料に対する様々な需要を満たすことができる広範囲かつ包括的な領域の研究用データベース

- 5年間、健康保険で請求された全ての患者を対象とし、長期間の標本データベースを構築
- 韓国統計学会、保健情報統計学会など専門家による諮問を予定。
- 2年単位で新規構築し、10年単位の標本データベースを構築することを目標。
- 時系列に基づき、患者の追加削除情報を統計学的便宜なく反映可能。

［図］情報提供の多様化の体系図

<table>
<tr><th colspan="4"></th><th>提供対象</th><th>提供方法</th></tr>
<tr><td rowspan="6">情報提供の多様化</td><td rowspan="4">標本資料</td><td rowspan="2">1年単位の標本資料</td><td>2009年 / 2010年 / 2011年 / 毎年アップデート</td><td rowspan="2">全ての研究者</td><td rowspan="3">直接提供</td></tr>
<tr><td>全体患者標本（NPS）、入院患者標本（NIS）、高齢者患者標本（APS）、小児・青少年患者標本（PPS）</td></tr>
<tr><td rowspan="2">臨床標本コホート</td><td>特定集団を選択した後、毎年追跡観察</td><td rowspan="2">公益目的</td></tr>
<tr><td>神経外科
産婦人科 ⟶ 特定領域コホートの拡大</td></tr>
<tr><td rowspan="2">大規模・長期間の標本データベース</td><td>5〜10年単位の長期間標本データベースの構築（500万人以上）</td><td rowspan="2">公益目的</td><td rowspan="2">遠隔サービス</td></tr>
<tr><td>5年間の標本データベース ⟶（2年単位でアップデート）⟶ 10年間の標本データベース</td></tr>
<tr><td rowspan="2">パーソナライズサービス</td><td rowspan="2">資料処理室</td><td>HIRAビッグデータデータベースの活用</td><td rowspan="2" colspan="2">協議事項</td></tr>
<tr><td>需要者の研究目的に合った資料の構築</td></tr>
<tr><td rowspan="2">外部連携</td><td rowspan="2">連携サービスデータベースの構築（共有）</td><td>調査の設計方法に応じて異なる</td><td rowspan="2" colspan="2">協議事項</td></tr>
<tr><td>国家機関の調査資料連携サービス
国民健康栄養調査、患者調査資料、がん登録資料など</td></tr>
</table>

4）標本資料構築のためのワーキンググループの運営

○ 臨床分野別に、研究テーマに対する需要の把握およびデータセット構築アイデアのための目的で3つの臨床分野（産婦・小児・青少年科、外科、内科）および医薬品分野とワーキンググループを結成

- 提供資料に対する研究利用の活性化方案および改善事項の把握、需要者中心の研究データセットへの拡張方案に関する議論を行う。
- これらを通して、専門家の意見を反映した保健医療情報研究ガイドラインを設定。
- 保健医療データの需要調査に応じて、研究者らに体系的な研究資料を支援。
- 診療情報データセットを構築し、研究者の参加拡大により研究資料を提供。

○ 関連専門家と意見交流を行うため、持続的に学会を代表する担当諮問委員が必要

※ 学会を代表する学会長に変動があった場合、標本資料に関する内容を把握できていな

いという問題が発生。

- 標本資料の開発や活用の拡大方案などに関する学会の意見の取りまとめ、および協力体系の維持。

○ ワーキンググループを通して、傷病コードの不正確性による診断名の妥当度に対する問題と主要疾病の基礎統計情報（発生率/有病率）算出の必要性を強調

○ 他機関とのデータの連携、個人（患者）単位から患者の家族単位への拡大の必要性などが分野別に共通するワーキンググループの主要争点として把握されている

○ 持続的なワーキンググループ会議を通して、こうした問題に対する解決方案を摸索する予定

개 요

1장. 연구 배경 및 목적

1. 연구배경

○ 의료·IT기술 발달 및 의료사용량 증가에 따라 보건의료 관련 데이터 급증
- 우리나라는 삶의 질 향상 및 고령화 사회진입에 따라 고품질 의료서비스 수요증가하고 의료IT 기술력의 부상 등의 요인으로 보건의료 관련 데이터는 급증

○ 공공기관이 보유한 빅데이터는 미래의 핵심가치로써 중요성이 급부상 하고 있음
- 데이터 활용에 대한 사회적 관심 증가와 더불어 의료정보제공이나 지원활동 등 서비스 확충 요구의 지속적 증가
- 공공기관이 보유한 다양한 데이터는 국가정책 수립 및 국민의 건강증진에 관련한 연구에 기초자료로 활용되어질 수 있음

○ 공공 빅데이터를 원하는 사회적 요구 증대
- 자료의 활용 활성화를 위한 제공채널의 다양화 및 연구자 지원체계의 확대 등
- 정보제공 인프라 구축을 위해서 범정부적인 '인텔리전스'를 확보할 수 있는 보건의료관련 '빅데이터' 플랫폼을 구축 방안 검토 필요

○ 최근 정부 정책은 국민맞춤형 서비스를 위한 공공빅데이터의 적극 공유·공개의 방향
- 정부부처 및 산하기간은 공공빅데이터를 활용한 국민맞춤형서비스 및 과학적 행정구현 등에 대한 공공기관 간 협력에 대한 공감대를 형성해 나가는 과정에 있음

○ 건강보험심사평가원은 다양하고 방대한 데이터를 구축·보유하여 보건의료 발전을 위해 적극 공개하고자 함
- 건강보험심사평가원은 전 국민의 98%가 등록된 우리나라 보건의료를 대표할 수 있는 건강보험청구데이터 뿐만 아니라 보건의료와 관련한 다양한 데이터를 운용하고 있음

2. 연구목적

○ 건강보험심사평가원의 "보건의료 빅데이터 플랫폼" 구축방안 검토
- 건강보험심사평가원의 미래 핵심 가치 창출을 위한 정보화 방안은 R&D 강화를 위한 IT인프라구축, 연계서비스를 통한 데이터의 다양성 확대, 수요자중심의 맞춤형 컨텐츠 제공을 통한 고객서비스 극대화 실현을 들 수 있음
- 보건의료 빅데이터 플랫폼의 정의를 IT인프라구축 뿐만 아니라 데이터의 다양성 확대, 서비스 개선으로 확장하여 실행방안 검토

○ 빅데이터 플랫폼 실현을 위한 IT인프라 구축
- 보건의료 빅데이터플랫폼 실현을 위해 빅데이터DB, 빅데이터분석시스템, HIRA-dream포탈 구축

○ 빅데이터플랫폼 실현을 통한 자료연계서비스 환경 구축으로 제공 가능한 정보의 다양성 확대
- 공공데이터의 공유·공개를 위한 기관간의 자료연계시스템 구축
- 수요자 중심의 정보제공 뿐만 아니라, 정부의 보건의료 정책 발전과 보건의료분야 및 기타 학술연구의 활성화에 목적을 두고 다양한 정보를 제공

