資料2

レセプト情報等のデータ構造について

平成25年1月17日

厚生労働省保険局総務課
保険システム高度化推進室

目次

目次

コンビニにおける商品管理

コード	商品	単価	販売個数	売上高
7654	缶ビール	600	1	600
2345	オレンジジュース	100	1	100
1104	ポテトチップス	150	2	300
4308	おでん（がんもどき）	80	1	80
4309	おでん（はんぺん）	70	1	70
4312	おでん（昆布）	50	1	50
	合計			1200

個々の商品毎に、コード、単価、個数等の情報が入力される。

電子レセプトのデータ構造にまつわる問題

現行の電子レセプトでは

【レセプト上の記載】

再診料
地域医療貢献加算
明細書発行体制加算　　73点　　×　２

【電子レセプト上の記録】

…,112007410（再診料）,,
…,112015670（地域医療貢献加算）,,
…,112015770（明細書発行体制加算）,73,2

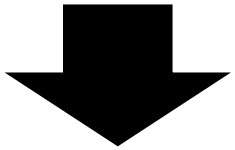
CSV(Comma Separated Value)形式で記録されている。

電子レセプトのデータ構造にまつわる問題

現行の電子レセプトでは

【レセプト上の記載】

再診料
地域医療貢献加算
明細書発行体制加算　　73点　　×　２



…,112007410,,
…,112015670,,
…,112015770,73,2

CSV(Comma Separated Value)形式で記録されている。

電子レセプトのデータ構造にまつわる問題

現行の電子レセプトでは

【レセプト上の記載】

再診料
地域医療貢献加算
明細書発行体制加算　　73点　　×　２

【電子レセプト上の記録】

…,112007410（再診料）,,
…,112015670（地域医療貢献加算）,,
…,112015770（明細書発行体制加算）,73,2

CSV(Comma Separated Value)形式で記録されている。

電子レセプトのデータ構造にまつわる問題

このCSVデータをエクセルで開くと

【電子レセプト上の記録】

…,112007410（再診料）,,
…,112015670（地域医療貢献加算）,,
…,112015770（明細書発行体制加算）,73,2

【エクセルファイル】

コード（診療行為）	点数	回数
112007410（再診料）		
112015670（地域医療貢献加算）		
112015770（明細書発行体制加算）	73	2

→ 再診料等の点数・算定回数は空欄、明細書発行体制加算（1点）は、73点として集計されてしまう。

実際の点数および回数はこのようになっている

コード（診療行為）	点数	回数
112007410（再診料）	69	2
112015670（地域医療貢献加算）	3	2
112015770（明細書発行体制加算）	1	2

電子レセプトのデータ構造にまつわる問題

薬剤データでも同様の問題がみられる

【レセプト上の記載】

ムコダイン錠500mg　　3錠
セルベックスカプセル50mg　　3カプセル
ロキソニン錠　60mg　　3錠　　１５×７

【エクセルファイル】

コード（薬剤品名）	数量	点数	回数
610407447（ムコダイン錠500mg）	3		
612320346（セルベックスカプセル50mg）	3		
620098801（ロキソニン錠　60mg）	3	15	7

実際の点数および回数はこのようになっている

コード（薬剤品名）	数量	点数	回数
610407447（ムコダイン錠500mg）	3	5	7
612320346（セルベックスカプセル50mg）	3	6	7
620098801（ロキソニン錠　60m）	3	4	7

- 本事例では各薬剤の点数の合計と当初ファイルでの合計点数が一致しているが、多剤投薬の場合はあらかじめ薬価を合計したうえで点数へと換算するため、実際には端数処理の影響によっては、個々の薬剤の薬価を換算した点数の合計と最下行の合計点数とが一致する上記のような場合ばかりとはいえない。

電子レセプトのデータ構造にまつわる問題

複数患者のレセプトから、セルベックス処方のデータ（該当する行）を抽出すると

患者ID	コード	（薬剤品名）	数量	点数	回数
1	612320346	セルベックスカプセル50mg	3		
2	612320346	セルベックスカプセル50mg	3	15	3
3	612320346	セルベックスカプセル50mg	3	18	3
4	612320346	セルベックスカプセル50mg	6		
5	612320346	セルベックスカプセル50mg	6	30	5
6	612320346	セルベックスカプセル50mg	6	45	6

