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　５ 年　 ６ 月 ２２ 日現在

機関番号：１５２０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2022
課題番号：１８Ｋ１１４３８

研究課題名（和文）ファジィ状態区間関係に基づく時系列医療データマイニング法の開発

研究課題名（英文）Development of a time-series medical data mining method based on fuzzy ranged relations

研究代表者
平野　章二（Hirano, Shoji）

島根大学・学術研究院医学・看護学系・准教授

研究者番号：６０３３３５０６

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）：本研究では，疾患と治療の時間進展を描出可能な時系列医療データマイニング法の開発に取り組んだ。従来の区間関係に基づくマイニング法に対して，提案法では新たにファジイ化した期間の概念を導入したことで，いくつかの状態が数週間同時に継続する，あるいは数日後に生じるなど，抽象的な期間表現を含む時系列頻出パターンの抽出が可能となった。また，提案法ではファジイ性の導入によって一つの系列が複数の区間関係へ所属できることから，区間関係を細分化した場合においても支持度の低下を抑制できることを人工データを用いた実験により示した。

研究成果の学術的意義や社会的意義
本研究で開発した手法により，イベント発生に至るまでの患者病態・治療等の特徴的推移をMPTP(Minimal Predictive Temporal Pattern)として診療データベースから抽出することができ，背景理解，治療計画立案，アウトカム予測等への活用が期待される。また，プロセスマイニングの一手法として医療以外の様々な分野へも応用可能である。

研究成果の概要（英文）：In this research, we have developed a time-series medical data mining method that can describe the temporal course of disease and treatment. In contrast to conventional mining methods based on interval relations, the proposed method introduces the concept of fuzzy ranged relations, allowing the extraction of frequent patterns that include abstract periods such as several weeks or a few days. Furthermore, experiments on a synthetic dataset demonstrated that the introduction of fuzziness allows for a single sequence to belong to multiple relations, making it possible to suppress the decrease of support values when we define many relations.

研究分野：パターン認識，データマイニング

キーワード：時系列データマイニング　ファジイ状態区間　医療データ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

電子カルテとして日々蓄積される電子健康記録には，処方，検体検査，画像検査，各種記載（所見や看護記録，退院時要約）などが時間情報とともに記録されている。これらを構造化・標準化した形で集積し，データマイニング，機械学習等の手法で分析することで，患者状態の時間推移や投薬・処置等のタイミングをアウトカムと相互に関連づける知識の生成や，アウトカム予測などの診断支援，症例対照研究への応用など医療の質向上に資することが期待される。

時系列医療データの主な特徴として，多次元性，質的・量的情報の混在，収集期間や間隔の不規則性が挙げられる[1]。その性質を踏まえたマイニング法が種々提案されているが，中でもBatal らの方法[2]は，多次元・不等間隔の時系列を入力にとることができ，また投薬など質的情報を組み込める点でも優れた手法といえる。Batal らの方法では，まず検査値の高低や投薬等の状態を状態区間として記述し，それらを時間関係（before もしくは co-occur with）と組み合わせてパターンを構成する。それにより，「PLT 低値 co-occur with 投薬 A⇒血栓発生」など時間関係を含む相関ルールが得られる。しかしながら，この方法で生成されるパターンは，状態の前後関係や共起関係を反映する一方，状態の持続時間や間隔の長短に関する情報を含まない。すなわち，PLT 低値かつ薬 A 投与の状態がどの程度の期間に渡って継続し，それからどのくらい経過して血栓を生じるかを知ることはできない。このように，イベントに至るまでの，疾患の時間進展を表す知識の獲得が難しいことが課題であった。

２．研究の目的

本研究では，Batal らの方法を基礎として，状態の継続期間と間隔をパターンへ組み込んだ新しい時系列データマイニング法を開発する。まず，before, co-occur with 等の区間関係（Allen's Interval Relations）[3]を拡張し，日，週，月など数段階に区分した期間を付与した「期間付き状態区間関係」を新たに定義する。続いて，各関係をファジイメンバシップ関数として定義することで関係にファジイ性を導入し，関係への所属度を元に頻出パターンを抽出する。これにより，一つの系列が所属度に応じて複数の関係に属することを許容し，関係を細分化することに伴う支持度の低下を抑制するとともに，'PLT 低値 co-occurs weeks with 投薬 A'など抽象化された期間の持つ曖昧さを加味したパターン抽出を目指す。

