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　 ６ 月 ２１ 日現在

機関番号：１２６１１
研究種目：若手研究
研究期間：2021～2023
課題番号：２１Ｋ１７７４６

研究課題名（和文）Secure, Precise and Fast Sequential Pattern Mining with Learning Data Distribution

研究課題名（英文）Secure, Precise and Fast Sequential Pattern Mining with Learning Data Distribution

研究代表者
Ｌｅ　Ｈｉｅｕ・Ｈａｎｈ（Le, Hieu Hanh）

お茶の水女子大学・文理融合 AI・データサイエンスセンター・准教授

研究者番号：６０８１３９９６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：シーケンシャルな医療データのセキュアな解析が研究されている。具体的に複数の病院から取得したシーケンスバリアント（SV）を解析するための高速な方法が検討されており、SVにおける分岐の要因の推定などが含まれる。本研究で提案された方法は複数の疾患に関する実際の病院の医療データを用いて評価された。
シーケンスの頻度を推定する際にプライバシーを確保するために、元の頻度に適切な量のノイズを加える方法を検討した。データの分布と医学的な意義に基づいて、関連するデータのみを解析対象とした。最後に、データアクセス制御が充分に管理されたクラウドを使用したセキュアな実験環境が提案され、セキュアなデータ解析を可能とした。

研究成果の学術的意義や社会的意義
この研究は、セキュアなシーケンシャルデータ解析に大きな波及効果をもたらす。これにより、医療や小売業など多くのビジネスにおいて、安全にカスタマイズ可能なツールやサービスを提供するアプリケーションの範囲が拡大できる。顧客向けのサービスや製品を安全にカスタマイズするだけでなく、産業企業内のサプライチェーン管理を効率的に最適化する可能性を見せることができる。

研究成果の概要（英文）：Secure sequential data analysis of sequential medical data has been extensively studied. In detail, several methods to analyze the sequence variants (SV) from more than one hospital have been designed, such as comparing SVs, identifying factors that led to branches in SVs, etc. The proposed methods were rigorously evaluated using real hospital medical data on multiple diseases.
Then, an appropriate amount of noise is added to the original frequency to ensure privacy when estimating the frequency of the sequences. Only related data is added to the analysis based on data distribution and medical meaningfulness. Finally, a secure experimental environment using the cloud in which the data access control is carefully managed has been suggested for a secure data analysis.

研究分野：データ工学

キーワード：シーケンス解析　電子カルテ　データ保護

## １．研究開始当初の背景

シーケンシャルパターンマイニング（SPM）は、現在データマイニングおよび知識発見において広く使用されている。SPM は医学、電子商取引、ワールドワイドウェブなどの様々なアプリケーションドメインからシーケンスデータベース内の頻出パターンを発見する。同時に、データをマイニングする際に個人のプライバシーを保証する必要性が、学術界と社会の両方から多くの注目を集めている。差分プライバシー（DP）は、データプライバシーの事実上の標準とみなされている。DP では、データ分析者が各ユーザーから個人データを収集し、プライバシーを強化するためにノイズを加えた出力を生成する。

SPM におけるプライバシーの保護に焦点を当てた研究は存在しているが、アイテムの順序を考慮するため、アイテム数とシーケンス数が多い場合、中間候補の数は膨大なる。これにより膨大な量のノイズが追加され、データ分析の精度が低下してしまう。

## ２．研究の目的

本研究のでは高度でセキュアな SPM 手法を提案することである。特に、シーケンシャルデータベース（SDB）として電子カルテデータを対象に、ある疾患に対する典型的なクリニカルパスでの分岐の要因推定を行い、更に分析時間を短縮するためにシーケンシャルの識別子を保持する SPM を提案する。次に、医療機関の特徴を発見するために、複数医療機関の SPV の比較方法の検討を行う。最後に、データ保護のために、セキュアな実験環境と適切なノイズを加えるメカニズムを検討する。

## ３．研究の方法

（1）クリニカルパスでの要因推定：SDB から、シーケンシャルの識別子を保持する SPM を適用しバリアントの抽出時間を短縮する。次に、抽出されたバリアントからシーケンシャルパターンバリアント（SPV）を生成し、SPV での分岐の要因を推定する。電子カルテデータに適用する際に、患者の病気履歴、年齢、性別、入院時期などの静的な情報と入院している間のバイタルサインや検体検査結果などの動的な情報を説明変数、分析している SPV を目的変数とし、有意差検定を行い、要因を推定する。

（2）セキュアなデータ解析：まず、分析対象となる医療データをインターネットから切り離されたクラウドサーバに保存する。各研究者は認証ありのリモートデスクトップにログインし、クラウドサーバにデータを取得して、クラウドサーバ析を行った。その際に、個人を特定できる情報を絶対に持ち出せないように厳密に管理された。次に、シーケンス数に対する出現頻度や医学的な観点で分析対象データを絞り、そのデータの出現頻度にノイズ生成メカニズムの検討を行った。

