

[別添 1]

厚生労働科学研究費補助金
臨床研究等 ICT 基盤構築事業・人工知能実装研究事業

リアルワールドデータの二次利用加速にむけた多施設データ処理
方式の導入の試行研究

令和 5 年度～令和 6 年度 総合研究報告書

研究代表者 黒田知宏

令和 7 年（2025）年 3 月

I.	総合研究報告 リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究	 1
II.	分担研究報告 リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究 (大阪大学における検証)	 5
	リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究 (国立がん研究センター東病院における検証)	 13
III.	研究成果の刊行に関する一覧	 15

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)

リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究

研究代表者 黒田知宏 京都大学 医学部附属病院医療情報企画部

研究要旨

RWD を用いた臨床研究を活性化させるために、各医療機関で医療系データベースの構造を統一させる取り組みがわが国でなされている。一方で、それらのデータベースからのデータ抽出に関しては、人的コストや、データ配布時のセキュリティ等、様々なリスクや課題が存在する。そこで本研究では、クラウド技術として知られているコンテナシステムを援用することで、同一のプログラムで複数機関からセキュアにデータ取得が可能となるデータ処理方式を開発する。本年度は、プロトタイプとして3機関に前記クラウドシステムを導入した。各機関で動作確認を実施し、クラウド内でデータ解析を実施した。これらの成果は論文誌として公表する予定である。

研究分担者

岩尾友秀 (京都大学医学部附属病院先端医療研究開発機構、特定助教)

A. 研究目的

医薬品等の有効性評価に当たって、従来の臨床試験に加えて、医用情報の二次利用によるいわゆるリアルワールドデータ（以下、RWD）の利活用が欧米では先行して進められている。特に既存の治療が十分でなく、患者数が少ないため頑健な臨床試験の実施が困難な希少疾患用医薬品等を迅速に開発するに当たって RWD の活用が不可欠である。欧米では、既に一部の医薬品で EHR (Electric Health Record) を利用した薬事申請が行われている。日本においても第2期健康・医療戦略期間に臨中ネットの実装を目指しており、その取組みを加速する観点から、先行する欧州で既に利用が始まっている仕組みの導入の可能性について緊急に確認することは重要性が高い。

これまで、わが国では AMED 事業において、臨床研究中核病院による RWD 利活用に向けた基盤構築の取組みである「臨中ネット」を進めてきた。臨中ネット事業においては、中核病院のみならず将来的には全国の主要な医療機関を対象として治験等の臨床研究に耐えうるデータベース（以下、臨中ネット共通 DB）を各医療機関で構築している。現状の計画で

は、臨床研究を実施する研究者が自ら SQL を作成し、それを受け取ったデータ提供側の医療機関の担当者が自機関の臨中ネット共通 DB で SQL を実行することでデータ抽出を実施することとなっている。しかしながら、データ抽出のみならず実際の抽出データの受け渡し作業に際し、研究者とデータ提供元の担当者にかかなりの負荷が生じることに加えて、抽出したデータ受け渡しの際に患者データの漏洩リスクが生じることが予想される。これらの課題を考慮すると、現状のままでは人的費用や堅牢性の観点から臨床研究に適した仕組みとは言い難い。例えば、米国の FDA が管理しているセンチネルデータベースにおいては、SQL 等の専門知識を有しない研究者でも利用可能なシステムをハーバード大学に導入しており、わが国においても同様の対応が今後必要になると考えられる。現時点では、複数機関のデータを共有した解析環境をクラウド上で構築した例は我が国では見られない。

そこで本研究では、この取組みを加速する観点から欧州で利用が開始されている、複数の医療機関からデータを抽出、共有する仕組みに関して、臨中ネット参加病院の一部で試行を行い、導入の可能性や全体への導入に当たっての課題の洗い出しを行うことで、将来的に効率良く臨床研究が実施できるようなシステムを開発する。

B. 研究方法

そこで本研究では、Google クラウドが提供するモジュールを援用することで、各医療機関の研究者がデータ提供元のリソースに接続し、効率的に分析可能な仕組みを構築する。また、患者データはダウンロードできないことをシステムで保証する。下記に、令和 5 年度、および令和 6 年度の計画と方法について説明する。

<令和 5 年度の計画・方法>

本研究は以下の項目から構成される。

・データ抽出システムに関する研究開発

(研究代表者：黒田、研究分担者：油谷、岸本、岩尾)

<令和 6 年度の計画・方法>

・データ抽出システムに関する各機関への導入及び検証

(研究分担者：岡田、小西、青柳、野村、寺尾、油谷、岸本)

