別紙３

令和３年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

大規模データの利活用研究の加速のための研究

研究代表者　木村　映善（国立大学法人愛媛大学 大学院医学系研究科　教授）

研究要旨
健康医療ビッグデータをこれまでの統計解析手法に加えてAIを適用させることにより、これまでに省みられることのなかった事象に光をあて、医療の質の向上・均てん化・診療支援と医療分野のイノベーションに貢献することが期待されている。このため、要配慮個人情報を収集し、匿名加工された医療情報を円滑に利活用する社会的仕組みとして、医療分野の研究開発に資するための次世代医療基盤法が施行された。この次世代医療基盤法の認定事業者から匿名加工医療情報を提供頂いてICT・AI技術を利用した研究や開発が進展することが期待されている。
しかし、認定事業者は稼働したばかりであり、医学研究に用いられている事例はまだない。また、機械学習手法は多数の変数を要求する傾向がありながら、匿名加工による変数削減とリスクのトレードオフの関係下に変数の数に制約を課せられる可能性があることから、匿名加工医療情報を用いた機械学習の研究について懐疑的な見方もでている。そこで、本研究では実際の認定事業者へのデータ提供依頼に始まり、研究者の研究体制への審査・監査・匿名加工医療情報の解析までを通した密着取材的なプロセス分析を通した有用性等を検証することにより、匿名加工医療情報がどのような研究に資するのか、またAI技術を用いた研究に関する技術的課題を明らかにし、認定事業者を利用した研究を加速する施策の提言につなげることを目的とする。
初年度は2つの認定事業者と契約し、多医療機関の電子カルテに対するシーケンス解析に関する研究、臨床現場で必要とされる説情報や因果関係等の説明を行う説明可能なAIの研究、安全なAIのプロトタイピングを提供する統計的特徴を維持した合成データ生成技術の開発、匿名加工医療情報に関する安全性と人工知能研究応用性の評価の4種類に着手した。また、大量のデータを安全に扱うために、認定事業者内にて安全に機械学習をする環境の構築と運用に関する申し合わせを行った。また、認定事業者の潜在的な利用者について認定事業者の啓発を兼ねた認定事業者を利用した研究に関する意識調査のアンケートの項目について検討した。予定されていた3つ目の認定事業者の認定が遅れていたため、次世代医療基盤法第25条下にかかるデータ融通の検討に切り換え、来年度で契約と事務局にかかる検討を実施することとした。来年度にかけて各研究活動を継続し、各研究活動の成果、アンケート結果、海外の認定事業者相当のデータ環境の調査、次世代医療基盤法第25条下にかかるデータ融通の検討を行い、認定事業者にかかる厚生労働省行政の施策提言にむけたとりまとめを行う予定である。

研究分担者氏名・所属研究機関名及び所属研究機関における職名

荒木　賢二　　国立大学法人宮崎大学・医学部附属病院・研究員
黒田　知宏　　国立大学法人京都大学・大学院医学研究科・教授
水島　洋　　国立保健医療科学院・研究情報支援研究センター・主任研究官
星　佳芳　　国立保健医療科学院・研究情報支援研究センター・センター長
渋谷　哲朗　　国立大学法人東京大学・医科学研究所・教授
佐々木　香織　北海道公立大学法人札幌医科大学・医療人育成センター・教授
伊藤　伸介　　中央大学・経済学部・教授
長島　公之　　一般財団法人 日本医師会医療情報管理機構・理事

A. 研究目的

健康医療ビッグデータをこれまでの統計解析手法に加えて AI を適用させることにより、これまでに省みられることのなかった事象に光をあて、医療の質の向上・均てん化・診療支援と医療分野のイノベーションに貢献することが期待されている。このため、要配慮個人情報を収集し、匿名加工された医療情報を円滑に利活用する社会的仕組みとして、医療分野の研究開発に資するための匿名加工医療情報に関する法律（以下、次世代医療基盤法）が平成 29 年に公布、平成 30 年に施行された。令和元年から次世代医療基盤法に基づく複数の認定匿名加工医療情報作成事業者（以下、認定事業者）が認定されている。この認定事業者から匿名加工医療情報を提供頂いて ICT・AI 技術を利用した研究や開発が進展することが期待されている。

しかし、認定事業者は稼働したばかりであり、AI を活用した医学研究に匿名加工医療情報が用いられている事例はまだない。また、機械学習手法は多数の変数を要求する傾向がありながら、匿名加工による変数削減とリスクのトレードオフの関係下に変数の数に制約を課せられる可能性があることから、匿名加工医療情報を用いた機械学習の研究について懐疑的な見方もでている。

そこで、本研究では実際の認定事業者へのデータ提供依頼に始まり、研究者の研究体制への審査・監査・匿名加工医療情報の解析までを通した密着取材的なプロセス分析を通した有用性等を検証することにより、匿名加工医療情報がどのような研究に資するのか、また AI 技術を用いた研究に関する技術的課題を明らかにし、認定事業者を利用した研究を加速する施策の提言につなげることを目的とする。

B. 研究方法

本研究班は、当初の研究計画において、第一号・二号業者において(1)匿名加工医療情報を活用した AI 研究、(2)匿名加工医療情報で機械学習する環境にかかる検討、(3)認定事業者にかかる意識調査、そして第三号業者認定後に(1)の研究の追加、あるいは第一号事業者における医用画像を活用した AI 研究を想定していた。

