ネステッドケースコントロール研究・ケースコホート研究のデザインと統計解析

野間 久史
情報・システム研究機構　統計数理研究所
2017年1月25日
第27回日本疫学会学術総会
疫学セミナー「追跡データ分析のA to Z」
e-mail: noma@ism.ac.jp
URL: http://normanh.skr.jp/

1

MRFIT試験

- Multiple Risk Factor Intervention Trial
- 1970~80年代 米国
- 冠動脈心疾患の予防プログラムをランダムに割り付けしたランダム化介入試験
- 対象者　12,866人，追跡期間　7年
- 研究にかかった費用は、1億ドル以上（1970~80年代当時）
  - すべての参加者に対して行った、栄養調査・血清サンプルの分析に、膨大なコストと労力がかかったといわれている

MRFIT Research Group (1982), Prentice (1986)　2

CHD Mortality

- Primary Endpoint: CHD Death
- 追跡終了時点（1982年2月）
  - 介入群　115/6,428 (1.8%)
  - 対照群　124/6,438 (1.9%)
  - Total　239/12,866 (1.9%)
- Primary Endpointが観測されたのは、全体の2%のみ！！
- 慢性疾患の疫学研究では、イベントの発生頻度はそれほど大きくないのが一般的である

3

統計解析における問題

- ログランク検定
- Coxの比例ハザード回帰モデル
  - 検出力に寄与するのは、イベントを発生した参加者が相対的に大きな割合を占め、打ち切りとなった参加者の寄与率はそれほど大きくない
- MRFIT試験では
  - 98%近くの対象者にはイベントは観測されず
  - 栄養調査・血清サンプル分析にかかった膨大なコストの大部分は、相対的に検出力に寄与していない（大きな効率の損失）

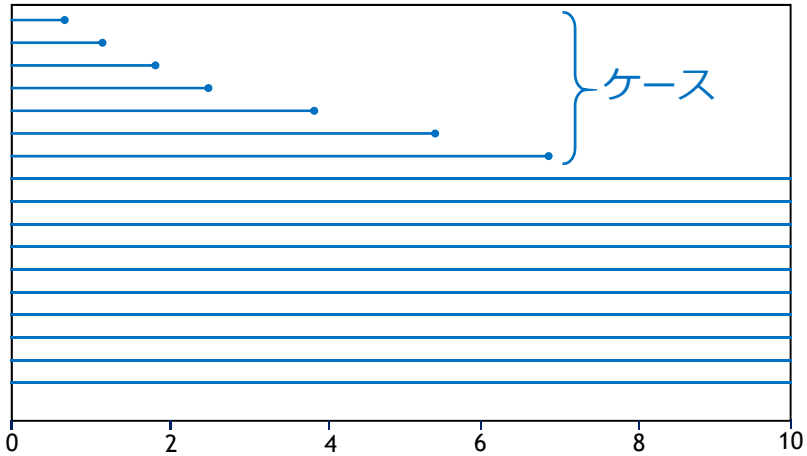
4

効率的な研究デザイン

- Nested Case-Control (NCC) 研究
- Case-Cohort 研究
  - 特に、測定コストの大きな共変量（栄養調査，バイオマーカー・遺伝子情報分析など）の測定コストの節減を目的として開発された研究デザイン
  - データを測定するのは、コホートにおける一部の対象者でよい
  - 統計的な精度・検出力を保持しつつ、研究のコスト・労力を大幅に節減することができる

