과를 보임 • 연간 종업원 수와 연간 매출액 1단 위 증가는 생존위험에 감소 효과를 보임	• 연계 가능한 식별정보 유무 • 결합 시 응용 가능한 일정 수 준의 수 • 분석에 유의한 변수를 포함 하는지 등
정찬미 (2021)	사회보장 정책 연구	주민등록전산정보+사회보장 수급 관련 데이터	주민등록전산 정보상의 개인정보	[1인 가구의 사회보장 수급현황 분석] • 청년은 소득수준 및 사회보험 가입 률은 높으나 자산 수준이 낮음 • 노인은 근로하는 노인은 많으나 양 질의 근로를 하지 못하는 것으로 나타남	• 개인 수급 현황을 단일 연도 기준으로 수집 시, 소득·자산 을 파악하지 못하는 한계가 있음

연구목록	가명정보 결합 목적	결합한 데이터 목록	매칭키	연구 결과	고려 및 주의사항
이현주 (2022)	사회보장 제도의 이해 및 제도의 효과성 검증	통합 행정데이터 (행정데이터+행정데이터)	생년월일, 성별, 이름을 가명처리 하여 결합키 생성	[사회보장 행정데이터의 효율적인 연 계·활용 방안] • 요청 정보의 범위에 대한 합리적 판단 기제 마련 • 데이터의 안정적 확보를 위한 협치 강화 • 여러 정보를 연계하기 위한 연계기 의 정확성 제고 • 통합데이터를 활용한 연구 주제 개 발과 분석 결과를 반영할 수 있는 기제 강화	• 통합데이터 구축을 위해 협 치 강화 필요 • 주민번호 활용한 결합키 보안
Pattaro et al. (2020)	연구 목적 (사회적 불평등 식별)	FRS(가족자원조사)+PSE-UK (빈곤과 사회배제조사) +IAS(무직수당)+IS(소득지원) +WPLS (일자리 및 복지 장기연구)	국민보험 번호 (National Insurance Number)	• 장기적인 행정데이터 측정값은 정 책 및 빈곤 연구뿐만 아니라 다양 한 응용 분야에서 잠재적으로 폭넓 게 활용됨을 확인	• 설문조사에서 누락되는 실제 여부에 대한 문제
Jenkins and Rios- Avila (2021)	연구 목적 (행정데이터 에서의 측정오차)	FRS-P14 (FRS(가족자원조사) +P14(연도별과세신고자료))	성명, 우편번호, 성별, 생년월일	• 전반적으로 FRS 데이터 수집이 P14 데이터 수집보다 신뢰성 있음 • 오차의 조합으로 인해 두 데이터 모 두 관측된 수집은 참여자의 불평등 을 상당히 과대 평가함 • 측정오차에서 발생하는 더 큰 잡음 (noise)을 상쇄하기에는 평균 회귀 가 충분하지 않음	• (조사의 오차 분산을 작게 하 는 방안) 설문조사 품질을 위 한 방법으로 설문 면접관에 게 급여명세서 제시

연구목록	가명정보 결합 목적	결합한 데이터 목록	매칭키	연구 결과	고려 및 주의사항
Ferreira-Santos et al. (2022)	연구 목적 (코로나19 관련 이환율 분석)	SINAVE (국립 역학 감시 정보 시스템) + SICO (국립 전자 사망 증명서 정보 시스템)	성별, 연령, 거주지, 사망일	• 일치 사례: 2,285건 남성: 53.9%, 평균 연령: 84세 • 질병률의 존재 범위: 6.3~62.5% • 10개 질환 비교 가능하며 동시에 보고된 질환도 있었음 • 이환율 유사성을 평가 결과: 최소 한의 일치 또는 미일치 • 결론: 데이터가 독립적으로 수집됨. 코로나19 관련하여 서로 다른 정보를 찾을 수 있었음	• 이환 정보 일부 불일치 사례 (ICD-10 개방형 문항으로 보고한 후 ICD-10를 사용하여 코딩) • 불일치 사례가 해당 주제 사망한 사례의 5% 이상일 경우 제외
NDB (진료 (조제) 보수 명세서 정보· 특정 건강진단 등 정보 데이터 베이스)	행정DB 구축 (의료비 적정화 제도의 작성, 실시 및 평가를 위한 조사나 분석 등에 이용)	진료(조제) 보수 명세서 정보+특정 건강 진단 등 정보 데이터 베이스	보험자 번호, 기호 번호, 성명, 생년월일, 성별	• 전국 행정 DB 구축 • 국가, 보험자, 연구 기관 등 정책 연구자에게 제3자 제공 • 정책 통계 추출 • 시계열 자료 구축 및 평가 • 건강검진 정보와 진료 정보 연계	• 결혼으로 성이 바뀌거나 이직, 전직 등으로 인해 적용되는 건강보험제도 또는 보험자가 변경되는 경우 헤시값 매칭 불가 • 헤시값을 2개 생성하여 매칭

자료: '3장 가명정보 결합을 활용한 통합 데이터 구축 및 연구 사례'를 기반으로 필자가 재구성하여 작성함.

제4장에서는 정보공개법, 통계법 등 7개의 법령⁴⁵⁾과 가명정보 관련 3개의 가이드라인⁴⁶⁾을 살펴보고, 한국보건사회연구원의 데이터 제공 절차에 대한 현행 규정⁴⁷⁾과 절차를 분석하였다. 「정보공개지침」에는 정보에 대한 가명처리 등 가명정보 제도에 대한 내용이 없는 것으로 나타났다. 이러한 종합적인 분석 결과를 토대로, 한국보건사회연구원의 익명·가명 데이터 제공 절차를 수정·보완하였고, 관련 지침을 새롭게 수립하여 제안하였다. 이에 따라 일원화된 데이터 제공 체계를 갖추게 되었고, 한국보건사회연구원의 다양한 데이터를 더 빠르게 제공할 수 있을 뿐만 아니라, 안전하게 제공할 수 있는 기반을 마련하였다고 생각한다. 특히, 이 연구에서 제안한 ‘데이터제공심의위원회’라는 일원화된 의사결정기구를 통해 다양한 데이터의 제공에 대해 통합적인 관리를 할 수 있으며, 데이터 제공으로 인한 개인 식별 등의 다양한 위험에 대해서도 효과적으로 고민하고 대응할 수 있을 것이다. 또한, 데이터 제공 체계에서 각 역할과 책임에 따라 담당 부서를 배치하는 방안도 마련하였다. 한국보건사회연구원의 현재 상황과 여건을 고려했을 때, 우선 ‘임시 전담 부서’로 조직 및 운영하도록 제안하였다. 향후 데이터 제공 건수가 많아지면 ‘내부 전담 총괄부서’로 조직 및 운영하는 것이 효율적이라고 생각한다. 이러한 개선은 한 번에 모두 이뤄질 수 없으며 우선순위에 따라 즉시 적용해야 할 내용과 점진적으로 적용해야 할 부분으로 구분하여 운영하는 것이 필요하다(〈표 6-2〉 참조).

45) 7개의 법령은 공공기관의 정보공개에 관한 법률(이하 ‘정보공개법’), 통계법, 공공데이터의 제공 및 이용 활성화에 관한 법률(이하 ‘공공데이터법’), 데이터기반 행정 활성화에 관한 법률(이하 ‘데이터기반행정법’), 개인정보 보호법, 생명윤리 및 안전에 관한 법률(이하 ‘생명윤리법’), 신용정보의 이용 및 보호에 관한 법률(이하 ‘신용정보법’)이 해당함.

46) 가명정보에 대한 가이드라인은 가명정보 처리 가이드라인, 보건의료데이터 활용 가이드라인, 공공분야 가명정보 제공 실무 안내서가 해당함.

47) 정보공개지침, 조사자료관리지침

〈표 6-2〉 우선순위에 따른 적용 내용

우선순위	내용
1	통합된 데이터 제공 지침 수립, 승인통계 제공 절차의 보완, 데이터 제공 관련 양식의 보완
2	데이터 제공과 관련한 임시 전담 조직을 구성하여 운영
3	데이터 제공 전담 부서를 구성하여 운영

자료: 필자 작성

제5장에서는 가명정보 결합 절차에 따라 실제로 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터를 통합하여 구축하였다. 그리고 구축된 통합데이터를 활용하여 두 개의 분석 주제에 대해 분석하였다.

먼저, 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터의 가명정보를 결합하기 위해 사전준비, 결합신청, 가명처리, 결합수행, 반출 및 관리 절차에 따라 결합을 실시하였다(〈표 6-3〉 참조). 가명정보 결합 절차⁴⁸⁾는 올해 3월부터 시작하여 10월 중순에 완료하였고, 그 이후부터 데이터 분석을 할 수 있었다. 가명정보를 결합한 통합데이터를 제공 받는 것에만 8개월 남짓 소요되었다. 1년 단위의 과제 수행에서는 매우 빠듯한 기간이라고 볼 수 있다. 보통 결합전문기관에 따라서는 1년 이상 소요되는 것이 다반사이기 때문에, 현재로서는 연구 기간이 충분하게 확보되었을 때 안정적으로 진행할 수 있다고 생각한다. 한편, 5계 단계 중에서 사전준비 단계가 5개월로 가장 오랜 시간이 걸렸다. 사전준비 단계에서는 한국보건사회연구원의 가명정보에 대한 내부 관리계획을 수정·보완하였고, 한국의료패널조사 데이터와 사회보장행정시스템의 행정데이터를 사용할 수 있도록 승인받는 과정 등을 거쳤다. 다음으로 가명처리 단계에 시간을 많이 할애한 것을 알 수 있다. 한국의료패널조사 데이터에

48) 가명정보 결합 절차를 본격적으로 진행하기 전인 1~2월에는 관련 제도, 선행연구 등을 검토하는 과정을 가졌음.

대한 적정성 검토 심사를 받기 위해 가명처리를 하였고, 적정성 검토 심사 결과에 따라 추가 가명처리를 하였다. 그리고 가명정보를 결합한 통합 데이터에 대한 적정성 검토 결과에 따라 다시 추가 가명처리를 실시하였다. 이렇듯 가명처리는 여러 번에 걸쳐 진행되었다.

〈표 6-3〉 한국의료패널조사와 사회보장정보시스템의 가명정보 결합 절차

단계	절차	일정 (월)										주요 내용
		3	4	5	6	7	8	9	10	11		
1	사전준비										- 한국의료패널조사의 가명정보 이용·제공 신청서를 제출하였고, 적합성 검토위원회에서 가명정보 제공 승인 통보받음 - 사회보장행정시스템의 행정데이터의 가명정보 활용을 위해 해당 정부 부처의 데이터 활용 승인 공문을 득함	
2	결합신청										- 가명정보결합종합지원시스템에서 가명정보 결합을 신청함 - 두 개 데이터 보유기관이 동일한 기준의 결합키를 생성하였고, 한국인터넷진흥원은 결합키만 가지고 결합을 실시함 - 결합률은 71%였는데, 이는 한국의료패널조사의 응답자 중에서 사회보장급여 수급자 4,673명 기준을 의미하며, 결합된 대상자는 3,320명임	
3	가명처리										- 한국의료패널조사 데이터에 대한 가명처리를 실시함 - 데이터심의위원회에서 적정성 검토를 심사함 - 적정성 검토 결과에 따라 한국의료패널조사 데이터에 대한 추가 가명처리를 실시함 - 한국의료패널조사 데이터를 한국사회보장정보원으로 전송함 - 사회보장행정시스템의 행정데이터도 동일한 과정을 거쳤음	
4	결합수행										- 한국사회보장정보원에서 가명정보 결합을 완료하여 통합데이터를 구축함 - 앞에서 진행했던 적정성 검토 결과에 대한 추가 가명처리를 실시함	
5	반출 및 관리										- 통합데이터 반출을 위해 반출심사위원회를 개최하여 심사하였고, 반출 승인을 득함 - 한국사회보장정보원 가명정보결합분석센터에서 연구 목적에 따른 데이터 분석을 실시함	

자료: '5장 가명정보 결합을 활용한 데이터 통합 실증 연구'를 기반으로 필자가 재구성하여 작성함.