本研究課題の交付申請時点(令和 3 年 4 月)で、次世代医療基盤法にもとづいた認定事業者は 2 社存在し、現在 1 社が受審中（註:令和 4 年 4 月に認定された）であった。令和元年 12 月に第一号が、令和 2 年 6 月に第二号が認定され、データ収集を開始している。第一号において十分なデータ収集実績を有するが、医用画像収集の事業は準備中であった。第二号においては、認定時から 1 年経過しておらず、医療情報を蓄積している状態であった。

本研究課題において可及的に国内全ての認定事業者の評価を実施することを求められていたので、当時受審していた第三号の候補事業者が認定されれば、第三号事業者において AI を利用した研究を追加すること、それが叶わなければ、第一号事業者で医用画像収集事業が始まり、医用画像が蓄積されていれば医用画像を利用した研究にすることとし、二段構えの研究計画を策定していた。令和 4 年 2 月時点で、第三号事業者の認定がなされていないこと、第一号業者において医用画像の収集事業の承認がなされていない状態であったことから、当初の研究計画の遂行は困難であると判断し、次世代医療基盤法第 25 条に基づくデータ融通に向けた検討をする研究に切り替えた。

以下に、以上の経緯を経た上での本研究班における各研究について記述する。

(1)匿名加工医療情報を活用した AI 研究

次世代医療基盤法の認定事業者の匿名加工医療情報を活用した AI に係る研究のために、第 1 号業者一般社団法人ライフデータイニシアティブ（LDI)、第 2 号業者一般財団法人日本医師会医療情報管理機構(J-MIMO)へ利用申請を行い、各々において利用の承認を得た。

LDI においては、「認定匿名加工医療情報作成事業者が保有する医療情報を活用した、匿名加工医療情報の作成に依らない AI 研究の実現可能性の検討」の名目において、「多医療機関の電子カルテに対するシーケンス解析に関する研究」（宮崎大学・ログビー)、「臨床現場で必要とされる説情報や因果関係等の説明を行う説明可能な AI の研究」（宮崎大学・東京工業大学)、「安全な AI のプロトタイピングを提供する統計的特徴を維持した合成データ生成技術の開発」（愛媛大学・NTT 社会研）の研究に着手した。

J-MIMO においては、「匿名加工医療情報に関する安全性と人工知能研究応用性の評価」の名目において、J-MIMO が保有する SS-MIX 標準ストレージベースの医療情報を匿名化した匿名加工医療情報において、差分プライバシーを考慮した匿名化技法の当該医療情報への適用可能性の検討および当該匿名加工医療情報の人工知能解析への応用可能性とさらなる安全性、利便性の向上の可能性の検討に着手した。

(2)匿名加工医療情報で機械学習する環境にかかる検討

AI にかかる研究で主流的に使われている深層学習は、多量・多種のデータを探索的にアクセスし膨大な計算を実施する。この多量・多種のデータを要求する背景に機械学習と特徴量の関係がある。

従来は分析対象データの中で予測に寄与する変数（特徴量）について、人間がこれまでの知見をもとにアプリオリに決定、調整していた。しかし、深層学習は自ら最適な特徴量を選択できる能力を獲得した。そのため、深層学習を使用する場合は、事前に使用する変数を決めずに幅広い変数、そして大量のデータを投与し、最終的にモデルに寄与する変数を自律的に決定させるというアプローチを取る。そして、このアプローチは匿名加工と相性が悪いのである。匿名化の安全性を高めるために、通常は使用する変数を必要最低限なものに抑えることを求めるからである。変数の種類が増えるほど匿名加工が困難になり、安全に倒れて匿名加工処理するとデータの質が落ちるというトレードオフがある。また、匿名加工の加工基準は一定ではなく、提供先のリスクを踏まえて検討するものである。つまり、利活用者に提供する場合は、利活用者の情報セキュリティ体制の状況も匿名加工の加工基準を検討する一要素となる。従って、AI にかかる研究に匿名加工医療情報を提供する際には、多くの変数が要求されること、利活用者の情報セキュリティ環境を考慮して、匿名加工の加工基準が高いものに設定され、匿名加工医療情報のデータの質に制約が加えられる可能性がある。そのため、外部への提供ではなく、認定事業者の安全な環境下のみに利用を限定することで匿名加工の加工基準を引き下げ、AI の研究に必要な多種多様な変数が含まれる状態を担保することが期待される。

そのため、LDI から提供される匿名加工医療情報を活用した3種類の研究を認定事業者内にて安全に実施する方法論を検討することとした。

(3)認定事業者にかかる意識調査

評価チームから各認定事業者へデータ保有状況、保有データ種に関する状況についてヒアリングを実施し、認定事業者を活用する AI 研究・制度環境に関するアンケート案を班会議にて策定する。現時点で、認定事業者を利用した AI 研究を志している研究者の数が極めて限定的であり、事前に対象者を選定することは困難であると考えられるため、認定事業者を利用した研究手法の啓発も兼ねて広く募る方法を検討する。

(4)次世代医療基盤法第 25 条下にかかるデータ融通の検討

認定事業者は次世代医療基盤法第 25 条において他の認定事業者に医療情報の提供を認められているが、そのデータの名寄せ・融通の実績はなく、その運用管理の規程も未整備である。そのため、研究班と認定事業者において、データ融通に関する実証実験の実施にむけて、データ融通にかかる規程の整備案、制度設計への提言、データ突合にかかる技術的課題を洗い出すためのワークショップを行う。

