平成 26 年度厚生労働省科学研究費補助金(循環器疾患･糖尿病等生活習慣病対策総合研究事業)
「追跡終了後コホート研究を用いた共通化データベース基盤整備とその活用に関する研究」
分担研究報告書

大規模コホートデータにおける一意性の検討

研究分担者　　祖父江友孝　　大阪大学大学院医学系研究科

研究要旨

個票データの開示を行う際には、一意性のあるデータは個人が同定される可能性があるので、一意性のあるデータがどの程度存在するかを検討しておく必要がある。今回、三府県コホートデータを用いて、標本数を変化させた場合にそれぞれどのような頻度で一意性が見られるかを比較した。100,629 **例全てを使用した場合と標本数を減らした場合、**複数の変数をそれぞれ組み合わせた場合の**分類数Kとユニークセル数S_1のパターンから、分類数の増加に伴い一意であるレコード数が増加するという分布の形状は似通っていたが、標本数が大きい場合ほど少ない分類数でユニークセルの割合が高率に達していた。コホートの規模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要がある。**

A. 目的

三府県コホートデータについて標本数を変化させた場合に、どのような頻度で一意性がみられるか比較し検討する。

B. 方法

三府県コホートデータを使用し 100,629 **例全てを使用した場合と、無作為抽出により標本数を** 1 **万、**1,000**、**100 **に変化させた場合について検討を行う。分析対象となる変数は昨年度と同じく、**226 **変数からなる個人レコードのうち、**ID **や数値化前データの変数、他と内容の重複する変数など** 22 **変数を除いた** 204 **変数とした。**

グループ化についても昨年度と同様に、**変数をその内容の近いもの同士で組み合わせてカテゴリ化し** 27 **のカテゴリを作成した。また、それらのカテゴリを内容から【個人特性】【追跡】【アンケート】の** 3 **グループに分けた。**

(1) 定義

対象（本研究の場合は三府県コホート 100,629 **例と、それより標本抽出された** 1 **万例、**1,000 **例、**100 **例）の個体が、数種類の変数の組み合わせに基づいていくつかのセルに分類されたとき、このとり得た分類数をKとする。さらに** 1 **つのセルに含まれる個体数がi**のセル数を$S_i(i = 1,2,\cdots,N)$**とする。このとき、$\sum S_i = K$となる。今回注目するのは個体数が** 1 **のセルの数であるユニークセル数S_1である。なお、個体自体を呼ぶときには一意という単語を用いるが、セルに対してはユニークセルという単語を用いる。**

(2) **検討内容**

[**検討** 1]

昨年同様、ベースとして【個人特性】と【追跡】のグループを考え、それらについて今後の解析に支障のないと考えられる範囲で可能な限りセルの併合（まるめの処理）を行う。今回は【個人特性】については昨年と同じ 2 パターンで変更なし、【追跡】については昨年の 4 パターンに新たに 2 パターンを追加した 6 パターンのサブグループを定義した。それらの分類数Kとユニークセル数S_1を求めた。

[検討 2]

100,629 例全てを使用した場合と、無作為抽出により標本数を 1 万、1,000、100 に変化させた場合について、21 のアンケートカテゴリに対し①アンケートカテゴリのみ、②【個人特性】とアンケートカテゴリをそれぞれ組み合わせた場合、③【追跡】とアンケートカテゴリをそれぞれ組み合わせた場合、④【個人特性】【追跡】の組み合わせに各アンケートカテゴリを組み合わせた場合、の全ての場合における分類数Kとユニークセル数S_1を求めた。

C. 結果

[検討 1]より、日付×転帰×死因からなる【追跡】グループでは、今回新たに検討した「追跡 4」（まるめの処理として ICD-9 コードを 17 の疾病大分類とする、かつ日付を月までにする）では分類数 5,229、ユニークセル数 2,083 であった。「追跡 5」（まるめの処理として ICD-9 コードを 17 の疾病大分類とする、かつ日付を追跡期間(単位：月)でみる）では分類数 1,593、ユニークセル数 439 であった。

昨年度の 4 パターンにこれらを追加したことで、最も大きいまるめの処理である「追跡 6」（昨年度の「追跡 4」に当たるもの：死因情報を除いて日付を追跡期間(単位：月)でみる）で一意性が消失するに至るまで、分類数とユニークセル数は漸減傾向を示した。（表 1）。

[検討 2] 100,629 例全てを使用した場合と、無作為抽出により標本数を 1 万、1,000、100 に変化させた場合について、①~④の組み合わせから得られた 461 パターンについて、分類数、ユニークセル数、分類数に占めるユニークセル数の割合S_1 / K、を示した（表 2）。

また 100,629 例全てを使用した場合と、無作為抽出により標本数を 1 万、1,000、100 に変化させた場合について、分類数Kを横軸、ユニークセル数S_1を縦軸にその分布を示した（図 1）。さらに、分類数Kを横軸、分類数に占めるユニークセル数の割合S_1 / Kを縦軸にその分布を示した（図 2）。100,629 例全てを使用した場合、分類数Kが増加するとともに、ユニークセル数S_1およびユニークセル数の割合S_1 / Kは増加するが、ユニークセル数の割合S_1 / Kについては、分類数Kが約 20,000 例になるまで急増し、次に 80%程度でプラトーに達し、分類数Kが 80,000 例あたりからさらに増加する、というパターンを示した。100,629 例全てを使用した場合と標本数を減らした場合を比較すると、分布の形状は似通っていたが、急増する部分の勾配が緩やかになり（100 例使用の場合は 40 例程度まで）、プラトーに達する部分が狭くなる傾向があった。

D. 考察

10 万人規模のコホート集団の場合、分類数が全対象者数の概ね 20,000 程度で、ユニークセルの割合が 80%に達していた。対象者数を少なくするにつれて、立ち上がりが緩やかになり、100 例規模のコホート集団では、分類数が 40 程度で、ユニークセルの割合が 80%に達していた。コホートの規模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要がある。

E. 結論

三府県コホートデータを用いて、いくつかの変数の組合せごとに一意性を検討した。10 万人規模のコ

ホート集団の場合、分類数が全対象者数の 20%程度で、ユニークセルの割合が 80%に達していた。100 例規模のコホート集団では、分類数が全対象者数の 40%程度で、ユニークセルの割合が 80%に達していた。コホートの規模にかかわらず、80%程度のレコードは一意性があるものとして対応する必要がある。

F. 健康機器情報
該当なし

G. **研究発表**
1. 論文発表
2. 学会発表
いずれもなし

H. 知的財産権の出願・登録状況
（予定を含む。）
1. 特許取得
2. 実用新案登録
3.その他
いずれもなし