다음은 가명정보 결합을 통해 구축된 통합데이터를 사용하여 실증 분석한 분석 결과를 요약하였다. 분석 주제는 (1) 조사데이터와 행정데이터의 정합성 검토와 (2) 사회보장 급여 수급자의 의료비 지출 특성 분석이었다. 한국의료패널조사는 의료비 지출을 포함한 각종 보건의료 정보를 포함하고 있는 조사데이터이다. 그런데 사회보장 급여는 세부적인 정보가 부족하거나, 응답자의 기억에 의존하여 응답을 받기 때문에 정확도가 낮아질 수 있다. 그러므로 가명정보 결합을 통해 구축된 통합데이터를 사용하면 조사데이터가 가지는 부족한 부분을 보완할 수 있다는 점이 가장 큰 강점이라고 볼 수 있다.

첫 번째 분석 결과를 살펴보면, 조사데이터와 행정데이터에서의 사회보장 급여를 수급 받는다고 응답한 비율은 유사한 편으로 나타났다. 그런데 조사데이터와 행정데이터 간 수급 여부가 일치하지 않는 경우를 보면, 기초연금은 조사데이터에서 수급 받는다고 했으나 행정데이터에서는 수급 받지 않는다고 한 가구의 비율이 높았던 반면에, 맞춤형 급여는 반대의 경우가 높았다. 한편, 맞춤형 급여 중에서 조사데이터에서는 수급 받지 않는다고 했으나 행정데이터에서 수급 받는다고 한 비율이 가장 높은 급여는 주거급여로 나타났다.

조사데이터와 행정데이터의 값이 일치하지 않는 요인은 다음의 것들을 생각해 볼 수 있다. (1) 조사데이터의 응답 시점과 행정데이터의 추출 시점 간의 차이 때문으로 볼 수 있다. 맞춤형 급여는 조사데이터와 행정데이터의 시점이 2019년 12월 31일로 동일하였다. 그러나 기초연금은 조사데이터의 경우 응답 시점이 2019년 1월 1일~12월 31일까지인 반면에, 행정데이터는 2019년 12월 31일이 추출 시점으로 동일하지 않다고 볼 수 있다. (2) 조사데이터의 경우 사회보장 급여 수급 여부를 가구 조사표에서 조사하고 있는데, 이는 대표가구원이 모두 응답하는 구조이므로

정확히 파악하지 못하는 경우가 있을 수 있다. (3) 행정데이터는 행정 처리 과정에서 생성되는 데이터이므로, 조사데이터와 같이 정제된 데이터가 아니다. 이러한 특성이 기인했을 가능성도 있다고 생각한다. (4) 두 데이터를 결합할 때, 결합키로 이름, 생년월일, 성별을 사용하였다. 이때 동일인이라도 주민등록 생년월일과 실제 생년월일(음력, 양력)의 차이, 또는 이름과 성별의 차이가 생길 수 있다.

조사데이터와 행정데이터의 사회보장 급여 수급 여부가 일치하지 않는 집단의 비율은 낮은 편이었으나, 이러한 차이를 가구 특성과 가구주의 인구사회학적 특성 분포를 통해 살펴보았다. 조사데이터에서 수급 받지 않는다고 했으나 행정데이터에서는 수급 받는다고 한 집단은 서울 및 경기도, 광역시에 거주하고 아파트에 거주하며, 연간 가구 총소득이 높다고 볼 수 있다. 가구주의 특성은 남성이며, 연령대가 젊은 편이고, 혼인 중이고, 최종 학력도 고졸 이상으로 높은 편이고, 경제활동을 하고 있는 것으로 나타났다. 이러한 결과는 한국의료패널조사의 표본이탈 경향 분석에서 나타난 패널가구 및 패널가구원의 특성과 유사하게 나타났다(이혜정 외, 2023). 즉, 조사에 참여하더라도 소극적으로 응답하는 경향이 있다고 볼 수 있다.

가명정보 결합으로 구축된 통합데이터의 정합성 검토를 통해 행정데이터의 활용 가능성을 확인하였다. 행정데이터와의 비교를 통해 조사데이터의 응답의 정확성을 파악할 수 있으며, 조사 수행 및 실사 관리에서 정확한 응답을 수집할 수 있도록 유용한 정보로 활용될 수 있다. 또한 향후 행정데이터의 정보를 연계하여 응답자의 부담을 낮추는 데 기여할 수 있다고 생각한다.

두 번째 분석 주제에 대한 결과를 살펴보면, 두 가지로 요약할 수 있다. 첫째, 맞춤형 급여의 수급 규모가 작은 가구일수록 의료비 지출 수준이

낮은 것으로 나타났다. 한편 본인부담금 수준이 높은 의료급여 2종 급여의 대상은 상대적으로 의료비 지출에 대한 부담 수준이 높은 것으로 나타났다. 이를 통하여 맞춤형 급여 수급액이 높은 저소득층과 근로무능력층의 의료비 지출이 의료서비스의 필요에 따라 충분히 이루어지고 있는지를 살펴볼 필요가 있음을 알 수 있다. 또한 각 수급 자격에 따른 본인부담금 수준과 의료비에 대한 부담 등을 다각적으로 살펴볼 필요가 있음을 알 수 있다.

둘째, 기초연금 수급 가구는 수급 자격에 따라 의료비 지출 특성이 상이하게 나타났다. 기초연금 ‘단독’ 수급 가구는 기초연금 ‘부부2인’ 가구에 비해 의료비 지출 규모가 작았으나 의료비에 대한 부담이 큰 것으로 나타났다. 또한 기초연금 ‘단독’ 수급 가구는 급여 수준이 낮을수록 의료비 지출 규모가 큰 것으로 나타났으며, 기초연금 ‘부부2인’ 가구는 급여 수준이 높을수록 의료비 지출 규모가 큰 경향을 보였다. 이를 통하여 급여 자격별로 의료비 지출 특성을 구분하여 분석하여야 함을 알 수 있다.

특히 단독 수급 가구의 경우, 기준연금액 100%를 수급하는 가구의 의료비 지출 수준이 상대적으로 낮다는 점에 주목할 필요가 있다. 이 집단은 연금 수급을 통한 소득이 높지만, 소득·재산의 수준이 상대적으로 낮을 가능성이 높다. 향후 이들의 건강상태나 의료비 지출에 대한 부담 등이 실제 의료비 지출에 대한 위축으로 작용하게 되는지 여부를 살펴볼 필요가 있다. 한편, 부부2인 가구의 경우 기준연금액 100%를 수급하는 가구의 의료비 지출이 상대적으로 높다는 점을 확인하였다. 이들은 부부가 모두 65세 이상이면서 기초연금 외의 소득이 없을 가능성이 높기 때문에 과도한 의료비 지출로 인한 가계부담 증대의 위험을 가지고 있다. 향후 부부2인 가구의 기초연금 소득의 대부분을 의료비로 지출하는 가구를 확인하여 이에 대한 보충적 지원 정책을 뒷받침할 수 있는 자료를 생성할

필요가 있다.

가명정보 결합을 사용하여 데이터를 통합할 때의 고려할 사항은 다음과 같다. 첫 번째는 결합하고자 하는 데이터의 정보 특성을 파악하여 제공할 수 있는 변수 수준에 대한 정확한 파악이 필요하다는 점이다. 조사데이터의 경우 코드북이 있지만, 행정데이터의 경우 코드북이 없기 때문에 변수에 대한 자세한 정보는 알 수가 없다. 그렇기 때문에 데이터 보유기관의 협조를 통해 결합하고자 하는 정보의 특성을 파악해야 한다. 두 번째는 가명정보를 결합하는 데이터의 연구 모집단의 대상, 분석 시점 등에 대한 검토를 해야 한다는 점이다. 이 연구에서는 연구 모집단의 대상을 맞춤형 급여와 기초연금 수급 대상으로 선정하였다. 기준 연도는 2019년으로 하였다. 한국의료패널은 설문 문항별로 기준 시점이 상이하다. 즉 2019년 12월 31일과 2019년 1월 1일~12월 31일이므로 확인이 필요하다. 사회보장정보시스템의 행정데이터는 추출 시점이 2019년 12월 31일이 기준 시점이다. 세 번째는 연구자와 데이터 보유기관 간 긴밀한 협의가 중요하다는 점이다. 그렇지 않으면 가명정보 결합률이 높지 않을 수 있거나, 분석 변수의 결측률이 높을 수 있거나, 가명처리 수준이 의도했던 것보다 강화될 수 있는 점 등으로 인해 연구 목적에 맞는 최적화된 통합데이터를 구축하는 것이 어려워질 수 있다. 네 번째는 행정데이터를 활용하고자 할 때는 데이터 보유기관의 승인뿐만 아니라 해당 정부 부처의 승인이 필요하다는 점이다. 사회보장정보시스템의 행정데이터를 활용하기 위해 한국사회보장정보원과 해당 정부 부처인 보건복지부, 국토교통부, 교육부의 승인을 공문을 통해 득한 것처럼 해당 부서의 승인을 모두 받아야 한다는 점을 고려해야 한다.

제2절 가명정보 결합 제도 활성화를 위한 방안

가명정보 결합을 통한 통합데이터 활용의 장점은 다음과 같다. 첫 번째는 조사데이터에 없는 추가적인 정보를 얻을 수 있으므로 연구의 확장성 측면에서 매우 강력한(powerful) 제도라고 볼 수 있다. 두 번째는 새로운 조사를 계획 및 수행하는 데 필요한 시간, 금전적 비용을 줄일 수 있다는 점이다. 세 번째는 다양한 정보를 수집하기 위하여 많은 설문 문항을 포함함으로써 발생할 수 있는 새로운 조사데이터에 대한 품질 저하의 위험이 크지 않은 편이라는 점이다. 그리고 조사데이터의 응답값을 행정데이터의 값으로 대체할 수 있다면, 조사데이터의 정확성을 기할 수 있어 품질 향상의 효과를 얻을 수 있다. 마지막으로, 조사에 참여하는 응답자 측면에서 보면 응답 부담의 가중을 줄일 수 있다는 점이다.

가명정보 결합을 통한 통합데이터의 한계점으로 다섯 가지를 들 수 있다. 첫 번째는 가명처리, 가명정보 결합 및 반출 등이 매우 복잡한 절차라는 점이다. 실제로 진행하였던 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터 결합에서는 두 개의 데이터를 활용하였다. 이때 각 데이터의 가명처리 수준에 맞춰 각 단계를 실시해야 하고, 단계별로 데이터에 대한 심의 과정을 거쳐야 하고, 결합전문기관에 가명처리한 데이터를 전송하고, 가명정보 결합 데이터를 반출하는 등 거쳐야 할 절차가 많고 복잡하였다. 두 번째는 가명정보 결합을 활용한 연구를 처음 수행했기에 익숙하지 않았던 것도 있지만, 대체적으로 가명정보를 결합하는 일련의 과정이 오래 걸린다는 점이다. 한국의료패널조사 데이터와 사회보장정보시스템의 행정데이터의 결합은 올해 3월에 시작하여 10월 중순에 완료하였고, 그 이후부터 데이터 분석을 할 수 있었다. 가명정보 결합 데이터를 제공받는 것에만 8개월 정도가 걸려서, 1년 단위의 과제 수행에서는 매우 빠듯한 기간이라고 볼 수 있다. 보통, 결합전문기관에 따라서는 1년

이상 소요되는 것이 다반사이기 때문에 현재로서는 연구 기간이 충분하게 확보되었을 때 안정적으로 진행할 수 있다고 생각한다. 세 번째는 가명정보 결합을 통해 구축된 통합데이터는 제한적으로 접근 가능하며, 일회성 사용이라는 점이다. 처음 연구 계획을 세울 때 등록한 연구진만 데이터를 활용할 수 있으며, 연구진을 추가할 수 없다. 또한, 연구 목적에 따라 연구가 완료되면 통합데이터는 폐기된다는 점이 아쉽다고 볼 수 있다. 네 번째는 연구자가 가명처리에 적합한 목적으로 가명정보 결합을 신청하더라도 데이터 보유기관이 가명정보를 제공해야 하는 의무가 없다는 점이다. 그렇기 때문에 연구 계획을 세웠어도 가명정보 결합 데이터를 활용한 연구를 진행하지 못할 수 있다. 마지막으로, 가명정보 결합을 통한 통합데이터 구축 시 가명정보에 대한 가명처리 수준이 높은 편이라고 볼 수 있다. 한국의료패널조사에서 의료비를 원 단위로 조사하여 제공하고 있으나, 가명정보 결합으로 인하여 식별 가능성이 증가할 수 있다는 데이터심의위원회의 결정에 따라, 의료비는 만 원 단위로 반올림 및 99.7% 이상은 상단 처리를 하였다. 의료비 이외에도 여러 변수들은 유사한 방법으로 가명처리하여 제공되었다. 이는 기존 조사데이터에서 제공하는 정보보다 축소 제공된 것이다.

