資料1

サンプリングデータセットの提供について

平成24年３月21日

厚生労働省保険局総務課

【目次】

① レセプト情報・特定健診等情報データベースの第三者提供の経緯について

（参考）「医療サービスの質の向上等のためのレセプト情報等の活用に関する検討会」（抄）
（平成20年2月７日）

４　国が行う分析の目的に関する考え方

(1)　医療費適正化計画の作成等に資する調査・分析を行うことが、高齢者医療確保法第１６条に基づきレセプトデータ及び特定健診等データを収集する一義的な目的である。

(2)　上記(1)の分析以外であっても、当該データを活用することが、新たに別途データを収集することと比較考量すれば、国民負担の軽減につながり、また迅速な分析、的確・適切な施策の迅速な実施により、行政サービスの向上、行政運営の効率化につながる場合もあると考えられる（例えば、感染症などの疾患の実態把握に基づく施策や、介護給付費と医療費の実態把握に基づく施策など）。
このため、所掌事務の遂行に必要な範囲内であることを前提とした上で、上記(1)の分析のほかにも、当該データの分析・活用が、上記(1)の分析目的と同様に、医療サービスの質の向上等を目指して正確なエビデンスに基づく施策を推進するに当たっての必要かつ有利となる場合についても、国が行う分析の目的に含めて考えることも必要と考えられる。

６　国以外の主体によるレセプトデータ等の活用のあり方

(2)　上記４(2)に示したような考え方を前提とするならば、国以外の主体が、国が収集したレセプトデータ及び特定健診等データを用いて、医療サービスの質の向上等を目指して正確なエビデンスに基づく施策を推進するに当たって有益となる分析・研究、学術研究の発展に資するような研究を行うことを一律に排除することは、国民負担の軽減、的確・適切な施策の迅速な実施という視点に立てば、かえって適切とは言えないと考えられる。
したがって、上記(1)により都道府県が活用する場合のほか、国以外の主体がこうした公益目的で国の収集データの提供を受けて分析・研究し、国において施策を検討する際にその分析・研究の成果を活用できるような仕組みも必要と考えられる。
　ただし、その際には、以下の点について十分留意する必要がある。

①　データの利用目的として公益性の確保が必要であることのほか、研究目的や研究計画、データの分析方法、データの使用・管理方法等について、個別に審査した上で、当該研究に必要な範囲内でデータを提供すること。

②　個別ケースごとの審査に当たって、公平・中立な観点から、データ利用の目的や必要性等について審査し、提供の可否等を決定する仕組みが必要であること。

③　個別ケースごとの審査の基準となる、第三者への提供に係る具体的なルールが別途必要であること。
当該ルールづくりに当たっては、新統計法における調査票情報等の利用及び提供のルール（現在総務省及び内閣府統計委員会において検討中）も踏まえて検討する必要があること。

④　上記③のルールに基づき国から適切にデータの提供を受けた者以外の者が、結果的に当該提供データをそのまま利用することのないよう徹底すること。
また、この点についても上記③のルールの中で必要な措置を講じておくこと。

⑤　レセプトデータ及び特定健診等データには、患者の病名等慎重に取り扱うべきデータが含まれていること等にかんがみ、上記③のルールに基づいて国がデータを提供する際にも、収集データをそのままの形で提供することは適当ではなく、当該データの一部（例えば患者等について原則として同一人物に同一に付される一連の番号、医療機関・薬局コード、一部の病名など）を加工するなどの対応が別途必要であること。
この場合の対応方針についても、上記③のルールの中でできるだけ明確に整理しておく必要があること。

新たな情報通信技術戦略工程表（抜粋）
（平成22年６月22日閣議決定）

２　１）iii）
レセプト情報等の活用による医療の効率化

短期（2010 年、2011 年）
〇レセプト情報等の提供のためのルールを整備し提供を開始する。また、膨大な関連情報の分析や活用のための技術等の研究開発を実施する。さらに医療効率化のためのデータ利用の在り方についての一次検討を実施し、各種データの一元的な利活用に向けた提供体制についても検討を実施する。また、匿名化やセキュリティ技術、大量データ分析・活用に向けた技術開発について検討を開始する。

厚生労働省：
2010 年度から各種データの一元的な利活用に向けた提供体制を検討
2010 年度中に有識者による検討会議の設立
2010 年度中にデータ活用のためのガイドライン策定
2011 年度早期にデータの提供開始
2011 年度から医療効率化のためのレセプトデータ等の利活用に関する
調査・検討を実施

②　審査方針について

事前審査の論点を踏まえた、事務局審査方針

研究内容・抽出について

① **「個人の識別可能性を下げる」という原則**に鑑み、**「対象者が極めて限定される可能性がある」**申出は、申出者にデータを渡した時点で個人が特定される可能性を否定できないため、事務局審査では提供依頼について不承諾とした。

- 「10未満のセルは空欄にする」「申出以外の分析は一切行わない」と明記されている申出であっても、今回は試行期間ということもあり、不承諾と考えた。
- 少数セルが多発する可能性が、研究を開始してみないとわからない、という意見もあった。今後検討予定の「基本データセット（仮称）」が構築されれば、その活用により事態が改善されると考えられる。
- 集計時に少数セルが頻発することが想定される申出のなかには、申出者が配慮すると明記されており、他の審査方針を満たしていたため、総合的にみて承諾と考えたものもある。

