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

令和　６ 年　 ５ 月 １７ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2018～2023
課題番号：１８Ｋ１１１９７

研究課題名（和文）異質な集団を含むデータに対する統計的学習理論を用いたモデル開発と臨床医学への応用

研究課題名（英文）Development of statistical models for data containing heterogeneous subgroups using statistical learning theory and its application to clinical medicine

研究代表者
林　賢一（Hayashi, Kenichi）

慶應義塾大学・理工学部（矢上）・准教授

研究者番号：７０６１７２７４

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：異質な部分集団によって構成されるデータに対し，予測力と解釈可能性を兼ね備えた統計的方法の構築に寄与することを目指した本研究の主な成果は，(1)予測力の改善指標となるIDIの改良版であるodds-IDIの開発，(2)臨床試験における諸種の状況を考慮した統計的解析法の提案，(3)異質性を考慮した生存時間回帰モデルの提案である．(1)では既存の指標を性能面で優越する結果が得られ，提案指標の理論的側面と解釈可能性を研究した．(2)では欠測などによって異質性が特徴づけられる場合の統計的解析法を研究した．(3)は潜在的な治癒群と未治癒群の混在を想定した場合の生存時間回帰モデルについて研究した．

研究成果の学術的意義や社会的意義
本研究の成果の学術的意義は，従来の統計手法では十分に対応できなかった異質性を含むデータに対する新たな視点からの解析法を提示したことである．予測力と解釈可能性を兼ね備えた新たな統計的手法の開発は，大量・複雑になるデータの特徴を人間が理解する上で重要な意義をもつ．この課題に対し，疫学や臨床試験を想定する諸種のデータについて予測精度のより高い統計モデルを開発し，またそれらのモデルの評価指標を提案した．これらの成果は，様々な分野におけるデータ分析の精度向上と新たな知見の創出に貢献し，社会全体の利益に資する可能性をもつと考えられる．

研究成果の概要（英文）：Our research aims to contribute to the development of statistical methods with both predictive power and interpretability for data consisting of heterogeneous subpopulations. The main results are (1) the development of an improved version of the IDI, the odds-IDI, (2) the proposal of statistical analysis methods that take into account various situations in clinical trials, and (3) the proposal of a survival time regression model that takes heterogeneity into account. In (1), the proposed index outperforms existing indices in terms of performance, and the theoretical aspects and interpretability of the proposed index were studied. In (2), statistical analysis methods were studied for the cases where heterogeneity is characterized by missing data and so forth. In (3), a regression model for survival outcomes in the case of a mixture of potentially cured and uncured groups was studied.

研究分野：統計科学

キーワード：医学統計学　IDI　ROC　欠測値　因果推論　二値回帰モデル　生存時間解析

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

大量のデータが横溢する昨今，これらを活用した技術・産業が隆盛を誇っている．この背景には機械学習や人工知能など，活発な研究分野の発展がある．これらは，極めて予測精度の高い方法を提供する．その一方で，医学などでは，データ解析から疾病など現象の機序を知ることは依然として重要視される．そのため，予測や説明がブラックボックスの方法論には適用の限界がある．

２．研究の目的

本研究の目的は，異質な部分集団が混在するデータに対し，高い予測力（または説明力）と解釈可能性が両立する統計モデルの構築と応用である．具体的には，分割規則で表現できる領域をもつクラスターを構成し（解釈可能性の向上），クラスターごとの要約や（主に線形）回帰関数の当てはめを考える（予測力の改善）．さらに，増加したパラメータを統計的学習理論に基づき推定する．さらに，提案手法を臨床医学の実データに適用し，適切な知識獲得を通じて当該分野と社会への貢献を目指す．