2020년 8월 5일 데이터 3법의 개정·시행으로 가명정보 결합의 제도적 기반이 마련되었으나, 데이터 이용자와 데이터 보유기관 등에게 가명정보는 아직 생소하고 어려운 개념이라고 볼 수 있다. 한편, 가명정보를 활용하는 데 불편한 점을 해소하기 위하여 관련 제도들은 지속적으로 수정·보완 및 개선되고 있는 상황이다. 이에 가명정보 결합 제도가 활성화되기 위한 방안을 제안하고자 한다.⁴⁹⁾ 가명정보 이용자, 데이터 보유기관(결합 전문기관 등), 가명정보 결합에 사용되는 결합키, 가명정보 결합 절차의

49) 전문가 자문 의견 결과를 통합적으로 정리하였음.

측면으로 구분하여 살펴보고자 한다.

가명정보 이용자에 대한 지원 강화가 필요하다. 이용자가 데이터 보유 기관이 제공할 수 있는 데이터에 대한 정보에 쉽고 편리하게 접근할 수 있어야 한다. 이러한 정보를 한 곳에서 볼 수 있는 플랫폼이 있다면 유용할 것이라고 생각한다. 이용자에게 데이터 보유기관의 가명정보를 이용하는 절차를 자세하게 안내하는 것이 필요하다. 현재는 기관별 안내 수준이 상이할 뿐만 아니라 세부적인 절차 안내 미흡, 협의 담당자의 확인 불가 등과 같이 다양한 문제점이 상존하고 있다. 가명정보를 활용하거나 결합하기 위한 절차를 처음 접하는 데이터 이용자를 지원하는 제도가 유지 및 확대되어야 한다. 현재 개인정보위원회, 한국인터넷진흥원, 한국지능정보사회진흥원 등을 중심으로 선도 사례 발굴, 경진대회, 가명정보 활용 지원 서비스⁵⁰⁾ 등이 이루어지고 있는 상황이나, 더 지속적으로 널리 확대될 필요가 있다고 생각한다. 특히 가명정보 이용자가 가명정보 활용 지원 서비스를 받을 수 있도록 적극적인 안내 및 홍보를 하는 것도 중요하다고 생각한다.

데이터 보유기관에 대한 지원 강화가 필요하다. 데이터 보유기관이 가명정보 데이터를 제공해야 할 의무는 없다. 아직 대부분의 데이터 보유기관은 가명정보 제공에 대한 인식 및 제공 절차에 대한 이해가 부족할 수 있다. 개인정보보호위원회가 발간한 「가명정보 처리 가이드라인」에 세부 절차가 일부 서술되어 있지만, 그 내용은 관련 지식이 없는 일반 업무 담당자가 이해하기 어렵다고 볼 수 있다. 한편으로 가명처리 등 관련 지식을 보유한 전문가가 없을 가능성이 높다고 볼 수 있다. 결합전문기관 등과 같이 가명처리 관련 조직을 운영하지 않는 기관이 가명처리 전문가를

50) 가명정보 활용 지원 서비스에는 가명·익명 처리 관련 법·제도 등 교육, 데이터 활용에 대한 맞춤형 컨설팅, 가명정보 결합 지원 등 안전한 가명정보 활용을 위한 기술지원 서비스 등이 해당함.

조직 내에 별도로 배치하는 것은 현실적으로 어렵다고 볼 수 있다. 이에 가명처리 단계의 핵심이자 최종 단계라 할 수 있는 적정성 검토 절차를 가명처리 관련 전문성이 확보되지 않은 기관에서 수행하는 것은 현실적으로 어렵다고 볼 수 있다. 또한, 이러한 상황을 고려했을 때 데이터 보유 기관에 다음과 같은 지원을 제공하는 것을 생각해 볼 수 있다. 데이터 보유기관을 대상으로 한 컨설팅 서비스 제공, 가명정보 관련 지식이 없는 일반 업무 담당자에 대한 교육 실시, 가명정보 전문가 인력 확보, 관련 업무를 수행할 수 있도록 예산 지원, 가명정보 제공에 대한 인센티브제도 마련 등을 고려해 볼 수 있다.

한편, 가명정보 결합률을 높일 수 있도록 데이터 보유기관과 결합키 관리기관의 적극적인 역할이 필요하다. 결합키 생성 시 개인 식별 정보를 많이 추가한다고 결합률과 결합의 정확도가 높아지는 것은 아니다. 그렇기 때문에 최적의 결합률을 달성할 수 있도록 결합키를 생성해야 할 것이다. 보통 결합키로 이름, 생년월일, 성별을 조합하여 사용하고 있는데, 다양한 결합키 조합에 대한 연구가 필요하다고 생각한다. 또한, 데이터 보유기관은 가명정보 결합 사례별로 결합키 생성 시 투입되는 개인정보의 속성을 확인해야 한다. 결합키 관리기관은 결합률 향상을 위해 적극적으로 대응하고 사례 연구 결과를 공유하는 것을 고려해 볼 수 있다.

가명정보 결합 절차의 측면에서 살펴보면 다음과 같다. 가명정보 결합 절차 간소화, 가명정보 결합 신청 양식의 일원화, 가명처리 수준 완화 등이 필요하다. 현재 가명정보의 분야에 따라 가명정보 결합 절차가 다르고 복잡하다고 볼 수 있다. 그리고 가명처리 적정성 검토가 완료된 가명정보에 대해 결합을 실시하나, 결합이 완료된 가명정보의 식별 가능성으로 인해 마지막 단계로 반출심사를 실시한다. 반출심사에서는 적정성 검토 시 검토했던 사안을 재심사하는 등의 불필요한 과정이 반복된다고 볼 수 있

다. 또한, 특정 가명정보 데이터를 활용하고자 하는 경우 특정 결합전문기관을 선택해야 한다. 그런데 선택할 수 있는 결합전문기관이 한정되어 있어 특정 결합 전문기관으로 쏠림 현상이 발생하게 된다. 이로 인해 데이터를 제공받는 시간이 더 오래 소요되어, 원활한 연구 수행이 어려워질 수 있다. 가명정보 결합 신청을 작성하는 양식도 결합전문기관에 따라 다르게 되어 있어 효율적이라고 볼 수 없다. 한편, 결합전문기관이 제공하는 분석 공간에서 분석을 하는 경우에는 식별 가능성이 현저히 낮다고 볼 수 있다. 연구 목적을 달성하기 위해 항목별 가명처리 수준을 조정하는 게 합리적이라고 볼 수 있으나, 관련 가이드라인의 부재로 과도한 가명처리 수준을 요구하고 있다.

가명정보 결합으로 구축된 통합데이터 사용 및 활용에 대한 접근성이 높아져야 할 것이다. 보통 가명정보 결합을 신청한 결합전문기관의 분석 센터에 방문하여 데이터를 사용해야 하는데, 지리적 문제로 인하여 사용 및 활용에 번거로움이 있을 수 있다. 이러한 문제들은 ‘개인정보 안심구역’을 통해 많은 부분이 해결될 수 있다고 생각한다. 개인정보 안심구역은 제로 트러스트 보안모델 기반 안전조치, 사전·사후적 데이터 처리과정 통제 등 환경적 안전성을 갖추면, 기존에 사실상 제한했던 다양한 데이터 처리가 가능하도록 지원하는 제도다(전자신문, 2023.10.31.). 개인정보 안심구역 내에서는 데이터 분석 시 환경의 안정성 강화 수준에 비례하여 가명처리 수준을 적정 수준으로 완화하거나, 연계정보 일부 등 다양한 결합키 활용을 허용할 수 있다. 또한, 시계열 분석 등 지속적·반복적 연구를 위한 가명정보의 장기간 보관과 제3자 재사용 등의 활용성이 강화될 수 있을 것이다. 현재는 개인정보 안심구역을 시범으로 운영할 계획이나, 가명정보를 더 유연하게 활용할 수 있도록 개인정보 안심구역 제도가 정착되고 활성화될 필요가 있다.

제3절 향후 과제

데이터를 개방하는 추세에 따라 데이터가 다량으로 개방되고 있고, 데이터를 활용할 수 있는 기회가 점점 많아지고 있다. 하지만 여전히 개별적으로 분산된 채 데이터가 제공되고 있다. 여러 개의 데이터를 서로 연계하고 통합한 데이터를 활용할 수 있는 방안을 지속적으로 모색해야 한다. 1차 연도 연구에서는 조사데이터와 조사데이터 간 데이터를 통계적 매칭 방법을 사용하여 통합해 보았다. 2차 연도 연구에서는 조사데이터와 행정데이터 간 데이터를 가명정보 결합 방법을 사용하여 통합해 보았다. 2개 연도 연구 결과를 종합하여, 여러 개의 데이터를 연계하고 통합할 수 있다는 점과 통합데이터의 분석 결과도 의미가 있다는 점을 확인하였다.

의미 있고 활용도 높은 통합데이터를 구축하기 위해서는 데이터 통합에 관한 연구 주제가 잘 설정되어야 한다. 연구 주제에 대한 수요를 파악하기 위해 1차 연도에 이어 2차 연도에도 보건복지 등 다양한 분야의 학계 및 실무 전문가 50명을 대상으로 설문조사를 실시하였다. 조사기간은 2023년 8월 1일부터 8월 20일까지였으며, 조사 방법은 이메일을 통한 서면 설문조사였다. 조사 내용은 데이터 통합 관련 연구 주제에 활용할 수 있는 기준데이터와 제공데이터의 목록, 분석 변수, 기대효과에 대한 것이었다. 설문조사 결과는 <부표 1>에 정리하였는데, 다출처 데이터 연계, 통합에 대한 수요가 많다는 것을 다시 한번 확인할 수 있었다. 그럼에도 불구하고 통합데이터 구축을 둘러싼 여러 가지 어려움으로 인해 활성화되지 못하고 있는 것이 현실이다.

다출처 데이터를 연계, 통합하여 연구할 수 있는 주제를 지속적으로 발굴하고, 양질의 데이터를 사용하여 통합데이터를 구축하고, 통합데이터

를 분석할 수 있어야 한다. 이러한 일련의 과정이 체계적으로 잘 이루어진다면, 다출처 데이터를 연계, 통합하여 활용한 기초 연구로 근거 기반 정책을 수립하는 토대를 마련할 수 있으리라고 생각한다.



가명정보의 결합 및 반출 등에 관한 고시, 개인정보보호위원회고시 제2022-7호 (2022).

개인정보 보호법 시행령, 대통령령 제33723호 (2023).

개인정보 보호법, 법률 제19234호 (2023).

개인정보보호위원회(2022). 가명정보 처리 가이드라인.

개인정보보호위원회. (2023. 5. 4.). 가명정보 결합·활용 현장의 목소리 청취. 개인정보보호위원회 보도자료. <https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=8831>. 2023. 11. 30. 인출함.

개인정보보호위원회. (2023. 7. 20.). 가명정보 활용 확산으로 데이터 신경제 선도 추진. 개인정보보호위원회 보도자료. <https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9044>. 2023. 11. 30. 인출함.

개인정보보호위원회·보건복지부(2021). [2021년] 보건의료 데이터 활용 가이드라인.

개인정보보호위원회·보건복지부(2022). [2022년] 보건의료 데이터 활용 가이드라인.

개인정보보호위원회·한국인터넷진흥원. (2023). 가명정보결합 종합지원시스템. <https://link.privacy.go.kr/>에서 2023.6.22. 인출함.

