令和２年度　厚生労働科学研究費補助金
政策科学総合研究事業（臨床研究等ICT基盤構築・人工知能実装研究事業）

総括研究報告書

# 多施設 SS-MIX2 標準化データベースからの臨床的表現型クラスタリングとその臨床エビデンス創出手法の開発研究

研究代表者　大江和彦（東京大学 医学部附属病院・教授）

研究要旨

目的：電子カルテ由来の多施設 SS-MIX2 標準化ストレージの臨床データを使用して、教師なしクラスタリング等の手法により、臨床的表現型をクラスタリングし、その臨床的特徴集団を抽出してその特徴を考察することを目的とする。また臨床的リアルワールドビッグデータから教師なしクラスタリング手法により未知の特性集団を得る手法について課題を検討する。

方法：8 大学病院の SS-MIX2 標準化ストレージから 3 年分の病名登録データと検体検査データを取得し、それぞれにおいて、a)血液系疾患 (D50-D77)、b)免疫系疾患(D80-D89)、c)内分泌代謝系疾患(E00-E87) 、d)高血圧疾患(I10-I15)、e)心不全 (I50)、f)炎症性関節炎(M05-M14)、g)結合織障害(M30-M36)、h)腎糸球体・腎機能障害疾患(N00-N19)に 7 領域に分けたデータセットを作成した。これらのデータセットに対して、非階層的クラスター分析法として、1) PAM(Partitioning Around Modroids)法、金剛分布モデルにもとづくクラスタ分析法として、2) EM (expectation maximization) アルゴリズムによりパラメータとクラスラベル推定を行う混合分布モデル (VII：球型、異なる体積) に基づいたクラスタ分析を行い、それぞれについて分布プロットにより可視化し、考察した。

結果と考察：多施設 SS-MIX2 標準化ストレージの臨床データからの解析データセット構築ができ、それに対して非階層的クラスタリングを実施した。対象となった検査件数の元データ規模は、各施設で 1400 から 9100 万件と大規模で、疾患領域ごとでも大きな領域では 24 万件以上の規模であった。クラスタ解析では、当初目的とした臨床的に特徴が明確な小規模クラスタの検出はできなかった。しかし ICD 分類の 3 桁、4 桁目を横断する複数の大きなクラスタとは別に小さなクラスタの存在が示唆された。これらのクラスらの臨床的意味付けはそれに所属するここのレコード抽出をして検討が必要であり、今後そのクラスタごとの薬剤治療別の予後や経過分類の違いを分析する必要がある。また、大きなクラスタに所属するケースを除外した多施設統合データを再解析するなどの手法の必要性も示唆された。一方、領域ごとにみても施設間で含まれるケースの多様性に大きな違いがあり、ビッグデータのままで統合することはかえって少数からなる特性集団を埋れさせる可能性が考えられた。

研究分担者[1]

| | | | |
|---|---|---|---|
| 中山雅晴・東北大学 | 教授 | 松村泰志・大阪大学 | 教授 |
| 近藤克幸・秋田大学 | 理事 | 津本周作・島根大学 | 教授 |
| 白鳥義宗・名古屋大学 | 病院教授 | 中島直樹・九州大学 | 教授 |
| 木村通男・浜松医科大学 | 教授 | 関　倫久・東京大学 | 助教 |

1 本研究は、研究分担者のデータ統合によりひとつの研究テーマをとして実施したため、分担研究者ごとの個別の報告書は作成しない。

## A.研究目的

**背景：**臨床エビデンスは、「高血圧合併２型糖尿病」のように特定の特性を有する患者集団を事前規定し、「阻害薬が有効」のようにその集団における別の臨床特性の存在を確認することで得られる。クリニカルクエスチョン(CQ)を思いつかなければ事前に集団を規定できず、存在を確認すべき臨床特性が不明で研究デザインができない。臨床の場では、患者の臨床特性で規定される集団が、別のどのような臨床特性を有するかを知りたいことが多いが、具体的なCQを思いつかないことが多く、DB駆動型のCQ自動生成、エビデンス示唆を得る手法の開発が必要である。