３．研究の方法

本研究は，主に 3 つの観点から統計的方法論の確立を行った．これらは，臨床医学を想定しており，それぞれ(1) 臨床試験における諸種の状況を考慮した統計的解析法の提案，(2) 主に疫学を想定した，異質性を考慮するための生存時間回帰モデルの提案，(3) 回帰モデルに対する予測力の改善指標である IDI の改善，と要約することができる．(1)では，臨床試験における治療効果の推定について，諸種の状況に応じてより精度の高い方法を模索した．具体的には，データ取得のデザインによってアウトカムの欠測が構造的に生じる場合，無視のできない欠測が生じる場合，過去の対照群データ(ヒストリカルコントロールデータ)が利用可能な場合などである．これらは，欠測の有無に依存して異質性が生じたり，試験データと過去のデータに質的な差が生じたりすることなどが想定される．これらの問題について，反教師あり学習や非線形な回帰モデル（教師あり学習）などの機械学習に端を発する方法を援用し，欠測の影響を緩和し推定のバイアスと精度の改善を試みた．(2)は，疫学研究における生存時間データに，異質なデータの存在を想定した方法論の開発である．具体的には，すでに治癒している（と想定される）集団とそうでない集団の混在を想定した．ここで前者を治癒群，後者を未治癒群とよぶ．さらに典型的な Cox 比例ハザードモデルは制約が強いモデル（比例ハザード性の仮定）のため，一般化ガンマ分布を用いた加速時間モデルを基礎においた混合治癒モデルによるモデリングを検討した．また，生存時間データとともに経時測定データが存在する場合に用いられる同時モデリング（joint modeling）について，異質な集団が存在している場合に単一のモデルを当てはめるのは不適切である．そこで，決定木に基づくモデルによる部分集団への再起分割法を考え，部分集団ごとにモデルのあてはめを行った．(3)は，過去に提案した power-IDI の問題を解決することが主題である．Power-IDI (Hayashi and Eguchi, 2019)は，IDI が Fisher 一致性という重要な性質をもたないことによる欠陥を修正した，モデルの予測改善を測るための指標である．Power-IDI は Fisher 一致性をもつ一方で，ハイパーパラメータの値を決定する必要があり，それを決める方法に明確な基準がないという問題が存在する．さらに指標を構成するためにモデルを二度変換する必要がある．これらが実用と解釈の困難さを残している．そこで，オッズ（確率を p としたとき，p/(1-p)がオッズである）を指標とした指標である odds-IDI を提案した．オッズがすでにモデルをある意味で変換していると理解することができるが，オッズは臨床研究で頻繁に用いられる指標であり，解釈の余地があると考えられることが重要である．

４．研究成果

「３．研究の方法」で述べた(1)-(3)の順に，成果を述べる．　(1)では，すべての研究において既存の指標を性能面で優越する結果が数値実験を通じて得られた．とくに，ヒストリカルコントロールデータが存在する場合における治療効果の推定については，ヒストリカルコントロールデータにより構成した予後スコアとよばれる量によって共変量調整を実施し，かつ治療との交互作用を含めた場合に推定量の分散が最も小さくなるという理論的結果を得た．(2)では，未治癒群の中にも異質性を想定するためにフレイルティモデルを用いた．これによって個人によりベースラインハザードが異なる場合を表現できる．しかし，通常用いられるガンマフレイルティモデルは治癒確率を過大評価する方向に働き，当該研究で用いた混合治癒モデルに適用することは不適切であることが判明した．そこで，ガンマ分布を正方向に移動させたシフトガンマ分

布を用いてフレイルティモデルを一般化することにより，モデルのあてはまりが改善することが示唆された．さらに実際のデータ（健診データ）を用いて，従来のモデルよりも提案モデルがよくあてはまることが実証できた．生存－経時同時モデルにおける再起分割法では，分割の規則を二つの部分集団間におけるパラメータの差によって定める．これによって，異質な部分集団ごとのモデリングが可能であることを数値的に示した．(3)ではオッズから定式化される IDI のような指標が Fisher 一致性をもつための条件を導き，odds-IDI を構成した．この量は単に Fisher 一致性をもち，既存の指標より精度の高い判断を与えるだけでなく，機械学習における指数ロス関数や多変量解析論における一般化分散との関連を示した．これは解釈可能性に寄与する重要な知見である．