患者の処方回数情報が消滅している場合が多く、セルベックスの総処方錠数（Σ（数量×回数））は、計算不能。

電子レセプトのデータ構造にまつわる問題

その他の課題

- 複数の傷病名コードが存在するため、医療資源が最も投入された傷病名一つを選択することは困難。（主傷病も一つとは限らない）

電子レセプトのデータ構造にまつわる問題

その他の課題

- 複数の傷病名コードが存在するため、医療資源が最も投入された傷病名一つを選択することは困難。（主傷病も一つとは限らない）

- 傷病名、医科診療行為、医薬品等各種マスターが頻繁に更新されているため、一定期間以上にわたってレセプト情報同士の紐付けを行う場合、必要とする情報にどのコードが関連するのかについて、レセプトが作成、発行された時期に応じて慎重な確認が求められる。

電子レセプトのデータ構造にまつわる問題

その他の課題

- 複数の傷病名コードが存在するため、医療資源が最も投入された傷病名一つを選択することは困難。（主傷病も一つとは限らない）

- 傷病名、医科診療行為、医薬品等各種マスターが頻繁に更新されているため、一定期間以上にわたってレセプト情報同士の紐付けを行う場合、必要とする情報にどのコードが関連するのかについて、レセプトが作成、発行された時期に応じて慎重な確認が求められる。

- 本データベースでは、テキスト文字で入力された情報は削除されている。

  例）

  - 未コード化傷病名（コード番号：0000999、コード自体は残っている）
  - フリーコメント（コード番号：810000001、コード自体は残っている）

目次

ハッシュ関数の採用

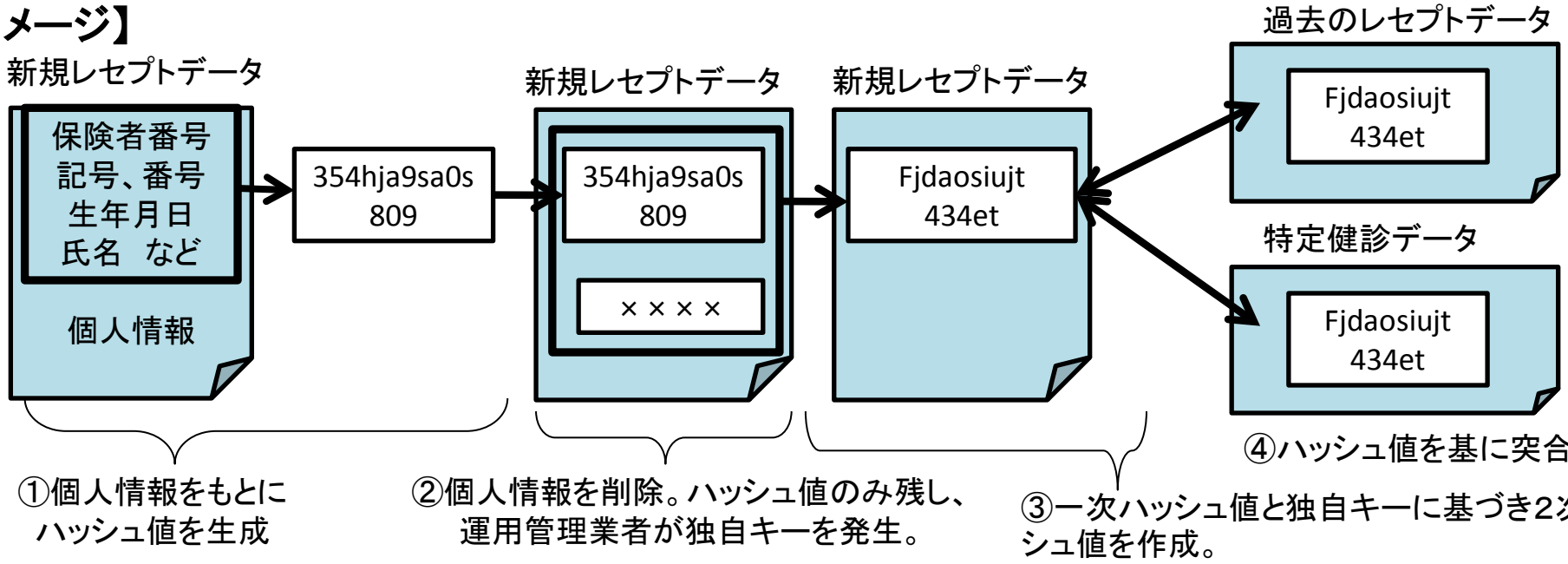
以下の特徴を持つ「ハッシュ関数」を用いることで、個人の直接的な識別情報を削除(「匿名化」)した上で、同一人物の情報であることを識別できるようにし、データベースへ保管している。

