[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

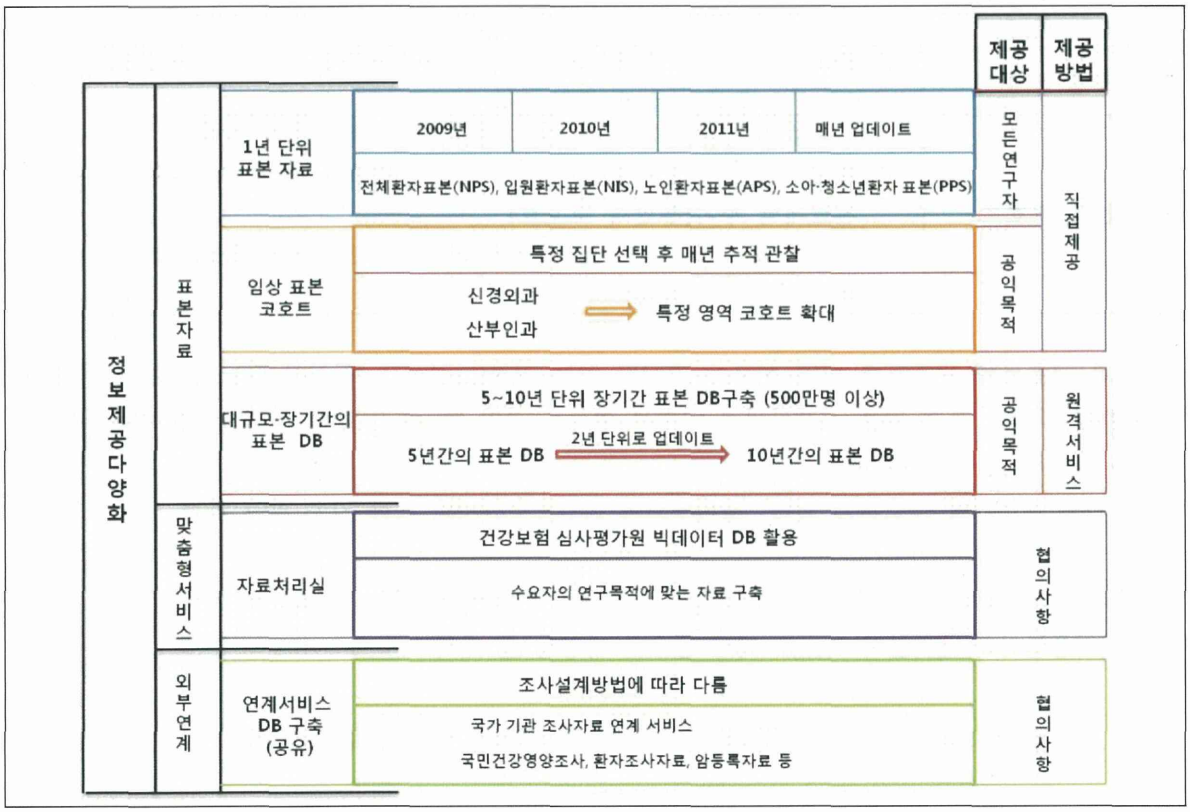
3) 대규모 환자의 장기간 표본 DB(원격서비스) 구축 계획

○ 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。

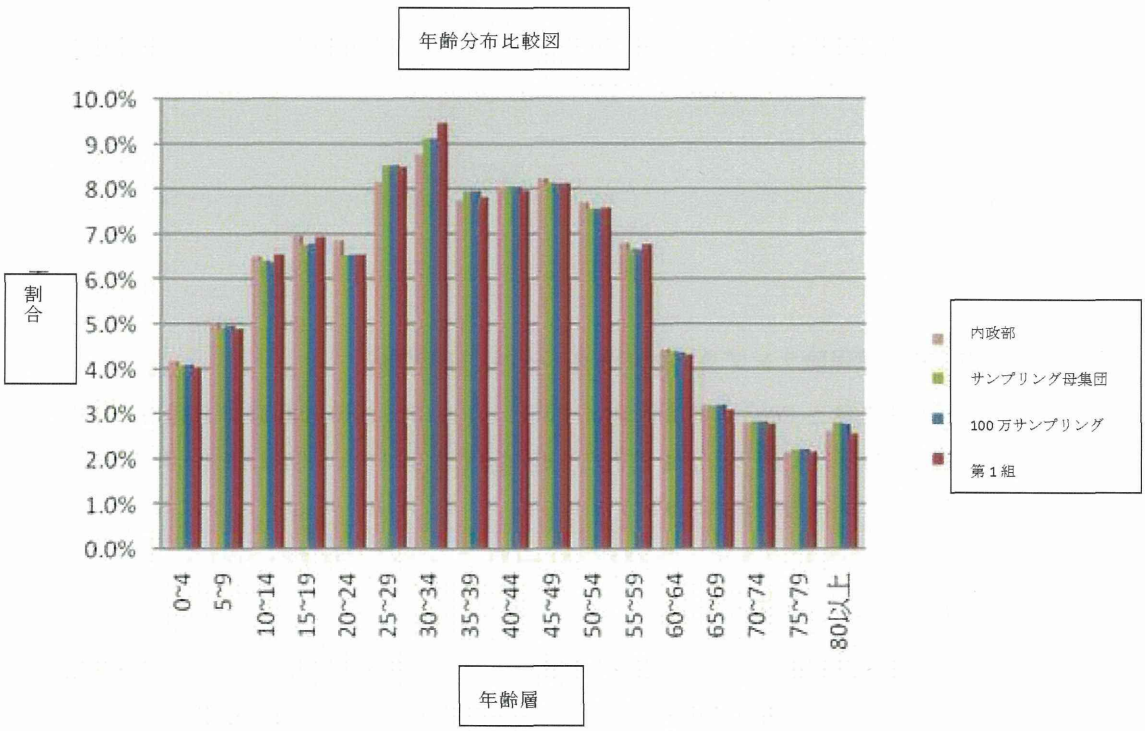
4. 健康保険サンプリングデータベースの構築

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（$\chi 2=0.067$, 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

5-3. 毎年出生人数分布-100 万人サンプル及びそのサンプリング第 1 組 4 万人データの毎年出生人数分布を統計し、カイ二乗により分析した場合、そのサンプリング母集団との差異は全て 5%以下で、目立った差異ではありませんでした。

5-4. 保険費用-100 万人サンプル及びそのサンプリング第 1 組 4 万人データの平均保険金額を統計し、normal test により分析した場合、そのサンプリング母集団との差異は全て 5%以下で、目立った差異ではありませんでした。

注：全民健康保険研究データベースは軍部門保険加入者のデータは含みません（100-9 版コードブック A-58 ページの健康保険データ説明参照）。

2005 年健康保険サンプリングデータベース、LHID2005

1. データ内容： 2005 年健康保険データベース中の「2005 年保険加入者」よりランダムに抽出された 100 万人の各年度受診データを取得し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

中央健康保険署の提供する 2005 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、25,678,998 人のデータを、データ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から、保険加入者 22,717,053 人のデータをサンプリングの母集団として選出します。2005 年健康保険データベースに登録された最終異動日は 2006 年 1 月 1 日であるため、私たちは「保険加入者」を「2006 年以前に出生し、2005 年 1 月 1 日から 2006 年 1 月 1 日の間に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 22,717,053 人に通し番号をつけて、乱数発生器（random number generator）を使って最低でも 100 万個の乱数（random number，実質 1,073,891 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 64 個）を取り除いて、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て実行しています。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）を提供し、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 22,717,053 の間に 110 万個の乱数を発生させ、重複する乱数（計 25677 個）を取り除き、合計 1,073,891 個の乱数を取得しました。

