

短期集中！臨床疫学の研究デザイン講座： ネステッドケースコントロール研究と ケースコホート研究

野間 久史

統計数理研究所

2022年11月6日

日本臨床疫学会 第5回 年次学術大会

e-mail: noma@ism.ac.jp

URL: <https://www.ism.ac.jp/~noma/>

1

慢性疾患の疫学研究

- ▶ 十分な精度・検出力を達成するために必要なサンプルサイズはしばしば大きくなってしまう
 - ▶ 特に、調査・実験のコストや労力は大きい
 - ▶ 血清バイオマーカーの分析
 - ▶ 食物摂取頻度調査
 - ▶ 遺伝子情報の解析
 - ▶ 組織学的標本の分析
- などなど

2

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) 3

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！！
- ▶ 慢性疾患を対象とした臨床研究・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

4

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！）

5

本講座では

- ▶ 臨床研究・疫学研究において、測定のコスト・労力の問題は、しばしば悩ましい問題になる
- ▶ 仮に、イベントを起こした少数の参加者（MRFITでは2%）を中心に、コホート内のごく少数の参加者についてだけ、測定を行うだけで、全員に測定を行った場合とほとんど同じ解析結果が得られる方法があるとしたら、いかがでしょうか？
- ▶ 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は、そのようなことを実現する研究デザインである
- ▶ 本講座では、それらの研究デザインと統計解析の方法について、解説を行います

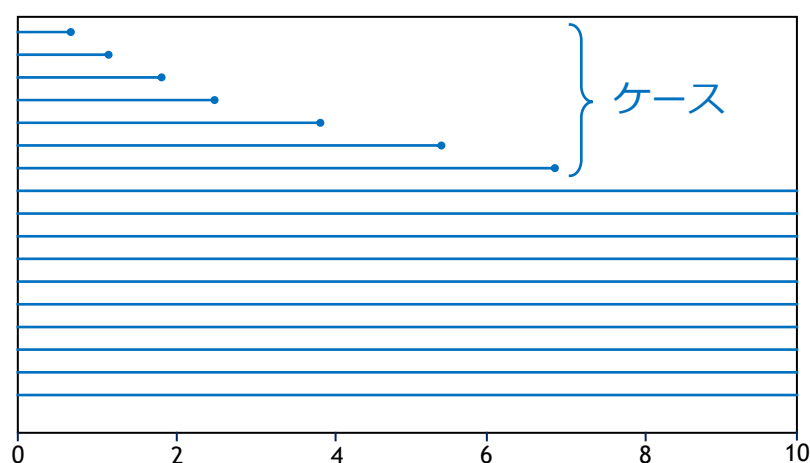
6

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

7

コホート研究



特定の集団（コホート）の追跡履歴情報を分析する

8

コホート研究

- ▶ 特定の集団（コホート）の追跡履歴情報をもとに、関心のあ
る要因と疾病の発生状況の関連を調べる研究デザイン
- ▶ 前向き（prospective）コホート研究
 - ▶ 「原因→結果」の時間的・因果的順序に沿って、「曝露-疾
病」の関連を評価するデザイン
- ▶ 後ろ向き（retrospective）コホート研究
 - ▶ データベースなどの記録に基づいて、時間的順序としては
「後ろ向き」にコホート研究を再現し、同様の分析を行う
研究デザイン

9



Atherosclerosis Risk in Communities Study Description

This website is intended for ARIC investigators, researchers, participants, and the scientific community. The [ARIC Participant](#) website is maintained for the ARIC participants, health professionals, and the general public.

The Atherosclerosis Risk in Communities Study (ARIC), sponsored by the [National Heart, Lung, and Blood Institute \(NHLBI\)](#) is a prospective epidemiologic study conducted in four U.S. communities. ARIC is designed to investigate the causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and date. To date, the ARIC study has published over 2,000 articles in peer-reviewed journals. ARIC includes two parts: the [Cohort Component](#) and the [Community Surveillance Component](#).