項目 1 は、研究代表者が所属する京都大学医学部附属病院（以下、京大病院）で試行的に実施する。本研究で想定しているシステムを図 1 に示す。はじめに、臨中病院側のシステムについて説明する。図 1 の右上で示すとおり、RDB として各医療機関では独自の臨中ネット共通 DB を持つ。また同時に、クラウドシステムにおいて計算リソースを確保し（図中、Airflow, Kubernetes モジュール）ている。分析者は、自機関に用意された端末からデータ提供元の計算リソースに接続し、RDB から ODBC 接続で所望のデータを抽出することができるものとする。なお、ODBC は、各データベースシステムで異なる SQL 命令の差異を吸収し、基本的な SQL 命令をアプリケーション経由で受け付けるためのデバイスドライバのことであり、アプリケーション側でのシームレスに計算処理が実行できるように導入する。

一方で分析者は、自機関の端末から、下記の手順で操作を実行し、データの収集やデータ分析を実行する。

1. 分析用のプログラム (Algorithm) を準備し、分析者の管理下にあるレジストリ (Container Registry) に配置する。
2. 分析結果やデータを収集するプログラムを準備し、1 と同様にレジストリに配置する。
3. 分析処理用のパイプライン設定の DAG を 1 と 2 で作成したコンテナ (Train) を設定して作成する。

4. 3 で作成した DAG をデータ提供元の Airflow に転送し、分析を開始する。

5. 1 と 2 のデータ収集、分析処理を実行した結果で得られた内容を含んだコンテナが分析者の管理下になるレジストリに戻る。以上述べた通り、本研究では、はじめに、先ほど述べた京大病院で試行的にクラウドシステムを用いたシステムを開発し、動作確認等を実施する。システム開発においては、京大病院に加えて他機関の研究分担者も参加し、各機関によるシステムを導入する。

次に、項目 2 に関して説明する。

項目 2 では、京大病院で開発したクラウドシステムを、大阪大学医学部附属病院と国立がん研究センター東病院に導入することで、実際に動作するか否かを評価する。京大病院では、開発したデータ検索・抽出用アプリケーションの動作確認を実施する。さらに、大阪大学医学部附属病院と国立がん研究センター東病院においては、別個に医学研究のリサーチクエスチョンを作成し、本システムのクラウド内でデータ解析を実施する。

各医療機関では、ODBC を使ってデータの収集が行える環境を構築し、その上で京大病院にて実施した検証と同じ検証を実施する。

なお、ODBC 利用環境を構築するためには、以下のような設定を実施する。

□各医療機関のデータベースと Google Cloud との間に FW を設置する

□各医療機関のデータベースと Google Cloud との通信には VPN を利用する

以上が本研究の計画・方法である。

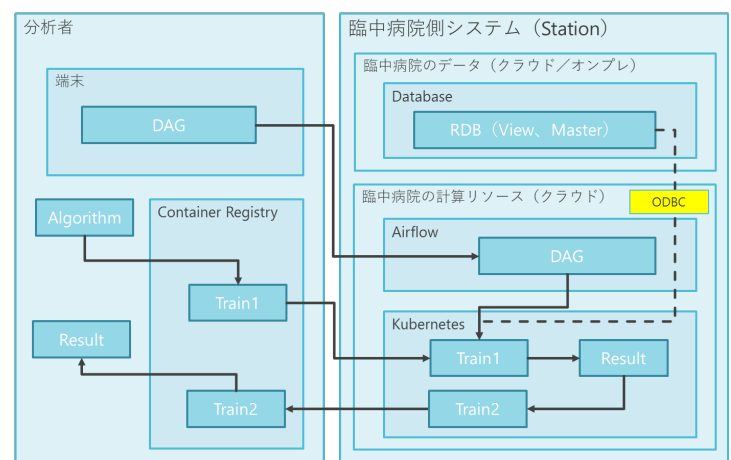


図 1 データ抽出環境の概要

C. 研究結果

B. 研究方法で述べた通り、データ抽出システムのクラウドを3機関に導入した。また、図2に示す通り、データ抽出・探索解析用のソフトウェアのインターフェイスを設計した。ユーザーが直感的にGUI操作できる仕様で、傷病名、医薬品、臨床検査値を指定することで、所望のデータを抽出することができる。また同時に、病名や医薬品を指定した患者人数計数機能を完備しており、検索結果をウィンドウに表示する機能がある。さらに、直接SQLを入力できるインターフェイスを備えることで、利用者の多様な抽出要望を満たすことが可能となった。