その他にも，インデックスモデルにおけるロバストな方法の模索や異常検知のための研究など，異質なデータが存在することに由来する問題にも取り組んだ．

５．主な発表論文等

〔雑誌論文〕　計12件（うち査読付論文　11件／うち国際共著　0件／うちオープンアクセス　4件）

| | |
|---|---|
| １．著者名<br>Taguri Masataka、Takahashi Kunihiko、Komukai Sho、Ito Yuri、Hattori Satoshi、Funatogawa Ikuko、Shinozaki Tomohiro、Yamamoto Michio、Hayashi Kenichi | ４．巻<br>44 |
| ２．論文標題<br>Advancements of Biometrics in the Field of Epidemiology | ５．発行年<br>2024年 |
| ３．雑誌名<br>Japanese Journal of Biometrics | ６．最初と最後の頁<br>129～200 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.5691/jjb.44.129 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>志村 重輔、林 沙織、岡安 悟志、板垣 昌幸、林 賢一 | ４．巻<br>12 |
| ２．論文標題<br>k 近傍法を用いたリチウムイオン電池の微小内部短絡検出 | ５．発行年<br>2023年 |
| ３．雑誌名<br>データ分析の理論と応用 | ６．最初と最後の頁<br>1～15 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.32146/bdajcs.12.1 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Takai Keiji、Hayashi Kenichi | ４．巻<br>6 |
| ２．論文標題<br>Model Selection with Missing Data Embedded in Missing-at-Random Data | ５．発行年<br>2023年 |
| ３．雑誌名<br>Stats | ６．最初と最後の頁<br>495～505 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.3390/stats6020031 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| １．著者名<br>Aida Haro、Hayashi Kenichi、Takeuchi Ayano、Sugiyama Daisuke、Okamura Tomonori | ４．巻<br>10 |
| ２．論文標題<br>An Accelerated Failure Time Cure Model with Shifted Gamma Frailty and Its Application to Epidemiological Research | ５．発行年<br>2022年 |
| ３．雑誌名<br>Healthcare | ６．最初と最後の頁<br>1383～1383 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.3390/healthcare10081383 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| 1．著者名<br>Hara, M., Hayashi,K., Kitamura, T., Honda, M., Tamaki, M. | 4．巻<br>84 |
| 2．論文標題<br>A nationwide randomised, double-blind, placebo-controlled physicians’ trial of loxoprofen for the treatment of fatigue, headache, and nausea after hangovers | 5．発行年<br>2020年 |
| 3．雑誌名<br>Alcohol | 6．最初と最後の頁<br>21-25 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.alcohol.2019.10.006 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

| | |
|---|---|
| 1．著者名<br>Inoue, K., Hikoso,S., Masuda, M., Furukawa, Y., Hirata, A., Egami, Y., Watanabe, T., Minamiguchi, H., Miyoshi, M., Tanaka, N., Oka, T., Okada, M., Kanda, T., Matsuda, Y., Kawasaki, M., Hayashi, K., Kitamura, T., Dohi, T., Sunaga, A., Mizuno, H., Nakatani, D., Sakata, Y. | 4．巻<br>23 |
| 2．論文標題<br>Pulmonary vein isolation alone vs. more extensive ablation with defragmentation and linear ablation of persistent atrial fibrillation: the EARNEST-PVI trial | 5．発行年<br>2020年 |
| 3．雑誌名<br>EP Europace | 6．最初と最後の頁<br>565-574 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1093/europace/euaa293 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| 1．著者名<br>Hayashi, K., Eguchi, S. | 4．巻<br>38 |
| 2．論文標題<br>The power‐integrated discriminant improvement: An accurate measure of the incremental predictive value of additional biomarkers | 5．発行年<br>2019年 |
| 3．雑誌名<br>Statistics in Medicine | 6．最初と最後の頁<br>2589-2604 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1002/sim.8135 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| | |
|---|---|
| 1．著者名<br>Dohi, T., Nakatani, D., Inoue, K., Hikoso, S. Oka, T. Hayashi, K., Masuda, M., Furukawa, Y., Kawasaki, M., Egami,Y., Kashiwase, K., Hirata, A., Watanabe, T., … Sakata,Y. | 4．巻<br>74 |
| 2．論文標題<br>Effect of extensive ablation on recurrence in patients with persistent ptrial fibrillation treated with pulmonary cein isolation (EARNEST-PVI) trial: design and rationale | 5．発行年<br>2019年 |
| 3．雑誌名<br>Journal of Cardiology | 6．最初と最後の頁<br>64-68 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1016/j.jjcc.2019.01.010 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| 1．著者名<br>Hayashi, K. | 4．巻<br>12 |
|---|---|
| 2．論文標題<br>Asymptotic comparison of semi-supervised and supervised linear discriminant functions for heteroscedastic normal populations | 5．発行年<br>2018年 |
| 3．雑誌名<br>Advances in Data Analysis and Classification | 6．最初と最後の頁<br>315-339 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1007/s11634-016-0266-6 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| 1．著者名<br>林賢一 | 4．巻<br>61 |
|---|---|
| 2．論文標題<br>統計学は錬金術ではない | 5．発行年<br>2018年 |
| 3．雑誌名<br>心理学評論 | 6．最初と最後の頁<br>147-155 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>無 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| 1．著者名<br>Hayashi, K., Shimizu | 4．巻<br>10 |
|---|---|
| 2．論文標題<br>Estimation of a Concordance Probability for Doubly Censored Time-to-Event Data | 5．発行年<br>2018年 |
| 3．雑誌名<br>Statistics in Biosciences | 6．最初と最後の頁<br>546-567 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1007/s12561-018-9216-5 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

