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】
目的と方法：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年目：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。２年目：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。
結果と考察：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、教師なし機械学習による自動クラスタリング等の手法で分析するための、臨床的表現型において共通特性をもつデータセットを、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)に分けて、作成方針の確立と自動作成環境の開発を行った。試験データの抽出では、1 病院分の検体検査件数で 800 万件以上、検体検査種別で 300 以上、１患者あたりの件数は多いもので 1700ー2000 件であった。この１施設分の分析用データセットを教師なし機械学習のクラスタリング手法 K-Means++により７つの疾患グループに、さらに疾患グループごとにその ICD10 の 4 桁目（細分類）を想定して８グループ程度を設定してクラスタリングを行った。K-Means++法以外にも Elbow 法等を用いたクラスタ数の決定を行いその効果を検討するとともに、クラスタごとの臨床的特性を取得する。

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

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、どのような別の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。

研究経緯：申請者が代表のAMED「医用知能情報システム基盤の研究開発」(2015.10〜2019.3)において統計解析可能な多施設臨床DB（120万症例以上）を構築済みで、倫理審査が完了しており申請者や分担者らが利用可能となっている。また同研究では津本(分担者)らが時系列データマイニングにより時間経過情報を含めて類型化する手法を開発してきた。

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年目：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。

２年目：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。

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、このデータベースを使用し、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ため、まず研究代表者の所属する１施設分のデータを用いたパイロット的なデータ分析を経て、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。すなわち、病名データで以下のICD10 コードの確定診断を有する 7 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。

その上で、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（約120項目を対象として探索的に決定）に含まれる血液検査結果定量値を抽出した上で、同一患者ごとに連続した６ヶ月ウインドウ期間における各検

査値の平均値、最小値、最大値、分散を変数値とし、その６ヶ月ウインドウを当該患者ごとの疾患存在期間内でずらして作成し異なる患者状態とみなした。次に同じ期間における投与医薬品のＡＴＣ分類コード（医薬品の国際的な効能成分分類）粒度のデータを抽出した。なお検査値については上記の定量値、医薬品については投与の有無のデータに変換した。参考までに付録１に検査結果抽出に関係する Python プログラムの主要部分を示す。

以上の分析用データセットの生成プログラムを Python 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が構築できた。本報告作成時点では、この手法による分析用データセットの作成は研究代表者の所属する１施設分で行った。今後、他の７施設のデータでも実施した上で、これらのデータを疾患 ICD10 ごとに統合する。

２）教師なし機械学習のクラスタリング

上記の１施設分の分析用データセットを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、前記全データについて血液検査結果だけで７つの疾患グループに、さらに疾患グループごとにその ICD10 の 4 桁目（細分類）を想定して８グループ程度を設定してクラスタリングを行った。初期値をいくつか変えて何度か実施して生成されるクラスタリングの結果と、疾患グループおよび ICD の細分類とを比較した場合の一致性がどの程度見られるかについてまず検討した。

Ｃ．結果と考察

研究代表者の病院分での 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)、各検体検査件数は約 800 万件であった。また患者別の検体検査実施件数は 1700ー2000 件あるものが見られた。

検体検査の項目数はまれに検査するものを含めると３００項目を超えるため、１）末梢血血液検査、血糖関係、凝固系、２）生化学、３）免疫系、４）ウイルスマーカ、５）血液ガス、などの区分に分け、区分ごとにデーセットを分割する必要があると考えられた。図 1 に、１）末梢血血液検査、血糖関係、凝固系の場合の項目セットを示す。

本研究は、多施設臨床 DB を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、共通特性をもつクラスタを多数自動生成し、自動的にその集団における未知の臨床特性を得ることによって、思いつかない CQ(Clinical Question)やエビデンスの生成の鍵を得ることができる点が特色である。

大規模臨床 DB から、気づかれていない臨床的表現型クラスタを自動識別し、臨床経過等の特性と確率情報を臨床エビデンスとして自動創成し、臨床現場で実際の患者にリアルタイムに近い適用する手法が開発される。これにより、たとえば標準臨床ガイドラインの策定において、１基礎疾患と２つ程度の合併疾患を有するような比較的シンプルなケースだけでなく、より多数の臨床的パラメータによる複雑な臨床的表現型の