【ハッシュ関数の特徴】

①与えられたデータから固定長の疑似乱数(ハッシュ値)を生成する。
②異なるデータから同じハッシュ値を生成することは極めて困難。
③生成された値(ハッシュ値)からは、元データを再現することは出来ない。

※ 個人情報(氏名、生年月日等)を基にしてハッシュ値を生成し、それをIDとして用いることで個人情報を削除したレセプト情報等について、同一人物の情報として特定することが可能。

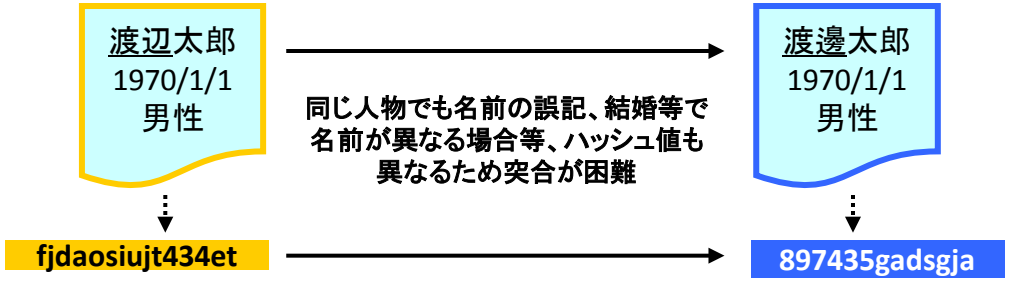
【イメージ】



ハッシュ関数についての留意点

ハッシュ関数自体、及びそのインプットとなる個人情報の管理状況から、同一人物の情報の紐付けを完全には行うことが困難なため、分析目的に応じた考慮（不良データの許容度、修正方針等）が必要。

①個人情報（保険者番号、記号番号、生年月日、性別、氏名）をもとにハッシュ値を生成するため、これらの情報に変化があった場合、突合が困難

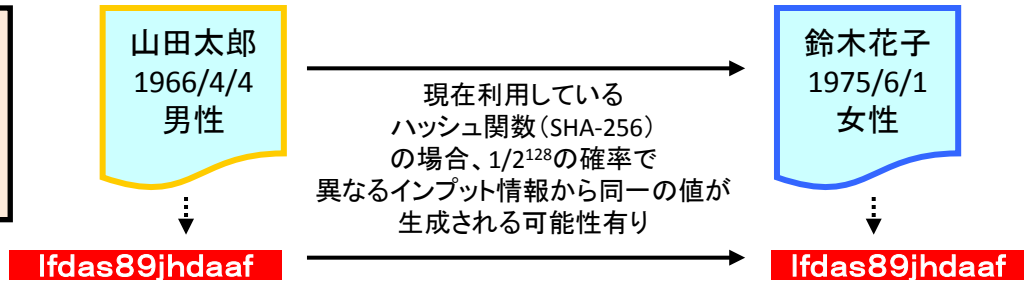


②レセプト情報と健診・保健指導データでは氏名の記載ルールが異なる

■レセプト　：　漢字氏名
■健診・保健指導　：　カナ氏名

→ インプットが異なるためハッシュ値も異なる

③ハッシュ関数の技術的特性として、極めて小さい確率ではあるが、異なる入力情報から同一のハッシュ値が生成される可能性がある。



留意点への対応

前ページの留意点に対応するため、現在、情報に変化のある「保険者番号、記号・番号」及び「氏名」について、それぞれ別のハッシュ関数を生成させ、データの突合の精度を向上させている。