**目的：**本研究では、電子カルテ由来のSS-MIX2標準化多施設臨床データベース（DB）を使用して、
①教師なし機械学習による自動クラスタリング等の手法により、臨床的表現型において共通特性をもつ集団（クラスタ）を多数自動生成し、②得られたクラスタの他の臨床情報特性を時系列変化を含めて類型化の手法を検討する。③その臨床的特性の出現確率等の統計的特性やその臨床的意味付けを分析し、④診療中の患者の電子カルテデータから上記クラスタに自動分類し、その結果にもとづいた臨床的特性を可視化することの臨床的有用性を評価する、ことを目指した。

## B．研究方法

１）分析用データセットの作成環境の構築

本研究では、8大学病院のSS-MIX2標準化ストレージに蓄積されている傷病名データと検体検査結果データを使用し、初年度の教師なし機械学習による自動クラスタリングを実施するための分析用データセットの作成手法を確立することが必要である。1年目の1施設でのデータでパイロット的にクラスタリングを実施した結果にもとづき、以下の手順で分析用データセットを作成することとした。

1-1）疾患対象の絞り込み

病名データで以下のICD10コードの確定診断を有する８つの患者集団をICD10コードとともに抽出した。

a)血液系疾患(D50-D77)、b)免疫系疾患(D80-D89)、c)内分泌代謝系疾患(E00-E87) 、d)高血圧疾患(I10-I15)、e)心不全(I50)、f)炎症性関節炎(M05-M14)、g)結合織障害(M30-M36)、h)腎糸球体・腎機能障害疾患(N00-N19)。

これらを選択したのは、これらの疾患群では疾患相互および疾患内の血液検査結果のパターンだけによってもその集団特性を表現できる可能性があるのに対して、感染症、腫瘍性疾患、精神疾患、消化管炎症性疾患、外傷等はこの可能性が低いという理由による。

1-2)対象期間と期間ウインドウの設定

対象期間は、2018年1月から2020年12月まで36ヶ月間とし、この期間中に前記ICD10コードの病名が新たに登録された病名レコードを対象とした。36ヶ月の期間中に同一のICD103桁コードの新規登録が複数回出現する患者では、その最初の出現レコードだけを対象とした。これは、現在の電子カルテでは、患者の病名出現状況によらず、治療期間中に何度も同じ病名を登録することがあるためと、期間中の最初の登録だけを対象としても本研究目的にはデータ量としては十分であることによる。また最初に病名登録がなされた時期が、もっともその疾患による影響が検体検査結果等に現れていると考えたためでもある。

次にこの36ヶ月を3ヶ月ごとの期間ウインドウに時間的粒度を粗くし、期間ウインドウ番号1から12を割り当てた。

1-3）検体検査結果データセットの作成

検体検査結果は1年目の調査にもとづき、8施設すべて標準コードが振られていて、全体件数としてそれぞれの患者集団における個々の疾患存在期間（診断開始日ー終了日）内において、全体で10万件以上の検査実施数がある検査項目を抽出した。

最終的に、抽出対象となった検査項目はJLAC10標準コード項目として90項目であった(表1､末尾掲載)。

その上で、この対象項目で前記36ヶ月に実施された血液検査結果を抽出した上で、同一患者ごと、各ICD10コードごとに前記3ヶ月ウインドウ期間における各検査値の平均値を求め、この期間の検査結果の代表値とした。

次に、前項の新規登録のあった病名レコードのウインドウ期間中の検査データだけをデータセットに

含め、それ以外の期間の検査代表値は含めないこととした。
以上の分析用データセットを R 言語（Ver3.6）で作成し、施設を指定して自動的に分析用データセットを生成する環境を構築した。

1-4）欠損値の取り扱い方法
教師なしクラスタリングでは、欠損値が存在していると処理が困難となる。そこで、新規登録のあった病名レコードのウインドウ期間中に検査代表値が存在しない（すなわちその 3 ヶ月ウインドウ内に検査が 1 度も行われていない）ウインドウ期間においては、同じ患者の他のすべてのウインドウ期間（すなわち 36 ヶ月間）での検査があればその平均値を、それがない場合には、当該施設での今回の抽出データセットにおけるその検査の平均値（すなわち施設平均値）を、代替値として設定した。
それでもなお欠損値となる検査項目についてはその ICD10 グループデータからは削除した。

