解析しやすい医薬データにするためのSASによるデータ加工

医薬データの加工で苦労した点
－問題提起－

京都大学　疫学研究情報管理学
大庭幸治

医薬データの加工で苦労した点

- 簡単な自己紹介と実例紹介
  - 日本動脈硬化縦断研究（JALS）
  - 今回の発表にあわせて実物とは異なる変数も加えています
- 実際に医薬データの加工で苦労した点
  - 高田さん・林さん・松丸さんへバトンタッチ

学生時代のお話

- 大橋 靖雄先生との出会い
  - 2003年～ 東京大学 生物統計学/疫学予防保健学
  - 卒業論文作成

- 日本動脈硬化縦断研究（JALS: Japan Arteriosclerosis Longitudinal Study）との出会い
  - 卒業論文のテーマ振り分け
  - 「日本全国からデータを集めるから、すごい研究ですよ！」
  - なんだか楽しそう

当時の私なんて…

- 学生
- プログラミング初心者
- SASは授業で聞いたことがあるくらい
  - 読んだことがあるのは、SASによるデータ解析入門くらい

- つまりは、今日の対象者

データをまとめてください

- データ収集、データクリーニング、変数の統一

- JALS0次研究
  - 循環器疾患を対象として過去に行われた17地域コホート・4職域コホートの疫学データ
  - 1985〜1999年までにベースラインの測定が行われ、日本各地で既に追跡が終了した65,435人のデータを個人単位で統合
  - 主要な循環器疾患のリスクファクターと総死亡・動脈硬化性疾患イベントとの関連について検討

データをまとめてください

- 基本的に自由なデータ形式で送付
  - 既に終了しているデータだったから
- 各コホートに以下の項目に関してデータ提供を依頼
  - ベースラインデータ
  - イベント・最終転帰データ
  - 繰り返し測定データ

ベースラインデータ

- 主要な背景因子・リスクファクターに関する情報が入ったデータ
- 各コホートで用いている形式で郵送
  - 別途、登録票という形で変数名やコーディングを調査
- 背景因子・リスクファクターの数は大体30～40項目程度
  - 性、年齢、臨床検査値、生活習慣…
  - 多いところでは100を超えるものも

ベースラインデータの例1

- コホートA

ID	性	年齢	検査日	総コレステロール	喫煙本数/日	…
1	1	56	1986/12/11	180	20	
2	2	61	1986/7/7	.	0	…
3				220	10	…

変数が日本語

数値（コード）で入力

ベースラインデータの例2

- コホートB

ID	SEX	AGE	checkdate	TC	SM_No	…
S_1	男	56	1985年2月11日	180	20	
S_2	女	61	1984年8月15日	.	0本	…
S_2			84年10月21日	220	10本	…

変数が英語

文字で入力

イベント・最終転帰データ

- 脳卒中・心筋梗塞などの心血管系イベントが発症したかどうかの情報、また対象者の最終転帰に関する情報が入ったデータ
- ID, イベント発症の有無, イベント発症日, 最終確認時点の状態 (生存 or 死亡), 最終生存確認日 (死亡日), 死因…
- 繰り返しイベントを発症する場合もあり

イベント・最終転帰データの例1

- コホートA

ID	脳卒中発症	脳卒中発症日	心筋梗塞発症	…	死亡	死亡日	死因
1	1	1990/12/11	1	…	1	1990/12/20	I63.1
2	0	イベントが横に並ぶ		…	0		
3	1	1995/6/21					

対象者が1行づつ縦に並ぶ

同じイベントが複数起きた場合
「脳卒中発症1」、「脳卒中発症2」
…となる

イベント・最終転帰データの例2

- コホートB

対象者がイベント数分、縦に並ぶ

ID	イベント	イベント発症日	…	死亡	死亡日	死因
1	脳梗塞	1990/12/11	…	1	1990/12/20	脳梗塞
1	心筋梗塞	1995/6/21	…	1	1990/12/20	脳梗塞
3	脳出血	1998/7/20	…	1	1999/10/10	肺癌

繰り返し測定データテーブル

- 繰り返し測定が可能な臨床検査値に関して、繰り返し測定分の情報が入ったデータ
  - 繰り返し測定が可能な臨床検査値、または対象者の生活習慣に関する変数など
  - IDに関しては必ず共通のものを使ってもらうようお願い

繰り返し測定データテーブルの例1

- コホートA

ID	検査日1	総コレステロール	収縮期血圧	…	検査日2	総コレステロール
1	1986/12/11	180	140	…	1987/12/11	175
2	1986/7/7	測定項目・測定日時が横に並ぶ				220
3	1986/6/21	220	120	…	.	.

