原 著

Multiple Imputation 法による 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解析

Multiple Imputation Analysis of Nested Case-Control and Case-Cohort Studies

野間久史^{*1}・田中司朗^{*2}・田中佐智子^{*3}・和泉志津恵^{*4}
Hisashi Noma^{*1}, Shiro Tanaka^{*2}, Sachiko Tanaka^{*3} and Shizue Izumi^{*4}

^{*1}統計数理研究所データ科学研究系

^{*2}京都大学大学院医学研究科社会健康医学系専攻薬剤疫学分野

^{*3}京都大学大学院医学研究科 EBM 研究センター

^{*4}大分大学工学部

^{*1}Department of Data Science, The Institute of Statistical Mathematics

^{*2}Department of Pharmacoepidemiology, Kyoto University School of Public Health

^{*3}EBM Research Center, Kyoto University

^{*4}Faculty of Engineering, Oita University
e-mail:noma@ism.ac.jp

The nested case-control and case-cohort designs are common means of reducing the cost of covariate measurements in large failure-time studies. Under these designs, complete covariate data are collected only on the cases (i.e., subjects whose failure times are uncensored) and some matched controls selected using risk-set sampling or a subcohort randomly selected from the whole cohort. In many applications, certain covariates are readily measured on all cohort members, and surrogate measurements of the expensive covariates may also become available. Using the covariate data collected outside the selected samples, the relative risk estimators can be improved substantially. In this study, we discuss a unified framework for the analysis of these designs using the multiple imputation method, which is a well-established method for incomplete data analyses. The multiple imputation method is currently available in many standard software, and is familiar to practitioners in epidemiologic studies. In addition, this multiple imputation method uses all the data available and approximates the fully efficient maximum likelihood estimator. We also discuss parametric and nonparametric approaches for modeling the distributions of missing covariates: the Markov Chain Monte Carlo method for the Cox regression model by Chen et al. (*Biometrika* 2006; 93: 791–807) and the approximate Bayesian bootstrap. Simulation studies demonstrated that in realistic settings, the multiple imputation estimators had greater precisions than existing estimators. Illustrations with data taken from Wilms' tumor studies are provided.

Key words: Nested case-control studies, Case-cohort studies, Incomplete data, Missing at random, Multiple imputation.

1. はじめに

疫学研究におけるコホート研究では、しばしば規模の大きな集団を、長期間に渡って追跡する必要があり、コホートにおけるすべての対象者において、完全な共変量情報を収集するためには、膨大な費用と労力が必要とされる。疫学研究の理論は、1970年代以降に大きな発展を遂げ、これらの問題に対処するため、古典的なコホート研究、ケースコントロール研究に加え、さまざまな研究デザインが提案されてきた(Rothman et al., 2008)。ネステッドケースコントロール研究(nested case-control studies)、ケースコホート研究(case-cohort studies)は、このような共変量の測定のコストや労力を節減するために開発された研究デザインである。

ネステッドケースコントロール研究(Thomas, 1977)は、イベントの発生時点でのリスクセットから、コントロールをサンプリングし、ケース・コントロール間の比較を行うことによって、曝露効果の推定を行う研究デザインである。特に、測定に高価な費用を要する共変量がある場合、選択されたケース・コントロールにのみ、その測定を行えばよいため、測定の費用を大幅に節減することができる。同様に、ケースコホート研究(Prentice, 1986)は、ケースを含めたコホート全体からのランダムサンプル(サブコホートという)によってコントロールを構成する研究デザインである。ケースコホート研究では、単一のサブコホートによって、母集団における共変量(遺伝子型など)の頻度や分布を推定することが可能であり、また、複数のアウトカムについてのリスク評価を行うことができるという利点もある。

古典的な統計解析の方法では、コホートからサンプリングされたケース、コントロール(サブコホート)を解析対象集団として、曝露効果の指標を偏りなく推定するための方法が議論されてきた(Kalbfleisch and Lawless, 1988; Oakes, 1981; Prentice, 1986; Self and Prentice, 1988; Thomas, 1977)。しかし、多くの研究では、コホートからサンプリングされなかった対象者にも、有用な共変量情報が測定されており、これらの情報を曝露効果の指標の推定に役立てることも、重要な統計的問題となる。この精度の改善の問題については、層別サンプリング・層別解析に基づく方法が利用されており(Borgan et al., 2000; Langholz, 2005)、例えば、ネステッドケースコントロール研究では、カウンターマッチング(Langholz and Borgan, 1995)などのコントロール選択のオプションも提案されている。

近年では、これらのデザインが、母集団からの2段階のサンプリングに基づくデザインであることが着目されるようになり、不完全データとしての定式化に基づく、解析の方法が議論されるようになっている(Breslow and Wellner, 2007ab; Qi et al., 2005)。すなわち、母集団からのランダムサンプリングによって、源泉集団のコホートが得られる過程を1段階目、そこからケース・コントロール(サブコホート)を選択する過程を2段階目と見なすことによって、2段階目で測定される共変量が観測されるか否かは「デザインによって生じる欠測」と見なすことができる。この欠測のメカニズムは、1段階目のコホートで観測された情報によって完全に規定されるMAR(Missing at random)での欠測であり、しかも、その欠測メカニズムは既知である。これにより、不完全データの解析の理論で確立した、推定の最適性の議論が可能となる。特に、セミパラメト

リック推測における最適性の議論 (Robins et al., 1994; Tsiatis, 2006) をもとに, 局外要因となる共変量の分布に, 特定のパラメトリックな仮定を置かないセミパラメトリックモデルのもとでの漸近有効な方法が広く議論されている (Breslow and Wellner, 2007ab; Kulich and Lin, 2004; Qi et al., 2005). しかし, このセミパラメトリック有効な推定方式を実行するためには, 困難な計算上の問題があり, 実践で用いるためには, さらなる検討が必要とされる (Breslow and Wellner, 2007a; Nan, 2004; Zeng and Lin, 2007). このため, 最近では, 漸近有効性は持たないものの, 精度を改善するための代替的な方法として, 標本調査理論における Calibration 法 (Deville and Sarndal, 1992) の応用などが議論されている (Breslow et al., 2009ab; Lumley et al., 2011; Støer and Samuelsen, 2012).

一方で, MAR のもとで漸近有効性を持つ代表的な方法としては, Multiple Imputation 法 (Rubin, 1987) も挙げられる. Multiple Imputation 法は, 当初はベイズ流の枠組みにおいて提案された方法であるが, 近年の研究により, 一定の正則条件のもとで, 頻度論的な漸近有効性を有する推定量を与えることが示されている (Wang and Robins, 1998; Robins and Wang, 2000). その理論的な詳細については, 3.1 節で解説する. Multiple Imputation 法は, 単純な原理と計算によって実行することができる手法であり, 既に, 多くの統計ソフトウェアで標準的なモジュールとして実装されている. 医学研究の実務家の間でも, 広く普及しており (Ibrahim et al., 2012; Klebanoff and Cole, 2008; Little et al., 2012; Sterne et al., 2009; Wood et al., 2004), 実践的にも, 医学研究者の理解を得られやすい方法であると思われる. Multiple Imputation 法のケースコホート研究への応用については, Marti and Chavance (2011) によって, 単純な回帰モデルによる補完方法が議論されているが, 限定的な設定と応用についての議論しか行われていない. Multiple Imputation 法の本質的な有用性は, より一般的に, 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別や, 層別・カウンターマッチングなどのコントロールの選択のオプションによらず, 統一的な枠組みでの取り扱いが可能であるという点にある.

本稿では, これらの 2 段階デザインの統一的な解析方法として, Multiple Imputation による漸近有効な推定の方法を与える. 補完値の生成方法には, パラメトリック・ノンパラメトリックな 2 つの方法を議論する. パラメトリックな方法には, 共変量に MAR での欠測があるもとでの Cox 回帰モデルにおける事後予測分布からのサンプリング方法として, Chen et al. (2006) によるマルコフ連鎖モンテカルロ法を利用した方法を与える. ノンパラメトリックな方法としては, 近似ベイズ流ブートストラップ法 (Approximate Bayesian bootstrap; ABB) を用いた方法を議論する. ABB による補完値の生成では, 局外要因である共変量の分布に特定のパラメトリックな仮定を必要とせず, セミパラメトリック推測の方法 (Breslow and Wellner, 2007ab; Kulich and Lin, 2004; Qi et al., 2005) と同程度の仮定のもとで, 曝露効果の推定が可能である. また, 数値実験・実データの解析を通して, 実践的な設定のもとでの既存の方法との有用性の比較を行う.