| 1．著者名<br>Hayashi, K., Eguchi, S. | 4．巻<br>印刷中 |
|---|---|
| 2．論文標題<br>The power‐integrated discriminant improvement: An accurate measure of the incremental predictive value of additional biomarkers | 5．発行年<br>2019年 |
| 3．雑誌名<br>Statistics in Medicine | 6．最初と最後の頁<br>－ |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1002/sim.8135 | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスではない、又はオープンアクセスが困難 | 国際共著<br>－ |

〔学会発表〕　計12件（うち招待講演　1件／うち国際学会　6件）

１．発表者名
Yuki Itaya, Jun Tamura, Kenichi Hayashi, Kouji Yamamoto

２．発表標題
Asymptotic properties of the Matthews correlation coefficient

３．学会等名
The 8th Japanese-German Symposium on Classification（国際学会）

４．発表年
2023年

１．発表者名
名取京太朗，林賢一

２．発表標題
生存－経時同時モデルに対する異質な部分集団への再帰分割法

３．学会等名
2023 年度 日本分類学会大会

４．発表年
2023年

１．発表者名
林賢一，吉牟田迪弥

２．発表標題
無作為化比較試験における半教師あり学習 を用いたリスク差の推定

３．学会等名
日本計量生物学会年会

４．発表年
2022年

１．発表者名
志村 重輔，林 沙織，岡安 悟志，板垣 昌幸，林 賢一

２．発表標題
k近傍法を用いたリチウムイオン電池の微小内部短絡検出

３．学会等名
日本分類学会

４．発表年
2022年

１．発表者名
林賢一，江口真透

２．発表標題
A new integrated discriminant improvement index via odds

３．学会等名
日本計量生物学会年会

４．発表年
2021年

１．発表者名
林賢一，江口真透

２．発表標題
二値回帰モデルに対するオッズに基づいた予測改善指標

３．学会等名
日本計算機統計学会 第35回シンポジウム

４．発表年
2021年

１．発表者名
Kenichi Hayashi, Shinto Eguchi

２．発表標題
Odds-based predictive improvement index for binary regression models

３．学会等名
The 11th Conference of the IASC-ARS The Asian Regional Section of the International Association for Statistical Computing（招待講演）（国際学会）

４．発表年
2022年

１．発表者名
会田晴郎，林賢一

２．発表標題
シフトガンマフレイルティを伴う加速時間治癒モデルとその疫学研究への応用

３．学会等名
統計関連学会連合大会

４．発表年
2020年

１．発表者名
Hayashi, K., Taguri, M.

２．発表標題
Estimation of causal effects in the presence of noncom- pliance without actual treatment information

３．学会等名
The 40th Annual Conference of the International Society for Clinical Biostatistics（国際学会）

４．発表年
2019年

１．発表者名
Hayashi, K., Takai, K.

２．発表標題
A model selection criterion for missing data with MAR mechanism but NMAR for its subsets

３．学会等名
International conference Data Science, Statistics & Visualisation (A Satellite Conference of the 62nd World Statistics Congress promoted by IASC)（国際学会）

４．発表年
2019年

１．発表者名
Tajima, F., Hayashi, K.

２．発表標題
A variable selection criterion for competing risk data with pseudo-observations

３．学会等名
The 6th International Society for Biopharmaceutical Statistics（国際学会）

４．発表年
2019年

１．発表者名
Taguri, M., Hayashi, K.

２．発表標題
A new composite estimand for regulatory clinical trials with dropouts

３．学会等名
The 29th International Biometric Conference（国際学会）

４．発表年
2018年

〔図書〕　計1件

| | |
|---|---|
| 1．著者名<br>林賢一 | 4．発行年<br>2020年 |
| 2．出版社<br>講談社サイエンティフィク | 5．総ページ数<br>352 |
| 3．書名<br>Rで学ぶ統計的データ解析 | |

〔産業財産権〕

〔その他〕

―

6．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|