

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)  
総合研究報告書

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（愛媛大学医学部 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じた密着取材的なプロセス分析を通じ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（現行では 27 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。

次年度から最終年度にかけて、英国、特にスコットランドにおける認定事業者相当の事業の渡航調査、認定事業者にかかる意識調査を実施し、解析をした。また、次世代医療基盤法の第 25 条下にかかるデータ融通の WG の成果を取りまとめた。そして本研究班の研究活動の総括・報告の場として、2023 年 11 月に開催された医療情報学連合大会において、シンポジウム「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 AI 研究の実現可能性の検討」において代表研究者、分担研究者による研究成果の発表を行った。また、英国でのリアルワールドデータの活用について、キーノートレクチャー「英国とマレーシアにおける医療情報等の利活用について」において、英国 Health Data Research の CTO である Emily Jefferson 教授に講演頂いた。

これらの研究から得た知見を踏まえて最終的な総括として、認定事業者にかかる施策への提言を取りまとめた。

## 研究分担者

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ため、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（2023 年現在では改正され

ており、第 27 条となっている）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 のプロトタイプ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アリに決定、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した。認定事業者を利用した AI 研究を志している地方公共団体や研究者を事前に特定することは困難であると考えられたため、地方自治体は一定基準以上の人口を擁し、部署の名称からリアルワールドデータを活用する可能性が伺えるものを抽出し、アン

ケート協力の依頼文面を郵送した。学術団体は医学系の学会から公開されている会員数が多いものから抽出し、学会会員に対してアンケート依頼文面のメール配信を依頼した。

#### (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（2023年現在は第27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こととした。定例会議、エンジニアWG、契約WG、審査WGを延べ30回開催し、次世代医療基盤法第25条下におけるデータ交換にかかる技術的課題、複数の認定事業者間にまたがる契約、審査の課題について検討し、最終提案にむけて資料を取りまとめた。

#### (5) 海外の医療情報の収集・分析にかかる制度・事業の調査

2022年9月に渡航し、英国で医療情報の収集、リンケージ、研究者のデータ利活用に供している組織として、NHS Digital (England); Scottish Safe Havens; Public Health Scotland; University of the West England; Imperial College London; Small Area Health Statistics Unit; Clinical Practice Research Datalinkを訪問し、聞き取り調査を実施した。またその聞き取り調査及び提供された資料をベースに文献調査を実施した。また、論文2編を執筆中に不明の点があれば当時ヒアリングした担当者にメールで照会した。

#### （倫理面への配慮）

本研究は、認定事業者より提供された匿名加工医療情報を利活用した研究や、個人情報収集しない形態でのアンケートを実施しており、倫理委員会の付議対象ではない。但し、認定事業者の利用目的等審査委員会による審査を受け、承認された上で匿名加工医療情報を提供頂いている。また、本研究に先立ってプロトタイプ開発のために研究者

が所属する医療機関のデータ等を利用した場合には、個別に倫理委員会の承認を得て実施している。

また、LDIより提供をうけた匿名加工医療情報については、安全なデータセンター内で匿名加工医療情報を機械学習しその結果のみをLDIの確認のもとセキュリティ便で研究者に送付、確認する体制を構築し、情報セキュリティを確保した。J-MIMO提供の匿名加工医療情報の解析については東京大学医科学研究所において高度に管理されたスパコンおよび、研究者のみがアクセスできる、施錠された部屋に設置されたPCのみを用いて行った。

## C. 研究結果

### (1)匿名加工医療情報を活用した研究

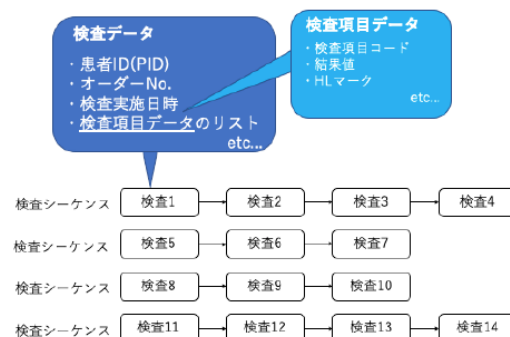


図 1 抽出する検査データのイメージ

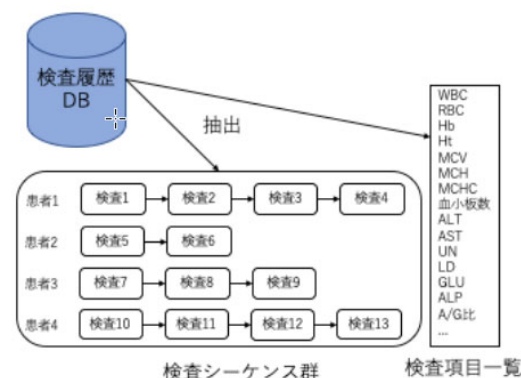


図 2 検査データの抽出のイメージ

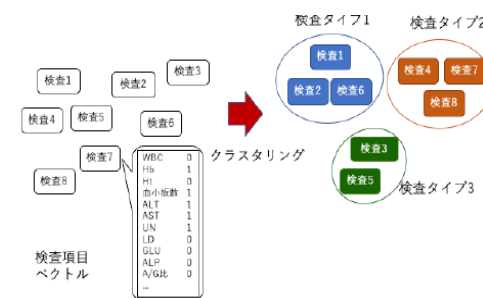


図 3 検査データの分類のイメージ

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的研究を実施した。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た。

(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ーのシーケンス解析を行い、診療オプションとなるバリエーションの安全性、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ー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ー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ーのデータは第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とオーダー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。抽出したレコードから、患者ID、オーダーグループ、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2)。グルーピングされた検査オーダー項目は各検査タイプとして類型化する(図3)。

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ー、検査オーダー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。

学習したモデルに検査結果を入力し、オーダー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、再現率を評価した。

また2023年には、複数の医療機関における頻出医療指示パターンを比較することを通して、それぞれの医療機関における特徴を把握することを目指した<sup>1</sup>。各医療機関の医療指示データに対して、SPM(SequentialPatternMining)を適用し、minsupを満たす頻出医療指示パターン(SequenceVariant, 以下SV)を抽出する。次に、SV間の距離を改めて定義し、算出する。その距離を用いて階層的なクラスタリングによって対象となる医療機関群をいくつかのクラスタに分類し、そのクラスタ内でLongestCommonSubsequenceVariant(以下、LCSV)を算出し、MergedSequenceVariant(以下、MSV)を生成することで、特徴が近い医療機関のSVにおける共通点と相違点を示した。

結果として、併合シーケンスバリエーションMSV生成の際に、最長距離法とワード法によるクラスタリング分類が良好な分類を示していた。PCIの治療において実施すべき医療指示(「ECG」、「HR」、「CRP」)が共通部分として抽出できていたことを確認した。しかし、今回の研究において、疾患ベースで絞り込まなかったため、PCI治療は狭心症や心筋梗塞等の複数の疾患に適用される治療法であったこと、SPMアルゴリズムに使われるminsupの値の適正性の問題、PCIに関する治療パスの特徴を表さない一般的な医療検査からの影響を排除していないという課題が残されている。今後の課題としては、適切なminsupを決める方針を示す必要がある。更に、分析対象とした疾患の治療の特徴を表さない医療指示の除外などの前処理を行うことで、より制度の高い分析が可能になると思われる。

<sup>1</sup> Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 横田 治夫. 千年カルテの匿名加工医療情報を利用した多医療機関の電子カルテに対する

シーケンス解析. 医療情報学 43(Suppl). 2023:309-14.

(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

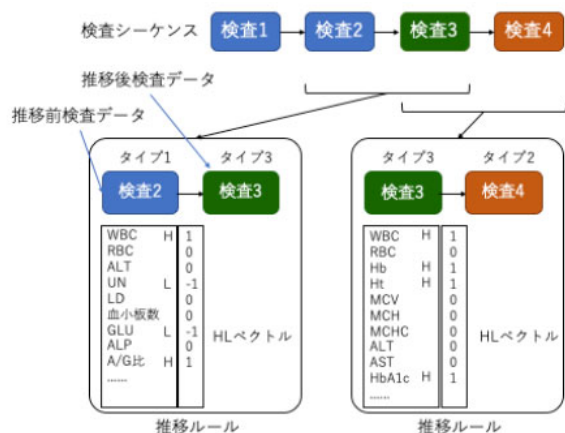


図 4 検査シーケンスのマイニングのイメージ

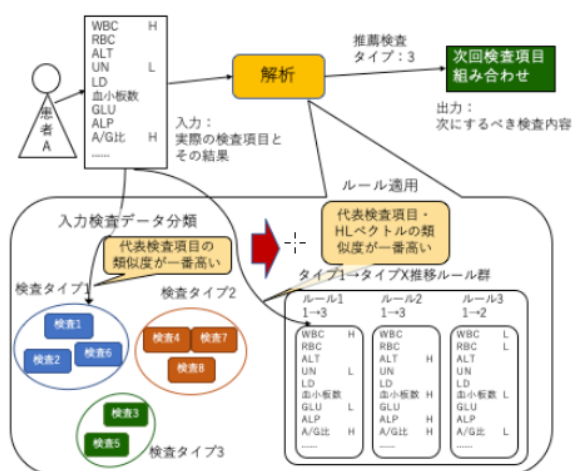


図 5 検査項目組み合わせの推薦のイメージ

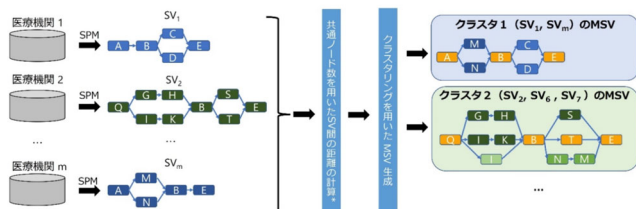


図 6 SPM、クラスタリング手法を介した MSV 生成

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 の研究を行った<sup>2</sup>。

事前にコンテナ型の仮想環境を構築し、前処理、学習、推論を行うプログラムの開発をした。そして、その Docker イメージを認定事業者のデータセンター内に持ち込み、プログラム実行、データ統計やログ・エラーの取得を委託し、結果はセキュリティ便や電子メール等で提供頂いた。その結果をもとにプログラムの改修し、その改修内容を認定事業者に提供し Docker イメージを更新して再実行するというプロセスを繰り返した。最終的に開発した機械学習モデルの評価結果と特徴量重要度をセキュリティ便で受領した。

認定事業者との電子メールのやりとりは 100 回近くに及んだ。時間を要したプロセスを上位順に記載する (表 1)。

予測精度の評価として当該研究グループは 2020 年度に発表した RF を用いた敗血症の予測モデルと本研究における敗血症の予測モデルの評価指標の比較をした (表 2,3)。

全体的に 2020 年度研究に比して今回の研究の方がよい評価指標となっている。留意点として、2020 年度の研究と本研究では用いた特徴量やデータの母数などの前提条件が異なる。特徴量の違いとして、2020 年度の研究では、DPC だけでなく DWH からバイタル、検体検査、実施オーダー情報(薬剤処方、手術、放射線)等も利用し、特徴量としては 358 個を用いた。本研究では DPC を元にした表データのみから 166 個の特徴量を用いた。また、2020 年度の研究では宮崎大学医学部附属病院のみの前処理済データ 13,527 件を利用したのに対し、本研究では千年カルテの 20 の医療機関の前処理済データ 68,593 件を用いた。2020 年度の宮崎大学医学部附属病院のみのデータ数に対して、本研究の 20 の医療機関のデータ数が 5 倍程度と、施設数の比率の 20 倍に対して小さかった 1 つの要因として、2020 年度では前処理で欠損値のデータ補完を行っ

<sup>2</sup> 松田 敦義, 池田 遼太郎, 荒木 賢二, 松尾 亮輔. 次世代医療基盤法に基づく匿名加工医療情報を利用した入院後敗血症の予測モ

デルの開発における課題と対策の検討. 医療情報学 43(Suppl). 2023:321-4.

たのに対し、本研究では欠損があるデータはすべて削除したことが考えられる。

表 3 に、敗血症の発症なし、発症ありで分けた場合の評価指標を示す。各指標において発症なしの値は非常に大きい、発症ありの各指標の値が小さい事がわかる。このことから、表 2 の本研究の評価指標は発症なしの評価結果により値が高められているといえる。発症ありの時の精度を高めることが今後の課題である。

予測モデル(RF)の特徴量重要度について、重要度が大きい順で上位 10 個を表 4 に示す。これを見ると体重、身長、年齢が上位に来ている。

多施設の匿名加工医療情報を用いる効果 2020 年

表 1 時間がかかった開発プロセス

|   | 開発プロセス                           |
|---|----------------------------------|
| 1 | プログラム実行で発生したエラーの確認と対応            |
| 2 | データ授受等、研究の進め方の確認                 |
| 3 | データベース内のテーブルの概要確認と、対象テーブルの修正     |
| 4 | データベース接続情報の確認と、接続情報の修正           |
| 5 | Docker イメージでインストールされるライブラリの確認と承認 |
| 6 | 千年カルテから取得するデータ種類の確認と承認           |

表 2 2020 年度の研究と本研究の評価指標の比較

| 研究      | AUC   | F 値   | 適合率   | 再現率   |
|---------|-------|-------|-------|-------|
| 2020 年度 | 0.727 | 0.424 | 0.374 | 0.490 |
| 本研究     | 0.769 | 0.528 | 0.536 | 0.523 |

表 3 本研究における敗血症の発症別の評価指標

| 敗血症  | F 値   | 適合率   | 再現率   |
|------|-------|-------|-------|
| 発症なし | 0.997 | 0.996 | 0.998 |
| 発症あり | 0.058 | 0.075 | 0.048 |

表 4 本研究における予測モデルの特徴量重要度

|    | 特徴量                     | 重要度   |
|----|-------------------------|-------|
| 1  | 体重                      | 0.150 |
| 2  | 身長                      | 0.130 |
| 3  | 年齢                      | 0.077 |
| 4  | 併存症_真菌症                 | 0.041 |
| 5  | 併存症_原虫疾患                | 0.036 |
| 6  | 併存症_悪性新生物<腫瘍>           | 0.028 |
| 7  | 併存症_高血圧性疾患              | 0.024 |
| 8  | 併存症_食道/胃/十二指腸の疾患        | 0.023 |
| 9  | 併存症_凝固障害/紫斑病及びその他の出血性病態 | 0.023 |
| 10 | 性別                      | 0.022 |

<sup>3</sup> Miura T, Kimura E, Ichikawa A, Kii M, Yamamoto J. Evaluating Synthetic Data Generation Techniques for Medical

度の研究と本研究の評価指標の比較から、多施設のデータを利用することにより、単一の医療機関のデータを用いた機械学習モデルよりも予測精度を向上できる可能性が示唆された。本研究の結果は特徴量の選定や欠損値の補完などができていない上で得られたもので、学習前のプロセスをより精密に行うことで、更に精度を高めていく事ができると考えられる。多施設のデータを用いてデータの母数を増やす事により、これまで難しかった大規模な予測モデルの開発や、多施設間での臨床データの比較、希少疾患の分析への活用などの用途が期待される。

(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った<sup>3</sup>。

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)、AIM、MEWM-PGM、深層学習の GAN ベースの CTGAN、岡田らによる統計量ベース(STAT)、GCoupla 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、属性ごとの L1 距離、L2 距離、Hellinger 距離、Wasserstein 距離の統計距離を用いて評価した。

なお、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）

機械学習等を実行した結果、愛媛大医学部附属病院 DPC データセットを用いた実験は、事前段階は int 型として準備していたデータが実際は float 型であったなどデータ型に関するエラーなどが生じた。他にも様々なエラーが生じたが、最終的に 150 往復程の情報のやり取り、プログラムの修正ののち、評価まで行うことができた。LDI の匿名加工医療情報を用いた実験についても、想定したデータが出力されない動作不具合が発生し、その原因解析および対処を試みたが、最終的に不具合解決に至らず、実験を中断することとした。同様にメールにして 30 件ほどの情報のやり取りがあったが、現在もエラーの原因を解析中という結果になった。

統計的な類似度については、各属性、DP 非適用 ( $\epsilon=\infty$ ) の時は元データと近い合成データが作成されているとみなせる。しかし、DP 化を施すことでどの手法も有用性が大きく落ちることが確認された。一方で、AIM や MWEM-PGM といったグラフィカルモデルベースの手法は比較的有用性の高い結果になった。利活用観点での有用性については、DPC データセットについて、目的変数を入院日数とし、合成データで学習させた LightGBM による回帰モデルを、元データをテストデータとして推論させたときに、RMSE が 17.4428 であり、これをベースラインとして評価した。結果として、統計的距離の時と同様に AIM や MWEM-PGM などのグラフィカルモデルベースの手法が比較的良い RMSE を提示する結果となった(図 7)。

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従来型の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。

#### (b) 第 2 号業者の匿名加工医療情報を用いた研究

k-匿名化は匿名化データを医療データ等から作成するにあたって最も重要な概念の一つである。k-匿名化することによって、データソース中のある特定のデータを特定困難にすることができる。しかしながらこの k-匿名化は攻撃者がデータの背景知識をまったく持っていないことを仮定している。一方、差分プライバシーは背景知識を持った攻撃者からも安全な強力なノイズ加工手法である。しかしながら、この手法は k-匿名化で考慮しているようなデータからの個人の識別特定に関しては考慮していない。そのため、その k-匿名化と差分プライバシーの双方を考慮した手法の開発が必要となっている。その双方を考慮した既存の手法の多くはランダムサンプリングに基づく手法となっており、データ全体を公開することを考えた手法は知られていない。そこで、k-匿名化と差分プライバシーの双方を満たしながらデータソース全体を匿名化して公開する手法について検討し、その実効性の匿名加工医療情報を用いて検証することとした。

その際、差分プライバシーの分析について議論がなされ、個人の識別特定にかかる活動に相当しないと判断されたので、実験データ(一般財団法人日本医師会医療情報管理機構、2021-3)の利用に関

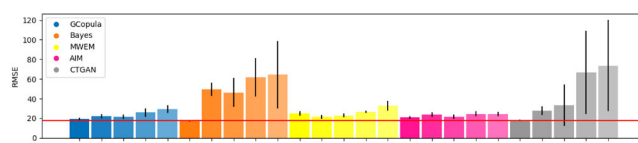


図 7 LightGBM モデル回帰精度(RMSE)

して一般財団法人日本医師会医療情報管理機構匿名加工医療情報提供審査委員会の承認を得て実施

した。第2号事業者の主なデータソースは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。東京大学医科学研究所ヒトゲノム解析センタースパコンシステム「SHIROKANE」を利用した。