The Cohort Component began in 1987, and each [ARIC field center](#) randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community, to receive extensive examinations, including medical, social, and demographic data. Follow-up also occurs semi-annually, by telephone, to maintain contact and to assess health status of the cohort.

In the Community Surveillance Component, the four communities are investigated to determine the long term trends in hospitalized myocardial infarction (MI) and coronary heart disease (CHD) deaths in approximately 470,000 men and women aged 35-84 years.

Objectives of the Study

1. Examine the ARIC cohort to characterize heart failure stages in the community, identify genetic and environmental factors leading to ventricular dysfunction and vascular stiffness, and assess longitudinal changes in pulmonary function and identify determinants of its decline.
2. Cohort follow-up for cardiovascular events, including CHD, heart failure, stroke, and atrial fibrillation; and for the study of risk factors related to progression of subclinical to clinical CVD.
3. Enhance the ARIC cohort study with cardiovascular outcomes research to assess quality and outcomes of medical care for heart failure and heart failure risk factors.
4. Community surveillance to monitor long-term trends in hospitalized MI, CHD deaths, and heart failure (inpatient and outpatient).
5. Provide a platform for ancillary studies, training for new investigators, and data sharing.

[ARIC Participating Institutions](#)

[ARIC Study Design](#)

[More . . .](#)

Participants	▶
About ARIC	▶
Study Organization	▶
Obtain/Submit Data	▶
Training	
Cohort	▶
Community Surveillance	▶
Ancillary Studies	▶
Publications	▶
Directory	
Contact the CC	
Site Map	

Download Acrobat Reader® [here](#)

Username *

Password *

<https://sites.csc.unc.edu/aric/desc>

10

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
- ▶ 前向きなフォローアップによって、CHD, 脳卒中などのアウトカムについての情報が集められた
- ▶ さまざまなリスク要因の評価も行われており、高価な測定を要する変数の評価も行われた
 - ▶ 遺伝子多型, 炎症バイオマーカー
- ▶ ケースコホート研究が有効に活用された有名なコホート研究の事例

Ballantyne et al. (2004), Breslow et al. (2009) 11

薬害サリドマイド事件

- ▶ 1957年 西ドイツで鎮静・催眠薬として開発
 - ▶ 動物実験で副作用が認められなかったことから、「妊婦や小児が安心して飲める薬」として世界中で使われるように
- ▶ 世界中で、妊婦のサリドマイド服用による子どもの四肢欠損症（サリドマイド胎芽症）が報告される



12

1962年 Lenz警告

母親のサリ ドマイド服 用	奇形の有無	
	あり	なし
あり	90	2
なし	22	186
合計	112	188

ケース：奇形のあった子ども，
コントロール：奇形のなかった子ども

13

ケースコントロール研究

- ▶ 新生児奇形のように稀なイベントが対象
 - ▶ コホート研究による調査では、膨大なサンプル数と長期間の追跡が必要となる
 - ▶ 非効率的、しばしば実現不可能
- ▶ 「イベントを起こした対象者」「イベントを起こさなかった対象者」から、曝露効果を推定することができれば、稀な疾病であっても、少数のサンプルで効率よく調査を行うことができる
- ▶ 早期の公衆衛生的な対策を行うこともできる