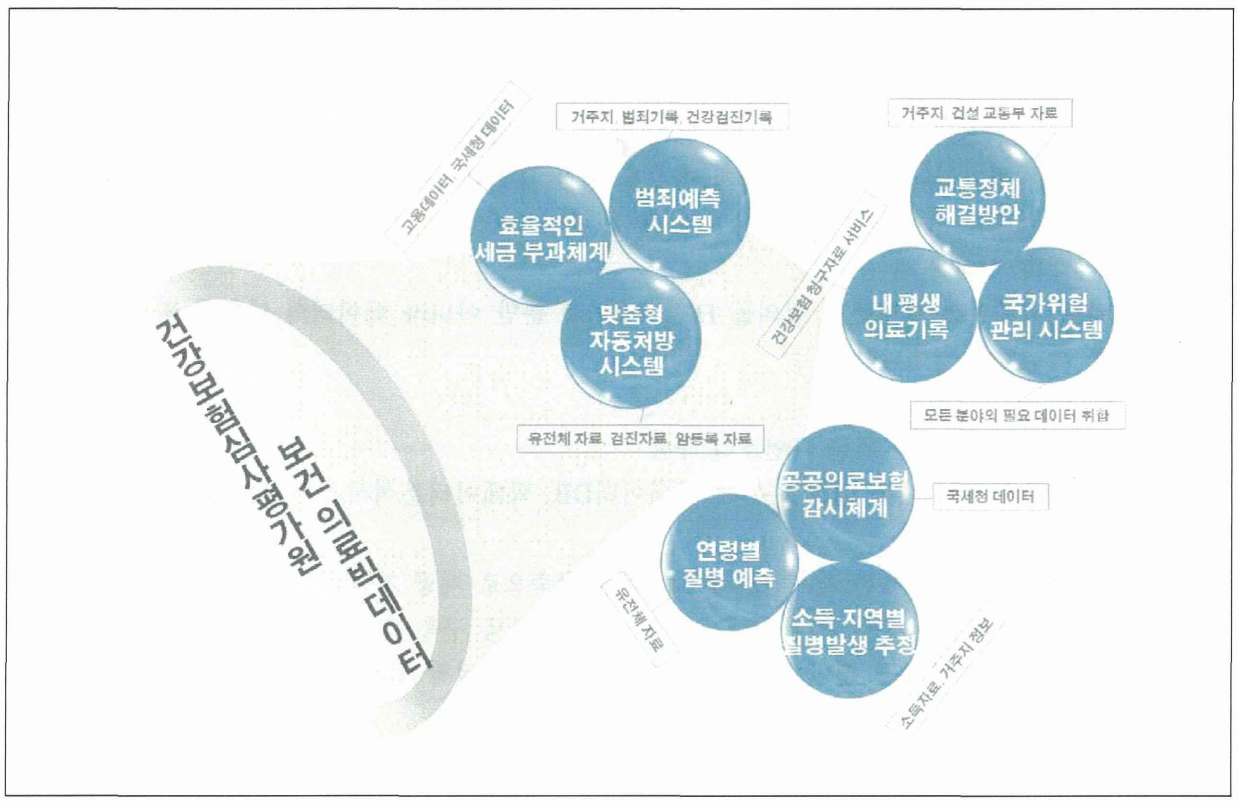
○ Onestop Service를 통한 고객서비스 극대화 실현
- DATA취득과 분석, 이용절차 등 손쉬운 자료이용 접근을 위한 Onestop Service를 통하여 연구 환경을 조성하고, 수요자의 상시 이용이 가능하도록 정보제공에 필요한 제도적인 기반을 조성

○ 맞춤형 건강정보 콘텐츠를 개발
- 심사평가원이 보유한 건강정보 데이터를 체계적으로 분석하여, 수요자의 눈에 맞춘 맞춤형 건강정보 콘텐츠를 개발하고 제공하는 방안을 검토

○ 의료 빅데이터 분석 전문가를 양성
- 보건의료분야 빅데이터의 지속적이고 체계적인 활용을 위한 의료 빅데이터 분석 전문가 양성 프로그램 개발

○ 보건의료분야의 다양한 정보제공을 통한 R&D활성화
- 다양한 정보를 수요자 중심의 맞춤형서비스로 제공하여, 공공기관이 보유한 정보를 기반으로 새로운 사업수요 창출 및 보건의료산업의 생산성을 증대
- 민간경제 활성화를 위한 공공부문의 역할이 강화
- 전 국민 대상인 건강보험청구자료가 보건의료분야의 국가정책 수립 및 학술연구의 기초자료로 활용됨에 따라, 보건의료정보의 근거와 분석을 기반으로 가치를 창출하고 보건의료분야의 국가정책 수립 및 의사결정을 지원할 수 있을 것으로 기대

<그림> 빅데이터 플랫폼 구축을 통한 다양한 가치 창출



2장 보건의료 자료의 현황 및 개요

1. 건강보험심사평가원 보유자료 개요

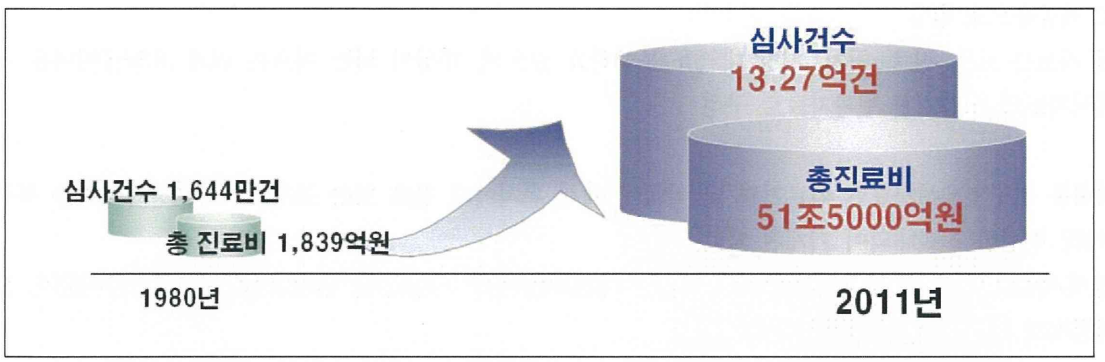
○ '건강보험 청구자료'란 의료기관에서 환자의 진료비용 중 '국민건강보험'이 부담하는 부분에 대해 지급의뢰를 하기 위해 '건강보험심사평가원'에 보험급여청구를 하면서 발생되는 정보

- '건강보험심사평가원'은 의료기관에서 청구하는 '청구명세서'에 대한 진료비의 적정성을 심사하고 그 결과를 '국민건강보험공단'에 전달하여 지급을 요청

○ 우리나라의 1년간 '건강보험 청구 환자 수'는 2011년 기준 45,804,866명으로 주민등록인구 50,734,284명의 90.3% 에 해당함

- 건강보험 청구에 대한 심사건수와 청구 진료비 총액은 꾸준히 증가하여 2011년 기준으로 심사건수는 약 13억 건, 총 진료비는 약 51조 5천 억원에 달함

