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において、どのような頻度で一意性が見られるかを確認した。変数を 1 つずつ個別に見た場合の一意性は小さかったが、全変数を組み合わせた場合、一意であるレコード数は対象者の約 99.98%であった。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は急増した。また、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った。一意性は容易に避けられるものではなく、利用の際には一意性があるものと考えて対応することが必要と考えられた。

A. 目的

三府県コホートデータについて、どのような頻度で一意性がみられるか検討する。

B. 方法

三府県コホートデータを使用し 100,629 例全てについて検討を行う。各個人レコードは 226 変数からなるが、そのうち 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 変数を分析対象とした（表 1)。

検討に際し**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（表 2)。**

(1) 定義

対象の個体（本研究の場合は 100,629 例）が数種類の変数の組み合わせに基づいて K 個のセルに分けられたとき、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,…,N)とする。つまり、∑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。

検討内容
[検討 1]
204 変数それぞれ単変数についての、分類数 K とユニークセル数 S_1 を求めた。
[検討 2]
全体（204 変数すべてを組み合わせた場合）の分類数 K とユニークセル数 S_1 を求めた。
[検討 3]
ベースとして【個人特性】と【追跡】のグループを考える。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）を行い、【個人特性】については 2 パターン、【追跡】については 4 パターンのサブグループを定義し、それらの分類数 K とユニークセル数 S_1 を求めた。
[検討 4]
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。

さらに、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、前述の解析と同様、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。
[検討 5]
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、単体の変数で一意である個体が存在するのは、「v0502（10 年観察終了日）」「v0600（死因 ICD-9 コード 4 桁）」「v1200（身長（cm））」「v1201（体重（kg））」「v1610（初経年齢）」「v1612(自然閉経年齢)」「v1613(手術閉経年齢)」「v1615（出産人数）」「v1616（初産年齢）」「v2101（喫煙開始年齢）」「v2102（喫煙本数/日）」「v2103（禁煙年齢）」「v2801（転入何年前か）」「v2940（最も長く就いた仕事）」「v2950（従事年数）」の 15 変数であった。（表 2）

[検討 2]より、204 の全ての変数を組み合わせた場合に一意となる個体の数は 100,605 であった。

[検討 3]より、性別×年齢×居住地の情報からなる【個人特性】グループにおいて、まるめの処理を行わない「個人特性 1」では分類数 673、ユニークセル数 19 であったのに対し、年齢を 5 歳階級とし 85 歳以上はまとめた「個人特性 2」では、分類数 120、ユニークセル数は 0 と、一意性が消失した（表 3）。
追跡に関する日付×転帰×死因からなる【追跡】グループでは、処理を行わない「追跡 1」では分類数 20,176、ユニークセル数 16,631 であったのに対し、まるめの処理として、死因 ICD-9 コードを 3 桁までとする、かつ日付を月までにする（「追跡 2」）ことによりユニークセル数は約半分、同じく死因コード 3 桁かつ日付を追跡期間(単位：月)でみる（「追跡 3」）ことによりさらに半分になり、一意性は減少した。さらに死因情報を除いて日付を追跡期間(単位：月)で見た場合（「追跡 4」）では分類数が 243、ユニークセル数が 0 になり一意性が消失した（表 3）。

[検討 4] ①~④の組み合わせから得られた 329 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。また分

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）。分類数が小さい時には分類数に占めるユニークセル数の割合も80%以下に分布するが、分類数の増加とともにユニークセルの割合が急増し、概ね分類数が20,000を超えると80%以上に分布した。すなわち、100,629例全体に対して16,000例程度（16%程度）が一意性のある個体数となり、分類数の増加に比例して、一意性のある個体数が増加した。

[検討5] 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3）。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4）。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

各変数のユニークセル数の確認より、一意性には、変数v0501（10年観察終了日）のように、分類数が大きいことでそれぞれに振り分けられる個体数が少なくなるため生じる一意と、変数v1615（出産人数）において出産人数が20人というように、疫学的にまれな属性の個体が存在したために生じる一意の大きく2パターンが考えられた。前者に対しては例えば日付データを月までにするなどにより分類数を減らすことで一意性を減少させることが可能であり、後者に対しては一定値以上（以下）については直接表示せず、無限までの片側区間で表示するといった方法により一意性の減少が図られる。

しかしながら今回すべての変数を組み合わせた場合の一意である個体の数は100,605であり、これは全レコード数の約99.98%にあたる。このように大規模なコホートデータにおいては、変数が多くなる（質問項目が多い）ことによる一意性は容易に避けられるものではない。また、本研究に利用した10万人規模のデータであるからまるめ処理などによりある程度の一意性の減少がみられるが、規模が小さくなると一意性が上がる可能性も高い。

分類数とユニークセル数の関係から、コホートデータにおいて、変数が増えるほど分類数は増大し、概ね分類数が20,000を超えると一意である個体の数も分類数の80%以上に分布した。一意性を上げないためには、一つのファイルに含む項目数を増やさないよう、ファイルを分けて保管することなどが考えられるが、通常、一意性があるものとの前提で対応する必要がある。

死因に関しては、簡単分類を参考とした丸めの方法なども検討する必要がある。

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より、各変数、全変数あるいはいくつかの変数の組合せごとに一意性を検討した。三府県コホートデータのような 10 万人規模のデータの場合、分類数が概ね 20,000 を超えると一意性のある個体数は分類数の 80%以上となり、一意性があるものとの前提で対応を考える必要がある。

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

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