2. 2段階サンプリングと不完全データとしての定式化

2.1 モデルと表記法

1段階目のコホートは、母集団からのランダムサンプリングによって得られた N 人の対象者によって構成されるものとする。 i 番目の対象者に対して ($i = 1, \dots, N$)、関心のあるイベントが起るまでの時間を T_i^* 、潜在的な打ち切りの時間を C_i として、 $T_i = \min(T_i^*, C_i)$ 、 $\delta_i = I(T_i \leq C_i)$ とする。1段階目の対象者すべてで測定される共変量を X_i 、2段階目の対象者のみに測定される共変量を Z_i とする。 T_i^* は、 (X_i, Z_i) を与えたもとの、 C_i とは条件付き独立であるとし、 (X_i, Z_i) 所与のもとの T_i^* の条件付き分布は Cox (1972) による比例ハザードモデル:

$$h(t | x_i, z_i) = h_0(t) \exp(\beta_1 x_i + \beta_2 z_i) \quad (1)$$

であると仮定する。関心のあるパラメータは、回帰係数 $\beta = (\beta_1, \beta_2)$ であり、 $h_0(t)$ はベースラインハザード関数である。

2.2 ネステッドケースコントロール研究

ネステッドケースコントロール研究における2段階目のサンプリングでは、1段階目のコホートを源泉集団として、イベントが起きた各時点のリスクセットから、コントロールのサンプリングを行う。これを、リスクセットサンプリング (risk set sampling) という。リスクセットサンプリングは、イベントが起きた時点ごとに、イベントを起こしたケースに対応するコントロールをマッチングさせたケースコントロールサンプリングと解釈することができる。通常のケースコントロール研究と同様、既知の交絡要因がある場合、マッチングをとることで、交絡の影響を調整した β の推測が可能である。これは、2段階目のサンプリングを、層別に行っていることに対応し、適当な層別変数を用いることで、統計的な精度の改善を図ることもできる。カウンターマッチング (Langholz and Borgan, 1995) は、関心のある曝露変数の代替変数を用いた層別サンプリングの方法である。

ネステッドケースコントロール研究から得られるデータは、1段階目のコホート全体を解析対象集団とすると、 Z_i が部分的に欠測した不完全データと見なすことができる。イベントの各発生時点で、それぞれ m 人のコントロールが選択されるものとして、イベント発生時点 $T = T_i$ におけるリスクセットの人数を R_i とすると、このリスクセットのそれぞれの対象者がコントロールとしてサンプリングされる確率は $m/(R_i - 1)$ となる ($i = 1, \dots, N$)。したがって、 i 番目の対象者が、追跡期間全体で、2段階目にサンプリングされる確率は、

$$p_i = \begin{cases} 1 & (\delta_i = 1) \\ 1 - \prod_{j: T_j < T_i} \left(1 - \frac{m}{R_j - 1} \delta_j\right) & (\delta_i = 0) \end{cases} \quad (2)$$

となる (Samuelsen, 1997)。すなわち、共変量 Z_i の欠測メカニズムは、 T_i の順序と δ_i 、層別変数によって完全に説明することができる。層別サンプリングの場合にも、コントロールの選択が層ごとに行われることを反映することで、同様に2段階目のサンプリング確率を得ることができる。

2.3 ケースコホート研究

ケースコホート研究では, 前節で述べた通り, コントロールのサンプリングを, ケースも含めた 1 段階目のコホート全体から行う(これを, サブコホートという). ケースコホート研究では, このケースとサブコホートのサンプリングを 2 段階目のサンプリングと見なす. ネステッドケースコントロール研究では, コントロールは対象となるイベントのケースにマッチングさせて選択されるため, 複数のアウトカムを対象とした研究では, アウトカムごとにコントロール集団をサンプリングする必要があるが, ケースコホート研究では, 1 つのサブコホートで, 複数のアウトカムについてのリスク評価を行うことが可能である. また, ネステッドケースコントロール研究では, 原則として, コホート内のケースはすべてサンプリングすることになるが, ケースコホート研究では, サンプリングするケースの割合も, 解析者が任意に定めることができる.

ケースコホート研究から得られるデータも, 1 段階目のコホート全体を解析対象集団とすると, Z_i が部分的に欠測した不完全データとなる. N 人のコホートの対象者のうち, イベントを起こしたケースの数を N_1 , イベントが観測されず, 打ち切りを起こした対象者の数を N_0 とする. N_1 人のケースからケースサンプルが選択される確率を r_1 , コホート全体からサブコホートのサンプルがランダムに選択される確率を r_0 とする. i 番目の対象者が 2 段階目にサンプリングされる確率は,

$$p_i = \begin{cases} r_0 & (\delta_i = 0) \\ r_0 + r_1 - r_0 r_1 & (\delta_i = 1) \end{cases} \quad (3)$$

となる. すなわち, 共変量 Z_i の欠測メカニズムは, r_0, r_1 によって説明することができる. 層別サンプリング(Borgan et al., 2000)の場合も同様に, 層ごとのケースサンプル, サブコホートの選択確率によって, それぞれの対象者が 2 段階目にサンプリングされる確率は規定される.

いずれのデザインも, 1 段階目のコホート全体を解析対象集団とすると, 比例ハザードモデル (1) のもとで, 共変量 Z_i が部分的に欠測した不完全データと見なすことができる. Z_i の欠測のメカニズムも, (2), (3) で与えた通り, 完全に既知であり, MAR のもとでの不完全データの解析方法によって, β についての推測を行うことが可能である.

3. Multiple Imputation 法

3.1 原理と方法

Multiple Imputation 法(Rubin, 1987; Schafer, 1997)は, MAR の欠測のもとでの最も代表的な不完全データの解析方法である. この方法では, 不完全データの欠測値にランダムな補完値を代入した複数のデータセットを生成し, それぞれに得られる解析結果を統合する. 古典的な単一の補完値を用いた方法でも, MAR の欠測のもとでの偏りのない推定量を得ることは可能であるが, 一般的に, 推定量の分散の評価が困難であることが問題となる(Barnard et al., 2005). Multiple Imputation 法では, 複数の補完値に基づく結果を統合することにより, 補完値の不確実性を含めた統計的な誤差の評価を行う明快な枠組みを与えることができる. 本節では, 一般論としての議論を行うため, 他節とは異なる表記法を用いる. 補完値の生成回数を M 回とし, 関心のあるパラメータを θ とする.

Multiple Imputation は, 以下の 3 段階のステップによって, 推定量を構成する.

Step 1. (補完値の生成)

欠測値に代入する補完値のデータセットを, M 組, 適当な方法によって生成する. 後に示す通り, Multiple Imputation 法は, ベイズ流の枠組みでの事後推測を近似するものと解釈することができ, この補完値のデータセットは, θ について適当な事前分布を仮定したもとでの事後予測分布からのサンプルとして生成されることを想定している. 具体的な方法については, 3.2 節に示す.

Step 2. (M 組の擬似的な完全データの解析)

Step 1 によって生成した M 組の補完値のデータセットを欠測データに代入することによって, M 組の擬似的な完全データを作る. この M 組のデータセットをそれぞれ独立に解析することによって, パラメータ θ の推定値 $\hat{\theta}_m$ とその分散の推定値 $\hat{V}_m$ を得る ($m = 1, 2, \dots, M$).

Step 3. (M 組の解析結果の統合)

Step 1, 2 によって得られた, M 組の推定値を, 以下の公式によって併合する.

$$\hat{\theta}_{\text{IM}} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m$$

$$\hat{V}(\hat{\theta}_{\text{IM}}) = W_{\text{IM}} + (1 + M^{-1})B_{\text{IM}}$$

ただし,

$$W_{\text{IM}} = \frac{1}{M} \sum_{m=1}^M \hat{V}_m, \quad B_{\text{IM}} = \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta}_{\text{IM}})(\hat{\theta}_m - \hat{\theta}_{\text{IM}})^T$$

である. $W_{\text{IM}}, B_{\text{IM}}$ は, それぞれ補完内・補完間での分散である. 検定・信頼区間の構成においては,

$$\nu = (M-1)(1 + \rho^{-1})^2,$$

を自由度とする t 分布を参照分布に用いればよい (Rubin and Schenker, 1986). ただし, $\rho = (1 + M^{-1}) B_{\text{IM}}$ である.

先に与えた, Multiple Imputation 法による推定量 $\hat{\theta}_{\text{IM}}, \hat{V}(\hat{\theta}_{\text{IM}})$ は, ベイズ流の枠組みで, 明快な解釈が可能である (Rubin, 1987). ここでは, 解析対象となるデータのうち, 観測データを D_{obs} , 欠測データを D_{mis} と表記する. MAR の欠測メカニズムのもとでは, 欠測データ D_{mis} を周辺化した, 観測データ事後分布 $p(\theta | D_{\text{obs}})$ によって推測は行われる. その確率密度関数は

$$p(\theta | D_{\text{obs}}) = \int p(\theta | D_{\text{obs}}, D_{\text{mis}}) p(D_{\text{mis}} | D_{\text{obs}}) dD_{\text{mis}}$$

と表され, $(D_{\text{obs}}, D_{\text{mis}})$ の完全データに基づく事後分布を, 欠測データの予測分布 $p(D_{\text{mis}} | D_{\text{obs}})$ で積分して得られるものと解釈できる. θ とその分散の推定量は, 事後平均 $E[\theta | D_{\text{obs}}]$, 事後分散 $V[\theta | D_{\text{obs}}]$ となる.