[그림 1] 2011년 심사건수 및 총 진료비

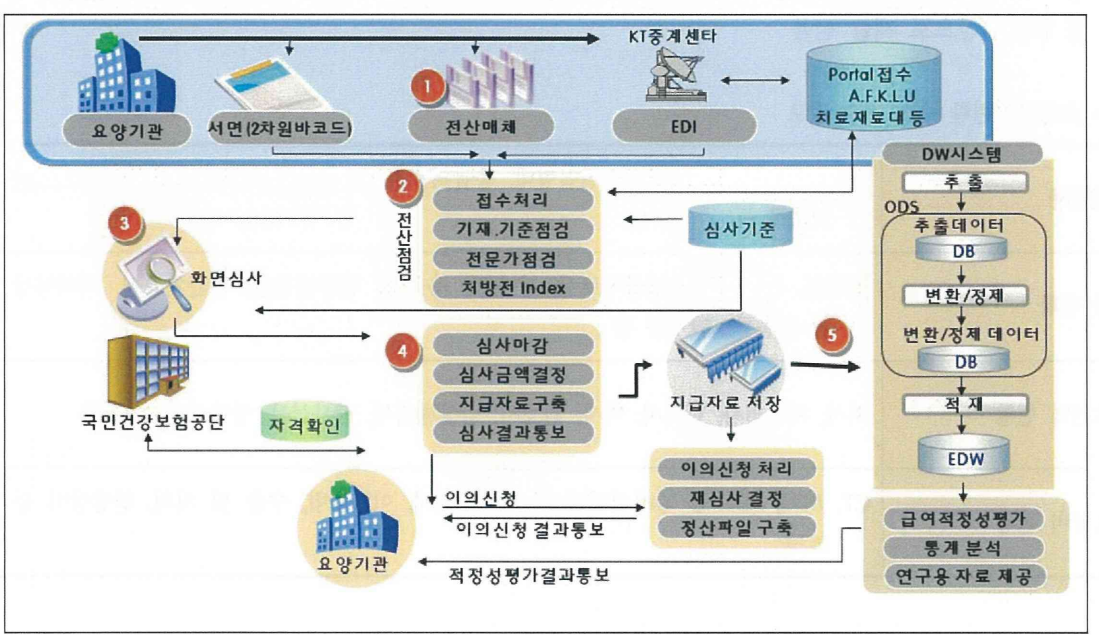


○ 건강보험에 등록된 요양기관 현황은 1980년대 7,728개소 에서 2011년 기준으로 82,948개소로 증가

[그림 2] 2011년 건강보험 등록 요양기관 현황

○ 병원에서는 환자를 진료하고 급여상환을 위해 건강보험심사평가원으로 [그림 3]의 절차에 따라 요양급여비용을 청구
- 요양기관으로부터 진료비 청구 명세서가 접수되면, 전산점검 및 심사 단계를 거쳐 데이터 누적

[그림 3] 심사처리절차 흐름도



○ [그림 3]과 같이 축적된 데이터는 DW(Data Warehouse) 시스템으로 저장되어, 급여적정성평가나 통계 분석, 연구자료 제공용으로 활용

- 수집된 자료는 모든 진료, 처치, 처방 내역을 포함하고 있으며, 대상이 되는 자료는 크게 요양급여비용 청구명세서자료와 요양기관 현황자료로 구성

○ 요양급여비용 청구명세서자료란 의료기관 및 약국 등에서 환자에게 진료 또는 조제한 후 요양급여비용 청구방법에 따라 작성한 진료내역이 기재된 자료

- 청구명세서는 EDI, 전산매체(디스켓, CD). 또는 서면으로 청구 가능하며, 건강보험환자, 의료급여환자, 보훈국비환자의 진료비 청구 내용을 모두 포함

<표 3> 건강보험 청구자료 주요 변수정보

일반 사항	분류코드	수진자 성명, 수진자 대체키, 가입자 번호, 사업장 번호, 요양기관 분류코드
	세부사항	상병, 수술여부, 진료과목, 요양개시일, 입원일수, 외래방문일수, 처방전수, 처방일수, 초진 및 재진 회수, 진료결과, 요양급여비용 총액, 본인일부부담금, 보험자부담금 등
진료세부내역		진료비 항목별 내역(행위, 의약품, 치료재료별로 세분화하여 기재) 처방조제 상세내역(개별 약품별 처방 및 조제내역)

○ 제출받은 요양기관 현황자료에는 요양기관의 일반현황, 병상 현황, 의료인력 현황, 의료장비 현황이 포함

- 요양기관 현황자료는 청구명세서자료와 함께 심평원 데이터베이스에 저장 관리되며 통계자료로 활용
- 요양기관 현황자료는 심평원에 청구한 요양급여비용을 심사•평가하는데 필요한 기초자료로 활용하기 위하여 요양기관으로부터 법정 서식인「요양기관현황통보서」 및 「요양기관변경사항통보서」를 통해 최소 월 1회 전산으로 제출 받음

<표 > 요양기관현황 주요 변수 정보

일반현황	개설자의 인적사항, 주소, 설립형태, 응급의료기관, 개방병원 등 운영현황 및 개설진료과목 등
병상 현황	입원실, 특수진료실(집중치료실, 수술실, 응급실, 인공신장실, 무균치료실, 격리병상 등 12종류), 낮병동 등
의료인력 현황	의사, 치과의사, 한의사, 약사, 간호사, 임상병리사, 방사선사, 영양사 등 24분류
의료장비 현황	CT, MRI, PET 등 방사선진단 및 치료, 검사, 이학요법, 수술 및 처치, 한방장비 등 167종

