様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２０　日現在

機関番号：　３２６１２
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１５９５０

研究課題名（和文）複雑な生命事象データにおける特異な部分集合の探索的同定に関する研究

研究課題名（英文）A study on exploratory identification of responsive subgroups in complex survival event data

研究代表者

林　賢一（Hayashi, Kenichi）

慶應義塾大学・理工学部（矢上）・講師

研究者番号：７０６１７２７４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：複雑な生命事象と関連する因子（バイオマーカーなど）の関係を正しく評価する指標と，得意な反応を示す部分集団を特定するための方法論を研究した．前者については，区間打ち切りデータとよばれる生存時間データについて，C-indexとよばれる指標の推定方法を開発した．また，IDI（integrated discriminant improvement）とよばれる指標のもつ問題点を解消するために，その修正も行った．後者については，一般化平均を用いたクラスタワイズ回帰モデルの拡張を行った．このモデルは，過去に提案されたクラスタワイズ回帰モデルを特殊な場合として含む．また，数値実験によりその有用性を示した．

研究成果の概要（英文）：This project aims to develop measures of statistical model for survival outcomes and related variables such as biomarkers, and methods that detects responsive subgroups in data. As for the first issue, an estimation method of C-index for interval censored data is proposed. Further, a modification of the IDI (integrated discriminant improvement) is developed to overcome critical concerns in the original IDI. As for the second issue, a cluster-wise linear model is extended based on the generalized mean, which is called Kolmogorov-Nagumo average. The extended model involves several models proposed in the previous study as special caes. Through numerical experiments, the superiority of the proposed model is implied.

研究分野：　統計科学

キーワード：　医学統計学　区間打ち切りデータ　C-index　IDI　NRI　ROC解析　power-IDI　クラスタワイズ回帰モデル

１．研究開始当初の背景

科学的根拠に基づいた医療（EBM）の確立は，臨床医学研究における喫緊の課題である．統計解析はこの課題に資する有用な方法論であり，創薬，治療ガイドラインの策定，公衆衛生政策の決定などに役立てられている．特に，ある介入（治療など）の効果が他と比べて著しい部分集団を規定する因子の探索や，疾病・死亡リスクの際立って高い部分集団の同定は，EBM の創出に向けての重要な問題である．

本研究では，特異な部分集団を同定するための統計学的方法論の構築に取り組み，これを通じて様々な疾患領域における EBM の確立に貢献することを目指す．本研究では，大きく分けて 2 つの問題に取り組む．それは，(1) 複雑な生命事象に対する正確なバイオマーカー評価法の確立と，(2) 特異な部分集団を高精度で同定する方法論の構築である．

(1) については，着目する生命事象（死亡や疾病の発症など）が複雑な形式で表現され，バイオマーカーなどの因子との関係を統一的な観点から評価する方法が存在しない点が問題である．(2) については，事象との関連が強い因子や，介入の効果が特に高い患者背景（バイオマーカーや遺伝的要素を含む）の探索的同定が課題である．部分集団を特定するための研究は，臨床試験の文脈で近年精力的に研究されているが，その大部分は関連する因子を「事前に」かつ「少数」特定できているという前提がある．しかしこの前提は，多くの因子を候補とする場合には一般にみたされない．さらに，よく利用される樹木構造モデルは，部分集団の同定精度の確保と第 1 種の過誤確率の制御が両立しないという問題もある．以上が研究開始当初の背景である．

２．研究の目的

本研究の目的は，複雑な生命事象とそれに関連する因子の関係を正しく評価し特異な部分集団を精確に同定する方法論を構築することである．具体的に取り組む事項は，(1) 事象に関連する因子（バイオマーカーなど）の正確な評価法の開発，(2) 特異な部分集団を同定する方法論の構築，(3) 臨床医学データへの適用の三点である．これらは，互いに欠くべからざる重要な問題である．

(1) では，諸種の複雑な生存時間データに対して，望ましい性質をもつ C-index の推定量の構成が目的である．C-index（concordance index）は，因子を B,事象が発生するまでの時間を T とするとき，Pr[B1>B2 | T1<T2]で表される量である．ここで(B1,T1)と(B2,T2)は，同じ分布に従う独立な確率変数の組である．近年，より精度が高い指標として提案された Net Relassification Index など（Pencina et al., 2008）が誤った結論を下しうることが明らかにされた（Hilde, Gerds, 2013）．ゆえに，解釈の容易さやロバストネスの観点から，C-index のより多い利用が期待される．推定量の構成は，「usable pair」とよばれる症例の組のみを用いるというアイデアに基づく．

