

## 科学研究費助成事業 研究成果報告書

令和 5 年 6 月 12 日現在

機関番号：12608

研究種目：基盤研究(B)（一般）

研究期間：2020～2022

課題番号：20H04192

研究課題名（和文）複数項目の値の変動に依存し動的に出現が変化する項目の予測の実現と評価

研究課題名（英文）Implementing and Evaluating Prediction of Dynamical Item Appearance Depending on Fluctuation of Multiple Item Values

研究代表者

横田 治夫（Yokota, Haruo）

東京工業大学・情報理工学院・教授

研究者番号：10242570

交付決定額（研究期間全体）：（直接経費） 13,400,000円

研究成果の概要（和文）：大量の履歴データのシーケンス中の複数項目の値の変動に依存して項目の出現が動的に変化する場合、次に発生する項目を根拠とともに予測することは、従来の手法では十分ではなかった。本研究では、同時に発生する複数の項目の値の変動を解析し、次に発生する項目をその根拠とともに予測する手法を提案し、実際に蓄積された電子カルテの履歴に適用し、複数の検体検査項目と、その値の変動から、次に実施する検査の項目や、医療指示の内容の予測を行う手法を実現し、評価を行った。さらに、シーケンス解析の手法を拡張し、複数の医療機関や、時期による頻出医療指示のシーケンスの違いを数値的に比較し可視化する手法に発展させ評価を行った。

研究成果の学術的意義や社会的意義

深層学習による予測や推薦の研究が盛んに行われているが、医療等の場面では予測や推薦の根拠を示すことが求められ、深層学習では根拠を示せないため説明可能深層学習の研究も行われているがまだ十分とは言えない。一方、シーケンシャルパターンマイニングでは、パターンの発生頻度を根拠とした予測や推薦が可能となるが、項目の出現がシーケンス中の多数項目の値の変動に依存する場合にはこれまで対応できていなかった。本研究では、クラスタリングと特性ベクトルを用いてその課題を解決し、実際の電子カルテデータを用いてその効果を示すとともに、複数医療機関や時期の違いについても解析する手法を提案し、その学術的・社会的意義は高い。

研究成果の概要（英文）：When the occurrence of items dynamically changes depending on the variation of values of multiple items in a sequence of large amounts of historical data, predicting the next occurring items with evidence is not sufficient with conventional methods. In this research, we propose methods to analyze the sequential patterns in the values of multiple items that occur at the same time, and to predict the items that will occur next along with their statistical basis. Based on the specimen inspections and their value fluctuations, we realized and evaluated a method for predicting the specimen inspections to be performed next and the content of medical orders. In addition, we extended the sequence analysis method to develop a method that numerically compares and visualizes the differences in the sequences of frequent medical orders between multiple medical institutions and depending on the time period.

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 項目推薦 医療検査 値予測

様式 C - 19、F - 19 - 1、Z - 19 (共通)

1. 研究開始当初の背景

ビッグデータという言葉が表すように、我々の身の回りには大量のデータが生成・蓄積され、解析の対象となり、多様な用途に供されている。中でも、様々な履歴のデータは、蓄積することで単調にデータ量が増加して大容量化するとともに、履歴中のシーケンスを解析することで傾向等を知ることが可能となり、様々な予測や推薦等の広範な需要があり、重要度が増している。

近年は、深層学習による予測や推薦の研究が盛んに行われているが、医療等の場面では予測や推薦の根拠を示すことが求められる。現状の深層学習では根拠を示せないため説明可能深層学習の研究も行われているがまだ十分とは言えない。一方、シーケンシャルパターンマイニング (SPM) を用いたアプローチでは、パターンの発生頻度を根拠とした予測や推薦が可能となるが、項目の出現がシーケンス中の多数項目の値の変動に依存する場合にはこれまで対応できていなかった。例えば、医療における実際の検体検査の項目は 1,400 を超え、それぞれの検査結果に依存して次の検査項目が決まるが、そのようなパターンの頻度を出すことはできていなかった。