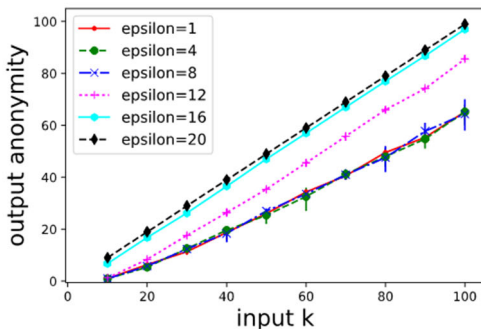


図8 アルゴリズム1下の匿名性

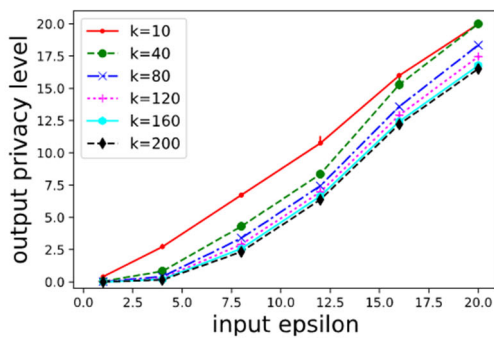


図9 アルゴリズム2下の匿名性

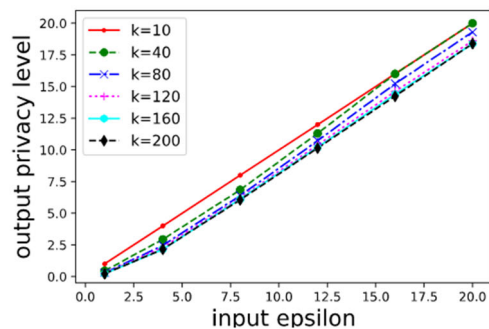


図10 アルゴリズム3下の匿名性

山本が考案した  $k$ -匿名化と差分プライバシーの双方を満たしながらデータソース全体を匿名化して公開する手法<sup>4</sup>は、Randomized Response と  $k$ -匿名化を組み合わせている。この手法を評価するた

めに、3種類の匿名加工アルゴリズムを比較した。1番目は、 $k$ -匿名化を行った後 Randomized Responseを行うものである。2番目は、Randomized Response後に  $k$ -匿名化を実施する。1番目と異なり出力は  $k$ -匿名性を満たす。3番目が提案手法であり、まず  $k'(<k)$ -匿名化を行い、Randomized Responseを適用し、さらにランダム化されたタプルに対して  $k$ -匿名化を適用するものである。

図8の結果から出力の匿名性は入力  $k$  にほぼ比例し、匿名性の増加率は  $k$  に依存しないことがわかる。 $\epsilon$  が大きいほど高い匿名性が確保される。しかし、 $\epsilon$  が大きくなるにつれて差分プライバシーの保証は低下し、個人の存在に関する情報が曝露されやすくなる。従ってアルゴリズム1は匿名性とプライバシーレベルに強いトレードオフがある。

アルゴリズム1と異なり、アルゴリズム2、3は  $k$ -匿名性が担保されており、さらに入力  $\epsilon$  に従ってより強いプライバシー保証を提供する。アルゴリズム3はアルゴリズム2と比較して、どの  $\epsilon$  に対してもプライバシーレベルは安定しており、アルゴリズム3の精度はアルゴリズム2より高くなることが推察された。

次にアルゴリズムの出力精度を評価するために、元の分布と提案手法の出力の差を測定する手段として KL divergence を用いた。K 値が大きくなると入力値がグループ化されやすくなり精度が低下する。アルゴリズム2においては  $\epsilon$  が小さいとプライバシー保障が強くなりすぎるため、出力精度が悪かった。アルゴリズム1では、出力されるプライバシーのレベルが  $\epsilon$  から変化しないため最も高い精度が得られるが、匿名性は低下している。一方でアルゴリズム3は匿名性、プライバシー保証、出力精度の全てにおいて高品質な結果を提供できることが確認された(図11)。

本研究では、 $k$ -匿名性と  $\epsilon$ -差分プライバシーの両方を満たす、データ共有のための新しいプライバシー保護手法を提案した。本研究で提案する手法は、データのサンプリングを前提とせず、元デー

<sup>4</sup> Akito Yamamoto, Eizen Kimura, and Tetsuo Shibuya, (e, k)-Randomized Anonymization:  $\epsilon$ -Differentially Private Data

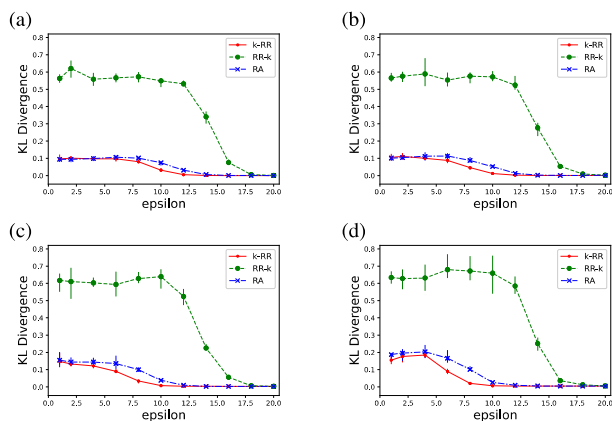


図 11 アルゴリズム間の KL Divergence の比較  
(a)k=100,(b) k=400,(c)k=800,(d)k=1,600 アルゴリズムは、  
それぞれ 1(k-RR)、2(RR-k)、3(RA)

タに含まれる全ての情報を公開できるという利点を持つ。特に、我々の第 3 の方法である  $(\epsilon, k)$ -ランダム化匿名化は、 $k$ -匿名性を真に満たしながら、 $\epsilon$  よりも強いプライバシー保証を達成できる新しい方法である。また、実データを用いた実験の結果、 $(\epsilon, k)$ -ランダム化匿名化により、元データに近い精度の高い結果が得られることが示された。今後の重要な課題は、無作為化応答手法との統合に適した  $k$ -匿名化手法の開発である。最適化された手法は、データの使用法や分析目的に強く依存する可能性があるため、我々はこの問題を継続的に探求する予定である。

## (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典型的には医療情報を保管するデータセンター(図 12 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 12 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。この運用に関するコンセプトを学会で報告した<sup>5</sup>。

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、認定事業者は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

<sup>5</sup> 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.

この仕組みを用いて、3種類のAIに関する研究が行われたが、一通り遂行できたのは2種類であった。中断に至った研究<sup>6</sup>については、事前のプロトタイプの開発では正常に稼働していたが、LDI内に設置した時にデータ型などに関するエラーや想定したデータが出力されない不具合があり、30件ほどやりとりしたが原因特定に至らずエラーの原因を解析中に中断することとなった。この原因として、データセットの様子を直接視認することができないため、エラーログからデータセットの様子を想像することができず、また実施状況をコンソールからリアルタイムに確認出来ずセキュリティ便やメールでやりとりするという大きなタイムラグが発生したことが円滑な遂行の妨げとなった。また分析を完遂した研究からもこの運用方式についての課題が提示された<sup>7</sup>。すなわち、機械学習モデルの精度の観点では、データを確認しながら特徴量の内容(例えば検査値の単位やノイズによる影響など)や表記ゆれなどの確認ができないため、データが正しい前提で開発を行う必要があり、実際に何かデータの問題を内包していた場合には、精度が落ちる等の問題が考えられる。研究にかかる時間の観点では、データの統計量や結果は認定事業者を通して確認する必要があるため、データを直接確認しながら機械学習モデルの開発を行うよりも時間が多くかかった。教師あり機械学習モデルを開発する上では、様々なデータを集計し、適切な特徴量を選定するプロセスが必要である。精度

[illegible]

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。アンケート依頼を送付した地方公共団体の部署は 539 部署となった。それらの部署からの総回答数 461 件中有効回答 233 件(有効回答率 51.5%)であり、アンケート回答率は 43.2%であった。31 学術団体の会員からは総回答数 713 件があり、そのうち有効回答として 474 件(有効回答率 66.5%)が得られた。全体的

11

に次世代医療基盤法・認定事業者に対する認知度は低かった。主成分分析、順序カテゴリカル解析を通して、認定事業者の認知度の向上と認定事業者の利活用を推進するという 2 つの施策を推進する必要性が見いだされた<sup>8</sup>。

#### (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、次世代医療基盤法第 25 条(2023 年現在は 27 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 13)。

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。

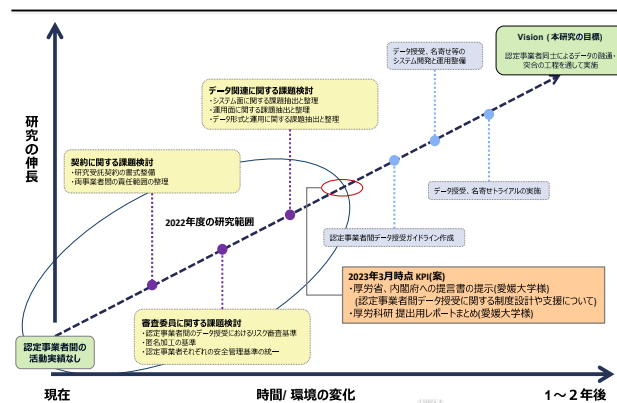


図 13 次世代医療基盤法第 25 条下のデータ融通・突合にかかる研究の全体像

##### (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

<sup>8</sup> 木村 映善, 高田 春樹, 荒木 賢二, 黒田 知宏, 星 佳芳, 佐々木

香織, 伊藤 伸介. 次世代医療基盤法の認定事業者の認知状況, 活用への意識に関する社会調査. 保健医療科学 2023;72(5):454-63.

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3分野に対応するWGを運営し、30回以上に亘る活動の結果、別途【資料1-1】【資料1-2】に掲載するように、次世代医療基盤法第25条による認定事業者間の医療情報の相互融通の検討にかかる報告として取りまとめた。

#### (5) 海外の医療情報の収集・分析にかかる制度・事業

英国において、匿名加工情報を研究者に提供する代わりに、データ加工、分析環境も内包したTrusted Research Environment(TRE)というコンセプトを推進しており、TREによってはリモートアクセスも許可していることが確認された。このTREの運用にあたって英国のOffice for National Statisticsが考案したFive Safe Models、すなわち”安全な利用者”、”安全なプロジェクト”、”安全な施設/環境”、”安全なデータ”、”安全な分析結果”という5つの安全管理対策の軸にそった事業展開は、わが国でも認定事業者にTREのコンセプトを導入し事業拡大する際の参考になると思われた<sup>9</sup>。次に、データのリンケージにあたって、直接識別子による名寄せとデータそのもののリンクの役務をIndexerとLinkerという独立した組織に担わせることで個人情報露出を最低限におさえた秘匿性の高い運用を実現していることが確認された<sup>10</sup>。本邦が学ぶべき点として、TRE、Five Safe Models、Indexerと

Linkerの分離、利用者への教育体制の拡充が確認された。この渡航調査に関する成果は現在論文投稿、査読中である。

また、公的統計や行政記録情報を扱う視点から、イギリスにおける学術研究目的のための公的統計の二次利用や行政記録情報の活用状況について調査した。Digital Economy Actに基づく法整備やADR UKの支援、またマイクロデータの利活用に関する方法的な枠組みであるFive Safes Model下にイギリスで展開されている大規模データの利活用状況について調査した<sup>11</sup>。

## 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に掛かる研究を遂行した。この経験から認定事業者を利用する意義等についての考察を述べる。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いうことになった。提言については、別添「【資料2】次世代医療基盤法 認定事業者にかかる政策等への提言」にて記述している。

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

<sup>9</sup> 木村 映善. Scottish Safe Haven における Five Safes の解釈と運用. 2023 年度統計関連学会連合大会. 2023:3CPM02-1.

<sup>10</sup> 木村映善, 佐々木香織, 伊藤伸介. 「英国における行政記録情報と医療情報をリンケージした利活用状況について～英国における機械学習での利用を前提とした医療情報の利活用状況の調査から～」『大規模データの公開におけるプライバシー保護の理論と応

用』. 公的統計マイクロデータ研究コンソーシアム/統計数理研究所. 2022.

<sup>11</sup> 伊藤 伸介, 木村 映善, 佐々木 香織. イギリスにおける大規模データの二次利用の新たな展開 ―スコットランドの事例を中心に―. 統計研究彙報. 2024;81:89-108.

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く、このメリットについて啓発することが望ましいと思われる。

リアルワールドデータに限定すれば、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。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い。潤沢な研究資金を有する民間企業にとっては、スケールメリットと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。利用コストについての理解は先行して取り組んでいる英国においても理解を頂くのに苦労してきたと伺っている。

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、適切なサンプリング手法、データ構成を担保し、基本的な統計量等について研究者に適切に開示できる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り、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。2024 年現在、改正次世代医療基盤法では仮名加工医療情報の概念が新たに追加され、認定

作成事業者が認定利用業者に対して仮名加工医療情報を利活用させる枠組みに関する議論がなされている。我々が本研究で取り組んできたことは、II型認定下に認定利用事業者が仮名加工医療情報を自分の所に保存せず、認定利用事業者が用意した解析環境下に限定して利用する状況に相当する。但し、認定作成事業者内の解析環境相当の環境には一切アクセスできず、認定事業者側の担当者と解析のためのプログラム引き渡しとその結果をオフラインでやり取りするという一段の手間がかかり、それが機械学習にかかる研究の円滑な遂行の支障になることを経験してきた。改正次世代医療基盤法下に、認定作業事業者の管理区域内に設置された解析環境に対して、同じく認定事業者内の取扱区域あるいはセキュアなネットワークを通じて解析環境にリモートアクセスをする環境の検討がなされているが、本研究班の経験からもこれらの環境の早期の実現を強く要望するものである。

## E. 結論

本研究の目的として、実際の認定事業者へのデータ提供依頼から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の密着取材的なプロセス分析を通じた有用性等の検証により、匿名加工医療情報がどのような研究に資するか、またAI技術を用いた研究の技術的課題を明らかにすることを挙げていた。

このテーマについては4種類の研究テーマを実施し、全て発表につながった。そのうち1種は最後まで完遂できなかったが、匿名加工医療情報に帰する問題ではなく、解析環境を認定事業者に設置してオフラインで運用したことによる研究速度・進捗の問題である。この問題については仮名加工医療情報の概念の導入及び認定利用事業者のための利用環境整備によって解決できる課題であると考えている。総じて、本研究の範囲では匿名加工医療情報がAI技術を利用する研究に貢献することが確認された。

また、認定事業者を利用する可能性のある者から次世代医療基盤法の認知度、認定事業者への期

待等の調査、及び海外の認定事業者の調査ともあわせて、認定事業者を利用した研究を加速する施策の提言に繋げることを目的としていた件については、認定事業者の認知度にかかるアンケートや英国の渡航調査を通して、認定事業者の認知度の向上と認定事業者の利活用を推進するという2つの施策を推進する必要性が見いだされた。

以上の経験を踏まえて、認定事業者を利用する研究を支援する施策の提言として、次世代医療基盤法の認知度を高めるために啓発活動を行うこと、認定事業者に医療等情報を提供する事業者の負担軽減を検討すること、認定事業者を利用した研究を支援するファンドや研究助成事業を設定すること、認定事業者に「信頼できる研究環境」の概念を導入しそれを支えるためのクラウド基盤の活用を積極的に支援すること、改正次世代医療基盤法に準拠した運用の早期確立への支援、相互運用性に寄与する標準規格の導入支援をすること、の6つの提言に集約した。

本研究班の研究は可能な限り成果を公開のものとすることとし、結果として24編の論文・学会発表、及び2件のシンポジウムの開催として結実した。

本研究班による取り組みが、次世代医療基盤法の認定事業者を活用したリアルワールドデータの研究を推進するために必要な施策のための知見につながれば幸いである。

## 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。但し、匿名加工医療情報を用いた機械学習のプロトタイプを開発するために研究者が所属した機関のデータを利用する際には、その研究に関して倫理委員会の承認を得た上で実施している。

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.Miura T, Kimura E, Ichikawa A, Kii M, Yamamoto J. Evaluating Synthetic Data Generation Techniques for Medical Dataset. HEALTHINF 2024. 2024;315-22.
- 2.Tetsuo Shibuya. Privacy preserving technologies toward safe medical science. The 47th Annual Meeting of the Japanese Society for Investigative Dermatology (JSID), December 2-4. 2022.
- 3.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. 2021;41(5):225-36.
- 4.Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫. 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. 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. 2023;1-5:1-8.
- 5.Yamamoto A, Kimura E, Shibuya T, editors.  $(\epsilon, k)$ -Randomized Anonymization:  $\epsilon$ -Differentially Private Data Sharing with  $k$ -Anonymity. HEALTHINF; 2023.
- 6.木村 映善, 高田 春樹, 荒木 賢二, 黒田 知宏, 星 佳芳, 佐々木 香織, 伊藤 伸介. 次世代医療基盤法の認定事業者の認知状況,活用への意識に関する社会調査. 保健医療科学 2023;72(5):454-63.
- 7.伊藤 伸介, 木村 映善, 佐々木 香織. イギリスにおける大規模データの二次利用の新たな展開 — スコットランドの事例を中心に —. 統計研究彙報. 2024;81:89-108.

### 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). 2021:364-7.
- 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. 2021;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:271-4.
- 4.木村映善, 佐々木香織, 伊藤伸介. 「英国における行政記録情報と医療情報をリンケージした利活用状況について～英国における機械学習での利用を前提とした医療情報の利活用状況の調査から～」『大規模データの公開におけるプライバシー保護の理論と応用』. 公的統計マイクロデータ研究コンソーシアム/統計数理研究所. 2022.
- 5.横田治夫, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二. 医療データのシーケンス解析とその課題. 第12回日本医療情報学会「医用人工知能研究会」人工知能学会「医用人工知能研究会」合同研究会予稿集, 2022. 2022;.
- 6.趙 子泰, Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 荒木 賢二, 横田 治夫. COVID-19の異なる医療機関と時期における頻出医療指示パターンの比較. 医療情報学42(Suppl). 2022:887-92.
- 7.Emily Jefferson. The UK's progress towards enabling secure, researcher access to sensitive health data at a UK population scale. 第43回医療情報学連合大会. 2023:120.
- 8.Le Hieu Hanh, 松尾 亮輔, 山崎 友義, 横田 治夫. 千年カルテの匿名加工医療情報を利用した多医療機関の電子カルテに対するシーケンス解析. 医療情報学43(Suppl). 2023:309-14.
- 9.三浦 堯之, 紀伊 真昇, 市川 敦謙, 山本 充子, 木村 映善. 合成データ生成技術の医療データへの適用と課題. 第27回日本医療情報学会春季学術大会. 2023:68-9.
- 10.三浦 堯之, 紀伊 真昇, 市川 敦謙, 山本 充子, 木村 映善. 実医療データを用いた合成データ生成技術の有用性検証の課題. 医療情報学43(Suppl). 2

023:pp.315-20.