(2) では，特異な部分集団を同定するために機械学習理論を応用し，同定精度の向上とその限界についての理解が目的である．既存の手法は，部分集団の検出能力が弱い，部分集団の解釈が難しいなどの短所がある．申請者は，これらの問題を同時に解決するため，機械学習において盛んに研究されているスパース推定の応用を考える．

(3) では，本研究で構築する方法を臨床医学データに適用し，その有用性を実証することが目的である．統計解析責任者として参画している疫学研究や臨床研究は，類例の少ない貴重なものであり，EBM の確立や意思決定において意義深い結果が期待される．

３．研究の方法

(1)　C-index の推定量を，区間打ち切りデータに対して適用できる形式で構成する．まず，区間打ち切りデータに対する usable pair を定義する．このとき，先行研究とは異なり，T1 と T2 の存在範囲がオーバーラップする場合が考えられる．この問題を，usable pair である確率を推定し，その重み付き U-統計量として推定量の構成を行うことにより解決する．さらに，既存の推定量との関連性や，一致性・漸近正規性などの統計学的性質を検討し，より望ましく実用的な推定量への洗練を行う．理論的解析が困難になる場合は，ブートストラップ法などの計算機統計学的方法を援用し，信頼区間の構成やバイアス補正を行うことでの対処を考える．

(2) 特異な部分集団を同定するための方法に関する先行研究を整理し，それらを包含するような機械学習的方法の枠組みを構築する．具体的には，高次元回帰モデルを基に，回帰モデルの応答変数が本研究で想定する複雑な生命事象データに適用できるよう一般化する．また，因子（バイオマーカーや遺伝子情報など）B を Bc=I{B<c}などと変換することにより，樹木構造モデルの推定も可能であることを示す．ここで，I{A}は，事象 A が真であるときに 1，偽であるときに 0 となる指示関数である．このとき，スパース学習における罰則または制約を課す．

(3) (1)と(2)で構築した方法論を臨床医学データに適用し，その有用性を実証する．

４．研究成果

(1) C-index の推定量を，区間打ち切りデータ対して適用できる形式で構成した．区間打ち切りデータは，事象の発生時間がある区間内に存在するという形式で表現される．その

ため，2 つの区間が重なる確率という形式で usable pair を一般化し，確率を重みとする U 統計量として推定量を構成した．また，IDI の問題点を数理的に吟味し，その問題を修正する power-IDI を提案した．これは，二項回帰モデルに対する評価指標である．具体的には，power-IDI は二値応答の予測の応力を改善する候補となる因子を含めたモデルの，それらを含まないモデルからの予測脳の改善の程度を定量化したものである．power-IDI は，Bayes リスク一致性だけでなく Fisher 一致性も有する．IDI は Fisher 一致性をもたないため，提案指標は改善の誤検出を防ぐ点において優れている．このことは数値実験によっても明確に示された．

(2) 特異な部分集団を精確に同定する方法論について，当初の予定とは異なる，クラスタワイズ回帰モデルの擬似線形関数によるアプローチを試みた．クラスタワイズ回帰モデルは，観測個体の背後に複数の潜在クラスタを想定し，それぞれのクラスタに異なる回帰関数を当てはめる方法である．擬似線形関数は，Kolmogorov-Nagumo 平均ともよばれる，狭義単調増加関数とその逆関数を用いて定義される関数である．提案したモデルでは，個体のクラスタ所属確率が Rose(1990)と同様の形で与えられ，パラメータを調整することによりハードクラスタリングモデルも含まれる．その結果，多くの既存モデルが包摂される．本モデルに対し，各クラスタにおける回帰関数のパラメータとクラスタ中心を逐次推定する方法と同時推定する方法を提案・比較した．また数値実験の結果，提案モデルが混合エキスパートモデルなど既存の線形モデルよりも高い予測力を与えることを示した．

(3) 大阪急性冠症候群研究会（OACIS）データにおける，循環器疾患に対するバイオマーカーの予測・診断性能を評価した．OACIS データは，阪神地区における心筋梗塞とその関連する疾患についての二次予防データである．

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　10 件）

(1) Hayashi, K., Shimizu, Y. Estimation of a concordance probability for doubly censored time-to-event data, Statistics in Biosciences, 2018, accepted. 査読有.

(2) Hayashi, K. Asymptotic comparison of semi-supervised and supervised linear discriminant functions for heteroscedastic normal populations, Advances in Data Analysis and Classification, 2017, in press. 査読有.

(3) Hayashi, K., Takai, K. Finite-sample analysis of impacts of unlabelled data and their labelling mechanisms in linear discriminant analysis, Communications in Statistics - Simulation and Computation, 2017, 46, pp.184-203. 査読有.