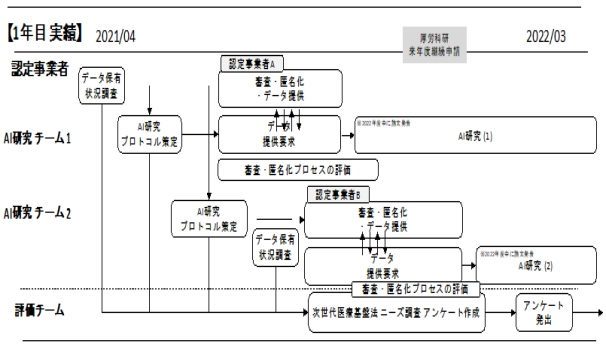

図 1. 本年度における研究活動の概要

C. 研究結果

現在、認定事業者を利用した研究は 4 テーマで行われており、それぞれの研究は終了後に研究論文として発表予定である。また、次世代医療基盤法と認定事業者の運用にかかる課題の検討、アンケート結果については、整理ののち、学会でのシンポジウムや最終レポートを通して提言する予定である。また、認定事業者の連絡協議会とも連動し、認定事業者の運用管理についてのガイドラインや契約方法等の策定にも寄与していく。

(1)匿名加工医療情報を活用した研究

第 1 号業者においては、3 つのテーマが設定され、後述する機械学習環境下において匿名加工医療情報を機械学習にかける研究を実施している。第2号事業者においては差分プライバシーとデータの有用性に関する検討をしている。

(a) 第 1 号業者

(a-1) 多医療機関の電子カルテに対するシーケンス解析に関する研究

これまで電子カルテ中の医療オーダのシーケンス解析を行い、診療オプションとなるバリアントの安全性、効率性の評価や分岐理由の推定の手法を開発して来た。また、検体検査項目をクラスタリングして検査タイプを抽出し、検査タイプと検査結果のシーケンスから次の検査項目を推薦する手法等を開発して来た。これらの研究の知見を活かし、ライフデータイニシアティブが収集した、異なる医療機関における医療オーダや検体検査のシーケンスの違いを解析する手法、検体検査結果と医療オーダとの関係をより詳細に解析する手法等の研究を行う。全体の流れとして、1、検査のデータを抽出、2、同時出現検査のグルーピング、3.検査データの分類、4.検査シーケンスのマイニング、5.検査項目組み合わせの提案の 5 段階のプロセスで行う。

検査結果と検査オーダのデータは第1号業者から提供された匿名加工医療情報から変換された患者 ID とオーダ No、検査日時ごとに対応する検査項目のレコードを抽出する。

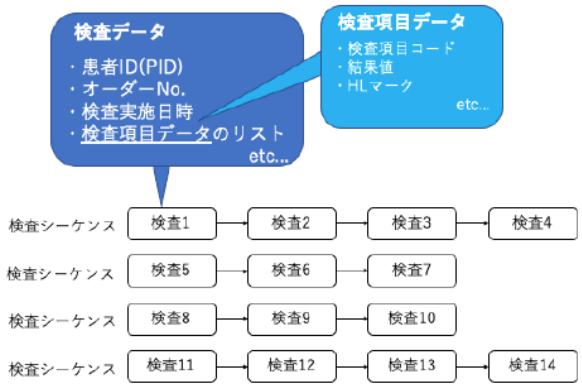

図2.抽出する検査データのイメージ

抽出したレコードから、患者 ID、オーダグループ、検査日時ごとに検査項目をグループし、同時に出現した検査項目のグルーピングを行う(図3)。グルーピングされた検査オーダ項目は各検査タイプとして類型化する(図 4)。

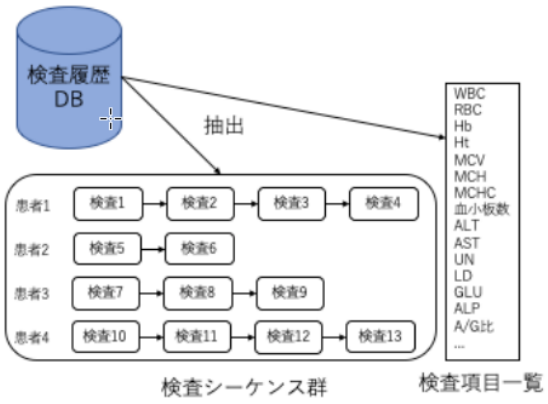

図 3.検査データの抽出のイメージ

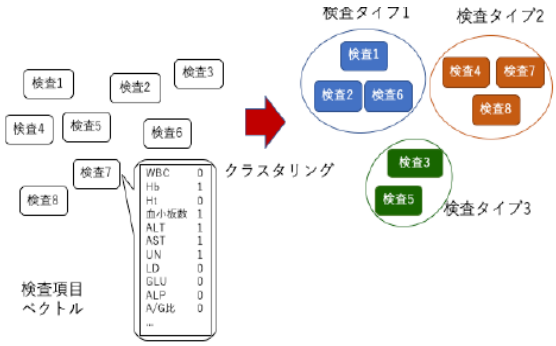

図 4.検査データの分類のイメージ

検査結果と検査グループを時系列に整理し、シーケンス分析のマイニングを行う。これにより、検査結果と検査オーダ、検査オーダによる検査結果の変化の関係性についてモデルの学習がなされる。