Step 2 でのパラメータの推定が最尤法によって行われる場合, m 回目の D_{mis} の補完データ $D_{\text{mis}}^{(m)}$ が予測分布 $p(D_{\text{mis}} | D_{\text{obs}})$ によって生成されたとすると, これを代入した擬似的な完全デー

タから得られる推定量 $\hat{\theta}_m, \hat{V}_m$ は, それぞれ $E[\theta | D_{obs}, D_{mis}^{(m)}], V[\theta | D_{obs}, D_{mis}^{(m)}]$ を近似するものと解釈できる. したがって, Multiple Imputation 推定量は, 十分な繰り返し回数のもと,

$$\lim_{M \rightarrow \infty} \hat{\theta}_{IM} = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m \rightarrow E[E[\theta | D_{obs}, D_{mis}] | D_{obs}] = E[\theta | D_{obs}]$$

となり, 事後平均 $E[\theta | D_{obs}]$ に収束する. 同様に, 事後分散も,

$$V[\theta | D_{obs}] = E[V(\theta | D_{obs}, D_{mis}) | D_{obs}] + V[E(\theta | D_{obs}, D_{mis}) | D_{obs}]$$

であり, それぞれの項は,

$$\begin{aligned} \lim_{M \rightarrow \infty} W_{IM} &= \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \hat{V}_m \rightarrow E[V(\theta | D_{obs}, D_{mis}) | D_{obs}] \\ \lim_{M \rightarrow \infty} B_{IM} &= \lim_{M \rightarrow \infty} \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_m - \hat{\theta}_{IM})(\hat{\theta}_m - \hat{\theta}_{IM})^T \rightarrow V[E(\theta | D_{obs}, D_{mis}) | D_{obs}] \end{aligned}$$

に対応する.

頻度論的観点からは, Multiple Imputation 法による推定量は, 最尤推定量と漸近的な精度が等しいことが知られている. 実際, Step 2 でのパラメータの推定が最尤法によって行われる場合, 一定の正則条件のもと, $\hat{\theta}_{IM}$ の漸近分布は $M \rightarrow \infty$ のもとで, 観測データ尤度関数に基づく最尤推定量の漸近分布に一致することが示されている (Robins and Wang, 2000; Wang and Robins, 1998). Cox 回帰の場合には, Step 2 で最大部分尤度法 (Cox, 1972) が用いられることになるが, Chen et al. (2006), Sinha et al. (2003) の議論により, Cox の部分尤度は, 累積ベースラインハザード関数に無情報な事前確率過程を仮定したもとで, β の事後分布と等価となるため, 同様の漸近有効性を有する推定量が得られる.

繰り返し回数 M は, 欠測の占める割合がそれほど大きくない場合には 5~10 程度でも十分であると言われてきた (Rubin, 1987). しかし, 近年の議論では, より多くの繰り返し回数が必要であるとされており, 例えば, White et al. (2011) では, 少なくとも不完全なデータが占める割合のパーセンテージ以上は繰り返しを行う必要があると述べている (完全データの占める割合が 80%である場合, 20 回以上の繰り返しが必要). また, Royston and White (2011) は, Multiple Imputation 法が, そもそもモンテカルロ法に基づく方法であることを考慮すると, 十分な精度でモンテカルロ誤差を制御するためには, 100~1000 回の繰り返しは必要であると述べている. 現在の計算機環境では, 繰り返し回数 M を大きくしても, 計算量はそれほど負担にならないので, 十分な回数での繰り返しを行うことが望ましい.

3.2 補完値の生成方法

前節で行った説明の通り, Multiple Imputation 法は, 一定の正則条件のもとで, 漸近有効な一致推定量を与える方法となる. ただし, そのためには適切な補完値を用いることが前提となり, 補完値の生成の問題 (Step 1) は, Multiple Imputation 法の最も重要な要素のひとつとなる. 本節では, この補完値の生成についての方法を議論する.

3.2.1 マルコフ連鎖モンテカルロ法(MCMC)

前節で説明した通り, Multiple Imputation 法は, バイズ流の枠組みのもとで, 関心のあるパラメータについての事後推測を近似したモンテカルロ法に基づく推定方法である. これにより, ひとつの率直なアプローチとして, 適当なバイズモデルのもとで, 欠測データの事後予測分布 $p(D_{mis} | D_{obs})$ から乱数を発生させ, それを補完値に用いるという方法が考えられる. 実践的な設定のもとでは, ほとんどのバイズモデルのもとで, 予測分布 $p(D_{mis} | D_{obs})$ を解析的に表現することは不可能であるが, マルコフ連鎖モンテカルロ法(MCMC)を用いることによって, この分布からの乱数を直接発生させることは可能である (Schafer, 1997; 1999).

2 節で定義した表記法を用いて, 比例ハザードモデル (1) のもとで, 共変量 Z_i が MAR で欠測した, バイズ流の推測を考える. 一般性を失うことなく, $T_1 \leq T_2 \leq \dots \leq T_N$ とし, このうち, タイを除いた(distinct な)イベント発生時点を $\tilde{T}_{(1)} < \tilde{T}_{(2)} < \dots < \tilde{T}_{(\tilde{N})}$ と表記する. 以下では, 累積ベースラインハザード関数 $H_0(t)$ の事前確率過程として, ガンマ過程 (gamma process) $\mathcal{GP}(c_0 H^*, c_0)$ を仮定する. H^* は, 既知の単調増加かつ微分可能な関数であり, $c_0 > 0$ である. ガンマ過程は, Cox の比例ハザードモデルのもとで, 累積ベースラインハザード関数のノンパラメトリックな事前確率過程として広く利用されている (Kalbfleisch, 1978; Sinha et al., 2003). ガンマ過程では, 累積ベースラインハザード関数の増分 $h_i = H_0(\tilde{t}_{(i)}) - H_0(\tilde{t}_{(i-1)})$ がガンマ分布 $\text{Ga}(c_0 h_{0i}, c_0)$ に従うことを仮定している. $h_{0i} = H^*(\tilde{t}_{(i)}) - H^*(\tilde{t}_{(i-1)})$ であり, これは, 累積ベースラインハザード関数 H_0 の増分過程 $\{h_i: i = 2, 3, \dots, \tilde{N}\}$ の平均を表し, c_0 はその精度パラメータとなる.

打ち切り時間 C_i の分布は, 欠測した共変量に依存しないものとし, 共変量の欠測メカニズムは MAR であるとする. イベントの起こる時間にタイを許したもとで, 比例ハザードモデル (1) の全尤度関数は, 以下のように書ける (Chen et al., 2006).

$$L_t(\beta) = \prod_{i=1}^N \exp \left\{ c_0 H^*(t_i) \log \left(\frac{c_0 + A_{i+1}}{c_0 + A_i} \right) \right\} \left[-c_0 h^*(t_i) \log \left\{ 1 - \frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + A_i} \right\} \right]^{\delta_i} \quad (4)$$

ただし, $h^*(t) = dH^*(t)/dt$, $A_i = \sum_{l \in \mathcal{R}(t_i)} \exp(\beta_1 x_l + \beta_2 z_l)$, $\mathcal{R}(t_i)$ は $T = t_i$ におけるリスクセットの添え字集合である. $Z_i | X_i$ の条件付き分布の密度関数は, $f(z | x, \alpha)$ とする (Z と X が独立である場合には, Z の周辺分布を $f(z | \alpha)$ とする). 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での応用については, サンプリングの構造 (2), (3) を, 特に考慮する必要はなく, 共変量の分布 $f(z | x, \alpha)$ を正しく特定できれば, 事後予測分布からのサンプルを得ることができる. パラメータ (α, β) の事前分布は, $p(\alpha, \beta) = p(\alpha)p(\beta)$ であるとし, β と α は独立であるとする.

尤度関数 (4) の型から明らかであるが, (α, β) の事後分布や Z_i の事後予測分布を解析的に得ることはできない. また, MCMC を用いて, これらの分布からの乱数を発生させることも, 一般的に容易ではない (Chen et al., 2006). 本稿では, Chen et al. (2006) によって与えられた Gibbs サンプリング (Gelfand and Smith, 1990) のアルゴリズムを用いて, 事後予測分布からのサンプルを生成する.

いま, タイを表す指示変数として, $H^*(\tilde{t}_i) = H^*(\tilde{t}_{i-1})$ ならば 0, $H^*(\tilde{t}_i) > H^*(\tilde{t}_{i-1})$ ならば 1 をとる確率変数 ν_i と, 以下に与える 2 つの潜在変数 $w = (w_1, \dots, w_N)$, $q = (q_1, \dots, q_N)$ ($0 < w_i <$

1, $q_i > 0; i = 1, \dots, N$) を導入する. 以下の 2 つの関係式が成り立つ.

$$\begin{aligned}
 & \left[-\log \left\{ 1 - \frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + A_i} \right\} \right]^{\delta_i} \\
 &= \int_0^1 \left[\left\{ \frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + A_i} \right\} / \left\{ 1 - w_i \frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + A_i} \right\} \right]^{\delta_i} dw_i \\
 &= \int_0^1 \left[\frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + (1 - w_i) \exp(\beta_1 x_i + \beta_2 z_i) + \sum_{j \in \mathcal{R}(t_i)} \exp(\beta_1 x_j + \beta_2 z_j)} \right]^{\delta_i} dw_i \\
 &= \int_0^1 \int_0^\infty \exp\{\delta_i(\beta_1 x_i + \beta_2 z_i)\} \times \exp\left(-q_i \left[(1 - \delta_i) \right. \right. \\
 &\quad \left. \left. + \delta_i \left\{ c_0 + (1 - w_i) \exp(\beta_1 x_i + \beta_2 z_i) + \sum_{j \in \mathcal{R}(t_i)} \exp(\beta_1 x_j + \beta_2 z_j) \right\} \right] \right) dq_i dw_i \\
 & \prod_{i=1}^N \exp \left\{ c_0 h_{0i} \log \left(\frac{c_0}{c_0 + A_i} \right) \right\} \\
 &= \prod_{i=1}^N \int_0^\infty \exp(-\nu_i h_i A_i) \frac{c_0^{c_0 h_{0i} + (1 - \nu_i)} h_i^{c_0 h_{0i} + (1 - \nu_i) - 1}}{\Gamma\{c_0 h_{0i} + (1 - \nu_i)\}} \exp(-c_0 h_i) dh_i
 \end{aligned}$$