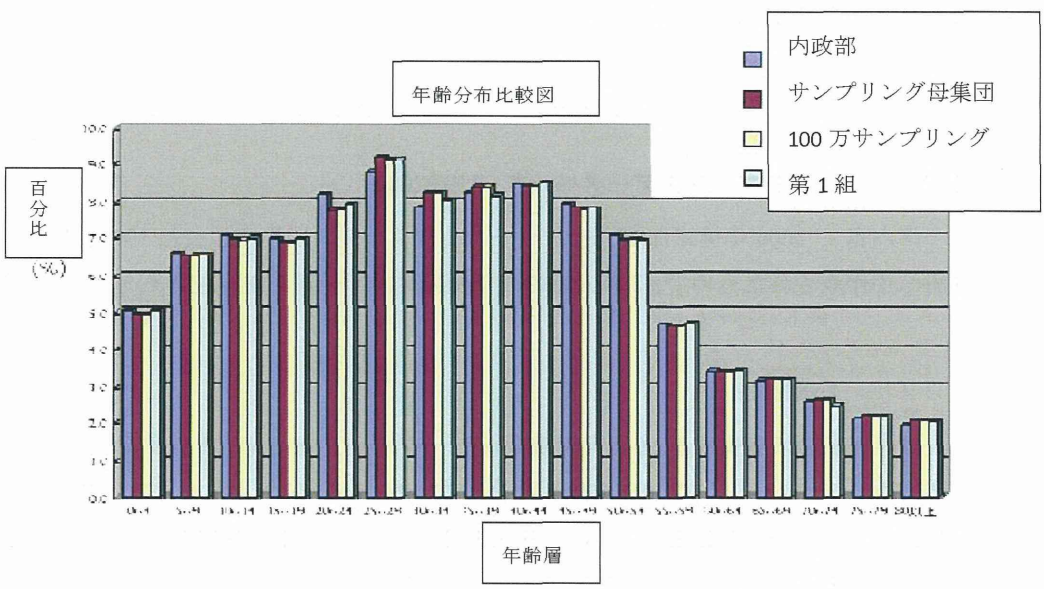
4. 健康保険サンプリングデータベースの構築

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、1996-2006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LHID2005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. 年齢分布-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、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98：100で、サンプリング母集団男女比と同様でした。その内の第1サンプリング組4万人データの男女比は99：100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は差異がありませんでした（χ2=0.008, df=1, p-value=0.931）。内政部の公告する2005年人口データの男女比は103：100で、サンプリング母集団と差異が見られました。詳細は表1の通りです。

表1、性別分布表

データ別	性別比	2005年人口数			
	男 ： 女	総計	男	女	性別不詳
サンプリング母集団	98 ： 100	22,717,053	11,262,470	11,454,582	1
100万サンプリング	98 ： 100	1,000,000	495,816	504,184	0
第1組4万人	99 ： 100	40,000	19,877	20,123	0
内政部人口統計	103 ： 100	22,770,383	11,562,440	11,207,943	0

5-3. 毎年出生人数分布-100万人サンプル及びそのサンプリング第1組4万人データの毎年出生人数分布を統計し、カイ二乗により分析した場合、そのサンプリング母集団との差異は全て5%以下で、目立った差異ではありませんでした。

5-4. 保険費用-100万人サンプル及びそのサンプリング第1組4万人データの平均保険金額を統計し、normal testにより分析した場合、そのサンプリング母集団との差異は全て5%以下で、目立った差異ではありませんでした。

2000年健康保険サンプリングデータベース、LHID2000

1. データ内容： 2000年健康保険データベースよりランダムに抽出された100万人の各年度受診データを取得して作成されたもので、5万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

中央健康保険署の提供する2000年健康保険データベースをサンプリング母集団としています。2000年健康保険データベースには中央健康保険署設立より2000年12月末までの合計約5,806万件の保険対象の累積性過去データが含まれており、保険対象の身分の変化や職場の異動に従って転出、転入の記録がされ、全て健康保険データベースの中に保存されます。重複する「身分証番号」を取り除くと、約2,372万口のデータがあります。

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。

3. サンプリング方法： 2000年健康保険データベースの保険対象を母集団として、ランダムに抽出を行います。

私たちは健康保険データベースの「身分証番号（暗号化済）+誕生日+性別」に基づいて一つの保険対象の身分を定義し、23,753,407人分を取得し、母集団としました。母集団の観察を通し、同じ身分証番号（以下ID）を異なる二人が共同保持する（即ちIDは同じでも、対応する「誕生日＋性別」が異なる）状況が存在しました。2000年母集団の中で、このようなID共用比率は約千分の一でした。詳細は添付表1を参照。

サンプリング方法は母集団の全ての保険対象に通し番号をつけ、これら異なる保険対象から更にランダムに抽出します。乱数生成関数（random number function）によって乱数（random number）を生成させ、重複しない乱数と健康保険データベース母集団保険対象の通し番号を連結させて、必要な保険対象サンプルをランダムに抽出します。

乱数生成作業に関して、私たちはSun Work Shop C 5.0の関数機能aを採用して乱数を生成しています。当関数はKnuth（1981）及びPark and Miller（1998）の方法（詳細は参考文献を参照）に基づき、linear congruential random number generationのテクニックを採用してbとなりました。当生成方法は数値が1と2,147,483,646の間の乱数を取り出すことが可能です。私たちはプログラムを運用して設定した取り得る値の範囲1と23,753,407の間より110万個の乱数値を取り出し、この110万個の乱数値の中から重複する乱数値（約2～3万個）を取り除き、最後に残りの値の中で20回に分け、毎回順に従って5万件の資料を一組として取り出し、合計で100万個の乱数を取得、母集団の中より100万個の乱数値と同じ通し番号を選出し、必要な保険対象サンプルをランダムに抽出しました。100万人とサンプリング各組サンプルのID共用比率を計算すると約千分の二でした。詳細は添付表2を参照。