개인정보보호위원회·한국인터넷진흥원. 가명정보 활용 종합지원플랫폼. <https://data.privacy.go.kr/mtcsprtInfo/mtcsprtSvc.html>에서 2023.8.4. 인출함.

공공기관의 정보공개에 관한 법률, 법률 제19408호 (2023).

공공데이터 관리지침, 행정안전부고시 제2021-70호 (2021).

공공데이터의 제공 및 이용 활성화에 관한 법률 시행령, 대통령령 제33842호 (2023).

공공데이터의 제공 및 이용 활성화에 관한 법률, 법률 제19408호 (2023).

과학기술정보통신부. (2022.2.18.). 개인정보위, 2기 가명정보 결합 선도 사례
중점 추진. 과학기술정보통신부 보도자료. <https://www.msit.go.kr/bbs/list.do>에서 2023.11.30. 인출함.

교육부·개인정보보호위원회(2020). 교육분야 가명·익명정보 처리 가이드라인
구인회·강성호·박형준·손병돈·우해봉·이원진·최옥금·함선유·한경훈·이지완
(2021). 행정데이터를 활용한 다층보후소득보장체계 심층분석, 보건복지
부·서울대학교 산학협력단.

국세청. 국세통계센터. <https://datalab.nts.go.kr/websquare/websquare.html?w2xPath=/ui/main/index.xml>에서 2022.8.4. 인출함.

금융위원회. (2023. 7. 19.). 빅데이터 생태계의 토대인 데이터 결합의 저변을 확대
해나가겠습니다.- 8개 기관 데이터 전문기관 추가지정. 금융위원회 보도자료.
<https://www.fsc.go.kr/no010101/80431?srchCtgry=&curPage=&srchKey=sj&srchText=%EB%B9%85%EB%8D%B0%EC%9D%B4%ED%84%B0+%EC%83%9D%ED%83%9C%EA%B3%84%EC%9D%98+%ED%86%A0%EB%8C%80%EC%9D%B8+%EB%8D%B0%EC%9D%B4%ED%84%B0+%EA%B2%B0%ED%95%A9%EC%9D%98+%EC%A0%80%EB%B3%80%EC%9D%84+%ED%99%95%EB%8C%80%ED%95%B4%EB%82%98%EA%B0%80%EA%B2%A0%EC%8A%B5%EB%8B%88%EB%8B%A4&srchBeginDt=&srchEndDt=>에서 2023. 11. 30. 인출함.

금융위원회·금융감독원. (2022). 금융분야 가명·익명처리 안내서.

데이터기반행정 활성화에 관한 법률, 법률 제19408호 (2023).

보건복지부 생명윤리정책과(2020.8.5.). (생명윤리정책과-2605) 개인정보 보호
법 개정에 따른 「생명윤리법 관련 기관 운영지침」 일부 개정 (2020.8.4.).
https://www.mohw.go.kr/board.es?mid=a10409020000&bid=0026&act=view&list_no=359711에서 2023.12.20. 인출함.

보건복지부. (2020). 2020년 기초연금 사업안내. 보건복지부.

보건의료기본법, 법률 제17966호 (2021).

- 사회보장기본법 시행령, 대통령령 제33629호 (2023).
- 사회보장기본법, 법률 제18215호 (2021).
- 생명윤리 및 안전에 관한 법률, 법률 제17783호 (2020).
- 신용정보의 이용 및 보호에 관한 법률, 법률 제19234호 (2023).
- 우리은행. 우리은행 대표 홈페이지. 가명정보 처리 현황. <https://spot.woori.bank.com/pot/Dream?withyou=CQSCT0133>에서 2023. 8. 4. 인출함.
- 이현주, 오욱찬, 이윤경, 이원진, 성재민, 이길제, 박형존, 이병재. (2020). 사회보장정책 효과성 분석을 위한 행정데이터 연계·활용 방안. 세종: 한국보건사회연구원, 사회보장위원회.
- 이현주. (2022). 사회보장제도 근거 강화와 사회보장 행정데이터 구축: 사회보장 행정데이터 연계·활용을 위한 과제. 보건복지포럼(2022.11), pp.8-23.
- 이혜정, 변종석, 황안나, 염아림. (2023). 표본조사 표본이탈 경향 분석과 추가표본 추출에 관한 연구. 세종: 한국보건사회연구원.
- 이혜정, 이기호, 안수인, 임종호, 이상혁, 조용찬. (2022). 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구. 세종: 한국보건사회연구원.
- 임성희, 김진옥. (2021). 개인 창업사업체 생존분석: 2010~2018 행정데이터를 중심으로. 통계연구, 제26권 제2호, pp.1-30.
- 장인성, 이지은, 정현상, 김민석, 장지연, 홍민기, ..., 황규성. (2022). 패널자료 품질개선 연구(XIII), 한국노동연구원.
- 전자신문. (2023.10.31.). ‘개인정보 안심구역’ 시범운영기관 공모… “데이터 처리 안정성·활용성↑” <https://www.etnews.com/20231031000083>에서 2023.11.30. 인출함.
- 정찬미. (2021). 1인 가구의 사회보장 수급현황 및 정책 방향: 행정데이터를 활용하여. 사회과학논총, 제24집, pp.211-228.
- 통계법 시행령, 대통령령 제33700호 (2023).
- 통계법, 법률 제17339호 (2020).
- 통계청. (2020). 가계금융복지조사에서의 조사자료와 행정자료의 통합방법 이

해. 통계청.

통합데이터. (2023). 한국의료패널조사 2019년 연간데이터와 사회보장정보시스템의 행정데이터의 가명정보 결합으로 구축된 통합데이터[데이터파일]. 한국사회보장정보원 가명정보결합분석센터에 방문하여 데이터 분석함.

포르투갈 보건국. (2012). Sistema de Informação dos Certificados de Óbito(국립 전자 사망 증명서 정보 시스템(SICO) 시행절차). 포르투갈. <https://docplayer.com.br/19840316-Sistema-de-informacao-do-s-certificados-de-obito.html>에서 2023.04.13. 인출함.

포르투갈 보건국. (2013). Guia prático para a utilização do SICO - Médicos/Instituições de Saúde(SICO 사용을 위한 실용 가이드 - 의사/의료 기관). 포르투갈. <http://www.arsalentejo.min-saude.pt/arsalentejo/novidades/Documents/Guia%20pr%C3%A1tico%20para%20a%20utiliza%C3%A7%C3%A3o%20do%20SICO%20v.%2011.10.pdf>에서 2023.04.13. 인출함.

포르투갈 보건국. (2016). RELATÓRIO DE MONITORIZAÇÃO DO SISTEMA DE INFORMAÇÃO NACIONAL DE VIGILÂNCIA EPIDEMIOLÓGICA (SINAVE)(국가역학감시정보시스템(SINAVE) 모니터링 보고서). 포르투갈. <https://diariodarepublica.pt/dr/detalhe/despacho/15385-a-2016-105574339>에서 2023.04.13. 인출함.

한국보건사회연구원. (2023). 생명윤리위원회 운영규정

한국보건사회연구원. (2023). 정보공개 지침.

한국보건사회연구원. (2023). 조사자료 관리 지침.

한국보건사회연구원. 한국복지패널 사이트(www.koweps.re.kr)

한국보건사회연구원. 한국의료패널 사이트(<https://www.khp.re.kr>)

한국보건사회연구원·국민건강보험공단. (2019). 2기 한국의료패널조사 2019년 연간데이터[데이터파일]. <https://www.khp.re.kr:444/>에서 2023. 5. 9. 인출함.

한국보건사회연구원·국민건강보험공단. (2022.5.). 2기 한국의료패널 2019년

- 연간데이터 유저가이드.
- 한국사회보장정보원. (2023). 결합신청서 및 첨부서류 양식.
- 한상필, 제호영, 김중훈, 김진정, 박상필, 김성현, ..., 김태영. (2023.5.). 가명정보 결합전문기관 이용 안내서. 서울: 한국사회보장정보원.
- 행정안전부. (2023. 7. 20.). 국민생활 밀접한 데이터, 묶음으로 개방 국민 편의 높이고 데이터 경제 활성화한다. 행정안전부 보도자료.
https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_0000000000008&nttId=101914에서 2023.11.30. 인출함.
- 행정안전부·개인정보보호위원회(2021). 공공분야 가명정보 제공 실무안내서
- Bailey, N., and Bramley, G. (2018). Introduction. In N. Bailey & G. Bramley (Eds.), *Poverty and social exclusion in the UK. The dimensions of disadvantage* (pp. 1-23). Bristol: Policy Press.
- Department for Work & Pensions. Family Resources Survey: background information and methodology(updated 21 July 2023), UK, retrieved from 2023.10.02. <https://www.gov.uk/government/statistics/family-resources-survey-financial-year-2021-to-2022/family-resources-survey-background-information-and-methodology#linking-frs-data-to-administrative-data>
- Dex, S., and McCulloch, A. (1998). The reliability of retrospective unemployment history data. *Work, Employment & Society*, 12(3), pp.497-509.
- Feteira-Santos, R., Camarinha, C., de Araujo Nobre, M., Elias, C., Bacelar-Nicolau, L., Costa, A. S., ... & Nogueira, P. J. (2022). Improving morbidity information in Portugal: Evidence from data linkage of COVID-19 cases surveillance and mortality systems. *International Journal of Medical Informatics*, 163, 104763. <https://doi.org/10.1016/j.ijmedinf.2022.104763>

- Gordon, D. (2017). PSE-UK 2012 Survey Producing an 'objective' poverty line in eight easy steps.
<https://www.poverty.ac.uk/editorial/producing-pse-poverty-line>
e retrieved from 2023.10.02.
- Jenkins, S. P., & Rios-Avila, F. (2023). Reconciling reports: modelling employment earnings and measurement errors using linked survey and administrative data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(1), 110-136.
- Kapteyn, A. and Ypma, J. Y. (2007). Measurement error and misclassification: a comparison of survey and administrative data. *Journal of Labor Economics*, 25 (3), 513-551.
- McKay, S (2012). Evaluating approaches to Family Resources Survey data linking. Department for Work and Pensions Working Paper 110. retrieved from 2018.11.23.
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/222871/WP110.pdf
- Pattaro, S., Bailey, N., and Dibben, C. (2020). Using Linked Longitudinal Administrative Data to Identify Social Disadvantage. *Social Indicators Research*, 147, pp. 865-895.
- 厚生労働省 老健局・保険局(후생노동성 노건국·보험국). (2018.9.6.). NDB、介護DB等の役割と解析基盤について(NDB, 개호 DB 등의 역할과 해석 기반에 대해서). 일본. 2023.9.4. 인출함.
<https://www.mhlw.go.jp/content/12401000/000350567.pdf>
- 厚生労働省保険局医療介護連携政策課保険データ企画室(후생노동성 보험국 의료 개호 제휴 정책과 보험 데이터 기획실). (2023.10). NDBの利用を検討している方へのマニュアル(NDB의 이용을 검토하고 있는 분을 위한 매뉴얼). 일본. 2023. 12. 20. 인출함. <https://www.mhlw.go.jp/content/12400000/001162218.pdf>



[부도 1] 적정성 검토위원회에 제출할 필요 서류⁵¹⁾

가명정보 이용·제공 신청서				
접수번호	ZZANG-20220522001		접수일	2022년 5월 22일
신청자	조직/부서명	A 유동회사 구매부		
	담당자 직위	과장	담당자 성명	이유동
	전화번호	02-1234-5678	이메일 주소	gamyseongzzang@utong.com
처리 목적	☒ 통계작성 ☐ 과학적 연구 ☐ 공익적 기록보존 코로나 이전과 코로나 이후의 상품군별 판매 추이에 대한 통계 작성을 위한 가명정보의 제공			
활용 형태	☐ 내부이용 ☒ 제3자 제공 ☐ 결합전문기관을 통한 결합			
이용 주제	☐ 동일 개인정보처리자 ☒ 제3자(제공받는 자) 제공			
처리 환경	☐ 내부 ☐ 외부			
처리 장소	☒ 폐쇄 환경 ☐ 제한 없음			
반복 제공 여부	☒ 1회 제공 ☐ 시계열 분석 등을 위한 반복 제공(회 예정)			
제공 방법	☐ 온라인 ☒ 오프라인			
제공 받는 자	(내부) ☒ 동일부서 ☐ 타부서 (외부) ☐ 제3자, 결합전문기관 제공			
신청 명세	가명처리 대상 명칭	자사의 고객 구매정보		
	데이터 내역	구매부에서 보관하고 있는 자사 고객 중 회원가입이 2019년 1월 이전인 고객의 구매정보 중 2019년 1월 ~ 2019년 12월, 2021년 1월 ~ 2021년 12월까지의 구매정보데이터 전체 222컬럼, 전체 대상인원 820만명 중 50% 무작위 샘플링한 410만명의 정보		
	가명정보 이용 기간	2022년 4월 1일 ~ 2022년 10월 31일(7개월)		
위와 같이 가명정보 이용·제공을 신청합니다.				
2022년 5월 22일				
신청인(부서장명) 이 유 동 (서명 또는 인)				
첨부서류	1. 개인정보 유형 분류표 2. 활용데이터 요구 수준표(선택) 3. 데이터 처리 위탁 계약서(위탁 계약의 경우) / 추후 제출 4. 데이터 제공 계약서(제3자 제공의 경우) / 추후 제출			

51) 자료: 개인정보보호위원회. (2022.4). 가명정보 처리 가이드라인.