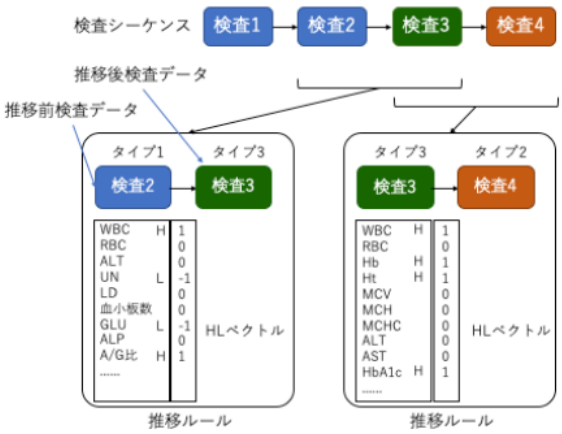

図5.検査シーケンスのマイニングのイメージ

学習したモデルに検査結果を入力し、オーダする検査項目を推薦させる。医師の判断と照合して入力検査タイプの推薦検査項目の適合率、再現率を評価する。

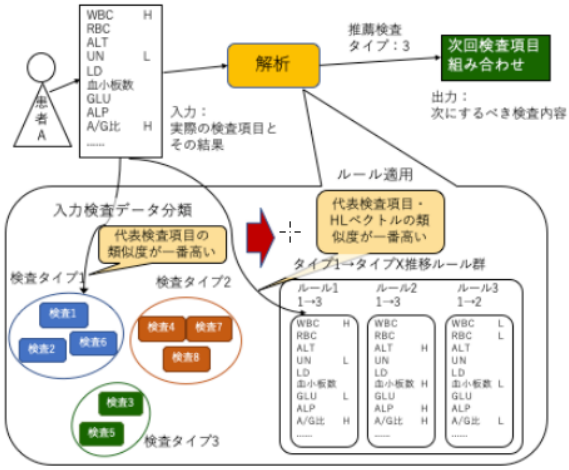

図 6.検査項目組み合わせの推薦のイメージ

(a-2) 臨床現場で必要とされる説情報や因果関係等の説明を行う説明可能な AI の研究

これまで説明可能な AI（XAI）を用いて入院後合併症の発症予測と特徴量の寄与度の可視化等の研究を行ってきた。XAI の研究において、医師等の視点から臨床現場で必要とされる説明情報について定性ヒアリングを行った事例や、因果関係の説明まで行なっている事例は少ない。これまでの研究を発展させ、臨床現場で必要とされる情報や因果関係等の説明を行う XAI の研究を行う。以下にアウトプットイメージを示す。以下の図は研究班が過去に発表した「複数の入院後合併症に対する時系列予測モデルの開発と説明可能な AI を用いたリスク要因の比較」で公表した図である。術後の敗血症に関して、出現回数が多い特徴量と転帰への影響度を図示したものである。このような図を各疾患、想定シナリオごとに、さらに医療機関ごとに比較することで、全般的な疾患への寄与因子の特定と、医療機関によるバイアスや治療方針の違いなどを総合的に比較検討できるようなモデルの開発を検討している。

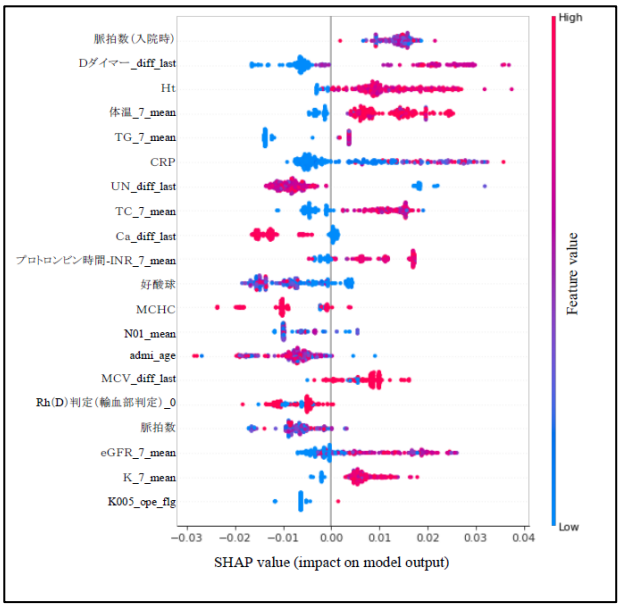

図7.敗血症の予測に重要な特徴量（手術）

(a-3) 安全な AI のプロトタイピングを提供する統計的特徴を維持した合成データ生成技術の開発

AI に関する取り組みにおいて、様々な障害があるがその上位に位置づけられているのがデータアクセスの困難性である。プライバシー保護のために、個人の特定可能性をなくすための匿名加工の加工基準を巡っての協議やデータ提供先への厳重な情報セキュリティ体制の要求など、迅速に研究を進めるにあたっての障壁となっている。認定事業者は丁寧なオプトアウトを前提として要配慮個人情報の収集の敷居を下げ、様々なデータソースの突合を可能にする制度である。しかし出口は匿名加工医療情報の提供であり、匿名加工に関わる課題は依然として残る。

そこで、安全な AI のプロトタイピングを提供するためのツールとして、統計量を維持した合成データに期待が持たれている。これまでの研究を発展させ、臨床研究の検討と加速に貢献する合成データの研究を行う。上記 2 研究のデータを事例として合成データを生成し、オリジナルデータとの統計的差異や品質について評価する。匿名加工医療情報のデータを実際利用する前の段階として、愛媛大学の DPC データを使用し、ベイジアンネットワーク(BN)、深層学習の GAN ベースの CTGAN、岡田らによる統計量ベース(STAT)によるデータ合成手法の比較を行い、合成データの品質評価として、Hellinger 距離、Kolmogorov-Smirnov 距離、相関係数を用いて評価した。中間結果として、差分プライバシー(DP: Differential Privacy)を適用した場合に合成データの誤差が拡大する傾向が確認され、元データの分布状況に誤差の発生状況が左右されることが確認された。結果として差分プライバシー非適用の場合は相対的に良い結果が得られた一方で，差分プライバシーを適用した場合は各手法のデータの品質が大きく落ちることが明らかになった。