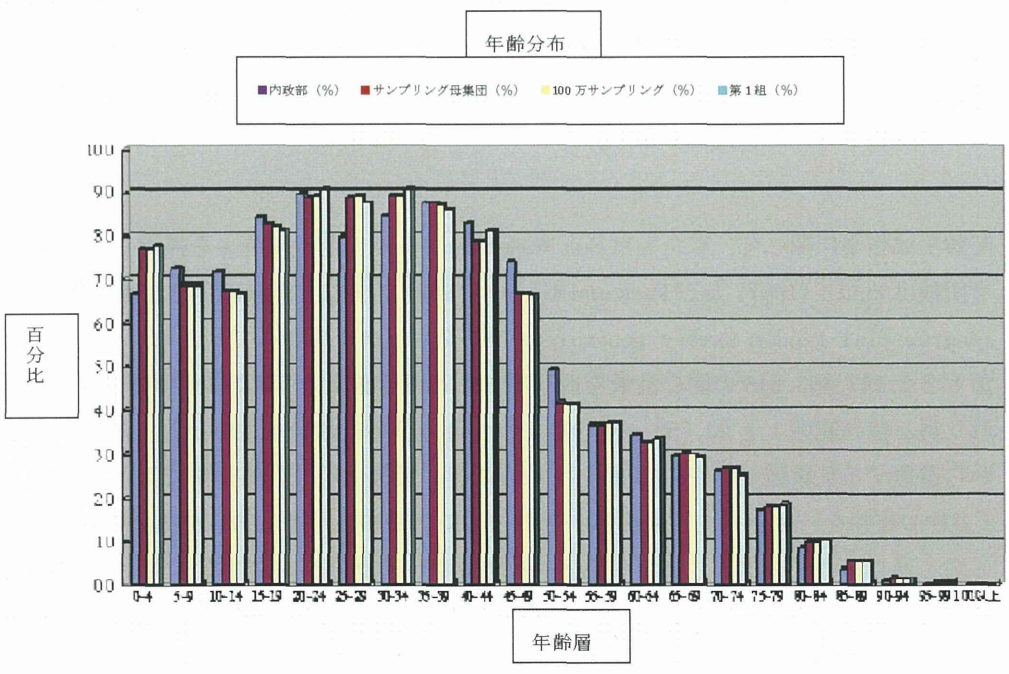
4. 健康保険サンプリングデータベースの構築：健康保険データベースの保険対象と健保データベースを連結。

ランダムに抽出した100万人の保険対象IDと健保データベースを連結させ、該当する100万人の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が得られます。保険対象の身分証番号を用い5万人1組の計20組に分け、健保データベースの歴年全ての受診データと連結させます。連結されるデータには：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）、そしてこれら保険対象の原始健康保険データが含まれ、毎年更新されます。その内第1～4組20万人のデータは2002年発行、残りの80万人（第