가명정보에 대한 안전조치 이행 약속서

본 기관은 「개인정보 보호법」에서 규정하고 있는 가명정보에 대한 안전조치의무 등(제28조의4) 및 가명정보에 대한 안전성 확보 조치(시행령 제29조의5)를 성실히 이행하고 기타 관련 법령을 준수하였습니다.

아울러, 이를 이행·준수하지 아니하여 발생하는 관련 법적 책임을 부담할 것을 약속합니다.

2022년 5월 22일

신청기관

B통계회사 (직인)

통계작성 계획서

통계명	2022_COVID19_SalesTrend_410	
대표 참여진	소속	B 통계회사
	담당자명	이동재
통계작성 배경 및 목적	COVID19는 전세계적으로 팬데믹을 불러왔다. 이에 따라 각종 산업이 위축되고 개인에 대한 거리두기로 인해 다양한 산업에서 변화를 가지고 왔으며 특히 개인의 소비 생활에도 많은 영향을 미치게 되었다. 이에 따라 코로나 이전과 코로나 이후의 개인별 판매 데이터를 분석하여 COVID19로 인해 개인에게 미치는 영향에 대한 파악 등 다양한 방면에 사용하기 위한 통계를 작성하려 한다.	
통계작성 대상자 수	410만명(2019년 1월~2021년 12월까지 구매내역이 있는 고객 중 50%를 무작위 샘플링)	
통계작성 계획 및 방법	동일 집단의 COVID19발병 이전과 발병이후의 제품의 판매 추이를 통해 COVID19가 개인의 소비 패턴의 변화에 어떤 영향을 주었는지를 파악하기 위한 통계를 작성	
기대효과 및 활용방안	코로나와 같은 질병이 발생할 때 소비의 변화에 따라 지원금의 처리 방법, 지원 규모와 생산 산업군에 대한 영향 등을 파악하여 이를 보완하기 위한 정책 연구 등에 사용	
붙임. 상세 통계작성 계획서 등		

개인정보 유형 분류표

순번	항목명	개인정보유형	비고
1	고객ID	개인식별정보	고객의 개인별 구분값
2	나이	개인식별가능정보	생년월일에서 추출한 정보
3	주소	개인식별가능정보	배송지 주소
4	성별	개인식별가능정보	
5	2019년 1월 여행용품구매액	개인식별가능정보	2019년 1월 여행용품 군의 구매 총액
6	2019년 1월 식품류 구매액	개인식별가능정보	2019년 1월 식품 군의 구매 총액
7	2019년 1월 의류 구매액	개인식별가능정보	2019년 1월 의복 군의 구매 총액
8	2019년 1월 취미용품구매액	개인식별가능정보	2019년 1월 취미용품 군의 구매 총액
9	2019년 1월 생활용품구매액	개인식별가능정보	2019년 1월 생활용품 군의 구매 총액
10	2019년 1월 유아용품구매액	개인식별가능정보	2019년 1월 유아용품 군의 구매 총액
11	2019년 1월 기타 구매액	개인식별가능정보	2019년 1월 6가지 주요 제품군이 아닌 제품의 구매 총액
12	2019년 1월 구매 총금액	개인식별가능정보	2019년 1월 구매 총 금액
13	2019년 1월 선호 제품군	개인식별가능정보	2019년 1월 가장 많이 구매한 제품군
~	~		
217	2021년 12월 생활용품구매액	개인식별가능정보	2021년 12월 생활용품 군의 구매 총액
218	2021년 12월 유아용품구매액	개인식별가능정보	2021년 12월 유아용품 군의 구매 총액
219	2021년 12월 기타 구매액	개인식별가능정보	2021년 12월 6가지 주요 제품군이 아닌 제품의 구매 총액
220	2021년 12월 구매 총금액	개인식별가능정보	2021년 12월 구매 총 금액
221	2021년 12월 선호 제품군	개인식별가능정보	2021년 12월 가장 많이 구매한 제품군
222	2021년 고객 등급	개인식별가능정보	2020년 까지의 구매내역을 통해 산정한 2021년 고객의 등급 P,G,S,B,F

활용데이터 요구 수준표

순번	항목명	요구 수준	비고
1	고객ID	각 고객이 서로 다른 사람이라는 구분만 가능하면 됨	
2	나이	중학생 이상 90세 미만까지 분석 예정 단 90세 이상의 경우에도 다른 대조군과의 비교를 위해 90세 이상 표기 요청	
3	주소	시군구 단위의 분석 예정이며 이에 따라 시군구 단위의 주소 필요	
4	성별	분석목적에 남녀의 차이에 대한 분석이 포함되어 필요	
5	2019년 1월 여행용품구매액	최소 1만단위 이상의 값에 대해서는 정확한 값이 필요	
6	2019년 1월 식품류 구매액		
7	2019년 1월 의류 구매액		
8	2019년 1월 취미용품구매액		
9	2019년 1월 생활용품구매액		
10	2019년 1월 유아용품구매액		
11	2019년 1월 기타 구매액		
12	2019년 1월 구매 총금액		
13	2019년 1월 선호 제품군	제품군이 명확하게 명시되어야 함	
~	~		
217	2021년 12월 생활용품구매액	최소 1만단위 이상의 값에 대해서는 정확한 값이 필요	
218	2021년 12월 유아용품구매액		
219	2021년 12월 기타 구매액		
220	2021년 12월 구매 총금액		
221	2021년 12월 선호 제품군	제품군이 명확하게 명시되어야 함	
222	2021년 고객 등급	5단계의 고객등급이 구분되어야 함	

식별 위험성 검토 결과보고서

가명정보 활용목적	B통계업체는 코로나 이전과 코로나 이후의 상품군별 판매 추이에 대한 통계작성을 위해 당사의 구매정보 데이터 중 2019년 1월부터 12월까지의 주요 상품군별 판매액 정보와 2021년 1월부터 12월까지의 주요 상품군별 판매액 정보를 나이와 성별, 시군구 단위의 주소별 데이터를 분석	
가명처리 대상 데이터 항목	- 고객ID, 나이, 주소, 성별, 2019년 1월~12월, 2021년 1월~12월까지의 여행용품, 식품류, 의류, 취미용품, 생활용품, 유아용품, 기타의 7개 범주의 구매금액의 월별 합계액, 월별 구매 총 금액, 월별 선호 제품군, 각 년도의 고객 등급(전체 222개의 월별) - 전체 고객 820만명 중 50%를 무작위 샘플링하여 구성한 410만명에 대한 데이터	
데이터 위험성	식별성 유무	'고객ID'는 개인식별정보임 '나이', '주소', '성별'은 조합했을 때 개인의 식별이 가능한 개인식별 가능정보임
	특이정보 유무	각 범주별 구매금액의 경우 특이정보로 인한 개인 식별성이 발생할 수 있음
	재식별시 영향도	단순 고객의 구매데이터로 재식별 시 영향도는 크지 않을 것으로 판단됨
처리 환경 검토	이용 및 제공 형태	- 제3자 제공 - 데이터 제공 계약을 체결하여 데이터를 제공 - 데이터 제공 계약에는 재제공 금지, 목적 달성 후 삭제, 재식별 금지 및 재식별 시 조치에 관한 사항들이 포함되어 있음 - B통계업체는 개인정보(가명정보)처리시스템에 대한 ISMS-P인증을 취득하고 있음
	처리 장소	- B통계업체에서 가명정보는 인터넷에 접근할 수 없는 차단된 별도의 분석 PC에서 분석 예정 - 분석PC가 있는 환경은 별도의 분석실로 내부적인 출입통제를 적용하는 것으로 파악됨
	다른 정보와의 결합 가능성	B통계업체는 다양한 통계를 생성하는 업체로 유사 업종에 대한 통계정보 등 결합 가능성이 있는 정보를 보유하고 있음
최종 검토의견*	- 해당 연구는 자사의 데이터를 B통계업체에 제공하는 것으로 데이터 자체 위험성과 처리 환경 위험성을 검토할 때 다음과 같은 조치가 필요함 - 통계전문업체의 특성 상 다른 정보의 결합가능성이 있으나 처리 장소와 개인정보 보호 수준을 검토할 때 결합에 대한 시도는 거의 없을 것으로 판단됨 '고객ID'는 개인식별정보로 개인식별 가능성이 매우 높으며 이에 따라 다시 원래의 정보로 대체할 수 없는 Salt값이 포함된 해시처리 등의 기법의 적용이 필요 '나이', '주소', '성별'은 그대로 사용하는 경우 조합에 의한 개인식별 가능성이 있으며 이에 따라 다음과 같은 처리가 필요 '나이': 물품의 주 구매대상이 아닌 20세 미만의 경우 삭제처리가 필요하며 그 외의 나이에 대해서는 일반적인 구매 분석 통계에 사용되는 10살 단위로 제공하며 80세 이상의 나이에 대해서는 80세 이상으로 처리하는 것이 필요 '주소': 동단위와 상세 주소의 경우 통계목적에 필요하지 않기 때문에 삭제하며 시군구 단위의 주소까지만 사용하는 것이 필요 '성별': 성별은 분석목적에 필요하므로 그대로 사용 - 구매액 관련 정보들은 구매금액별 특이정보를 검토하여 구매 금액에 대한 적절한 수준의 상단 코딩을 적용(19년 8월 취미용품 114,562,000원 → 1억원 이상)하고 금액에 대해서 라운딩을 적용하는 것이 필요 - 고객등급의 경우 식별성이 높은 VIP, S, A를 하나로 B, C를 하나로 D, E, F를 하나로 묶을 필요가 있음	

항목별 가명처리계획

순번	항목명	개인정보유형	제3자 제공	
			처리 방법	처리 수준
1	고객ID	개인식별정보	대체	-일련번호 대체
2	나이	개인식별가능정보	범주화	-10살 단위 범주화
			상하단 코딩	-20세 미만 삭제 -80세 이상은 80세 이상 경계치 입력
3	주소	개인식별가능정보	부분삭제	-동단위 이하 삭제
4	성별	개인식별가능정보	처리 없음	
5	2019년 1월 여행용품구매액	개인식별가능정보	범주화	-상단 99.9%를 초과하는 경우 경계치로 변경 -금액은 다음과 같이 범주화 적용 -0원:0원 -10만단위 미만:1만단위로 라운드 업 -1,000만 단위 미만:10만 단위로 라운드 -1,000만 단위 이상 100만 단위로 라운드
6	2019년 1월 식품류 구매액	개인식별가능정보		
7	2019년 1월 의류 구매액	개인식별가능정보		
8	2019년 1월 취미용품구매액	개인식별가능정보		
9	2019년 1월 생활용품구매액	개인식별가능정보		
10	2019년 1월 유아용품구매액	개인식별가능정보		
11	2019년 1월 기타 구매액	개인식별가능정보		
12	2019년 1월 구매 총금액	개인식별가능정보	범주화	-각 구매액과 동일한 처리
13	2019년 1월 선호 제품군	개인식별가능정보	그대로 사용	
~	~	~	~	
214	2021년 12월 식품류 구매액	개인식별가능정보	범주화	-상단 1.5IQR을 넘는 값은 평균값으로 상단 코딩 -금액은 다음과 같이 범주화 적용 -0원:0원 -10만단위 미만:1만단위로 라운드 업 -1,000만 단위 미만:10만 단위로 라운드 -1,000만 단위 이상 100만 단위로 라운드
215	2021년 12월 의류 구매액	개인식별가능정보		
216	2021년 12월 취미용품구매액	개인식별가능정보		
217	2021년 12월 생활용품구매액	개인식별가능정보		
218	2021년 12월 유아용품구매액	개인식별가능정보		
219	2021년 12월 기타 구매액	개인식별가능정보		
220	2021년 12월 구매 총금액	개인식별가능정보	범주화	-각 구매액과 동일한 처리
221	2021년 12월 선호 제품군	개인식별가능정보	그대로 사용	
222	2021년 고객 등급	개인식별가능정보	범주화	-다음과 같이 범주화 처리 -VIP, S, A → 1등급 -B, C → 2등급 -D, E, F → 3등급