② **同じく「個人の識別可能性を下げる」という原則**に鑑み、**多数の項目を用いた探索的研究**や、SYの「傷病名コード」、SIの「診療行為コード」、IYの「医薬品コード」（DPCの場合にはBU, SB, CD）どれかひとつでも**「全て求める」**という要望の申出は、事務局審査では提供依頼について不承諾とした。

- 多変量解析、propensity score分析、重回帰分析など多くの項目を必要とする申出も**探索的研究**と考え、今回は不承諾とした。
- 何らかの抽出を経たあとのサンプルに対しても、上記のような申し出に対しては同様の判断とした。
- この方針は**特定健診**の情報を用いた研究にも適用した。
- 上記のように、研究手法によってはこうした抽出条件とならざるを得ないものもあり、これも今後検討予定の「基本データセット（仮称、後述）」構築によって対応することが考えられる。
- 申出のなかには、傷病名コードを全て求めてはいたものの、「2次医療圏単位の集計」に限定されているなど個々人の特定可能性が極めて低いため、承諾と考えたものもある。

③ **「1申出1審査」**という原則に鑑み、**「複数の研究」**が1申出に盛り込まれている場合も、審査にあたっては他の申出内容も加味したうえで、提供についてはより慎重な評価を行った。

- 事務局から内容の照会を行ったことにより再提出において改善がみられたものもあったが、「利用目的」ごとの申出書作成を原則としている（ガイドラインp.9）ことを踏まえた評価を行った。
- 細分化された研究が並んでいた申出のなかには、全体が大きな目的で統一されており、その他の審査方針も含めて総合的に鑑みて承諾と考えられるものもあった。

④ **研究には公益性が求められるため、**研究に際して抽出項目の措定や目的と抽出項目との関連において、事務局審査にて**「定義が不十分」**な箇所があると思われる申出は、相対的にみてより慎重な評価を行った。

- 抽出項目が研究内容からみて最小限とはいえなかったり、関連性が不明、または説明不十分と判断された申出がこれにあてはまる。
- 公表形式の具体例が研究内容に鑑みて非常に限定されていたり、公表形式から想定される必要項目をはるかに超える抽出項目を要求したりする申出のなかには、研究内容の「定義が不十分」と思える申出もみられた。

⑤ **その他**、以下のような事例に対しても、事務局審査において慎重な評価を行った。

- 都道府県単位など、**地域に限定したデータ提供**の申出については、国が保有する全国レベルのデータベースという位置づけに鑑みて、その公益性について留意した。
- 「集計表情報」の提供を求める申出のなかには、集計表作成が複雑であり学術研究の領域に踏み込んでいると思われるものもあった。集計表情報作成は、**簡略な操作にて作成できるもののみを対象**とした（単純なクロス集計など）。

事前審査の論点を踏まえた、事務局審査方針

セキュリティ要件について

①「情報セキュリティマネジメントシステム（ＩＳＭＳ）の、申出者個々の研究環境に応じた合理的な対応」の実践を求めていることに鑑み、**独自のセキュリティ規程が一部、もしくは全て欠けている事例**は、事務局審査にて提供依頼について不承諾とした。

- 事務局から内容の照会を行ったことにより再提出において改善がみられた申出が複数みられたが、その際にも、「運用フロー図」「リスク分析・対応表」「運用管理規程」「自己点検規程」の整合性について、他の申出と同様に評価を行った。
- 情報資産の安全をチェックする頻度が、上記書類の相互で区々になっている申出など、セキュリティ対策の書類相互において齟齬が著しいものについては、より慎重な評価を行っている。
  - ✓ 散見された例として、
    - ✓ 情報資産の特定不十分：USBや帳票が、「使われる」ことになっている書類と「使われない」ことになっている書類がある。
    - ✓ チェック体制の不備：自己点検規程には点検頻度が記載されているが、運用管理規程ではそのことについて言及がなされていない。
- 再提出時にそれらが修正されている場合にはその点も考慮した。一方、改善が不明瞭な場合は不承諾とした（②以下も同様）。

②入退室の管理が不十分であったり、利用者以外のアクセスが可能な場所でレセプト情報等が利用される事例についても、より慎重な評価を行い、なかには不承諾としたものもある。

- 研究室や情報管理室が共有スペースとなっており、他者が容易に解析機器に接触できる事例については、それらに対するセキュリティの強度や透明性の確保をどの程度整備できているのかについて留意した。
  - ✓ ISMSの観点からすれば、ハード面でのセキュリティ確保以上に、申出者それぞれの環境に適合した実践可能なセキュリティ対策の整備のほうが、より重要視されている。
  - ✓ たとえ「レセプト情報を管理する部屋には研究者のみ入ることができる」と記載されていても、その部屋が新たに準備された部屋と明記されているか、共用スペースと認識されかねない名称であるかで、記載内容の信頼度は変わってくる。研究環境が共用施設である場合には、記載の一貫性が保たれているかどうかについても評価した。

③研究者や所属施設、研究施設が複数（多数）にまたがる事例については、セキュリティ対策実践の難易度が上がると想定されるため、その対応について慎重に評価を行い、不承諾とした申出もある。

- 施設間での業務分担が不明瞭な申出については、不承諾とした。
- 申出のなかには申出者が多数となるものもみられたが、研究環境に即した現実的な対応が提案されていたため、総合的に鑑みたうえで提供依頼について承諾としたものがある。