5

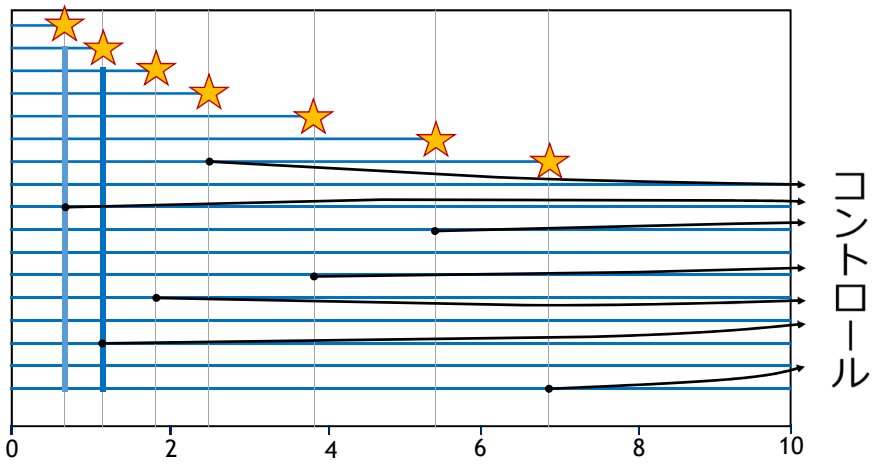
Nested Case-Control 研究



コホート研究の中で、ケースコントロール研究を行う

6

Risk Set Sampling



イベント発生時点ごとにコントロールをマッチング

7

MRFIT試験でのNCC研究

- 1990年代 追跡終了後：MRFIT試験の中で、実際にNCC研究が行われている
- 冷凍保存していた血清サンプルを分析し、C-Reactive Protein を測定
- NCCを用いることで、全員分の血清サンプル分析を行う必要はなく、コストは大幅に節減された
- ケースと選択された一部のコントロールのみ、データを測定すればよい

Kuller et al. (1995)

8

C-Reactive ProteinとCHDの関連の評価

- MRFIT試験でのRisk-set sampling
  - Cases: CHD死亡（148人）
  - Controls: 施設，介入群などでマッチング（1:2マッチング; 296人）

測定のコスト・労力は 444/12,866 に！

Kuller et al. (1995) 9

統計解析の方法

- Cox回帰モデル
  - $h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$
- 当然、一部の対象者しかサンプリングされないので、通常の部分尤度によるハザード比の推定はできない
- Risk Set Samplingは、コホートの履歴の中で、時点ごとにケースコントロールサンプリングを行うサンプリング方法であった
- 時点ごとのケース・コントロールの組を、Matched Case-Controlサンプルと見なして、層別解析すればよい

10

ハザード比の推定量

- ロジスティック回帰の条件付き尤度

$$\mathcal{L}(\beta) = \prod_j \left[\frac{\exp(\beta x_j)}{\sum_{k \in C_j} \exp(\beta x_k)} \right]$$

  - C_j: ケース，コントロールの組
- β の推定量は、対数ハザード比の一致推定量となる
- Matched Case-Control Studiesの解析コードをそのまま適用して、条件付きロジスティック回帰モデルで解析ができる

Thomas (1977), Goldstein and Langholz (1992) 11

MRFIT試験の統計解析

- 条件付きロジスティック回帰モデル
  - Risk Set Samplingでの時点ごとのケース・コントロールに結果変数 {1, 0} を割り付け
  - 年齢，喫煙本数（/Day），拡張期血圧，トリグリセリド、HDL/LDLコレステロールを交絡要因として調整
- 追跡期間中、早い時点で一度コントロールとしてサンプリングされた参加者が、後からイベントを起こし、ケースとしてサンプリングされることもあるが、重複サンプリングは無視する（あくまでも Matched Case Control 研究として解析する）

Kuller et al. (1995) 12

MRFIT試験の解析結果

コホートのHazard Ratioに一致！

Quartile of CRP (mg/l)	Cases (No., %)	Controls (No., %)	OR (95%CI)	Score Test
1 (0.2-1.2)	26 (18%)	94 (32%)	1.0 (reference)	
2 (1.3-1.9)	28 (19%)	66 (22%)	1.6 (0.8-3.1)	
3 (2.0-3.2)	47 (32%)	69 (23%)	2.7 (1.4-5.2)	
4 (3.3-79.0)	47 (32%)	67 (23%)	2.8 (1.4-5.4)	< 0.001
Total	148	296		

Kuller et al. (1995)