２）クラスタリング
1 年目の予備分析では教師なし機械学習のクラスタリング手法である K-Means++によりクラスタリングの試行を Python scikit-learn ライブラリを用いて実施した。K-Means++は最初にクラスタ数を設定する必要がありこと、各クラスタのデータ件数を同じに近づけようとする特性があること、今回のように次元数が 70 から 90 に達するデータセットでは必ずしも適切でないことなどから、今回は 1）PAM(Partitioning Around Modroids)法、2) EM (expectation maximization) アルゴリズムによりパラメータとクラスラベル推定を行う混合分布モデル（VII：球型、異なる体積）に基づいたクラスタ分析を行い、それぞれについて分布プロットにより可視化した。

PAM(Partitioning Around Modroids)法は、Kaufman&Rousseeuw により提案された非階層的クラスタリング手法で、最適な分類を求めるアルゴリズムは k-means 法に似ているが、入力データが三角不等式を満たさない非類似係数でも使えるという利点があり、採用した。

モデルに基づいたクラスター分析は、観測データが異なる分布の混合分布であると仮定し、個体が属するクラスのラベルをも隠れ変数として推定する。混合分布のパラメータおよびクラスのラベルの推定は EM（expectation maximization）アルゴリズムがよく用いられている。

本研究では、PAM 法には R 言語のパッケージ cluster の pam 関数を、 また、混合分布モデル（VII：球型、異なる体積）に基づいたクラスタ分析には、R 言語のパッケージ mclust の mclustBIC 関数を用いてモデルの BIC 値を求めクラスタ数の目標を定めた。関数 mclustBIC は、最大尤度推測法を用いた EM アルゴリズムでパラメータを推定する際に必要となる、ガウス混合分布モデルの情報量規準 BIC（Bayesian Information Criterion）値を求めるものである。

3)結果の可視化
PAM にもとづくクラスタリングは、結果を R 言語パッケージの cluster plot により実施した。また、Silhouette Plot を行い、各クラスタに割当てられた件数を出力した。
モデルに基づいたクラスター分析では、約 70 個の変数を比較的関係の深い、または臨床上で検査第分類と考えられる項目の組み合わせを意識して set1 から set8 に分類し、それぞれの項目間での散布図におけるクラスタリング状況を可視化することとした。なお、クラスター数は、BIC 値がクラスター数 4 以上はほぼ変化がない傾向を示したため、少し大きめの 5 を採用し、モデルは"VII"（球型、異なる体積）を採用した。

## C．研究結果

1）8 施設ごとのデータセット構築
1-1）病名データ
SS-MIX2 標準化ストレージの病名データは PPR メッセージ中の PRB セグメントに記述されるが、病名開始日の意味を持つフィールドは、PRB-7（プロブレム設定日付：診断日）と PRB-16(プロブレムの発生日付：開始日)の 2 つが存在し、区別がつけがたい。今回の 8 施設中、4 施設では PRB-7 にのみ開始

日付が格納され、PRB-16 は欠損値出会ったのに対して、残りの 4 施設は逆であった。
ICD10 コードはいずれの施設も全件で格納されており問題なかった。
1-2) 検査結果単位
検査結果の単位は、2 施設で一切格納されておらず同一ベンダーの同一システムであった。システムベンダーの担当者に調査依頼したところ、実装ミスで出力されていないことが判明したが、今回は研究期間中の修正はできなかった。
1-3) テスト患者データの削除
SS-MIX2 標準化ストレージにはテスト患者のデータが混じっている施設があった。特異な外れ値を示す患者の検出などで削除可能であるが、実データに近いテストデータを設定されたテスト患者については統計的な手法での検出は困難である。

2) 8 施設ごとのデータセットのプロファイル
表 2 にデータセット全体の規模を示す。8 施設全体を統合した場合のデータセット件数は、免疫系疾患(D80-D89) が 7,027 件が最も少なく、内分泌代謝系疾患(E00-E87) が 234,634 件と最も多かった。

3)各疾患領域ごとの 8 施設におけるクラスタリング
各疾患領域ごとの 8 施設分の PAM にもとづくクラスタリングを図 1 から図 8 に示す。各図には施設 1 から 8 でのプロットが示されている。各施設の下段に表示されているように、いずれも 2 合成成分の説明率は 10 から 15%と低かった。

またモデルに基づいたクラスター分析（5 クラスタ）の結果を、散布図でで検査項目セット 1 から 8 にわけてクラスタを色別にプロットした。施設 1 の場合を図 9 に示す。

