

3 基本データセットについて


厚労科研研究班による基本データセットについて

【提案されたデータセットについて】

- 今回提案されたデータセットは、患者単位で、各年度毎あるいは1回の入院エピソードを単位として、医療費や診療日数等の限定的な情報のみのものから、診療行為等の情報を含むものまで、申出に応じて一定程度提供できるものである。
- 抽出率については、現在一律で、保険者レベルで約25%抽出を行った後、患者レベルで20%の抽出を行っているため、全体で約4-5%程度の患者数のみとなっている。
- データセットDにおいては、申出時の要望に応じてさまざまな情報を付与することが可能となっている。現在は、「傷病名」「診療行為」「医薬品」「特定器材」といった項目を予定しているが、その他の情報を付与することも可能である。

【現在準備中のデータセット】

- 今回提案された基本データセットは入院事例に絞ったデータセットであるが、それに加えて、現在外来事例に絞ったデータセット、ならびに特定の疾患について整備した疾患別データセット等を準備中である。

- 
- 基本データセットは、現在提供しているデータセット「サンプリングデータセット」とは異なり、「年度単位で紐付けされたレセプト情報」「入院エピソードごとの情報」といった特徴があり、その他の情報については相当程度削除したデータセットといえることから、今後順次利用者に提供していくこととしてはどうか。
 - 患者の抽出率や、データセットDにおける付与情報の上限数などといったデータの内容に関する事項については、今後利用者からの意見を集めながら、適宜変更を加えていくこととしてはどうか。
 - 外来版データセット、疾患別データセット等についても、準備が終了次第、提供を検討することとしてはどうか。

入院データセット B, C, Dの利用に当たってのセキュリティ環境について

- ✓ 入院データセット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 | 入院データセットB | 入院データセットC | 入院データセットD | |
| レセプトの 各種属性情報 | <ul style="list-style-type: none"> 性別、5歳刻み年齢階級別 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歳刻み年齢階級別 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） | サンプリングデータセットでの処理に加えて、 | | | |
| 傷病名 | <ul style="list-style-type: none"> 希少なものは匿名化 | <ul style="list-style-type: none"> 食事療養、生活療養の情報を全て削除 | <ul style="list-style-type: none"> 食事療養、生活療養の情報を一部削除 | <ul style="list-style-type: none"> 食事療養、生活療養の情報を全て削除 | |
| 診療・医薬品・ 特定器材コード | | <ul style="list-style-type: none"> 主傷病名1つのみ なし | <ul style="list-style-type: none"> 入院エピソードごとに、主傷病名1つのみ なし | <ul style="list-style-type: none"> 複数病名（申請者の要望により設定可） 設定可能、上限あり（全項目数で256未満） | |
| 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"> 入院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、入院エピソード毎に「回数」「点数」情報のみが提供され、日付情報は含まれない | |

- 入院データセット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。データセットDは、ある程度詳しい診療情報が設定可能であるが、入院エピソードごとにまとめられたデータであるために、依然として匿名性が高いと考えられる。
- 入院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。