2. 研究の目的

本研究では、SPM において、同時に出現する複数の項目をグループとして扱い、その要素となる各項目の値をクラスに分け、クラスの列をグループの特性ベクトルとし、グループのシーケンスへの特性ベクトルの変動による依存関係を検出することで、次のグループの変化の予測を行い、推薦につなげるアプローチの提案と、その実証とその展開を目的とする。

3. 研究の方法

提案アプローチの適用例として、実際の電子カルテデータベースに含まれる検体検査項目の履歴を対象に、複数の検査結果の値の変動による次の検査項目の予測および推薦の実験を行い、その効果の実証を試みると同時に、医療オーダーシーケンス解析にも展開する。

非常に簡略化した検体検査のシーケンスの例として、以下の図 1 に肝機能の血液検査の例を示す。検査項目の -GTP、ALP はそれぞれ基準値範囲 [9-109]、[117-350] 内で、GOT (AST)、GPT (ALT) がそれぞれの基準値範囲 [13-33]、[8-42] より高い値に変化した場合に、HBs 抗原、HC 抗体の検査を行うべきであるが、-GTP も高い場合には追加検査ではなく飲酒の指導を行うべきであるとする。その場合、追加で行われるどの検査項目、あるいは不要となる検査項目 (例では LDL) が、それ以前のどの検査項目の値に依存しているのかを抽出する必要がある。これは、GOT、GPT、HBs といったそれぞれの検査項目のシーケンスの出現頻度だけ見ても抽出できないし、それぞれの検査項目の値の変動だけ見ても抽出できない。

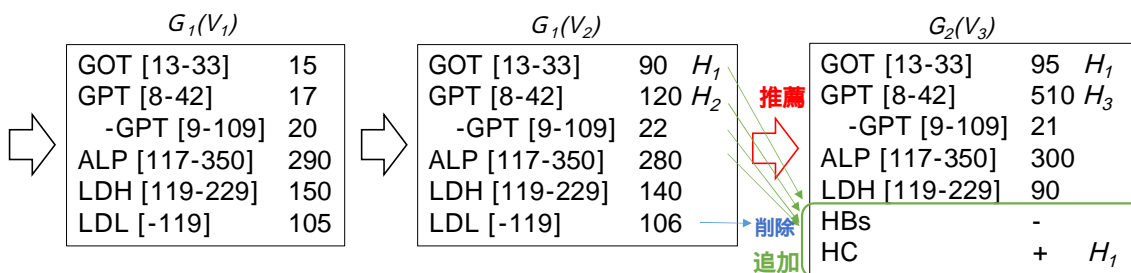


図 1 複数の値の変動により項目の出現が動的に変化する場合のパターンと推薦の例

本研究では、まず、同時に発生する複数の項目を項目のグループとして扱うと同時に、グループの要素となる各項目の値を大まかなクラスに分け、そのクラスの列をそのグループの特性ベクトルとして、グループのシーケンスと値のグループの特性ベクトルの変動の依存関係の抽出を行い、予測の評価を試みる。

具体的に、図 1 の医療検査の例を用いて本研究のアプローチを考える。同時に出現している GOT、GPT、-GTP、ALP、LDH、LDL といった血液の検査項目をグループ  $G_1$  とし、値のクラスで構成されるグループの特性ベクトルを  $V_1=(N, N, N, N, N, N)$  とする。ここで、正常値は  $N$  で、異常値は基準値からの相対的な差によって、高い場合には  $H_1$ - $H_4$ 、低い場合には  $L_1$ - $L_4$  といったクラスとして表し、特性ベクトルを構成する。図 1 に示した例では、同じ  $G_1$  のグループで、GOT が  $H_1$ 、GPT が  $H_2$ 、それ以外の項目は正常値の範囲とする特性ベクトル  $V_2=(H_1, H_2, N, N, N, N)$  となる。このグループ  $G_1$  の特性ベクトルの変化 ( $V_1 \rightarrow V_2$ ) に依存して、新たに HBs と HC の検査項目を加え、不要となる LDL の検査項目を外す場合、それに対応したグループ  $G_2$  が生成され、対応する特性ベクトルが  $V_3=(H_1, H_3, N, N, N, N, H_1)$  となる。

