

科学研究費助成事業 研究成果報告書

平成 29 年 6 月 28 日現在

機関番号：14201

研究種目：基盤研究(C) (一般)

研究期間：2013～2016

課題番号：25330039

研究課題名(和文) 欠測を伴うコホートに対するケース・コホートデザインの適用と解析方法の開発

研究課題名(英文) Development of statistical method for case-cohort design when measurements of interests are known to be missing

研究代表者

和泉 志津恵(大久保志津恵)(Izumi, Shizue)

滋賀大学・データサイエンス教育研究センター・教授

研究者番号：70344413

交付決定額(研究期間全体)：(直接経費) 3,700,000円

研究成果の概要(和文)：ケース・コホート研究では、コホート全体ではなく一部の選択された対象者のみから高価なゲノム情報を測定することにより、研究のコスト・労力が大幅に節減できる。しかし、対象者から観測されるデータには頻りに欠測値が含まれるため、データ解析に困難を伴う。本研究では、欠測を伴うコホートデータに対応した、ケース・コホート研究の新規な理論的枠組みを開発した。そして、大規模なゲノム疫学研究への提案方法の応用を検討した。数値実験により検証した結果から、コホート内の大部分の対象者のアウトカムが欠測となる場合に、既存の方法でサブコホートの抽出やデータ解析を行うと、興味ある変数の効果にバイアスが生じることが示唆された。

研究成果の概要(英文)：In the case-cohort study, costly and laborious research can be greatly saved by measuring expensive genome information only from selected subjects, not the entire cohort. However, since data observed from the target frequently includes missing values, data analysis becomes difficult. In this study, we developed a novel theoretical framework for case-cohort studies when measurements of interests are known to be missing. We also examined the application of the proposed method to large-scale genomic epidemiological studies. From results verified by numerical experiments, when the outcomes of the majority of the subjects in the cohort are missing, the sub-cohort extraction or data analysis performed by the traditional method leads bias in the effects of the interesting variable.

研究分野：総合領域

キーワード：医薬生物 ゲノム統計解析 ビッグデータ活用 ゲノム疫学

1. 研究開始当初の背景

(1) 国内の研究動向

近年の分子生物学の発展により、膨大なゲノムデータが測定され、疾病に関連する遺伝子の探索が可能になっている。たとえば、大規模コホート研究のひとつである Japan Collaborative Cohort Study for Evaluation of Cancer Risk (JACC Study) では、玉腰(当該課題の研究分担者)が中心となり、疾病に関連する遺伝子や生活習慣を探索し、顕著な研究成果を発表している[引用文献 1-4]。

ケース・コホート研究では、コホート全体ではなく一部の選択された対象者のみから高価なゲノム情報を測定することにより、研究のコスト・労力が大幅に節減できる[図 1]。そのため、ゲノム疫学、遺伝疫学、薬剤疫学などの幅広い分野において、ケース・コホート研究が活発に実施されている[引用文献 5]。しかし、対象者から観測されるデータには、疾病の有無などの結果変数、ゲノム情報や喫煙などの説明変数に欠測値が含まれる場合、既存の方法ではデータ解析に困難を伴う。それゆえに、欠測を伴うコホートデータに対応した、ケース・コホート研究の新規な理論的枠組みの開発が、国内外の研究現場では期待されている。

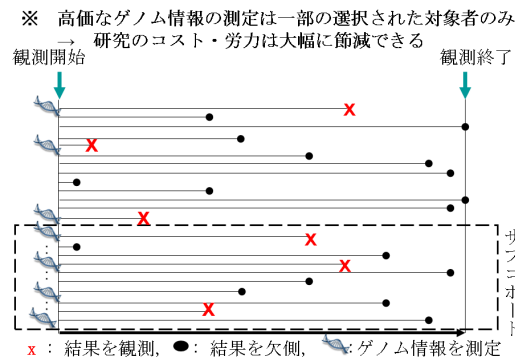


図 1. ケース・コホート研究の概念図

(2) 国外の研究動向

欠測を伴わないコホートに対するケース・コホート研究では、two-phase study デザインとしてとらえ[引用文献 6]、サブコホートの抽出確率の逆数を重みとした Inverse Probability Weight (IPW)法によりデータ解析を行う[引用文献 7-8]。一方、欠測データに対応した、ケース・コホート研究の理論的枠組みは開発されていない。