３．研究の方法

3.1 準備：Batal らの MPTP 抽出法

本節では，提案法の基礎となる Batal らの方法について述べる。まず，検査値の高低や投薬の有無などの状態を表す記号Eと，その状態の開始時点b，終了時点eを組み合わせ，状態区間$S=(E,b,e)$を構成する。例えば血小板数が低値の状態(PLT=L)が時点 1 から時点 2 まで生じている場合，$S=('PLT=L',1,2)$となる。このような状態区間Sを開始時刻順に並べることで，任意の時系列を状態系列$<S_1,S_2,\dots,S_k>$として記述できる。パターンはこの状態系列を元に構成されるが，各々の状態が区間を伴うため，区間同士の関係を定義しなければならない。Batal らは，Allen の論文で定義される 13 種類の関係の中から，$'E_i$ before E_j' $(e_i<b_j)$ と $'E_i$ co−occurs with E_j' $(b_j\le e_i)$の 2 種類を利用している。この関係を，状態系列に含まれる全ての状態組$\{(S_i,S_j):i<j\}$に対して割り付けることで，長さkのパターンP_k(k−pattern という)を，k個の状態とその関係マトリクスRを用いて$P_k=(<S_1,S_2,\dots,S_k>,R)$と表現する。関係マトリクス$R$は$(k-1)\times(k-1)$の三角行列で，例えば図 1 左に示す 3 状態$S_1$，$S_2$，$S_3$の場合は同図右に示すものとなる。なお，本稿では簡単のため，記号「|」を行の区切り，記号「.」を列の区切りとして関係マトリクスを 1 行に展開した形式で，$S_1-S_2-S_3\,|\,b.b\,|\,c$ のようにパターンを表現する。

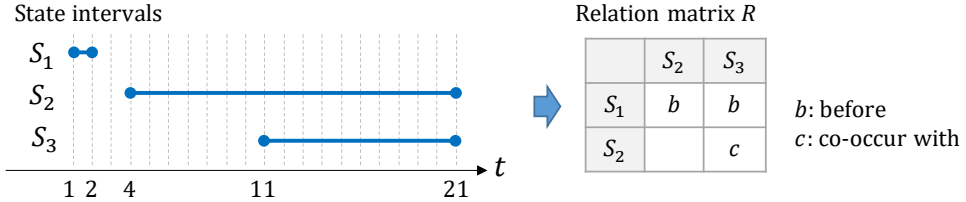

図 1: 状態系列と関係マトリクスの例

頻出パターンの抽出は，候補の生成と，支持度に基づく候補選別の 2 段階で行われる。候補の生成では，長さkの頻出パターンの先頭に頻出 1−pattern を 1 つ挿入することで，新たな候補$(k+1)$−pattern を生成する。また，追加した 1−pattern と 2 番目以降のk個の状態との関係を格納する新たな 1 行を関係マトリクスの先頭に挿入し，そこへ関係bとcの組み合わせのうち時間矛盾を生じないものを生成して代入する。生成された候補のうち，正例クラス内の支持度が所

与の最小サポート閾値より高いものが頻出パターンとなる。また，自身のサブパターンのどれよりも有意に予測能力が高い頻出パターンを MPTP(Minimal Predictive Temporal Pattern)として抽出する。その評価はサブパターンの支持度の最大値を母数とする片側二項検定により行い，有意確率が所与の水準より小さい場合に MPTP と判定する。

3.2 期間付き状態関係の導入とファジイ化

前節の方法では状態の前後関係及び共起関係をパターンにより記述できるが，その関係の時間的な長さについて表現することはできない。本研究では，before 関係および co-occur 関係を拡張した「期間付き状態関係」を新たに考案し，前節のプロセスに組み込む。さらに，関係をファジイ化することで一つの系列が所属度に応じて複数の関係に属すること，すなわち，一つの系列が同時に複数の区間関係を支持することを可能にする。