④技術的対策が不十分（ＩＤ管理、外部ネットワークとの接続など）な事例については、より慎重な評価を行った。

- ID管理をする際の本人確認など、透明性の確保の程度に応じて、管理体制の評価を行い、承諾可否の判断根拠の一助とした。
- やはりISMSの観点により、どれだけ既存の環境のなかで現実的な研究実施体制を構築できているかについて、着目した。
  - ✓ たとえば、外部ネットワークに接続しない機器を用いると記されている場合でも、それが新品であると明記されていたり他の研究への使用を禁じている事例と他の事例とでは、事務局審査において評価に差が生じた。

審査基準の明確化①～優先順位について～

○今後、データ提供の申出が増加していくことも考えられるが、申出が非常に多くなり、公益性やセキュリティ要件の面で審査を了する申出が増えた場合、全ての申出についてデータ抽出を行うことは困難となることも考えられる。

（参考）第１回申出における医政局指導課のデータ抽出では、全国の半年間における医科・DPCレセプト4億7,000万件の抽出作業に約200時間を要した（営業日で２週間程度）。

○こうしたことから審査の基準とは別に、各申出にデータ提供の優先順位をつけ、順位の高いものから順番に対応できるものまでデータ提供を行うこととしてはどうか。評価の方法については、有識者会議の委員に協力を仰いで行う部分と事務局において行う部分とを設け、事前に各項目について点数化した上で有識者会議の審査に諮ることとしてはどうか（イメージは次ページ）。

【評価項目】

以下のうち、①と②については、有識者会議の一部の委員があらかじめ評価を行い、それ以外については事務局において、評価を行った上で有識者会議に諮ることとしてはどうか。

①研究内容の簡潔さ・明解さ（複雑・難解なものとなっていないか）

・複雑な研究内容や仮定を置いている研究であるかどうかについて判断することとして、理解が得やすいか否か。

②学術的な期待度

・申出された研究により、学術的に有意義な結果が得られる期待度が高いかどうか。

③具体的な政策への反映を想定しているか（単なる基礎資料か否か）

・具体的な政策への反映を想定しているものかどうか。

（例）医療計画の策定の基礎資料とするため都道府県への提供を想定　など。

④地域の範囲（全国か、地域限定の研究か）

・全国規模のデータベースという性質を活かす観点から、地域限定をした研究よりは、全国規模の研究を行うものを優先してはどうか。

⑤活用するデータ量・規模（大量のデータを使用するものか否か、抽出に要する見込み時間で判断）

・上記を第１回申出における処理時間約200時間を一応の標準的な処理時間として、データ抽出に要する時間についての評価を行うこととしてはどうか。

（参考）優先順位付けのイメージ

【点数付けの考え方】
〇事務的な作業が膨大であったとしても極めて学術的に有益な研究であれば、データ提供の優先順位上考慮する必要があることから、有識者委員の学術面での評価点数と事務局による実務面での評価点の最高点を同一とする。

	項目	評価
通常審査部分	研究内容の公益性	
	セキュリティ要件	
	・・・・	
優先順位評価部分	①学術的な期待度	□低い（1） □やや低い（2） □普通（3） □ある程度高い（4） □高い（5）
	②研究内容の簡潔さ・明解さ	□難解（1） □ やや難解（2） □普通（3） □ ある程度明解（4） □明解（5）
	有識者委員評価点合計	**（2～10）**
	③具体的な政策への反映	□想定していない（0） □想定している（3）
	④研究地域の範囲	□地域限定（0） □全国（2）
	⑤活用するデータ量 ※200時間を一応の標準処理時間とする。概ね300時間超がデータ量が多い。100時間未満が少ない。	□データ量が少ない（～100時間程度）（5） □普通（100～300時間）抽出作業（3） □データ量が多い（300時間～）（0）
	事務局評価部分	**（0～10）**
	優先順位部分の評価点合計	**〇〇点（0～20）**

有識者委員が評価

審査基準の明確化③～集計表情報について～

〇統計法のオーダーメード集計においては、集計対象となる項目をあらかじめ限定した上で、集計方法についてもその対象項目を2次元又は3次元までで集計するなどの縛りを設けている。

〇単純なクロス集計であっても、集計単位が複層化していく場合、複雑さが増すと共に個人の特定可能性も高まることが想定されるため、レセプト情報等についても一定の明確な基準が必要。

〇例えば集計対象項目は、レセプト毎の傷病名コード、診療行為コード、医薬品コード、特定器材コード等に限定し、集計方法については、性別、年齢階級別、都道府県別の集計を念頭において、原則、3次元までとすることとしてはどうか。

（参考）医療施設調査のオーダーメード集計の様式

3 集計対象項目

〇病院票
・施設数
・病床数（許可病床数、特殊診療設備、LDR、緩和ケア病棟）
・患者数（特殊診療設備、検査等の実施状況、緩和ケア病棟、緩和ケアチーム）
・設置台数（検査等の実施状況、手術等の実施状況）
・実施件数（在宅医療サービス、手術等の実施状況）
・従事者数（診療録管理専任従事者、分娩取扱従事者）

〇一般診療所票
・施設数
・病床数（許可病床数）
・患者数（検査等の実施状況）
・設置台数（検査等の実施状況、手術等の実施状況）
・実施件数（手術等の実施状況）
・従事者数

〇歯科診療所票
・施設数
・病床数（許可病床数）
・従事者数

5 オーダーメード集計提供項目
利用可能な集計区分は、集計対象項目ごとに分類一覧に示す区分となり、集計区分の組み合わせ（クロス数）は合計が3次元までとなります。ただし、「病床の規模」（病院票）、「病床の有無」（一般診療所票）及び「診療科目（重複計上）」（病院票及び一般診療所票）を含む組み合わせの場合は5次元まで可能となります。