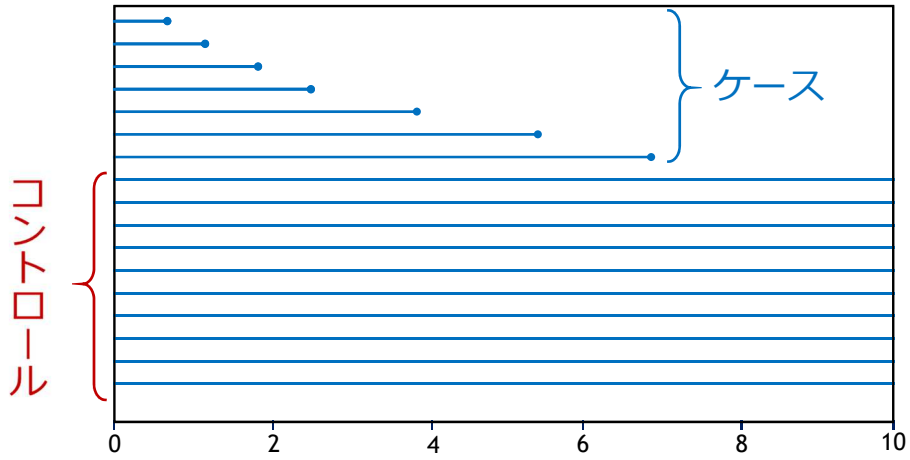
13

NCC研究の限界

- MRFIT試験のような大規模なコホート研究では、研究者が関心のあるアウトカムは必ずしも1つではない
- せっかく多大な時間と労力をかけて行う追跡調査なので、可能な限り、多くの情報を得たいというのが人情
- NCC研究では、複数のアウトカムに関心がある場合、アウトカムごとにケース・コントロールをとることに
- 研究のコストは倍々算で増えることに！！
- すべての参加者に高価な測定を行うよりは効率的であるが、もっとうまいデザインはないか？？

14

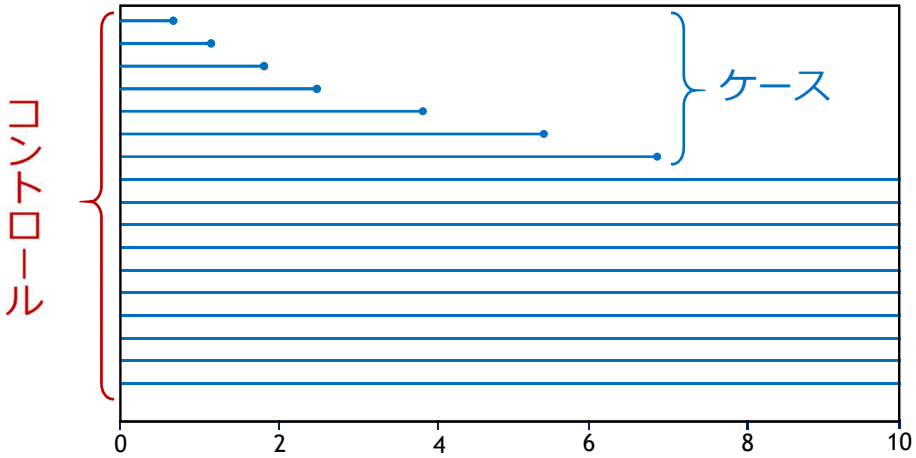
古典的なCase-Control研究



コントロールはイベントを起こしていない対象者

15

思い切って、コントロールを変えてみよう



コントロールをケースも含めた全集団からサンプリング？？

16

Case-Cohort研究

- Nested Case-Control研究の “Multiple Outcomes” の問題点を克服するために考案された
- 「コントロールを、ケースも含めた全集団からサンプリングしたケースコントロール研究」というデザイン
  - サブコホート（Subcohort）
- コントロール集団は、イベントの種類に依存しないため、複数のアウトカムについての解析を行いたいという場合にも、共通のコントロールとして利用できる

Prentice (1986) 17

生存時間解析

- Coxの比例ハザード回帰モデル

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\beta_1 x_1 + \cdots + \beta_p x_p)$$
$$= h_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})$$

- NCC研究と同じく、Case-Cohort Samplesは、コホートからのランダムサンプルとは見なせないので、通常の部分尤度に基づく推測では、妥当なハザード比の推定量が得られない

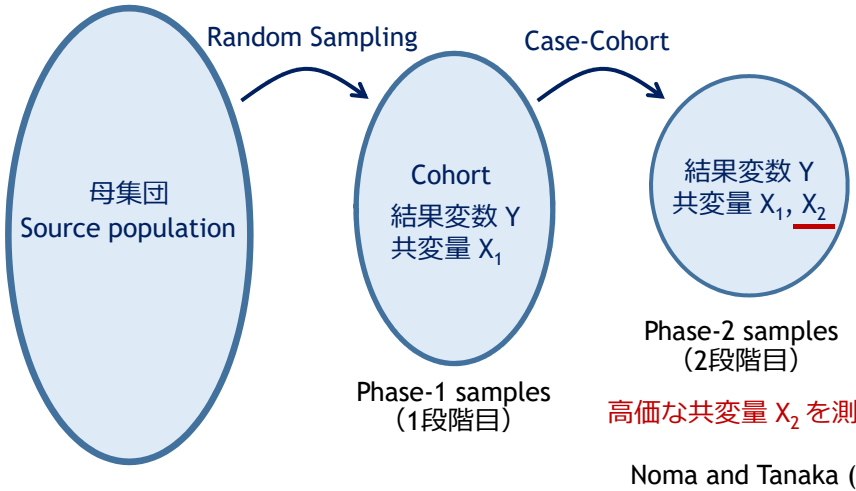
Cox (1972) 18

2段階サンプリングとしての定式化

- ケースコホート研究は、コホート研究の中で、ケースコントロール研究を行うというデザイン
- 対象となったコホートが、母集団（Source Population）からのサンプリングで得られたサンプルで、そこから得られるCase-Cohort Samplesは、母集団を起点に考えると、2段階のサンプリングを経て得られたサンプルと見なすことができる