11.佐々木 香織. e-Health, big-data時代の生政治 : Big-Brotherイメージの超克. 立命館生存学研究. 2023;7:13-7.

12.佐々木 香織, 木村 映善, 伊藤 伸介. Scottish Safe HavenにおけるTREとFive Safes Modelから学ぶ -改正次世代医療基盤法に基づく「仮名加工医療情報」の利活用に向けて-. 医療情報学43(Suppl). 2023:pp.325-8.

13.山本 章人, 木村 映善, 渋谷 哲朗. 医療情報における差分プライバシーなk匿名化について. 医療情報学43(Suppl). 2023:pp.306-8.

14.木村 映善. Scottish Safe HavenにおけるFive Safesの解釈と運用. 2023年度統計関連学会連合大会. 2023:3CPM02-1.

15.松田 敦義, 池田 遼太郎, 荒木 賢二, 松尾 亮輔. 次世代医療基盤法に基づく匿名加工医療情報を利用した入院後敗血症の予測モデルの開発における課題と対策の検討. 医療情報学43(Suppl). 2023:321-4.

### 3. その他

1.岡 響, 木村 映善. AIファースト・ヘルスケア : 医療現場におけるAIアプリケーションの利用: オライリー・ジャパン,オーム社 (発売); 2022.

2.木村 映善. PHRと医療健康情報の標準化. Precision Medicine. 2021;4(3):22-5.

### H. 知的財産権の出願・登録状況

該当なし

## 次世代医療基盤法第 25 条による認定事業者間の医療情報の 相互融通の検討にかかる報告

### 背景

次世代医療基盤法（正式名称：医療分野の研究開発に資するための匿名加工医療情報及び仮名加工医療情報に関する法律）の第二十五条には、「第三十条第一項の規定により医療情報の提供を受けた認定匿名加工医療情報作成事業者は、主務省令で定めるところにより、他の認定匿名加工医療情報作成事業者からの求めに応じ、匿名加工医療情報の作成のために必要な限度において、当該他の認定匿名加工医療情報作成事業者に対し、同項の規定により提供された医療情報を提供することができる。」とある。この条文が意図しているところは、以下の通りである。我が国においては、認定匿名加工医療情報作成事業者（以下、認定事業者）が複数あり、かつ医療情報取扱事業者との契約は任意である。また、契約する対象が主に地方自治体であったり、あるいは特定機能病院、臨床研究中核病院を中心とした大規模病院が主体であったりと認定事業者の特性がでることもある。すなわち、単一の認定事業者が全ての医療等情報を保有する状態にはならないということである。そのため、匿名加工医療情報の利活用者からみると、利用したい医療等情報は複数の認定事業者に分散している可能性がある。そのため、利活用者の要望に応じて、認定事業者間で必要な医療情報を融通して統合された匿名加工医療情報として提供できる道を拓く次世代医療基盤法第二十五条は有効活用が期待されるものである。しかしながら、法律上の条文としては存在するが、認定事業者の実務経験はなく、契約、利用にかかる審査、認定事業者間のデータの融通にかかる技術的課題について未検討の状態であった。そこで、大規模データの利活用研究の加速のための研究（21AC1001）の研究班と認定事業者でこの議論を引き受けて、その検討成果を公開し、認定事業者の活動方針を提言することとした。

### 検討の進め方

愛媛大学と、2022 年当時活動状態にあった認定事業者である一般社団法人ライフデータイニシアティブ(以下、LDI)、一般財団法人日本医師会医療情報管理機構（以下、J-MIMO）並びに関係者で複数回にわたり定例会議と契約 WG、審査 WG、技術的検討を行うエンジニア WG と、3 つの WG を並行して開催した。2023 年 11 月時点で各 WG において形成された合意について以下に記す。

### § 契約 WG

#### 基本方針

- ・ 認定事業者と利活用者間での契約書面は、各認定事業者の現行の契約書を踏襲する。

共通の契約書を検討するにあたり、現状の各認定事業者と利活用者との契約の比較検討を実施した。その結果、各認定事業者と利活用者間の契約条文に大きなギャップは確認されなかった。しかし、認定事業者の事業の独自性を保つために、各認定事業者と利活用者間の契約書面の共通化は行わないこととした。一方、第 25 条にもとづく円滑な医療情報交換を実現するために、認定事業者間の契約のひな型を作成することとした。

| 利活用者                                              | 提供を受ける認定事業者                                             | 医療情報を提供する認定事業者                                                                                 |
|---------------------------------------------------|---------------------------------------------------------|------------------------------------------------------------------------------------------------|
| <p>提供依頼</p> <p>*研究目的、必要となる医療情報、研究者一覧、契約金額等の調整</p> | <p>研究計画の調整</p> <p>*承認</p> <p>審査委員会</p>                  | <p>研究計画の調整</p> <p>*医療情報提供依頼</p> <p>*契約者名、研究内容、研究体制、医療情報の内容（疾患、期間）等</p> <p>審査委員会</p> <p>*承認</p> |
|                                                   | <p>医療情報の受領</p> <p>*医療情報の統合、クレンジング等</p> <p>審査委員会（承認）</p> | <p>該当データ抽出</p> <p>*情報セキュリティ便による提供</p>                                                          |
| <p>匿名加工医療情報の取得</p> <p>*研究着手</p> <p>支払い</p>        | <p>匿名加工医療情報提供</p> <p>審査委員会報告</p> <p>請求/支払い</p>          | <p>審査委員会報告</p> <p>*個別契約に基づく請求/支払い</p> <p>請求/支払い</p>                                            |

19

査委員会に諮る。審査委員会の結果は医療情報を提供した認定事業者にも共有される。利活用者は複数の認定事業者が保有される医療情報を連結、匿名化した結果としての匿名加工医療情報を受け取る。利活用者からみて契約・審査は最初に利用申請を提出した認定事業者に一本化されている。

認定事業者間にかわす契約は、次世代医療基盤法第 25 条に準拠した運営をしていくにあたって、法律、ガイドラインに準拠した運営をしていくことを相互に約束する「基本契約」と、利活用者からの申請に対して個別に締結していく「個別契約」の 2 種類とする。それぞれの契約に含める事項は下記の通りである。

#### 基本契約に含める事項

- ・ 目的外利用の禁止
- ・ データ取扱者の限定
- ・ 安全管理措置の管理
- ・ 違反時の制裁
- ・ 機密保持
- ・ 終了時のデータ取扱
- ・ 仕様と異なるデータがあった場合の対応

#### 個別契約に含める事項

- ・ 契約金額
- ・ 利用目的
- ・ 利用期間
- ・ データ仕様
- ・ データ提供方法

#### § 審査 WG

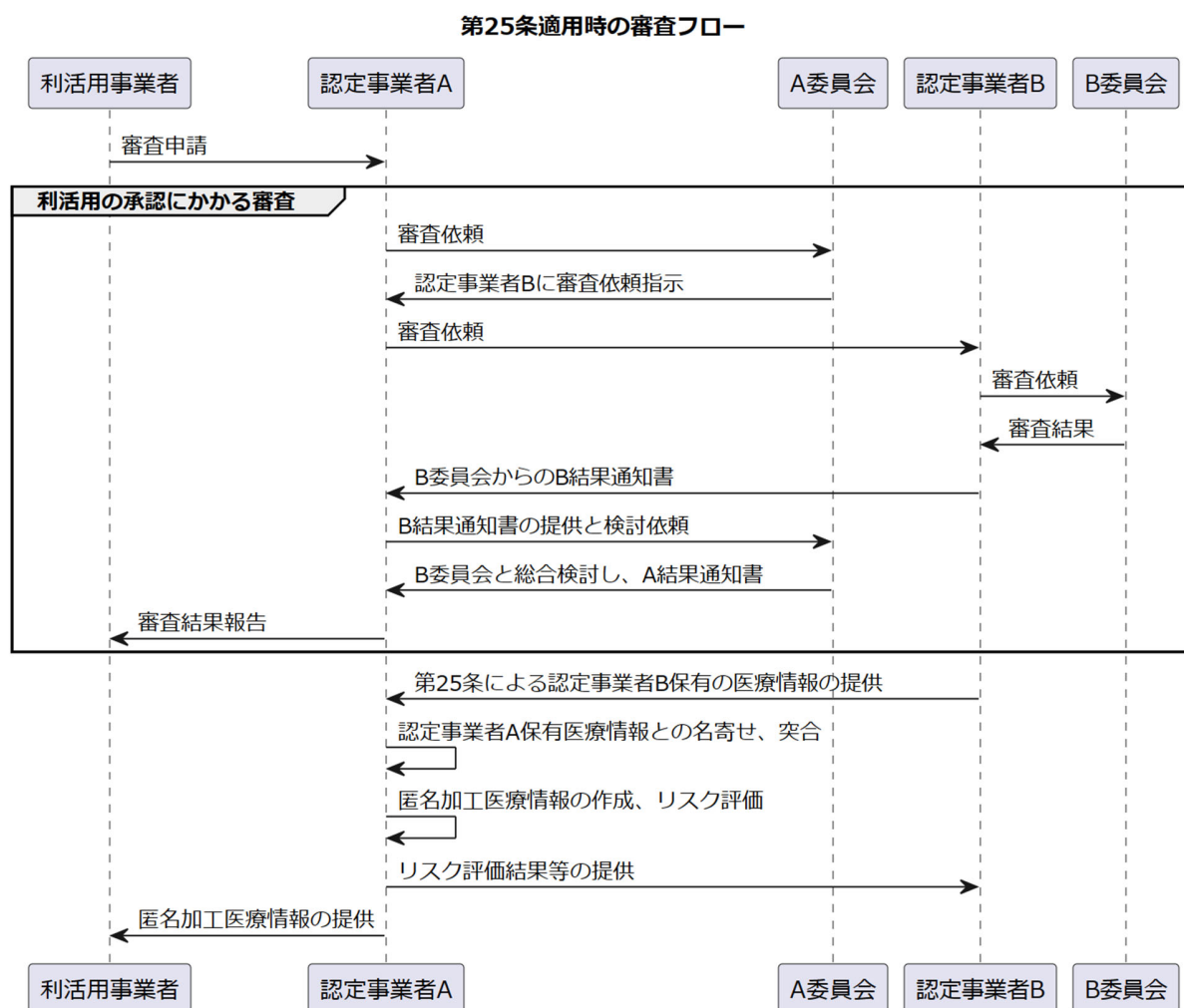
##### 全体の方向性

1 回目の審査で、利活用目的、利活用の詳細及び利活用事業者体制等、他の認定事業者からの提供に関する採否を審査する。その審査結果をもとに、提供の依頼を受ける側の認定事業者にデータ提供を申請し、依頼を受けた認定事業者は、依頼元の認定事業者の審査委員会の審査結果をもとに審査依頼（または報告：詳細は今後検討）を行う。

また、提供を受けた認定事業者は、自社の保有するデータの追加、突合を行ったのち、再度匿名加工医療情報に焦点を当てた審査を実施する事で、匿名加工医療情報に関する依頼案件との適合性の確認、質と安全性の担保が可能となる。

利活用事業者側のメリットとして、申請を行った認定事業者のみの対応でよいため、重複対応の回避が可能となる。また、認定事業者側のメリットとして、提供に関する審議結果をもとにデータ提供が可能となり、審査工数の削減と提供までの期間短縮が可能となる。2回目のリスク評価の審議結果の共有を受けることで、提供後のデータの取り扱い、事例共有を得ることが可能となる。

以下の審査フローは代表研究者がこれまでのWGの議事録より書き起こした審査フローの概要である。(あくまでもWGで想定されているワークフローの全容を視覚的に把握するための資料であり、最終的な運用がこの通りに行われるというものではないことに留意されたい。)



## § エンジニア WG

### 全体的な方針

- ・次世代医療基盤法の趣旨に従い、認定事業者間のデータ授受は積極的に融通しあうことを前提とする。

### 1.医療情報の提供項目

#### 方針

- ・医療情報取扱事業者より収集した医療情報に関して、認定事業者間での共通データ項目を整理し、認定事業者間で共有される「データ定義書」を作成する。このデータ定義書は、審査委員会審議時に必要データ項目の確認、認定事業者契約書において契約対象となるデータ項目を指定・確認、認定事業者間でデータを交換する際の対象項目を指定するのに使用される。
- ・認定事業者において定義が異なる項目については変換定義をまとめる。
- ・オプトアウトを行使した方をデータから除外した状態でデータ交換を行う。

#### アウトカム

- ・LDI は MML、レセプト、DPC を、J-MIMO は電子カルテデータ、SS-MIX の標準ストレージをそれぞれ医療情報取扱事業者より授受しており、これらのデータソースを元にした「データ定義書」を構成した。なお、厚生労働省への報告時点ではデータ定義書は非公開情報としているが、利用者が認定事業者へ利用申請する際に利用したいデータを指定するための資料とできるようなバージョンを作成し、公開することを検討する。
- ・現時点では上記のデータソース以外についての検討は行っていない。

### 2.情報の提供形式

HL7、MML、CSV 等、元データは様々なフォーマットがあるが、認定事業者間でのデータの授受の際には、「データ定義書」に準拠した CSV ファイル形式で授受することとした。

### 3.提供方法

- ・医療データの授受手段をセキュリティ上で授受することとする。データ授受に際して 7-Zip の暗号化モード(AES256)によるデータ暗号化、及び SHA-256 のハッシュ値を授受し、受領したデータのハッシュ値の一致を確認する。
- ・暗号化のパスワードは別途安全なチャネルを利用して送付する。
- ・提供媒体は、基本的に可搬記録媒体とする。

#### 4.提供媒体の管理

- ・提供に使われた可搬型媒体は各々の認定事業者において適切に保管・管理する。
- ・確実なデータ交換と保管・管理をするために認定事業者間で共通のファイル命名規則を決定した。ファイル命名規則には年度、提供年月日、案件通番、提供元・提供先名称を含めることとし、ファイル名単独で最低限のファイルの素性確認ができることとした。

#### 5.データ精度の基準について

- ・現時点では、検査コード、検査名称、単位、値の表現形式等について、統一基準は作成せずに元データのまま授受し、受取側の認定事業者で利活用者の要望とあわせてデータクレンジングを行うこととする。
- ・データソースの種類によって項目が同じであっても意味合いや解釈が異なる可能性があり、データクレンジングやマッピングが必要になることがある。例えば MML 由来の情報、SS-MIX 由来の情報とでは診療科の項目は同じであっても解釈が異なることがある。各事業者が各データソースの特性を踏まえ、データ定義書に留意事項やデータクレンジングに必要な情報を記述することとする。

#### 6.提供したデータの保存期限

- ・受領側では提供された医療情報を使用後は削除する。
- ・提供側では個別契約、仕様書に定められた保存期間まで保存することとする。

#### 7.提供記録の管理方法

- ・データ提供側認定事業者は、受領側認定事業者に提供したデータ並びに提供記録を管理する。

- ・データ受領側認定事業者では、提供を受けたデータ及び利活用者に提供した匿名加工医療情報の提供記録を管理する。利活用者からの削除証明の取得後は提供側事業者に報告する。
- ・オプトアウトを行使した患者にかかる情報を認定事業者間で共有する。

なお、各々の記録については以下のガイドラインに準拠する。これは 2023 年 11 月時点のものであり、法律、ガイドラインが改正等されれば、随時更新、追加されるものである。

※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  
法律についてのガイドライン（次世代医療基盤法ガイドライン）上の  
該当箇所【令和 4 年 4 月改定】

- ①認定事業者間でデータを授受する際に残す記録  
2-2-15 他の認定匿名加工医療情報作成事業者との間での  
医療情報の授受（法第 25 条）
- ②認定事業者→利活用者に提供する際に残す記録について  
認定匿名加工医療情報作成事業者に対する医療情報の提供をした  
医療情報取扱事業者の記録事項（規則第 34 条第 1 項）
- ③医療機関からの問い合わせ対応に残す記録について  
8-1-2 医療情報の提供に係る記録の対象となる事項  
（規則第 34 条及び第 37 条）

## 8.依頼～提供までのスケジュール

データ授受に関して運用フローを作成することで依頼から提供までのスケジュールを見積もることができる。しかし、まだ実際にデータ授受が行われていないため、運用フローが確立していない。現時点ではデータ授受にかかるタスクを分解・整理した。今後の運用を通して運用フローとタスクの整理を行い、それにもとづいて定型業務への移行、スケジュールの策定ができるような体制としていく。

表 データ授受にかかるタスク

| No. | 概要タスク              | 実施主体         |                       |
|-----|--------------------|--------------|-----------------------|
|     |                    | 提供元<br>データ提供 | 提供先<br>データ受領<br>※一次受け |
| 1   | 利活用者からの依頼受領        | —            | ○                     |
| 2   | 一次受け⇔利活用者との要件調整    | —            | ○                     |
| 3   | 一次受け側での利用目的等審査委員会  | —            | ○                     |
| 4   | データ量の事前調査（提供申請）    | —            | ○                     |
| 5   | 提供元での利用目的等審査委員会    | ○            | —                     |
| 6   | 見積提示/個別契約書の提示      | ○            | —                     |
| 7   | 提供元でのデータ作成         | ○            | —                     |
| 8   | 提供元でのオプトアウト実施      | ○            | —                     |
| 9   | 提供先でのデータ受領         | —            | ○                     |
| 10  | データ受入検査            | —            | ○                     |
| 11  | 提供先でのデータクレンジング     | —            | ○                     |
| 12  | 提供先でのデータマート作成      | —            | ○                     |
| 13  | 名寄せ                | —            | ○                     |
| 14  | 利活用者に応じた医療データの匿名加工 | —            | ○                     |
| 15  | 利用目的等審査委員会の実施      | —            | ○                     |
| 16  | 匿名加工医療情報の提供        | —            | ○                     |

## 9.名寄せの仕様

認定事業者間で個別の名寄せの手法を確立している。名寄せの手法は共通化せず、各自で行うこととする。なお、認定事業者ごとに名寄せに使用する項目とアルゴリズムが異なることを確認した。

認定事業者 A が名寄せに使う項目

- ・生年月日、氏名（漢字・カナ）、性別、保険番号（保険者番号、記号、番号）、住所等

認定事業者 B が名寄せに使う項目

- ・患者 ID（ローカル ID）・医療機関 ID、姓名(カナ)、姓名(漢字)、生年月日、性別、電話番号、郵便番号、名(カナ)、保険番号（保険者番号、記号、番号）

## 10.授受データの期間

作成したデータ定義書に定義するデータ項目に関して保有期間・件数もあわせて管理し、データ交換を行う認定事業者間において共通して提供できる期間を特定できるようにする。その上で、利活用者と提供データの対象期間を協議、決定する。