ハッシュ値を2つ生成させる

① 保険者番号・記号番号・生年月日・性別からハッシュ値①を生成させる。

保険者番号
記号
番号
生年月日
性別
→ fjdaosiujt434et

② 氏名・生年月日・性別からハッシュ値②を生成させる。

氏名
生年月日
性別
→ 897435gadsgja

対応可能なケース

ケース①（記号・番号変更）

転職などで保険者番号、記号・番号が変更になった場合

ハッシュ値②により紐付けが可能

※ ただし、年月日・性別・氏名について同一の人物がいた場合、紐付けが不可能となる。

ケース②（氏名変更）

氏名の記載ミス、結婚などで氏名が変更になった場合

ハッシュ値①により紐付けが可能

※ ただし、生年月日、性別について同じ人物が同一記号・番号内に2名以上、存在した場合、紐付けが不可能となる。（双子など）

ケース③（レセプトと健診・保健指導データの紐付け）

氏名の記載ルールが異なるレセプトと健診・保健指導データを紐付ける場合

ハッシュ値①により紐付けが可能

※ ただし、生年月日、性別について同じ人物が同一記号・番号内に2名以上、存在した場合、紐付けが不可能となる。（双子など）

対応不可能なケース

記号・番号と氏名ともに変更があった場合

・結婚などで保険者が変更、氏名が変更になった場合
・転職などで保険者が変更、氏名の記載ミスがあった場合

目次

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。

サンプリングデータセット：匿名化処理

➢ 基本的な匿名化処理の方針

- 傷病名や診療行為といった患者に関する情報で、レセプトに出現する回数の少ないコードがそのまま記載されていると、患者の特定可能性に留意する必要が生じる。一方で、出現回数の少ないコード情報を含むレセプトをすべて削除してしまうと、母集団の性質が反映されないサンプルとなる恐れがある。
- したがって、出現回数の少ないコード情報を**特定のコードで代替（ダミー化）**することで匿名化処理を行う。
  ※匿名化の手法については、第8回有識者会議での議論（本資料P10-13）も参照。

➢ 匿名化処理の対象

- マスターのあるコード分類のうち患者の特定可能性を下げる観点で必要と思われる以下について匿名化を行う。

傷病名マスター　**医科診療行為マスター**　**医薬品マスター**

- 「特定器材マスター」「コメントマスター」「調剤行為マスター」「修飾語マスター」については**匿名化を行わない**

➢ 匿名化処理の基準

- 「医科入院」「DPC」「調剤」「医科入院外」各レセプト種別において、それぞれのマスターごとに、何回コードが出現しているかを算出する。
- これを全てのレセプトで合計し、総出現回数を求める。
- 出現回数の少ないコードから順に、総出現回数の**0.1%**に達するまで、匿名化を行う（**「0.1%ルール」**）。

※ＤＰＣについて（詳細）

- DPC診断群分類に対しても、「0.1%ルール」に沿って匿名化を行う。また、傷病名（SB）、診療行為及び医薬品のコーディングデータ（CD）、出来高部分の傷病名（SY）、診療行為（SI）、医薬品（IY）等、各コードについても「0.1%ルール」を適用する。