## D．考察

SS-MIX2 標準化ストレージからのデータ取得では、病名開始日の格納フィールドが施設によって異なるなど、意味的な解釈の違いをあらかじめ整理して把握しておくことが必要であった。また、検査結果の単位の確実な把握と、1 施設内のデータであっても途中で単位が変わるケースがあるため、これらをクレンジング段階で検出し、データ変換しておくことが必要であった。

また、検査結果の単位情報の欠落は、今回の研究では多施設データを 1 データに統合してからの分析は行わなかったので支障はないが、多施設統合データを作成する場合には、単位の整合性をとることが必須であり、このような単位情報の欠落は大きな問題となりうると考えられ、情報の整備が必要である。

クラスタ分析については、次元数が 70 から 90 と比較的大きく、小さなクラスタが大きなクラスタのなかに埋もれてしまっていると考えられる。
疾患領域ごと、施設ごとに図 1 から図 8 を見ると、比較的均質の集団で構成されている施設と、非常に多様な疾患が混在している施設があることがわかる。そのため、今回は全施設統合データでの解析は、かえって小さな特徴を隠してしまうと考えられたので行わなかったことは妥当であると思われた。

一方、今回は多施設 SSーMIX2 のデータストレージが一斉にシステムリプレースの時期と重なってしまい、データを再取得できるようになったのが 2 年目の後半となったため、データセットを整備してクラスタ解析結果を、もとの電子カルテデータと個々に付き合わせる分析まではできなかった。散布図ではいくつか特徴的な小さなクラスタが存在していることが窺えるため、今後は、それを分離するための手法を開発していきたい。

手法としては、今回はクラスターを 5 に設定して解析したが、むしろ 10 ないし 15 程度に設定した上で、集団規模の大きい上位 5 程度のクラスターに所属するケースを削除した上で、残りのデータセットを多施設統合してから、再度クラスター解析を行うなどの手法が考えられる。

データ処理上の課題としては、今回、テラバイトサイズの十分なメモリーを確保した上で R 言語パッケージによる処理を行ったが、それでも糖尿病疾患領域などデータ件数が多い領域では、クラスタ解析後のプロットでエラーとなることがあり、規模の大きいデータを対象とした処理環境には、課題があると考えられた。

## E．結論

多施設の SS-MIX2 標準化ストレージから 8 疾患領域について検査結果値データセットを整備し、非階層的クラスタリングとモデルに基づいたクラスター分析を試みた。未知の特徴的な検査結果パターンを示すクラスターは検出できなかったが、主たるクラスタから外れる小さなクラスタが存在していることが示唆された。今後が、それに該当するケースを個々に抽出して、時系列的な検査結果値を追加して解析することが課題と考えられた。