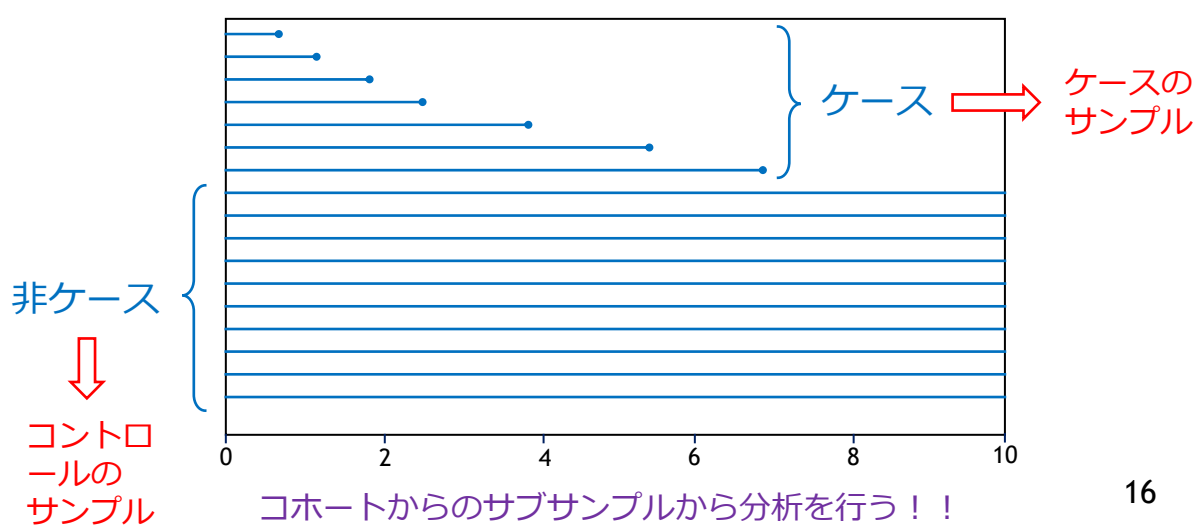
14

ケースコントロール研究

- ▶ ケースコントロール研究は、コホート研究の結果を、コホートの一部の対象者のデータから、後ろ向きに再現するための研究デザイン
- ▶ ケースコントロール研究が定義されるためには、その母体となるコホート研究が明確に定義されなくてはならない
- ▶ ケースコントロール研究は、その母体となるコホート研究でのオッズ比を、一部のサンプリングされたサブサンプルから推定するという研究デザインである

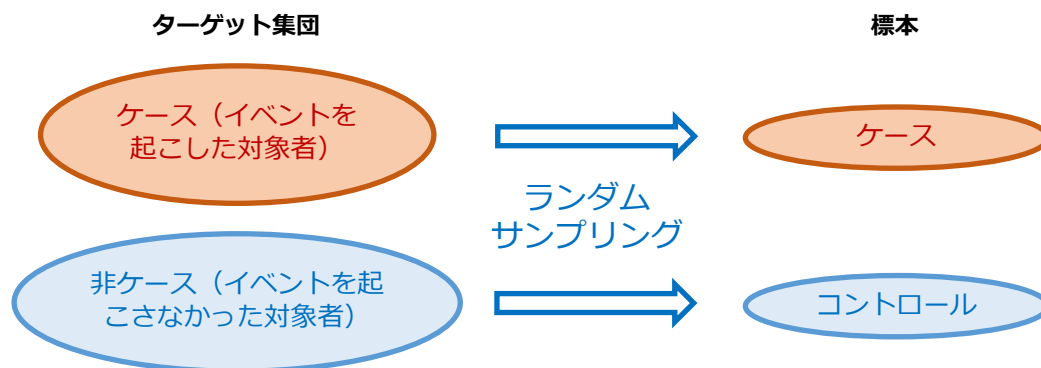
15

コホート研究とケースコントロール研究



16

ケースコントロール研究



- ▶ ケースコントロール研究では、母集団のケース、非ケースから、ケース、コントロールの集団が選択されるため、選ばれた対象者は、母集団のランダムサンプルではない
- ▶ それぞれ、母集団のケース、非ケースのランダムサンプルとは見なすことができる 17