2. 研究の目的

近年、膨大なゲノムデータを用いて、疾病に関連する遺伝子の探索が可能になっている。ケース・コホート研究では、コホート全体ではなく一部の選択された対象者のみから高価なゲノム情報を測定することにより、研究のコスト・労力が大幅に節減できる。しかし、対象者から観測されるデータには頻繁

に欠測値が含まれるため、データ解析に困難を伴う。本研究では、欠測を伴うコホートデータに対応した、ケース・コホート研究の新規な理論的枠組みの開発を行い、提案方法の特性を数値実験により検証し、汎用アプリケーションを開発し、国内外の大規模なゲノム疫学研究への提案方法の応用を検討することを目的とする。

3. 研究の方法

(1) サンプルデザインを開発する。コホートでの欠測の発生メカニズムを Missing at random と仮定する。その仮定のもと、Breslow & Chatterjee (1999)により提案された Two-phase study の枠組みを拡張して Multi-phase 抽出法についての研究を行う。

(2) 統計解析の方法を開発する。Borgan ら (2000)による方法を拡張し、新たなサンプリングデザインを考慮して、IPW 法に基づく統計解析の方法を開発する。

(3) 数値実験により有用性を評価する。JACC Study などの大規模コホート研究のゲノムデータを用いて、統計解析向けプログラミング言語の Matlab や R において数値実験を行い、実践における有用性の評価を行う。

(4) 標本数の設計方法を構築する。Izumi & Fujii (2010)による方法を拡張して、新規なサンプリングデザインと新たに開発した統計解析の方法に基づく、標本数の設計方法を構築する。

(5) 数値実験により標本数の設計方法を検証する。統計解析向けプログラミング言語の Matlab や R において数値実験を行い、設計方法の特性を調べる。

(6) アプリケーションを開発する。提案方法に対する汎用アプリケーションを Matlab や R において開発し、IT を活用して Web 上で研究内容やアプリケーションに関する情報を発信する。

(7) ゲノム疫学研究へ提案方法を応用する。数値実験による検証結果に基づいて、国内外の大規模なゲノム疫学研究への提案方法の応用を検討する。

(8) 研究成果を考察する。本研究において開発する方法と数値実験による検証結果をまとめる。研究成果を統計関連学会連合大会や国際的な学術会議において発表する。さらに、国際学会論文誌に研究論文を投稿し、本研究を総括する。

4. 研究成果

(1) 欠測データに対応した、ケース・コホート研究の新規の理論的枠組みの開発を行

った。まず、初年度には、コホートでの欠測の発生メカニズムを Missing at random と仮定した。その仮定のもと、Breslow & Chatterjee(1999)により提案された two-phase sampling の枠組みを拡張して研究効率の高い、新規なサンプリングデザインを検討した。子どもの健康と環境に関する全国調査（環境省エコチル調査）の元となった北海道研究（北海道スタディ）のデータを参考にして、欠測のパターンを分類し、Multi-phase 抽出法について検討した。研究分担者や連携研究者から得られたコメントに基づき、サンプリングデザインの修正や改良をさらに検討した。

次年度には、乳幼児を対象とした北海道スタディやエコチル調査のデータの特徴を参考にして、コホートでの欠測の発生メカニズムを、初年度に仮定した Missing at random の場合に加えて、コホート内の対象者全員から得られる追跡開始時の共変量に依存する場合、さらに、サブコホート内の対象者のみから得られる共変量にも依存する場合の3通りに拡張して検討した。

(2) 解析方法として、抽出確率の逆数を重みに用いた部分尤度法、推定された重みを用いた部分尤度法、多重代入法のうち Multiple Imputation by Chained Equations (MICE)について検討した。加えて、因果推論的なアプローチである Structural Mean Models, 多段階発がん過程に基づく数理モデル、経時データに対する変化係数を用いたモデルについても検討した。

(3) JACC Study や北海道スタディのデータの特徴を参考にして仮想データを作成し、R や SAS などの統計解析向けプログラミング言語を用いた数値実験を行い、欠測のパターンごとの母数の推定値のバイアスについて評価した。