## ４．研究成果

（1）クリニカルパスでの要因推定

本研究では、ある病院の電子カルテシステムに記録されている、2015 年 1 月 1 日から 2018 年 4 月 20 日の範囲の実際に使用された医療行為データと検体検査の履歴データを対象とした。このデータ個人情報保護の鑑定より患者を一意に特定でいるような情報を含まない。対象データセットのシーケンス数、シーケンス内の最大・平均医療行為数と最大・平均入院日数の情報を表１に示す。

表１：対象データセット

| | 経尿道的膀胱腫瘍切除（Tur-Bt） | 非ホジキンリンパ（NHL） |
|---|---|---|
| シーケンス数 | 394 | 99 |
| 最大医療行為数 | 179 | 1,491 |
| 平均医療行為数 | 49.79 | 121.22 |
| 最大入院日数 | 20 | 185 |
| 平均入院日数 | 7.40 | 24.81 |

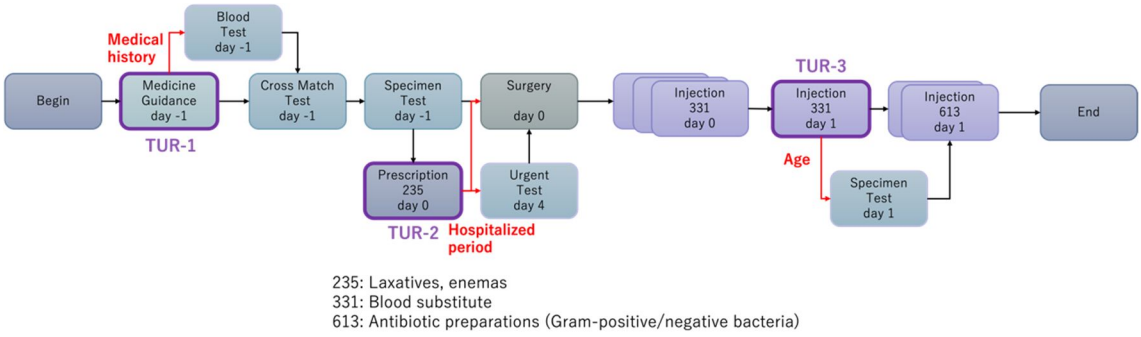


図１：TUR-Bt の要因推定結果

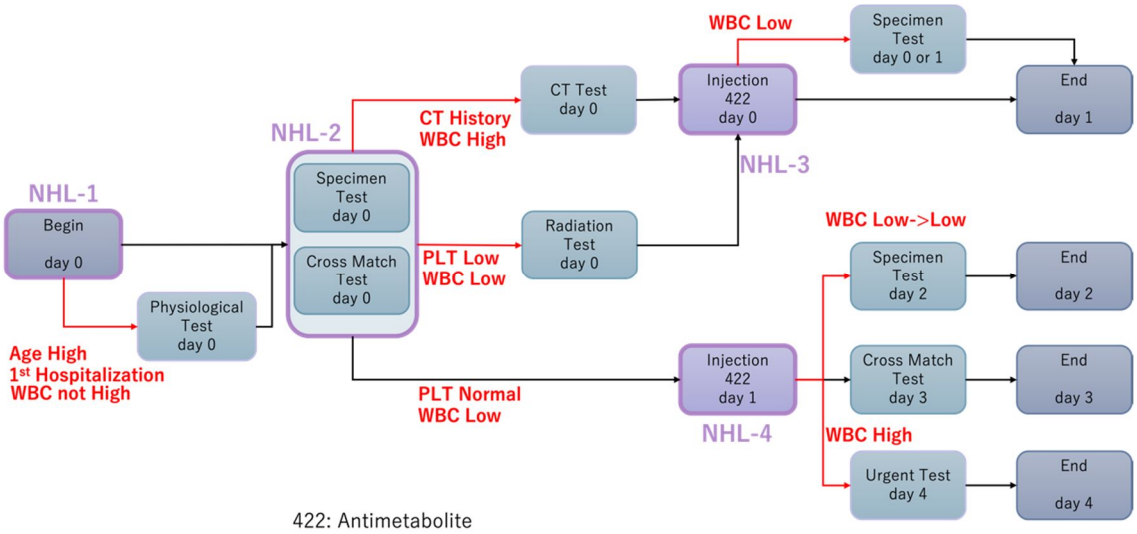


図２：NHL の要因推定結果

図１は TUR-Bt に対するバリアントにおいての要因推定の結果を示す。各ノードは医療行為と医療の実施相対日を表し、注射（Injection）には、215，331，613 などの薬効コードが付与されている。ここでは、４つの分岐の内３つの分岐に対して要因推定ができた。TUR-1 は血液検査を受けた患者と受けていない患者を区別する要因として、今回の入院が初回かどうかを示す入院履歴である。血液検査結果がない初回入院の患者は、過去に TUR-Bt で入院したことがある患者よりも、血液検査を受ける可能性が高いことを示す。この結果は、医療従事者が以前の入院時に得た検査結果を再利用する可能性が高いという事実とよく一致している。TUR-2 と TUR-3 では、それぞれ入院時期と年齢が、緊急検査（Urgent Test）と検体検査（Specimen Test）を受けるか受けないかの要因と推定できた。2013 年 7 月ごろからよく使われるようになった緊急検査は 2013 年 1 月以前あまり実施されなかった。そして、年寄り上の患者が若い患者よりも注射後の検体検査を受ける可能性が高いとわかった。

図２は NHL に対するバリアントにおいての要因推定の結果を表示する。このバリアントは年齢・入院履歴と患者の年齢以外にも、検体検査の結果とその結果の推移を分岐の要因として推測できた。NHL-2 では、ここまで CT テストを受けたことがある患者の方がさらに CT テストを受ける傾向があった興味深い結果を得た。放射線検査（Radiation Test）より詳細の症状を観察できる CT テストを受けたことがある患者は一般的に健康的に問題がある可能性が高いため、再検査が必要である。その他、白血球数（WBC）の高い患者もさらに CT テストを受ける傾向があった。逆に、白血球数が低いまたは血小板数（PLT）が低かったら、CT テストの代わりに放射線検査のみで良いと判断する。最後に、血小板数が正常のであれば、CT テストと放射線検査が不要で、直接に注射（NHL-4）に進むことができる結果を得た。しかし、ここで、直近の 2 回の白血球数が低いのままであれば、患者の状況を詳しく見るために検体検査が必要となった。

（2）セキュアなデータ解析：

本研究は最長で 1,000 を超えた電子カルテデータのような長さの長いシーケンスのデータにノイズを加えて分析した時に、出現頻度・出現分布と医学的な有用性を考慮したが大部分の場合は分析が完了せずに、実行時間が非常にかかるとわかった。

一方、本研究は医療データを用いて分析を行うため、セキュアな実験環境の検討結果として、分析対象となる医療データをインターネットから切り離されたクラウドサーバに保存し、各研