まず，開始時刻で昇順ソートされた状態系列において，2 つの状態$\{(S_i, S_j): i < j\}$の時間差d_{ij}を次式により定義する。

$$d_{ij} = \begin{cases} e_i - b_j & \text{if} \quad e_i < e_j \\ e_j - b_j & \text{otherwise} \end{cases} \qquad (1)$$

ここで，e_iは状態S_iの終了時点，b_jとe_jは状態S_jの開始時点と終了時点である。S_jはS_iより後にあるため，$d_{ij} < 0$の場合は before 関係，$d_{ij} \geq 0$の場合は co-occur 関係となる。$e_i \geq e_j$の場合，後から始まったS_jがS_iよりも先に終わることから，共起の期間はS_jの開始から終了までとなり，$e_j - b_j$と定義する。このことは，Allen の関係におけるS_j during S_iを期間付きの co-occur 関係として取り扱うために必要となる。図 1 の例では，S_2とS_3の時間差は，$d_{23} = 21 - 11 = 10$となる。

次に，この時間差d_{ij}に基づく期間付き状態関係を定義する。関係の数および期間は任意に定めることができるが，ここでは表 1 に示す 6 種類の関係を定義する。例えば関係 *cd* (co-occur days with)は，2 つの状態が数日間にわたって共起している関係を表す。さらに，数日という期間のもつ曖昧さを組み入れるために，状態関係を図 2 に示すようにファジイメンバシップ関数として定義する。ここでは，以降の処理を簡単化するため，メンバシップ関数の形状を台形とし，重なりは最大で 2 関係までとした。台形メンバシップ関数の形状パラメータとして，L0(左下)，L1(左上)，R1(右上)，R0(右下)の各点と対応する日数差d_{ij}を表 1 に示すように各関係について定義する。なお，$d_{ij} = 0$の点は 2 状態が同日に生じていることから co-occur 関係に含める。

Relations		Function params (days)	
Name	Meaning	[L0, L1]	[R1, R0]
cd	co-occur days	[0, 0]	[5, 14]
cw	co-occur weeks	[5, 14]	[31, 62]
cm	co-occur months	[31, 62]	[inf, inf]
bd	days before	[-14, -5]	(0, 0)
bw	weeks before	[-62, -31]	[-14, -5]
bm	months before	[-inf, -inf]	[-62, -31]

表 1: 期間付き状態区間関係

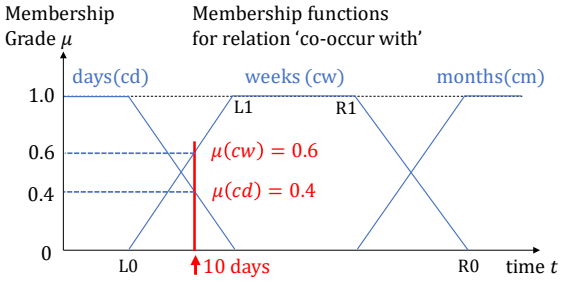

図 2: 状態区間関係とメンバシップ関数

図 1 の例の場合，状態S_2とS_3の日数差$d_{23} = 10$であり，図 3 に示すとおり関係*cd*と関係*cw*に対してそれぞれ所属度約 0.4 と約 0.6 で同時に属することとなる。同様に，S_1とS_3は$d_{13} = -9$であることから関係*bd*と*bw*にそれぞれ所属度約 0.6 と約 0.4 で属する。

続いて，複数の関係とそれらへの所属度を表現できるように，状態間の関係マトリクスRを拡張したファジイ関係マトリクスRFを次式により定義する。

$$RF_{ij} = \{r: \mu(r)\}, \quad \forall r \text{ s.t. } \mu(r) > 0 \qquad (2)$$

ここで，rは表 1 に示す*cd*，*cw*等の期間付き状態関係，$\mu(r)$は関係rに対する状態S_iとS_jの所属度である。所属度が 0 を上回る関係について，(関係：所属度)のペアをファジイ関係マトリクスRFの要素へ格納する。前出の例の場合，RFは右のとおりとなる。