サンプリングデータセット：匿名化処理

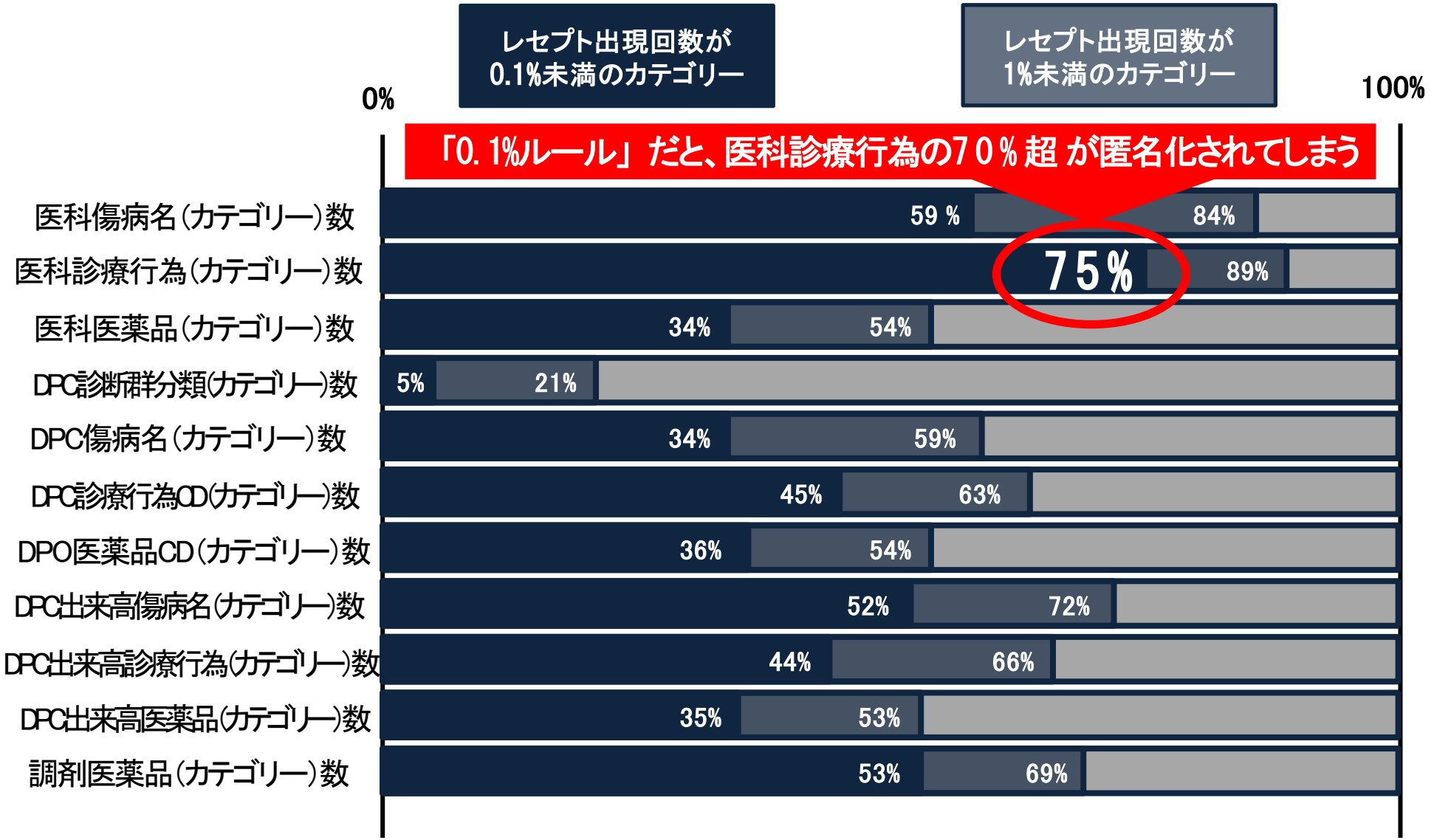
➢ 匿名化処理の基準：「医科診療行為マスター」における例外的な扱い

- 「医科診療行為マスター」においては、以下のような論点がある。

  - 「再診」「処方せん料（その他）」「明細書発行体制等加算」など、数千万件単位で算定されている入院外診療行為があるため、「0.1%ルール」を適用すると、**レセプト出現回数が2,000程度**に達する診療行為でも、匿名化されてしまう。
    匿名化される診療行為例：往診（深夜）加算、胃洗浄、腹腔鏡下胆嚢摘出術など。
  - 入院中に実施される診療行為、とくに手術の多くが匿名化されてしまう。
  - 「0.1%ルール」が適用された場合、他のマスター（傷病名、医薬品（医科、調剤））においては、レセプト出現回数が**おおよそ100～200程度**のコードが匿名化されている。

- したがって「医科診療行為マスター」においては、「0.1%ルール」をさらに緩和してはどうか。すなわち、「レセプトでの出現回数」が**全出現回数の0.01%以下（レセプト出現回数が100～200程度までのコードが匿名化される水準）**の診療行為コードについて匿名化してはどうか。

ある月のレセプトごと各コードの出現回数比



サンプリングデータセット：対象月の基礎情報

- **サンプリングデータセット対象レセプト情報（平成２３年１０月診療分）**

	レセプト総枚数	データ総容量	1レセプトあたりファイル容量
医科入院	1,402,187枚	12.1GB	8.6KB
医科入院外	77,559,281枚	175.6GB	2.2KB
DPC	915,517枚	16.7GB	18.2KB
調剤	48,513,258枚	87.4GB	1.8KB

（※）本データは、平成24年2月現在において格納されているレセプトデータの総数である。

サンプリングデータセット：その他の処理

➢ 匿名化したコードの点数情報について

- 「医科診療行為マスター」「医薬品マスター」においてコードを匿名化する際には、それらコードの点数情報についても匿名化する。ただし他の行で合算されている場合にはそのままとする。
- 「記録されている点数から匿名化した**コードを推定してはならない**」という約束を明記する。