差分プライバシーにおいて品質が大きく落ちる原因としては、対策のために繰り返しノイズを付与する手法を付加するために、ノイズが大きくなることが確認された。また、本検証段階では深層学習ベースよりも機械学習ベースの有用性が高い場合が多く見られたこともあり、他の深層学習ベースでの合成データ手法の研究開発状況を調査しつつ、より有用性と安全性を高いレベルで両立させる手法について引き続き探索することとした。なお、本研究については前述したとおり匿名加工医療情報の活用の前段階としてのプロトタイピングの為に愛媛大学の DPC データを利用したため、「人を対象とする医学研究に関する倫理指針」に基づき，愛媛大学医学部の倫理審査委員会の承認を得て実施した．研究課題名「統計的特徴を維持した合成データ生成手法の品質評価」（承認番号 2012001）来年度では匿名加工医療情報をベースに検証を進める予定である。

(b) 第 2 号業者

検討の過程において、差分プライバシーの分析は個人の識別特定にかかる活動に相当しないという整理がなされた。第 2 号事業者の主なデータソースは全て SS-MIX 標準化ストレージであり、SS-MIX 内に格納されている HL7 2.x メッセージの各項目をレビューし、分析対象となるデータ項目及び匿名加工医療情報の匿名加工の基準に関する方針について決定した。東京大学への匿名加工医療情報を提供頂いて、年度末から研究分担者による分析が開始されている。

(2)匿名加工医療情報で機械学習する環境にかかる検討

現行の認定事業者の運用は類型的には医療情報を保管するデータセンター(図 8 左)とデータセンターから匿名加工医療情報を抽出する接続用端末が設置されているデータセンター（図 8 右）の二拠点（エリア）に分かれており、通常は医療情報の抽出、匿名加工処理、そして匿名加工医療情報を利活用者にセキュリティ便でお渡しするまでをトレーサビリティを確保した状態で運用がなされる。

次世代医療基盤法及びガイドラインを検討した結果、データセンター内において医療情報を研究者の研究計画を踏まえた匿名加工処理を施し、認定事業者の倫理委員会にて承認された匿名加工情報のみをセキュリティルーム外に出す取扱いしか認められないことを確認し、データセンター内でのセキュリティルーム（図 8 右）内で匿名加工医療情報を安全に取り扱う運用について検討した。

具体的には、Graphics Processing Unit(GPU)という深層学習を高速に処理する演算装置を搭載したコンピュータをセキュリティルーム内に設置する。

そして、そのコンピュータ内には Docker という仮想環境をホストする仮想基盤（以下、実行環境）を整備する。この Docker は「コンテナ」という単位で仮想的に複数のコンピュータ環境が稼働するのを支援する仕組みである。
なお、このコンピュータはセキュリティルーム内に設置し、物理的保護対策を施した上でいかなるネットワークにも接続されない状態で管理される。
利用者は機械学習のためのプログラムをコンテナ上に導入・構築し、認定事業者にコンテナのイメージファイルを提出する。認定事業者はコンテナイメージをセキュリティルーム内の Docker 環境にホストし、接続用端末から HDD 経由で Docker 環境に研究に必要な医療情報をコピーし、Docker 環境を稼働させて機械学習や分析を実行させる。実行した結果として、ログ、機械学習モデル、統計解析結果など個人特定可能性がないデータが生成されるが、セキュリティ上の万全を期して、通常の匿名加工医療情報の提供プロセスと同じ手順を踏んだ上で研究者に提供される。研究者は機械学習の分析を実行して生成された統計情報の結果から機械学習の成否を判断し、必要に応じてコンテナ内のプログラムを改修する。
研究中は、このプロセスを数回繰り返すことになるが、認定事業者に外部から入るのは研究者より提出されたコンテナのイメージファイル、外部に出されるのはコンテナ内プログラムで処理された結果、そしてそれらの一連の動きはエアーギャップ下に行われる。なお、実行環境にロードしたコンテナイメージやデータは研究者には一切返還されず、研究者から実験結果の提供をうけて改修したコンテナイメージを受け取り再度実行環境に読み込むという手順をとる。また、実行環境のデータやプログラムが保存される内部・外部ストレージは研究終了後に物理的に破壊することで情報が漏洩する経路を形成しないように細心の注意を払って運用管理を設計した。
また、これら一連の検討に派生して、匿名加工医療情報を用いて学習した機械学習モデルの扱いや匿名加工医療情報の要件についての確認がなされた。