○ 건강보험심사평가원은 건강보험청구자료와 요양기관현황자료 외에도 다양한 자료를 <표>와 같이 보유

<표> 심평원 보유자료 현황

자료명	자료내용	자료건수 (용량)	활용내용
합계 : 15종			
요양급여비용청구 명세서 DB	요양기관(병·의원, 약국)의 진료비 청구명세서(일반내역, 상병내역, 진료내역, 원외처방내역)	7,581백만건 (33TB)	요양급여비용심사, 평가지표산출 및 각종 통계생산
요양기관현황 DB	요양기관현황정보(요양기관기호, 요양기관명, 소재지, 전화번호, 우편번호, 설립일자, 종별구분, 의사수 등 인력현황, CT장비수 등 장비현황, 병상수 등)	8만건 (0.1GB)	요양기관별 일반현황 및 청구자료 연계 분석 활용
의약품처방조제 정보 DB	요양기관에서 수진자에게 처방·조제하는 병용금기,연령금기 등 처방전정보	116억건 (6.5TB)	의약품 안전성관련 정보 의약사에게 실시간 제공 및 DUR 운영 현황 모니터링
의약품 생산실적 DB	의약품생산내역(제조사,제조년월,생산량,생산금액 등)	50만건 (3GB)	국내의약품 생산현황 분석
의약품 수입실적 DB	의약품수입내역(수입사,수입년월,수입량,수입금액 등)	8만건 (1GB)	국내의약품 생산현황 분석
의약품 공급실적 DB	의약품공급내역(공급업체,구입업체,공급일자,공급량,공급금액 등)	12억건 (6TB)	의약품 유통현황 및 의약품 공급·청구 비교분석
안전상비의약품 편의점 정보 DB	안전상비의약품 판매처 정보	17천건 (0.2GB)	안전상비의약품판매처 관리
의약품 RFID Tag 정보 DB	RFID tag부착의약품 정보	10백만건 (24GB)	RFID부착의약품정보제공 및 관리
요양기관종합정보 DB	일반(의료관명, 기호, 대표자)인력(성명,주민번호,자격,근무)장비(종류,코드)	22백만건 (4TB)	요양급여비용지급 차등제 등에 활용
통계자료 DB	건강보험통계연보(건강보험 청구건수, 진료비현황 등) 진료비통계지표(건강보험/의료급여 심사실적 등)	20만건 (84MB)	통계발간 보건의료연구 정보공개
질병통계 DB	상병소분류 통계 (3단상병의 환자수 및 진료비 등) 상병세분류 통계 (4단상병의 환자수 및 진료비 등)	186만건 (478MB)	통계발간 보건의료연구 정보공개
행위통계 DB	행위별 통계(수가, 보험등재약, 치료재료 등의 진료코드에 따른 실시횟수 및 금액)	50만건 (262MB)	통계발간 보건의료연구 정보공개
병원평가정보 DB	급성심근경색증 / 급성기뇌졸중 / 고혈압 / 당뇨병 / 혈액투석 / 의료급여 정신과 / 수술의 예방적 항생제 / 수술별 진료량 / 진료결과 / 제왕절개분만 / 관상동맥우회술 / 항생제 처방률 / 유소아 중이염 항생제 / 항생제 처방률 / 주사제 처방률 / 약품목수 / 처방약품비 / 요양병원 / 대장암 평가결과	4억건 (50GB)	정보공개 보건의료연구 요양기관 의료 질향상관리
요양병원환자평가표 DB	요양병원 환자평가표(요양기관, 환자 인적사항, 입원일, 혈압, 인지기능, 신체기능, 배설기능, 질병진단, 건강상태, 피부상태, 특수처치)	12백만건 (6GB)	요양병원정액수가 산정 및 요양병원 심사시 활용
사망의심자자료	사망자료(인적사항, 사망일자, 시도코드, 복지대상여부)	1백만건 (100MB)	행정정보공동이용센터 사망정보 연계

○ 건강보험 청구자료 활용 시 주의사항으로 급여가 인정된 의료이용 내역만 포함되어 있기 때문에 비급여 내역 또는 처방전 없이 구입할 수 있는 아스피린 등의 일반의약품에 대한 정보는 청구자료에서 확인 불가

○ 진단명의 정확성에 대한 연구자의 고려가 필요
- 진단명의 정확성은 외래보다는 입원환자, 다빈도 경증질환자 보다는 위중한 환자에서 높게 나타나며 의원급보다는 종합병원급 요양기관에서 더 높은 경향이 있음
- 진단명 및 시술에서 의사의 개인차, 관습적 요인을 완전히 배제하기 어렵기 때문에 자료의 특성, 환자의 의료이용행태와 질병의 고유특성, 의사의 진료과정과 임상환경, 병원의 전산망과 청구과정, 건강보험급여제도 등을 충분히 파악해야 올바른 해석이 가능
- 진단명의 타당도 등을 주기적으로 평가하여 청구자료의 지속적인 품질관리를 통해 신뢰성을 높이는 것도 중요한 과제

2. 국내 보건의료자료 보유 현황

○ 기관별 보유자료원 형태와 특징에 따라 포본방식이 아닌 전국을 대상으로 하는 모집단의 자료원은 건강보험청구자료(건강보험심사평가원), 중앙암등록자료(국립암센터), 자격자료(국민건강보험공단), 전국다문화가족실태조사(한국보건사회연구원), 영아모성사망조사(한국보건사회연구원), 사망원인자료(통계청), 신생아자료(통계청), 거주지 및 사망자료(안전행정부), 소득수준자료(국세청) 정도로 구분이 가능
- 표본단위 자료는 자료연계 시 필요한 사전동의를 획득이 용이하고, 그 용량이 작아 기관간의 동의가 따르면 USB메모리만으로도 이동 가능

<표> 기관별 보유 보건의료 자료

자료명	수집 방식	수집 주기	생산 기관
건강보험청구자료	전국민 대상	매일 수시로 데이터 발생	건강보험심사평가원
중앙암등록자료	전국(암환자 대상)	최초 발생 이후 의료기관 이동 혹은 완치 시 갱신	국립암센터
국민건강보험공단 건강검진자료	당해 검진대상자	2년에 1회 혹은 1년에 1회 검진	국민건강보험공단
자격 자료	전국민 대상	변동사항 발생 시	
국민구강건강실태조사	학교표본 (어린이 집 및 초•중•고 학생 대상)	매 3년마다 1회	보건복지부
국민건강영양조사	가구표본(전국민 대상)	매년 1회	질병관리본부
지역사회건강조사	가구표본(전국민 대상)	매년 1회	
청소년건강행태온라인조사	학교표본(중1~고3학교 학생 대상)	매년 1회	
퇴원손상심층조사	병원표본(퇴원환자 대상)	매년 1회	
노인실태조사	가구표본(노인 대상)	매3년 마다 1회	한국보건사회연구원
베이비부머의 생활실태 및 복지욕구	가구표본(1951~1964년생 대상)	2010년 1회	
장애인실태조사	가구표본(장애인과 비장애인 대상)	2000년, 2005년, 2008년, 2011년 연도별 1회	

전국다문화가족 실태조사	전국(다문화 가정 대상)	2009년 1회	
전국 출산력 및 가족 보건복지실태조사	가구표본(기혼 및 미혼남녀 대상)	매 3년마다 1회	
영아모성사망조사	전국(영아모성사망자 대상)	약 3~4년 주기 1회	
차상위계층 실태조사	가구표본(차상위 계층 대상)	2007년 1회	
환자조사	의료기관표본(전국환자대상)	매년 1회	
한국복지패널	가구표본(전국민 대상)	2006년 기준 연단위 추적 조사	
한국의료패널	가구표본(전국민대상)	2008년과 2009년 반기별 추적 조사	
고령화연구패널	인구표본(45세 이상 국민 대상)	2005년 기준 2년단위 추적 조사	한국고용정보원
한국노동패널	가구표본(전국민 대상)	1998년 기준 연단위 추적 조사	
사회조사	가구표본(전국민 대상)	매년 1회	통계청
통계청 사망원인자료	전국(사망자 대상)	사망 시 발생	
통계청 신생아 자료	전국(신생아 대상)	출산 시 발생	
거주지 및 사망자료	전국(이주자 및 사망자 대상)	거주지 이동 및 사망시 발생	안전행정부
소득수준 자료	전국민 대상	변동사항 발생	국세청

○ 전국을 대상으로 하는 모집단의 자료원 중 특정계층(영아, 암환자 등)이 아닌 전국민을 대상으로 하는 자료원은 "건강보험청구자료", "소득수준 자료" 정도로 다시 제한

- 전국단위·전국민대상의 자료원에서 자격자료와 소득수준 자료의 경우, 수시가 아닌 변동사항이 있을 때에만 업데이트 하므로 데이터 양 자체가 그리 많지는 않으나, 건강보험 청구자료의 경우 전국민을 대상으로 수시로 생성되고 방대한 용량을 가지기 때문에 상당한 초기 투자비용[1](약 700억 원)과 유지·보수 비용 발생(연 60억 원 이상, 전산직 160명 이상)