(4)(1) や (2) の結果に基づいて、標本数（サブコホートサイズ）の設計方法を構築した。

(5) 実データの特徴に基づく仮想データを作成し、その仮想データを用いて、数値実験により標本数（サブコホートサイズ）の設計方法を検証した。

(6) 統計解析向けプログラミング言語の R や SAS を用いて、提案方法に対する汎用アプリケーションのプロトタイプを開発した。

(7)(1) から (6) までの結果を用いて、ゲノム疫学研究への応用を試みた。分子疫学研究へ既存の方法に併せて提案方法を解説し、大規模なゲノム疫学研究におけるサンプリングデザインを含む研究計画を作成した。その際、乳幼児を対象とした北海道スタディやエコチル調査のように、コホート内の

多数の対象者における反応変数の値が欠測となると事前に分かっている場合、提案方法の代わりに、反応変数の値を補完する多重代入法 (Multiple Imputation) の方が適切となる可能性がでてきた。そこで、多重代入法を用いるアプローチのもとで、サンプリングデザインを検討した。

(8)(1) から (7) までの研究成果の考察を行った。乳幼児を対象とした北海道スタディやエコチル調査のように、コホート内の多数の対象者における反応変数の値が欠測となると事前に分かっている場合、幾つかのシナリオのもとで、反応変数が欠測となる場合の影響を補正し、共変量の効果を推定する方法を用いた研究デザインを提案した。その成果を国際学会にて報告し、論文にまとめた。

この他、本研究に間接的に関係するものとして、2 段階ケース・コントロール研究のデザイン、疫学研究への統計的方法に関する成果も研究発表欄にリストしている。

<引用文献>

- 1 . Tamakoshi A, et al., for the JACC Study Group. (2012). Multiple roles and all-cause mortality: the Japan Collaborative Cohort Study. *European Journal of Public Health*, 23 (1): 158-164.
- 2 . Tamakoshi A, et al. (2011). Effect of coffee consumption on all-cause and total cancer mortality: findings from the JACC study. *European Journal of Epidemiology*, 26(4): 285-93.
- 3 . Tamakoshi A, et al, for the JACC Study Group. (2010). Impact of smoking and other lifestyle factors on life expectancy among Japanese: findings from the Japan Collaborative Cohort (JACC) Study. *Journal of Epidemiology*, 20(5): 370-376.
- 4 . Tamakoshi A, et al. for the JACC Study Group. (2010). Relationship of sFas with metabolic risk factors and their clusters. *European Journal of Clinical Investigation*, 40(6): 527-533.
- 5 . 久保田潔 . (2006). ケース・コホート研究と PMS . *薬剤疫学*, 11(1): 23-34.
- 6 . Breslow NE, Chatterjee N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C*, 48(4): 457-468.
- 7 . Barlow WE, Ichikawa L, Rosner D, Izumi S. (1999). Analysis of case - cohort designs. *Journal of Clinical Epidemiology*, 52: 1165 - 1172.
- 8 . Borgan Ø, Langholz B, Samuelsen SO, Goldstein DR, and Pogoda J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6: 39-58.

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

〔雑誌論文〕(計20件)

野間久史、田中司朗、田中佐智子、和泉志津恵、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、計量生物学、33、101-124、2013、査読有。

Tanaka S, Fukinbara S, Tsuchiya S, Suganami H, Ito YM, Current practice for the prevention and treatment of missing data in confirmatory clinical trials: A survey of Japan-based and foreign-based pharmaceutical manufacturers in Japan, Therapeutic Innovation & Regulatory Science, 48, 717-23, 2014, 査読有。DOI: 10.1177/2168479014530974.

Iguchi Y, Ito YM, Kataoka F, Nomura H, Tanaka H, Chiyoda T, Hashimoto S, Nishimura S, Takano M, Yamagami W, Susumu N, Aoki D, Tsuda H, Simultaneous analysis of the gene expression profiles of cancer and stromal cells in endometrial cancer, Genes Chromosomes Cancer, 53, 725-737, 2014, 査読有。DOI:10.1002/gcc.22182.