ケースコントロール研究の問題点

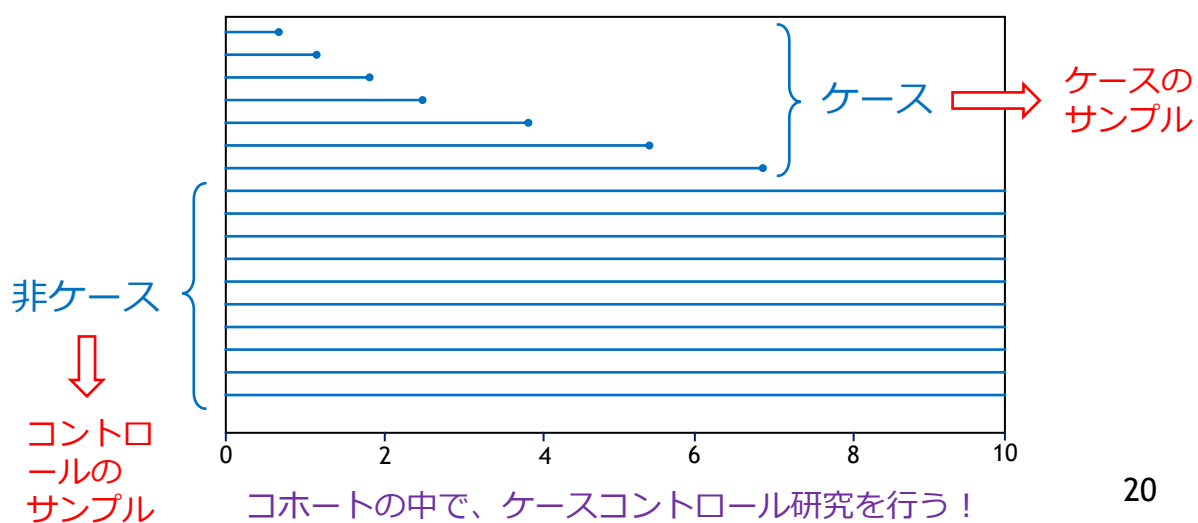
- ▶ コントロールの選択バイアス
 - ▶ 「イベントを起こさなかった患者」から、後ろ向きにコントロールの対象者を選ぶ際、母体となるコホートの非ケースから偏りなく対象者を選ぶことは非常に難しい (e.g., 同じ病院に来院した別な病気の患者から選ぶ、など)
- ▶ 測定バイアス
 - ▶ 後ろ向きに、不完全な記録や記憶をさかのぼって、曝露・交絡の情報を測定しようとする、深刻な欠測や記憶の誤りが起こることがある

しかし、ケースコントロール研究の利点

- ▶ コホートの中のごく少数の対象者について、データを測定するだけで、曝露効果の推定は行うことができる
- ▶ 前頁のバイアスの問題さえ解決されれば、ケースコントロール研究の分析を行うことで、少数の対象者（サブサンプル）によって、コホート研究と同等の結果を得ることができるのでは？
- ▶ コホート研究を前向きに行い、曝露・交絡要因の測定も前向きに行うが、曝露効果の推定は、コホート内の選ばれたケース・コントロールから行う...そんなことができるのか??

19

コホート内ケースコントロール研究



20

コホート内ケースコントロール研究

- ▶ 通常のケースコントロール研究：仮想的に、母体となるコホートを想定して、そこからのサブサンプルとして、ケース・コントロールが得られたと「仮定」して解析を行う
- ▶ コホート内ケースコントロール研究：実際にコホート研究を行って、コホート全体について、曝露・交絡・アウトカムの情報を測定する。そこから、ケースコントロールサンプリングを行って、サブサンプルから曝露効果の推定を行う。
- ▶ ケースコントロール研究の問題点を解決し、しかも、少数のサブサンプルで分析を行うことができる！！

21

コホート内ケースコントロール研究

- ▶ ただし！測定すべき曝露・交絡の情報が、年齢・性別・BMIなど、すべて特別な労力・コストをかけずに得られるものであれば？
- ▶ わざわざサブサンプルに絞って、サンプルサイズを減らして（検出力は下がる）、分析を行う意味はないのでは？
- ▶ その通り！！
- ▶ 冒頭のMRFIT試験の事例のように、測定に、特別な労力・コストを必要とする変数がある場合に、その変数の分析を行うときに用いることで有効なデザインとなる！！