가명정보 처리 기초자료 명세서

신청 기관 정보			
기관명	B동계전문회사		
주소	서울시 은평구 통일로 2368		
데이터명	D2468-20210325.txt	평가 목적	통계작성목적의 데이터 제3자 제공
데이터 수집	2019년 1월부터 12월, 2021년 1월부터 12월까지의 고객의 구매정보를 수집		
이용 방법	코로나 이전과 코로나 이후의 상품군별 판매 추이에 대한 통계작성을 위해 당사의 구매정보 데이터 중 2019년 1월부터 12월까지의 주요 상품군별 판매액 정보와 2021년 1월부터 12월까지의 주요 상품군별 판매액 정보를 나이와 성별, 시군구 단위의 주소별 데이터를 분석		
이용기간	2022년 2월 1일 ~ 2022년 3월 31일		

데이터 명세		
번호	구분	검토사항
1	데이터 특징	전체 고객의 구매정보를 7가지 제품군(여행용품, 식품류, 의류, 취미용품, 생활용품, 유아용품, 기타)으로 분류하여 각 군별 구매금액을 계산한 데이터
2	데이터 생성 방법	전체 고객의 구매정보 820만명 중 50%를 무작위 샘플링하여 410만명을 추출하여 생성 각 제품의 구매금액을 위의 7가지 제품군으로 분류하고 각 제품군의 월별 구매금액의 합계를 계산하여 생성
3	데이터 제공방법	제3자 제공 계약을 통해 제공에 대한 근거를 마련 데이터 제공은 1회성으로 제공하며 제공 시 암호화와 관련 보호조치를 적용한 HDD에 데이터를 복사하여 제공할 예정, 암호화는 한국인터넷진흥원의 안전한 암호 가이드라인에 따라 안전한 방식으로 암호화 하며 암호화에 사용한 키는 별도의 경로를 통해 제공할 예정임 데이터 제공 후 당사의 데이터는 B동계업체의 초기 확인을 거쳐 문제가 없는 경우 바로 절차에 따라 파기할 예정
4	데이터 관리 환경	제공한 가명정보는 인터넷에 접근할 수 없는 차단된 별도의 분석 PC에서 분석 예정 분석 PC가 있는 환경은 별도의 분석실로 내부적인 출입통제를 적용 B동계전문업체는 개인정보(가명정보)처리 시스템에 대한 ISMS-P 인증을 취득 B동계전문업체는 개인정보보호법의 관련 조항에서 명시한 개인정보 보호 조치를 수행하고 있으며 내부 관리계획에 가명정보 관련 항목을 추가하여 안전한 관리를 수행하고 있음

예시 원본 데이터 세부 항목별 명세

컬럼명	명세내용
고객ID	EC.XXXXXX
나이	10살 ~ 97살
주소	시(도)군구 도로명 번지
성별	M/F
2019년 1월 여행용품구매액	2019년 1월 여행용품 구매 금액 합계
2019년 1월 식품류 구매액	2019년 1월 식품류 구매 금액 합계
2019년 1월 의류 구매액	2019년 1월 의류 구매 금액 합계
2019년 1월 취미용품구매액	2019년 1월 취미용품 구매 금액 합계
2019년 1월 생활용품구매액	2019년 1월 생활용품 구매 금액 합계
2019년 1월 유아용품구매액	2019년 1월 유아용품 구매 금액 합계
2019년 1월 기타 구매액	2019년 1월 위의 범주에 포함되지 않는 구매 금액 합계
2019년 1월 구매 총금액	2019년 1월 구매 금액 합계
2019년 1월 선호 제품군	위의 7가지 범주 중 가장 높은 구매 금액 합계액의 제품군
~	~
2021년 12월 식품류 구매액	2021년 12월 식품류 구매금액 합계
2021년 12월 의류 구매액	2021년 12월 의류 구매 금액 합계
2021년 12월 취미용품구매액	2021년 12월 취미용품 구매 금액 합계
2021년 12월 생활용품구매액	2021년 12월 생활용품 구매 금액 합계
2021년 12월 유아용품구매액	2021년 12월 유아용품 구매 금액 합계
2021년 12월 기타 구매액	2021년 12월 위의 범주에 포함되지 않는 구매 금액 합계
2021년 12월 구매 총금액	2021년 12월 구매 금액 합계
2021년 12월 선호 제품군	위의 7가지 범주 중 가장 높은 구매 금액 합계액의 제품군
2021년 고객 등급	2021년 고객 등급(MIP, S, A, B, C, D, E, F)

280 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구(II)

예시 원본 데이터

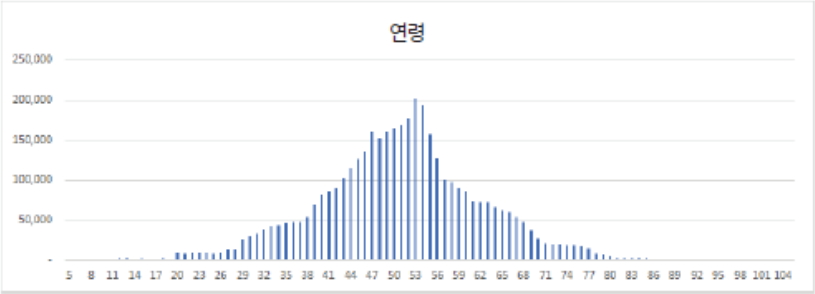
컬럼명	예시
고객ID	EC-232578
나이	36
주소	서울시 은평구 연서로 29길 245
성별	M
2019년 1월 여행용품구매액	0
2019년 1월 식품류 구매액	248,240
2019년 1월 의류 구매액	180,000
2019년 1월 취미용품구매액	740,000
2019년 1월 생활용품구매액	562,500
2019년 1월 유아용품구매액	360,000
2019년 1월 기타 구매액	715,000
2019년 1월 구매 총금액	2,805,740
2019년 1월 선호 제품군	취미용품
~	~
2021년 12월 식품류 구매액	362,150
2021년 12월 의류 구매액	0
2021년 12월 취미용품구매액	24,300
2021년 12월 생활용품구매액	478,130
2021년 12월 유아용품구매액	452,300
2021년 12월 기타 구매액	782,350
2021년 12월 구매 총금액	2,099,230
2021년 12월 선호 제품군	기타
2021년 고객 등급	S

예시 개인정보 속성

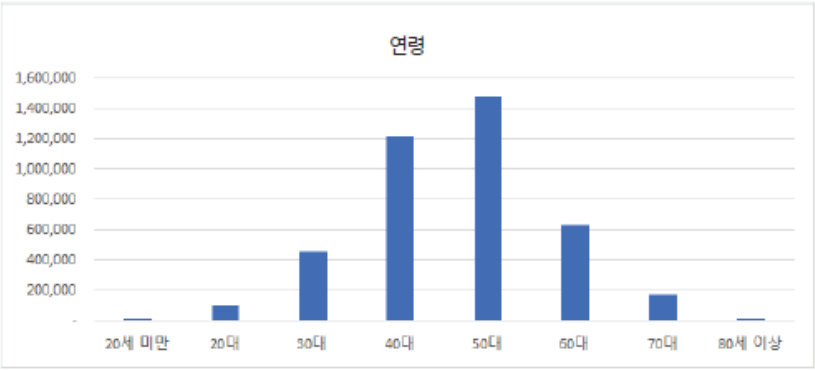
컬럼명	예시
고객ID	개인식별정보(ID)
나이	개인식별가능정보(QI)
주소	개인식별가능정보(QI)
성별	개인식별가능정보(QI)
2019년 1월 여행용품구매액	개인식별가능정보(NSA)
2019년 1월 식품류 구매액	개인식별가능정보(NSA)
2019년 1월 의류 구매액	개인식별가능정보(NSA)
2019년 1월 취미용품구매액	개인식별가능정보(NSA)
2019년 1월 생활용품구매액	개인식별가능정보(NSA)
2019년 1월 유아용품구매액	개인식별가능정보(NSA)
2019년 1월 기타 구매액	개인식별가능정보(NSA)
2019년 1월 구매 총금액	개인식별가능정보(NSA)
2019년 1월 선호 제품군	개인식별가능정보(NSA)
~	~
2021년 12월 식품류 구매액	개인식별가능정보(NSA)
2021년 12월 의류 구매액	개인식별가능정보(NSA)
2021년 12월 취미용품구매액	개인식별가능정보(NSA)
2021년 12월 생활용품구매액	개인식별가능정보(NSA)
2021년 12월 유아용품구매액	개인식별가능정보(NSA)
2021년 12월 기타 구매액	개인식별가능정보(NSA)
2021년 12월 구매 총금액	개인식별가능정보(NSA)
2021년 12월 선호 제품군	개인식별가능정보(NSA)
2021년 고객 등급	개인식별가능정보(NSA)

예시 평가대상 데이터 분포

-연령의 일반 분포(원본)



-연령의 일반 분포(가명처리 후)



(부표 1) 보건복지 분야 전문가 대상 기명정보 결함을 활용한 연구 주제 관련 설문조사 결과

분류	번호	항목	내용
복지	1	연구 주제	가계부채와 삶의 질
		활용 데이터	[법원] 전국 개인파산 개인회생 신청자 데이터: 인구학적, 사회경제적 특성, 채무 상황 [신용회복위원회] 전국 채무조정제도 신청자 데이터: 해당 여부(개인워크아웃, 프리워크아웃, 신속채무조정), 인구학적, 사회경제적 특성, 채무 상황
		연구 내용	가계부채 및 채무조정 신청자와 제도 이용에 따른 삶의 질 변화
		기대 효과	- 채무조정제도의 효과성 평가 - 가계부채가 삶의 질에 미치는 영향 분석
	2	연구 주제	북한이탈주민 취약계층 발굴 및 지원 방안 연구
		활용 데이터	[남북하나재단] 북한이탈주민실태조사 [국민건강보험공단] 건강보험DB: 북한이탈주민의 건보료 납입 기간, 의료지출 내역 등 [한국고용정보원(고용노동부)] 고용보험DB: 북한이탈주민의 실업급여 수급내역
		연구 내용	- 보건·복지 영역을 중심으로 북한이탈주민 취약계층 발굴을 위한 지표 구축 - 보건·복지 영역의 북한이탈주민 취약계층에 대한 실제 거구지 중심의 실질적 지원 방안 연구
		기대 효과	- 주민등록 주소지가 아닌 실제 거주지 중심의 지원 체계 구축 가능 - 북한이탈주민 취약계층의 보건·복지 분야에서의 위기 상황에 선제적 대응 체계 구축 기대
	3	연구 주제	기초생활보장제도 탈수급, 탈빈곤, 재수급 현황 및 탈빈곤 요인에 관한 연구
		활용 데이터	[통계청] 사망통계 [한국사회보장정보원] 사회보장정보시스템: 기초생활보장-차상위 수급 이력, 등록장애인 여부, [국세청] 국세DB: 상시 및 일용근로소득(연간) [고용노동부] 고용보험DB: 실업급여 수급 여부 [교육부] 교육DB: 고등학교 졸업 여부 [한국장학재단] 국가장학금 수급 여부