Yamaguchi S, Terasaka S, Kobayashi H, Asaoka K, Motegi H, Nishihara H, Kanno H, Onimaru R, Ito YM, Shirato H, Houkin K, Prognostic factors for survival in patients with high-grade meningioma and recurrence-risk stratification for application of radiotherapy, PLoS One, 9, e97108, 2014, 査読有。DOI: 10.1371/journal.pone.0097108.

Kawamoto T, Nitta N, Murata K, Toda E, Tsukamoto N, Hasegawa M, Yamagata Z, Kayama F, Kishi R, Ohya Y, Saito H, Sago H, Okuyama M, Ogata T, Yokoya S, Koresawa Y, Shibata Y, Nakayama S, Michikawa T, Takeuchi A, Satoh H, Rationale and study design of the Japan environment and children's study (JECS), BMC Public Health, 14, 25, 2014, 査読有。DOI: 10.1186/1471-2458-14-25.

Ohishi W, Cologne JB, Fujiwara S, Suzuki G, Hayashi T, Niwa Y, Akahoshi M, Ueda K, Tsuge M, Chayama K, Serum interleukin-6 associated with hepatocellular carcinoma risk: a nested case-control study, International Journal of Cancer, 134, 154-163, 2014, 査読有。DOI: 10.1002/ijc.28337.

高橋邦彦、和泉志津恵、竹内文乃、位置情報を用いた疫学研究とその統計的方法、統計数理 特集号「疫学研究のデザインとデータ解析：最近の理論的展開と実践」、62、3-24、2014、査読有。

野間久史、ケースコホート研究の理論と統計手法、統計数理 特集号「疫学研究のデザインとデータ解析：最近の理論的展開と実践」、62、25-44、2014、査読有。

竹内文乃、野間久史、観察研究におけるバイアスの感度解析、統計数理 特集号「疫学研究のデザインとデータ解析：最近の理論的展開と実践」、62、77-92、2014、査読有。

和泉志津恵、佐藤健一、川野徳幸、経時的に観測されたテキストデータに対する変化係数モデルに基づく統計的な分類方法と視覚化について、計算機統計学、28、81-92、2015、査読有。

Lin Y, Nishiyama T, Kurosawa M, Tamakoshi A, Kubo T, Fujino Y, Kikuchi S; JACC Study Group, Association between shift work and the risk of death from biliary tract cancer in Japanese men, BMC Cancer, 15, 757, 2015, 査読有。DOI: 10.1186/s12885-015-1722-y.

Lin Y, Obata Y, Kikuchi S, Tamakoshi A, Iso H; JACC Study Group, Helicobacter pylori infection and risk of death from cardiovascular disease among the Japanese population: a nested case-control study within the JACC Study, Journal of Atherosclerosis and Thrombosis, 22, 1207-1213, 2015, 査読有。DOI: 10.5551/jat.27987.

Michikawa T, Nitta H, Nakayama SF, Ono M, Yonemoto J, Tamura K, Suda E, Ito H, Takeuchi A, Kawamoto T; Japan Environment and Children's Study Group, The Japan Environment and Children's Study (JECS): A preliminary report on selected characteristics of approximately 10,000 pregnant women recruited during the first year of the study, Journal of Epidemiology, 25, 452-458, 2015, 査読有。DOI: 10.2188/jea.JE20140186.

Izumi S, Sakata R, Yamada M, Cologne J, Interaction between a single exposure and age in cohort-based hazard rate models impacted the statistical distribution of age at onset, Journal of Clinical Epidemiology, 71, 43-50, 2016, 査読有。DOI: 10.1016/j.jclinepi.2015.10.004.

Satoh K, Tonda T, Izumi S, Logistic regression model for survival time analysis using time-varying coefficients, American Journal of Mathematical and Management Sciences, 35, 353-360, 2016, 査読有。DOI: 10.1080/01966324.2016.1215945

Goudarzi H, Miyashita C, Okada E, Kashino I, Kobayashi S, Chen CJ, Ito S, Araki A, Matsuura H, Ito YM, Kishi R, Effects of prenatal exposure to perfluoroalkyl acids on prevalence of allergic diseases among 4-year-old children, Environment

International, 94, 124-132, 2016, 査読有 . DOI:10.1016/j.envint.2016.05.020.