➢ 高額レセプトの扱い

- 保険局で行っている医療給付実態調査において、点数階級分布で使用している**「入院診療700,000点以上」「入院外診療50,000点以上」**に該当するレセプトを**最初に削除**したうえで抽出を行う。すなわち、該当レセプトは母集団から削除される。
- 上記のレセプトを最初から削除する理由は、医薬品など点数情報が別の行で合算されている場合、点数情報を匿名化することが難しく、「高額群」として一括りにしたレセプトの点数が、他の情報から推定できてしまう恐れがあるためである。

➢ その他削除した項目

- 公費医療レセプトは、公費医療であることを確認できる情報をすべて除いたうえで、抽出を行う。レセプト数が多いことから、レセプトそのものを**抽出前に削除することは行わない予定**である。
- 医科及びDPCレセプトで、移植医療を受けた患者のレセプトに含まれる臓器提供者関連情報はすべて削除する。
- その他、以下に該当する項目は削除する。

保険者に関する情報　　**医療機関コード**　　**都道府県情報**

➢ 各種マスターにないコードの扱い

- いずれのレセプトにおいても、データとして残っているコードが、同時期の「マスター」では確認できない事例がある。
  →平成23年10月のマスターと照合し、マスターにないコードの情報は削除する。

（参考）ある月における高額レセプトの状況

累計点数	出現頻度	
	医科入院	DPC
100,000～	4.5%	14.04%
200,000～	0.50%	2.57%
300,000～	0.16%	0.80%
400,000～	0.072%	0.33%
500,000～	0.036%	0.16%
600,000～	0.016%	0.063%
700,000～	0.008%	0.022%

累計点数	出現頻度	
	医科入院外	調剤
30,000～	0.40%	0.038%
40,000～	0.15%	0.021%
50,000～	0.042%	0.014%

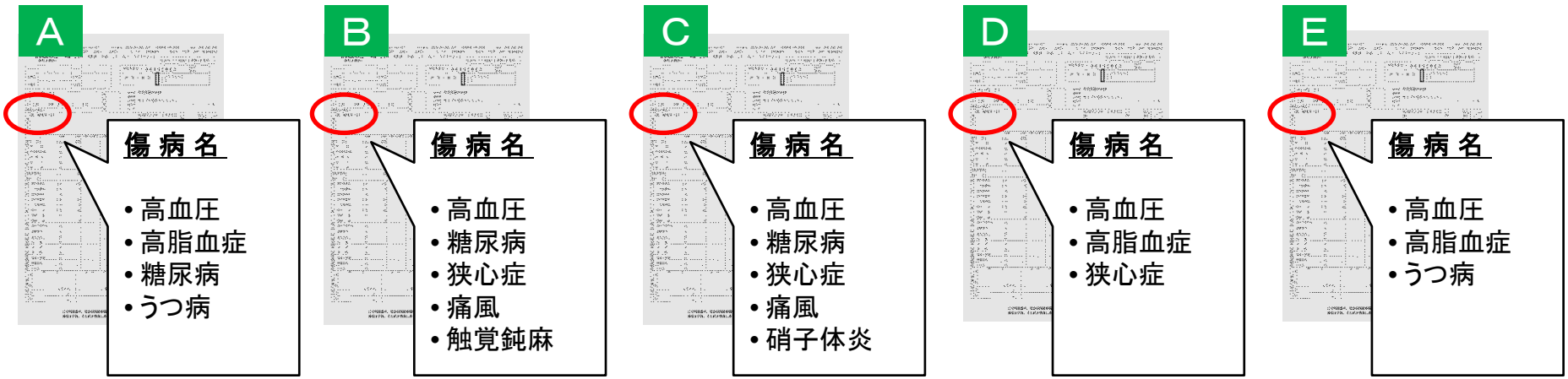
「サンプリングデータセット」の具体案：データ内容

4. 匿名化処理をどう行うか？

- レセプトに出現する回数が少ない情報（たとえば「傷病名」、「診療行為」、「医薬品」コード）が含まれていると、それらの情報から個人が特定されてしまう可能性が高くなる。このため、レセプトに出現する回数が少ないコードについては、**一定の割合で匿名化処理を行う**こととしてはどうか。
- マスターが用意されている各コード（「傷病名」「診療行為」「医薬品」など）において出現回数の低いものを一定数匿名化すると仮定する。その際、レセプトに出現する回数を基準にして匿名化の基準を定めるとなれば、どの程度の数の傷病名コードを匿名化することになるだろうか？