今後の提供の論点②－公表形式について－

＜審査にあたっての課題＞

○申出全般として、公表形式が不明確、又は、研究内容の広範さと比して、明らかに一部しか示されていないような申出が多かった。個人の特定可能性が問題となるのは、公表形式であるため、今後の審査にあたっては、公表形式の明示についてより厳格に対応する必要があるのではないか。

○現在、第１回審査で承諾された申出について、順次データの提供作業を行っているが、特に医療機関の集計単位が3未満（２以下）となってはならない、とする部分については、地域の医療提供体制の研究にあっては柔軟に考える必要もあるのではないか。

（注）第６回の検討会において、「データ提供にあたっての一応の基準」として、医療機関の属性に関する情報を集計することにより、特定の集計単位に該当する医療機関が２以下となる場合には、公表不可との考え方を示している。一方で、地域の実情を特に勘案する必要がある場合には、例外もありうるとの議論もあった。

＜今後の対応＞

① 一旦、研究成果として公表されたものについては、それを目にした者がその公表された成果物とその他の様々な情報とを照合することについて制限を加えることができない。したがって公表形式の確認は厳格にする必要がある。

今後の審査にあたっては、公表形式の明示を求めるとともに、公表前の成果物の報告が確実に行われるようにする必要がある。

② 医療機関の集計単位が3未満となってはならない、との基準については、地域の医療提供体制の分析・調査を行う場合で、自治体への提供を目的としているなどにおいては、柔軟に運用する必要がある場合もあるのではないか。

ただし、個別の医療機関が明示されないことなど、こうした観点から公表形式を事前に報告を受け確認する運用を徹底する必要がある。

③　サンプリングデータセット作成の背景

第一回申出の結果

- 第一回申出では、合計で43例の申出のうち承諾となった申出は6例にとどまった。

- 審査方針のなかでも、「研究内容・抽出について」の項目②に該当するために不承諾となった事例が多くみられた。

- 一方、第一回申出では申出書提出後に申出内容について文書で申出者と事務局とで情報交換する機会があった。その過程で、セキュリティ要件に対する申出者の意識の高まりが認められた。結果として、審査方針のうち「セキュリティ要件について」の項目に照らし不承諾となった申出は、少数にとどまった。

第一回申出審査を踏まえて

抜粋資料

➢『「事前審査の論点を踏まえた、事務局審査方針(研究内容・抽出について②」』

同じく「個人の識別可能性を下げる」という原則に鑑み、**多数の項目を用いた探索的研究**や、SYの「傷病名コード」、SIの「診療行為コード」、IYの「医薬品コード」(DPCの場合にはBU, SB, CD)どれかひとつでも**「全て求める」**という要望の申出は、事務局審査では提供依頼について不承諾とした。

- 多変量解析、propensity score分析、重回帰分析など多くの項目を必要とする申出も**探索的研究**と考え、今回は不承諾とした。
- 何らかの抽出を経たあとのサンプルに対しても、上記のような申し出に対しては同様の判断とした。
- この方針は**特定健診**の情報を用いた研究にも適用した。
- 上記のように、研究手法によってはこうした抽出条件とならざるを得ないものもあり、これも今後検討予定の「基本データセット(仮称、後述)」構築によって対応することが考えられる。

➢『「今後の予定について」「基本データセット(仮称)の作成について」』

- 「傷病名」「診療行為」「医薬品」コードにおいて「全数希望」の申出が今回複数みられたが、個人が特定化される可能性、というリスクを考えると、データベースから全ての情報を提供することは困難である。
- 一方で、レセプト情報等の活用が最善でありかつ公共性の高い研究については、個人が特定されてしまう可能性のみを根拠として一概に研究の門戸を閉ざすべきではない、という意見もある。
- 次回以降の有識者会議で、ナショナルデータベースから一定の抽出を行い匿名性を高めた「基本データセット(仮称)」の構築について検討していくこととしたい。
- 「基本データセット(仮称)」の活用により、研究デザイン等の理由からやむをえず項目の全数が必要となる研究においても、個人が特定される可能性を可能な限り低めることができうると考えられる。

➢『「今後の予定について」「その他」』

- 申出者や研究者が、レセプト情報等でどのような研究ができるのか等の理解の一助となるよう、ダミー値のみで構成された架空のデータセットの作成についても、検討を行っていく。

➢探索的研究に対するニーズ

- 現在の審査方針では、「傷病名」「診療行為」「医薬品」コードのいずれかすべてを必要とする申出は、事務局審査方針「研究内容・抽出について②」に照らせば、承諾することができなくなる。
- 一方、**第一回申出で不承諾とした申出の複数**がこうした申出に該当しており、探索的研究に対する研究者のニーズが高いことをうかがわせる。こうした状況を考えると、探索的研究に対しても門戸を開くことができるよう、海外の先進事例も踏まえ、何らかの対応を検討する必要がある。

整備する基本データセットの類型

➢ **台湾でのデータセット提供の事例を参考として**

種別	概要	作成方法等
基本資料データ	特定項目の集計表情報	入院レセプト、外来レセプトの 特定項目を集計したもの
系統抽出データ	一定の割合で抽出をかけ、 匿名性を高めた月単位のデータセット	入院レセプトの5%を抽出 外来レセプトの0.2%を抽出
特定主題データ	疾患に応じ、母集団の構成等を 調整し抽出したデータセット	16種類の特定疾患等（悪性新生物等） に沿って抽出したデータファイル
ランダム抽出データ	まとまった数のレセプトを抽出し 一定期間紐付けたデータセット	2000年は20万人（1996年～2007年分）、 2005年は100万人分を抽出
教育用データ	研究者向けの練習用ツール	1000人分のランダムサンプリング 教育用に無償提供