	S_2	S_3
S_1	$\{bd: 1.0\}$	$\{bd: 0.6, bw: 0.4\}$
S_2		$\{cd: 0.4, cw: 0.6\}$

以上の手続きにより，状態系列Sとファジイ関係マトリクスRFにより表現された事例集合を構成する。続いて，次節の方法により候補パターンと各事例とのマッチングを行う。

3.3 候補パターンとのマッチング

候補パターンと事例とのマッチングは，(1)状態系列のマッチング，(2)関係マトリクスのマッチング，の 2 段階で行う。状態系列のマッチングは，候補パターンの状態系列が事例の状態系列に順序を保って含まれる場合に真となるもので，元手法と同様である。2 段階目の関係マトリクスのマッチングは状態系列のマッチングが真となった場合に限り行う。期間の導入によって関係マトリクスの種類が増加し，関係マトリクスは異なるが状態系列は等しい候補パターンが多

数生成されることから，このように 2 段階のマッチングを行うことで，状態系列がマッチしない事例については関係マトリクスのマッチングを全て省略し，探索を効率化することができる。

関係マトリクスのマッチングは，候補パターンが通常の関係マトリクスR，事例がファジイ関係マトリクスRFであることから，以下の方法により所属度を考慮して行う。まず，所属度に対する閾値th_μ $(0 < th_\mu \leq 1)$を導入する。そして，次の条件が満たされる場合に$R \subseteq RF$，すなわち候補パターンの関係マトリクスRが事例の関係マトリクスRFに含まれるとみなす。

$$R \subseteq RF \quad \text{if} \quad \min\{\mu(r) \mid \forall r \in R\} \geq th_\mu \qquad (3)$$

ここで，rはRの各要素を示す。RFにおいて，対応する状態組が同じ関係rに属し，かつそれらの所属度$\mu(r)$が全てth_μ以上であるとき，$R \subseteq RF$となり，最終的に候補パターンが事例とマッチすると判断する。

3.4 候補パターン生成時の時間制約

3.1 節で述べたように，候補パターンを生成する際，先頭に追加した 1-pattern と残りの状態との関係として可能な組み合わせを列挙して関係マトリクスの空行を埋める必要がある。before 関係と co-occur 関係には時間制約があり，同一行において before より後に co-occur が入ることはない。期間付き状態関係においても同様に時間制約があり，関係マトリクスにおいて，いずれかの before 関係が先行する場合，それ以降，同じ行では before 関係のみが許容される。また，$bd < bw < bm$の関係性から，後ろに来る before 関係は前の before 関係と同等以上の長さである場合のみ許容される。これらの制約に基づき，不要な候補パターンを除外する。

４．研究成果

人工的に生成した系列データセットに対して提案法を適用し，ファジイ化した期間付き状態区間関係に基づく MPTP の抽出を試みた。データセットの生成方法は以下のとおりである。まず，各事例の状態系列を 3 つの状態(S_1, S_2, S_3)からなるものとし，各状態の開始基準日を次のとおり定めた。これらは，状態S_1を中心におき，S_2とS_3の発生順を両クラスで入れ替えたものである。

class P（正例）： $S_1 = 3,\ S_2 = 6,\ S_3 = 0$
class N（負例）： $S_1 = 3,\ S_2 = 0,\ S_3 = 6$

この開始基準日に対して，区間[0, 7]の一様乱数をオフセットとして加えたものを，各事例における各状態の開始日とした。これにより，例えばクラス P の系列における状態S_2は，開始日が第 6 日から第 13 日の範囲で一様分布する。同様に，状態の継続期間を区間[1, 14]の一様乱数とした。データ数はクラス P，クラス N ともに 500 とした。生成した系列セットにおける状態の順序分布を表 2 に示す。なお，同表及び以降の各図表においては簡便のため各状態S_i $(i = 1,2,3)$の記号Sを省略しインデックスの数字(1, 2, 3)により表記する。例えば同表の 3-1-2 は，状態がS_3，S_1，S_2の順で発生する系列を示す。同表のとおり，いずれのクラスにおいても状態の発生順序が開始基準日と同順であるのものが約 6 割，異なる順序であるものが約 4 割である。