クラスタに属するケース（いくつもの合併疾患を有し、いくつかの治療経歴を有する複雑な臨床経過をたどったケース群など）ごとに細分化した臨床ガイドラインを生成することができる可能性がある。集団の細分化をすることが、ガイドラインを適用すべき患者集団の規定を詳細化することに繋がり、個別化医療に近づくことになる。細分化された基準では、患者がどの集団に属するかの判定をコンピュータシステムに委ねなければ判定できない状況になることによりガイドライン準拠を推進する新たな ICT 手段を有することになり、標準的臨床ガイドラインの適用のあり方と普及推進に関する施策に貢献できる可能性がある。また、臨床中核拠点病院における臨床データベース駆動型の臨床研究の推進施策や、電子カルテデータの二次利用データの品質改善・管理に関する研究事業の推進にも貢献するとともに、さらに機械学習による厚労省標準 SS-MIX2 の新しい活用事例になると考えられる。

１年目の成果は、本報告作成時点では分析用データセットの作成方針の確立と自動作成環境の開発、およびその試行による教師なし機械学習のクラスタリングの探索的実施が、研究代表者のデータベースを使用して可能となったところまでである。

今後、得られたクラスタごとの結果分析を行い、その結果によっては、分析用データセットの作成方法の修正が必要と考えられれば修正を行うとともに、あらかじめクラスタ数を設定する K-Means++法以外のクラスタリング手法として、階層的クラスタリングも合わせて実施して結果を比較するなどを実施する必要がある。

２年目の計画としては、

１）クラスタの他の臨床情報特性の時系列変化を含めた類型化

各クラスタにおける他の臨床検査値の陽性率、疾患特異性の高い医薬品投与状況、臨床経過の類型細分化、重症度の時間的推移、既知の診断情報などを、記述統計や時系列データマイニング等の手法により、３年間程度の期間について解析し、クラスタごとの臨床的特性を取得する

２）クラスタにおける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

上記統計的特性とともに多次元ベクトル情報に変換し、上記臨床的特性と、それに対応する疾患・病態を検討し、論文や診断治療基準での既知臨床エビデンスと比較し、臨床医とディスカッションを行い、今回自動的に得られた臨床エビデンスの特徴や課題を明らかにする。

３）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システムの開発とその臨床的有用性の評価を行う。

といった手順を検討する。

Ｅ．結論

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、教師なし機械学習による自動クラスタリング等の手法で分析するための、臨床的表現型において共通特性をもつデータセットを、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)に分けて、作成方針の確立と自動作成環境の開発を行った。試験データの抽出では、1病院分の検体検査件数で 800 万件以上、検体検査種別で 300 以上、１患者あたりの件数は多いもので1700ー2000 件であった。この１施設分の分析用データセットを教師なし機械学習のクラスタリング手法 K-Means++により７つの疾患グループに、さらに疾患グループごとにそのICD10の４桁目(細分類)を想定して８グループ程度を設定してクラスタリングを行った。K-Means++法以外にもElbow 法等を用いたクラスタ数の決定を行いその効果を検討するとともに、クラスタごとの臨床的特性を取得する。

Ｆ．健康危険情報

Ｇ．研究発表

１．論文発表

1. K Yamada, M Itoh, Y Fujimura, M Kimura, K Murata, N Nakashima, M Nakayama, K Ohe, T Orii, E Sueoka, T Suzuki, H Yokoi, C Ishiguro, Y Uyama on behalf of MID‐NET project group: The utilization and challenges of Japan’s MID‐NET® medical information database network in postmarketing drug safety assessments: A summary of pilot pharmacoepidemiological studies. Pharmacoepidemiology and Drug Safety 28(5),601-608, May.
2. Hayakawa M, Imai T, Kawazoe Y, Kozaki K, Ohe K. Auto-Generated Physiological Chain Data for an Ontological Framework for Pharmacology and Mechanism of Action to Determine Suspected Drugs in Cases of Dysuria. Drug Safety. 2019,42
3. Kagawa R, Shinohara E, Imai T, Kawazoe Y, Ohe K. Bias of Inaccurate Disease Mentions in Electronic Health Record-based Phenotyping. International journal of medical informatics. 2019;124: 90-96.
4. Nakashima N, Noda M, Ueki K, Koga T, Hayashi M, Yamazaki K, Nakagami T, Ohara M, Gochi A, Matsumura Y, Kimura M, Ohe K, Kang D, Toya Y, Yamagata K, Yokote K, Ikeda S, Mitsutake N, Yamamoto R, Tanizawa Y.: Recommended configuration for personal health records by standardized data item sets for diabetes mellitus and associated chronic diseases: A report from Collaborative Initiative by six Japanese Associations.J Diabetes Investig. 2019 May;10(3):868-875.
5. Seki T, Tamura T, Suzuki M, SOS-KANTO 2012 Study Group. Outcome prediction of out-of-hospital cardiac arrest with presumed cardiac aetiology using an advanced machine learning technique. Resuscitation 141 128-135 2019
6. 大江 和彦 ビッグデータと人工知能技術による診療支援システム Dementia Japan(1342-646X)34 巻 1 号 Page70-75(2020.01) ,解説