(4) Hara, M., Hayashi, K., Kitamura, T. Outcomes differ by first documented rhythm after witnessed out-of-hospital cardiac arrest in children; an observational study with prospective nationwide population-based cohort database in Japan, European Heart Journal - Quality of Care and Clinical Outcomes, 2017, 3, pp.83-92. 査読有.

(5) Hara, M., Hayashi, K., Kitamura, T. Outcomes differ by first documented rhythm after witnessed out-of-hospital cardiac arrest in children; an observational study with prospective nationwide population-based cohort database in Japan, European Heart Journal - Quality of Care and Clinical Outcomes, 2017, 3, pp.83-92. 査読有.

(6) Shirane, R., Tang, H., Hayashi, K., Okuno, Y., Iso, H., Asada, H., Yamanishi, K., Mori, Y. Relationship between cell-mediated immunity to varicella-zoster virus and aging in subjects from the community-based Shozu Herpes Zoster study, Journal of Medical Virology, 2017, 89, pp.313-317. 査読有.

(7) Yamada, K., Adachi, T., Mibu A., Nishigami, T., Motoyama, Y., Uematsu, H., Matsuda, Y., Sato, H., Hayashi, K., Cui, R., Takao, Y., Shibata, M., Iso, H. Injustice experience questionnaire, Japanese version: cross cultural factor-structure comparison and demographics associated with perceived injustice, PLoS One, 2016, 11(8): e0160567. doi:10.1371/journal.pone.0160567 査読有.

(8) Taniguchi, T., Ohtani, T., Nakatani, S., Hayashi, K., Yamaguchi, O., Komuro, I., Sakata, Y. Impact of body size on inferior vena cava parameters for estimating right atrial pressure: a need for standardization?, Jounral of the

American Society of Echocardiography, 2015, 28, pp.1420-1427. **査読有**.

(9) Hara, M., Hayashi, K., Hikoso, S., Sakata, Y., Kitamura, T. Different impacts of time from collapse to first cardiopulmonary resuscitation on outcomes after witnessed out-of-hospital cardiac arrest in adults, Circulation: Cardiovascular Quality and Outcomes, 2015, 8, pp.277-284. **査読有**.

(10) Yamamoto, M., Hayashi, K. Clustering of multivariate binary data with dimension reduction via L1-regularized likelihood maximization, Pattern Recognition, 2015, 48, pp.2959-3968. **査読有**.

〔学会発表〕（計　9 件）

(1) **田栗正隆，林賢一．中間事象を伴う臨床試験における複合ストラテジーの検討，日本計量生物学会年会，東京，2018 年 3 月．**

(2) **田島史啓，林賢一．擬似値に基づく競合リスクイベントデータ解析における変数選択基準, 日本計量生物学会年会, 東京, 2018 年 3 月．**

(3) Hayashi, K. Semi-supervised learning for normal populations: a perspective from statistical missing data analysis, The 2017 Conference of International Federation of Classification, Tokyo, Japan, 2017.

(4) Hayashi, K. Cluster-wise regression models via a quasi-linear function, The International Association of Statistical Computing (Asian Regional Section) Joint Conference, Austin, New Zealand, 2017.

(5) Hayashi, K., Eguchi, S. Modification of integrated discrimination improvement by beta divergence, The 28th International Biometric Conference, Victoria, Canada, 2016.

(6) **林賢一, 高井啓二．MAR データにおける変数の部分集合に対する情報量規準，統計関連学会連合大会, 金沢, 2016 年 9 月．**

(7) **林賢一，江口真透．** Integrated discrimination improvement **の問題点とその修正，統計関連学会連合大会，金沢，2016 年 9 月．**

(8) Hayashi, K. Computing C-statistic with interval censored data, The 8th International Conference of the ECRIM WG on Computational and Methodological Statistics. London, UK, 2015.

(9) Takai, K., Hayashi, K. An information criterion for a subset of MAR data, The 8th International Conference of the ECRIM WG on Computational and Methodological Statistics. London, UK, 2015.

〔図書〕（計　1 件）

(1) **ピーター フラッハ（著），竹村 彰通（監修），田中 研太郎（翻訳），小林 景（翻訳），兵頭 昌（翻訳），片山 翔太（翻訳），山本 倫生（翻訳），吉田 拓真（翻訳），林 賢一（翻訳），松井 秀俊（翻訳），小泉 和之（翻訳），永井 勇（翻訳），朝倉書店, 機械学習 データを読み解くアルゴリズムの技法，**2017，pp.192-227.

〔産業財産権〕

○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
林 賢一（Hayashi Kenichi）
慶應義塾大学・理工学部・講師
研究者番号：70617274

研究者番号：

(2)研究分担者

(　　)

研究者番号：

(3)連携研究者

(　　)

研究者番号：

(4)研究協力者

(　　)