ここでも, 観測されたデータをまとめて D_{obs} と表記し, 共変量 Z のうち, 欠測した対象者のものについては Z_{mis} と表記するものとする. また, $h = (h_1, \dots, h_N)$ とする ($h_i > 0; 1, \dots, N$). これらを潜在変数としてモデル化することにより, $(\beta, \alpha, z_{mis}, w, q, h)$ の同時事後分布は, 以下の型で与えられる.

$$\begin{aligned}
 p(\beta, \alpha, z_{mis}, w, q, h \mid D_{obs}) &\propto \prod_{i=1}^N \left\{ \exp\{\delta_i(\beta_1 x_i + \beta_2 z_i)\} \right. \\
 &\quad \left. \exp\left(-q_i \left[(1 - \delta_i) + \delta_i \left\{ c_0 + (1 - w_i) \exp(\beta_1 x_i + \beta_2 z_i) + \sum_{j \in \mathcal{R}(t_i)} \exp(\beta_1 x_j + \beta_2 z_j) \right\} \right] \right) \right. \\
 &\quad \left. \times h_i^{c_0 h_{0i} + (1 - \nu_i) - 1} \exp(-h_i \{c_0 + \nu_i A_i\}) f(z_i \mid x_i, \alpha) \right\} p(\beta) p(\alpha)
 \end{aligned}$$

この密度関数の型から, 条件付き事後予測分布 $p(z_{mis} \mid \beta, \alpha, w, q, h, D_{obs})$ において, 共変量 Z が 3 つの変数 (w, q, h) 所与のもとで, すべての対象者で条件付き独立となることがわかる. これにより, 以下の Gibbs サンプリングのアルゴリズムを構成することができる.

Gibbs サンプリング (Chen et al., 2006)

(i) $p(\beta \mid z_{mis}, w, q, h, D_{obs})$ からのサンプリング

確率密度関数は, 以下の型で与えられる.

$$\begin{aligned}
 p(\beta \mid z_{mis}, w, q, h, D_{obs}) &\propto \prod_{i=1}^N \left\{ \exp\{\delta_i(\beta_1 x_i + \beta_2 z_i)\} \right. \\
 &\quad \left. \exp\left(-q_i \delta_i \left[(1 - w_i) \exp(\beta_1 x_i + \beta_2 z_i) + \sum_{j \in \mathcal{R}(t_i)} \exp(\beta_1 x_j + \beta_2 z_j) \right] \right) \exp(-h_i \nu_i A_i) \right\} p(\beta)
 \end{aligned}$$

これにより, $p(\beta)$ が log-concave な形状を持つ場合, 条件付き事後分布 $p(\beta \mid z_{mis}, w, q, h, D_{obs})$ も

log-concave な形状を持つ。これは、非正則な無情報事前分布を仮定する場合にも成り立つ。したがって、例えば、Gilks and Wild (1992) の Adaptive Rejection サンプルングを利用して、この分布からの乱数を得ることができる。

(ii) $p(\alpha | z_{mis}, D_{obs})$ からのサンプルング

確率密度関数は、以下の単純な型で与えられる。

$$p(\alpha | z_{mis}, D_{obs}) \propto \prod_{i=1}^N p(z_{mis}, D_{obs} | \alpha) p(\alpha)$$

非正則な無情報事前分布を仮定する場合には、大標本のもとでの近似的なサンプルングの方法として、現時点の z_{mis} の値を代入した擬似的な完全データに対して、 Z の分布のモデル $f(z | x, \alpha)$ における最尤推定値 $\hat{\alpha}$ を求め、この $\hat{\alpha}$ を平均、観測情報行列の逆行列を共分散行列とした多変量正規分布からサンプルングを行うという方法もある。

(iii) $p(z_{mis} | \beta, \alpha, w, q, h, D_{obs})$ からのサンプルング

i 番目の対象者の Z_{mis} を $Z_{i,mis}$ とする。 $Z_{i,mis}$ の条件付き事後予測分布の密度関数は、

$$p(z_{i,mis} | \beta, \alpha, w, q, h, D_{obs}) \propto \exp \left[\delta_i (\beta_1 x_i + \beta_2 z_i) - \left\{ q_i \delta_i (1 - w_i) + h_j \nu_j + \sum_{j: i \in \mathcal{R}(t_j)} (q_j \delta_j + h_j \nu_j) \right\} \exp(\beta_1 x_i + \beta_2 z_i) \right] f(z_{i,mis} | x_i, \alpha)$$

として与えられる。先にも述べた通り、潜在変数 (w, q, h) を導入する利点は、この $Z_{i,mis}$ の条件付き事後予測分布が、条件付き独立になる点にあり、これによって、欠測した共変量 $Z_{i,mis}$ のサンプルングを容易に行うことが可能となる。

(iv) $p(w | \beta, z_{mis}, D_{obs})$ からのサンプルング

Lin (1994) による Collapsed Gibbs Sampler の方法を応用し、 q による周辺化を行った後の条件付き事後分布 $p(w | \beta, z_{mis}, D_{obs})$ からのサンプルングを行う。 $(\beta, z_{mis}, D_{obs})$ を与えたもとで、 w_i は条件付き独立であり、累積分布関数は、以下のように書くことができる。

$$\begin{aligned} \Pr(w_i \leq \omega | \beta, z_{mis}, D_{obs}) \\ = \begin{cases} \log \left\{ 1 - \frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + A_i} \omega \right\} / \log \left\{ 1 - \frac{\exp(\beta_1 x_i + \beta_2 z_i)}{c_0 + A_i} \right\} & (\delta_i = 1) \\ \omega & (\delta_i = 0) \end{cases} \end{aligned}$$

これにより、累積分布関数の逆関数による変換を利用して、 $p(w | \beta, z_{mis}, D_{obs})$ からの乱数を生成することができる。

(v) $p(q | \beta, z_{mis}, w, D_{obs})$ からのサンプルング

$(\beta, z_{mis}, w, D_{obs})$ を与えたもとでの q_i は条件付き独立となり、その条件付き事後分布は、

$$\left[(1 - \delta_i) + \delta_i \left\{ c_0 + (1 - w_i) \exp(\beta_1 x_i + \beta_2 z_i) + \sum_{j \in \mathcal{R}(t_i)} \exp(\beta_1 x_j + \beta_2 z_j) \right\} \right]^{-1}$$

を平均に持つ指数分布となる。

(vi) $p(h \mid \beta, z_{mis}, w, D_{obs})$ からのサンプリング

$(\beta, z_{mis}, w, D_{obs})$ を与えたもとで, h_i は条件付き独立となり, その条件付き事後分布は, ガンマ分布 $\text{Ga}(c_0 h_{0i} + (1 - \nu_i), c_0 + \nu_i A_i)$ となる.

理論的, 技術的な詳細については, Chen et al. (2006) をご参照いただきたい.

適当な Burn-in 期間のあと, このアルゴリズムによって生成される Z_{mis} の系列は, 条件付き事後分布 $p(z_{mis} \mid D_{obs})$ からの乱数と見なすことができる. これを補完値に用いることにより, Multiple Imputation 法による, β の近似的な事後推測を行うことができる.

3.2.2 近似ベイズ流ブートストラップ (ABB)

MCMC は, 欠測値を含む共変量 Z の分布型にパラメトリックなモデルを仮定している. しかし, 我々が実際に興味を持っているのは, 曝露効果の指標をモデル化した比例ハザードモデル (1) のパラメータ β であり, 残る共変量の分布 $Z \mid X$ は局外要因である. 実際, 近年の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解析方法の議論では, この Z の分布に特定のパラメトリックなモデルを仮定しないセミパラメトリックな方法論が議論されている (Breslow and Wellner, 2007ab; Kulich and Lin, 2004; Qi et al., 2005).

そこで, 可能な限り分布型についての仮定を必要としない方法として, 近似ベイズ流ブートストラップ (Approximate Bayesian bootstrap; ABB) が提案されている (Rubin, 1987; Rubin and Schenker, 1986; Schafer, 1997; 1999). ABB は, 観測されたデータをもとに定められる適当な層ごとに, Z_{mis} の補完値を, 観測された Z の値 Z_{obs} の経験分布関数から 2 段階のリサンプリングを行うことによって得るという方法である.

Step 1. (Z_{obs} の経験分布関数からのリサンプリング)

適当に定められた層ごとに, 観測データ Z_{obs} の経験分布関数からリサンプリングを行い, Z_{obs} と同数のサンプル Z_{obs}^* を得る.