22

デザインの理論の研究の歴史

- ▶ コホート内で、ケースコントロール研究を行うという構想は、多くの統計学者・理論疫学者の関心を引き、1970~80年代にかけて、さまざまなケース・コントロールのサンプリング方法が考案された
- ▶ 現在では、忘れ去られたデザインもあります...
- ▶ その中でも、ケース・コントロールのサンプリングを行う際に、交絡要因のマッチングを行うことができるなどの種々の利点がある方法として、ネステッドケースコントロール研究がスタンダードな方法として、現在まで普及していった

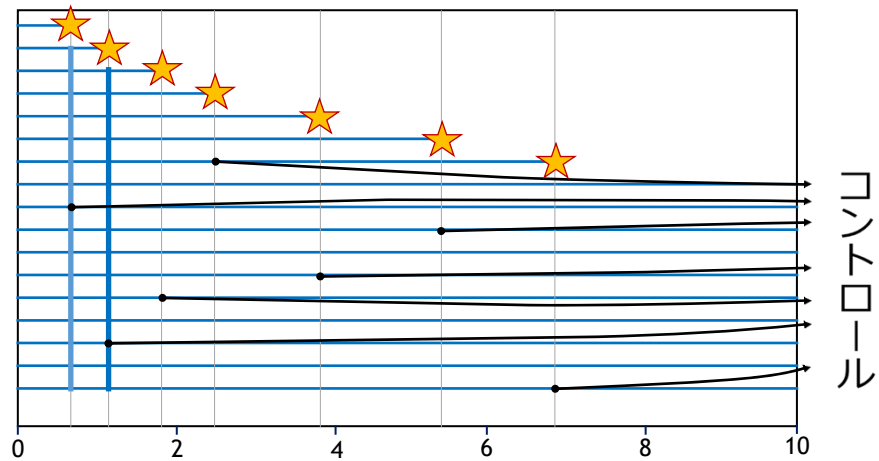
23

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
- ▶ 通常のケースコントロールサンプリングではなく、コホートの追跡履歴の中から、時点でマッチングしたケースコントロールサンプリング（リスクセットサンプリング）を行う
 - ▶ 交絡要因のマッチングを、同時に行うことができる
 - ▶ オッズ比ではなく、ハザード比を、直接的に推定することができる
 - ▶ カウンターマッチングなどのマッチングの技術を用いて、検出力をより高くすることができる

24

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

25

Risk Set Samplingの考え方

- ▶ 通常のケースコントロールサンプリングとは異なり、ケースの発生時点で、その時点の非ケースの集団から、コントロールをマッチングさせる
- ▶ コホートの追跡履歴の中で、Matched Case-Control研究を行うという考え方
- ▶ マッチングの際に、交絡要因が同じコントロールにのみサンプリングを限定すると、交絡要因のマッチングを同時に行うこともできる（マッチングによる交絡調整）

26

統計解析の方法

- ▶ 条件付きロジスティック回帰モデル

$$\Pr(Y = 1) = \frac{\exp(\beta_{0i} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_{0i} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

- ▶ Matched Case-Control研究での多変量解析の方法
- ▶ β_{0i} は、マッチングを取った組ごとに異なる切片項で、マッチングを取った交絡要因の効果は、この項に吸収されて、自動的に調整される（尤度から消えてくれる）
- ▶ 推定されるのは、通常は、オッズ比のはずが・・・。

Thomas (1977), Goldstein and Langholz (1992) 27

統計解析の方法

- ▶ Cox回帰モデル

- ▶ $h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$
- ▶ 母体となるコホートにおいて、比例ハザード性の仮定が成り立っており、Cox回帰モデルが正しいモデルである場合には
- ▶ 実は、前頁の条件付きロジスティック回帰分析で得られる回帰パラメータの推定値は、このCox回帰のハザード比の推定値になることが理論的に証明されている！！
- ▶ イベントは稀でなくても、相対リスクの指標として解釈することができる！！

