

3 基本データセットにおける 匿名化基準等について

データセットにおける希少傷病名の匿名化について

データセットにおいては、傷病名を主傷病名などから1つに限定して提供することとしているが、その傷病名が非常に希少な傷病名の場合、患者が特定される可能性が高まる恐れがある

- 希少傷病名の取扱いについては、安全性を重視した場合と、利便性を重視した場合とで以下に分けられる。
 - ◆ 安全性を重視した場合…希少傷病名は匿名化して提供する
 - ◆ 利便性を重視した場合…特段の対応を行わない
- 匿名化を行った場合、サンプリングデータセットと異なりデータセットの傷病名は1つに限定されるため、匿名化されたレセプトからは傷病名情報が一切消失してしまい、傷病名を元にした分析が行えなくなる。
- もし匿名化を一切行わないとすると、一定の条件で抽出されたデータ(保険者単位で約25%、その後患者単位で20%抽出)であるとはいえ、非常に希少な傷病名情報が付与されたデータが提供される可能性があり、その患者の特定可能性が高まることは否定できない。

- 最低限の匿名化として、抽出されたデータを年度毎に全て俯瞰し、**1回**しか出現しない傷病名については、匿名化して提供することとしてはどうか。

参考:「1回しか出現しない傷病名のみの匿名化」についての概念整理→以下の推定をもとにした匿名性の比較

- ・ 保険者単位で25%抽出、その後患者単位で20%抽出される→**およそ5%抽出**されることとほぼ同等と考えられる
- ・ 5%抽出下で、1回しか出現しない傷病名を匿名化する→**全数データで20回以下しか出現しない傷病名が匿名化**されることとほぼ同等と考えられる

	匿名化される傷病名の割合	併存傷病名情報について
サンプリングデータセット	59%(第9、17回有識者会議資料参照)	匿名化されていない併存傷病名については情報あり
基本データセット	25%(研究班試算)	併存傷病名はなく、傷病名は1つに限定されている

ICD-10コードによる傷病名代替のイメージ

データセットにおいては、傷病名を主傷病名などから1つに限定して提供することとしているため、傷病名を匿名化したデータにおいては傷病名情報が消失してしまう

- 傷病名が匿名化されたデータにおいては、傷病名の代わりとして、匿名化された傷病名コードを基として、**アルファベットを含めて3桁レベルのICD-10コード番号を提供することとしてはどうか**

- ✓ ICDとは、正式には「疾病及び関連保険問題の国際統計分類：International Statistical Classification of Diseases and Related Health Problems」と言い、異なる国や地域から、異なる時点で集計された死亡や疾病の記録、分析、比較を行うために国際的に統一した基準で設けられた分類のこと。
- ✓ 日本では現在、1990年に実施された第10回修正 (ICD-10)、およびその後2003年までの一部改正分を盛り込んだ、ICD-10

- (2003年版) 準拠を使用している。
- ✓ ICD-10コード1桁目がアルファベット、2桁目以降が数字で示されるコードであり、疾病統計においては大分類、中分類、小分類という体系で分類されている。

(詳細は、「疾病、障害及び死因分類の正しい理解と普及に向けて」(<http://www.mhlw.go.jp/toukei/sippe/dl/fukyuubon.pdf>)を参照)

※ データの傷病名がICD-10コードで代替されるイメージ
(あくまでも一例であり、実態を反映するものではない)

傷病名: 若年性境界型高血圧症
(4019017)
診療行為: 再診料 ○回
処方料 ○回…
医薬品: ノルバスクOD5mg ○回
セルベックスカプセル50mg ○回…

匿名化

傷病名: ~~若年性境界型高血圧症~~
~~(4019017)~~
診療行為: 再診料 ○回
処方料 ○回…
医薬品: ノルバスクOD5mg ○回
セルベックスカプセル50mg ○回…

傷病名での分析ができなくなる

ICD-10
にて代替

傷病名: I10
本態性(原発性<一次性>)高血圧(症)