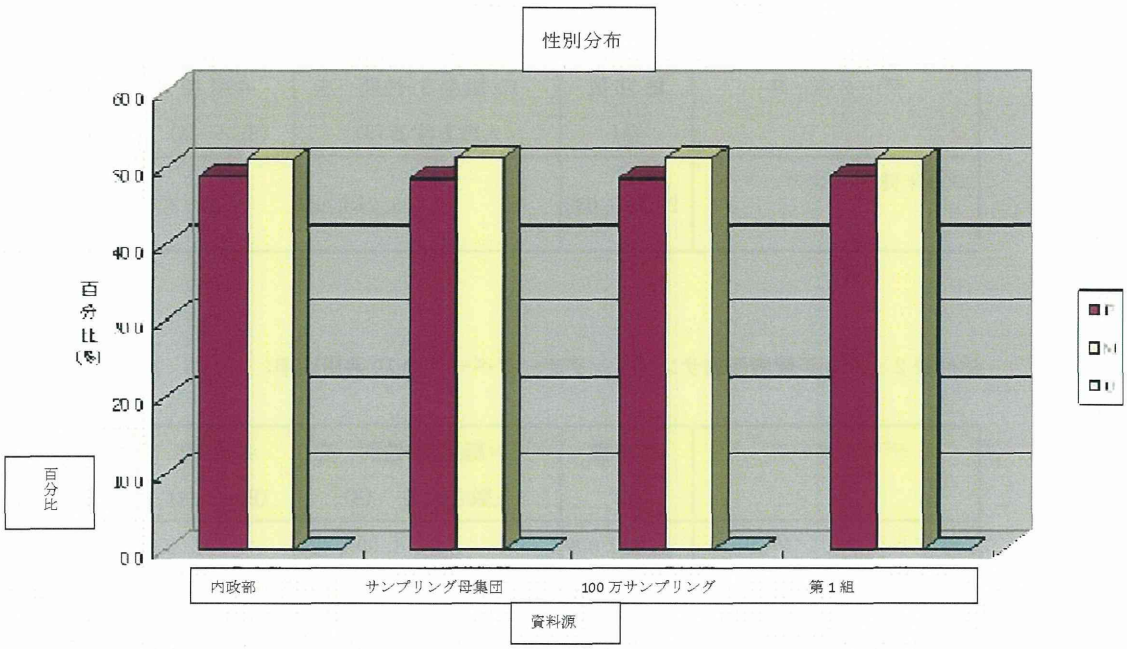
5～20 組）のデータは 2009 年発行のものです。

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、100 万サンプル ID 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 5 万人 1 組で使用するため、私たちもその内の 1 組である 5 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 年齢分布--5 歳毎（例：0-4 歳）に一つの年齢層として組分けし、100 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 5 万人のデータを比較し、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。



5-2 性別分布- 100 万人サンプル統計による男女比は 106：100 で、サンプリング母集団男女比と同様でした（以前 4 組 20 万人に行ったカイ二乗検定結果とも同様でした：$\chi 2=1.74$, df=1, p-value=0.187）。その第 1 サンプリング組 5 万人データの男女比は 105：100 で、比の値が極めて近くなりました。内政部の公告する 2000 年人口データの男女比は 105：100 で、第 1 組の比の値と一致し、サンプリング母集団、100 万人サンプルの性別比の値とも極めて近いものでした。男女人数が人口総数に占める割合を比較する場合、100 万人及びその第 1 組サンプル、サンプリング母集団、そして内政部公布データとも、全て女性は 49%、男性は 51%を占めています。



5-3 毎年出生人数分布--100 万人サンプル統計及びそのサンプリング第 1 組 5 万人データの毎年出生人数分布を、カイ二乗により分析した場合、そのサンプリング母集団との差異は全て 5%以下で、目立った差異ではありませんでした。

5-4. 保険費用--100 万人サンプル統計及びそのサンプリング第 1 組 5 万人データの平均保険金額を、 により分析した場合、そのサンプリング母集団との差異は全て 5%以下で、目立った差異ではありませんでした。

添付表 1：2000 年健康保険データベースの ID 共用比率:

データベース	総 ID 数 (A)	「ID+誕生日+性別」を人数とする(B)	共用 ID (B) - (A)	共用 ID (%) (B-A)/B
2000 年健康保険データベース	23, 725, 083	23, 753, 407	28, 324	0. 12%

添付表 2：2000 年健康保険サンプリングデータベースの ID 共用比率:

データベース	総 ID 数 (A)	「ID+誕生日+性別」を人数とする (B)	共用 ID (B) - (A)	共用 ID (%) (B-A)/B
100 万サンプリング ID	1, 000, 000	1, 002, 420	2, 420	0. 24%
第 1 組（2002 年発行）	50, 000	50, 109	109	0. 22%
第 2 組（2002 年発行）	50, 000	50, 143	143	0. 29%
第 3 組（2002 年発行）	50, 000	50, 122	122	0. 24%
第 4 組（2002 年発行）	50, 000	50, 147	147	0. 29%
第 5 組（2009 年発行）	50, 000	50, 108	108	0. 22%
第 6 組（2009 年発行）	50, 000	50, 147	147	0. 29%
第 7 組（2009 年発行）	50, 000	50, 099	99	0. 20%
第 8 組（2009 年発行）	50, 000	50, 119	119	0. 24%
第 9 組（2009 年発行）	50, 000	50, 111	111	0. 22%
第 10 組（2009 年発行）	50, 000	50, 103	103	0. 21%
第 11 組（2009 年発行）	50, 000	50, 116	116	0. 23%
第 12 組（2009 年発行）	50, 000	50, 121	121	0. 24%
第 13 組（2009 年発行）	50, 000	50, 127	127	0. 25%
第 14 組（2009 年発行）	50, 000	50, 141	141	0. 28%
第 15 組（2009 年発行）	50, 000	50, 127	127	0. 25%
第 16 組（2009 年発行）	50, 000	50, 112	112	0. 22%
第 17 組（2009 年発行）	50, 000	50, 113	113	0. 23%
第 18 組（2009 年発行）	50, 000	50, 106	106	0. 21%
第 19 組（2009 年発行）	50, 000	50, 134	134	0. 27%
第 20 組（2009 年発行）	50, 000	50, 115	115	0. 23%