<표> 모집단의 자료원

자료명	수집 대상	수집 시기
건강보험청구자료	전국민 대상	매일 수시로 데이터 발생
자격자료	전국민 대상	변동 사항 발생시
중앙암등록자료	암환자 대상	(최초 발생 이후) 의료기관 이동 혹은 완치 시 갱신
통계청 사망원인자료	사망자 대상	사망 시 발생
통계청 신생아 자료	신생아 대상	출산 시 발생
거주지 및 사망자료	이주자 및 사망자 대상	거주지 이동 및 사망시 발생
소득수준 자료	전국민 대상	변동 사항 발생시
전국다문화가족 실태조사	다문화 가정 대상	2009년 1회
영아모성사망조사	영아모성사망자 대상	약 3~4년 주기 1회

○ 건강보험 청구자료는 빅데이터 구축에 중요한 부분

- 건강보험 청구 자료는 개인정보 뿐만 아니라 의료이용 정보를 시간에 흐름에 따라 확인할 수 있어 빅데이터 구축에 중요한 부분으로 작용

1) 건강보험심사평가원 청구자료 관리를 위한 투자 및 유지·보수 비용 기준

○ 건강보험 청구자료를 분석하고 관리하는 데에는 막대한 유지보수 비용이 소요

- 매일 새롭게 생성되고 그 용량이 방대하여 막대한 유지보수 비용이 발생하고, 자료 유출시 개인정보를 침해할 위험성이 높음
- 건강보험 청구자료는 기관간의 이동이 효율적이지 못하며, 중복투자의 우려가 존재

2. 건강보험 심사평가원의 보유 정보 개방 다양화 방안

○ 사회적인으로 공공정보의 활용 요구 증가에 따른 심평원 보유 공공정보의 적극 개방 및 공개
- 정보공개를 통한 국민의 알권리 충족 및 공공데이터 활용의 활성화 지원
- 정부 3.0의 취지에 발맞추어 다양한 채널을 통한 심평원 보유 자료의 단계적 개방

○ 심평원 자료 제공 서비스의 다양화 요구 (다양한 채널, 데이터 연계)
- 영역별, 단계별 데이터 확대 구축
- 자료 제공 영역의 다양화를 통하여 연구자의 자료 활용성 극대화
- 자료 처리실: 원내에 연구자 전용 통계분석 자료 처리실 제공
- 원격 접속 서비스: 우리원 가상 PC를 접속하여 통계 분석할 수 있는 환경 제공
- 표본자료 확대 개방: 표본 자료 다양화를 통하여 연구자의 자료 접근 및 활용성 극대화

○ 수요자 중심의 정보 개방(접근성, 편의성 확대)
- 수요자의 편의성과 접근성을 고려하여 다양한 정보를 효율적으로 제공하는 구체적인 방안 마련 필요
- 비용 대비 효과적인 양적·질적 데이터 제공으로 새로운 가치를 창출 가능
- 보건의료 정보의 적극 개방 및 공개로 국민의 알권리 충족
- 공공데이터의 민간 활용 활성화

○ 보건의료분야의 연구 활성화를 위한 보건자료 제공 인프라에 대한 필요성 강조
- 심사·평가 혁신 및 보건의료 정책결정(또는 정책연구) 시 개방된 공공데이터 분석을 통하여 과학적 근거를 제공
- 임상학회와의 워킹 그룹 회의를 통하여 수요자 중심의 연구데이터 셋으로 확장
- 비용 효과적인 보건의료 데이터 제공으로 보건의료 분야 연구 활성화 지원

○ 현재 심평원의 정보 제공방법으로, 환자표본자료 제공 및 자료처리실 이용을 통해 자료를 제공
- 심평원 내에서 연구자 전용으로 통계분석을 위한 자료처리실 15석의 사용 공간을 보유하고 있고, 수요 증가 시 자료처리실 확대 계획
- 일반 연구자를 대상으로 제공 되는 환자표본자료의 다양화를 통해 연구자의 자료 접근 및 활용성을 극대화

○ 연구자들의 편의성을 위해 심평원의 가상 PC를 접속하여 통계분석할 수 있는 원격 접속 서비스를 2013년 11월 제공계획

<그림> 정보제공 다양화 방안

1) 자료처리실을 통한 맞춤형 자료 제공

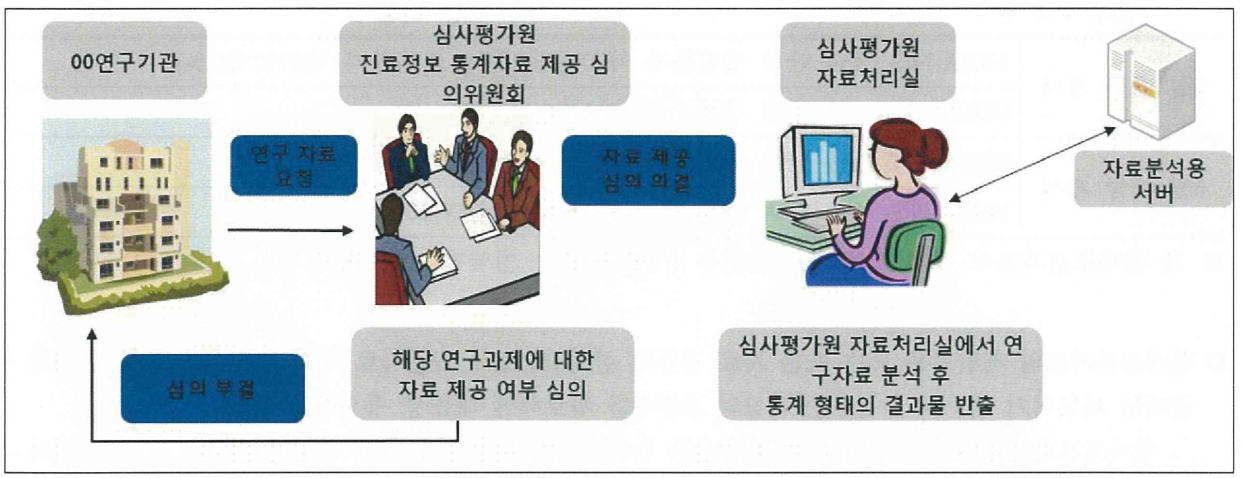
○ 상세하고 구체적인 자료 제공 가능(대규모 자료)
- 연구자에 연구 목적에 맞는 자료를 맞춤형으로 제작하여 직접 제공
- 개인정보 보호가 취약한 방대한 자료 제공 및 개인정보 연계가 필요한 코호트자료 제공
- 전국민을 대상으로 한 정보이므로 발생이 적은 희귀질환 분석 가능
- 자료 산출 가능 기간 : 최근 심사결정분(5개년도 기준)

○ 국책사업 및 공공연구 지원(외부기관 조사자료와 건강보험 청구자료 연계)
- 비영리 학술·연구 또는 공익기관에서 학술·연구 등 목적의 경우 제공(국가기관, 공공연구기관, 대학, 병원, 연구센터, 학회 등)
- 국가·행정기관에서 업무수행과 관련하여 요청하는 경우