前回の有識者会議における議論を踏まえれば、我が国において整備の緊急性および必要性が高いものは、上記のデータセット種別のうち

系統抽出データ　　**教育用データ**

となる。

基本データセット整備の概要

➢ 早期に整備するデータセット

✓ **サンプリングデータセット**（前頁の「系統抽出データ」）

- 探索的研究へのニーズに対応し、安全性に十分配慮したデータセットを、今後改善していくことを前提として試行的に提供する。
- 平成24年4月の予定としている次回の申出受付では、第一回申出で不承諾となった申出者に限定し、審査のうえで提供を決定する。

✓ **練習用データセット**

- 申出予定者や研究者が実際に格納されたレセプト情報等の把握に役立つよう、架空のデータセットとして作成する。

➢ その他のデータセット

- 24年度の厚生労働科学研究（「汎用性の高いレセプト基本データセット作成に関する研究（ 24010201）」）の成果などを踏まえ、整備を進めていく予定としている。

④　サンプリングデータセットの内容

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。
- 作成レセプトは **医科入院レセプト** **医科入院外レセプト** **DPCレセプト** **調剤レセプト**
- 「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）

- 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。

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。

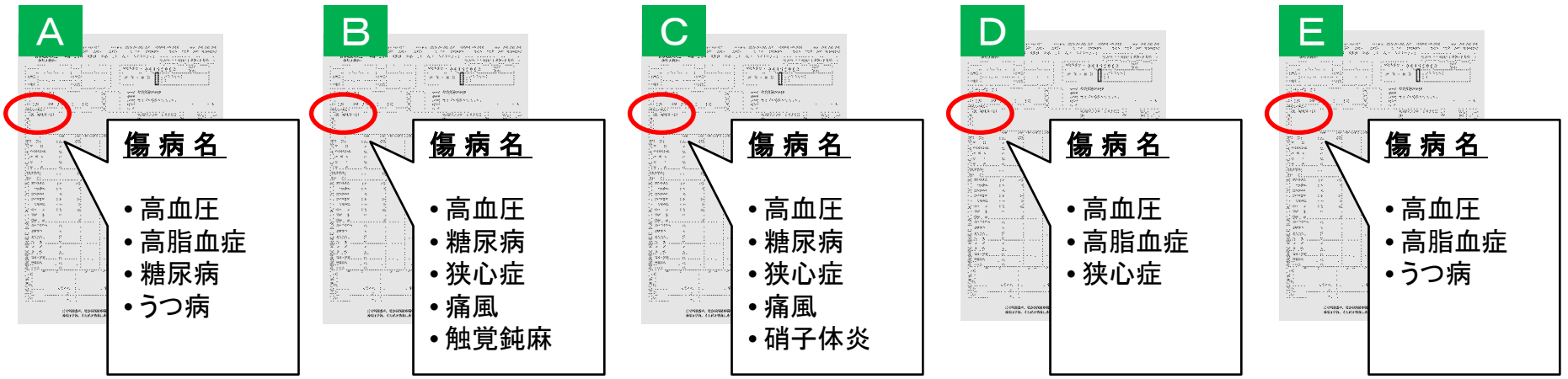
匿名化処理の例：傷病名

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**こととする。
- 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。