参考資料 4. 米国 AHRQ のデータセンターで利用できるデータファイル

認可された研究プロジェクトにかかわる研究者（利用者）は、メリーランド州ロックビルの AHRQ データセンターにおいて、機密性を理由に一般に公開されていない機密データファイルにアクセスすることができる。その他有資格研究者についても、米国センサス・リサーチ・データセンター（RDC）のネットワーク（http://www.census.gov/ces/dataproducts/index.html）を通じて、機密データファイルにアクセスすることが可能である（RDC の研究提案プロセスと利用可能なデータセットについては、詳しくは AHRQ 国勢調査局の機密 MEPS データへのアクセスに関する取り決めを参照）。

機密データファイル

機密データファイルは、世帯要素・保険要素リンクファイル（1996～1999 年および 2001 年分が利用可能）である。多くの世帯に関しては、世帯構成員の雇用者を通じて加入できる健康保険が主たるデータソースとなっている。ただし、雇用者側が提供する健康保険制度に関する詳細のほとんどは、世帯回答者から入手することは難しい。こうしたデータの欠落を補う目的で、保険要素に関する調査を通じて、世帯要素の就業者を雇用する雇用者に問い合わせを行っている。雇用者への質問は、健康保険の提供内容、保険料、保険料に占める被雇用者の負担額、その他全体的な制度の詳細についてである。こうした情報は、データファイルを介して、就業者に関する世帯調査で収集されたデータとリンクされ、就業者の保険オプションに関するより完全な全体像を把握することができる。調査の無回答項目は、リンクデータを使用しての全国的な推定の算出の妨げとなるため、ファイルに含まれた個人サンプル以上の結果を導き出すことはできない。1996～1999 年の調査ファイルに関する詳細資料については、MEPS-IC の調査票とコードブックに対する横断的な変数も含めて、ダウンロードすることができる。ただし今後数年間は、こうしたデータを収集する計画はない。

ナーシングホーム要素ファイルは（1996 年のみ利用可能）は、ナーシングホームと、1996 年（暦年）のいずれかの時点でナーシングホームに居住、または居住を許可された者が調査対象となっている。調査においては、ナーシングホーム居住者の人口統計的特性、居住歴、機能的健康状態、サービスの利用度、処方薬の利用度、医療費に関する情報が収集される。またナーシングホームの管理者と指名されたスタッフからも、施設規模、所有権、認可状況、提供サービス、収益と経費、その他施設の特徴に関する情報を収集している。さらにコミュニティ・アンケートを通じて、近親者またはコミュニティ内のその他の情報所有者から、収入、資産、家族関係に関するデータ、ならびに抽出されたナーシングホーム居住者の介護に関する情報が収集される。調査ファイルに関する詳細資料については、ナーシングホーム調査票とコードブックに対するリンクも含めて、ダウンロードすることができる。ただし、今後数年間は、こうしたデータを収集する計画はない。

医療提供者要素ファイルは、病院、医師、在宅医療提供者、薬局から詳細な料金および支払いデータを収集し、世帯要素の回答者から得た医療費に関する情報を補足することにある。収集された課金データには、入頭払いであったかどうか、診療所および外来診療部門での各処置（CPT4）に対する課金、支払い源ごとの額等に関するデータが含まれる。来院および入院に関する診断コード、処方薬に関するNDCコードについても収集されている。医療提供者要素については、これ単独で全国的な推定を算出することを意図したものではないが、支払い、処置コード、診断コードに関する詳細情報は、世帯要素データだけでは不可能な各種分析を補足するものである。

エリア・リソース・ファイル（ARF）は、郡に特化した医療資源データファイルで、政策担当者、研究者、ならびに国の医療提供制度と米国の健康状態および医療に影響を及ぼし得る要因に関心を持つその他専門家による利用を意図したものである。データベースとして、米国の各郡に関する7,000以上の変数を収録している。その収録情報は、医療施設、医療職、資源不足に関する程度、健康状態、経済活動、医療研修プログラム、社会経済的および環境的特性に関するものである。ARFのデータはこれまでも必要に応じて、AHRQデータセンターで使用されているMEPSファイルとマージされてきた。現在もマージは可能であるが、ARFファイルのコピーを保健資源事業局（HRSA）から入手した旨を証明する必要がある。

2年間の2パネル（Two-Year, Two-Panel）ファイル。このファイル（2YP）は、年間統合ファイル、補足変数ファイル、縦加重ファイルのデータをプールする個人レベルの調査ファイルである。パネルごとに1つのファイルがあり、各ファイルには1年次と2年次のパネルに基づいた個人および変数ごとの記録が収録されている。当該の年間PUF（HC-012、HC-020、HC-028、HC-038、HC-039）のいずれかに記録を有する、パネル1から4の全個人が母集団となる。年間変数については、プールされたパネルに基づく調査を簡易化する目的で、各変数名が全パネルで一貫するように名称が修正されている。パネルをプールすることにより、人種的・民族的マイノリティー、障害者、農村居住者、特定の健康状態の患者、利用頻度の低いサービスといったより小さな母集団に関する分析効果を高めるのに有用であると思われる。