Step 2. (Z_{obs}^* の経験分布関数からのリサンプリング)

Step 1 で得られた Z_{obs}^* の経験分布関数からリサンプリングを行い, 欠測した共変量 Z_{mis} の補完値 Z_{mis}^* を得る.

この方法は, Z_{obs} の分布に, 観測値の各点に値をとる多項分布モデルを, 事前分布に, 自然共役な Dirichlet 分布を仮定しており, 上記の手順によって, 事後予測分布からの近似的なサンプルを得ることができる (Rubin, 1987; Schafer, 1997; 1999).

この方法では, Z の分布について, 特定のパラメトリックな分布型を仮定する必要がなく, 必要となるのは, イベントの発生時間の分布に対しての比例ハザードモデル (1) の仮定だけである. 一方, この方法も, 対象となる Z_{obs}, Z_{mis} が交換可能であるという仮定は必要とする. しかし, この仮定は, 先述のセミパラメトリック推測の方法論 (Breslow and Wellner, 2007ab; Kulich and Lin, 2004; Qi et al., 2005) にも必要であり, その意味で, ABB による Multiple Imputation 法は, これと同等の仮定を用いた方法と解釈することができる. ただし, 既存のセミパラメトリック有効な β の推定方法に比べると, 計算の問題は, 大きく軽減されており, 上記のような単純な手順で実行することができる.

実践的には、予測分布 $p(z_{mis} | D_{obs})$ からのサンプルを得るために、観測データの中で Z の分布に関連のある共変量がある場合には、それらの共変量の値に応じて分布の交換可能性を仮定し、ABB を行うことが理に適っている。Lavori et al. (1995) は、Rosenbaum and Rubin (1983) による傾向スコア (propensity score) を利用した層別を行い、層ごとに ABB を行うという方法を提案している。この方法は、適当な 2 項回帰モデルから「 Z が欠測する確率」を推定し、その推定値の大きさによって層別を行うというものである。傾向スコアの性質から、欠測に関連する共変量の周辺分布が、欠測の生じた対象者と生じていない対象者の間でバランスするように層別が行われる。ただし、2 段階サンプリングにおける共変量 Z の欠測メカニズムは完全に既知であり、 Z の観測確率は、先述の (2), (3) で与えられる。特に、ケースコホート研究では、 Z の観測確率は 2 段階目のサンプリングの層ごとのケース、打ち切りを起こした対象者の間でそれぞれ均一である。したがって、これらの層ごとに、ABB を行ってもよい。

4. シミュレーション実験

本節では、シミュレーション実験によって、Multiple Imputation 法の推定量の性能の評価といくつかの従来法との比較を行う。4.1 節で、ネステッドケースコントロール研究、4.2 節で、ケースコホート研究についての評価を行う。

4.1 ネステッドケースコントロール研究

シミュレーションの設定は、5 節の事例と Marti and Chavance (2011) のシミュレーションを参考にして定めた。まず、1 段階目の対象者として、10000 人の対象者が母集団からランダムにサンプリングされたとする。イベントまでの時間は、指数回帰モデル

$$h(t | x_1, x_2, z) = \lambda_0 \exp(\beta_1 z + \beta_2 x_1 + \beta_3 x_2) \quad (5)$$

に従うものとする。 $\lambda_0 = 0.04$ とした。打ち切りの時間は、区間 (0, 5) における連続一様分布により、生成されるものとした。 X_1, X_2 は、1 段階目の対象者全員に観測されている共変量であるとして、それぞれ標準正規分布に従うものとする。 Z は 2 段階目でのみ観測される 2 値変数であり、

$$\Pr(Z = 1 | x; \gamma) = \frac{\exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 \nu)}{1 + \exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 \nu)} \quad (6)$$

にしたがって生成されるものとする。 V は、 $\Pr(V = 1) = 0.10$ である Bernoulli 変数であり、これも 1 段階目の対象者全員に観測される共変量とする。また、 V は Z と強い相関を持つ代替変数であることを想定しており、 Z 所与のもとでは T^* と条件付き独立であると仮定する。モデル (6) のパラメータは、 $\gamma_0 = -3.5$, $\gamma_1 = 0.50$, $\gamma_2 = 4.5$ とし、周辺的には $\Pr(Z = 1) \approx 0.10$ となるように設定した。 V と Z の関連の大きさと、それぞれの周辺分布は 5 節の事例を模したのとなっている。また、このモデルから、 X_1 は Z と関連がある共変量となり、 X_2 は Z とは独立な共変量となる。

2 段階目のリスクセットサンプリングは、 Z と関連のある変数 X_1 の tertiles で区分される 3 つの水準と V によって分けられる 6 つの層ごとに行った。イベントの発生時点ごとに、対象となるケースと同じ層におけるリスクセットから、1 名のコントロールをサンプリングした。回帰パラメータ $(\beta_1, \beta_2, \beta_3)$ については、以下の 5 つの設定を用いた。① (0.00, 0.30, 0.20), ② (0.60, 0.30, 0.20),

表 1. シミュレーション実験における, 層ごとのデータの要約(1200 回の試行における平均).

設定	$V = 1$					$V = 0$		
		X_1 ; 1st tertile	X_1 ; 2nd tertile	X_1 ; 3rd tertile		X_1 ; 1st tertile	X_1 ; 2nd tertile	X_1 ; 3rd tertile
1 段階目の対象者 (Phase-I samples; $N = 10000$)								
①	Cases (%)	948 (9.5%)	27 (8.1%)	31 (9.3%)	37 (11.1%)	242 (8.1%)	281 (9.4%)	330 (11.0%)
	Total	10000	334	333	333	2999	3000	3001
②	Cases (%)	1017 (10.2%)	41 (12.3%)	48 (14.5%)	57 (17.1%)	246 (8.2%)	287 (9.6%)	338 (11.3%)
	Total	10000	333	332	334	3000	3001	3000
③	Cases (%)	1127 (11.3%)	62 (18.6%)	74 (22.2%)	88 (26.4%)	254 (8.5%)	297 (9.9%)	352 (11.7%)
	Total	10000	333	334	333	3000	2999	3001
④	Cases (%)	1043 (10.4%)	49 (14.7%)	50 (14.9%)	50 (15.0%)	297 (9.9%)	298 (9.9%)	299 (10.0%)
	Total	10000	334	335	333	2999	2998	3001
⑤	Cases (%)	1042 (10.4%)	34 (10.2%)	48 (14.4%)	66 (19.8%)	212 (7.1%)	288 (9.6%)	394 (13.1%)
	Total	10000	333	333	334	3000	3000	3000
Subcohort size	1590	60	ケースコホート研究(4.2 節)におけるサブコホート			60	120	300
						450	600	

③ (1.20, 0.30, 0.20), ④ (0.60, 0.00, 0.60), ⑤ (0.60, 0.60, 0.00). 表 1 に, それぞれの設定のもとでのイベント数の分布をまとめた. それぞれの設定のもとで, 1200 回の反復を行い, $(\beta_1, \beta_2, \beta_3)$ の推定量の平均と標準偏差(SD)の評価を行った.

比較対象としては, 以下の 5 つの方法を検討した.

(a) 1 段階目の完全データに基づく最大部分尤度推定量(Whole cohort; WC)

コホートにおけるすべての情報を利用した推定量であり, 比例ハザードモデル (1) のもとで, 漸近有効な推定量となる (Andersen and Gill, 1982). 以下の解析手法によって, どの程度, 精度が回復されるかを評価するための基準とした.

(b) Thomas (1977) による条件付き最尤推定量(Thomas 法)

Thomas (1977) は, 条件付きロジスティック回帰の尤度関数によって β の推定を行うことを提案しており, これにより, 回帰係数 β の一致推定量を得ることができる. 一般的に, Thomas (1977) の推定量は漸近有効性を持たないが, 一定の正則条件のもとで, それに近い精度を達成することが示されている (Borgan et al., 1995; Goldstein and Langholz, 1992).

(c) Samuelsen (1997) による重みつき推定量(Samuelsen 法)

Samuelsen (1997) は, ネステッドケースコントロール研究を, 共変量 Z が MAR で欠測した不完全データであると見なし, それぞれのデータが観測される確率 (2) の逆数を重みとした, 重みつき部分尤度方程式によって, 回帰係数 β を推定する方法を提案している. いわゆる逆確率重みつき法 (Inverse probability weighting; IPW 法) であり, 簡便な計算によって, β の一致推定量が得られる.

(d) Calibration 法 (Støer and Samuelsen, 2012)

Calibration 法は, 古典的な Horvitz-Thompson (1952) 推定量の精度を改善する方法として, 標本調査法の領域で開発された (Deville and Särndal, 1992). Breslow et al. (2009ab) は, ケースコ

ホート研究における IPW 推定量の改善の方法として, この Calibration 法を応用することを提案しており, Støer and Samuelsen (2012) は, ネステッドケースコントロール研究にも, この方法が有用であることを示している. Calibration 法は, IPW 法の重み (2) を, 1 段階目のコホート全体で測定されている共変量 X の情報を用いた制約 (Calibration 方程式) によって, 修正するという方法である.