7. 大江 和彦　AIとビッグデータのための医療情報の標準化 医療機器学 (1882-4978)89 巻 6 号 Page545-551(2019.12)

２．学会発表

1. 早川 仁, 関 倫久, 河添 悦昌, 大江 和彦 パスウェイデータベースを利用したグラフ畳み込み深層学習による悪性腫瘍の診断分類性能の検討 医療情報学連合大会論文集(1347-8508)39 回 Page352(2019.11)
2. 関 倫久, 河添 悦昌, 大江 和彦 SS-MIX2標準化ストレージを用いた入院後の死亡退院リスク予測モデルの開発 医療情報学連合大会論文集(1347-8508) 第39回 Page249(2019.11), 国内, 口頭

H．知的財産権の出願・登録状況

なし

表１　末梢血血液検査、血糖関係、凝固系の項目セット**の例**

JLAC10	項目名	JLAC10	項目名
2B140000002206201	D. Dダイマー_定量値__血漿	2A160000003460461	血液像_前骨髄球__血液塗抹標本
3D046000001920402	ＨｂＡ１ｃ（ＮＧＳＰ）__全血(添加物入り)	2A160000003460456	血液像_単球__血液塗抹標本
3D010000002226101	グルコース_定量値__血漿	2A160000001966256	血液像_単球__全血(添加物入り)
4Z271000002202301	ヒト脳性Na利尿ポリペプチド_定量値__血漿	2A160000001930957	血液像－リンパ球__全血(添加物入り)
2B100000002231101	フィブリノーゲン_定量値__血漿	2A160135201930957	血液像－リンパ球__全血(添加物入り)
2B030000002231157	プロトロンビン時間_INR値__血漿	2A160000001930955	血液像－好塩基球__全血(添加物入り)
2B030000002231153	プロトロンビン時間_PT活性(%)__血漿	2A160135201930955	血液像－好塩基球__全血(添加物入り)
2B030000002231155	プロトロンビン時間_PT比__血漿	2A160000001930954	血液像－好酸球__全血(添加物入り)
2B030000002231151	プロトロンビン時間_被験血漿PT時間(秒)__血漿	2A160135201930954	血液像－好酸球__全血(添加物入り)
2B020000002231100	活性化部分トロンボプラスチン時間___血漿	2A160000001930951	血液像－好中球__全血(添加物入り)
2A160000003460477	血液像_その他__血液塗抹標本	2A160135201930951	血液像－好中球__全血(添加物入り)
2A160000003460457	血液像_リンパ球__血液塗抹標本	2A160000001930956	血液像－単球__全血(添加物入り)
2A160000001966257	血液像_リンパ球__全血(添加物入り)	2A160135201930956	血液像－単球__全血(添加物入り)
2A160000003460458	血液像_異型リンパ球__血液塗抹標本	2A160000001930106	血液像－破砕赤血球%__全血(添加物入り)
2A160000003460460	血液像_後骨髄球__血液塗抹標本	2Z010000001792001	血沈_定量値__血液(含むその他)
2A160000003460455	血液像_好塩基球__血液塗抹標本	2A990000001992057	末梢血液一般検査_ＭＣＨ__全血(添加物入り)
2A160000001966255	血液像_好塩基球__全血(添加物入り)	2A990000001992058	末梢血液一般検査_ＭＣＨＣ__全血(添加物入り)
2A160000003460454	血液像_好酸球__血液塗抹標本	2A990000001992056	末梢血液一般検査_ＭＣＶ__全血(添加物入り)
2A160000001966254	血液像_好酸球__全血(添加物入り)	2A990000001992054	末梢血液一般検査_ヘマトクリット__全血(添加物入り)
2A160000003460451	血液像_好中球__血液塗抹標本	2A990000001992053	末梢血液一般検査_ヘモグロビン__全血(添加物入り)
2A160000001966251	血液像_好中球__全血(添加物入り)	2A990000001992055	末梢血液一般検査_血小板数__全血(添加物入り)
2A160000003460453	血液像_好中球分葉核__血液塗抹標本	2A990000001992051	末梢血液一般検査_赤血球数__全血(添加物入り)
2A160000003460452	血液像_好中球桿状核__血液塗抹標本	2A990000001992052	末梢血液一般検査_白血球数__全血(添加物入り)
2A160000003460462	血液像_骨髄芽球__血液塗抹標本	2A110000001966202	網赤血球数_構成比__全血(添加物入り)
2A160000003460459	血液像_骨髄球__血液塗抹標本		