Ochi N, Yoshinaga K, Ito YM, Tomiyama Y, Inoue M, Nishida M, Manabe O, Shibuya H, Shimizu C, Suzuki E, Fujii S, Katoh C, Tamaki N, Comprehensive assessment of impaired peripheral and coronary artery endothelial functions in smokers using brachial artery ultrasound and oxygen-15-labeled water PET, Journal of Cardiology, 68, 316-23, 2016, 査読有 . DOI: 10.1016/j.jcc.2015.10.006.

富田哲治, 佐藤健一, 和泉志津恵, 広島平和宣言における単語出現頻度に基づく広島市の平和観の経時変化について, 長崎医学雑誌, 91, 176-179, 2016, 査読有 . Taguri M, Izumi S, A global goodness-of-fit test for linear structural mean models, Behaviormetrika, 44, 253-262, 2017, 査読有 . DOI: 10.1007/s41237-016-0003-7.

Taguri M, Izumi S, Erratum to: A global goodness-of-fit test for linear structural mean models, Behaviormetrika, 44, 263, 2017. DOI: 10.1007/s41237-016-0012-6.

[学会発表](計21件)

伊藤陽一, 欠測データに対する解析手法の概説, 2013年度日本計量生物学会年会・特別セッション(招待講演), 2013年05月23日~2013年05月24日, パルセイロいざか(福島県福島市).

Noma H, Tanaka S, Tanaka S, Izumi S, Multiple imputation analysis of nested case-control and case-cohort studies, The 46th Society for Epidemiologic Research (SER) Annual Meeting, 2013年06月18日~2013年06月21日, Boston Park Plaza Hotel (Boston, Massachusetts, U.S.A.).

Ito YM, Some examples of the treatment of missing data in investigator-initiated clinical trials, Joint Conference of the Fifth Annual International Symposium on the Evaluation of Clinical Trials (Methodologies and Applications) and the Fourth East Asia Regional Biometric Conference 2013 (招待講演), 2013年07月04日~2013年07月07日, Renmin University of China (Beijing, China).

和泉志津恵, 佐藤健一, 松浦陽子, 川野徳幸, 経時的テキストデータにおける出現パターンの抽出と分類, 2013年度統計関連学会連合大会, 2013年09月08日~2013年09月11日, 大阪大学(大阪府豊中市).

和泉志津恵, 大瀧慈, 合原一幸, 放射線への連続曝露の場合の多段階発がん数理モデルの構築と超過リスクの定量的評価, 2013年度統計関連学会連合大会, 2013年09月08日~2013年09月11日, 大阪大学(大阪府豊中市).

野間久史, 田中司朗, ケースコホート研究の統計解析: 2段階ケースコントロール研究との等価性と漸近有効な推定方式, 日本行動計量学会第41回大会, 2013年09月03日~2013年09月06日, 東邦大学(千葉県船橋市).

和泉志津恵, 永田大貴, 伊藤陽一, 2段階ケース・コホート研究における標本サイズの設計, 2014年度日本計量生物学会年会, 2014年5月23日-24日, 統計数理研究所(東京都立川市).

永田大貴, 和泉志津恵, 伊藤陽一, 2段階ケース・コントロール研究における検出力の算出とその評価, 2014年度日本計量生物学会年会, 2014年5月23日-24日, 統計数理研究所(東京都立川市).

Ito YM, Izumi S, Comparison of sampling schemes for a case-cohort design in Hokkaido cohort study when some outcomes of interests are known to be missing, The XXVII International Biometric Conference (IBC2014), 2014年07月05日~2014年07月11日, Firenze Fiera (Firenze, Florence, Italy).

Taguri M, Izumi S, A goodness-of-fit test for linear structural mean models, The XXVII International Biometric Conference (IBC2014), 2014年07月05日~2014年07月11日, Firenze Fiera (Firenze, Florence, Italy).

Izumi S, Satoh K, A new approach for statistical classification and visualization for longitudinal text data, The XXVII International Biometric Conference (IBC2014), 2014年07月05日~2014年07月11日, Firenze Fiera (Firenze, Florence, Italy).