例：循環器内科外来に通院する方の以下AからEの5枚のレセプトにおいて、個人が特定される可能性を下げるため、これら5枚のレセプトに記録されている傷病名を、出現回数を基準として少ないものから１０％匿名化するとしたら？

※ この事例は架空の設定にもとづいたものであり、必ずしも実態を反映したものではない。

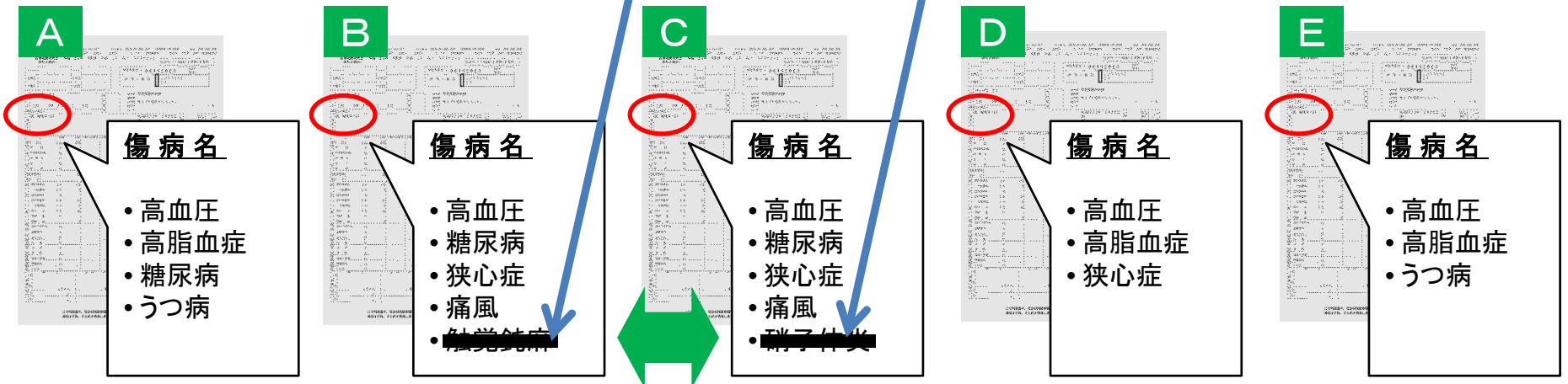


「サンプリングデータセット」の具体案：データ内容

集計結果

	1	2	3	4	5	6	7	8	
傷病名	触覚鈍麻	硝子体炎	うつ病	痛風	糖尿病	高脂血症	狭心症	高血圧	合計
出現回数	1	1	2	2	3	3	3	5	20
レセプト	B	C	A, E	B, C	A, B, C	A, D, E	B, C, D	A,B,C,D,E	
全出現回数に占める割合	5%	5%	10%	10%	15%	15%	15%	25%	100%