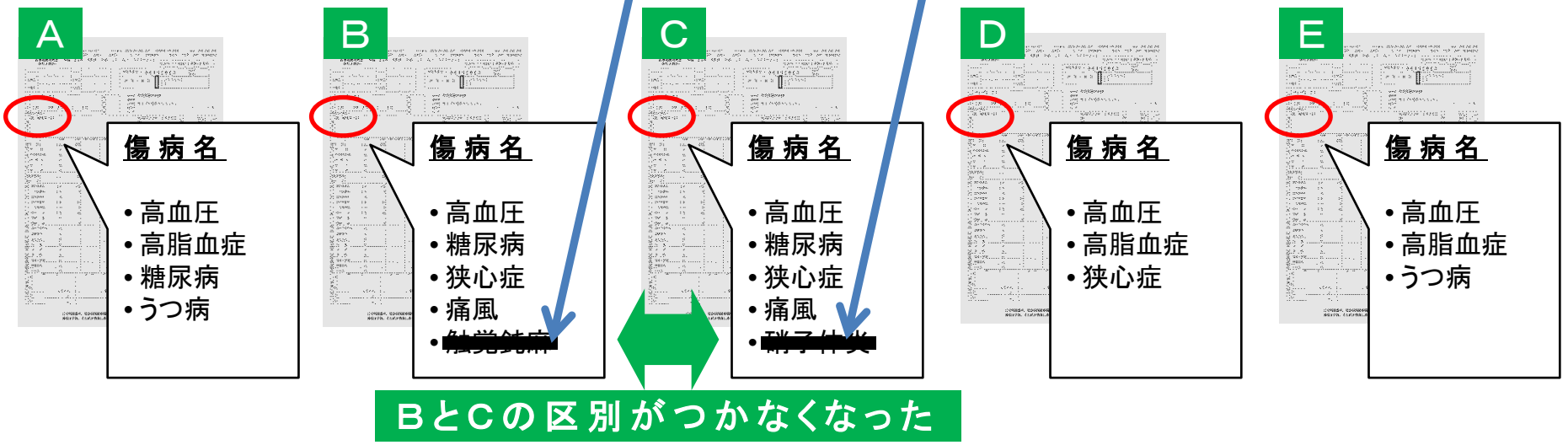


匿名化処理の例：傷病名

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

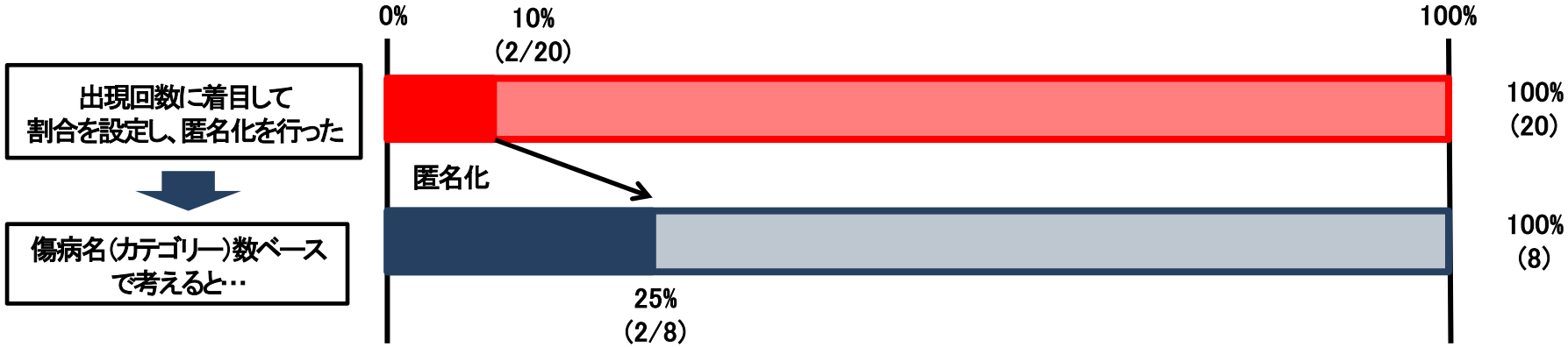


BとCの区別がつかなくなった

匿名化処理の例：傷病名

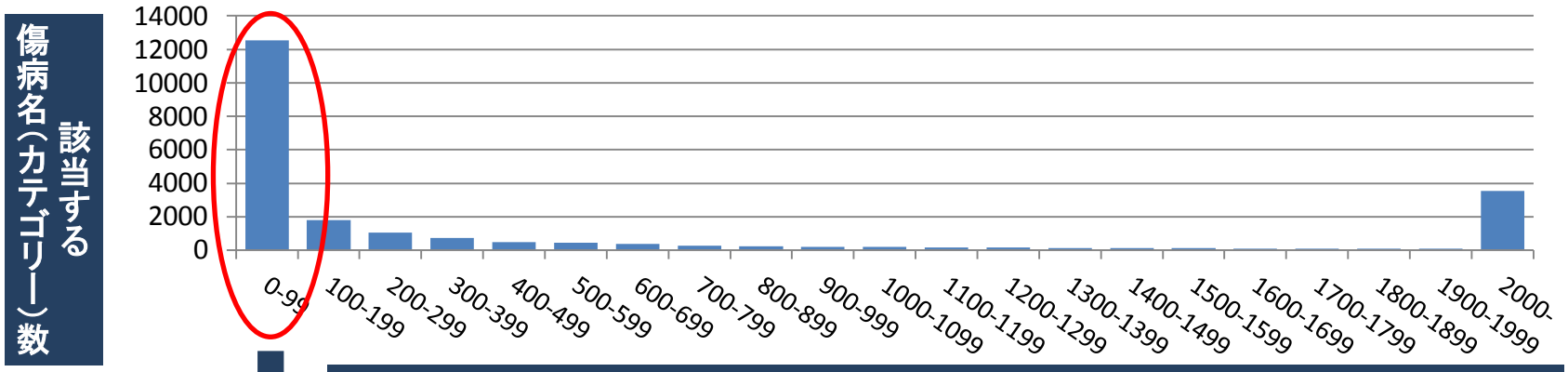
傷病名（カテゴリー）数とレセプト出現回数の関係

- 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。



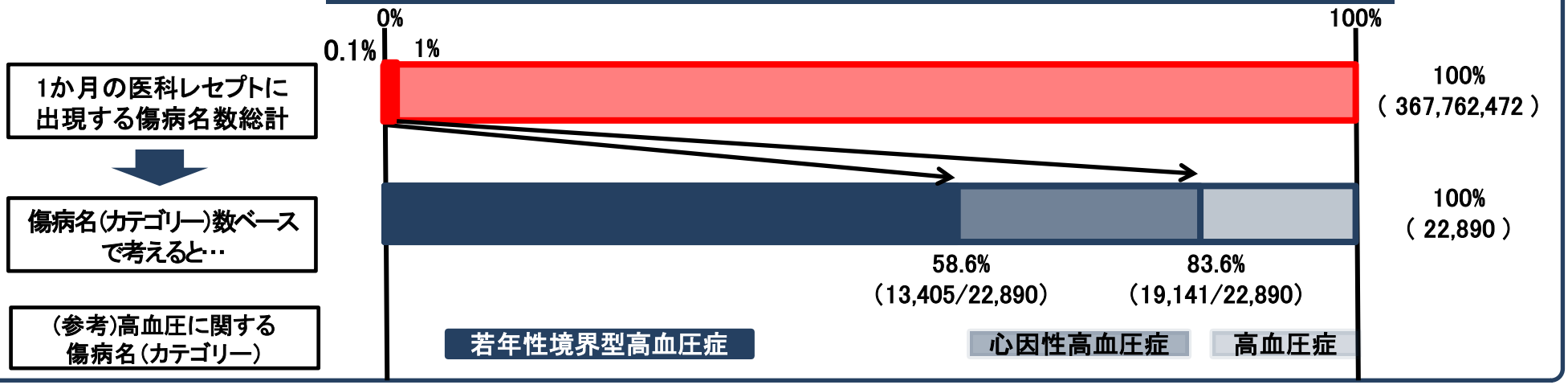
匿名化処理の例：傷病名

（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。



匿名化処理：医科診療行為マスター