なお、基準値やクラス分けは、適用分野によって異なるため、適用分野の専門家の意見を取り入れて設定することとするが、一度設定した後は基本的には変更はないものとする。また、場合によってはクラス分けをせず、元の値をそのまま使った方がよい場合がある可能性もあるが、本研究を進めて行く中で特性ベクトルの構成についての検討を深める。さらに、同様のシーケンスを生成する特性ベクトルの要素は、必ずしも全く同一とは限らない。例えば、GOT や GPT のクラスが H3、あるいは H4 といったクラスになる場合もあるかもしれないし、ALP が異常値を示す場合もある。つまり、特性ベクトルとして同一ではなくても、類似した特性ベクトルの変動によるシーケンスの遷移の頻度を見る機能を取り入れる。

上述のアプローチを実現するためには、まず、同時に出現する複数の項目をグループ化する。そのため、同時に出現する項目の頻度を求め、クラスタリングを行い、項目グループを構成する。次に、項目グループを単位とした SPM を行い、項目グループの推移の頻出パターンを抽出する。その際、同一あるいは類似の項目グループが連続して現れる場合には無視する。続いて、項目グループの推移における特性ベクトルの依存性を検出する。この時、前述したように、類似特性ベクトルをまとめる機能も実現する。この抽出結果に基づき、項目グループの推移と特性ベクトルの変動から、次に発生すべき項目グループの予測を行い、推薦する。流れを図 2 に示す。

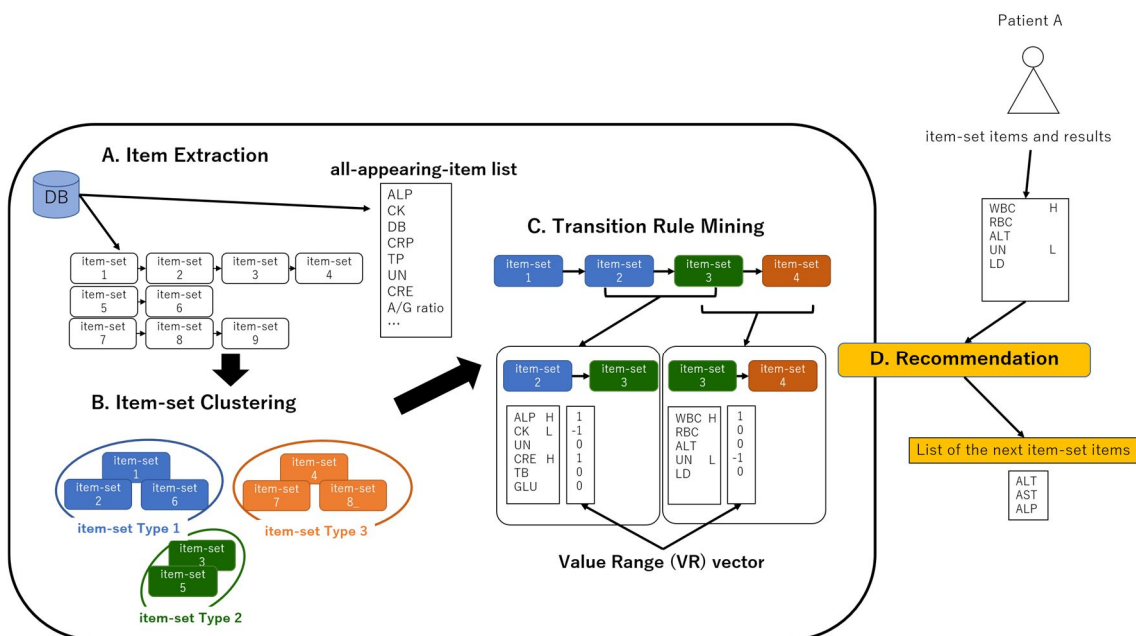


図 2 検体検査項目推薦の流れ

#### 4. 研究成果

##### 4.1 検体検査項目推薦

###### 4.1.1 対象データ

本研究では宮崎大学医学部附属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検体検査の履歴データを対象とした。この検査履歴データは宮崎大学医学部附属病院で使用されている電子カルテシステム WATATUMI によって取得されており、個人情報保護の観点より患者を一意に特定でき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、検査項目数等の情報を表 1 に示す。