**究者は認証ありのリモートデスクトップにログインし、クラウドサーバにデータを取得して、クラウドサーバ析を行った。**

**以上、本研究において医療支援を行うために、医療情報のシーケンスを解析するセキュアで高度な手法を提案し、実際の電子カルテデータに適用することで有効性を確認することができた。この成果は、医療情報だけでなく、広くシーケンス解析に適用することができる。**

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　1件）

| | |
|---|---|
| １．著者名<br>Hieu Hanh Le, Tatsuhiro Yamada, Yuichi Honda, Takatoshi Sakamoto, Ryosuke Matsuo, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota | ４．巻<br>4, issue 1, no. 3 |
| ２．論文標題<br>Methods for Analyzing Medical-Order Sequence Variants in Sequential Pattern Mining for Electronic Medical Record Systems | ５．発行年<br>2023年 |
| ３．雑誌名<br>ACM Transactions on Computing for Healthcare | ６．最初と最後の頁<br>1～28 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1145/3561825 | 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>Yuqing Li, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫 | ４．巻<br>1, no.5 |
| ２．論文標題<br>シーケンスバリアントの比較と電子カルテの分析への応用 | ５．発行年<br>2023年 |
| ３．雑誌名<br>日本データベース学会データドリブンスタディーズ論文誌 | ６．最初と最後の頁<br>1-8 |
| 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>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>Yuqing Li, Hieu Hanh Le, Ryosuke Matsuo, Tomoyoshi Yamazak, Kenji Araki, Haruo Yokota | ４．巻<br>13427 |
| ２．論文標題<br>Comparison of Sequence Variants and the Application in Electronic Medical Records | ５．発行年<br>2022年 |
| ３．雑誌名<br>Proceeding of the 33rd International Conference on Database and Expert Systems Applications (DEXA2022), Part 2 | ６．最初と最後の頁<br>117～130 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>10.1007/978-3-031-12426-6_10 | 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15件（うち招待講演　0件／うち国際学会　4件）

| |
|---|
| １．発表者名<br>松尾亮輔、Le Hieu Hanh、山崎友義、小口正人、横田治夫 |
| ２．発表標題<br>複数流行波からなる感染症の流行波における特異的な重症化マーカーの検出方法 |
| ３．学会等名<br>第２８回医療情報学会春季学術大会 |
| ４．発表年<br>2024年 |