Goldstein and Langholz (1992) 28

注意！サンプリングの重複

- ▶ ある時点でのケースに対してのコントロールとしてサンプリングされた人が、後の時点でイベントを起こし、ケースとなることがあり得る
- ▶ また、ある時点でコントロールになった人が、別な時点で、別な人のコントロールにもなることがあり得る
- ▶ 同じ人が、異なるマッチドセットに存在することに！！
- ▶ その場合は、重複が起これないように、1回サンプリングされた人は、2度目のサンプリングが起これないように除外すべきなのか？

29

注意！サンプリングの重複

- ▶ No!! コホートの追跡履歴の中から、Unbiased な Matched Case-Controlサンプリングが行われることで、ハザード比の推定が正確になることが保証されている
- ▶ 重複しないような操作をしてしまうと、バイアスの入ったサンプリングになってしまう！
- ▶ 逆に、ハザード比の推定にバイアスが入ってしまうので、むしろ重複が起こるサンプリング方式が、妥当な（バイアスのない）サンプリング方式となる

30

ここまでのまとめ

- ▶ Nested Case-Control研究は、コホートの追跡履歴の中で、Matched Case-Control研究を行うデザイン
- ▶ 統計解析は、通常のMatched Case-Control研究と同様、条件付きロジスティック回帰分析でOK（Rのclogit関数などで、簡単に実行可能）
- ▶ 回帰パラメータの推定値は、ハザード比の推定値として解釈することができる（稀なイベントの仮定がなくても、相対リスクの指標として解釈することができる）

31

事例：MRFIT試験

- ▶ 1990年代 追跡終了後：MRFIT試験の中でも、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を測定し、CHD死亡との関連を評価
- ▶ NCCを用いることで、全員分の血清サンプル分析を行う労力を回避することができ、コストも大幅に節減された
- ▶ 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) 32

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) 33

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、回帰モデルの説明変数として調整

Kuller et al. (1995) 34

MRFIT試験の解析結果

Quartile of CRP (mg/l)	Cases (No., %)	Controls (No., %)	HR (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		

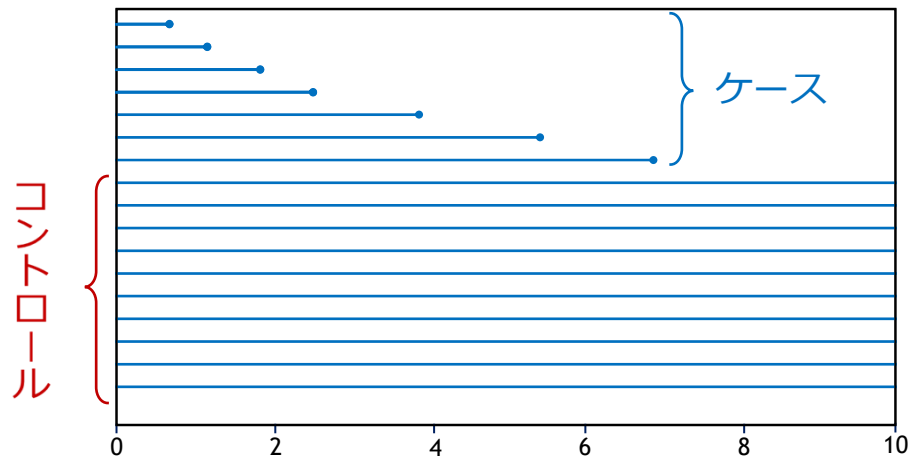
Kuller et al. (1995) 35

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なる
- ▶ 測定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、もっとうまい研究デザインはないか??