MEPS（医療費支出パネル調査）一般利用データファイルは、MEPSウェブサイトからダウンロードできるすべてのデータファイルを、AHRQデータセンターでも利用することができる。

<u>機密ファイルの変数</u>

ICD-9コード：病名

産業・職業コード：就業者が従事する産業および職業をより高い専門性をもって特定することができる。

州・郡FIPSコード：このコードを使用して、エリア・リソース・ファイルや州・郡レベ

ルのデータをMEPSデータにマージすることができる。

国勢調査区・区画グループコード：このコードを使用して、米国国勢調査局や調査区・区画グループレベルのデータをMEPSデータにマージすることができる。

機密用途データ要素：調査票の直接の特定が不可能なデータに関するデータ要素ではあるが、現在のところ編集も公開もされていない（資産情報と帰属NDCコードに関する）。

連邦・州限界税率：全米経済研究所（NBER）のTAXSIMパッケージに基づく税シミュレーションが、1996年～2002年にかけて（HC［世帯要素］の通年の母集団を対象に）実施された。連邦、州およびFICA（連邦保険拠出法）税に関して税額と限界税率が算出されている。財産税、売上税と市および郡の所得税については、算出されていない。

申請手順と書式

データ利用を希望する研究者は、研究企画書を含む、申請書を提出しなければならない。申請書は委員会が審査する。申請書式については、HTMLフォーマットとPDFフォーマットが使用できる。AHRQ データセンターへの申請書は継続的に受理され、月単位で審査され承認されている。研究者には、まずプロジェクト案の実現可能性について、AHRQ データセンターのスタッフと協議するために、申請パッケージ一式を、米国医療研究・品質調査機構の調査運用部門／CFACTのAHRQデータセンター・コーディネーター宛てに郵送することが推奨されている。AHRQ データセンターの責任者は、内容に不備がないかどうかを検討し、必要に応じて研究者から説明を受けた上で、AHRQ の資金調達・アクセス・コスト動向研究センター（Center for Financing, Access and Cost Trends）のディレクターに向けて承認の是非に関する提言を行う。また、適切な申請パッケージの作成に向けて、提案内容について研究者およびデータセンター長と協議することもある。プロジェクトが承認された場合は、必要に応じて、要求のサービスの費用見積について、申請者である研究者と検証を行う。

以下の書類が作成され、プロジェクトごとに、AHRQ データセンター・コーディネーターにより詳細が明記される。1)AHRQデータセンター取り決め。研究者およびデータセンターが実施するプロジェクト、データおよびサービスの範囲、両当事者の義務ならびに経費について明記される。2)プログラミングサポートに関するタスクオーダーおよび請求に関する取り決め、3)必要に応じて、機密保持誓約書。

申請料金

承認されたデータセンター・プロジェクトに関しては、技術サポート、簡単なファイル構築、最長4時間のプログラミングサポートを対象として、300ドルのユーザー料金が請求される。学位論文、その他学位要件に取り組むフルタイムの大学院生、ならびに連邦政府機関については、この料金は免除される。国勢調査局のリモート・データセンター（RDC）の利用についても、データセンター料金は免除されるが、申請者にはRDCでの作業に際し

て、国勢調査局の求める追加料金を支払う義務がある。

提案審査

AHRQ データセンターの責任者が、提案ごとに審査の調整にあたる。審査に際しては、以下の点が考慮される。プロジェクトに関する既存データの実現可能性、つまり、利用可能な情報に基づき、調査を実施できるかどうか。当初から、サンプルが意図する分析をサポートし得ないことが明らかとなる場合もある。たとえば、MEPS は小さな国または一部の条件に関する推定を裏づけ得るものではない。

機密情報の開示の危険性、つまり、あらゆる回答者（個人、雇用者、保険会社、病院、医師）に保証する機密性を脅かすことなく、分析を実施できるかどうか。

AHRQ データセンターにおけるプロジェクト支援リソースの利用可能性。現時点では、現場において提供できる技術サポートに限りがある。したがって、データセンターを来訪するまでに、最大限の努力を払って MEPS データに習熟しておくことが推奨される。