Izumi S, Ohtaki M, Aihara K, Innovative mathematical modeling for the effects of chronic exposure to radiation on cancer risk, The XXVII International Biometric Conference (IBC2014), 2014年07月05日~2014年07月11日, Firenze Fiera (Firenze, Florence, Italy).

和泉志津恵, 田栗正隆, 線形構造平均モデルにおける適合性検定についての検討, 2014年度統計関連学会連合大会, 2014年09月13日~2014年09月16日, 東京大学(東京都文京区).

Izumi S, Tonda T, Satoh K, Statistical inference for linear varying coefficients in Cox proportional hazard model, Kyoto International Conference on Modern Statistics in the 21st Century, 2014年11月17日~2014年11月18日, Kyoto International Conference Center (Kyoto, Japan).

Cologne J, Izumi S, Sakata R, Yamada M, Interaction between exposure and age in cohort-based risk models: effect

modification or age dependence of the excess rate?, 応用統計学会 2015 年度年会、2015 年 03 月 14 日、京都大学医学部創立百周年記念施設芝蘭会館(京都府京都市)。
Izumi S., Ohtaki M, Aihara K, Innovative mathematical modelling for the effects of chronic exposure to radiation on cancer risk, 15th International Congress of Radiation Research (ICRR2015) (国際学会)、2015 年 05 月 25 日 ~ 2015 年 05 月 29 日、Kyoto International Conference Center. (Kyoto, Japan) .

Izumi S., Tonda T, Satoh K, Construct a simultaneous confidence interval for linear time varying coefficients in Cox proportional hazard model, Joint Statistical Meetings (JSM) 2015 (国際学会)、2015 年 08 月 08 日 ~ 2015 年 08 月 13 日、Washington State Convention Center. (Seattle, U.S.A.).

Izumi S., Cologne J., How does the statistical interaction between a single point exposure and attained age imply the shape of age-at-onset distribution?, East Asia Regional Biometric Conference (EAR-BC) 2015 (国際学会)、2015 年 12 月 20 日 ~ 2015 年 12 月 22 日、Kyushu University (Fukuoka, Japan) .

Izumi S., Sato T, Ito YM., Estimating the effects of exposure in a case-cohort design of Hokkaido Cohort study when some binary outcomes of interests are known to be missing, XXVIIIth International Biometric Conference (IBC2016) (国際学会)(国際学会)、2016 年 07 月 10 日 ~ 2016 年 07 月 15 日、Victoria Convention Centre. (Victoria, Canada) .

佐藤健一、富田哲治、和泉志津恵、生存時間データにおけるロジスティック回帰モデルを用いたオッズ比の推測、第 27 回日本疫学会学術総会、2017 年 01 月 25 日 ~ 2017 年 01 月 27 日、ベルクラシック甲府 (Yamanashi, Japan) .

- 21 Izumi S., Big Data and Health - From Data Science Viewpoint -, International Workshop on Bioethics Governance 2017 (基調講演) (国際学会)、2017 年 3 月 11 日、滋賀大学大津サテライトキャンパス (Shiga, Japan) .

〔その他〕

ホームページ等

<https://www.ds.shiga-u.ac.jp/faculty/>

6 . 研究組織

(1)研究代表者

和泉志津恵 (IZUMI, SHIZUE)

滋賀大学・データサイエンス教育研究センター・教授

研究者番号 : 70344413

(2)研究分担者

玉腰暁子 (TAMAKOSHI, AKIKO)

北海道大学・大学院医学研究科・教授

研究者番号 : 90236737

伊藤陽一 (ITO, YOICHI)

北海道大学・大学院医学研究科・准教授

研究者番号 : 10334236

野間久史 (NOMA, HISASHI)

統計数理研究所・データ科学研究系・助教

研究者番号 : 70633486

(平成 26 年度より外れる)

(3)連携研究者

竹内文乃 (TAKEUCHI, AYANO)

慶應義塾大学・医学部・専任講師

研究者番号 : 80511196

Cologne John (COLOGNE, JOHN)

放射線影響研究所・統計部・研究員

研究者番号 : 50344411

(4) 研究協力者

末永聡史 (SUENAGA, SATOSHI)

永田大貴 (NAGATA, DAIKI)