(e) Multiple Imputation 法

補完値の生成方法には, 3 節で与えた MCMC, ABB による方法と, Marti and Chavance (2011) の単純な回帰モデルによる方法 (Naïve 法) を用いた. 補完値の生成回数 M は, 共通して 200 回とした. MCMC の Burn-in 期間は, 100 回とした. ABB による方法では, 先に挙げた, 6 つの層ごとにリサンプリングを行った. MCMC における Z の分布についての仮定は, ロジスティック回帰モデル (6) をそのまま用いた. Naïve 法では, X_1 と層を表すダミー変数, イベントの指示変数を共変量としたロジスティック回帰モデルによって, 補完値を生成した.

結果は, 表 2 に示した. それぞれ, 1200 回の試行での推定値の平均と SD を示している. すべての方法で, 回帰係数の推定量には一貫性が成り立つことから予想される通り, バイアスはほとんど生じなかった.

まず, 既存手法の Thomas 法と Samuelsen 法を比較すると, 全般的に Samuelsen 法のほうが高い精度を示しており, Calibration 法では, さらにその推定精度は改善された. これは, Støer and Samuelsen (2012) のシミュレーションなどでも, 同様の結果が報告されている. 一方, Multiple Imputation に基づく方法では, さらなる精度の改善が認められた. まず, 2 段階目でのみ観測される Z に対応する回帰係数 β_1 の精度は, MCMC, ABB, Naïve 法のいずれにおいても, 一貫して改善が認められた. 特に, MCMC による方法が最も精度が高く, ABB や Naïve 法は, それに比べて若干精度は落ちるという結果になった. Naïve 法は, モデルの仮定としては, MCMC と同様に Z の分布にパラメトリックなモデルを仮定しているが, 推定精度では, ABB と同程度, あるいは劣る, といった結果が得られた. β_2, β_3 の推定精度については, Multiple Imputation 法では, 全てのシナリオを通して, WC での完全データを用いた解析の推定精度と同程度にまで, 精度は改善された. これは, Multiple Imputation 法による解析が, X_1, X_2 に関する対象者全体の情報を利用しているからであると考えられる. Thomas 法, Samuelsen 法は, 2 段階目にサンプリングされた対象者のデータをもとに推定関数を構成しており, 2 段階目にサンプリングされなかった対象者のデータは, Samuelsen 法の重みにしか貢献しない. 一方, Calibration 法では, Multiple Imputation 法と同様に, 1 段階目の対象者の共変量情報を β の推定に利用しているので, 同様の推定精度の改善が認められる.

4.2 ケースコホート研究

パラメータの設定と, 1 段階目のコホートの生成は, 4.1 節と同じ方法で行った. 2 段階目のケース, サブコホートのサンプリングも, 4.1 節と同じく, X_1 の tertiles で区分される 3 つの水準と V によって分けられる 6 つの層ごとに行った. サブコホートのサイズは, 表 1 下段のものとした.

表 2. シミュレーション実験の結果: ネステッドケースコントロール研究 (1200 回の試行の平均と SD).

設定	Multiple Imputation 法							
		WC	Thomas 法	Samuelson 法	Calibration 法	MCMC	ABB	Naïve 法
① $\beta_1=0.0$	mean	-0.001	-0.009	0.005	0.000	0.002	-0.017	-0.001
	SD	0.104	0.217	0.140	0.141	0.129	0.130	0.131
	$\beta_2=0.3$	mean	0.303	0.305	0.296	0.300	0.292	0.299
		SD	0.066	0.213	0.074	0.069	0.066	0.062
	$\beta_3=0.2$	mean	0.201	0.202	0.195	0.202	0.195	0.199
		SD	0.064	0.093	0.089	0.070	0.065	0.065
② $\beta_1=0.6$	mean	0.599	0.603	0.604	0.610	0.599	0.600	0.583
	SD	0.084	0.194	0.111	0.112	0.101	0.109	0.104
	$\beta_2=0.3$	mean	0.304	0.299	0.301	0.300	0.301	0.299
		SD	0.063	0.195	0.074	0.063	0.065	0.061
	$\beta_3=0.2$	mean	0.201	0.200	0.203	0.200	0.199	0.201
		SD	0.061	0.092	0.088	0.067	0.062	0.065
③ $\beta_1=1.2$	mean	1.198	1.214	1.191	1.236	1.199	1.199	1.151
	SD	0.069	0.177	0.095	0.099	0.084	0.093	0.087
	$\beta_2=0.3$	mean	0.302	0.302	0.299	0.303	0.295	0.300
		SD	0.059	0.196	0.075	0.064	0.063	0.061
	$\beta_3=0.2$	mean	0.201	0.202	0.204	0.199	0.203	0.199
		SD	0.059	0.087	0.083	0.065	0.058	0.060
④ $\beta_1=0.6$	mean	0.601	0.607	0.605	0.609	0.599	0.593	0.577
	SD	0.085	0.195	0.111	0.112	0.107	0.107	0.104
	$\beta_2=0.0$	mean	-0.003	-0.012	0.004	-0.001	0.001	0.001
		SD	0.062	0.195	0.075	0.065	0.064	0.060
	$\beta_3=0.6$	mean	0.600	0.595	0.601	0.596	0.599	0.598
		SD	0.061	0.096	0.093	0.066	0.060	0.062
⑤ $\beta_1=0.6$	mean	0.596	0.606	0.591	0.613	0.598	0.596	0.588
	SD	0.083	0.200	0.111	0.109	0.100	0.107	0.103
	$\beta_2=0.6$	mean	0.600	0.606	0.597	0.596	0.603	0.602
		SD	0.064	0.197	0.076	0.065	0.063	0.063
	$\beta_3=0.0$	mean	-0.001	-0.004	0.000	0.001	0.000	0.001
		SD	0.061	0.089	0.088	0.067	0.060	0.064

比較対象としては, 以下の 4 つの方法を用いた.

(a) 1 段階目の完全データに基づく最大部分尤度推定量 (Whole cohort; WC)

コホートにおけるすべての情報を利用した推定量であり, 以下の解析手法によって, どの程度, 推定精度が回復されるかを評価するための基準とした.

(b) Borgan et al. (2000) による重みつき推定量 (Borgan 法)

ケースコホート研究の解析方法では, 古くから, 2 段階目のサンプリングのメカニズムを考慮

表 3. シミュレーション実験の結果: ケースコホート研究(1200 回の試行の平均と SD).

設定			WC	Borgan 法	Calibration 法	Multiple Imputation 法		
						MCMC	ABB	Naïve 法
①	$\beta_1=0.0$	mean	-0.001	-0.003	-0.003	0.000	-0.010	0.000
		SD	0.104	0.130	0.124	0.119	0.121	0.116
	$\beta_2=0.3$	mean	0.303	0.302	0.302	0.302	0.299	0.304
		SD	0.066	0.075	0.064	0.064	0.065	0.066
	$\beta_3=0.2$	mean	0.201	0.202	0.201	0.201	0.199	0.200
		SD	0.064	0.088	0.067	0.067	0.067	0.066
②	$\beta_1=0.6$	mean	0.599	0.600	0.596	0.603	0.584	0.589
		SD	0.084	0.109	0.103	0.096	0.098	0.095
	$\beta_2=0.3$	mean	0.304	0.300	0.302	0.298	0.301	0.304
		SD	0.063	0.071	0.062	0.062	0.062	0.063
	$\beta_3=0.2$	mean	0.201	0.203	0.201	0.198	0.200	0.200
		SD	0.061	0.083	0.066	0.065	0.066	0.062
③	$\beta_1=1.2$	mean	1.198	1.206	1.200	1.200	1.166	1.169
		SD	0.069	0.096	0.091	0.083	0.084	0.086
	$\beta_2=0.3$	mean	0.302	0.299	0.302	0.300	0.301	0.304
		SD	0.059	0.071	0.062	0.060	0.059	0.061
	$\beta_3=0.2$	mean	0.201	0.200	0.201	0.207	0.201	0.196
		SD	0.059	0.083	0.066	0.060	0.060	0.061
④	$\beta_1=0.6$	mean	0.601	0.602	0.595	0.604	0.583	0.588
		SD	0.085	0.111	0.105	0.102	0.093	0.094
	$\beta_2=0.0$	mean	-0.003	0.002	0.002	0.002	-0.001	0.002
		SD	0.062	0.074	0.063	0.058	0.063	0.063
	$\beta_3=0.6$	mean	0.600	0.602	0.601	0.599	0.600	0.599
		SD	0.061	0.087	0.066	0.056	0.061	0.062
⑤	$\beta_1=0.6$	mean	0.596	0.603	0.596	0.599	0.582	0.594
		SD	0.083	0.106	0.101	0.095	0.098	0.095
	$\beta_2=0.6$	mean	0.600	0.599	0.602	0.598	0.598	0.603
		SD	0.064	0.074	0.062	0.062	0.061	0.065
	$\beta_3=0.0$	mean	-0.001	0.000	0.002	0.001	-0.003	-0.003
		SD	0.061	0.081	0.064	0.059	0.060	0.062

した重みつき部分尤度方程式による推定方法(IPW 法)が議論されてきた(Barlow et al., 1994; Lin and Ying, 1993; Prentice, 1986; Self and Prentice, 1988). Borgan et al. (2000) は, これらの方法を層別サンプリングの場合に拡張しており, ここでは, Lin and Ying (1993) の方法を拡張した, Borgan らの Type-II 推定量を比較対象に用いた.