○ 정보보호 조치 후 제공
- 성명, 요양기관 명칭 및 요양기관기호 등 개인정보와 개별 법인·단체의 정보는 식별 불가능한 형태로 변형하여 제공 ⇒ 주민등록번호 및 요양기관기호는 별도의 대체번호를 부여하여 중복성 배제
- 심사평가원 내 설치된 자료처리실에서 분석하도록 제공

○ 건강보험 청구명세서 기재사항 범위 내 자료 연계가능
- 환자의 포괄적 사전 동의 획득이나, 건강보험청구자료와 연계에 대한 동의 획득시 연계 가능
- 비급여, 전액본인부담 내역 산출 불가능

[그림] 자료처리실 운영 절차



1) 1년 단위 표본자료 개발 제공

○ 사전정보공개의 일환으로 일반연구자에게 중복된 통계자료 제공업무를 줄이고 자료를 효율적으로 사용할 수 있도록 사전에 가공·공표된 자료제공(표본자료)

○ "건강보험심사평가원"에서 "자료처리실"을 운영하고 있으나, 직접 내방하여 이용해야 하는 번거로움으로 접근성과 편의성 측면에서 한계가 존재하여 직접제공이 가능한 환자표본자료 개발

- 연간 약 10억 건 이상의 방대한 용량의 자료는 사용자의 저장용량, 처리속도 등 수용능력의 한계로 인하여 연구자로 하여금 시의적절한 자료 확보를 불가능하게 함
- 다양한 수요층에 대한 접근성과 편의성, 즉시성의 확보를 위한 대안의 하나로 우리나라의 건강보험 청구자료에 대한 입원환자표본자료(HIRA-NIS)를 2010년 12월 개발하여 2012년 6월 부터 제공 서비스 개시

○ 2013년 6월 현재 3가지 종류의 표본자료를 추가로 개발하여 제공 중

- 전체 환자표본자료를 통하여 중증질환과 같은 입원진료를 연구하기에는 대표성이 부족
- 표본자료의 종류를 전체환자표본(NPS), 입원환자표본(NIS), 노인환자표본(APS), 소아·청소년환자표본(PPS)으로 나누어 제공
- 환자특성과 대표성을 반영하도록 특정 계층에 대해 별도로 표본을 추출함으로써, 그 특정 계층만이 지니고 있는 환자의 특성과 대표성을 높여 연구에 대한 활용도를 제고

<표> 표본자료 종류 및 산출기준

표본자료 종류		산출 기준
4월 제공 개시	HIRA-NIS	1년 단위 입원환자 약70만명(13%), 외래환자 약40만명(1%)
	HIRA-NPS	1년 단위 전체 환자 약140만명(3%)
6월 제공 개시	HIRA-APS	1년 단위 65세 이상 환자 약100만명(20%)
	HIRA-PPS	1년 단위 20세 미만 환자 약110만명(10%)

※ 각 환자표본자료의 표본 한계치는 환자수 150만명 또는 영역별 20%이내를 기준

○ 환자표본자료의 제한점은 모든 표본 자료 공통의 한계점으로서 표본자료 내의 관측치는 확률에 의해 추출되는 자료이기 때문에 적정수준 이상의 표본수를 확보해야 대표성, 유의성을 보장
- 환자표본자료에서 특정 연령대의 희귀질환 발생빈도의 경우 표본추출 빈도가 너무 적어 대표성과 설명력이 떨어질 수 있음
- 표본자료의 설명력은 다빈도 상병 일수록 커지며, 상병의 발생 빈도가 떨어지면 감소하게 됨

○ 환자표본자료의 제한점을 해결하기 위한 방안으로 임상표본코호트자료 개발과 원격서비스 제공 방안을 개발
- 임상표본코호트는 1년 단위 환자표본자료를 확장하여 환자의 의료사용 이력을 시간의 흐름에 따라 확인 가능하여 활용 가능한 연구의 범위가 넓음
- 원격서비스의 경우 건강보험청구 자료의 대부분을 개인정보보호 처리 후 큰 제약 없이 사용 가능

2) 임상 표본 코호트 구축 계획

○ 임상학회와의 워킹그룹을 통하여 개인정보가 보호되는 수준에서 다년간의 표본추적코호트 자료 구축
- 1년 단위 환자표본자료로는 분석이 불가능한 희귀질병 등 환자들의 장기간 follow-up이 필요한 보다 전문적이고 임상적인 영역 대상
- 개인정보 보호를 위해 환자식별 대체키를 연구자마다 다르게 부여하여 자료의 이동 경로를 추적할 수 있고, 연구자들 간의 데이터 연계가 불가능하도록 개인정보 보호에 초점을 두고 개발 중

○ 2013년도 12월 까지 척추수술환자를 5년간 추적한 척추 수술 코호트 자료와 산모의 출산일을 기준으로 과거 1년부터 출산 이후 5년간 추적한 산부인과 코호트 자료 구축

<표> 임상 표본 코호트 종류

코호트 명칭	내 용
척추 수술 코호트	○ 2007년도 정형외과 및 신경외과 척추수술(진단)환자를 대상으로 5년간의 추적 코호트 자료 구축
산부인과 코호트	○ 2008년도의 출산한 산모 49만 명에 대한 표본 산모 추출 ○ 2008년도에 출산한 산모의 출산일을 기준으로 과거 1년간의 진료내역 추가(후향적 코호트) ○ 산모의 출산일 기준으로 산모와 신생아의 5년간의 진료내역 추가(전향적 코호트)

○ 산부인과 코호트를 구축하게 되면 산모가 가지고 있는 질환(고령산모, 임신성 당뇨 등)으로 인해 신생아 주별 사망률 등과 같은 신생아에게 미치는 영향을 파악 할 수 있으며, 희귀난치성 신생아가 가져오는 가계부담률 파악이 가능