表 2：テストデータにおける状態の順序分布

	3-1-2	1-3-2	3-2-1	1-2-3	2-3-1	2-1-3	total
class P	289	100	88	10	9	4	500
class N	3	12	5	81	87	312	500
total	292	112	93	91	96	316	1,000

状態の開始日と継続期間から，任意の 2 状態(S_i, S_j)間の関係が定まる。図 3 に状態S_i，S_jの日数差d_{ij}の分布を示す。差d_{ij}が< 0である場合にいずれかの before 関係，≥ 0である場合にいずれかの co-occur 関係となる。同図上側に日数差とファジイ区間関係との対応をあわせて示す。本実験で使用したファジイ区間関係は表 1 で定義したものと同一である。テストデータには計 1,500 組の状態組が存在する。そのうち，所属度 1.0 で関係bdもしくはcdへ完全に属するもの($|d_{ij}| \leq 5$であるS_i, S_jの組)は，クラス P で約 74.8%，クラス N で約 73.3%であった。一方，関係bdとbw，もしくは関係cdとcwなど，複数の関係へ同時に所属するもの($|d_{ij}| > 5$であるS_i, S_jの組)はクラス P で約 25.2%，クラス N で約 26.6%であった。これらの状態組を含む系列では，所属度閾値th_μに応じて内包する関係マトリクスが変化するため，抽出されるパターンの支持度や確信度に影響を及ぼし得る。

所属度の閾値th_μを 0.0 から 1.0 まで 0.1 刻みで変化させ，頻出パターン及び MPTP を抽出した結果を図 4 に示す。同図において，k-FP は当該閾値において抽出された長さkの頻出パターン数を，k-MPTP は同じく長さkの MPTP 数をそれぞれ表す。なお，本データセットにおいては状態S_1，S_2，S_3が全ての系列に含まれるため，1-FP は常に 3 個となり，そのいずれもクラスに特異的では無いため 1-MPTP は存在しない。同図から，$k \geq 2$においては FP 数，MPTP 数ともに閾値 0.0 で最大値をとり閾値の増加とともに単調減少する傾向が見られた。