(c) Calibration 法(Breslow et al., 2009ab)

4.1 節で解説した通り, Breslow et al. (2009ab) によって提案された方法である. IPW 法の重み (3) を, 1 段階目のコホート全体で測定されている共変量 X の情報を用いて修正した, 重みつ

き部分尤度による推定量である.

(d) Multiple Imputation 法

Multiple Imputation 法は, 4.1 節と同じ方法で実行することとし, 補完値の生成回数 M は 200 回とした. ABB による方法では, 先に挙げた 6 つの層ごとにリサンプリングを行った. MCMC における Z の分布についての仮定は, ロジスティック回帰モデル (6) をそのまま用いた. Naïve 法の補完値の生成には, X_1 と層を表すダミー変数, イベントの指示変数を共変量としたロジスティック回帰モデルを用いた.

結果は, 表 3 に示している. それぞれ, 1200 回の試行での推定値の平均と SD を示した. いずれの方法も, 推定量の一致性が成り立つことから, バイアスはほとんど生じなかった.

ケースコHORT研究の設定でも, Multiple Imputation に基づく方法によって, 従来法よりも, 精度の高い推定値が得られた. まず, 2 段階目でのみ観測される Z に対応する回帰係数 β_1 の精度は, 設定によるが, MCMC, ABB, Naïve 法のいずれにおいても, 従来法と同等か, それ以上の精度が確認された. 特に, MCMC による方法が, 高い精度を持つ傾向が認められた. β_2, β_3 の推定精度についても, Multiple Imputation 法では, 全てのシナリオを通して, WC での完全データを用いた解析の推定精度と同程度の精度が得られた. これは, 4.1 節同様, Multiple Imputation 法による解析が, 対応する共変量 X_1, X_2 の情報を, 1 段階目の対象者全体で利用した推定関数を利用しているからであると考えられる. Calibration 法と Borgan et al. (2000) の Type-II 推定量を比較すると, 1 段階目の対象者全体での付加的な共変量情報を用いているため, Calibration 法のほうが高い推定精度を示した.

5. National Wilms Tumour Studies

Breslow et al., (2009b), Kulich and Lin (2004), Liu et al. (2010) は, 米国の National Wilms Tumor Study (NWTs) において, 1980 年から 1994 年の間に Wilms 腫瘍と診断された患者を対象とした 3915 名のコHORTを利用して, 仮想的なネステッドケースコントロール研究, ケースコHORT研究を行い, 種々の解析手法の性能の評価を行っている. Wilms 腫瘍患者では, 「unfavourable histology (UH)」とされる特殊な細胞型の腫瘍を持つ少数の患者が, 「favourable histology (FH)」と診断される患者に比べ, 治療後の再発や死亡が起こりやすいことが報告されている (Beckwith and Palmer, 1978). NWTs では, 100 以上の参加施設で組織診断が行われた後, 中央の研究施設に検体が送られ, 一人の病理専門家によって中央診断が行われている. 両者の診断結果にはある程度, 不一致があることが確認されており, 施設での診断結果の精度は, 全体で感度 74%, 特異度 98%であったとされている. 施設診断に比べ, 中央診断のほうが診断の誤差は小さいと考えられるが, 一方で追加コストが必要となる. 先に挙げた一連の研究では, 中央での診断結果を 2 段階目で測定される共変量, 施設ごとの診断結果をその代替変数とした評価が行われている. 本節では, 同様に, この NWTs によるコHORTを 1 段階目のコHORTとしたネステッドケースコントロール研究を行い, Multiple Imputation 法による解析を行う.

2 段階目のリスクセットサンプリングは, 施設病理診断の結果 (UH, FH), 進行度 (Stage I, II or

III, IV), 年齢(1歳未満, 以上)に基づく8つの層による, 層別サンプリングによって行う. それぞれのイベント発生時点において, 対応する層から, 1人のコントロールをサンプリングした(1:1のマッチング). 仮想的な2段階目のサンプリングを1000回行い, それぞれの試行で, 4.1節と同じく, WCの完全データを用いた解析, Thomas (1977) 法, Samuelsen (1997) 法, Calibration 法(Støer and Samuelsen, 2012), Multiple Imputation 法による解析を行った. 解析には, 共変量に中央病理診断の結果(UH, FH), 年齢(区分線形関数による適当な変換を行っている; Age0 が1歳未満, Age1 が1歳以後に対応する), 進行度(Stage I, II or III, IV), 腫瘍径(Diameter)と, 3つの交互作用項(進行度×腫瘍径, 中央病理診断×Age0, 中央病理診断×Age1)をモデル化したCox回帰モデルを用いた. 2節の表記法に対応させると, 施設病理診断の結果, 年齢, 進行度, 腫瘍径が X , 中央病理診断の結果が Z となる.

Multiple Imputation 法のMCMCによる補完値の生成では, 2段階目の変数 Z に対応する中央病理診断のモデルに, ロジスティック回帰モデルを用いた. 説明変数には, 施設病理診断(UH, FH), 年齢, 進行度, 腫瘍径と進行度×腫瘍径の交互作用項を用い, Burn-in 期間は100回とした. ABB法でのリサンプリングを行う層は, 4節と同じく, 2段階目のサンプリングで用いた8つの層をそのまま用いた. Naïve法の補完値の生成には, 同様に, 層を表すダミー変数, イベントの指示変数と年齢, 腫瘍径, 進行度×腫瘍径の交互作用項を説明変数としたロジスティック回帰モデルを用いた. それぞれ $M = 200$ とした.

表4に, それぞれの解析方法による1000回の回帰係数の推定値の平均と標準偏差(SD), 漸近的な標準誤差(asymptotic standard error; ASE)の推定値の平均を示している. ただし, WCの完全データの解析結果には, 1段階目のコホートの対象者全員の完全データに対してのCox回帰の最大部分尤度による推定値と, そのSEの推定値を示している. Thomas法では, 進行度は, マッチング変数として用いられているため, モデルの説明変数には含めていない. 結果としては, Samuelsen法, Calibration法, Multiple ImputationのMCMC, ABBでは, 推定値の平均がWCの推定値に概ね近い値をとったが, Thomas法, Multiple ImputationのNaïve法では, 一定のバイアスが認められた. Thomas法のバイアスについては, Liu et al. (2010)の解析でも, 同様の傾向が認められている. Thomas法は, 推定精度(SD, ASE)も, 他の方法に比べて大きく劣っている.

特に関心がある, 中央病理診断についての回帰係数の推定精度を比較すると, やはりMultiple Imputationのほうが小さなSDを示した. Naïve法は, WCの結果から, 0.50ほどのバイアスがあるが, SDは小さい. ABBは, MCMCやNaïve法, また, Calibration方程式にパラメトリックな制約を用いるCalibration法に比べると, SDは大きい, それでも, Thomas法, Samuelsen法に比べ, 大きな改善が認められた. ただし, Naïve法のASEの推定値は, SDの大きさに比べると相対的に大きく, 誤差を過大に評価している可能性がある.

一方, 1段階目で測定される共変量についての主効果のパラメータの推定値は, 全体を通して, Thomas法以外は, 概ねWCの結果と一致している. 推定精度については, Multiple Imputationでは, WCの結果とほぼ同等のASEの推定値が得られており, SDも小さい. 従来法のThomas法, Samuelsen法は, かなりの精度の損失が認められ, これは, 4節と同様, Multiple Imputation

表 4. NWTS データの解析結果: ネステッドケースコントロール研究 (1000 回の試行における平均).

		Multiple Imputation 法						
		WC	Thomas 法	Samuelson 法	Calibration 法	MCMC	ABB	Naïve 法
Histology	推定値	4.042	2.719	4.065	4.078	4.265	3.887	3.513
	SD	—	0.867	0.343	0.178	0.102	0.257	0.163
	ASE	0.413	1.111	0.628	0.540	0.435	0.479	0.645
Age ₀	推定値	-0.661	-0.213	-0.689	-0.653	-0.625	-0.652	-0.683
	SD	—	0.517	0.250	0.041	0.018	0.031	0.020
	ASE	0.326	0.709	0.452	0.331	0.327	0.328	0.331
Age ₁	推定値	0.104	0.116	0.106	0.105	0.106	0.103	0.105
	SD	—	0.022	0.018	0.006	0.002	0.001	0.001
	ASE	0.017	0.029	0.024	0.017	0.016	0.016	0.017
Stage	推定値	1.346	—	1.336	1.333	1.241	1.424	1.381
	SD	—	—	0.240	0.084	0.055	0.056	0.048
	ASE	0.244	—	0.350	0.271	0.253	0.249	0.266
Diameter	推定値	0.069	0.058	0.069	0.070	0.068	0.071	0.070
	SD	—	0.016	0.015	0.005	0.002	0.001	0.002
	ASE	0.014	0.021	0.021	0.015	0.014	0.014	0.014
Stage×Diameter	推定値	-0.076	-0.070	-0.075	-0.054	-0.072	-0.083	-0.080
	SD	—	0.023	0.020	0.045	0.004	0.003	0.003
	ASE	0.019	0.030	0.028	0.049	0.020	0.020	0.021
Histology×Age ₀	推定値	-2.635	-1.527	-2.642	-2.680	-2.731	-2.561	-2.039
	SD	—	0.857	0.401	0.245	0.151	0.246	0.166
	ASE	0.464	1.099	0.698	0.607	0.500	0.535	0.693
Histology×Age ₁	推定値	-0.058	-0.025	-0.056	-0.075	-0.074	-0.043	-0.070
	SD	—	0.055	0.045	0.006	0.022	0.012	0.014
	ASE	0.034	0.065	0.053	0.021	0.042	0.038	0.042

法のほうが, 一貫して高い精度を示している. 交互作用項についても, 同様に Multiple Imputation 法による精度の改善が認められた. ただし, Naïve 法では, 少しバイアスの大きな項が見受けられた.