表 1 対象検体検査データセット

マイニング用シーケンス数	9,344
マイニング用検査データ数	42,099
テスト用シーケンス数	290
テスト用検査データ数	1,854
検査項目の種類	1,474
代表検査項目の種類	995

###### 4.1.2 検査結果による次の検査項目推薦結果

表 1 の検体検査履歴データから抽出したデータのうち 5% のシーケンスをランダムに選んでテスト用データとし、残りの 95% のデータでクラスタリングし、パターンマイニングを行った後、テスト用のシーケンスの各データを入力として推薦を行い、適合率、再現率の評価を行った。

クラスタリング手法として DBSCAN を用い、パラメータを Eps=1.42, MinPts =30 としたところ、検査データの分類の結果、検査タイプ 0-58 の全 59 タイプに分類された。この検査タイプを用いて提案手法によりシーケンスを解析し、推薦を試みた。このとき、推移ルールは入力デー

タと類似度が高いもの上位3つを選び、推移後検査タイプの重複を省いて、類似度が高い順に推薦した。入力検査タイプ毎の推薦検査タイプ適合率・再現率・F値を図3に示す。検査タイプにより推薦の精度が高いもとの低いものがあることがわかる。これは、検査のタイプにより、パターンとして次に行う検査項目が決まるものと、一般的な検査タイプで次に様々な検査が行われるものがあることを示している。