## 11.文字コード

UTF-8 に統一し、非 UTF-8 コードで記述されたデータは UTF-8 に変換した上で管理する。

## 12. 匿名加工

- ・ 認定事業者間におけるデータの授受は匿名加工前のデータ形態で行う。
- ・ 匿名加工方法の共通化は現時点では行わない。理由として研究テーマによって配慮すべき事項、リスクが変化しており、定型的に匿名加工手法を決定するプロセスを確立できていないためである。
- ・ 利活用者と直接協議するデータ受領側認定事業者が最終的な匿名加工に関して決定、責任を負う。



# 認定事業者間データ連携の在り方に関する研究

---

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  
大規模データの利活用研究の加速のための研究（21AC1001）

2023/11/16

最終報告書

## 認定事業者間の突合・データ統合の検証 ～認定事業者間連携の在り方\_検討の経緯～

### 目的

次世代法第25条では認定事業者同士によるデータの融通・名寄を通したデータ間の突合処理を認めている。しかし、現時点でこの条項にもとづいた認定事業者の活動実績はなく、また突合可能であることを確認、担保されたうえでデータ収集・蓄積が進められているわけではない。

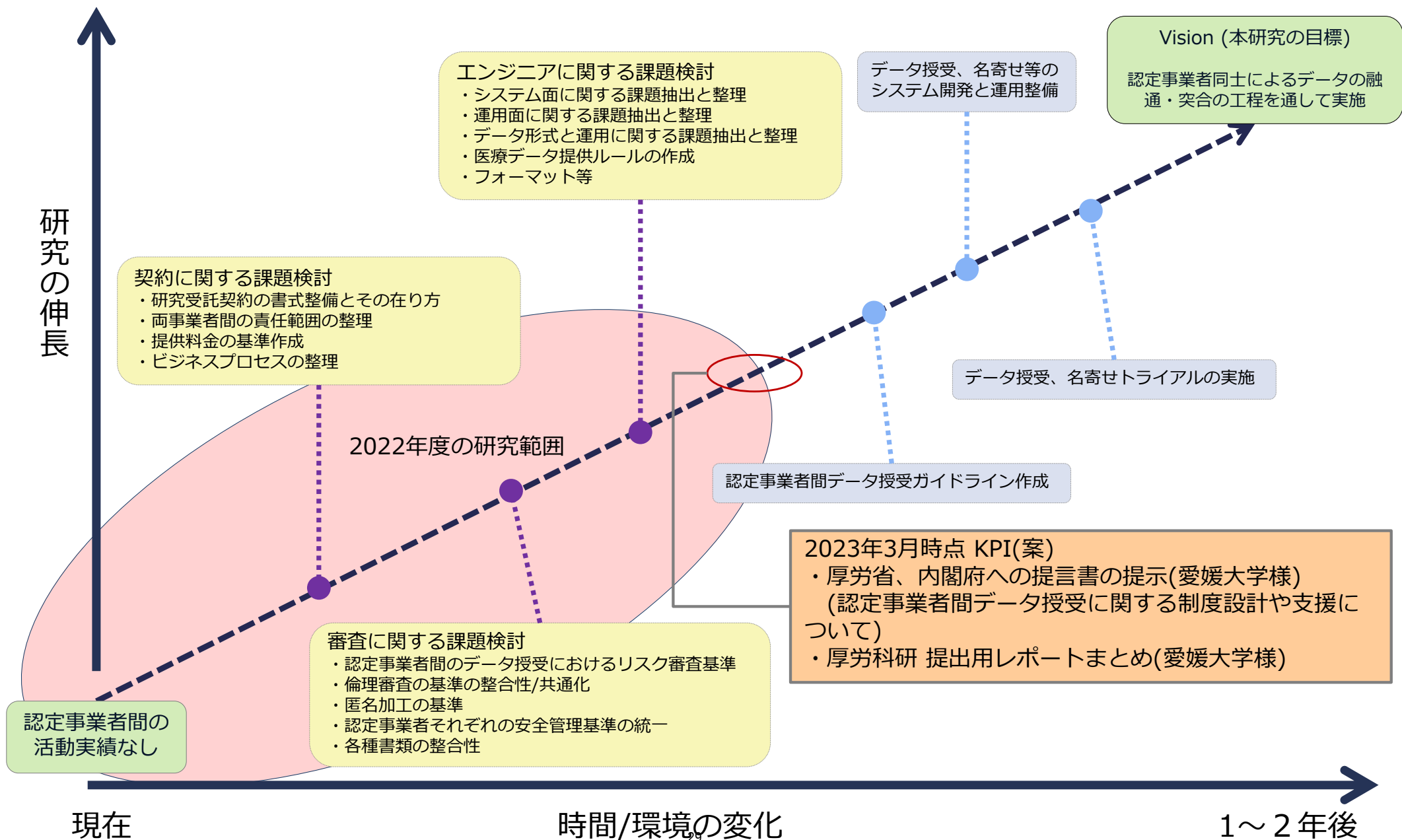
そこで、認定事業者同士によるデータの融通・突合を実際に全て通して実施することで、**データの融通・突合にかかる課題を洗い出し、また今後データ収集における留意点やデータの融通・突合に関する認定事業者の申し合わせ、ガイドラインの検討**を行う。

### 仮説

現在、認定事業者に医療情報を提供している協力医療情報取扱事業者は内閣府にて公表されている。比較的地理的に近い位置に所在する医療情報取扱事業者であれば、同じ患者が受診している（≡突合ができる対象者が存在する）可能性が高いと考える。

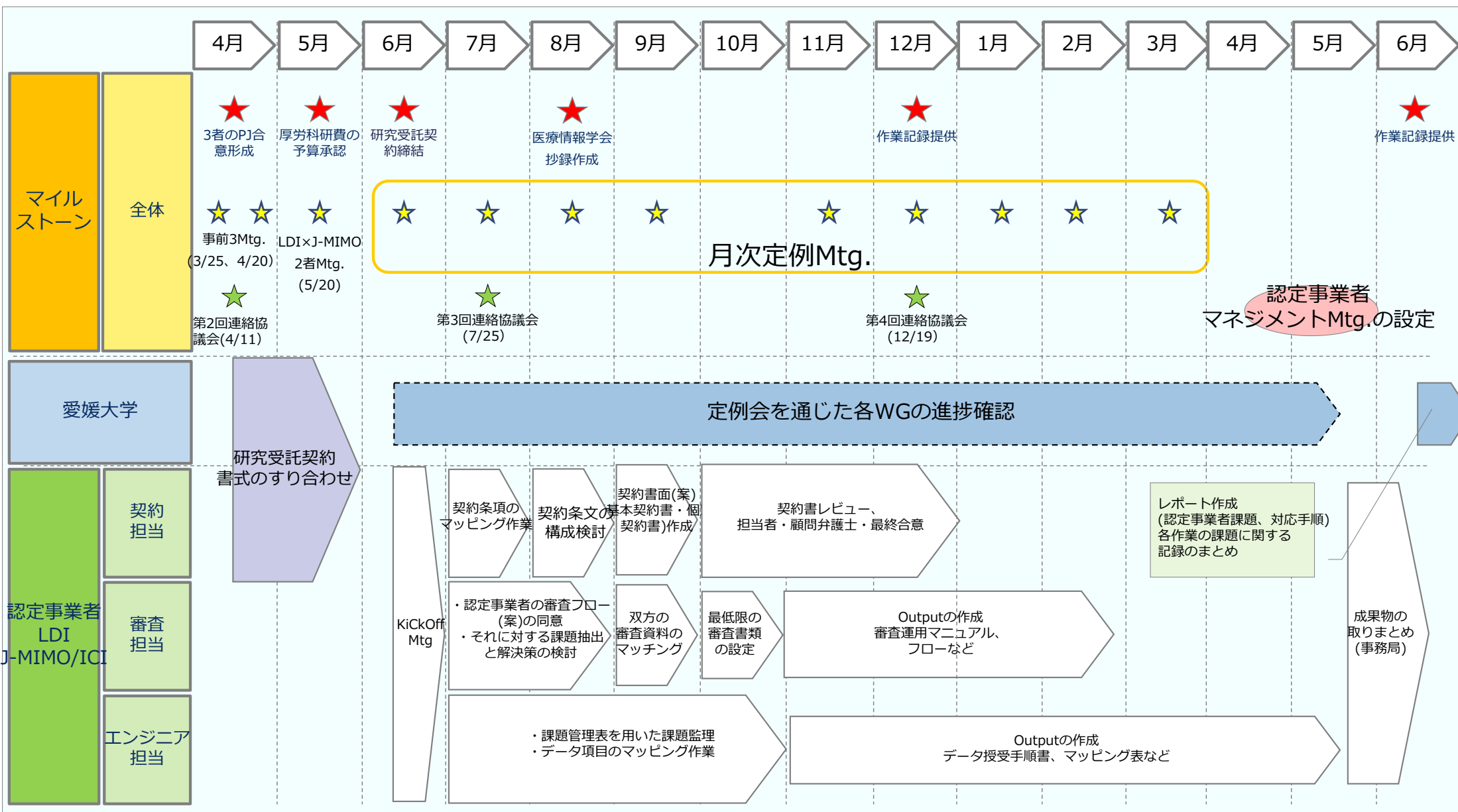
# 本プロジェクトを通じて目指すゴールの設定

## ～認定事業者間連携の在り方\_目標設定～



# 本プロジェクトを通じて目指すゴールの設定

## ～認定事業者間連携の在り方\_スケジュールと体制について～



# 全体定例会、各WGの開催実績

|      | 全体定例会 | 契約WG | 審査WG | エンジニアWG | 合計  |
|------|-------|------|------|---------|-----|
| 開催回数 | 8回    | 6回   | 6回   | 10回     | 30回 |

| No. | 開催日        | 時間       | 開催 ミーティング名  | ステータス   | 備考                      |
|-----|------------|----------|-------------|---------|-------------------------|
| 1   | 2022/6/22  | 9時半-10時半 | 第1回定例会      | 議事録完成   |                         |
| 2   | 2022/6/27  | 13時-14時  | 第1回審査WG     | 議事録完成   |                         |
| 3   | 2022/6/28  | 15時-16時  | 第1回エンジニアWG  | 議事録完成   |                         |
| 4   | 2022/6/29  | 15時-16時  | 第1回契約WG     | 議事録完成   |                         |
| 5   | 2022/7/11  | 17時-18時  | 第2回定例会      | 議事録完成   |                         |
| 6   | 2022/7/12  | 15時-16時  | 第2回エンジニアWG  | 議事録完成   | エンジニアWGは隔週火曜日15時より開催    |
| 7   | 2022/7/19  | 13時-14時  | 第2回審査WG     | 議事録完成   |                         |
| 8   | 2022/7/26  | 15時-16時  | 第3回エンジニアWG  | 議事録完成   |                         |
| 9   | 2022/8/2   | 16時-17時  | 第2回契約WG     | 議事録完成   |                         |
| 10  | 2022/8/9   | 15時-16時  | 第4回エンジニアWG  | 議事録完成   |                         |
| 11  | 2022/8/23  | 17時-18時  | 第3回定例会      | 議事録完成   | 定例会は毎月第4週目の火曜日17時～18時開催 |
| 12  | 2022/8/23  | 15時-16時  | 第5回エンジニアWG  | 議事録完成   |                         |
| 13  | 2022/8/30  | 10時-11時  | 第3回審査WG     | 議事録完成   |                         |
| 14  | 2022/9/6   | 15時-16時  | 第6回エンジニアWG  | 議事録完成   |                         |
| 15  | 2022/9/8   | 10時-11時  | 第3回契約WG     | 議事録完成   |                         |
| 16  | 2022/9/20  | 15時-16時  | 第7回エンジニアWG  | 議事録完成   |                         |
| 17  | 2022/9/20  | 17時-18時  | 第4回定例会      | 議事録完成   |                         |
| 18  | 2022/10/4  | 15時-16時  | 第8回エンジニアWG  | 議事録完成   |                         |
| 19  | 2022/10/6  | 10時-11時  | 第4回契約WG     | 議事録完成   |                         |
| 20  | 2022/11/10 | 17時-18時  | 第5回契約WG     | 議事録完成   |                         |
| 21  | 2022/11/22 | 17時-18時  | 第5回定例会      | 議事録完成   |                         |
| 22  | 2022/12/8  | 14時-15時  | 第6回契約WG     | 議事録完成   |                         |
| 23  | 2022/12/13 | 9時-10時   | 第4回審査WG     | 議事録完成   |                         |
| 24  | 2022/12/20 | 17時-18時  | 第6回定例会      | 議事録完成   |                         |
| 25  | 2023/1/13  | 13時-14時  | 第9回エンジニアWG  | 議事録完成   |                         |
| 26  | 2023/1/17  | 10時-11時  | 第5回審査WG     | 議事録完成   |                         |
| 27  | 2023/1/24  | 17時-18時  | 第7回定例会      | 荒天の為、中止 |                         |
| 28  | 2023/2/2   | 15時-16時  | 第10回エンジニアWG | 議事録完成   |                         |
| 29  | 2023/2/16  | 10時-11時  | 第6回審査WG     | 議事録完成   |                         |
| 30  | 2023/2/28  | 17時-18時  | 第8回定例会      | 議事録完成   |                         |

### 本WGの目的

他の認定匿名加工医療情報作成事業者との間での医療情報の授受（法第25条）に基づき、次世代医療基盤法に基づき基本的なルールを定めるものとする。

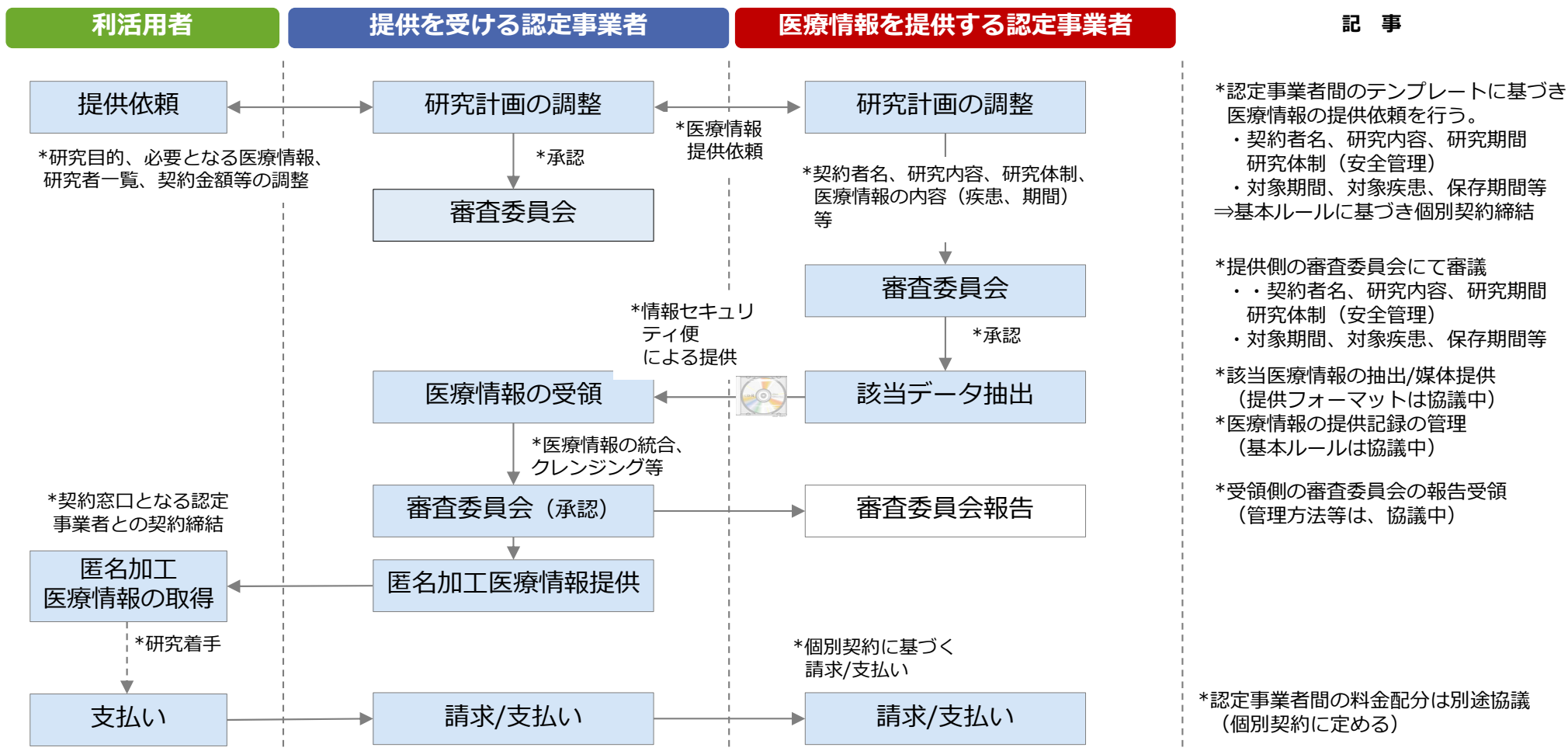
**本WGの取り組みは、2023年度の運用開始を目標に検討を進める。**

### 認定事業者間の医療情報の授受に関する基本的な考え方

- ✓ **利活用者との契約、審査等のルールは、窓口となる認定事業者のルールに基づき実施する。**
  - 医療情報の提供に関する契約金額、安全管理基準遵守等の基準、契約終了時の匿名加工医療情報の削除、等々
  - 医療情報の提供に関する審査委員会は、窓口とする認定事業者で実施する。（匿名加工医療情報の作成も提供を受ける側で実施）  
（提供側の認定事業者の審査委員会は、窓口とする認定事業者への提供の報告を基に実施し記録に残す。）
- ✓ **認定事業者間には、データ提供形式、運用基準等を定めた「基本契約＋個別契約」にて実施する。**
  - 認定事業者間における契約金額、データ授受の方式、データ提供範囲（疾患、期間等）、利活用者の安全管理基準等を定義
  - 認定事業者間のデータ授受に関する記録は、法律に基づき両事業者で管理する。（管理方法、期間は今後協議する。）
- ✓ **実現方式は、次世代医療基盤法に準拠し、効率性、迅速性、正確性を基準に検討する。**

**現在、各WGにて具体的な検討を進めており、今後、検討の過程で見直しとなる可能性がある。**

### 基本的な考え方に基づいた業務フロー（案）



**契約WGでは上記のフローに基づき、認定事業者間の契約事項を定める。**

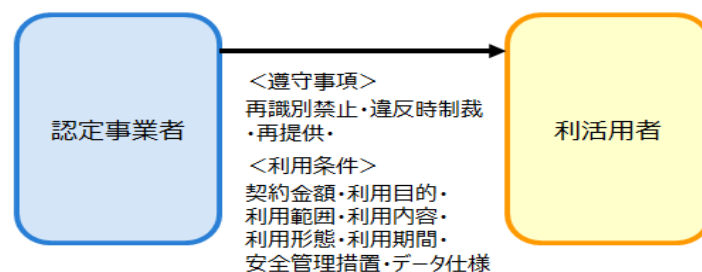
### 基本方針

- ・ 認定事業者と利活用者間での契約書面は、各認定事業者の現行の契約書を準拠する。
- ・ 認定事業者間の契約形態は、認定事業者間のデータ授受に係る基本的なルールを定めた基本契約書と、案件毎の詳細を定めた個別契約書で運用をしていくこととする。