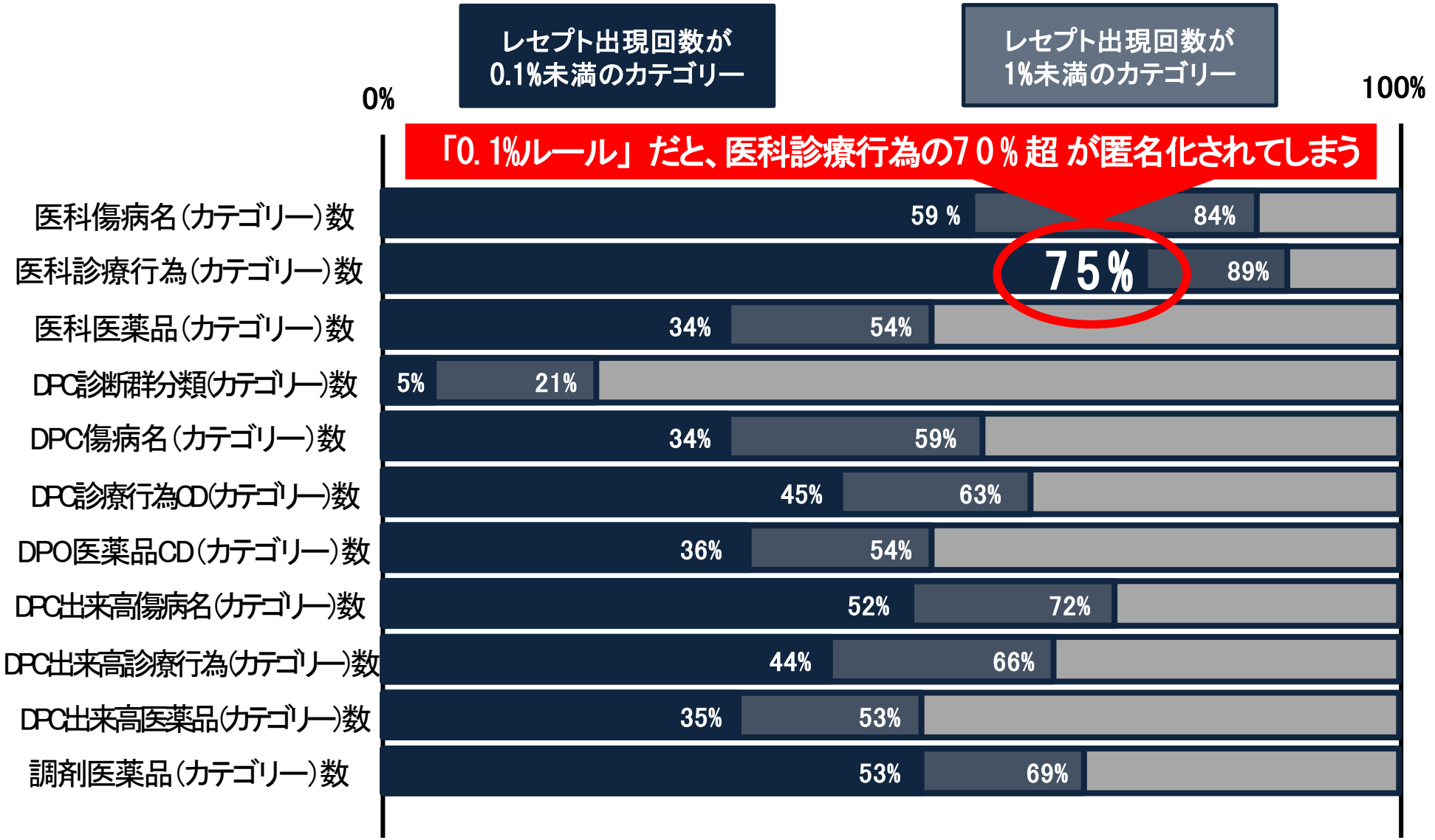
➢ 匿名化処理の基準：「医科診療行為マスター」における例外的な扱い

- 「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を行う。

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。

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

- 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。
- 以下に関連する項目は削除する。

保険者等に関する情報 | **医療機関等に関する情報** | **公費医療に関する情報** | **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%