36

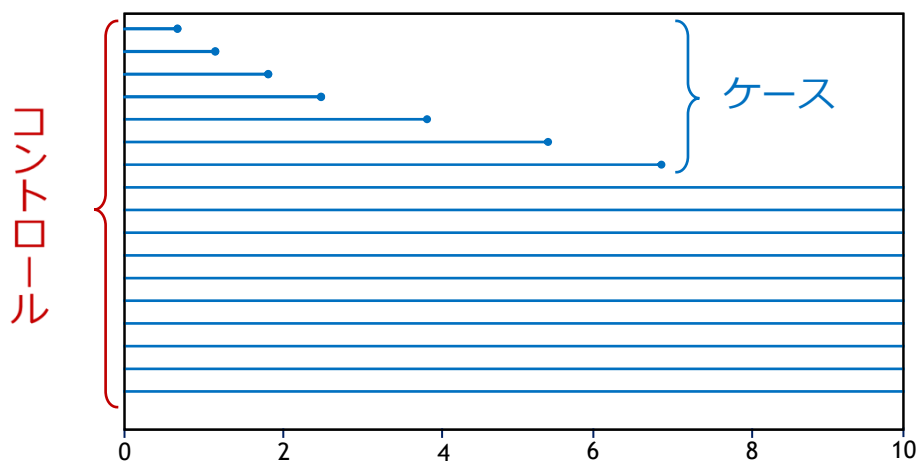
古典的なCase-Control研究



コントロールはイベントを起こしていない対象者

37

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??

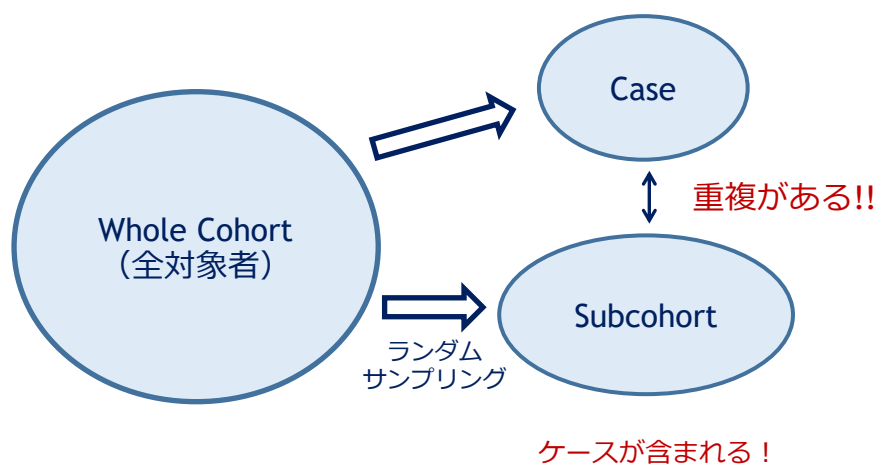
38

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) 39

妥当な評価は可能？



40

生存時間解析

- ▶ Coxの比例ハザード回帰モデル

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \cdots + \beta_p x_p)$$

- ▶ 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で、通常の部分尤度に基づく分析からは、妥当なハザード比の推定量が得られない
- ▶ ケース・サブコホートでの重複もある！！
- ▶ しかし、発想を変えると・・・