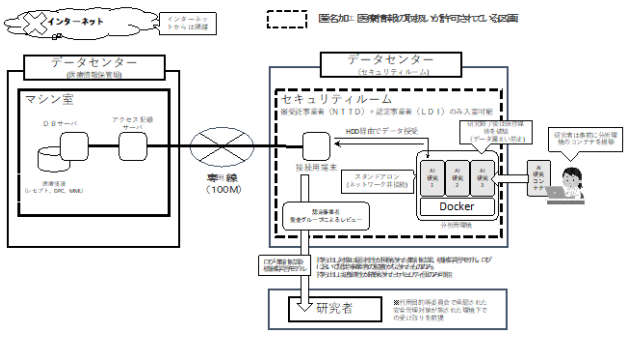

図 8. AI にかかる研究のための認定事業者の環境整備

(3)認定事業者にかかる意識調査
初年度において各認定事業者へデータ保有状況、保有データ種に関する状況ヒアリングを実施したことを踏まえ、認定事業者を活用する AI 研究・制度環境に関するアンケート案について研究班で数回にわたり検討した（アンケート案は別添）。アンケート実施の対象者について日本医学会に関連する学会に所属する研究者、民間企業、地方自治体の保健福祉部を対象とした。現時点で、認定事業者を利用した AI 研究を志している研究者の数が極めて限定的であり、事前に対象者を選定することは困難であると考えられた。また、研究班及びオブザーバーより、認定事業者や次世代医療基盤法の認知度が必ずしも高くないことを踏まえ、アンケートの設問への解答を通して、次世代医療基盤法や認定事業者について理解が深まるような啓発も兼ねてアンケートの回答中に認定事業者や制度に概要を理解していくステップを組みこむことを検討した。
最終年度において最終的なアンケート案のブラッシュアップ後、Web アンケートサイトを構築し、学会事務局や地方自治体にアンケート依頼を発出し、回答を収集・分析する予定である。

(4)次世代医療基盤法第 25 条下にかかるデータ融通の検討
第一号業者(LDI)、第二号業者（J-MIMO）と愛媛大学において 3 回対談を実施し、次世代医療基盤法第 25 条下にかかるデータ融通の検討範囲について協議・検討した。将来的に認定事業者同士によるデータの融通・突合の行程を実施できるようになることを最終目標と設定し、来年度の研究範囲について設定した（図 8）。
認定事業者間のデータの授受にあたって、データの授受・名寄せの技術的な検討と、データ授受に関わる法制度、契約、運用管理規程の整備が必要である。
そこで、検討する WG を契約担当、事務局担当、技術担当の 3 つの WG を構成して本プロジェクトに取り組む方向性で整理がなされた。それぞれの WG にかかる問題意識と想定しているタスクは以下の通りである。

(a)契約に関する検討
事業者間におけるデータ融通の実績がないこと、また認定事業者は今後も増加する可能性があるため、当初から三者以上の契約を想定して契約書案を検討する必要がある。現時点では、最終的な成果物としての匿名加工医療情報の提供をうける利活用者と代表的な窓口となる認定事業者との二者間契約の背景に、匿名加工医療情報の提供に必要な医療情報を名寄せ・融通しあう複数の認定事業者間で都度契約することを想定している。事業者間の契約形態、契約形式、契約条文についてのプロトタイプ的な検討を行う。

(b)事務局に関する検討

現在は、利活用者への提供にあたっての判断は認定事業者内の利用目的等審査委員会が行っている。しかし、相互にデータを融通、名寄せするにあたっては、実名データを他の認定事業者に提供し、かつ他の認定事業者での匿名加工処理を行うため、リスク基準、評価プロセスが異なるとデータの融通が困難になる。相互の安全管理基準の確認、審査プロセスの共通化、あるいは審査基準の統一、そしてそれらを受けて審査規程の改正や審査委員への教育といった派生的な業務・見直しが発生する可能性がある。認定事業者の独自性と共通化すべき部分の境界の設定についても検討が必要である。

(c)技術に関する検討

現在、LDI は MML（Medical Markup Language)、J-MIMO は HL7 2.x メッセージを中心とした医療情報交換規約で医療情報を収集している。また、認定事業者が設立されてから、匿名加工医療情報提供のケースを蓄積しはじめている。しかし、現状は医療情報の提供はデータの内容を事前バリデーションするプロファイルによる検証を伴わずに行われているため、二次利用性は必ずしも担保されていない。例えば、項目は標準化されていても、独自の項目の使い方、インハウス・コード（標準的なコードではなく、当該組織固有のコード）の使用、本来的な使い方から逸脱しているもの（検査結果の付帯情報に個人名を記載している等）があり、データ内容の確認やクレンジングに時間を取られている状況である。その中で、名寄せ・融通についての技術的課題は、大きく分けて(1)どの情報を名寄せの為のキー情報とするか、(2)名寄せの精度・信頼性をどのように担保するか、(3)融通する際のデータ形式はどうするか。相互運用性のある標準規格に変換して提供するのか、あるいは最終的に提供される匿名加工医療情報の形態にあわせて提供するか。(4)最終的な匿名加工医療情報にむけて匿名加工の基準をどのように設定するか、等、出口の成果も見据えた検討を行わなければならない。

これらの技術にかかる検討の実証実験にまで踏み込むことは、本研究班の当初の研究活動の範囲を超えるため、来年度においては(a)契約に関する検討と(b)事務局に関する検討までを行う。(c)技術的検討については、システム・運用面・データ形式に関する課題抽出と整理に止める。そして、本研究班の最終レポートにおいて、認定事業者間データ授受に関する提言や制度設計に向けた提案の形でまとめていくことを合意した。

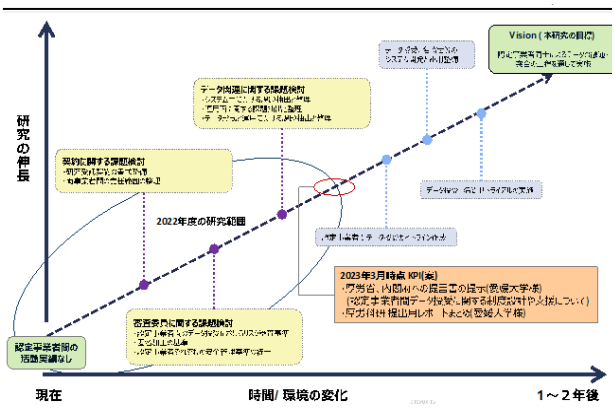

図 9.次世代医療基盤法第 25 条下のデータ融通・突合にかかる研究の全体像
(ICI 社による議事録より抜粋)