保険者等、医療機関等、公費医療に関する情報の処理

➢ 保険者等に関する情報

- 保険者等に関する情報は、保険者それぞれに対する被保険者の割合が大きく異なっているため、抽出に際して母集団と構成比率が変化しないような処理を行わないこともあり、法別番号等の情報を削除することとしている。
- この原則に則り、保険者や被保険者に関する直接的な情報に加え、間接的に把握しうる情報についても今回のサンプリングデータセットでは削除する。具体的には、以下の項目が削除対象である。
  - 「レセプト共通レコード」（レコード識別情報「ＲＥ」）における「レセプト種別」、「給付割合」、「一部負担金・食事療養費・生活療養費標準負担額区分」、「レセプト特記事項」
  - 「保険者レコード」（レコード識別情報「ＨＯ」）における「保険者番号」、「被保険者証（手帳）等の記号」、「被保険者証（手帳）等の番号」、「職務上の理由」、「証明書番号」、「負担金額」

➢ 医療機関等に関する情報

- 病床数（「レセプト共通レコード」における項目「病床数」）に関する情報については、全てのレセプトに情報が含まれているわけではないものの、その情報から個別の医療機関を推定することが可能となる。このため、たとえば「～200」、「200～399」、「400～599」、「600～799」、「800～999」、「1000～」といったように病床情報をカテゴリー化したうえで提供する。
- 同様に、任意記載となってはいるが「レセプト共通レコード」における「カルテ番号等」も削除する。

➢ 公費医療に関する情報

- 公費医療等に関する情報についても、保険者に関する情報と同様、抽出に際して母集団と構成比率が変化しないような処理を行っておらず、また、その種類によっては非常に受給者数が少ないものも含まれる。このため、負担者や受給者に関する直接的な情報に加え、間接的に把握しうる情報についても今回のサンプリングデータセットでは削除する。具体的には、以下の項目が削除対象である。
  - 「公費レコード」（レコード識別情報「ＫＯ」）における「負担者番号」、「受給者番号」、「任意給付区分」、「負担金額」

⑤　サンプリングデータセットの提供形式

申出資格等について

1. 申出資格等について

- 第一回申出で複数の申出者が該当した審査方針「研究内容・抽出について②」は、基本データセットの審査においては許容されるため、申出の多数が承諾となる可能性が高い。申出資格を制限しない場合、現行のデータ提供等の体制を考慮すると、データ提供が大幅に遅れるなどの懸念がある。
- 一方、第一回審査で不承諾となった申出者においては、審査の過程で事務局とのコミュニケーションが図られており、審査方針やセキュリティ要件を具備する意義についても既に理解していただいていると考えられる。
  →こうした背景を踏まえ、例えば初回のサンプリングデータセットの申出資格としては、**第一回審査で不承諾となった申出者に限定する**。
  →なお、提供するデータの期間を単月に限定した場合、複数月、複数年のデータ利用を前提とした第一回申出の研究内容についてはそのまま実施することは不可能となる。この場合、**申出内容（研究方法）を必要に応じて変更したうえで**申し出ていただく。

2. 申出書内容・セキュリティ要件について

- 通常の「レセプト情報等の提供」と同様の申出書内容、セキュリティ要件を求めるかどうか
  →匿名性が高められたデータとはいえ個票データであることから、基本的には同様のセキュリティ要件とする。
  →申出書の様式は基本的に通常の「レセプト情報等の提供」を使用する。ただし、探索的研究を見据えた匿名性の高いデータセットであることから、申出書に記載される公表形式には、**通常の「レセプト情報等の提供」ほどの具体性および網羅性を求めない**。

申出資格等について

3. 提供・審査体制について

- 今回の「サンプリングデータセット」提供については、平成24年4月に予定している第二回申出と、受付時期をほぼ同一とする予定である。ただし、基本データセットの申出審査と第二回申出審査の時期については、「サンプリングデータセット」の審査および提供を先に行い、**両者の審査期間が重ならないようにする**。
- 第一回申出で不承諾となった申出者においては「第二回申出」と「サンプリングデータセット」の両方を申出することができる。但しデータ提供においてはサンプリングデータセットも個票として扱われるものである。このため、

  ① 原則としてこれら情報は他情報との照合を認めていないこと、また
  ② できるだけ多くの方々に使用されることによりその課題を明らかにする必要があること

  から、**申出者あるいは利用者がレセプト情報等を用いた複数の研究に同時に関与することは認めない**。
- この考え方をあてはめれば、「サンプリングデータセット」の提供を承諾された申出者・利用者は、このデータを用いた研究を終了するまで、レセプト情報等の第三者提供を活用した他の研究が不可能となる。
- この考え方は、第一回申出で承諾を受けた研究者にも適用する。ただし模擬申出についてはガイドライン作成以前に行ったものであるため、この考え方を適用しない。

⑥　その他

その他

➢ ホームページの活用

- レセプト情報・特定健診等情報提供に関して、厚生労働省保険局総務課では

  「レセプト情報・特定健診等情報提供に関するホームページ」
  （http://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/）

  を立ち上げており、このホームページを通じて情報提供を行っている。
- このホームページを活用し、今後は「練習用データセット」や、提供されるデータ構造の項目を整理した「レコードフォーマット」等についても、情報提供を行っていく予定である。

➢ フィードバックのお願い

- このサンプリングデータセットは、抽出の手法や削除される項目、匿名化処理について、個人が特定化される可能性をできるだけ引き下げることに重点を置いて作成したものである。
- サンプリングデータセットの提供を承諾された申出者の方々には、このデータセットを活用していただくとともに、サンプリングデータセットの今後の改善につなげていくためにも、問題となる点や改善すべき点について、事務局まで積極的にご意見をお寄せいただきたい。