분류	번호	항목	내용
복지	4	연구 내용	- 기초생활보장 수급자에 대한 패널데이터를 구축하고 수급 이력을 통해 탈수급 및 재수급 현황 분석 - 국제정 및 고용노동부 자료를 연계하여 탈수급자의 탈빈곤 여부 분석 - 탈빈곤에 미치는 영향(특히, 교육의 매개효과) 분석
		기대 효과	- 기초생활보장 수급자의 장기 추적을 통해 탈수급 및 탈빈곤 현황을 파악하고 탈빈곤 요인을 분석하여 향후 탈빈곤을 지원하기 위한 시사점 도출
		연구 주제	이주배경 주민 기초보장 수급 현황
		활용 데이터	[보건복지부(한국사회보장정보원)] 사회보장정보시스템: 기초보장 수급 가구 현황(가구 구성원) [법무부(출입국외국인정책본부)] 등록외국인(거소신고+등록외국인), 귀화자 현황
		연구 내용	- 귀화자의 경우 이주배경을 가지고 있으나, 한국인으로 구분되어 별도로 관리되고 있지 않음. 그러나 현행 국제결혼 허용 규정상 결혼 전에는 배우자의 부모 재산이 합산되지만, 결혼 후 분가하여 기초생계수급으로 전락하는 경우가 많음. 사회보장은 가구 대상이므로 '외국인' 혹은 '이주배경'에 대해 별도로 데이터를 산출할 수 없는 구조임. - 법무부의 국적취득 당시의 정보를 보건복지부와 연계하여 기초보장 수급 가구에 외국인 혹은 귀화자가 포함된 가구, 혹은 국내 출생 자녀가 포함된 가구의 비율을 파악할 필요가 있음. - 또한 기초보장 수급 가구의 이주배경 가구의 최초 입국 시 체류자격을 파악하여 사회보장제도에 대한 매뉴얼 등을 수립할 필요가 있음.
	5	기대 효과	- 특정 체류자격에서 기초사회보장 수급 가구가 많이 발생하는 경우, 이러한 가구의 발생에 대한 출입국제도를 강화하거나, 반대로 대상 가구에 대한 사회적 취약계층으로 별도 관리, 지원하는 제도가 마련되어야 할 것임. - 최근 서울시 외국인주민·다문화가족 실태조사(박민정 외, 2022)의 결과 사회보장 수급의 비율이 일반 국민보다 상당히 높게 나타났고, 밀집 지역에 이러한 가구가 집중되는 경향이 나타나 해당 연구가 시급하다고 판단됨.
		연구 주제	복지사각지대에 놓인 복한이탈주민 파악
		활용 데이터	[통일부] 복한이탈주민 등록 센서스("복한이탈주민명부" 활용) [통일부] 복한이탈주민 사회통합조사: 경제적, 심리적 위기 가구 [보건복지부(한국사회보장정보원)] 사회보장정보시스템: 복지급여 대상자 데이터

분류	번호	항목	내용
복지	6	연구 내용	- 복지사각지대에 있는 탈북민 현황 파악 및 맞춤형 지원 - 경제적, 심리적으로 어려움을 겪고 있는 탈북민이지만, 적절한 복지 급여를 받지 못하고 있는 경우 파악
		기대 효과	- 통일부와 보건복지부의 데이터를 결합하여, 경제사회적 취약계층에 속하는 북한이탈주민 중, 적정 수준의 복지 급여를 받지 못하는 가구를 파악 - 이를 통해 복지사각지대에 놓인 북한이탈주민의 복지를 향상시킴.
		연구 주제	복지급여 전반에 대한 실태를 확인하고 노후보장을 위한 복지급여 정책 개선방향을 도출
		활용 데이터	[통계청] 퇴직연금통계 원자료(행정자료 결합 원자료) [국세청] 국세DB: 퇴직소득세 원자료 [근로복지공단] 중소기업퇴직연금기금제도 원자료
		연구 내용	- 고령화로 인한 노후보장과 관련한 관심이 커짐에 따라, 기존에 주요한 관심에서 소외되었던 퇴직급여에 대한 관심이 증가하고 있음. - 근로자퇴직급여 보장법에 의하면, 퇴직급여제도는 1) 퇴직연금제도, 2) 퇴직금제도, 3) 중소기업퇴직연금기금제도로 구성됨. - 현재, 여러 원자료를 취합해 구성해 매년 통계청에서 발표하는 퇴직연금통계 현황자료는 퇴직연금에만 초점을 두고 있어, 퇴직급여를 구성하는 또 다른 제도인 퇴직금과 중소기업퇴직연금기금제도는 포괄하지 못하고 있음. - 이를 고려한다면, 퇴직연금을 중심으로, 통계청의 퇴직소득세 원자료(별정퇴직금+명예퇴직금에 대한 소득세 정보), 근로복지공단에서 관리하는 중소기업퇴직연금기금제도 원자료를 결합하여 퇴직급여 전체 현황을 확인하는 것에 대한 중요성이 커지고 있다고 판단됨. - 행정원자료를 결합해, 노후보장으로서 퇴직급여의 현실을 더 정확히 인식하여 정책개선방향을 도출할 필요
	7	기대 효과	퇴직급여제도 개선과 관련한 데이터에 기반한 정책적 청사진을 제공하는 데 기여
		연구 주제	노인일자리 사업의 정책 효과성 분석
		활용 데이터	[한국사회보장정보원] 사회보장정보시스템: 성, 연령, 신청이력, 수급상태, 기초연금 급여 등 노인일자리 사업 참여자 정보 [한국노인인력개발원] 행정데이터: 노인일자리 참여유형, 노인일자리 참여 이력, 급여 등 [국민연금공단] 국민연금DB: 기초연금 여부, 국민연금 급여 상태 등 [국민건강보험공단] 건강보험DB: 노인일자리 참여자의 건강상태, 병원이용 등

분류	번호	항목	내용
복지		연구 내용	- 노인일자리사업 참여자의 참여 여부, 참여 이력 및 참여 유형에 따른 정책효과 분석: 소득 변화, 사회보장 급여 수준 변화, 건강 상태 등에 따른 변화 추이 분석
		기대 효과	- 노인일자리 참여에 따른 정책 효과를 확인하고 향후 노인일자리 사업의 정책 개선방안을 제시
		연구 주제	외국인 보건의료 실태
		활용 데이터	[국민건강보험공단] 건강보험DB: 등록외국인(거소신고자, 등록외국인) 건강보험 관련 데이터 [종합병원 등 의료기관] 외국인의 경우 등록번호 앞자리로 식별 가능하므로, 각 병원의 데이터를 가명처리하면 진료과목 및 이용실태에 대한 상세 정보 취득 가능
보건	8	연구 내용	- 체류자격별, 국적별 건강보험 가입 및 이용 실태 분석 : 단기체류(의료관광 포함)와 장기체류자 중 외국인노동자와 정주형 이민자의 이용실태, 주요 질병(진료과목), 환자 수, 서비스 접근성 등 다각적 분석 - 체류자격별, 국적별 전염병 감염 여부, 마약류 복용 및 중독경험, 정신질환 발생 등
		기대 효과	- 외국인의 경우 건강보험을 이용하여 부담 이득을 취한다는 실재와 다른 부정적 언론 보도에 대한 증거를 기반한 현황 제시 가능 - 또한 결혼이민자 부모(F-1-5)의 경우, 수술을 목적으로 한국에 입국한다는 등 외국인과의 관련된 건강보험 이용 관련 사실에 근거한 정보 제공이 필요함. 가명정보를 활용하여 정확한 데이터 기반 분석 결과를 제시할 수 있을 것임.
		연구 주제	건강의 사회경제적 결정요인 규명과 정책과제
		활용 데이터	[한국보건사회연구원·국민건강보험공단] 한국의료패널: 소득, 재산, 교육, 의료이용, 건강상태, 건강행태, 가족관계 등 [한국건강보험공단, 건강보험심사평가원] 건강보험 및 장기요양 의료/요양 행정데이터: 입원, 외래, 의약품 처방, 질병명, 장기요양 데이터 [한국사회보장정보원] 사회보장정보시스템: (소득, 재산, 복지수급 상태 등 [통계청] 사망원인 통계
	9	연구 내용	- 의료패널 조사대상자에 대해 사회보장정보원과 진보공단 '심평원의 행정데이터와 연결하여 시계열로 매칭된 정보를 결합하여, 대상자의 질환 발생과 의료이용, 장기요양수급, 사망에 이르는 경로를 추적함. - 소득/재산/주거형태/주거지역/직업 등 경제적 요인, 가족/사회관계 등 사회적 요인, 건강습관/영양 등 보건학적

분류	번호	항목	내용
보건	10		요인 등이 질병 발생과 의료이용, 사망에 미치는 인과관계를 추정함. 물론 유전적 요인을 추적하기는 쉽지 않음. 다만, 부모의 질병과 사망원인에 대한 조사항목이 있다면 활용이 가능함. - 코호트 혹은 패널 분석을 통해 건강에 미치는 영향 요인을 규명할 수 있고, 거꾸로 건강이 사회경제적 지위에 미치는 요인을 규명할 수 있음. - 분석을 통해 정부와 사회, 직장, 가계가 개인과 가족의 건강 유지와 증진을 위해 무엇을 할 수 있는지, 무엇을 어떻게 해야 하는지에 대한 개입수단을 발굴할 수 있음. - 또한 건강에 대한 투자가 개인과 가족의 사회경제적 지위와 삶의 질을 어떻게 개선할 수 있을지에 대해 규명할 수 있음.
		연구 주제	인구·사회·경제적 특성에 따른 의료이용 경로 및 의료비 지출 비교
		활용 데이터	[국민건강보험공단] 건강보험DB: 전 국민 의료이용 청구자료 [한국보건사회연구원·국민건강보험공단] 한국의료패널: 의료이용형태와 의료비 지출 규모에 관한 정보뿐 아니라 영향을 미치는 요인들을 포괄적으로 분석할 수 있는 데이터
		연구 내용	- 응답 가구의 소득, 부채 및 자산, 지출 자료와 의료비 지출 자료를 연계 분석하여 재난적 의료비 발생 추이를 모니터링하고 관련 요인 분석, 재난적의료비 지원사업의 효과 평가 - 의료서비스 접근성에 따른 의료이용 경로 및 의료비 지출 비교 - 상용치료원 보유 여부에 따른 의료이용 경로 비교 - 민간의료보험 가입 여부 및 가입 유형에 따른 의료이용 경로 및 의료비 지출 비교
		기대 효과	- 의료이용 형태 및 의료비 지출 정보를 행정데이터로 대체함으로써 조사의 효율성을 제고하고, 좀 더 정확한 자료의 활용으로 응답자의 회상에 의존하는 조사자료(패널데이터)의 한계를 극복 - 의료이용 및 의료비 지출과 관련된 인구사회경제적 요인을 다양하게 살펴봄으로써 국민의 건강 형평성 제고를 위한 정책적 개입 방안 마련
	11	연구 주제	예방적 건강행동이 의료이용, 의료비 지출, 건강 결과에 미치는 중단적 영향
		활용 데이터	[국민건강보험공단] 건강보험DB: 전 국민 의료이용 청구자료 및 건강검진 데이터 [한국보건사회연구원·국민건강보험공단] 한국의료패널: 의료이용형태와 의료비 지출 규모에 관한 정보뿐 아니라 영향을 미치는 요인들을 포괄적으로 분석할 수 있는 데이터