Zhao and Lipsitz (1992)　19

2段階サンプリングモデル



Noma and Tanaka (2016)　20

不完全データとしての定式化

- Phase-I Cohortの対象者集団を解析対象集団とすると？
- Case-Cohort Samplesに選ばれた対象者
  - (Y, X_1, X_2) がすべて観測されている
- Case-Cohort Samplesに選ばれなかった対象者
  - (Y, X_1) が観測されている
  - X_2は観測されていない（＝欠測と見なすことができる）
- 共変量 X_2 が部分的に欠測した不完全データとなる！！
- 不完全データの解析手法をそのまま適用することで、ハザード比の妥当な推定ができる！！

21

IPW法

- Inverse Probability Weighting (IPW)法
  - 観測確率の逆数で、推定関数を重みづけ
  - MARのメカニズムのもとで、一致推定量が得られる
- Phase-I Cohortのすべての対象者の X_2 の観測（欠測）確率はサンプリングデザインによって規定されるため、完全に既知
- 一般的な臨床研究で生じる欠測は、欠測確率が未知なので、その推定自体が難しい問題となるのだが、ケースコホート研究の応用では、欠測確率の真値が既知という前提のもとで解析を行うことができる！！

Robins et al. (1994) 22

Cox回帰モデルの部分尤度

$$\mathcal{L}(\beta) = \prod_i \frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j}}$$

- R_i: 時点 i でのRisk Set

ハザード比 $\boldsymbol{\beta}$ のバイアスのない推定量は、個々人のデータがそれぞれ等しく部分尤度に寄与することによって得られる
（標本平均 $\bar{x} = \sum_{i=1}^{n} x_i/n$ は最尤推定量の一種であるが、推定量に対する個々人の寄与率は、全員「1」で等しい）

ケースコホート研究は、ランダムサンプリングの仮定が成り立たないため、等しい寄与率ではバイアスが生じる！！

Cox (1972) 23

IPW法による修正部分尤度

$$\mathcal{L}_{IPW}(\beta) = \prod_i \frac{\omega_i e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i}}{\sum_{j\in R_i} \omega_j e^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j}}$$

- R_i: 時点 i でのRisk Set
- ω_i = N_1/n_1 (for cases, N_1:total case number, n_1:number of selected cases),
  = N/n_0 (for non-cases, N:cohort size, n_0:subcohort size)

Phase-2 Samplingのサンプリング割合の逆数で寄与率を重みづけした部分尤度によって妥当な推定量が得られる

Borgan et al. (2000) 24

層別サンプリング

- 重要な共変量の分布の偏りを防ぐために、共変量で層別をして、サブコホートのサンプリング確率を調整するデザイン
  - 層別サンプリング（Stratified Sampling）
- 関心のある変数 X_2 と相関の強い代替変数で層別を行うと…
  - IPW法による解析で、X_2のハザード比の推定精度が上がる !!
  - デザインの段階で層別をしていなくても、事後的に層別をしてIPW解析をしても推定精度が上がる
- 統計解析では、IPW法の重みを、層ごとのサンプリング割合の逆数に変更するだけでよい

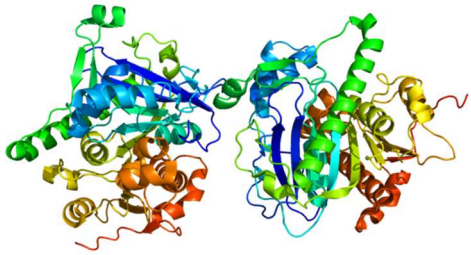
Borgan et al. (2000) 25

ARIC Study

- The Atherosclerosis Risk in Communities Study
- 1980-90年代、米国で行われた地域ベースのコホート研究
- CHD，脳卒中など複数のアウトカムを対象としたケースコホート研究で、多数のリスク要因の評価を効率的に行っている
- 高価な測定を要するリスク要因
  - 遺伝子多型，炎症バイオマーカー
- 母体となるコホートは、ARIC Cohort 1つだけであるが、Case-Cohort研究を有効に活用し、数十報の研究論文を公表している