対象者が1行づつ縦に並ぶ

繰り返し測定データテーブルの例2

- コホートB

対象者が測定回数分、縦に並ぶ

ID	検査日	総コレステロール	収縮期血圧	…
1	1986/12/11	180	140	…
1	1987/11/21	185	130	…
2	1986/7/7	200	180	…
3	1986/6/21	220	120	…

測定項目が横に並ぶ

データ整理の作業がとても大変

- 変数やコードがバラバラ
  - 変数名は日本語がほとんど
  - コードではなくテキストで入力されている場合もあった
- 統一するための変数名を考えるのも大変
  - その場で変数を考えるのにも限界が…
- 何か統一のフォーマットってないの？
  - どのデータベースでも標準化された変数やコードを利用していたらとても効率的（理想）

問題提起その1

- バラバラである変数名やコードをまとめる上で、参考とすべきモデルはないでしょうか？

- **CDISC (Clinical Data Interchange Standards Consortium)**
  - 医薬データの標準化
  - 松丸さん教えてください！

収集されたデータ形式のパターン

- データの形式がバラバラ
  - 扱う変数が多い場合、繰り返しデータが測定される場合に特に問題

- データ形式のパターンは、対象者について
  - 「項目」
  - 「時点」

  をどのような形式で入力しているかによって4つに大きく分けることができた

タイプ1

- 項目、時点ともに横に並んだデータ
  - 同一対象者のデータが、1つの行に全て入力されているデータ

ID	測定日1	項目A	項目B	…	測定日2	項目A	項目B	…
1	1986/12/11	180	0	…	1987/12/11	175	1	…
2	1986/7/7	200	0	…	.	.	.	…
3	1986/6/21	220	1	…	1987/7/21	200	1	…

タイプ2

- 項目が横に並び、時点が縦に並んだデータ
  - 同一対象者のデータが、繰り返し測定分、縦に積まれたデータ

ID	測定日	項目A	項目B	…
1	1986/12/11	180	0	…
1	1987/11/21	185	1	…
2	1986/7/7	200	0	…
3	1986/6/21	220	1	…
3	1987/7/21	200	1	…
…	…	…	…	…

タイプ3

- 項目が縦に並び、時点が横に並んだデータ

ID	項目名	測定内容1	測定日1	測定内容2	測定日2	…
1	A	180	1986/12/11	185	1987/11/21	…
2	A	200	1986/7/7	.	.	…
3	A	220	1986/6/21	200	1987/7/21	
…	…	…	…	…	…	…
1	B	0	1986/12/11	1	1987/11/21	…
2	B	0	1986/7/7	.	.	…
3	B	1	1986/6/21	1	1987/7/21	

タイプ4

- 項目、時点ともに縦に並んだデータ

ID	項目名	測定内容	測定日
1	A	180	1986/12/11
1	A	185	1987/11/21
2	A	200	1986/7/7
3	A	220	1986/6/21
3	A	200	1987/7/21
…	…	…	…

1	B	0	1986/12/11
1	B	1	1987/11/21
2	B	0	1986/7/7
3	B	1	1986/6/21
3	B	1	1987/7/21
…			

解析をお願いします

- データ整理およびクリーニングが無事終わると、解析作業に移る

- 解析の目的の応じて、データを加工する必要性
  - どのように加工すればよいのかも良く分からない
  - おおもとのデータテーブルでの単純集計の結果と、解析用に加工したデータテーブルでの単純集計の結果が異なったりすることもしばしば

主に行っていた解析事項

- ベースラインデータの検討
  - 単純集計
  - コホート間でのベースラインデータの違いを検討
- 総死亡・動脈硬化性疾患などのイベントとリスク因子との関連
  - ベースライン（性・年齢）別の発症率の検討
  - 多数の交絡要因を調整した解析
    - 層別解析・回帰分析
  - リスク因子の経時的な変動を考慮した解析
    - 年齢も疫学研究では大きなリスク因子

加工するにしても…

- SASプログラムの書き方など良く知らない
  - データテーブルの結合（set, merge）は何とか…
  - 残りはほとんど if文で対応

問題提起その2

- 解析の目的に見合ったデータ形式に加工する上で、参考となるようなプログラム例ってどこかにないでしょうか？
  - 実際、ほとんどない
- 加工プログラム例1：少数ステップでの処理
  - 高田さん教えてください！
- 加工プログラム例2： SASマクロを用いた処理
  - 林さん教えてください！