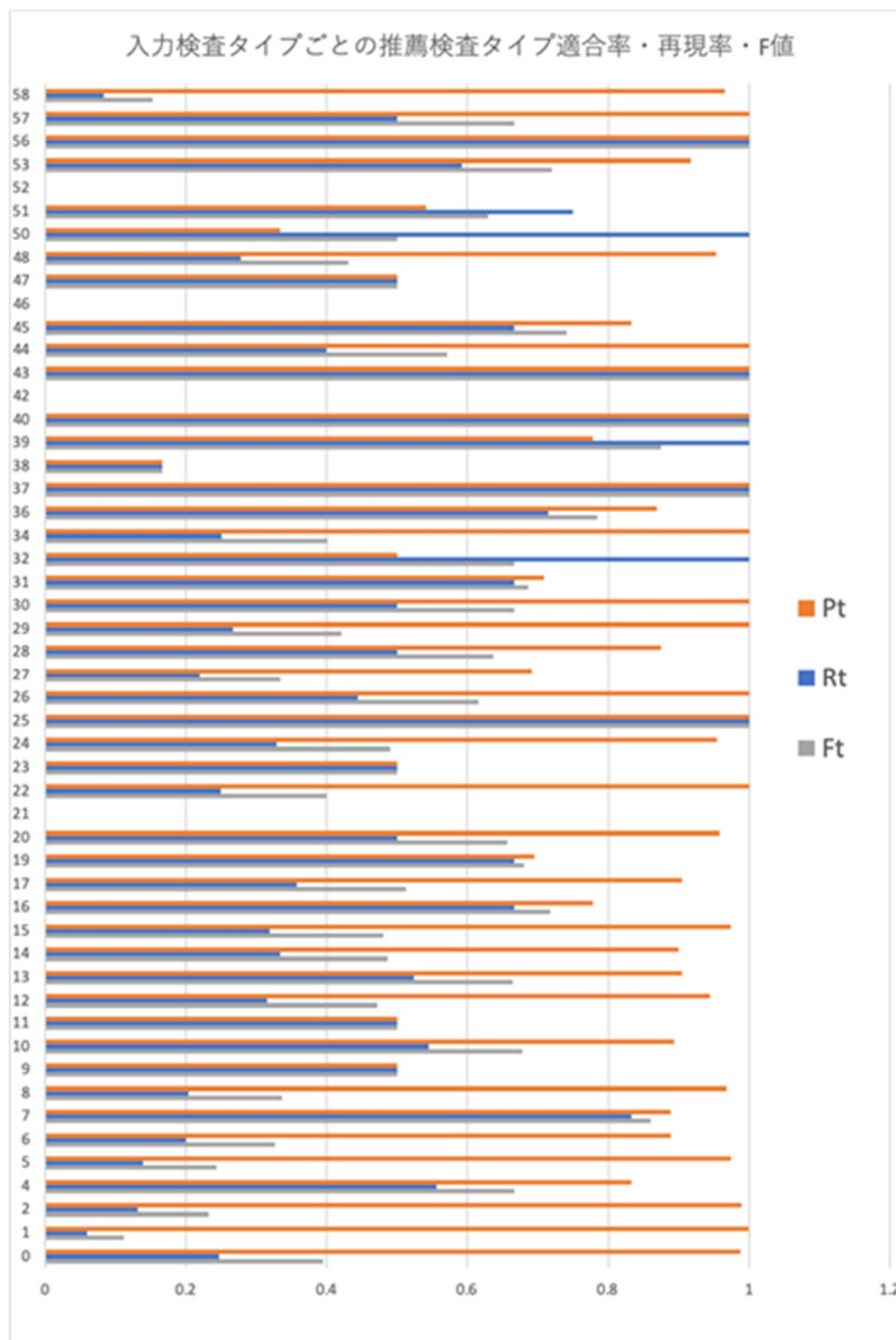


図3 入力検査タイプ毎の推薦検査タイプ適合率・再現率・F値

推薦の具体例として、入力検査タイプの type45 の検査結果に対して、type45 type34 というルールが適用され、実際に次を実施された検査が type34 であった例を図4に示す。これは、手術前に行う検査内容であり、正解との差分になっている項目(背景色を変更している)は心電図検査で、項目に含めないこともある検査項目であり、選択したテストセットでは含まれないパターンであった。医療関係者による確認では、現実に即した結果という判断を得ている。

input	applied rule		recommend	correct
type45	45 to 34		type34	type34
A1b	A1b		肺活量実効値	
UN	UN		肺活量予測値	肺活量予測値
UA	UA		1回換気量実効値	1回換気量実効値
CRE	H CRE	H	予備呼気量実効値	予備呼気量実効値
TB	TB		予備呼気量予測値	予備呼気量予測値
DB	DB		予備吸気量実効値	予備吸気量実効値
GLU	GLU		最大呼気量実効値	最大呼気量実効値
TC	TC		努力性肺活量実効値	努力性肺活量実効値
TG	H TG	H	努力性肺活量予測値	努力性肺活量予測値
Na	Na		一秒量実効値	一秒量実効値
K	K		一秒量予測値	一秒量予測値
C1	C1		一秒率ゲンスラー実効値	一秒率ゲンスラー実効値
Ca	Ca		一秒率ゲンスラー予測値	一秒率ゲンスラー予測値
IP	IP		一秒率テフノー実効値	一秒率テフノー実効値
AST	AST		中間呼気流量実効値	中間呼気流量実効値
ALT	ALT		中間呼気流量予測値	中間呼気流量予測値
LD	LD		エアトラッピング指数実効値	エアトラッピング指数実効値
γGT	γGT		ピークフロー実効値	ピークフロー実効値
A/G比	A/G比		ピークフロー予測値	ピークフロー予測値
血清情報-乳ビ	血清情報-乳ビ		50%肺活量流量実効値	50%肺活量流量実効値
血清情報-溶血	血清情報-溶血		50%肺活量流量予測値	50%肺活量流量予測値
血清情報-ビリルビン	血清情報-ビリルビン		25%肺活量流量実効値	25%肺活量流量実効値
eGFR	eGFR		25%肺活量流量予測値	25%肺活量流量予測値
CRP	CRP		V50V25比実効値	V50V25比実効値
TP抗体定性	TP抗体定性		25%肺活量流量/身長実効値	25%肺活量流量/身長実効値
TP抗体半定量	TP抗体半定量		25%肺活量流量/身長予測値	25%肺活量流量/身長予測値
HBs抗原定性	HBs抗原定性		塵肺判定(%VC)投薬前	塵肺判定(%VC)投薬前
HBs抗原半定量	HBs抗原半定量		塵肺判定(FEV1%-G)投薬前	塵肺判定(FEV1%-G)投薬前
HCV抗体(定性)	HCV抗体(定性)		塵肺判定(V25/H)投薬前	塵肺判定(V25/H)投薬前
HCV抗体(定量)	HCV抗体(定量)		安静時HR	
			安静時RR	
			安静時PR	
			安静時QRS	
			安静時QRS軸	
			安静時QT	
			安静時QtC	