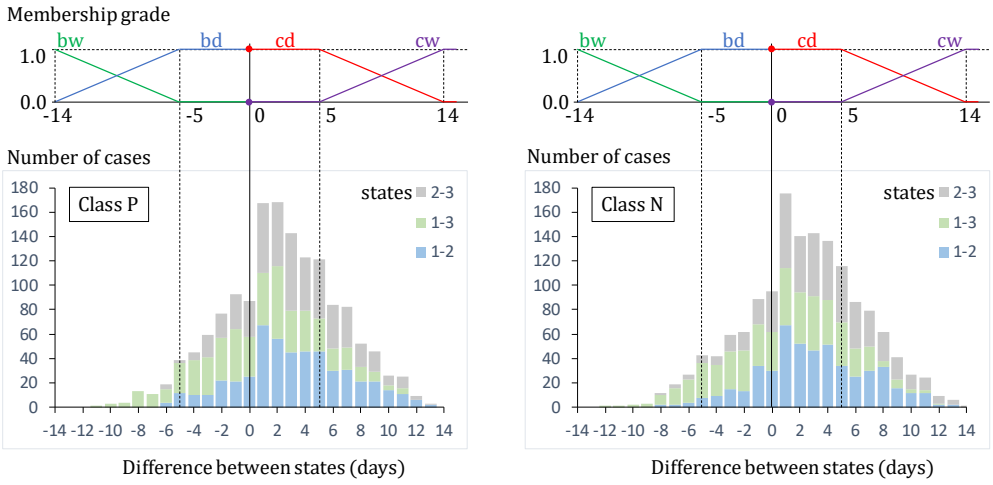

図 3: テストデータにおける状態間の日数差の分布

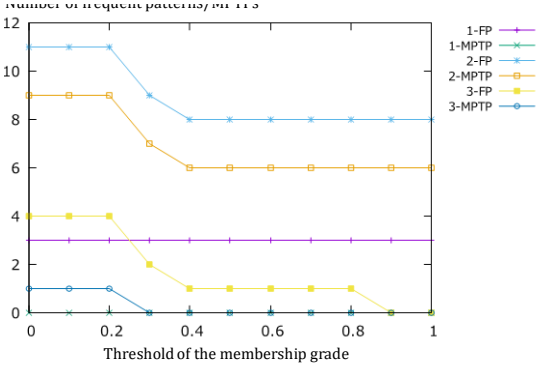

図 4: 所属度閾値と MPTP 数及び頻出パターン数

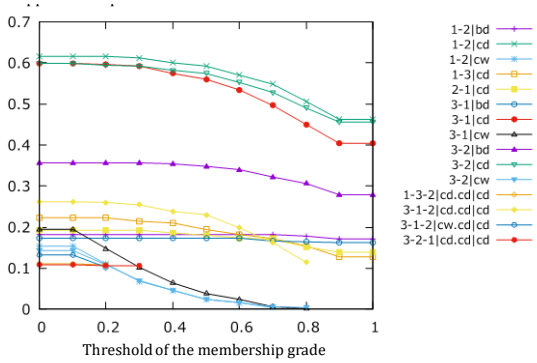

図 5: 所属度閾値と各 FP・MPTP の支持度

抽出された MPTP のうち，$k \geq 2$である 10 個とそれぞれが抽出された所属度閾値の範囲を表 3 に示す。表 2 のとおりクラス P には 3-1-2 を順序にもつ系列が多く含まれることから，その部分系列を中心とした MPTP が生成されている。状態順序が同一で関係の異なる 2 つのパターン $1-2|cd$と$1-2|cw$ を比較すると，cdが全ての閾値で MPTP として抽出された一方，cwは閾値が 0.2 までの範囲に限定されている。

表 3: MPTP が抽出されたときの所属度閾値

Pattern	Range of th_μ	Pattern	Range of th_μ
1-2|bd	0.0-1.0	3-1|cw	0.0-0.3
1-2|cd	0.0-1.0	3-2|bd	0.0-1.0
1-2|cw	0.0-0.2	3-2|cd	0.0-1.0
3-1|bd	0.0-1.0	3-2|cw	0.0-0.2
3-1|cd	0.0-1.0	3-1-2|cd.cd|cd	0.0-0.2

図 5 に所属度閾値と各 FP・MPTP の支持度の関係を示す。支持度についても全体として閾値の増加とともに単調減少する傾向が見られ，前出のパターン $1-2|cw$ は，所属度閾値が 0.2 を超えたところで支持度が最小支持度である 0.1 を下回り，MPTP ではなくなる。この事象は，状態S_1とS_2の共起期間が関係cwに属する長さではあるものの，図 3 の分布で 6-7 日付近にあたるケースに起因しており，cwへの所属度が低いため閾値th_μの上昇とともに早期にcwから外れ，$1-2|cw$ を支持するケースが減少することによる。$3-1|cd$ と $3-1|cw$ など他のパターンについても同様である。これらから，提案法によって，状態間の関係に含まれる days，weeks などの期間がファジイ化され，一つのケースが同時に複数の関係を支持できるようになること，また，関係を満たす度合いが所属度によって表現され，どの程度の所属度をもつ場合にパターンを支持させるかを閾値によって指定できることが示された。

参考文献

[1] Jensen PB, Jensen LJ, Brunak S.: Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet.13(6):395-405 (2012).

[2] Batal I, Valizadegan H, Cooper GF, Hauskrecht M.; A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data. ACM Trans Intell Syst Technol. 4(4) (2013).

[3] Allen JF.: Maintaining knowledge about temporal intervals. In: Communications of the ACM. 26 (1983).

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 平野章二，津本周作
２．発表標題 ファジイ区間関係に基づく時系列医療データからの頻出パターンマイニング
３．学会等名 第8回人工知能学会医用人工知能研究会
４．発表年 2019年

１．発表者名 Shoji Hirano, Shusaku Tsumoto
２．発表標題 Mining frequent temporal patterns from medical data based on fuzzy ranged relations
３．学会等名 IEEE BigData 2019（国際学会）
４．発表年 2019年

〔図書〕　計0件

〔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

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