- 장기간의 데이터를 구축하게 되면 소아질환이 성인으로까지의 질환 지속 파악 가능

○ 산부인과코호트 구축 시 산부인과·소아과에서 많은 연구가 활발히 진행 될 것으로 예상

○ 향후, 의약품분야와 내과분야 코호트 자료 등으로 다양화 할 계획

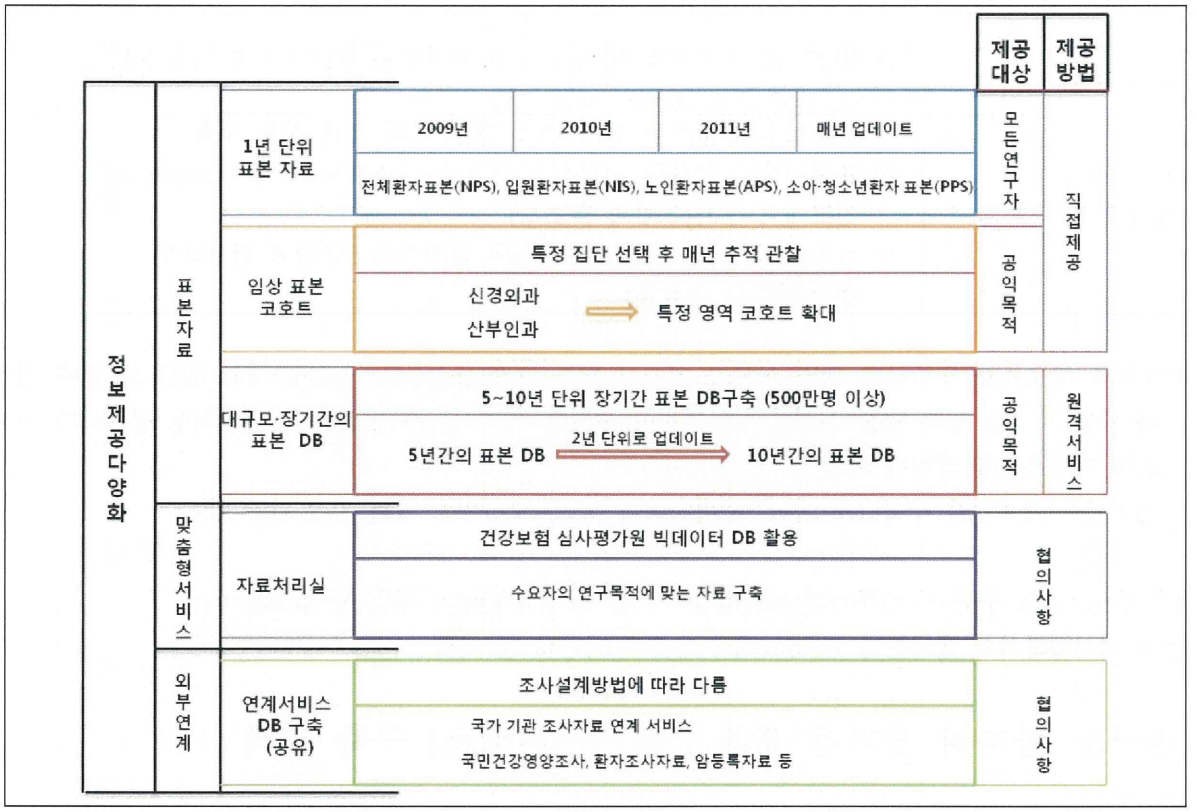
3) 대규모 환자의 장기간 표본 DB(원격서비스) 구축 계획

○ 모든 연구가 가능하고 표본자료에 대한 다양한 수요가 충족될 수 있는 광범위하고 포괄적인 영역의 연구용 DB구축 계획

○ 최소 500만 명의 대규모 환자의 장기간 표본 DB는 모든 연구가 가능하고 표본자료에 대한 다양한 수요가 충족될 수 있는 광범위하고 포괄적인 영역의 연구용 DB

- 5년간 건강보험으로 청구된 모든 환자들을 대상으로 하여 장기간의 표본 DB를 구축
- 한국통계학회, 보건정보통계학회 등을 통한 전문가 자문 예정
- 2년 단위로 새롭게 구축하여 10년 단위 표본 DB 를 구축하는 것을 목표
- 시간의 흐름에 따라 자료에 환자가 추가되고 누락되는 부분을 통계학적 편의 없이 반영가능

[그림] 정보 제공 다양화 체계도



4) 표본자료 구축을 위한 워킹그룹 운영

○ 임상분야별로 연구주제에 대한 수요 파악 및 데이터셋 구축 아이디어를 위한 목적으로 3개 임상분야(산·소아청소년과,외과,내과) 및 의약품 분야와 워킹그룹 결성
- 제공 자료에 대한 연구 이용 활성화 방안과 개선 사항 파악, 수요자 중심으로의 연구 데이터셋으로 확장 방안에 관한 논의 진행
- 이를 통해, 전문가의견을 반영한 보건의료 정보 연구 가이드라인을 설정
- 보건의료 데이터 수요 조사에 따라, 체계적으로 연구 자료를 연구자들에게 지원
- 진료정보 데이터셋 구축 연구자 참여 확대로 연구자료 제공

○ 관련 전문가와 의견 교류를 위해 지속적으로 학회를 대표할 수 있는 담당 자문위원 필요
※ 학회를 대표할 학회장 변동 시 표본자료와 관련한 내용을 알지 못하는 문제 발생
- 표본자료 개발 및 활용 확대 방안 등에 관한 학회 의견 취합 및 협력체계 유지

○ 워킹그룹을 통하여, 상병코드의 부정확성으로 인한 진단명 타당도에 대한 문제와 주요 질병의 기초통계 정보(발생률/유병률) 산출의 필요성 강조

○ 타 기관과의 데이터 연계, 개인(환자)단위에서 확대되어 환자가족 단위로의 구축 필요성 등이 분야별로 공통된 워킹그룹 주요 쟁점으로 파악 됨

○ 지속적인 워킹그룹 회의를 통해 이러한 문제에 대한 해결방안을 모색 예정

参考資料 3. 台湾における健康保険サンプリングデータ

2010 年, LHID2010 (2011 年発行 100 万人)

2005 年, LHID2005 (2007 年発行 100 万人)

2000 年, LHID2000 (2002 年発行 20 万人, 2009 年発行 80 万人)

■ 目的

全民健保データベースのデータ量は膨大であり、全てのデータを研究分析に提供するとなれば、相当大型なコンピュータシステムが必要なうえ、処理に時間が掛かり困難を極めるため、ミスや誤差が生じやすく、またプライバシーの保護の観点からも良いとは言えません。解決策の一つとして代表性を持つサンプリングデータを、ユーザーの研究分析に提供する方法があります。そのため、保険対象を基本サンプリング単位とするサンプリングデータベースを作る必要があり、歴年全ての受診データの収録、且つ追跡の継続によって出来上がったサンプリングデータベースは、研究学者の多様な研究に提供することが可能となります。

2002 年（中華民国 91 年）より当計画は 20 万人の健康保険サンプリングデータベース（LHID2000）を学界に提供し始め、研究者は健康保険サンプリングデータベースを取得すれば、それぞれの研究計画のニーズに応じ、縦断調査（longitudinal study）や断面調査（cross-sectional study）を行うことが可能となりました。当 LHID2000 は 2009 年には 80 万人分のサンプルデータ（第 5～20 組）を新たに追加提供し、計 100 万人のデータを発行しました。

LHID2000 は 2000 年以降に出生もしくは新規加入した保険対象を含まないことから、学者や専門家は 5 年ごとに新世代データにサンプリングを行うことを提案し、私たちは 2005、2010 年の健康保険データベース発行時に、再度サンプリングを行いました。健保データ研究が日増しに広まり、研究テーマもさらに広がってきたこともあり、研究者よりサンプリングデータ数増加の要求が提出されため、改めてサンプリングを行ったほか、サンプリング人数を 100 万人にまで拡大し、全ての受診データを取得して人数データを作成しました。100 万人の健康保険サンプリングデータベース（LHID2005、LHID2010）は統計的検定力（statistical power）を持つ前向き研究（Prospective study）や後ろ向き研究（Retrospective study）の可能性を大幅に引き上げる事が可能となりました。

2010 年健康保険サンプリングデータベース、LHID2010

1. データ内容： 2010 年健康保険データベース中の「2010 年保険加入者」よりランダムに抽出された 100 万人の各年度の受診データによって作成されたもので、4 万人毎の一年度の受診データを 1 単位として発行し、毎年更新されます。