### 認定事業者間\_契約の在り方について

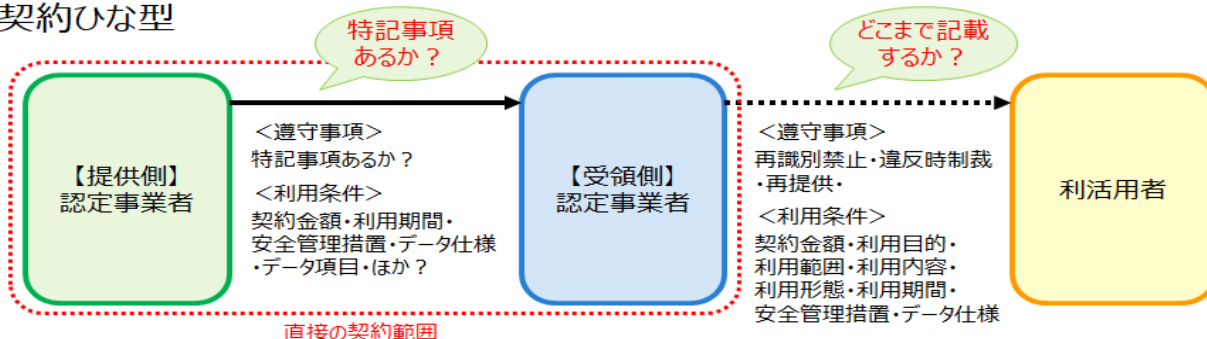
共通の契約書の検討に当たり、まずは、各認定事業者間の利活用者との契約書を比較検討を実施した。

#### 既存契約ひな型



- ✓ 各認定事業者と利活用者間の契約条文に大きなGapはないことを確認した
- ✓ 認定事業者と利活用者との契約/審査等は現行のルールに準拠することとする

#### 新契約ひな型



次世代医療基盤法で定められた事項を遵守されることを前提に“認定事業者間”の契約事項を定める

### <遵守事項の記載方針>

| 項目                 | 認定事業者間 | 対利活用者  |
|--------------------|--------|--------|
| 匿名加工医療情報である旨の明示    | ×      | ○ 基本契約 |
| 目的外利用の禁止           | ○ 基本契約 | ○ 基本契約 |
| データ取扱者の限定          | ○ 基本契約 | ○ 基本契約 |
| 安全管理措置の管理          | ○ 基本契約 | ○ 基本契約 |
| 再識別の禁止             | ×      | ○ 基本契約 |
| 違反時の制裁             | ○ 基本契約 | ○ 基本契約 |
| 再提供先の管理            | ×      | ○ 基本契約 |
| 機密保持               | ○ 基本契約 | ×      |
| 終了時のデータ取扱い         | ○ 基本契約 | ○ 基本契約 |
| 仕様と異なるデータがあった場合の対応 | ○ 基本契約 | ○ 基本契約 |
| 審査内容の遵守 ①法の目的      | ×      | ×      |
| 審査内容の遵守 ②科学的妥当性    | ×      | ×      |
| 審査内容の遵守 ③公表時の配慮    | ×      | ×      |
| 審査内容の遵守 ④利益の収受     | ×      | ×      |
| 審査内容の遵守 ⑤その他       | ×      | ×      |

次世代医療基盤法に準拠する前提であることを定義する

次世代医療基盤法に準拠する前提であることを定義する

認定事業者間のルールを定義する

認定事業者の契約に準拠する

## <利用条件の記載方針>

| 項目                         | 認定事業者間 | 対利活用者                            |
|----------------------------|--------|----------------------------------|
| 契約金額                       | ○ 個別契約 | ×                                |
| 利用目的                       | ○ 個別契約 | ×                                |
| 利用範囲<br>(責任者・研究者)          | ×      | △ 責任者のみ記載<br>(共同研究者／委託先も責任者のみ記載) |
| 利用範囲 (場所)                  | ×      | ×                                |
| 利用期間                       | ○ 個別契約 | ×                                |
| データ仕様<br>(種類、形式、対象期間)      | ○ 個別契約 | ×                                |
| データ提供方法<br>(通信／媒体送付／オンサイト) | ○ 個別契約 | ×                                |
| 終了時のデータ取扱い                 | ×      | ×                                |
| 安全管理措置                     | ×      | ×                                |
| 審査内容の遵守<br>④利益の収受          | ×      | ×                                |
| 第三者認証の取得状況                 | ×      | ×                                |
| 提供履歴                       | ×      | ×                                |

# 認定事業者間の契約条文の構成（案）

契約WG

## 認定事業者間\_基本契約

| NO.          | 契約条項 | 契約条文の内容                                                                                                                               |
|--------------|------|---------------------------------------------------------------------------------------------------------------------------------------|
| 1 前文         |      | J-MIMO、LDIとは、次世代医療基盤法第25条の定めに従い自ら収集した医療情報を相手側の依頼に基づき、該当する医療情報を提供することに関して、次の通り基本契約を締結する                                                |
| 2 利用条件、契約金額等 |      | 都度契約する個別契約に従う                                                                                                                         |
| 3 用語の定義      |      | 医療情報：<br>匿名加工医療情報：<br>.....                                                                                                           |
| 4 個別契約       |      | 案件毎に個別契約を締結することを定める                                                                                                                   |
| 5 準拠法/順守事項   |      | 次世代医療基盤法の定めに従い、認定事業者、利活用者共に遵守させることを定義する<br>・利活用者に求める安全管理基準<br>・目的外利用の禁止<br>・再識別の禁止<br>・違反時の制裁<br>・終了時の匿名加工医療情報/統計情報の扱い<br>・その他（項目を列挙） |
| 6 委託側の責務     |      | 委託側の責務（必要な情報の提供、受領した医療情報のクレンジング、利活用者に対する匿名加工処理等）を定義する                                                                                 |
| 7 受託側の責務     |      | 受託側の責務（審査委員会の実施、必要な医療情報の抽出・提供、提供した医療情報の記録義務等）を定義する。                                                                                   |

| NO.                | 契約条項 | 契約条文の内容                                                              |
|--------------------|------|----------------------------------------------------------------------|
| 6 機密保持             |      | 相手方から提供を受ける各種ドキュメント、利活用者等の機密保持を遵守する                                  |
| 7 知的財産の扱い          |      | 該当案件に関する知的財産権の扱い等を定義する                                               |
| 8 免責事項             |      | 該当案件に対する免責事項の扱いに等を定義する                                               |
| 9 権利義務等の譲渡禁止       |      | 本基本契約の第三者への譲渡の禁止を定義する                                                |
| 10 契約金額の支払い        |      | 認定事業者間での契約金額の支払い基準（請求のタイミング、支払いサイト等）を定義する                            |
| 11 端数整理            |      | 請求金額の端数整理の考え方を定義する                                                   |
| 12 支払い遅延損害金        |      | 支払い遅延損害金の扱いを定義する                                                     |
| 13 契約終了時の医療情報の取り扱い |      | 契約終了時での認定事業者間での医療情報の取り扱いに関して定義する                                     |
| 14 契約解除            |      | 基本契約の契約解除の条件を定義する                                                    |
| 15 反社会的勢力との関係排除    |      | 反社勢力との関係排除の誓約（必要性？）                                                  |
| 16 契約期間の延長         |      | 基本契約の自動契約延長を定義する                                                     |
| 17 損害賠償            |      | 損害賠償の考え方を定義する<br>不可抗力等の定義も合わせて行う                                     |
| 18 管轄裁判所           |      | 本契約に本契約に関る一切の紛争に関する提訴及び裁判所における調停は、被告の所在地を管轄する地方裁判所を第一審の専属的合意管轄裁判所とする |
| 19 協議              |      | 本契約に定めのない事項その他本契約の条項に関し疑義を生じたときは、甲乙協議のうえ円満に解決を図るものとする                |
| 37 20 存続条項         |      | 契約終了後も継続される条文を定義する                                                   |

### 認定事業者間\_個別契約

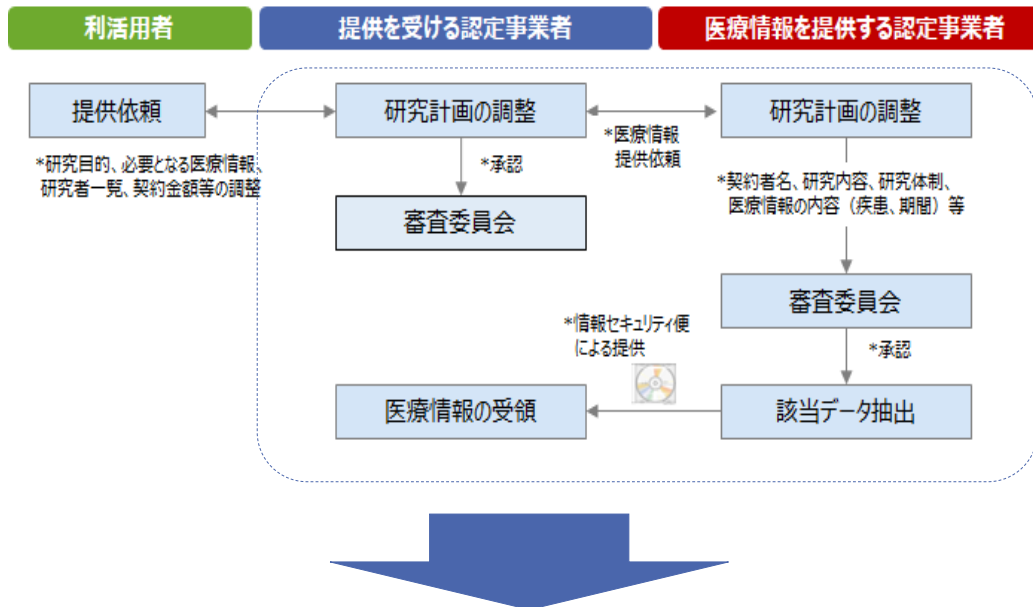
| NO.                                                    | 契約条項 | 契約条文の内容                                                                                           |
|--------------------------------------------------------|------|---------------------------------------------------------------------------------------------------|
| 1 前文                                                   |      | J-MIMO、LDIとは、次世代医療基盤法第25条の定めに従い自ら収集した医療情報を相手側の依頼に基づき、該当する医療情報を提供することに関して、基本契約●●状に基づき次の通り個別契約を締結する |
| 2 利用条件、契約金額等                                           |      | 利用条件：別添「仕様書」のとおり<br>契約金額：〇〇〇円（消費税相当額を含む。）                                                         |
| 3 利活用者の利用目的                                            |      | 別添「仕様書」のとおりとする。                                                                                   |
| 4 利用範囲（責任者/研究者）                                        |      |                                                                                                   |
| 5 利用期間                                                 |      |                                                                                                   |
| 6 保存期間                                                 |      |                                                                                                   |
| 7 対外発表の有無                                              |      |                                                                                                   |
| 8 提供データの仕様<br>・データの種類<br>・形式<br>・対象期間<br>・データ項目<br>・…… |      | ※フォーマットは、審査WG、エンジニアWGとの協議により決定する                                                                  |
| 9 データ提供の方法（媒体）                                         |      |                                                                                                   |
| 10 医療情報の納期                                             |      |                                                                                                   |
| 11 ……                                                  |      |                                                                                                   |

「基本契約」「個別契約」の各案は別添

認定事業者間のデータ交換の実施については、  
予め定めた「仕様書」「データ定義書」をベースに個別契約の締結を行う。

認定事業者間 データ定義書（共通）（案）

| 仕様書 |                |                                                                                                                                                               |
|-----|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | 本データの<br>利用目的  | 以下を目的とする研究開発のために、匿名加工医療情報取扱事業者に対し、本データに由来する匿名加工医療情報を提供する。<br>＜研究開発の目的＞<br>●●                                                                                  |
| 2   | 本データの<br>利用範囲  | 本データの利用範囲は、受領側認定事業者又は認定医療情報等取扱受託事業者の認定事業医療情報等取扱者とする。<br>また、本データに由来する匿名加工医療情報の利用範囲は、以下の者を研究開発の責任者として、匿名加工医療情報取扱事業者が受領側認定事業者に適正に申請した者とする。<br>＜研究開発の責任者＞<br>●●●● |
| 3   | 本データの<br>利用期間  | ●●年●●月●●日から<br>●●年●●月●●日まで                                                                                                                                    |
| 4   | 本データの<br>データ仕様 | ＜データ種類＞<br>●●<br>＜データ項目＞<br>●●<br>＜データ形式＞<br>●●<br>＜データ対象期間＞<br>●●年●●月●●日から●●年●●月●●日まで                                                                        |
| 5   | 本データの<br>提供方法  | ●●                                                                                                                                                            |
| 6   | 本データの<br>提供期限  | ●●年●●月●●日まで                                                                                                                                                   |
| 7   | その他            | ●●                                                                                                                                                            |



大枠の運用フローに基づき、研究計画の調整から認定事業者間での医療情報の提供/受領は、左図に掲示している「データ定義書」にて対応する。

- ・ 認定事業者間の個別契約
- ・ 各認定事業者の審査委員会の審議資料
- ・ 認定事業者間の医療情報の提供/受領

### 基本方針

- ・ 審査を2回に分けて、それぞれの認定事業者で審査を行う。
- ・ 利活用目的に対する審議を中心とし、基本的には相手方の判断基準を尊重する方針。
- ・ 認定事業者間のデータ授受は、双方が認定機関であることから、インフラ部分（事業体制・安全管理措置など）については審議対象から免除し、あくまで**利活用目的に対する審議を中心**とし、受領側認定事業者の審査結果の提出をもって、データ授受を許可することが妥当。

### 2段階審査の概要

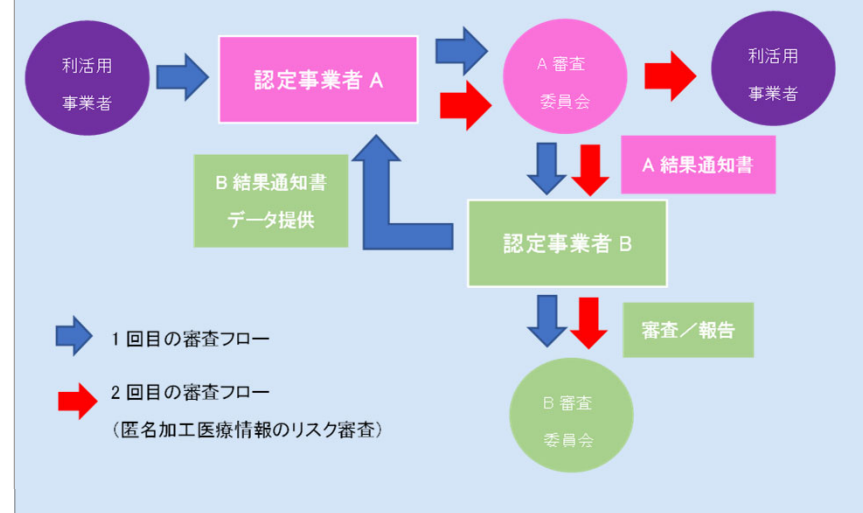
#### < 1回目の審査 >

- 利活用者の申請を受けた認定事業者 A が、利活用の目的、詳細及び事業体制等をもとに、認定事業者 B からのデータ受領の是非を審査し、認定事業者 B にデータ提供を申請。
- 認定事業者 B は、認定事業者 A の審査結果を参考に、自社の審査を実施し、認定事業者 A にデータを提供。

#### < 2回目の審査 >

- 認定事業者 A は、自社保有データの追加・突合を行った後、再度、匿名加工に焦点を当てた審査を実施し、認定事業者 B に報告。利活用者へ匿名加工データを提供。
- 認定事業者 B は、認定事業者 A の（2回目の）審査結果を自社の審査委員会に報告。

認定事業者間のデータ授受における審査フロー（イメージ図）



### <LDI> (2023年6月現在)

審査方法： 審議は2回に分けて実施

- 1回目：リスク審査（計画書中心の審議）
- 2回目：提供前審査（匿名化ルール・処理結果の確認）

審査実績： 匿名加工医療情報 6件

統計情報 18件

### <J-MIMO> (2023年6月現在)

審査方法： 審議は1回にまとめて実施（※2段階に分ける事も可能）

審査実績： 匿名加工医療情報 2件

統計情報 0件

### ＜利活用者のメリット＞

- 申請を行った認定事業者への対応だけでよいいため、重複対応を回避できる。

### ＜認定事業者のメリット＞

- データ提供側の認定事業者は、データ受領側の認定事業者の審査結果をもとにデータ提供が可能となり、審査工数の削減、データ提供までの期間を短縮できる。
- データ受領側の認定事業者が2回目の審査を実施することで、匿名加工の適正性の確認、データ品質や安全性を担保できる。
- 2回目のリスク評価の審査結果の報告を受けることで、データ提供側の認定事業者でも、提供後のデータの取扱い、事例の共有を得られる。

#### <申請フォーマット>

- 基本方針として、認定事業者間の審査資料の共有は最低限の範囲とし、審査資料として各認定事業者のフォーマットを利用する。

（注） 案件において特有の審査項目や、相手方の認定事業者が求める資料については、該当するものが出てくれば、追加作成することも検討。

- 利活用者から認定事業者（認定事業者A）への申請フォーマットにおいて、審査資料を他の認定事業者（認定事業者B）へ共有することの許諾を得ることを検討する  
当該フォーマット（NDA含む）については、各認定事業者の責任において、ひな型を作成する等によって対応する（統一のひな型等は作成しない）。  
また、当該フォーマット（NDA含む）により合意した内容は、提供側認定事業者（認定事業者B）には共有しない。

### 共有する資料

- 研究計画
- データに関する仕様書（テーブル定義書）
- 匿名加工の仕様に関する資料

※ 「データに関する仕様書（テーブル定義書）」については、エンジニアWGの成果物を利用する想定。

### 共有不要な資料（例）

- 匿名加工の適正性・安全性を担保する説明資料
- 研究者の名簿（実務経歴書等を含む）
- COIに関する資料

➡ 上記「共有する資料」について、両認定事業者のひな型を照合し、必要な項目が含まれていることを確認をおこなった。

※ 両認定事業者の審査書類と対応表は、別添資料①を参照

# 審査において共有する資料

審査WG

| 書類名                         | 項目                          |
|-----------------------------|-----------------------------|
| 研究計画                        | 研究の名称                       |
|                             | 申請機関の名称                     |
|                             | 研究実施体制（共同研究機関を含む機関名・代表者名）   |
|                             | 研究の目的                       |
|                             | 研究の方法（データの利用形態等を含む）・スケジュール  |
|                             | 研究期間（データ利用期間）               |
|                             | 研究結果の公表の有無及び倫理的配慮           |
| データに関する仕様書<br>匿名加工の仕様に関する資料 | 利用データの種類（カルテ・レセプト等）         |
|                             | 利用データの対象期間                  |
|                             | 利用データの範囲（特定の集団を抽出する場合の抽出基準） |
|                             | データ項目一覧                     |
|                             | データ項目名                      |
|                             | データ型                        |
|                             | 分類（識別子、準識別子、静的属性等）          |
|                             | 匿名加工の有無                     |
|                             | 匿名加工の手法                     |