図4 推薦の具体例 ( type45 の検査タイプに対し type34 を推薦した場合の差異)

#### 4.2 医療オータシーケンスへの展開

以上の単一医療機関の医療指示、検体検査のシーケンスの解析を発展させ、複数医療機関のシーケンスの比較を行うことで、医療機関による差異から、医療機関の特色や、他の医療機関のやり方を参考にした改善等につなげることも可能となる。千年カルテプロジェクトといった活動で、実際に複数医療機関の医療データも収集されつつある。

しかし、単一医療機関の中でもバリエーションが存在することからバリエーションを考慮した医療機関間のシーケンス比較が必要となる。本研究では、複数の医療機関のバリエーションの共通部分を抽出し、その共通部分からの差異を提示するために、最長共通サブシーケンスバリエーション(LCSV: Longest Common Subsequence Variant)と、それを用いた併合シーケンスバリエーション(MSV: Merged Sequence Variant)を提案している。LCSVは、従来の最長共通サブシーケンス(LCS: Longest Common Subsequence)をバリエーションを扱えるように拡張したもので、MSVはLCSVを用いてノードに医療機関のラベルを付け、共通部分とそれぞれの医療機関のシーケンスを区別できるようにしたものである。

実際に複数医療機関の経皮的冠動脈インターベンション(PCI: Percutaneous Coronary Intervention)を受けた患者およびCOVID-19の患者の入院期間中に行われた医療指示履歴を対象に、LCSVとMSVを求め、医療機関の違いによる医療オーダーの差異を可視化している。また、医療機関の差異だけでなく、COVID-19の感染波間の違いについても同様にLCSVとMSVを求め差異を可視化している。さらに、バリエーション間の差異を数値化し、その変動を見ることで、COVID-19に対する医療オーダーが変化した時点の特定することも可能にした。

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手法を明らかにした。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4件（うち査読付論文 4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 Hieu Hanh Le, Tatsuhiro Yamada, Yuichi Honda, Takatoshi Sakamoto, Ryosuke Matsuo, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota	4. 巻 4
2. 論文標題 Methods for Analyzing Medical-Order Sequence Variants in Sequential Pattern Mining for Electronic Medical Record Systems	5. 発行年 2023年
3. 雑誌名 ACM Transactions on Computing for Healthcare	6. 最初と最後の頁 1~28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3561825	査読の有無 有
オープンアクセス 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 -
1. 著者名 Yuqing Li, Hieu Hanh Le, Ryosuke Matsuo, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota	4. 巻 DEXA2022
2. 論文標題 Comparison of Sequence Variants and the Application in Electronic Medical Records	5. 発行年 2022年
3. 雑誌名 Proceeding of the 33rd International Conference on Database and Expert Systems Applications	6. 最初と最後の頁 117~130
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-031-12426-6_10	査読の有無 有
オープンアクセス 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 -
1. 著者名 李玉清, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫	4. 巻 1-5
2. 論文標題 シーケンスバリエーションの比較と電子カルテの分析への応用	5. 発行年 2023年
3. 雑誌名 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 1~8
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス 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 -
1. 著者名 Hieu Hanh Le, Yutaka Horino, Tomoyoshi Yamazaki, Kenji Araki, Haruo Yokota	4. 巻 CBMS2021
2. 論文標題 Sequential Pattern Mining of Large Combinable Items with Values for a Set-of-items Recommendation	5. 発行年 2021年
3. 雑誌名 Proceeding of the 34th IEEE International Symposium on Computer-Based Medical Systems	6. 最初と最後の頁 56-61
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/CBMS52027.2021.00017	査読の有無 有
オープンアクセス 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 -