2. サンプリング母集団

中央健康保険署が提供する2010年健康保険データベースは「身分証番号+誕生日+性別」をもって一人とし、27,378,403人のデータを、データの母体ファイルとすることが可能です。データ母体ファイルの中で、性別不詳者を取り除き、保険加入者23,251,700人のデータをサンプリングの母集団として選出します。2010年健康保険サンプリングデータベースに登録された最終保険加入日は2010年12月31日であるため、私たちは「2010保険加入者」を「2011年以前に出生し、2010年1月1日から2010年12月31日の一日でも保険に加入していた者」と定義し、年齢が非0－120歳の者を取り除きました。

3. サンプリング方法：

サンプリング母集団よりランダムに100万人のサンプルを抽出します。ランダム抽出方法はサンプリング母集団の23,251,700人に通し番号をつけ、乱数発生器（random number generator）を用いて最低でも100万個の乱数（random number，実質1,074,263個の乱数を取得）を発生させ、100万個の乱数と同じ通し番号を取得し、必要な保険対象サンプルをランダムに抽出し、身分証番号重複者（計24個）を取り除き、100万人のサンプルを取得するまで再抽出します。

乱数発生作業に関して、私たちはOracleのDBMS_RANDOMパッケージを採用し実行しています。DBMS_RANDOMパッケージは組み込み式の乱数発生器（Oracle’s internal random number generator）を用いて、8桁の整数の乱数を発生させることが可能です。私たちは1と23,251,700の間に110万個の乱数を発生させ、重複する乱数（計25,346個）を取り除き、合計1,074,263個の乱数を取得しました。

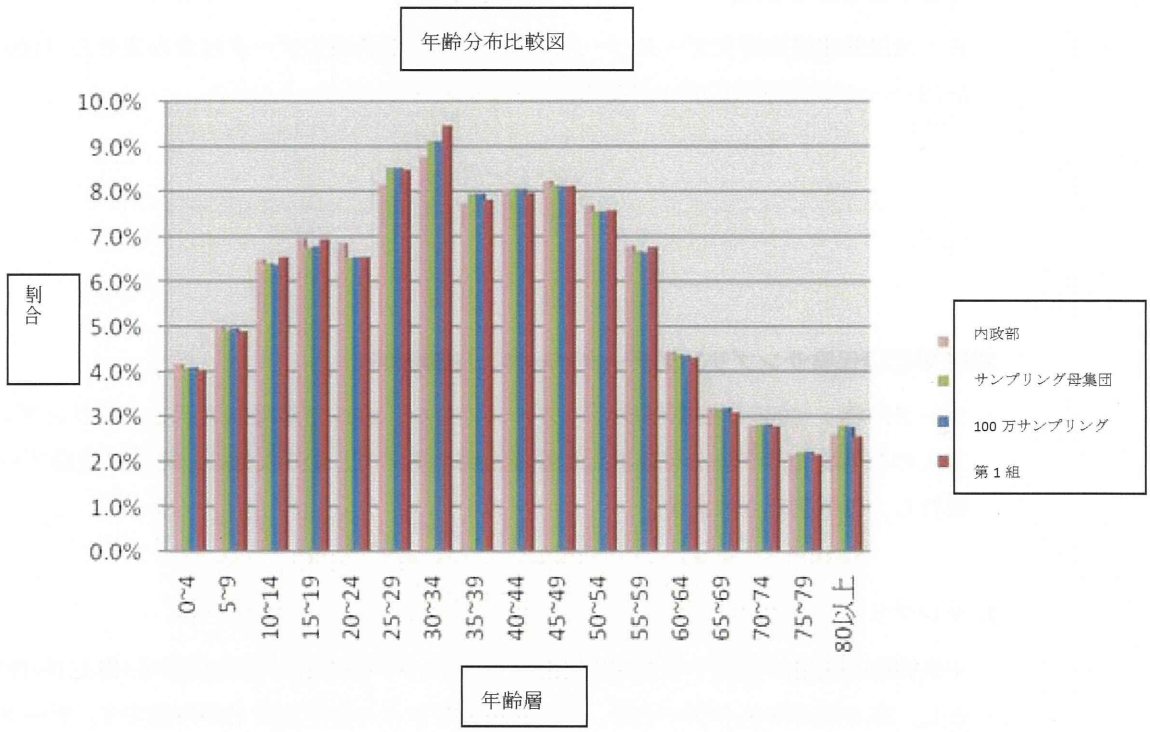
4. 健康保険サンプリングデータベースの構築

ランダムに抽出した100万人のサンプルを、身分証番号（暗号化済）を使って4万人1組の計25組に分け、健保データベースと連結させて、1996-2010年の該当する100万人の全民健保研究データベース内にある全ての受診データを取得すれば、100万人の健康保険サンプリングデータベースLHID2010が得られ、その後も毎年更新して、この100万人サンプル新年度の受診データを追加させます。

連結される受診データには：外来処方及び治療明細データ（CD）、外来処方医令明細データ（OO）、入院医療費用リスト明細データ（DD）、入院医療費用医令明細データ（DO）、特約薬局処方及び調剤明細データ（GD）、特約薬局処方調剤医令明細データ（GO）、そして原始健康保険データが含まれます。

5. 健康保険サンプリングデータベース代表性テスト：データ中の年齢、性別、毎年出生人数分布、及び平均保険金額を統計して、100万サンプルとサンプリング母集団間に差異の有無を比較し、同時に内政部の公布データ値と比較して、100万人サンプルのサンプリング母集団に対する代表性を分析します。健康保険サンプリングデータベースは4万人1組で使用するため、私たちもその内の1組である4万人のサンプルを選んで代表性の分析を行いました。分析方法は図、表及び統計的仮説検定を含み、詳細は以下説明の通りです。

5-1. 5 歳毎（例：0-4 歳）に一つの年齢層として組分けし、80 歳以上は一つの年齢層として、各年齢層人数が人口総数に占める割合を統計します。100 万人サンプルとそのサンプリング第 1 組 4 万人のデータを比較した結果、サンプリング母集団、内政部の公布データ、各年齢層人数が人口総数に占める割合の分布とほぼ一致しました。詳細は図 1 の通りです。



5-2. 性別分布- 100 万人サンプル統計による男女比は 97：100 で、サンプリング母集団男女比と同様でした。その内のサンプリング第 1 組 4 万人データの男女比は 96：100 で、比の値が極めて近く、カイ二乗検定による 100 万人サンプルとサンプリング母集団の男女比には差異がありませんでした（χ 2=0.067, df=1, p-value=0.796）。内政部の公告する 2010 年人口データの男女比は 101：100 で、サンプリング母集団と差異が見られました（注）。詳細は表 1 の通りです。

表 1、性別分布表

データ別	性別比率	2010 年人口数		
	男:女	総計	男	女
サンプリング母集団	97:100	23, 251, 700	11, 452, 740	11, 798, 960
100 万サンプリング	97:100	1, 000, 000	492, 423	507, 577
第 1 組 4 万人	96:100	40, 000	19, 627	20, 373
内政部人口統計	101:100	23, 162, 123	11, 635, 225	11, 526, 898