## 共有不要な資料

| 書類名               | 項目                          |
|-------------------|-----------------------------|
| 安全性を担保する説明資料      | 研究実施場所                      |
|                   | 安全管理措置                      |
| 匿名加工の適正性を担保する説明資料 | 匿名加工結果の評価                   |
| 研究者の名簿            | 研究者の名簿                      |
| COIに関する資料         | 研究開発に係る金銭その他の利益の収受及びその管理の方法 |

# 変更申請の取扱い

## 審査WG

### 共有する資料

| 書類名                         | 項目                          |
|-----------------------------|-----------------------------|
| 研究計画                        | 研究の名称                       |
|                             | 申請機関の名称                     |
|                             | 研究実施体制（共同研究機関を含む機関名・代表者名）   |
|                             | 研究の目的                       |
|                             | 研究の方法（データの利用形態等を含む）・スケジュール  |
|                             | 研究期間（データ利用期間）               |
|                             | 研究結果の公表の有無及び倫理的配慮           |
| データに関する仕様書<br>匿名加工の仕様に関する資料 | 利用データの種類（カルテ・レセプト等）         |
|                             | 利用データの対象期間                  |
|                             | 利用データの範囲（特定の集団を抽出する場合の抽出基準） |
|                             | データ項目一覧                     |
|                             | データ項目名                      |
|                             | データ型                        |
|                             | 分類（識別子、準識別子、静的属性等）          |
|                             | 匿名加工の有無                     |
|                             | 匿名加工の手法                     |

### 共有不要な資料

| 書類名               | 項目                          |
|-------------------|-----------------------------|
| 安全性を担保する説明資料      | 研究実施場所                      |
|                   | 安全管理措置                      |
| 匿名加工の適正性を担保する説明資料 | 匿名加工結果の評価                   |
| 研究者の名簿            | 研究者の名簿                      |
| COIに関する資料         | 研究開発に係る金銭その他の利益の収受及びその管理の方法 |

元々の審査項目に変更があった場合、申請を受けた認定事業者での取り扱いは、基本的に当初より審査対象になっていた**重要な項目は、変更申請を要する**。

一方、審査対象外の項目に関する変更については、不要とする。

尚、変更申請があった場合の、**手続きは各認定事業者の規程に従う**。

相手方認定事業者への)変更申請が必要なケースとして、「委託先の追加」は含まれるが、既に申請済みの組織内における研究者の異動に伴う(研究者)変更は申請不要とする。

### 共有する資料

| 書類名                         | 項目                          |
|-----------------------------|-----------------------------|
| 研究計画                        | 研究の名称                       |
|                             | 申請機関の名称                     |
|                             | 研究実施体制（共同研究機関を含む機関名・代表者名）   |
|                             | 研究の目的                       |
|                             | 研究の方法（データの利用形態等を含む）・スケジュール  |
|                             | 研究期間（データ利用期間）               |
|                             | 研究結果の公表の有無及び倫理的配慮           |
| データに関する仕様書<br>匿名加工の仕様に関する資料 | 利用データの種類（カルテ・レセプト等）         |
|                             | 利用データの対象期間                  |
|                             | 利用データの範囲（特定の集団を抽出する場合の抽出基準） |
|                             | データ項目一覧                     |
|                             | データ項目名                      |
|                             | データ型                        |
|                             | 分類（識別子、準識別子、静的属性等）          |
|                             | 匿名加工の有無                     |
|                             | 匿名加工の手法                     |

利活用者、受領側認定事業者を問わず、逸脱事象が発生した際には、制度への信頼性にも関わる事を踏まえ、「共有する資料」に含まれる重要な項目のみならず、審査対象に含まれない項目についても、原則として相手方事業者に報告すべきものとする。

（審査委員会にも報告するか否かは、各認定事業者の規程・判断に従う。）

（受領側認定事業者による逸脱例）

匿名加工のミス、申請（承認）されていない項目の提供、申請（承認）された項目の提供もれ

### 共有不要な資料

| 書類名               | 項目                          |
|-------------------|-----------------------------|
| 安全性を担保する説明資料      | 研究実施場所                      |
|                   | 安全管理措置                      |
| 匿名加工の適正性を担保する説明資料 | 匿名加工結果の評価                   |
| 研究者の名簿            | 研究者の名簿                      |
| COIに関する資料         | 研究開発に係る金銭その他の利益の収受及びその管理の方法 |

### 打合せの経緯

・エンジニアWGでは、認定事業者間におけるデータ授受を行うにあたり大きく二つの活動を通じて、WGの運営を進めた。一つは課題管理表を基にした認定事業者間のデータ授受に伴う課題などの協議。もう一つはデータ定義書の作成である。

### データ定義書の基本方針

- ・医療情報取扱事業者より収集した医療情報について認定事業者間の共通データ項目を整理する。（性別/転記事由などのテーブル情報を含む。）
- ・認定事業者間でのデータ定義書（共通）を用い各WG共通の用語として整理する。
- ・本作業は「エンジニアWG」が中心となって認定事業者間共通のデータ定義書を作成する。（認定事業者増の際は、本定義書をベースにマッピング作業を行う。）

### 認定事業者間 データ定義書（共通）

| NO. | 共通データ項目 | J-MIMO | LDI | 記事       |
|-----|---------|--------|-----|----------|
| 1   | 患者ID    | ○      | ○   |          |
| 2   | 所在地（住所） | ○      | ○   |          |
| 3   | 性別      | ○      | ○   | 変換テーブルあり |
| 4   | 生年月日    | ○      | ○   |          |
| 5   | 国籍      | ○      | ○   | 変換テーブルあり |
| 6   | 疾患名     | ○      | ○   |          |
| 7   | 疾患コード   | ○      | ○   | ICD-10   |
| 8   | 疾患開始日   | ○      | ○   | yyyymmdd |
| 9   | 疾患終了日   | ○      | ○   | yyyymmdd |
| 10  | 転記事由    | ○      | ○   | 変換テーブルあり |
| XX  | :       |        |     |          |
| XX  | :       |        |     |          |
| XX  | :       |        |     |          |

Image

### データ定義書の活用範囲

- ・審査委員会審議資料（必要データ項目）
- ・認定事業者契約書（個別契約書の添付資料）
- ・認定事業者間データ交換（共通のフォーマット）

### 医療情報のクレンジングの考え方について

認定事業者間の医療情報交換に関しては、受領側認定事業者のルールに基づきデータのクレンジングを行うこととする。なお、データクレンジングに必要な情報は、データ定義書作成において共有する。

### 医療情報の保存期間/提供記録の管理方法の考え方について

仕様書

|   |                |                                                                                                                                                               |
|---|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | 本データの<br>利用目的  | 以下を目的とする研究開発のために、匿名加工医療情報取扱事業者に対し、本データに由来する匿名加工医療情報を提供する。<br><研究開発の目的><br>●●                                                                                  |
| 2 | 本データの<br>利用範囲  | 本データの利用範囲は、受領側認定事業者又は認定医療情報等取扱受託事業者の認定事業医療情報等取扱者とする。<br>また、本データに由来する匿名加工医療情報の利用範囲は、以下の者を研究開発の責任者として、匿名加工医療情報取扱事業者が受領側認定事業者に適正に申請した者とする。<br><研究開発の責任者><br>●●●● |
| 3 | 本データの<br>利用期間  | ●●年●●月●●日から<br>●●年●●月●●日まで                                                                                                                                    |
| 4 | 本データの<br>データ仕様 | <データ種類><br>●●<br><データ項目><br>●●<br><データ形式><br>●●<br><データ対象期間><br>●●年●●月●●日から●●年●●月●●日まで                                                                        |
| 5 | 本データの<br>提供方法  | ●●                                                                                                                                                            |
| 6 | 本データの<br>提供期限  | ●●年●●月●●日まで                                                                                                                                                   |
| 7 | その他            | ●●                                                                                                                                                            |

認定事業者間の医療情報交換に関しては、個別契約/仕様書の定義に基づき、審査委員会/提供側認定事業者の提供範囲、保存期間等を定義する。

#### 1.受領側認定事業者

- 提供側認定事業者から提供を受けた医療情報を含め利活用者に提供した匿名加工医療情報並びに提供記録を保持する。
- \*保存期間終了後は、利活用者からの削除証明を取得し、提供側認定事業者に報告する。

#### 2.提供側認定事業者

- 受領側認定事業者へ提供した医療情報並びに提供記録のみ<sup>49</sup>を管理する。

## データカタログの共有検討

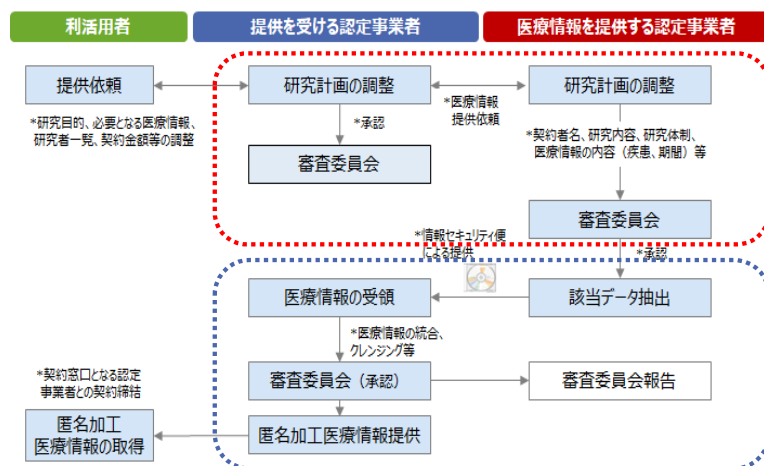
### データカタログの共有検討

**データ定義書 = データカタログ**として、両事業者間の共通バイブルとして保持する。

\*改定時等は、認定事業者連絡協議会にて検討する。

\*対外対応（広報）については、Topミーティングにて整理する。

### データ授受の範囲の検討



①受領側認定事業者から前述の「個別契約(仕様書)」を受領する。

-研究目的、利用範囲、対象疾患領域、必要となるデータ項目、データ利用期間、データ保存期間等を定義する。

（審査委員会で根拠となる資料）

②両事業者間の審査委員会にて審議し承認を得る。

③提供側認定事業者の承認が得られた際は、仕様書の定義に基づき医療情報（顕名）を指定のルールに基づき提供する。

（名寄せ、クレンジング等は受領側認定事業者で実施する。）

④受領側認定事業者で匿名加工医療情報を作成し、審査委員会の承認を得て利活用者に提供する。

## 次世代医療基盤法 認定事業者にかかる施策等への提言

2024 年 5 月 11 日

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） 大規模データの利活用研究の加速のための研究（21AC1001）研究班は、令和 4 年度から 6 年度にかけて AI や RWD 研究において、医療分野の研究開発をささえるために医療分野の研究開発に資するための匿名加工医療情報及び仮名加工医療情報に関する法律（以下、次世代医療基盤法）下の認定事業者の活用に関して検討いたしました。この取り組みを踏まえて、認定事業者を活用した研究に関して取り組むべき施策について提言致します。

### 背景

医療情報を円滑に利活用する社会的仕組みとして、次世代医療基盤法が平成 30 年に施行されました。その結果、令和元年から次世代医療基盤法に基づいて認定された複

数の認定匿名加工医療情報作成事業者（以下、認定事業者）から匿名加工医療情報が利活用に提供され、ICT・AI 技術を活用した研究や開発が進展することが期待されています。しかし本調査が始まった当初は認定事業者がまだ事業を開始したばかりであり、AI を用いた医学研究の事例はほとんどありませんでした。機械学習は AI 研究で主要な技術ですが、データから傾向を分析して有益な判断ができるモデルを構築するために多くの情報を要求する傾向があります。一方、個人の特定を防ぐために匿名加工するには、個人の特定につながる情報が削除されます。つまり、機械学習は多くの情報を必要とする一方で、匿名化によって情報が減少しトレードオフの関係が生じます。患者のプライバシーを保護しつつ、十分なデータを提供するという難しいバランスを取ることが要求されます。そのため、匿名加工医療情報と機械学習を用いた研究の実現性には懐疑的な見方もありました。

そこで本研究では、認定事業者から匿名加工医療情報を提供いただき、複数の研究者によって AI 技術を用いた研究を実施し、認定事業者と匿名加工医療情報を用いた研究を実施する際の課題を明らかにすることを目的としました。また、認定事業者を利用する可能性のある者からアンケート調査、および海外の認定事業者の調査も実施して、認定事業者を利用した研究を加速するために必要な取り組みについて検討することとしました。これらの調査研究に関する詳細は厚生労働科学研究成果データベース等で報告しております。また、研究期間中に次世代医療基盤法の改正法が公布されま

した。本稿では、これらの社会的状況・成果等を踏まえて、研究班として認定事業者の利活用を促進するために必要と思われる施策について提言いたします。

## 今後とるべき施策の提言

### 提言 1:

次世代医療基盤法の認知度を高めるために啓発活動を行うこと。

地方公共団体や学術団体へのアンケートを通じて、次世代医療基盤法や認定事業者に関する認知度が全体的に低いことが明らかになりました[1]。さらに、この認知度の低さに関連して、医療等情報の提供にむけた協力体制や準備、そして RWD 研究の支援に関して具体的な検討が不十分であることも判明しました。啓発のプログラムでは、次世代医療基盤法や認定事業者への認知度を高めるための取り組みや、認定事業者の利活用を支援するためのコンテンツ開発が求められます。後者については、利活用者が認定事業者から提供される匿名加工医療情報の詳細について、どのような種類のデータが提供され、どのような研究に役立てることができるのかを検討しやすいように、認定事業者が保有・提供している医療等情報に関する具体的な情報を開示することが望まれます。

## 提言 2:

**認定事業者へ医療等情報を提供する事業者の負担軽減を検討すること。**

事業者が認定事業者へ医療等情報を提供するにあたって地方公共団体、学術関係者から寄せられた要望の多くは、認定事業者への医療情報の提供にかかる患者への通知作業の代行・支援への要望でした。また、最も大きな協力動機は、認定事業者を活用することへの関心でした。前者に関しては、認定事業者から直接的な利益を得られないと感じている事業者にとって、避けたい事務的負担であると考えられます。特に、認定事業者による医療情報の不適切取得事案[2]以降、医療情報を提供する事業者の視点からは、リスクと利益の不均衡が指摘されています。認定事業者を活用する意欲があれば、そのリスク（非常に小さいと思われるが）を受け入れることもできるでしょうが、多くの事業者にとっては、そのリスクを受け入れるだけのメリットを感じられないでしょう。

先述した啓発活動によって、認定事業者への関心や利用への意欲を喚起するだけでなく、事業者が最も負担を感じている、医療情報の提供に関する患者への通知作業や責任を軽減する施策を検討することが重要です。

### 提言 3:

#### 認定事業者を利用した研究を支援するファンドや研究助成事業を設定すること

学術関係者が認定事業者を利用することを推進するためには、認定事業者の利用に伴う費用や分析環境の提供に関する問題に対処する必要があります（分析環境については後述の提言にて言及します）。アンケートでは、認定事業者の利用時の価格設定について、「前提からして高額すぎる」、あるいは「そもそもコストを設定すべきではない」という指摘もありました。研究班が先行事例として調査した英国でも、公益目的であっても事業である以上はデータ収集、維持管理のコストが発生し、それを一定程度利用者に負担していただくことについて理解を頂くのに時間がかかったということです。認定事業者の運営維持について利用者に理解を促すための啓発活動が必要です。一方で、学術関係者の多くは認定事業者から提供される匿名加工医療情報に対する費用を賄える予算を持っていません。提言 1 で述べたような取り組みによって認定事業者の事業への理解を深めていただき、コストへのご理解を頂くことができる可能性はあります。それでも、学術関係者が負担できる研究予算には限界があります。AI や RWD を利用した研究においては、認定事業者にデータを集積し、研究事業の持続性を確保することに配慮した研究計画を優先的に採択したり、予算や体制面を重点的にサポートしたりする施策が考えられます。また、認定事業者の匿名加工医療情報を

利用した研究には、認定事業者の利用料金の一部を支援する施策も考えられます。ただし、個別の対応に頼ると全体的に最適化された予算配分や研究計画の最適化が困難になる可能性があるため、認定事業者を利用した研究について省庁間で情報共有し、サポートする体制を構築することが望まれます。

そして提言 4 にも関連いたしますが、現在、市場競争原理にさらされていることで性能、安全性、コスト的にもこなれた情報処理技術を認定事業者が積極的に活用できる土壌を構成することで、認定事業者の運営維持の持続性を確保できるように配慮することが望まれます。

#### 提言 4:

認定事業者に「信頼できる研究環境」の概念を導入し、

それを支えるためのクラウド基盤の活用を積極的に支援すること

本研究班の AI を利用した研究により、認定事業者の保有する医療情報および匿名加工医療情報が AI の研究に貢献することが確認されました。しかし、今回の経験から、認定事業者が「信頼できる研究環境」として安全にデータを分析できる環境を構築し、利活用に提供することが強く求められました。また、その「信頼できる研究

環境」をホストするための情報システム基盤として、認定事業者がクラウド事業者を活用できるように制度を整備することも望まれます。「信頼できる研究環境」とは、研究班において文献調査・渡航調査[3, 4]したスコットランドにおいて、英国国家統計局 (Office for National Statistics: ONS)が開発した公的統計マイクロデータの利活用促進のために開発し、また行政記録データの利活用を整備においても採用した5つの安全模範(Five Safes Model:5SM)を背景として、行政記録データや医療健康データ等とリンケージされたデータを含む、データ安全な分析環境を提供している"Trusted Research Environment(TRE)"[5]を訳したものです。2023年の日本医療情報学連合大会では、このTREに寄与されてきた英国のHealth Data Research UKのCTOであるEmily Jefferson教授にキーノートスピーチを頂き、クラウド事業者を含む情報基盤上にFive Safes Modelを採用したTREの状況を解説頂きました。

