

サンプリングデータセットの概略について

サンプリングデータセットの**具体的な説明**につきましては第8回、第9回有識者会議資料
(<http://www.mhlw.go.jp/stf/shingi/2r98520000022d61-att/2r98520000022def.pdf>)
(<http://www.mhlw.go.jp/stf/shingi/2r98520000024d12-att/2r98520000024d5x.pdf>)、
サンプリングデータセットの**詳細な仕様およびレコードフォーマットの説明**につきましては下記資料
(<http://www.mhlw.go.jp/stf/shingi/2r9852000002911e-att/2r9852000002912v.pdf>)
(<http://www.mhlw.go.jp/stf/shingi/2r98520000029hp2-att/2r98520000029hso.pdf>)
をご参照下さい。なお、第一回提供時以降、データの仕様の変更はありません。

主な特徴

- 1カ月分のレセプトデータに対し抽出・匿名化処理等を行ったデータです。
- 研究目的の範囲内で、探索的研究を行うことができます。
- 各データセットとも、DVD1枚の容量におさまる情報量です。
- すでに抽出処理が済んでいるので、手続き終了から提供までの間、抽出処理に要していた時間が短縮されます。

対象レセプト

- **平成23年分～平成25年分**の「医科入院」「医科入院外」「DPC」「調剤」のデータセットがあります。(計12種類)
- 各年のデータセットとも、**10月診療分**、かつ11月・12月・翌年1月審査分のレセプトが対象となります。
- 「医科入院外」には、同年10月診療分、ならびに11月診療分の**調剤レセプトが紐付けられています。**

抽出方法

- 「医科入院」「DPC」は、レセプト全体の**10%を抽出**しています。
- 「医科入院外」「調剤」は、レセプト全体の**1%を抽出**しています。
- 性別、および5歳刻みの年齢層において、母集団と構成比率を変えないように抽出されています。

匿名化処理

- 「傷病名コード」「医科診療行為コード」「医薬品コード」「診断群分類」において、**出現頻度の低いものは匿名化処理**を行っています。

空欄化 もしくは削除 した主な情報

- 「医科入院」においては700,000点以上、「医科入院外」「調剤」においては50,000点以上のレセプトは削除しています。
- 「都道府県」に関する情報、「保険者」に関する情報は削除しています。
- 「DPC」レセプトにおいては、医療機関別係数は空欄化、その係数の影響を受ける情報は削除しています。

参考：第9回有識者会議資料より

サンプリングデータセット：対象・抽出方法

➤ 対象となるレセプト

- **平成23年10月診療分、単月**のレセプト情報とする。
 - 年末年始や年度変わり、学休期間、ゴールデンウィーク等祝日の多い月を回避し、10月とした。
- 「医科入院」、「DPC」、「調剤」は、それぞれ単月のみの情報とする。「医科入院外」は、月をまたいで処方薬を入手する事例があるため、**同一月および翌月の調剤レセプトを紐付ける。**
 - あらかじめ所定の割合で抽出を行ったうえで、ハッシュ値を用いて紐付けを行う。
 - ハッシュ値による紐付けのため、100%捕捉することはできない。

➤ 抽出方法

- レセプト種類毎に、次のように抽出を行う。(レセプト数、容量等はおおむねの推計)

ひと月あたりの集計(概算)		全レセプト数	抽出率	抽出後レセプト数	抽出後データ容量
入院	医科入院	140万	10%	14万	1.2GB
	DPC	92万		9万	1.6GB
入院外	調剤	4,851万	1%	49万	0.8GB
	医科入院外(+調剤)	7,756万		78万	1.8GB(+1.6GB)

- なお、**性別、5才刻み年齢別に母集団と構成比率が変化しないよう**、抽出を行う。

平成24年分、平成25年分のサンプリングデータセットもご用意しております。対象・抽出方法については平成23年分のサンプリングデータセットと同等です。