분류	번호	항목	내용
보건	12	연구 내용	- 복약준용도 등 만성질환 관리의 유형 분석, 각 유형별 의료이용 및 의로비 지출의 종단적 추이 모니터링, 심뇌혈관질환 이환율 비교 - 건강검진이 의료이용 및 의로비 지출에 미치는 영향, 건강검진 수검 여부에 따른 심뇌혈관질환 이환율 및 사망률 추적 관찰 - 건강생활 습관(신체활동, 흡연, 음주 등)과 의료이용 및 의로비 지출의 종단적 관계 분석, 사망률 비교
		기대 효과	- 의료이용형태 및 의로비 지출 정보를 행정데이터로 대체함으로써 조사의 효율성을 제고하고, 좀 더 정확한 자료의 활용으로 응답자의 회상에 의존하는 조사자료(패널데이터)의 한계를 극복 - 예방적 건강행동이 의료이용 및 의로비 지출에 미치는 영향을 장기적으로 추적 관찰함으로써 1차 예방관리의 효과 평가 및 정책적 근거 마련
		연구 주제	도시 20~40대 1인 가구의 식사 및 구매 식품 패턴과 건강 결과 간의 상관성
		활용 데이터	[통계청] 인구센서스: 대도시 및 광역시 거주자의 등록센터스 데이터 [BC카드, 삼성카드, 현대카드] 카드데이터: 고액정보, 업종별 구매(외식, 배달식 등) 정보 [유통사] 식품 구매 품목 데이터: 신선식품, 가공식품 포함 [국민건강보험공단] 건강보험DB: 건강검진 결과, 진료내역, 진단정보
		연구 내용	- 젊은 성인의 성별, 연령별, 광역시별 1인 가구의 형태를 구별하기 위한 분석 후 가구 형태 구분 - 1인 가구의 식사 관련 행동(외식 여부, 포장식 여부 등) 및 구매 패턴(신선식품, 가공식품, 밀키트 등)에 따른 건강 결과(빈혈, 저체중, 과체중, 비만, 복부비만, 대사증후군 등) 연계 분석
	13	기대 효과	- 통계청, 카드사, 유통사, 국민건강보험공단 데이터를 결합하여 1인 가구의 형태 및 식사 및 구매 식품 패턴과 수와 건강 결과 간의 상관성을 분석하여 국민의 건강과 삶의 질 향상을 위한 기초 데이터 제공 - 도시 거주 1인 가구의 식품 관련 행동과 건강 결과 간의 결과를 기반으로 이들의 건강을 향상하기 위한 지역사회회의 정책을 위한 기초자료 제공
		연구 주제	보건소 기반 취약계층 대상 보건사업의 경제성 평가
		활용 데이터	[한국건강증진개발원] 보건사업의 대상자, 소요 예산, 사업 효과 등 [국민건강보험공단] 건강보험DB: 사업대상자의 의료이용량(진수, 진료비, 질병코드 등) [한국사회보장정보원] 사회보장정보시스템: 사업대상자의 인적 사항, 의료이용량 등

분류	번호	항목	내용
보건		연구 내용	- 보건소 기반의 다양한 보건사업(예: 모바일헬스케어 사업, 방문건강관리사업, 영양개선사업 등)의 건강증진 효과, 의료비 절감, 고용창출 효과 등 경제성 평가
		기대 효과	- 보건사업의 결과에 대한 신뢰도 높은 추계 가능 - 보건사업 예산의 경제성 타당성 제시
	14	연구 주제	건강 및 삶의 질 요인 관련성
		활용 데이터	[통계청] 인구센서스: 인구·사회학적 변수 [국립암센터] 암DB: 암 등록자료, 각 병원 검진센터 [국민건강보험공단] 건강검진DB: 건강상태 [통신사] 통신 사용 현황 [한국도로공사] 도로 이용현황 [한국전력] 전기 사용 현황 [한국환경공단] 대기 및 환경 (오염)지수
		연구 내용	- 각 건강지수와 관련된 사회적 결정인자들의 관련성 파악 - 지수들의 원인-결과 관련성을 파악
		기대 효과	- 추후 스마트 건강 도시가 진행될 경우 건강 예측인자로서 매우 중요해짐
보건 및 복지	15	연구 주제	COVID 19 전후의 사망률 분석: 코로나19로 인한 영향 평가
		활용 데이터	[한국사회보장정보원] 사회보장정보시스템: 기초생활보장 및 차상위급여 수급 여부 [국민연금공단] 국민연금DB: 연금수급 여부, 연금액 [질병관리청] 코로나19 백신 접종 [국민건강보험공단] 건강보험DB: 건강보험 코호트 데이터, 건강보험 질병코드, 보험 유형 및 보험료 [통계청] 사망통계
		연구 내용	- 코로나19 전후로 사망률의 변화를 분석하고 코로나19의 직접적 영향으로 인한 사망률 증가와 사회적 거리두기 등 간접적 영향으로 인한 사망률 - 코로나19의 직접적 영향으로 인한 사망과 간접적 영향으로 인한 사망의 특성(지역, 연령, 성별, 사회경제적 자원 등) 차이 분석

분류	번호	항목	내용
보건 및 복지		기대 효과	- 코로나19와 같은 질병-재난의 직접적 영향 뿐 간접적 영향에 취약한 집단에 대한 파악을 통해 정책적 시사점 도출
		연구 주제	외국인에 대한 사회보장 확대 적용과 재정건전성에 관한 연구
		활용 데이터	[법무부(출입국·외국인정책본부)] 체류자격별 외국인 명부: 영주자격, 재외동포자격 등 [한국사회보장정보원] 사회보장정보시스템: 범정부 복지대상자 자격, 수급이력정보 [국민건강보험공단] 건강보험DB: 의료급여 내역 등, 건강보험 가입 정보 [한국고용정보원] 고용보험DB: 고용보험 사용자 정보, 고용보험 가입정보, 외국인고용관리시스템: 회원자료 [국민연금공단] 국민연금DB: 국민연금 가입 정보 [대법원] 전자가족관계시스템: 가족관계 등록자료 [보건복지부] 장애인정책과: 장애인 등록 명부 [근로복지공단] 산재보험DB: 산재보험 가입 정보 [국세청] 국세DB: 개인별 연말정산 자료(근로/사업/기타 소득 등), 과세 자료 등
	16	연구 내용	- 지역별, 체류자격별, 연령계층 국내 체류외국인 현황을 개관 - 귀화자 등 한국국적을 취득한 외국인 집단 및 외국국적 보유 외국인 집단의 복지 수급 이력정보를 파악 - 연령계층별, 소득수준별 외국인 구성의 변화에 따른 잠재적인 미래 사회복지 수요를 다각적으로 분석(사회보험/공공부조 및 사회서비스) - 유사한 연령대, 소득수준의 내-외국인 간 사회보장제도 접근성의 차이를 분석
		기대 효과	- 현 제도하에서 외국인에 대한 사회보장제도 운영의 재정적 부담 수준 파악 가능 - 외국인에 대한 사회보험, 공공부조 및 사회서비스 수급 범위 확대 시, 서비스의 범위, 수급대상 범위 등 확대 시 나리오에 따른 잠재적인 복지재정 수요 파악 가능 - 외국인 집단의 소득수준을 기반으로 한 사회보험 방식 및 보험료 산정 체계 마련을 위한 기초자료 작성 - 중앙 및 지방정부의 지속 가능한 재정건전성 및 재정자립을 고려한 외국인 사회보장 제도 설계를 위한 기초자료 제공

분류	번호	항목	내용
보건 및 복지	17	연구 주제	치매환자의 진단부터 사망(장례)까지 보건·복지 서비스 이용 실태 분석
		활용 데이터	[건강보험심사평가원] 건강보험 심사자료: 병원 내원 일시, 진단코드, 청구내용 [중앙치매센터] 치매안심센터 이용 정보 및 시점(누적) - 치매검진(선별검사, 진단검사, 감별검사 등) - 치매환자서비스(배회가능 어르신 인식표, 치매치료관리비 지원, 조호물품, 쉼터, 사례관리) - 치매가족지원서비스 등 [건강보험공단] 노인장기요양보험DB: 노인장기요양보험서비스 이용 - 지역, 일반 급여종류(6개 재가급여, 2개 시설급여), 치매관련급여 - 기관의 상세정보(기관의 기본정보, 평가정보, 정원/현원, 인력현황, 시설현황 등) - (재가급여 이용하는 경우)배치된 요양보호사, 기관 입소 및 퇴소 시점 [국민건강보험공단] 건강보험DB: (일부 이용자 분석 가능)노인코호트DB의 정보 - 자격(성, 연령, 지역, 가입자 구분, 소득분위 등 대상자의 사회경제적 변수 및 장애, 사망), - 진료(의과 보건의료(T1), 치과 한방(T2), 약국(T3)자료에 대한 명세서(20t), 진료내역(30t), 상병내역(40t), 처방전교부상세내역(60t) - 건강검진: 1차 일반건강검진 주요 결과 및 문진에 의한 생활습관 및 행태 - 노인장기요양: 장기요양 신청 및 판정 결과(T01), 인정조사사항(T02), 청구명세서(T04), 장기요양시설 현황 (T05) 변수(T01) 21개, (T02) 81개, (T04) 12개, (T05) 25개 변수 [한국사회보장정보원] 사회보장정보시스템: 노인맞춤돌봄서비스 등 노인복지관 서비스 이용 정보, (일부 이용자 분석 가능)취약노인지원시스템 이용 정보(취약노인지원시스템) [한국노인인력개발원] 노인일자리 및 사회활동지원사업 공익활동의 노노케어 사업 수혜 정보 [통계청, 행정안전부] 인구동향조사의 사망 - 사망자의 성명, 주민등록번호, 주소, 국적, 최종 졸업학교, 직업, 혼인상태 - 사망일시, 사망장소, 사망종류, 사망원인, 신체상황, 발병부터 사망까지의 기간 - 의사·한의사 등 진단자 구분, 의사명, 수술 및 해부소견, 병원명, 병원주소 - 사고사 등 외인사(外因死)인 경우 사고 종류, 의도성 여부, 사고일시, 사고지역, 사고장소 - 신고인 자격 및 사망자와의 관계 [보건복지부] e하늘장사정보시스템: 화장, 장사, 장사시설서비스 등

분류	번호	항목	내용
보건 및 복지	18	연구 내용	- 치매환자의 진단부터 사망까지 통합적 보건, 복지 서비스 이용 실태 - 치매환자 지원 서비스 간 연계 현황 파악 - 통합 사례관리 방안 - 치매환자의 보건복지 서비스 이용에 따른 건강 변화 - 치매환자의 총 의료비 지출 및 사회적 지원 금액 추정 - 치매환자의 거주 개적(집, 요양원, 요양병원, 병원, 호스피스 등) - 효과성 분석(치매환자): 치매안심센터 서비스, 노인장기요양보험서비스, 노인맞춤돌봄서비스, 노인맞춤돌봄서비스 응급안전안심서서비스, 노노케어 등 - 치매환자의 웰다잉: 사망장소, 장사·장례 등
		기대 효과	- 보건, 복지 서비스 간 연계성을 강화하여 이용자 중심에서 통합적 실태 파악 - 치매 진단, 치료 및 서비스 이용, 사망까지 치매 진행 경과에 따른 보건복지서비스 이용 변화 궤적 파악 - 치매와 관련된 개인, 가족, 국가 차원의 지출(비용) 추정 - 다양한 보건복지서비스 효과성 분석을 통해 효율적인 고품질의 서비스로 제도 개선 - 효과성 높고, 비용 절감할 수 있는 보건복지서비스 제공
		연구 주제	복지서비스 유형에 따른 노인(장애인)의 의료-복지 통합돌봄의 실효성 예측
	18	활용 데이터	[통계청] 인구센서스 [국민건강보험공단] 장기요양보험DB: 장기요양보험등급 신청 및 판정 데이터, 건강보험DB: 의료비 지급 현황 [한국사회보장정보원] 사회보장정보시스템: 행복e음 데이터
		연구 내용	- 노인(장애인)의 성별, 연령, 장애 유무, 지역 등에 따라 제공받은 복지 사업을 분석하고 국민건강보험공단 장기요양등급 신청 및 판정 실태 제시(각 연도) - 노인(장애인) 의료비 지출 현황 분석(각 연도 증감실태 분석) - 요양보호시설 입주 및 퇴원 여부, 현재 거주지와 매칭하여 분석
		기대 효과	- 제공받은 복지사업에 따라 노인(장애인)의 장기요양등급 신청과 판정 여부, 장기요양등급 재신청 여부에 대한 분석 가능 - 제공받은 서비스 유형에 따라 의료비 지출의 증감 부분도 확인 가능 - 노인(장애인)의 지역사회 지속 거주를 위한 수요자 중심 맞춤형 복지정책 수립 기여