このTREで実現されていることで、我々が倣うべきだと思われることは、(1)様々な情報をリンケージして非識別化されたデータを提供する制度と環境の整備、(2)研究者がリモートアクセスし分析できる環境の提供およびその柔軟な構成、(3)最終的に分析結果などを外部に持ち出す前に審査プロセスを実施すること、およびこれらの取り組みを統合して安全に実施できる環境の提供が含まれます。(1)と(3)については、現在の認定事業者でも実現可能ですが、(2)のような利用者がリモートアクセスして分析できる環境の整備については、現在内閣府のワーキンググループで議論がなされており

ます。このリモートアクセス可能な分析環境は、セキュリティに十分配慮した上で、事業継続性が確保できるように対費用効果の高いパブリッククラウド<sup>1</sup>サービスを含めて選定、構築することが望ましいと考えます。その理由は、AI や機械学習を利用した研究は多くの計算機リソースやソフトウェアを使用するからです。研究者によって要求される環境は多様であるため、現在の認定事業者のオンプレミスの計算機リソースでは対応できません。一方、クラウドサービスであれば、必要な計算機リソースや機能をオンデマンドで調達し、必要な研究プロジェクトに迅速に対応できますし、分析のための情報システムの構築および運用に関するコストを受益者が負担しつつ、様々な研究計画に効果的に対応することが期待されます。

なお、分析環境を構築するクラウドサービス基盤については、認定事業者、利用者の負担を減らしつつ、同時に最先端の技術的環境の恩恵下に効率的な研究環境を構築できるようにすることから、Pay for Service で利用でき、最新の計算機資源、AI 環境が利用できるよう大手クラウドベンダーを選択できるような、開かれたポリシーを策定することが望まれます。

---

<sup>1</sup> 本稿で言及している「パブリッククラウド」とは、クラウドサービス提供事業者が構築した環境を複数の顧客と共同利用する利用形態です。クラウドサービス事業者は複数のサービスを提供しますので、利用者は自身が使用するクラウドサービスのみを組み合わせ、使用した分だけがチャージされる従量課金制を採ることにより、設計次第では自身で計算機資源を保有するより安全かつ安価になるケースがあります。サービスやデータは顧客ごとに厳重に論理的に分類されており、他の顧客の環境が侵害を受けてもただちに影響をうけることはありません。「パブリック」という言葉から想像されるようにサービスやデータが一般公開されたり、他の利用者から見えたりするということはありません。またインターネットからではなく閉域網を利用して接続できるサービスを選択することで、インターネットとの接触点を完全になくす方法も採用することができます。また、「プライベートクラウド」の中でも論理的分離によってプライベートクラウドを構成するタイプが主流になっているという状況もあります。

## 提言 5:

### 改正次世代医療基盤法に準拠した運用の早期確立への支援

研究班による取り組みでは匿名加工医療情報でも AI の研究に資することを確認しましたが、希少疾患や特異な値を含むデータは匿名加工の過程で脱落する可能性があります、そのようなデータが極めて重要な意味を持つようなケースでは匿名加工医療情報での研究が困難となる可能性があります。これに対して改正次世代医療基盤法では、匿名加工医療情報を新設し、また認定利用事業者の安全管理措置に関する基本的な考え方の 1 つとして、認定利用事業者の概念を設けました。すなわち、利用事業者が自ら整備した環境に匿名加工情報を保存する I 型認定利用事業者、認定事業者等が整備した Visiting 環境での利用に限定される II 型認定利用事業者です。英国では 5SM の「認可された研究者(accredited researcher)」や認可された機関 (recognized organization) の概念を採用し、研究者への教育や審査・認定に関する制度を整備することを通して安全性を担保しつつ RWD 研究の活発化につなげています。我が国においても、このような取り組みを参考にし認定利用事業者の教育や「信頼できる研究環境」の整備について迅速に進められるよう支援する施策が望まれます。

## 提言 6:

### 相互運用性に寄与する標準規格の導入支援をすること

次世代医療基盤法第 25 条<sup>2</sup>では、認定事業者間での医療情報の融通が認められていますが、本研究開始時点では実績はなく、運用管理規程や医療情報交換の手順が定められていませんでした。このため、研究班と認定事業者において、第 25 条に関連して契約、審査、データ連系の 3 つのテーマについて検討しました。その中でデータ交換の際にデータ項目の標準化が継続的な検討事項となりました。認定事業者ごとに収集されるデータの種類や形式が異なるため、認定事業者間でデータ交換を行う前に、この際を埋める作業が必要であり、この作業に多くの工数、コストが発生することが予想されました。また、収集されるデータの非互換性や標準化されていないコードでの記述により、データを利用可能な状態にするためには多くの加工プロセスが必要であり、データ品質の維持も困難になる傾向があります。さらに、標準化の視点は収集するデータに限らず、分析対象となるデータセットにも重要です。RWD を使用した研究では、研究者による分析の再現性、ツールの共用化、組織間での分析を実現するため

---

<sup>2</sup> 研究当時は第 25 条であったが、令和 6 年 4 月の改正法では 27 条になっている。研究当時の状況を記述するために、敢えて第 25 条のまま記載していることに留意されたい。

に、共通化したデータベース構造(Common Data Model)や用語の統一が必要です。分析のためには、標準規格に準拠したデータベースを利用者自身が用意するか、認定事業者が提供する分析環境で用意する必要があります。そのため、認定事業者や医療等情報を提供する事業者において、相互運用性に寄与する標準規格の導入は非常に重要です。標準的医療情報システムに関する検討会（内閣官房健康・医療戦略室）および医療等情報利活用ワーキンググループ（厚生労働省）で取り上げられた保健医療情報分野の標準規格（厚生労働省標準規格）は、医療 DX の推進に関する工程表にて言及されている情報システムに取りこまれる予定ですが、これらの成果が認定事業者へ提供する医療等情報にも及ぶように支援する視点を持ち、事業を検討することが望まれます。

## 研究班 構成員

|        |                              |
|--------|------------------------------|
| 木村 映善  | 国立大学法人愛媛大学・医療情報学講座 教授        |
| 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      |

[1]木村 映善, 高田 春樹, 荒木 賢二, 黒田 知宏, 星 佳芳, 佐々木 香織, et al., 次世代医療基盤法の認定事業者の認知状況,活用への意識に関する社会調査, 保健医療科学, 72(5), 454-63,(2023).

[2]内閣府 健康・医療戦略推進事務局. 内閣府健康・医療戦略推進事務局.次世代医療基盤 法の認定事業者による医療情報の不適切取得事案に 対する指導について (2022)

[Available from: <https://www8.cao.go.jp/iryou/nintei/nintei/pdf/220920ldi.pdf>.

[3]佐々木香織, 木村映善, 伊藤伸介, スコットランドの Trusted Research Environment にみる医療健康データの利活用状況, 投稿・査読中.

[4]伊藤 伸介, 木村 映善, 佐々木 香織, イギリスにおける大規模データの二次利用の新たな展開 —スコットランドの事例を中心に—, 統計研究彙報, 81, 89-108,(2024).

[5]Jefferson E, Liley J, Malone M, Reel S, Crespi-Boixader A, Kerasidou X, et al., GRAIMATTER Green Paper: Recommendations for disclosure control of trained Machine Learning (ML) models from Trusted Research Environments (TREs), arXiv preprint arXiv:221101656, (2022).

本稿は第 43 回医療情報学連合大会における「Keynote レクチャー2 英国とマレーシアにおける医療情報等の利活用について」セッション(2023 年 11 月 24 日)において、Emily Jefferson 教授に“The UK’s progress towards enabling secure, researcher access to sensitive health data at a UK population scale”の題名にて講演頂いたものの抄訳・要約である。木村（愛媛大学）の文責にて記す。なお、個人情報の保護に関する用語は、我が国の使い方と異なる部分があるため、ミスを避けるために、訳さずに原語を記載している。

--

今朝は、英国における、研究者によるセンシティブな健康医療データへの安全なアクセスを可能にするための進捗状況についてお話しします。まず、講演の概要です。Health・Data・Research United Kingdom、通称 HDR UK についてお話しします。信頼される研究環境 (TRE:Trusted Research Environment)とは何でしょうか。TRE は Safe Haven とも言います。HDR は、非常に複雑な状況の課題にどのように取り組んでいるのでしょうか。私たちがこの領域で取り組んでいるのは、テクノロジー・エコシステムの構築です。Health Data Innovation Gateway、そして DARE UK もぜひご覧いただきたいと思います。今朝の構成は以上です。

私共 HDR UK とは何でしょうか？私共のビジョンは、英国の健康医療データを統合し、人々の生活を向上させる発見ができるようにすることです。そして、私たちのビジョンは、ビッグデータと高度な分析が、すべての患者、(医薬品の) 相互作用、臨床試験、そして公衆衛生を向上させるための生物医学的発見に利益をもたらすことです。これは素晴らしいことです。しかし、実際にそれを行うにはどうすればいいのでしょうか？ 私たちが主にフォーカスしている分野は、公共の利益のための信頼できるデータ利用に向けて、研究者の取り組みをいかに加速させるかということです。私たちは、自らが健康医療データ分析を行うのではなく、コミュニティ全体が健康医療データを迅速に分析できるように、体系的な改革を行おうとしています。そのため、私たちは非常に多くの異なるグループ、多くの異なるパートナーシップにまたがって活動し、英国内でのすべての健康医療データに迅速にアクセスできるように変革しようとしています。研究者が多様で質の高いデータを見つけやすくし、またアクセスしやすくすることをおして、私たちは健康状態を改善し、英国の科学を後押ししているのです。私たちは、研究者がビッグデータを安全かつセキュアに利用するために必要な革新的ソリューションを生み出すことで、困難な技

術的問題を解決するためのリーダーシップを発揮しています。そして、英国全土のデータ、人、組織をつなぐオープンな研究共同体を発展させることで、ヘルスデータサイエンスを加速し、合理化しています。では、TRE について、日本の皆様と同じようなコンセプトを持っているかどうかはよくわかりません。(TRE) はセキュアなデータ環境、あるいは Safe Haven とも呼ばれています。そこで、英国ではどのような状況なのか、少し背景を説明します。これは私の印象ですが、多くの異なるデータがサイロ化されています。研究のために健康医療データにアクセスすることは、その機密性の高さゆえに、他の多くの分野とは異なる配慮が必要です。私たちがよく扱うのは、いわゆる非同意データ(unconsented data sets)と呼ばれるもので、日常的な診療を通じて収集されるデータです。そのため、研究目的でデータを使用することに患者からの同意を得ているわけではありません。つまり、一般診療データ、プライマリーケアデータ、病院データ、検査データなどです。私たちは、このデータを使用するために患者から同意を得た、同意のある研究収集コホートも扱っています。一般的に非同意データは、TRE を通じてアクセスする必要があります。しかし、より多くの同意済みコホートも、TRE を通じてアクセスされています。英国では、たとえ pseudonymized されていたとしても、非同意データはデータ共有契約(data sharing agreements)を通して共有されるのではなく、TRE の中で閲覧され、分析されるというのが一般的な傾向です。このような運用モデルには、データリポジトリ(の機能)が適しています。データリポジトリには識別可能なデータが保存されているかもしれないし、異なる組織から複数の異なるデータリポジトリが提供されているかもしれません。pseudonymized されたデータのサブセットはこれらから抽出されて、研究者に割り当てられた分析環境に配置されます。インターネットへのアクセスはなく、すべて厳重に管理されています。この環境にはエアロック(airlock)があり、出し入れできます。持ち出せるのは、図か結果として要約された表だけです。

(担当者が) ファイルを開いて、行レベルの情報や開示の可能性のあるデータがないことを確認するという、開示コントロールのチェックプロセスを経て、fully anonymous information だけが提供されます。つまり、傾向をプロットしたグラフかもしれませんし、5 人以下や 10 人以下の数字を含まない表だったりします。TRE とは何か、その原則についてもう少し知りたい方は、この下にリンクがあります。TRE や Safe Haven の全体的な考え方は、5 つの原則(Five Safes)に従うということです。これらの原則は、かなり以前から存在しています。それは安全な人々(Safe People)です。つまり、有効な研究者だけがその環境にアクセスできます。安全なプロジェクト(Safe Project)は、公益に資するプロジェクトでなければなりません。つまり、企業がマーケティング目的でこの環境を利用することはできません。Safe Setting、この環境は本当にロックダウンされており、リスクのあるものを持ち込んだり持ち出したりすることはできないようになっています。安全なデータ(Safe Data)とは、そこにあるすべてのデータは、特定の科学的な疑問に答えるための最小限のデータのみで、pseudonomised されたものです。そして、安全な出力(Safe output)とは、集計や要約されたデータ以外のものを取り出すことができない、とい

うことです。個人レベルの（粒度である）データを取り出すことはできません。もし興味があれば、これは HDR が書いた論文です。前のスライドと同じリンクにあります。信頼できる研究環境の構築、原則とベストプラクティスについて書かれています。この論文では、信頼された研究環境を利用するための最善の方法を文書化しています。HDR が解決しようとしている問題は、現在、多くの研究者は、1 つ 1 つデータ管理者を渡り歩いていくことしかできません。つまり、信頼できる環境で特定の研究に取り組むために、1 つのデータソース（のみ）にアクセスするのです。なぜなら、適切なガバナンスコントロールを行い、アクセス権を取得し、これらの異なるもの（データ）をすべてセットアップすることは、実際には非常に困難だからです。ですから、私たちが本当に望んでいるのは、このプロセス全体をもっと合理化し、研究者が全国から複数の異なるデータセットにアクセスするのを非常に容易なものとすることです。比較的シームレスな方法で、連携データプール(federated data pooling)や連携分析(federated analysis)によってアクセスできるようにするのです。英国中の多くの異なるデータセットについて、たとえそれらが異なるグループによって保持されていたとしても、研究者が（それらのデータを統合して）分析を行うことが日常的になることを目指しています。これが私たちの真の目的です。私たちには 20 年後のビジョンがあり、それは 6,700 万人（コーホート）に関する研究を本当に日常的にできるようになることです。これは英国の人口です。おそらく 20 年後までには、6,700 万人よりもっと多くなっているでしょう。ですから、私たちは統合的なアプローチをとっています。私たちは信頼できるデータ利用を加速させています。研究者に力を与え、パートナーシップを促進しようとしています。そのため、国家研究データ戦略を実施し、研究とイノベーションのニーズに沿ったインフラとサービスを組み立てることで、データ利用を加速させます。ここにあるのは、国レベルでのアクセスを可能にするために私たちが取り組んでいる、英国内のデータセットです。また、テクノロジー・サービス・エコシステムという柱もあります。これは私が率いている柱です。この分野については、後半で少しお話しするつもりです。しかし、テクノロジーでこれらの課題をすべて解決できるわけではありません。そのため、私たちは信頼と透明性、データガバナンス、データへのアクセスの法的側面にも取り組んでいます。これには、国民のニーズを理解し、信頼できる方法でこれを行う方法について国民と協力する、膨大なプログラムも含まれます。有用なデータ(usable data)とは、データ標準規格に準拠しているものを中心としたものであり、すぐに利用でき、質が高く、臨床試験など目的に合ったフォーマットでデータにアクセスできるものです。また、キャパシティビルディング(capacity building)とは、この分野で活躍できる人材を拡大し、訓練することです。私たちは、科学的発見を前進させ、患者や公衆に利益をもたらすために、オープンサイエンスとチームサイエンスにコミットし、多様な視点からのスキルを持つ人々を評価することで、研究者と協力しています。そのため、研究者や研究プロジェクトのライフサイクルのさまざまな段階で、さまざまなグループが活動しています。また、データを直接分析する人たち（データアナリストと呼びました）だけでなく、事前に多くの作

業を行っている人たちとも連携しています。私たちがやろうとしているのは、ドライバー・プログラムを使って、インフラが提供できるものの限界を押し広げ、さまざまな解決策をテストする手助けをし、模範的なプロジェクトを使ってこの分野を変革することです。現在6つのプロジェクトがあり、今後5年間継続されます。非常に複雑な状況ではありますが、私たちは重要なパートナーシップを構築・維持し、インセンティブを整え、分断された状況全体の複雑さを軽減して、ヘルス・データ・サイエンスを合理化しています。また、私たちが協力しているすべての組織の雰囲気を感じていただくために（スライドを提示して）、ここにはすべての組織が載っているわけではありません。非常に複雑な状況であり、私たちが行っていることの大部分は、研究コミュニティ全体で取り組むことに関わっています。健康医療データにアクセスしようとしている英国中の何百、何千もの異なるグループと、そのカウンターにあるすべてのデータ提供者と協力し、この状況を改善していくことで、これら2つの異なるグループが協力しやすくなるようにしています。また、産業界だけでなく、国民保健サービスも含まれます。

私は最高技術責任者（CTO）であり、この話に技術的な面をかなり持ち込んでしまったのではないかと思います。それぞれ異なる信頼される研究環境は、非常に複雑な研究における一連のプロセス(journey)にわたって人々をサポートしなければなりません。私はこれらを4つの異なる領域に分類しました。つまり、データ発見、データアクセス、データ環境、データ分析です。そして、これらの異なる領域のそれぞれにおいて、特定のプロジェクトをサポートするために、様々な異なるソフトウェアプログラムやインフラをセットアップする必要があります。ですから、これらの異なるツールやプロセスについて一つ一つ説明するつもりはありません。しかし、それぞれの異なる Safe Haven や TRE は、これらの異なるステップのひとつひとつを実行するために、ソフトウェアを構築したり、ソフトウェアを使用したり、ソフトウェアを設定したりしなければなりません。つまり、人々がデータを発見できるような仕組みを提供しなければなりませんし、すべてのアクセスコンポーネントを設定しなければなりません。そして、特定のデータ研究に関連するデータセットを構築し、それを非識別化しなければなりません。研究者だけがアクセスできるように環境を設定し、その方法の手続きを定めて、例えばインターネットにアクセスできない環境であっても機能するようにしなければなりません。また、データのクリーニングやデータの標準化には、非常に多くの手順が必要です。特に、臨床ケアで収集されたデータを取得する場合は、非常に厄介なデータである可能性があります。これらのいくつかのステップは信頼された研究環境が行うこともできるし、他のステップは TRE 自身が行うこともできます。ここで重要なのは、多くの異なるグループがそれぞれ異なるソリューションを構築しなければならず、システム間での再利用や相互運用性がほとんどないことです。そのため私たちは、コミュニティ全体で協力して、この一連の作業をより効率的で合理的なものにしたいと考えています。そのために、私たちはテクノロジー・エコシステム全体で取り組んでいます。そして、この分野で働く人々のコミュニティをまとめようとしています。それは、この分野で働く研