どういった疾患の患者なのか
把握することが可能

ICD-10コードで代替することにより
複数ある傷病名コードのどれなのかを
特定することが困難になる

ICD10 コード	傷病名	傷病名 コード
I10	悪性高血圧症	8830212
	境界型高血圧症	8832479
	高血圧症	8833421
	高血圧緊急症	8842178
	高血圧切迫症	8842488
	高レニン性高血圧症	8842089
	収縮期高血圧症	8842500
	若年高血圧症	4019016
	若年性境界型 高血圧症	4019017
	低レニン性高血圧症	8842094
	本態性高血圧症	8840107

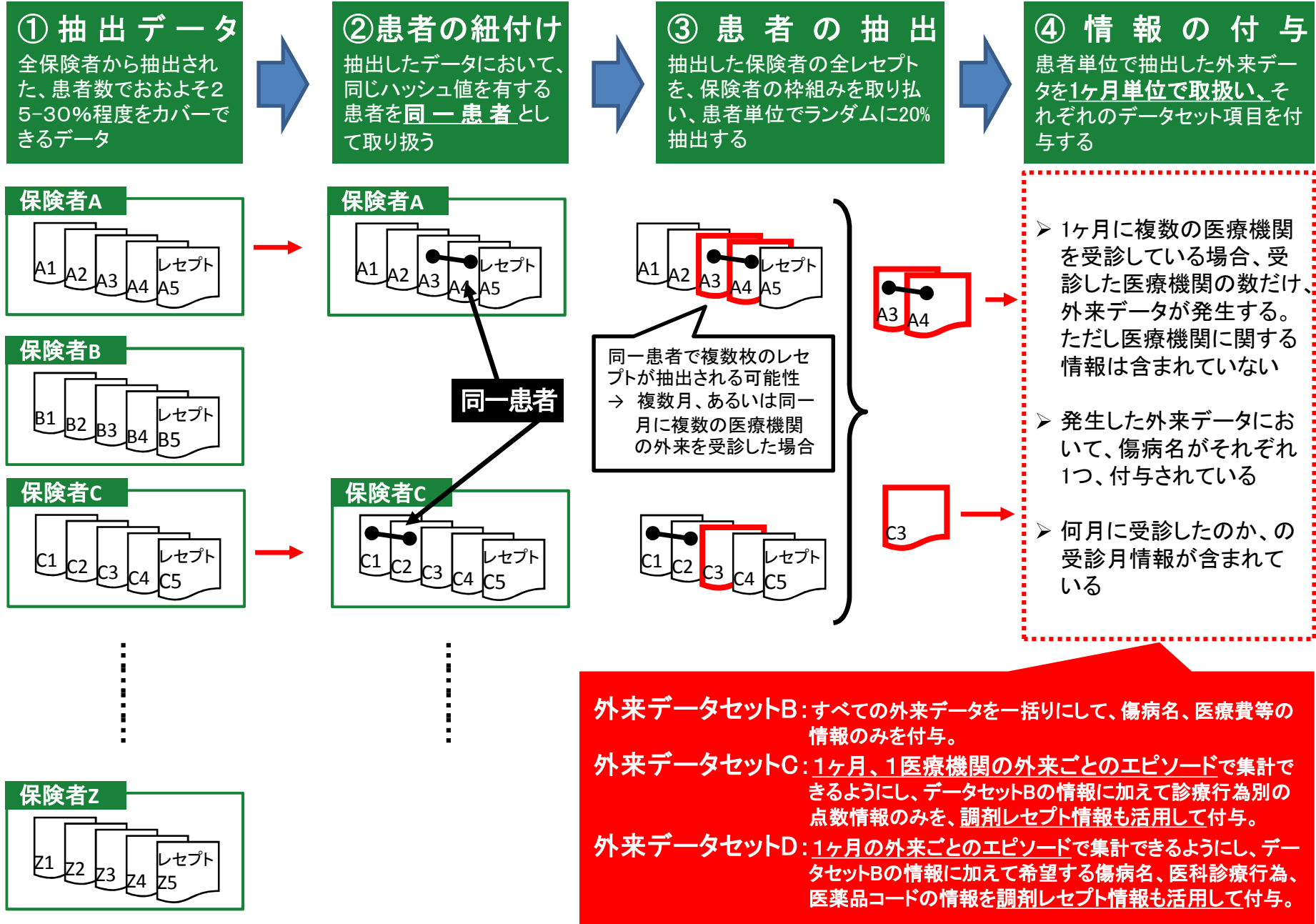
データセットDにおける、医薬品等特定のコード情報の提供について

データセットDにおいては、診療行為や医薬品に関する情報を申出者において指定することとしているが、サンプリングデータセットと同等のセキュリティ要件にて提供することを認めることとするのであれば、特定の医薬品情報等が提供されることは、個人の特定可能性を高めうる観点から慎重に検討する必要があるのではないか