１．発表者名
Hieu Hanh Le, Yuki Yasumitsu, Ryosuke Matsuo, Tomoyoshi Yamazaki, Haruo Yokota

２．発表標題
A Clustering-based Sequence Variants Analysis Method for Electronic Medical Records of Multimedical Institutions

３．学会等名
The 3d International Workshop on Big Datah in Healthcare in conjunction with IEEE MIPR 2024（国際学会）

４．発表年
2024年

１．発表者名
Zitai Zhao, Yuki Yasumitsu, Hieu Hanh Le, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota

２．発表標題
Analysis of Transitions in Differences between Frequent Medical-order Sequences for COVID-19

３．学会等名
The 36th IEEE International Symposium on Computer-Based Medical Systems（国際学会）

４．発表年
2023年

１．発表者名
Le Hieu Hanh, 松尾亮輔, 山﨑 友義, 横田 治夫

２．発表標題
千年カルテの匿名加工医療情報を利用した多医療機関の電子カルテに対するシーケンス解析

３．学会等名
第43回医療情報学連合大会（第24回日本医療情報学会学術大会）

４．発表年
2023年

１．発表者名
趙 子泰, Le Hieu Hanh, 山﨑 友義, 荒木 賢二, 横田 治夫

２．発表標題
COVID-19の電子カルテ履歴からの医療指示シーケンスパターン変化時期の抽出

３．学会等名
第27回日本医療情報学会春季学術大会

４．発表年
2023年

１．発表者名
Zitai Zhao, Yuki Yasumitsu, Hieu Hanh Le, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota

２．発表標題
Analysis of Transitions in Differences between Frequent Medical-order Sequences for COVID-19

３．学会等名
The 36th IEEE International Symposium on Computer-Based Medical Systems（国際学会）

４．発表年
2023年

１．発表者名
黒川 健人, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫

２．発表標題
動的に医療指示種類を変更したシーケンス解析における特徴的な治療パターン抽出

３．学会等名
第１５回データ工学と情報マネジメントに関するフォーラム

４．発表年
2023年

１．発表者名
Zhao Zitai, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫

２．発表標題
COVID-19に関する頻出医療指示パターンの時期による差異と差異発生時期の可視化

３．学会等名
第１５回データ工学と情報マネジメントに関するフォーラム

４．発表年
2023年

１．発表者名
安光 夕輝, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫

２．発表標題
クラスタリングを用いた多病院間の頻出医療指示パターン比較

３．学会等名
第１５回データ工学と情報マネジメントに関するフォーラム

４．発表年
2023年

１．発表者名
Zhao Zitai, Le Hieu Hanh, 松尾 亮輔, 山﨑 友義, 荒木 賢二, 横田 治夫

２．発表標題
COVID-19の異なる医療機関と時期における頻出治療パターンの比較

３．学会等名
第42回医療情報学連合大会

４．発表年
2022年

１．発表者名
横田治夫, Le Hieu Hanh, Li Yuqing, 松尾亮輔, 山﨑友義, 荒木賢二

２．発表標題
数医療機関間の頻出医療指示パターン比較手法

３．学会等名
第26回日本医療情報学会春季学術大会

４．発表年
2022年

１．発表者名
Hieu Hanh Le, Yutaka Horino, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota

２．発表標題
Sequential Pattern Mining of Large Combinable Items with Values for a Set-of-items Recommendation

３．学会等名
The 34 IEEE International Symposium on Computer-based Medical Systems (CBMS 2021)（国際学会）

４．発表年
2021年

１．発表者名
An Wang, Hieu Hanh Le, Ryosuke Matsuo, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota

２．発表標題
MERJ: Medical Entity-Relation Extraction System for Japanese Clinical Texts

３．学会等名
The 14th Forum on Data Engineering and Information Management (DEIM 202)

４．発表年
2022年

| |
|---|
| １．発表者名<br>Li Yuqing，Le Hieu Hanh，松尾亮輔，山崎友義，荒木賢二，横田治夫 |
| ２．発表標題<br>シーケンシャルパターンマイニングに基づく多病院間の頻出治療パターンの比較 |
| ３．学会等名<br>第１４回データ工学と情報マネジメントに関するフォーラム予稿集 |
| ４．発表年<br>2022年 |

| |
|---|
| １．発表者名<br>横田治夫，Le Hieu Hanh，松尾亮輔，山崎友義，荒木賢二 |
| ２．発表標題<br>医療データのシーケンス解析とその課題 |
| ３．学会等名<br>第１２回日本医療情報学会「医用人工知能研究会」人工知能学会「医用人工知能研究会」合同研究会 |
| ４．発表年<br>2022年 |

〔図書〕　計0件

〔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 |
|---|---|