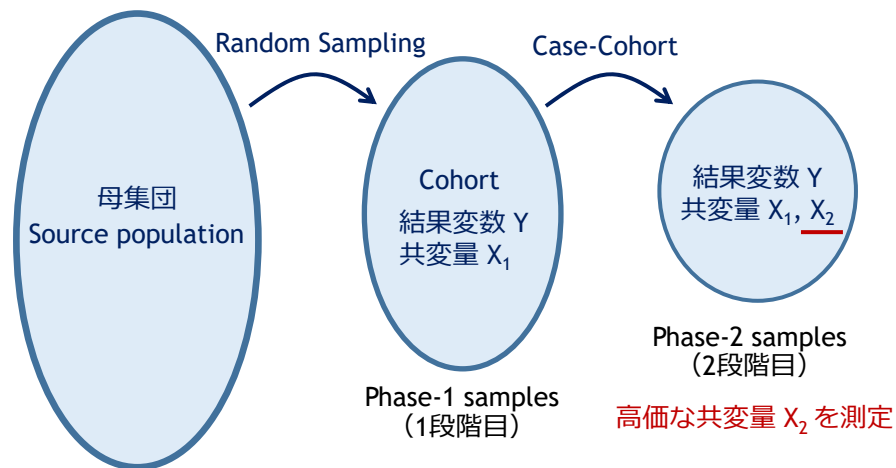
41

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も、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) 42

2段階サンプリングモデル



Noma and Tanaka (2016)

43

不完全データとしての定式化

- ▶ Phase-1 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！！
- ▶ しかも、欠測メカニズムは、MARであることが完全に保証されている！！

44

IPW法

- ▶ Inverse Probability Weighting (IPW) 法
 - ▶ 観測確率（1 - 欠測確率）の逆数で重みづけを行う
 - ▶ MARのメカニズムのもとで、ハザード比のバイアスのない推定量が得られる
- ▶ Phase-1 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が、ケースコホート研究では、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) 45

Cox回帰モデルの部分尤度

$$\mathcal{L}(\beta) = \prod_i \frac{e^{\beta^T x_i}}{\sum_{j \in R_i} e^{\beta^T x_j}}$$

- ▶ R_i : 時点 i でのRisk Set

ハザード比 β 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) 46

IPW法による修正部分尤度 (Lin-Ying法)

$$\mathcal{L}_{IPW}(\beta) = \prod_i \frac{\omega_i e^{\beta^T x_i}}{\sum_{j \in R_i} \omega_j e^{\beta^T x_j}}$$

- ▶ R_i : 時点 i でのRisk Set
- ▶ $\omega_i = N_1/n_1$ (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

Lin and Ying (1993) 47

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は上がる

Borgan et al. (2000) 48

IPW法による解析

- ▶ 層ごとに、Phase-2のサンプリング確率が異なる
 - ▶ = 層ごとに、 X_2 の欠測確率が異なる
- ▶ 所属する層の観測確率（1－欠測確率）の逆数を、各個人の重みにした IPW 法を用いればよい
- ▶ 実際は、観測割合（観測確率の推定値）の逆数を重みに用いることが多い
- ▶ Borganの Type-II 推定量という名前で知られている

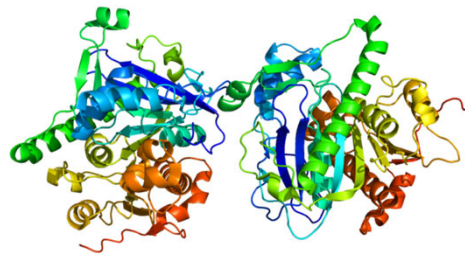
Borgan et al. (2000) 49

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、多くのリスク要因の評価が効率的に行われた
- ▶ 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) 50

リポタンパク質関連ホスホリパーゼA2 (Lp-PLA₂)



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-PLA₂ Studies Collaboration (2010) 51

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) 52

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)

53

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)

54

補足：NCC研究におけるIPW解析

- ▶ NCC研究でも、ケースコホート研究と同様に、2段階のサンプリングデザインと定式化することができる
- ▶ 不完全データと定式化して、IPW法での解析を適用することができる
- ▶ ただし、重み（欠測確率）の計算は複雑に
- ▶ 実は、条件付きロジスティック回帰分析よりも、IPW法のほうが、一般的には推定精度は高くなる
- ▶ 実習で、ご確認ください

55

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.
- ▶ 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.
- ▶ Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics* 20, 1903-1928.

56

- ▶ 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.
- ▶ Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*. 88:1341-1349.
- ▶ 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* 26: 691-706.

57

- ▶ 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. Liddell, 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.

58