プロジェクト案が授権立法（authorizing legislation）に規定された AHRQ の使命に適ったものであるかどうか。

申請が承認されたからといって、研究案の実質性、方法論、理論、方針としての妥当性、またはその科学的メリットを AHRQ が認めたのではない点に注意が必要である。承認は単に、当該研究が AHRQ の使命と広く一致した実行可能なものと判断されていること、および、回答者に保障され、または法により要求される機密性とプライバシーの保護の観点から、データが適切に使用されることを意味するに過ぎない。

データファイルへのアクセス

研究者は、必ずメリーランド州ロックビルの AHRQ データセンター内に所在する施設においてコンピュータを起動させなければならない。現時点においては、リモートアクセスは利用できない。

AHRQ データセンターをすでに訪れ、初期プログラミングを実行済みである研究者に対しては、SAS または Stata の更新プログラムを送付し実行できる、限定的なサービスを利用できる。これらのプログラムは研究者の代わりにデータセンター・コーディネーターが実行する。このサービスは、プログラム開発を目的としていないため、プログラミングサポートは対象に含まれない。プログラムのデバッグについては研究者がその責任を負うことになる。このサービスの目的は、統計モデルの変数の変更など、既存プログラムの簡単な修正をサポートすることにある。単独のプロジェクトに関する要望が多数あれば、別途料金による追加サービスの提供の交渉に応じる。

AHRQ データセンターでは、研究者が自身のデータを提供し、AHRQ データとマージすることを認めている。ユーザーからの提供データは、ユーザーが収集・保有する独自データから構成されている可能性がある。ユーザーには、マージを提案するデータに関する完

全な資料をデータセンター・スタッフに提供することが求められる。この資料には、記述変数と変数値ラベルについて記載することが推奨される。ユーザーには、データセンター・スタッフと情報をやり取りし、データがマージ可能なものであり、そのフォーマットに一貫性があることを保証する責任がある。地理的地域の特性をマージする場合に関しては、調査の抽出枠の詳細（つまり、対象となった、あるいは対象外となった郡）がユーザーに判明しないような形で実行することが求められる。データマージについては、ユーザーの来訪に先立ち、データセンターまたはその請負業者が実施する。データセットのリンクに使用するファイルおよび情報については、ユーザーは入手できない。方針として、AHRQはデータセンターで使用するすべてのデータ（研究者からの提供データを含む）の目録を作成し、その他研究者にも利用できるものとするが、既存データの利用取り決めによって要求される場合、特定データへのアクセスを制限する取り決めを策定することもできる。
このページの一番上に戻る

データセンターから持ち出しできるアウトプット

データセンターから持ち出されるすべてのデータは、開示承認を得なければならない。
研究者がデータセンターの責任者に検討を求め、許可された場合は、許可済みのテーブルに加えて、次のデータをデータセンターから外部メディアにダウンロードすることができる：プログラム、ワープロ文書、許可済みプリントアウトの電子版。

データセンターから持ち出しできないアウトプット

直接または推定の如何にかかわらず、回答者または限定された地理的地域が特定される可能性のあるアウトプットは、データセンターから持ち出すことはできない。
地理的地域に基づく従属変数モデルについても、変数値を暗号化する場合を除いては、データセンターから持ち出すことはできない。
データ主体の特定に使用できる、抽出単位に関する識別子も持ち出すことはできない。
一般に、直接または推定の如何にかかわらず、データファイルの一般利用に際して明示されない識別子については、データセンターから持ち出すことはできない。
抽出事例のプリントアウトについては、データセンターから持ち出すことはできない。
（SAS の）Proc Means を使用する場合、最小値と最大値が同じセル、またはセルサイズが100 未満のセルを持ち出すことはできない。Proc Univariate または Cross-Frequency を作成する場合、最小値と最大値が同じテーブル、最小値と最大値に 1 つの観測値しかないテーブル、または行と列サイズが 100 未満のテーブルを持ち出すことはできない。

AHRQ データセンターで利用できないデータファイル

認可プロジェクトにおいてデータファイルが要請される場合を除き、研究者はデータファイルにアクセスできない。研究者は氏名や住所といった直接識別子を含むファイルにアク

セスできない。研究者は、暗号化する場合を除いては、地理データ（州、郡、国勢調査区）にアクセスできない。地理的地域の特性へのアクセスが許可されるかどうかは、認可研究プロジェクトに応じて異なる。また場所に関するダミーコードを含むファイルへのアクセスは許可されるが、地名とコードの関連付けにつながるデコードについては許可されない。要望があれば、ファイル全体を（メディケイドの手厚さが高い／中程度／低い州に居住している等の）複数のカテゴリーにプリコードできる。

（注）以上の情報は、AHRQ のデータセンタに関するホームページ情報を翻訳したものである。http://meps.ahrq.gov/mepsweb/data_stats/onsite_datacenter.jsp