- 特定の医薬品情報に限らず、特定の診療行為や特定器材等の情報を提供することにより個人の特定可能性を高まりうることについて懸念が発生する余地があると言えるものの、申出にあたっては有識者会議審査分科会において有識者の審査を経ることとしていることから、データ提供にあたってはこの点に十分留意した上で、慎重な審査を行うことを以て対応することとしてはどうか。

外来データセットの大まかなイメージ



外来データセット B, C, Dについて

	外来データセットB	外来データセットC	外来データセットD
ファイル容量	<ul style="list-style-type: none"> 約2GB 	<ul style="list-style-type: none"> 約2GB 	<ul style="list-style-type: none"> 約4GB（要望内容により変動）
レコード数 (=患者数)	<ul style="list-style-type: none"> 保険者単位で抽出した3000万件の患者レコードから、さらに20%をランダムに抽出 	<ul style="list-style-type: none"> 左記患者の外来エピソード全て（年度単位で要望可） 	<ul style="list-style-type: none"> 左記患者の外来エピソード全て（年度単位で要望可）
抽出患者	<p>【保険者の抽出方法】</p> <ul style="list-style-type: none"> 都道府県単位の保険者（協会けんぽ等）は、全国8ブロックから1つずつ、人口25%分の保険者を抽出 全国にまたがる保険者（組合健保、共済等）は、全保険者から20%をランダムに抽出 <p>【抽出された保険者間での、患者抽出方法】</p> <ul style="list-style-type: none"> 抽出された保険者の全患者から20%分に該当する患者をランダムに抽出 		
対象期間	<ul style="list-style-type: none"> 1年 	<ul style="list-style-type: none"> 外来イベント全て データセットBと併用することで、年間の外来イベントを全て把握できる。 	<ul style="list-style-type: none"> 外来イベント全て データセットBと併用することで、年間の外来イベントを全て把握できる。
キーになるID	<ul style="list-style-type: none"> ID1とID2を用いて試行的に作成したID3を用いて、レセプト間、データセット間の紐付けが可能である。 		
患者属性情報	<ul style="list-style-type: none"> IDおよび性、5歳刻み年齢階級 		
項目	<ul style="list-style-type: none"> 外来合計点数 外来合計診療実日数 外来受診医療機関数 主傷病名 	<p>【外来データセットBに加え】</p> <ul style="list-style-type: none"> 診療識別情報 	<p>【外来データセットBに加え】</p> <ul style="list-style-type: none"> 申出者の要望により、診療行為、医薬品情報、傷病名等を提供
データ構造	<ul style="list-style-type: none"> データセットは固定 	<ul style="list-style-type: none"> データセットは固定 	<ul style="list-style-type: none"> データセットは要望に応じ可変
研究の例	<ul style="list-style-type: none"> 年間の医療機関受診動向 年齢階級別、性別、傷病別の医療費 等 	<p>【左記に加え】</p> <p>患者別の診療行為別医療費分析等（社会医療診療行為別調査の個票のイメージ）</p>	<p>患者別の診療行為・医薬品・特定器材の利用別分析等（研究に必要なデータ項目を含み個票）</p>

- ユーザーは、申請前に各ファイルのデータ項目を確認し、対象となる疾患数等の集計情報を参考にして、あらかじめ用意された数パターンの抽出ファイルから選択する。

外来データセット B, C, Dの利用に当たってのセキュリティ環境について

- ✓ 外来データセットB, C, D は、年度単位で経時的に紐付けられてはいるものの、傷病名や診療行為、医薬品といった情報については、相当程度削除されている。
- ✓ これらのデータセットを使用するにあたっての適切なセキュリティ環境について、どう考えるか。