究者であると同時に、これらのソリューションの一部を構築する技術者でもあります。また、能力やトレーニングの面でも、人々を集めようとしています。私たちは、コミュニティ全体の相互運用性の標準化に取り組んでいます。私たちはインフラストラクチャーの科学を重視しています。データガバナンスの要件を満たしながら、研究コミュニティのニーズも満たすような、この種のシステムをどのように拡張するかについては、それ自体、膨大な量の科学が必要です。また、私たちは医療以外の分野にも取り組んでいます。健康医療データだけを見ても、教育データや警察データ、その他の行政データとリンクさせる必要があるかもしれません。ですから、私たちが構築しているプロセスやツールの多くは、健康医療データだけでなく、あらゆる形態の機微なデータに対して機能します。そしてその中心には、数多くの技術的ソリューションがあり、エコシステムには HDR が構築していない、そして当然のことながら多くの異なるソリューションが含まれています。ですから、私たちはまだ要点しか見ていませんが、私たちが取り組んでいるのは、エコシステムをサポートするツールです。私たちは、これらの異なるコンポーネントを統合するためのソフトウェアを構築しようとしています。また、エコシステムの中で、一度作れば多くの人が使えるようなツールを作ろうとしています。つまり、これはすべて共同作業なのです。私たちはオープンソースプロジェクトで他の人々と協力し、信頼できるさまざまな研究環境が利用できるツールを構築しようとしています。そうすることで、より効率的になるわけです。これが私たちの非常にハイレベルなアーキテクチャです。ソリューション、標準、相互運用性に関して協力し合いながら、オープンで機敏なコミュニティ主導の環境を目指すというものです。私たちが、これらの取り組みを行うにあたって、何が安全であるかを定義する上で、一般市民や患者の参加を得ることは非常に重要です。個別対応では、私たちが望むような規模に到達することはできないでしょう。誰もがアクセスしたいすべてのデータが1つの場所にある単一のデータ・リポジトリは存在しないでしょう。ですから、私たちは連携しながら進める必要があります。また、国を超えた連携も必要です。そこで私たちは、さまざまな環境でさまざまな TRE が互いに会話し、API やサービスを利用することで、すべてのプロセスを効率化できるような、連携したサービス・エコシステムを構築しようとしています。私たちはテクノロジー・エコシステムのさまざまな部分に資金を提供してきました。これらは、私たちが資金を提供しているさまざまな戦略的分野です。ですから、今日はこれらの異なる分野すべてに目を通す時間はありませんが、皆さんが興味を持ちそうないくつかの分野に焦点を当てたいと思います。しかし、これらすべてを一緒に動かすことで、全体が部分の総和よりも大きくなるということです。相互運用可能な技術、コードの再利用、トレーニング、そしてこれらの異なる分野にまたがる能力開発です。もし、もっと詳しく知りたいという方がいらっしゃいましたら、私は喜んで、これらのさまざまなことについてお話させていただきます。中央の小さな緑色の塊は、信頼できるさまざまな研究環境間でソリューションの共有が行われていることを示しています。しかし、それらはあまり広く知られておらず、ソリューションもあまり広く共有されていません。そこで私たちがやろう

としているのは、ゲートウェイと呼ぶ技術的なソリューションを構築することです。このゲートウェイを利用することで、さまざまな人がさまざまなソリューションを見つけたり、データを見つけたり、出版物を見つけたり、人々の情報を見つけたりすることができます。つまり、本質的には有用なディレクトリなのです。そして、ディレクトリー機能だけでなく、私たちが提供できるいくつかのモジュールを構築してもらい、各 TRE が自分でそれを提供する必要がないようにしたいのです。いずれは、さまざまなグループがゲートウェイを利用し、相互運用できるようにしたいと考えています。現在のゲートウェイのバージョンはマーク（？）と呼ばれるもので、主な機能は以下の通りです。メタデータの辞書があり、900 を超えるデータセットがゲートウェイに登録されています（このスクリーンショットを撮った時点では 862）。そしてこれが、それぞれのデータセットに関する情報です。特定のデータセットをクリックすると、取得された変数やアクセス方法に関する情報、そのデータセットに関するすべての記述データを読むことができます。そして、データ・アクセスやリクエスト・フォームに入力することができます。

データセットを保有している様々な信頼できる研究環境やグループに問い合わせをし、もしそのグループが標準化されたデータアクセスフォームを導入していれば、そのプロセスで申請書を提出することができます。まだ、すべてのデータセットがそれに契約しているわけではありません。これは発展途上のツールです。また、別の方法もあります。特定のデータセットの情報にアクセスし、アクセスを承認して貰えるようデータ管理者とチャットで相談することもできます。このツールの目的は、研究者がすでにあるデータについてすべてを知ることができるようにすることです。研究者はここに来れば、新しいデータセットについて知ることができ、それぞれのグループへの連絡方法も知ることができます。また、データの利用登録や透明性の確保は私たちにとって非常に重要です。さまざまな研究プロジェクトでデータセットがどのように利用されたのか、なぜそれが公共の利益につながると見なされたのかという点で、一般の人々にとって透明性を確保するようにしています。特定のデータセットについて、そのデータセットを使った研究、その研究結果、研究発表の内容を理解することは非常に有益です。ゼロから始める必要はなく、他の人がすでに行ったことをベースにして、特定のデータセットの力を理解することができるからです。データセットの共有方法に関するデータ標準があります。これはデータの使用方法について記述するものであり、それに関するすべての情報をそこで見つけることができます。データディスカバリーとは、特定のデータセットに適用された可能性のあるソフトウェアやデータ解析スクリプト、また関連する出版物に関する情報を見つけることです。このように、ここにはさまざまなソリューションとのリンクがたくさんあります。繰り返しになりますが、これは研究者の業務を合理化し、1つの場所に行けば、英国のこれらの情報をすべて見つけることができるようにするものです。そしてコホート発見ツール。これは実現可能性分析(feasibility analysis)とも呼ばれます。特定のデータ管理者のところに行くと、あなたのプロジェクトに必要なデータセットがあるかどうかを調べる必要があることが多いです。例えば、25 歳から 35 歳で喘息があ

り、COVID ワクチン接種を受けたことのある人が何人いるか？というリサーチクエストを  
したいかもしれませんが、その質問は多くの異なるグループに対してする必要があります。200  
人しかいないという答えが返ってきて、基準を緩和したり、基準を変更したくなるかもしれませ  
ん。そのために何カ月もかかることもあります。そこで私たちがやろうとしているのは、そのプ  
ロセスを支援する自動化ツールを作ることです。そうすれば、そんなに時間はかかりません。私  
たちは CO-Connect<sup>1</sup>というプロジェクトを通じてこれを実現しましたが、このプロジェクトは  
もともと COVID データに焦点を当てていました。しかし、そのプロジェクトが終了した今、私  
たちはコミュニティ全体で、あらゆる異なる疾患領域にわたる他のデータセットに目を向け、  
これを集団規模で行えるようにするために取り組んでいます。少しビデオをお見せしましょう。

(ビデオ上映)

HDR UH コホート検索ツールは、Health Data Research Innovation Gateway のウェブサイト  
に組み込まれています。このツールを使えば、研究者は英国中の様々なデータセットから、研究基  
準に合致する人が何人いるかを調べることができます。例えば、研究者は各データセットの中  
で、PCR 検査が陽性で 40 歳未満の人が何人いるか？と問い合わせることができます。

ゲートウェイのウェブサイトは、anonymous data を使って研究者が尋ねた質問のみを処理し、  
受け取った質問に対して要約、anonymous 化された回答を提供します。そのため、身元が判明  
することはありません。このツールは、研究者がデータを保有する個々の組織に直接問いわ  
せることなく、利用可能なデータの概要を素早く知ることができることを意味します。基盤とな  
るデータは、それを保有する組織から移動することがないため、人々のデータは安全かつセキ  
ュアに保たれます。データ管理者は、健康医療データを保有し、アクセスを管理する組織です。  
各組織内には、個人を特定できるデータが保存されている場所とは別に、データ管理者の安全  
な環境内にある安全なコンピュータが設置されています。各組織のスタッフは、機密または識別  
可能であると思われるものを削除した関連データのコピーを作成します。例えば、氏名、住所、  
特定の生年月日、検査日、治療日などの情報は削除され、識別子は new pseudonymous codes に  
変換されます。その後、データ管理者はこの pseudonymous data を安全なコンピュータに転送  
します。この例では、Jane Doe についてデータ管理者によって転送されるのは、この人が 2020  
年 8 月第 2 週に PCR 検査で陽性であったという pseudonymous code と、その人の生年だけで

---

<sup>1</sup> COVID-Curated and open analysis and research platform <https://co-connect.ac.uk/> のことと  
思われる (木村)

す。データ管理者のセキュアな環境内のソフトウェアはゲートウェイにメッセージを送信し、データに対して実行する必要があるクエリを返します。例えば、データセットの中でPCR検査が陽性で40歳未満の人は何人いるか？とクエリを実行すると、ソフトウェアは、各クエリに対して要約された回答を返します。これは(クエリで指定した)基準を満たす人の数を示しています。研究者が最も有用なデータを発見するのに役立ちます。この例では、基準を満たした人は102人でした。ユニークな特徴量によって個人が特定されるのを防ぐため、各データ管理者は、結果を返さない時の最小限度を設定しなければなりません。ほとんどの場合において10人以上の結果のみが返されることになります。また、誰かが、個人を特定するために複数の質問をして、その差分から洞察することから保護するために、各データ管理者は結果を丸めることを有効にすることができます。それは決して正確な結果ではないけれども、研究者がデータセットの規模を知ることができる程度には正確であることは保証されています。この例では、データセットには正確には102人ではなく、110人いると言われるかもしれません。同じクエリがすべての組織にむけて同時に発出され、英国全土から(結果が戻って集計され)情報が提供されます。私たちがこの領域で何をしようとしているのか、少しは感じていただけたでしょうか。現在のゲートウェイは、本当に便利なツールです。パンデミックの最中に急いで構築されたものですが、基礎となるビルディング・ブロックです。最初の一步です。しかし、このゲートウェイを使ってできることの可能性はもっと大きいと思います。ですから、私たちは現在、来年の春か夏を目標に、大規模な新バージョンを立ち上げようとしているところです。このバージョンは、人々が本当にやりたいと思うことに対して、より将来を見据えた形で作られています。そして、オープンソースで、モジュール方式でリリースされる予定です。そして、人々がこのツールを採用したり、相互運用したり、素晴らしいコミュニティ・ソリューションを構築したいと思うのであれば、私たちは大歓迎です。このように、私たちはさまざまな方法でこのツールを強化しようとしています。私たちの重要な原則は、コミュニティが開発するということです。車輪の再発明はしたくありません。そのため、私たちは実際に既存のソリューションをいくつか組み込んでいます。そのため、他の主要な組織によって書かれた他のツールを使用し、強化し、これらのニーズを満たすために拡張したいと考えています。そして、模範となるようなものにしたいと考えています。ですから、私たちはコミュニティや模範的なプロジェクトと協力して、彼らのニーズを本当に理解し、そのニーズに合うようにします。標準化と相互運用性が鍵であり、自動化を進めることが第一です。そのため、できるだけ多くのオープンAPIを用意し、人々が手作業ではなく、自動化された方法でデータを送信できるようにしています。そして、人々のフィードバックに耳を傾け、それを確実にプロセスに組み込んでもらいたいと考えています。ですから、TREは素晴らしい環境だと思います。英国内では、10年以上前から存在しています。しかし、TREをより多く、より大きく、より良くしたいという要望が非常に強いのです。標準的な統計パッケージを使い、小規模なデータを用いた標準的な疫学研究には最適です。しかし、このような環境を本当に

拡張し、マルチオミクス解析や AI をサポートできるようにするには、まだかなりの道のりがあると思います。TRE の中にはそうなっているものもありますが、それが日常的に採用され、スケールできるようになるには、まだかなりの道のりがあります。ビッグデータをどう扱うか？セキュリティホールを作らないような安全な方法で、これらの環境内でプログラムによるアクセスを可能にするにはどうすればいいのか？すでに提供されているツールを使うのではなく、研究者が望むツールを自由にインストールできるようにするにはどうすればいいのか？環境やソフトウェアのコストを最小限に抑えるにはどうすればいいのか？また、ハイパフォーマンス・コンピューティングや GPU のようなものをどのように構成するのか？世の中にある優れたクラウド・ソリューションの多くは、必ずしもすぐに使える信頼できる研究環境ソリューションを持っているわけではありません。また、インターネットへのアクセスを遮断すると、すぐに多くのソフトウェアが機能しなくなります。そのため、スケーラブルなインフラを提供することは、必ずしも自明ではありません。また、AI や機械学習のために訓練されたアルゴリズムをどのようにテストするかという点についても、私たちは多くの仕事をしなければならないと思います。ハッキングされる可能性のある機密データを誤ってエンコードしていないことをどうやって確認するのでしょうか？これが、私が次世代 TRE と呼んでいるものです。私たちは DARE<sup>2</sup>と呼ばれるプログラムに取り組んでいます。DARE は、データとアナリティクスの研究環境に関するものです。HDR が主催していますが、さまざまなデータ領域で活動しています。健康医療データだけではありません。行政データ、教育データ、国勢調査データなどもあります。これらの同意のない大規模なデータセットは、TRE の中で同じように扱われる必要があります。DARE は、このような状況を一変させ、研究方法の標準化を図ろうとしています。また、プロセス全体が合理化され、スケーラブルになります。前のスライドにあったように、これまで 2 つの段階がありました。フェーズ 1 では、いくつかのスプリント・プロジェクトを検討し、それがフェーズ 1 の B につながりました。このプログラムはちょうど終了したところです。つまり、これらの異なる分野で活動する 5 つの異なるプロジェクトに資金を提供したわけです。まず TRE-FX プロジェクト<sup>3</sup>ですが、私はこのプロジェクトにかなり深く関わっていました。つまり、すべてのデータを 1 つのリポジトリに集約することができない多くのシナリオがあるということが課題なのです。複数のリポジトリにまたがって分析する必要がありますが、それは非常に難しいことです。分析ツールが多くの異なる TRE をまたいで機能することを望まれるかもしれませんが、各ツールキットには独自の交換標準があり、また各リポジトリは特注のソリューションを開発しています。各々のソリューションは、5 つの最も安全な原則に対して独立してテストされなければなりません。

---

<sup>2</sup> DARE UK (Data and Analytics Research Environments UK) <https://dareuk.org.uk/>

<sup>3</sup> TRE-FX のことと思われる (木村) <https://dareuk.org.uk/tre-fx-enabling-federated-analytics-for-better-healthcare-outcomes/> (木村)

ん。そして各ソリューションは、PPIE<sup>4</sup>、つまり公的な患者参加と関与の両方によって検証されなければなりません。そしてガバナンスです。つまり、連携プロジェクトを立ち上げるためのインフラを整え、承認を得るだけでも3年はかかるということです。そして、いったんセットアップが終われば、実際に分析を実行するのに10分もかからない、といったこともありえます。これでは効率的なやり方とは言えません。そこで私たちが試みたのは、中間的な標準を作れないかということです。これは、アレクシア<sup>5</sup>がすでに何年も前から発表しているリサーチ・オブジェクト(the research object)という標準をベースにしたものです。そこで私たちは、Five Safesを遵守するためにリサーチオブジェクトを強化しようと思いました。この考え方は、この安全なデジタル・リサーチ・オブジェクトを使って、サービス間の相互運用を行うための標準化されたメカニズムがあるということです。そして、サービスツールやTREのメタデータに対するクエリや結果が、Five Safesを満たしていることを表しています。そのため、この標準に準拠した複数の異なる分析ツール、複数の異なるTREを持つことができれば、新しい分析ツールを導入して、まったく新しいインフラを構築することなく、この環境の中で活動することが非常に容易になります。もうひとつ、SATRE<sup>5</sup>と呼ばれるプロジェクトに重点を置いていました。これは、TREのコミュニティ全体で取り組んでいました。TREのアーキテクチャがどのようなものであるかの標準を策定するもので、80種類くらいあったと思います。コミュニティ全体から意見を集めました。そして、TREがどのようなものであるかについての標準的な青写真を作成しました。これは、最初にお見せしたHDRが作成したFive Safesに関する文書よりもはるかに詳細なものです。そして私たちは、これをさらに前進させたいと考えています。Teleport<sup>6</sup>は、TRE-FXプロジェクトよりも、さらに連携分析に注目していました。このプロジェクトでは、連携したデータ・プーリングに注目していました。つまり、1つのTREに入ると、データはさまざまなTREに保管されているが、ユーザーにとってはすべて1つの環境内にあるように見える、というものです。これは、連携を実現するもう一つの方法です。これも大きな成功を収めました。SARA<sup>7</sup>は半自動リスクアセスメントを取り組み、それぞれのプロジェクトがどの程度リスクが高いかを調べ、より適切で割合の高いデータガバナンスを実現しようと思いました。SACRO<sup>8</sup>は、研究者のアウトプットを半自動でチェックするプロジェクトです。これもまた、私が目指していることに非常に近いものです。GRAIMATTER<sup>9</sup>のプロジェクトは、DARE フェーズ1のAプロジェクト

---

<sup>4</sup> Public involvement and engagement (PPIE)のことと思われる(木村) <https://www.hdruk.org/news/why-ppie-is-so-important-to-health-data-research-sail-databank/>

<sup>5</sup> SATRE: Standardized Architecture for Trusted Research Environments <https://dareuk.org.uk/driver-project-satre/>

<sup>6</sup> TELEPORT: Connecting researchers to big data at light speed <https://dareuk.org.uk/driver-project-teleport/>

<sup>7</sup> SARA: Semi-Automated Risk Assessment of Data Provenance and Clinical Free-Text in TREs <https://dareuk.org.uk/driver-project-sara/>

<sup>8</sup> SACRO: Semi-Automated Checking of Research Outputs <https://dareuk.org.uk/driver-project-sacro/>

<sup>9</sup> GRAIMATTER <https://dareuk.org.uk/sprint-exemplar-project-graimatter/>

で、その後 SACRO につながりました。GRAIMATTER は私が主導したプロジェクトで、信頼された研究環境の中で AI をどのように活用するか、また、取り出されるものが情報開示的でないことをどのようにチェックするかということがテーマでした。グラフやトレンドだけでなく、訓練されたモデルを取り出そうとする場合、その訓練されたモデルに行レベルの情報が含まれていないことをどうやって確認するのでしょうか？

仮に pseudonomised されていたとしても、誰かが誰であるかを知ることができる可能性はあります。そのため、さまざまなモデルが非常に多くの情報を含んでいる可能性があります。例えば、サポート・ベクトル・マシンは、学習データをすべてモデル内にエンコードすることができます。そのため、これを取り除こうとすると、情報開示管理に関するすべてのルールを破ることになります。では、TRE はどのようにそれを扱うのでしょうか？我々は、信頼された研究環境からの機械学習モデルの開示管理に関する推奨事項に関するグリーンペーパーを作成しました。そして、安全でセキュアな方法でこれを行うために必要な、さまざまな方法とさまざまなプロセスについて説明しています。というわけで、英国内で行っている様々なことについての私の紹介はこれで終わりです。