D. 考察

(1)認定事業者を活用した研究のありかた

認定事業者 2 社の協力を得て 4 つの AI に掛かる研究が進行中であり、現時点では研究内容そのものについての考察は不可能ではあるが、認定事業者を利用する過程についての考察を述べる。木村の個人的な所感としては、現時点ではアカデミアが自前の学術研究プロジェクトではなく、認定事業者を利用することのメリットは短期的視点では正当化しづらいと思われる。一方、長期的視点における利点を追求し、その実現をサポートするような厚生労働省行政の施策の提言をすべきであると考える。この長期的視点の利点にかかる検討については来年度実施するアンケートや海外事例調査を踏まえて裏付けのある形で検討したい。

まず、短期的視点では、現在認定事業者で収集している医療情報における多様性、データ量において、学会主導で研究目的を明確に設定して収集している大規模事業と比較して圧倒的な差別化を図りにくい時期であると思われる。この問題については、認定事業者が事業を継続してデータを蓄積し、またクライアントとして自治体をはじめとした様々なステークホルダーへ拡大することによって発展的に解消することが期待される。

また、臨床研究において、倫理委員会の審査手続やデータの管理、同意の取得等の研究にかかる事務作業の負担は、認定事業者の利用や次世代医療基盤法下の匿名加工医療情報提供にかかる利用目的等委員会の審議に代えることによって省力化されることが期待されている。しかしながら、認定事業者においても利用目的の審査の厳密性は倫理委員会の基準に互しており、審査にかかる事務負担については大きな変化はない、というのが実感である。一方、データの管理や患者からの同意の取得の業務から解放されることは非常に意義が大きいと思われる。

また、今回の研究事業においては、データのコストについて認定事業者に多大なるご協力を頂いた上で検討している。

自ら研究プロジェクトを立ち上げてデータ収集をするよりは遙かに低予算で実施することができるのではあるが、それでも匿名加工医療情報の提供に掛かる金額は決して小さなものではなく、文部科研費でいうところの基盤研究(A)から(S)にかけてのトップクラス級の研究予算を獲得できなければ利用が困難である。そういう意味において、潤沢な研究資金を有する民間企業にとっては、コスト的に優れた選択肢となるであろう。しかし、一般的な研究者や地方自治体にとっては、認定事業者を利用するコストを研究資金として外部から調達するプロセスまで含めて考えると未だ利用の敷居が高いと思われる。

また、本研究を実施して改めて確認されたことであるが、研究予算が年度単位で計上され、かつ契約終了後に支払うというような運用では複数年間の匿名加工医療情報の提供にかかる契約締結が困難であり、契約上の工夫が必要であった。

また、いずれの認定事業者もデータ件数に一定程度比例した料金体系を設定しているため、事業拡大におけるデータ件数の増加は認定事業者の価値を高める一方で、全件のデータの利用のコストも上昇することになるので利用者にとって敷居が高くなりつづける。従って、利用者のニーズに応じてデータをサンプリングして提供することによってコストを抑える方策も今後検討されるようになるだろうが、このサンプリングのプロセスに瑕疵があるとデータにバイアスが生じ、そのデータを用いた研究に疑問がつくことになる怖れがある。認定事業者側にもデータサイエンスについて十分な素養を持つ人材を配置し、またデータ構成、分布等の基本的な統計量等についても研究者に適切に開示するような体制を整えていくことが望まれる。

上記のような問題意識を踏まえると，長期的視点において、認定事業者の事業継続性をサポートしつつ、研究者が認定事業者を利用しやすいような研究ファンドの開発や、後述する認定事業者の計算機環境の整備をサポートするような制度設計への考察をしていくことが必要であると思われる。

(2)機械学習する環境にかかる検討

外部に提供できない要配慮個人情報を用いた研究のために Data Visiting あるいは Data Enclaving というコンセプトが提唱・運用されている。これは、データを保有している事業者が計算機資源を提供する、あるいはプライベートクラウドで立ち上げられた計算機インスタンスへの接続を提供し、その中で分析ができるように「外部から閉じられて研究活動が完結する環境」を提供することによって、研究者等への第三者提供の必要がないようにするものである。また研究者にとっても非常に高額な計算機環境を用意しなくてもクラウドの従量課金制下に柔軟に計算機資源の配置を設計できること、匿名加工医療情報の提供にかかる匿名化手法、リスク評価のプロセスを短縮できることで恩恵がある。

一方で、Data Visiting を提供する事業者側はデータマートや分析ツール等の環境整備の負担がかかる。

そのため、例えば米国の National COVID Cohort Collaborative (N3C)プロジェクトでは、各プロジェクト団体からの医療情報収集対象の Common Data Model(CDM)として、OHDSI、Pcornet、ACT、TriNetX の CDM を利用し、かつデータマートとして OHDSI の OMOP を利用するなどして、研究プロジェクトの迅速な展開と構築・運用管理の負担を少なくしている。認定事業者の認可より日にちが経っておらず、学術研究に関するノウハウの蓄積がなされているところであるが、本研究でも認定事業者における Data Visiting/Data Enclaving の運用を見据えて、現行の次世代医療基盤法やガイドラインの遵法のもと仮想環境の展開を試み、各法律やガイドラインの解釈や仮想環境の運用手法について重要な知見を蓄積した。