6. 考 察

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研究の歴史は, 1970 年代にまで遡り, 多くの理論的な成果の蓄積とともに, これらの方法論は, 既に, 実用性の高い水準にまで確立されたといえる. 近年では, 疫学研究の実践においても, 広く普及しつつあり (Breslow et al., 2009b; Cologne et al., 2012; Langholz and Richardson, 2009), 今後, ますます, 実践的な場面で, 重要な研究手法になるものと思われる.

本稿では, Multiple Imputation 法による, これらのデザインのもとでの有効な曝露効果の推定方法を与えた. 近年の研究では, 先述の通り, 局外要因である共変量 X, Z の分布に特定の分布型を仮定しないセミパラメトリック有効な方法が活発に議論されているが, Breslow and Wellner (2007a) は, これらの方法論の研究の現状に対して次のように述べている.

Most of these methods are relatively recent and involve sufficiently complex calculations, or sufficiently restrictive assumptions, that none have yet seen widespread use. These limitations are certain to decline with advances in computing hardware and software, making more efficient estimation methods more widely available. In the meantime, the WL (weighted likelihood) estimation procedure outlined at the end of section 6 offers a relatively simple and robust alternative. (Breslow and Wellner, 2007a; p.97)

その後, Breslow et al. (2009ab) は, 標本調査理論における Calibration 法の応用を議論しているが, Multiple Imputation 法は, セミパラメトリック有効な方法の代替的な方法としてではなく, 望ましい理論的な性質と実践的な利便性を兼ね備えた, 有用な方法であると考えられる. Multiple Imputation 法は, 長い研究の歴史を持ち (Rubin, 1996), 不完全データの解析における標準的な方法として, 既に, 医学研究の実務家の間でも広く普及している (Ibrahim et al., 2012; Klebanoff and Cole, 2008; Little et al., 2012; Sterne et al., 2009; Wood et al., 2004). その原理も, 「欠測値に適当な値を代入する」という直感的にわかりやすいものであり, 難解なセミパラメトリック推測の方法論に比べると, 医学研究の実務家にも理解しやすいものと思われる. また, ネステッドケースコントロール研究やケースコホート研究の設定では, 欠測は「デザインによって生じる」ものであり, MAR であることが保証されている. パラメトリックな補完値の生成方法では, モデルの誤特定によって, 推定量のバイアス・精度の問題は生じるが (Marti and Chavance, 2011), ABB 法などを用いることによって, セミパラメトリック有効な方法と同等の仮定のもとでの推測を行うこともできる. 解析を実行する上でも, SAS (SAS Institute, Cary, NC) の PROC MI, PROC MIANALYZE のように, 多くの標準的なソフトウェアに解析用のモジュールが付属しているという利点がある (Harel and Zhou, 2007; Royston and White, 2011; White et al., 2011).

また, Multiple Imputation 法は, 2 段階目で測定される共変量 Z が複数ある場合や, より複雑な多段階のサンプリング (Lee et al., 2010) が行われる場合にも, サンプリングの構造を「欠測メカニズム」として導入すればよいというだけで, 統一的な取り扱いが可能である. 特に, MICE (Multiple Imputation by Chained Equation; Royston and White, 2011; White et al., 2011) は, これらの設定のもとでも有効な方法であると思われる. これらの方法論については, 今後, 実践での応用を通して, その有用性の評価が行われることが望まれる. 一方で, 本稿における議論から, Multiple Imputation 法は, 漸近有効性を持ちながら, このような複雑なサンプリングデザインにも柔軟に対応することができる方法として, 実践でも広く有用な方法として利用することができると思われる.

謝 辞

本研究は, 科学研究費補助金 (課題番号: 24800081), 統計数理研究所共同利用研究 (課題番号: 24-共研-2047) の助成を受けた.

参考文献

- Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100–1120.
- Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology* **52**, 1165–1172.
- Barnard, J., Rubin, D. B., and Schenker, N. (2005). Multiple imputation methods. In *Encyclopedia of Biostatistics* (2nd edn), P. Armitage, and T. Colton (eds), pp. 3420–3428. Chichester: Wiley.
- Beckwith, J. B., and Palmer, N. F. (1978). Histopathology and prognosis of Wilms tumor. *Cancer* **41**, 1937–1948.
- Borgan, Ø., Goldstein, D. R., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Annals of Statistics* **23**, 1749–1778.
- 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. (2009a). Improved Horvitz-Thompson estimation of model parameters from two-phases stratified samples: Applications in epidemiology. *Statistics in Bioscience* **1**, 32–49.
- Breslow, N. E., Lumley, T., Ballantyne, C. M., Chambless, L. E., and Kulich, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* **169**, 1398–1405.
- Breslow, N. E., and Wellner, J. A. (2007a). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* **34**, 86–102.
- Breslow, N. E., and Wellner, J. A. (2007b). A Z-theorem with estimated nuisance parameters and correction note for 'Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression.' *Scandinavian Journal of Statistics* **34**, 86–102.
- Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2006). Posterior propriety and computation for the Cox regression model with applications to missing covariates. *Biometrika* **93**, 791–807.
- Cologne, J., Preston, D. L., Imai, K., et al. (2012). Conventional case-cohort design and analysis for studies of interaction. *International Journal of Epidemiology* **41**, 1174–1186.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.
- Deville, J. C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

- Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.
- Gilks, W. R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.
- Goldstein, D. R., and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Annals of Statistics* **20**, 1903–1928.
- Harel, O., and Zhou, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine* **26**, 3057–3077.
- Horvitz, D., and Thompson, D. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association* **47**, 663–685.
- Ibrahim, J. G., Chu, H., and Chen, M. H. (2012). Missing data in clinical studies: issues and methods. *Journal of Clinical Oncology* **30**, 3297–3303.
- Kalbfleisch, J. D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214–221.
- Kalbfleisch, J. D., and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* **7**, 149–160.
- Klebanoff, M. A., and Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology* **168**, 355–357.
- Kulich, M., and Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-control studies. *Journal of the American Statistical Association* **99**, 832–844.
- Langholz, B. (2005). Case-control study, nested. In *Encyclopedia of Biostatistics* (2nd edn), P. Armitage, and T. Colton (eds), pp. 646–655. Chichester: Wiley.
- Langholz, B., and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling. *Biometrika* **82**, 69–79.
- Langholz, B., and Richardson, D. (2009). Are nested case-control studies biased? *Epidemiology* **20**, 321–329.
- Lavori, P. W., Dawson, R., and Shera, D. A. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine* **14**, 1913–1925.
- Lee, A. J., Scott, A. J., and Wild, C. J. (2010). Efficient estimation in multi-phase case-control studies. *Biometrika* **97**, 361–374.
- Lin, D. Y., and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* **88**, 1341–1349.
- Lin, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* **89**, 958–966.

- Little, R. J., D'Agostino, R., Cohen, M. L., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* **367**, 1355–1360.
- Liu, M., Lu, W., and Tseng, C.-H. (2010). Cox regression in nested case-control studies with auxiliary covariates. *Biometrics* **66**, 374–381.
- Lumley, T., Shaw, P. A., and Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review* **79**, 200–220.
- Marti, H., and Chavance, M. (2011). Multiple imputation analysis of case-cohort studies. *Statistics in Medicine* **30**, 1595–1607.
- Nan, B. (2004). Efficient estimation for case-cohort studies. *Canadian Journal of Statistics* **32**, 403–419.
- Oakes, D. (1981). Survival times: aspects of partial likelihood. *International Statistical Review* **49**, 235–264.
- Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1–11.
- Qi, L., Wang, C. Y., and Prentice, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **100**, 1250–1263.
- 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.
- Robins, J. M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
- Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology* (3rd edn.). Philadelphia: Lippincot Williams & Wilkins.
- Royston, P., and White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software* **45**, Issue 4.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489.

- Rubin, D. B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* **81**, 366–374.
- Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control data. *Biometrika* **84**, 379–394.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3–15.
- Self, S. G., and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* **16**, 64–81.
- Sinha, D., Ibrahim, J. G., and Chen, M.-H. (2003). A Bayesian justification of Cox’s partial likelihood. *Biometrika* **90**, 629–641.
- Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal* **338**, b2393.
- Støer, N. C., and Samuelsen, S. O. (2012). Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Analysis* **18**, 261–283.
- 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.
- Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
- Wang, N., and Robins, J. M. (1998). Large sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 377–399.
- Wood, A., White, I. R., Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomised controlled trials. *Clinical Trials* **1**, 368–376.
- Zeng, D., and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B* **69**, 507–536.