〔学会発表〕 計15件（うち招待講演 1件 / うち国際学会 2件）

1. 発表者名 横田 治夫, Le Hieu Hanh, Li Yuqing, 松尾 亮輔, 山崎 友義, 荒木 賢二
2. 発表標題 複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 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論文集
4. 発表年 2022年

1. 発表者名 小林 うらら, Le Hieu Hanh, 横田 治夫
2. 発表標題 電子カルテと紐づけ解析可能とする生理計測データ集積のPKIとプロキシ再暗号化を用いた実現可能性確認
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 2022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 2022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 2022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 2022年

1. 発表者名 李 玉清, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫
2. 発表標題 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 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 2021年～2022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 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 2021年～2022年



1. 発表者名 Kim Byunghak, Le Hieu Hanh, 横田 治夫
2. 発表標題 暗号化したブロックチェーン上データの解析におけるオフチェーンDB改ざん検知
3. 学会等名 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 2021年～2022年

1. 発表者名 横田 治夫, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二
2. 発表標題 医療データへのシーケンス解析適用とその課題
3. 学会等名 人工知能学会第二種研究会資料, Vol. 2021, No. AIMED-012
4. 発表年 2021年～2022年

1. 発表者名 堀埜 裕, Le Hieu Hanh, 山崎 友義, 荒木 賢二, 横田 治夫
2. 発表標題 電子カルテ中の検体検査結果に基づく次の検査項目推薦の精度向上
3. 学会等名 第40回医療情報学連合大会
4. 発表年 2020年

1. 発表者名 坂本任駿, 小林莉華, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫
2. 発表標題 頻度と実施時刻によるグループ化を採り入れたシーケンス解析に基づく医療指示推薦
3. 学会等名 第13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 2021年

1. 発表者名 小林莉華, 坂本任駿, Le Hieu Hanh, 荒木賢二, 横田治夫
2. 発表標題 動的な患者情報を用いた医療行為推薦を支援するための医療シーケンスの可視化
3. 学会等名 第13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 2021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 34th IEEE International Symposium on Computer-Based Medical Systems (国際学会)
4. 発表年 2021年

1. 発表者名 Haruo Yokota
2. 発表標題 Information Technologies for the Secondary Use of Electronic Medical Records
3. 学会等名 The 15th International Conference on Ubiquitous Information Management and Communication (招待講演) (国際学会)
4. 発表年 2021年

〔図書〕 計1件

1. 著者名 横田治夫	4. 発行年 2022年
2. 出版社 共立出版	5. 総ページ数 232
3. 書名 電子カルテデータ解析: 医療支援のためのエビデンス・ベースド・アプローチ	

〔産業財産権〕

〔その他〕

-

## 6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究分担者	荒木 賢二  (Araki Kenji)  (70274777)	宮崎大学・医学部・教授    (17601)	
研究分担者	小川 泰右  (Ogawa Taisuke)  (60586600)	宮崎大学・医学部・助教    (17601)	
研究分担者	L e H i e u ・ H a n h  (Le Hieu Hanh)  (60813996)	東京工業大学・情報理工学院・助教    (12608)	
研究分担者	山崎 友義  (Yamazaki Tomoyoshi)  (50586609)	宮崎大学・医学部・研究員    (17601)	
研究分担者	串間 宗夫  (Kushima Munuo)  (00727414)	宮崎大学・医学部・研究員    (17601)	
研究分担者	松尾 亮輔  (Matsuo Ryosuke)  (30815931)	宮崎大学・医学部・研究員    (17601)	

## 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