記載項目 匿名化項目	匿名化や削除など、加工されている主な項目等			
	サンプリングデータセット	外来データセットB	外来データセットC	外来データセットD
レセプトの 各種属性情報	<ul style="list-style-type: none"> 性別、5歳刻み年齢階層別 元データの構成比率を保ち、1%もしくは10%の抽出 	<ul style="list-style-type: none"> 性別、5歳刻み年齢階層別 約25%に抽出した保険者から、被保険者をランダムに20%抽出 ハッシュ値1と2を用いて作成した新しいIDを付与 		
保険者情報	<ul style="list-style-type: none"> 保険者番号は削除 被保険者証等記号・番号は匿名化（ハッシュ値2） 	<ul style="list-style-type: none"> 左記に同じ 		
傷病名	<ul style="list-style-type: none"> 希少なものは匿名化 	<ul style="list-style-type: none"> 主傷病名1つのみ 	<ul style="list-style-type: none"> 1ヶ月分、1医療機関の外来エピソードごとに、主傷病名1つのみ 	<ul style="list-style-type: none"> 複数病名（申出者の要望により設定可）
診療・医薬品・ 特定器材コード		<ul style="list-style-type: none"> なし 	<ul style="list-style-type: none"> なし 	<ul style="list-style-type: none"> 設定可能、上限あり
その他	<ul style="list-style-type: none"> 高額レセプトは削除 DPCレセプトは医療機関別係数及びその影響がある項目は削除 患者や医療機関を特定できる可能性のあるその他記載項目（病床数等）はカテゴリー化 	<ul style="list-style-type: none"> 上記以外の項目は、年度を通じた合計点数、診療実日数、受診医療機関数等のみ 	<ul style="list-style-type: none"> 1ヶ月分、1医療機関の外来エピソード毎にデータがまとめられている データセットBに加え、診療行為別の点数情報が付加される 	<ul style="list-style-type: none"> 要望された情報については、1ヶ月分の外来エピソード毎に「回数」「点数」情報のみが提供され、日付情報は含まれない

- 外来データセットB, C については、サンプリングデータセットと比較しても相当程度、診療情報が削除されており、匿名性の高いデータとして位置づけられるのではないかと考えられる。
- 外来データセットの匿名性を鑑みて、サンプリングデータセットと同様のセキュリティ要件を適用してはどうか。

傷病名の特定化:「主傷病名」の使用、等

・第17回有識者会議での議論

- 基本データセットでは、各レセプトに対し記載されている傷病名から「主傷病名」をもとにした傷病名を入院・外来それぞれ1つを付与した、サマリー的なデータセットが提供可能である。
- 具体的な傷病名の選択方法としては、レセプトに付与されている傷病名のうち、主傷病名として格納されているものの中から最も新しい傷病名であり、複数レセプトの中で一番高い点数の病名を付与することを想定している。

・加えて、今回新たに提案する傷病名の特定化について

- レセプトの中には主傷病名がない場合もあり得る。この場合、最も新しい傷病名で、かつ一番最初に記載された傷病名を付与することを想定している。

・外来データセットにおける傷病名の特定化について

- 外来データセットにおいても、入院データセットにおける傷病名付与のルールを踏襲することとする。

・傷病名を1つに特定することの長所と短所

- レセプトには複数の傷病名が記載されていることが少なくないが、傷病名が1つに特定されることで、傷病名を基準とした分析が簡便に行えることが期待できる。
- 一方で、いわゆる「レセプト病名」の存在など、1つに特定した傷病名がどの程度実態を反映できているのかについては、しばしば議論の対象となっている。特に、医科入院外レセプトの場合、年余にわたり継続的に付与されている慢性疾患(例:高血圧症 等)の傷病名と、一時的に罹患した急性疾患(例:急性腰痛症 等)の傷病名とが混在している可能性が少なくなく、そのいずれをもって傷病名として代表させるのが良いかについては、意見が分かれるところである。
- 傷病名の付与については、当座はこれまで有識者会議で議論してきたルールに則って行っていくものの、今後データ提供を経るなかで、データ利用者からの意見も踏まえ、より適切な付与ルールの整備も含め、継続的に検証を続けていくこととしてはどうか。