希少疾病を指す新たなコードを付与する

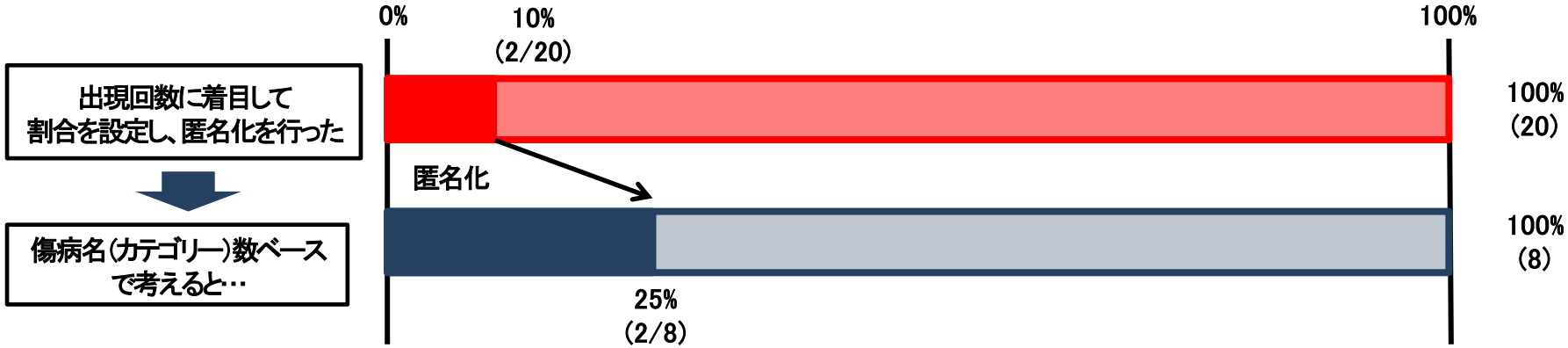


BとCの区別がつかなくなった

「サンプリングデータセット」の具体案：データ内容

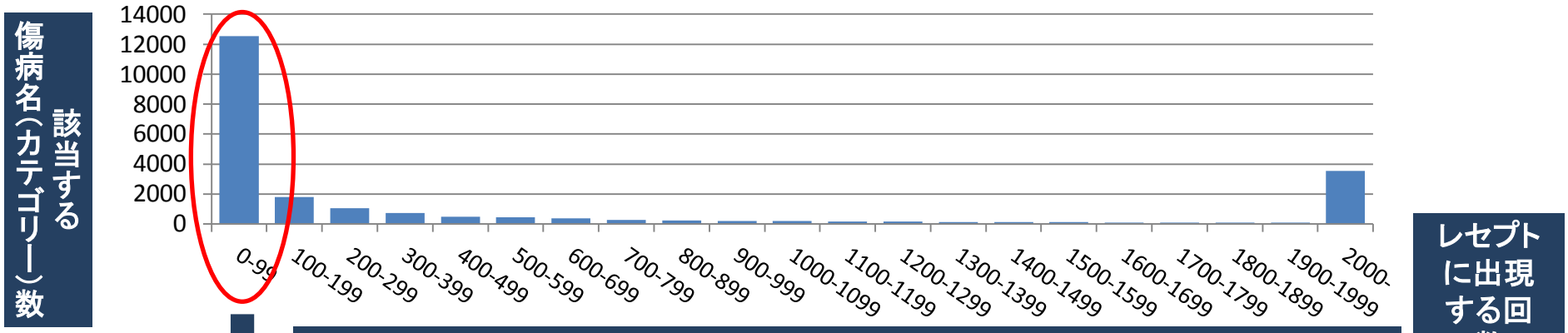
傷病名（カテゴリー）数とレセプト出現回数の関係

- この事例では5枚のレセプトの匿名性を高めるため、5枚のレセプトに出現する傷病名の**出現回数の少ないものから「10%」**を匿名化することを考えた。
- 集計結果から、1度しか出現しなかった**「触覚鈍麻」「硝子体炎」**を合計すると10%に達したためこれらを匿名化した。その結果、傷病名からは**[B]と[C]の区別がつけられなくなるなど、5枚のレセプトの匿名性を高めることができた。**
- しかし、「出現回数」を「10%」に設定することで匿名化した傷病名は「触覚鈍麻」と「硝子体炎」の**2傷病名（カテゴリー）**であり、これはこの5枚のレセプトに出現する全ての傷病名（**8傷病名（カテゴリー）**：「触覚鈍麻」「硝子体炎」のほか、「うつ病」「痛風」「狭心症」「高脂血症」「糖尿病」「高血圧」）のうち、**「25%」**に相当する。
- つまり、出現回数の少ない傷病名や出現回数の多い傷病名があるため、**傷病名（カテゴリー）数からみた匿名化の割合は、「出現回数」を基準にして設定した匿名化の割合よりも高い割合をとる**こととなる。これを帯グラフで表すと、以下のようになる。



「サンプリングデータセット」の具体案：データ内容

（参考）ある月の医科レセプトにおける各傷病名（カテゴリー）の出現回数から



ひと月に100回未満しかレセプトに出現しない傷病名（カテゴリー）が１２，０００以上と、傷病名（カテゴリー）全体（22,890として計算。傷病名（カテゴリー）数はマスターの更新時期によって変動する）の半分超を占めている。したがって、レセプトに出現してくる傷病名（カテゴリー）のほとんどは、出現回数の高い数10パーセント程度の傷病名（カテゴリー）でカバーされているのが実態である。

例：この月の場合、レセプトに記録される傷病名の出現回数のうち99％は、傷病名（カテゴリー）全体の16.4％、3,749の傷病名（カテゴリー）のみでカバーされている。下図参照。