Ballantyne et al. (2004), Breslow et al. (2009) 26

リポタンパク質関連ホスホリパーゼA2（Lp-PLA_2）



http://en.wikipedia.org/wiki/File:Protein_PLA2G7_PDB_3D59.png

- CHD，脳卒中などのリスク要因で、心疾患等のバイオマーカーにも利用される

The Lp-PLA_2 Studies Collaboration (2010)　27

ARIC Study Cohort

	Non-CHD Cases								CHD Cases	Totals
	Black				White					
	Female		Male		Female		Male			
	Age< 55	Age≥55	Age< 55	Age≥55	Age< 55	Age≥55	Age< 55	Age≥55		
Whole Cohort	1,133	719	598	393	2,782	2,213	1,959	1,818	730	12,345

層ごとのNon-Casesの分布はアンバランス

⇒ すべての層で均一にサンプリングを行うのは非効率であり、層別サンプリングを行っている

Breslow et al. (2009)　28

IPW法の重みの計算

	Non-CHD Cases								CHD Cases	Totals
	Black				White					
	Female		Male		Female		Male			
	Age< 55	Age≥55	Age< 55	Age≥55	Age< 55	Age≥55	Age< 55	Age≥55		
Whole Cohort	1,133	719	598	393	2,782	2,213	1,959	1,818	730	12,345
サンプル	59	54	42	71	88	154	117	147	604	1,334
割合	5.2%	7.5%	7.0%	18.1%	3.2%	7.0%	6.0%	8.1%	82.7%	10.8%
重み	19.2	13.3	14.2	5.5	31.6	14.4	16.7	12.4	1.2	

Breslow et al. (2009)

29

IPW法による解析の結果

	ハザード比	95%信頼区間	P-value
Age in years/10	1.533	(1.240, 1.868)	< 0.001
Male sex	2.143	(1.672, 2.746)	< 0.001
White race	1.038	(0.799, 1.347)	0.781
Former smoker	0.656	(0.483, 0.892)	0.007
Never smoker	0.576	(0.419, 0.792)	0.001
SBP/100	4.730	(2.440, 9.172)	< 0.001
LDL-C/100	2.175	(1.515, 3.122)	< 0.001
HDL-C/100	0.079	(0.029, 0.215)	< 0.001
Diabetes	1.772	(1.303, 2.409)	< 0.001
Lp-PLA_2 0.310-	1.053	(0.759, 1.462)	0.756
Lp-PLA_2 0.422-	1.177	(0.846, 1.637)	0.333

Breslow et al. (2009)

30

R プログラムについて

- NCC研究でも、同様に2段階デザインとしての定式化が可能であり、不完全データの枠組みのもとでのIPW解析が可能
- IPW解析のほうが、条件付きロジスティック解析分析よりも、推定精度（検出力）は高い
  - R package multipleNCCで実装できる
- ケースコホート研究のIPW解析は、R package survivalで実装することができる（他にも、Prentice法，Prentice-Self法などのプロシジャも利用可能）
- 詳細については、添付資料をご参照ください

31

参考文献

- Ballantyne, C. M., Hoogeveen, R. C., Bang, H., et al. (2004). Lipoprotein-associated phospholipase A2, high-sensitivity C-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study. Circulation 109, 837-842.
- Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. Journal of Clinical Epidemiology 52, 1165-1172.
- Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, D. R., and Pogoda, J. (2000). Exposure stratified case-cohort designs. Lifetime Data Analysis 6, 39-58.
- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009). Using the whole cohort in the analysis of case-cohort data. American Journal of Epidemiology 169, 1398-1405.
- Cox, D. R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 34, 187-220.

32

- Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. Annals of Statistics 20, 1903-1928.
- Kuller, L. H., Tracy, R. P., Shaten, J., et al. (1996). Relation of C-reactive protein and coronary heart disease in the MRFIT nested case-control study. American Journal of Epidemiology 144: 537-547.
- The Lp-PLA2 Studies Collaboration. Lipoprotein-associated phospholipase A2 and risk of coronary disease, stroke, and mortality: collaborative analysis of 32 prospective studies. The Lancet 2010;375:1536-1544.
- MRFIT Research Group: Multiple risk factor intervention trial; risk factor changes and mortality results. JAMA 1982;248:1465-1477.
- Noma, H., and Tanaka, S. (2016). Analysis of case-cohort designs with binary outcomes: Improving the efficiency using whole cohort auxiliary information. Statistical Methods in Medical Research, DOI: 10.1177/0962280214556175.

33

- Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73, 1-11.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression-coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846-866.
- Thomas, D. C. (1977). Addendum to a paper by F. D. K. Liddel, J. C. McDolad and D. C. Thomas. Journal of the Royal Statistical Society, Series A 140, 483-485.
- Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. Statistics in Medicine 11, 769-782.

34