今後の研究計画・予定

以上、今年度の研究成果と本研究班の当初の研究計画を踏まえ、今後の研究活動は以下の通り予定している。

(1) 認定事業者を活用する AI 研究・制度環境に関するアンケートの実施

初年度において検討したアンケート案について、さらに研究班、コンサルの支援下にブラッシュアップとアンケート依頼先の最終検討、アンケートシステムの構築を行う。最終年度後半までに学会事務局や地方自治体にアンケート依頼を発出し、回答を収集・分析する。

(2)海外の医療情報の収集・分析にかかる制度・事業の調査

海外において、わが国の次世代医療基盤法下の認定事業者の様に、現行の個人情報保護法の制限を超えて医療情報を収集するために特別な法律的措置や制度設計を行い、医療情報収集事業を行っている事業体に関する状況調査を文献中心に実施し、認定事業者のオンサイトリサーチセンターの拡張あるいはプライベートクラウド活用にむけた施策に向けた調査報告書を作成する。

(3)前年度から継続している AI 研究

第 1 号、第 2 号事業者における 4 種の AI 研究を継続し、次年度中に成果を出し論文投稿を行う。

(4) 次世代医療基盤法第 25 条下にかかるデータ融通の検討

第 25 条下におけるデータの名寄せ・融通について、認定事業者間の契約と事務局に関する検討を行い、認定事業者の事業計画、厚生労働行政への施策提言を行う。

(5)認定事業者にかかる厚生労働省行政の施策提言にむけたとりまとめ

認定匿名加工医療情報作成事業者が提供する匿名加工医療情報の種類、作成手法、国内における匿名加工医療情報のニーズ等の調査、及び認定事業者の匿名加工医療情報の提供を受けての AI 学習の試行、海外のオンサイトリサーチセンター事例調査を取りまとめ、匿名加工情報及び ICT/AI の技術革新を利用した研究の実現可能性の検証をし、匿名加工医療情報の利活用の技術的課題の抽出及びその解決策にむけた厚生労働行政の施策提言を行う。

E. 結論

当初の予定では、第 3 号事業者あるいは第 1 号事業者の医用画像収集事業の開始を想定していたが、いずれも実現せず、次世代医療基盤法第 25 条下のデータの名寄せと融通に掛かる検討に切り替えた以外は全体としての研究遂行は予定通りに進行している。通常の匿名加工医療情報の提供プロセスに加えて、匿名加工医療情報で機械学習する環境にかかる検討を入れたため、契約までに時間を要したが、結果的には各事業者の協力により、4 種類の AI にかかる研究を実施できており、当初の想定より研究範囲が広がっている。来年度より、認定事業者の潜在的利用者へのアンケート、海外事例の調査、次世代医療基盤法第 25 条下のデータ融通の検討も加わり、最終的に認定事業者にかかる厚生労働省行政の有意義な提言につながることが期待される。

F. 健康危険情報

本研究はリアルワールドデータ（日々の診療行為から発生したデータや、診療報酬、健診データ）の二次利用にもとづいた研究であり、侵襲性のある活動はない。また、次世代医療基盤法の認定事業者の匿名加工医療情報を用いる研究であるため、人を対象とする生命科学・医学系研究に関する倫理指針の対象外である。

認定事業者の潜在的利用者に対する認定事業者の認識にかかるアンケート調査についても、代表研究者が所属する愛媛大学において指針対象外であり IRB の付議対象とはならないという判断を受けている。

G. 研究発表

1.論文発表

1)木村 映善, 窪寺 健, 長瀬 嘉秀 : 健診標準フォーマット実装ガイドの開発 : 医療情報学 41,225-236,2022.

2)木村 映善 : リアルワールドデータを利用した国際的臨床研究への参加にむけて : 愛媛医学 40, 55-60,2021.

3)Tanaka, K., Kimura, E., Oryoji, K., Mizuki, S.-i., Kobayashi, T., Nishikawa, A., Yoshinaga, E.,Miyake, Y. : Hypertension and dyslipidemia are risk factors for herpes zoster in patients with rheumatoid arthritis: a retrospective analysis using a medical information database : Rheumatology International,1-7,2021.

4)Miyake, Y., Tanaka, K., Senba, H., Hasebe, Y., Miyata, T., Higaki, T., Kimura, E., Matsuura, B.,Kawamoto, R.: Education and household income and carotid intima-media thickness in Japan: baseline data from the Aidai Cohort Study in Yawatahama, Uchiko, Seiyo, and Ainan : Environmental Health and Preventive Medicine 26,1-8,2021.

2.学会発表

1) 木村 映善 : 認定事業者におけるリモートエグゼキューションの検討 : 2021 年度統計関連学会連合大会講演報告集,2021.

2) 木村 映善 : PHR の実装における課題 : 医療情報学 41(Suppl.),364-367,2021.

3) 木村 映善 : 観察研究に資する RWD 収集における CDM の意義 : 日本医療情報学会 第 41 回医療情報学連合大会 共同企画 13 包括的・重層的症例データベース J-CKD-DB により可能となった臨床研究と今後の発展への期待,271-274, 2021.

3. その他

1)木村 映善 : PHR と医療健康情報の標準化 : Precision Medicine 4,22-25,2021.

H. 知的財産権の出願・登録状況

該当なし