**Ｆ．健康危険情報**

特になし

**Ｇ．研究発表**

**1. 論文発表**


1. Ma X, Imai T, Shinohara E, Kasai S, Kato K, Kagawa R, Ohe K. EHR2CCAS: A framework for mapping EHR to disease knowledge presenting causal chain of disorders - chronic kidney disease example. Journal of Biomedical Informatics. 2021;115: 10369.
2. Seki T, Kawazoe Y, Ohe K. Machine learning-based prediction of in-hospital mortality using admission laboratory data: A retrospective, single-site study using electronic health record data. PLoS ONE. 2021;16(2): e0246640.
3. Okuda M, Yasuda A, Tsumoto S, An approach to exploring associations between hospital structural measures and patient satisfaction by distance-based analysis. BMC Health Serv Res. 21, 63-70, 2021.
4. Shusaku Tsumoto, Tomohiro Kimura, Shoji Hirano, Determination of Disease from Discharge Summaries, The Review of Socionetwork Strategies 15, 49-66,2021.
5. Sofue T, Nakagawa N, Kanda E, Nagasu H, Matsushita K, Nangaku M, Maruyama S, Wada T, Terada Y, Yamagata K, Narita I, Yanagita M, Sugiyama H, Shigematsu T, Ito T, Tamura K, Isaka Y, Okada H, Tsuruya K, Yokoyama H, Nakashima N, Kataoka H, Ohe K, Okada M, Kashihara N.Prevalence of anemia in patients with chronic kidney disease in Japan: A nationwide, cross-sectional cohort study using data from the Japan Chronic Kidney Disease Database (J-CKD-DB).PLoS One. 2020 Jul 20;15(7):e0236132.
6. Sofue T, Nakagawa N, Kanda E, Nagasu H, Matsushita K, Nangaku M, Maruyama S, Wada T, Terada Y, Yamagata K, Narita I, Yanagita M, Sugiyama H, Shigematsu T, Ito T, Tamura K, Isaka Y, Okada H, Tsuruya K, Yokoyama H, Nakashima N, Kataoka H, Ohe K, Okada M, Kashihara N.Prevalences of hyperuricemia and electrolyte abnormalities in patients with chronic kidney disease in Japan: A nationwide, cross-sectional cohort study using data from the Japan Chronic Kidney Disease Database (J-CKD-DB).PLoS One. 2020 Oct 15;15(10):e0240402.
7. Nakagawa N, Sofue T, Kanda E, Nagasu H, Matsushita K, Nangaku M, Maruyama S, Wada T, Terada Y, Yamagata K, Narita I, Yanagita M, Sugiyama H, Shigematsu T, Ito T, Tamura K, Isaka Y, Okada H, Tsuruya K, Yokoyama H, Nakashima N, Kataoka H, Ohe K, Okada M, Kashihara N. J-CKD-DB: a nationwide multicentre electronic health record-based chronic kidney disease database in Japan. Sci Rep. 2020 Apr 30;10(1):7351.
8. Izuhara M, Izuhara HK, Tsuchie K, Araki T, Ito T, Sato K, Miura S, Otsuki K, Nagahama M, Hayashida M, Hashioka S, Wake R, Kimura T, Tsumoto S, Saito Y, Inagaki M., Real-World Preventive Effects of Suvorexant in Intensive Care Delirium: A Retrospective ,Cohort Study, J Clin Psychiatry 81, 20m13362,2020.
9. Shusaku Tsumoto, Tomohiro Kimura and Shoji Hirano, Automated Dual Clustering for Clinical Pathway Mining, IEEE Big Data 2020, 5387-5396, 2020.
10. Shusaku Tsumoto, Tomohiro Kimura and Shoji Hirano, Order Trajectory Analysis in Hospital Information System, IEEE Big Data 2020, 5397-5404, 2020.
11. Nakayama M, Takehana K, Kohro T, Matoba T, Tsutsui H, Nagai R. Standard Export Data Format for Extension Storage of Standardized Structured Medical Information Exchange. Circulation Reports.

2(10); 587 - 616. 2020.

12. 宮本 恵宏, 竹村 匡正, 竹上 未紗, 興梠 貴英, 中山 雅晴, 的場 哲哉, 小室 一成, 斎藤 能彦, 安田 聡, 宍戸 稔聡, 西村 邦宏, 平松 治彦, 上村 幸司, 辻田 賢一, 宇宿功市郎, 中村 文明. 電子カルテ情報をセマンティクス(意味・内容)の標準化により分析可能なデータに変換するための研究 . 医療情報学. 40(1) 32 - 33. 2020.

**2. 学会発表**

1. 真鍋 史朗、服部 睦、山口 純司、波内 良樹、島井 良重、坂井 亜紀子、山本 征司、武田 理宏、小西 正三、松村 泰志.臨床研究支援システムを用いた NCD 登録システムの開発、医療情報学 40(Suppl)2020.11. 18、425-430、日本医療情報学会
2. 松村 泰志、小川 俊夫、村木 功、山田 裕一郎、査 凌、藤井 歩美、村田 泰三、祖父江 友孝レセプトデータを用いた5大がんの Phenotyping の精度評価、医療情報学 40(Suppl)2020.11. 18、714-716、日本医療情報学会
3. 和田 聖哉、武田 理宏、真鍋 史朗、小西 正三、松村 泰志, 形態素解析が日本語医学BERT モデルに与える影響、医療情報学 40(Suppl)2020.11. 18、738-743、日本医療情報学会
4. 古川 大記,大山 慎太郎,近藤 康博,馬場 智尚,小倉 高志,長谷川 好規,白鳥 義宗. 深層学習を用いた間質性肺炎の高精度予後予測アルゴリズム,日本呼吸器学会誌(2186-5876)9 巻増刊 Page117(2020.08)
5. 大江和彦:ICT、ビッグデータを活用した循環器診療の次のステージ.日本循環器学会学術集会抄録集 84 回 Page シンポジウム 26-1(2020.07)

**H. 知的財産権の出願・登録状況**

**1. 特許取得**

特になし

**2. 実用新案登録**

特になし

**3. その他**

特になし