研究方法で述べた通り、GCPを用いたクラウド内解析環境を3機関に導入した。また、データ抽出・探索解析用のソフトウェアのインターフェイス設計を実施し、GUIにより、ユーザーが直感的操作可能なソフトウェアを開発・導入した。

大阪大学と国立がん研究センター東病院では医学研究を立案し、本研究で開発したデータ抽出用ソフトウェア、及びDataTrainと名付けた解析環境のもとで、それぞれデータ解析を試みた。また、京都大学では、クラウド環境において大容量データを扱うSQLのテストを実施した。実施過程においてシステム上のトラブルはいくつか生じたものの、それらに対応しつつ医学研究を実施することができた。

クラウドシステムを各病院に導入するにあたり、様々な課題が見つかった。ひとつめは、病院ごとに細かなセキュリティルールが異なることに起因する課題である。3機関では、それぞれの接続方法が異なり、京大病院ではGoogle Cloudとの内部接続、大阪大学医学部附属病院ではオンプレのデータベースに対するVPN接続、そして、国立がん研究センター東病院ではMicrosoft Azureとの異なるクラウド間接続となった。今後、広く他の機関へ展開する際の主要な接続方式が網羅できたものと考えられる。

ふたつめは、データ抽出時の課題である。本研究ではデータベースシステムとしてPostgreSQLを用いたが、前記データベースに対してAirflowをはじめとした複数のソフトウェアを連携させてアクセスする仕組みであった。実際の抽出テストにおいては、データベースの計算・応答速度が遅い機関のデータを抽出する際に、エラーが頻発した。主に、PostgreSQLからの応答が遅いため、多層に実装されたいずれかのソフトウェアがタイムアウトに類似するエラーが多く出力されることとなった。

以上のべたとおり、本年度は、上記の内容を各機関に実装することでクラウド内で研究を完結することが可能な研究開発基盤を構築した。



図2 データ抽出ソフトウェアの操作画面

C. 考察

研究結果で述べたとおり、令和5年度はデータ抽出システムのクラウド環境を試み、データ抽出ソフトウェアのプロトタイプを開発した。複数病院を横断してデータを抽出するシステムは日本では類を見ない。また、複数機関に対して同じプログラムでアクセスすること可能であるため、各機関に限定した分析であれば、効率的に実施することが可能である。一方で、現状のシステムでは、各機関にアクセスするための権限付与を各機関の担当者が手動で実施する必要がある。今後は、倫理申請が通過した研究課題に関しては、自動的にアクセス権限が不要されるような仕組みが効率の面から良いと考えている。

本研究は、GCPクラウドシステムを用いてクラウド空間の中のみで、解析を実施することが可能であることが令和6年度の各医療機関での実証結果で示された。一方で、分析方法に関しては従来までのそれと同じ方法を用いている。そのため、分析時におけるユーザビリティという観点からは、課題が残されている。「各機関に対して個別にアクセスして各機関の分析結果を得る」という用途は、仮説構築や診療指針の確認といった探索的用途には優れるが、研究成果が求められるような医学研究には向いていない。なぜならば、多くの医学研究では、すべての機関のデータを統合した後に統計解析を実施する必要があるためだ。特に、従来までの臨床研究と比較して、RWDを用いた研究のデータ前処理は負荷が高いことが知られている。一般に、集計や統計解析においては各患者と変数が一対一対応している必要がある。しかし、電子カルテ等のRWDでは同一患者に対して複数回同一の検査や処方がなされているケースが多く、その場合各患者IDに対して検査や処方をひとつだけ「選択」する必要がある。さらに、複数テーブルの全探索が必要になるケースもある。例として、「病名テーブルに記載されているⅡ型糖尿病の診断日から、30日以内に医薬品テーブルにおいてインスリンが投与されている患者」を集計する場合を

考える。このケースでは、診断日は患者ごとに異なるため、患者ごとに医薬品テーブルを全探索する必要がある。単純な人数集計であるが、RWD を用いる場合は相応の手間と時間が必要になる。

上記で述べたことを考慮すると、RWD の集計処理は、大きく分けると①単一ファイルのみで選択の必要がない項目、②単一ファイルのアクセスで済むが選択処理が必要になる項目、③複数ファイルへの全探索アクセスが必要になる項目、に分類することができる。本研究で開発したシステム上に、これらの対策を施すことができれば、患者情報が漏洩する可能性が少ないという本システム特有の利点に加えて、利用者のユーザビリティの双方を改善することが期待できる。

次に、研究結果で述べたデータ抽出時にエラーが頻発した件について考察する。本研究で実装したクラウドシステム特有のタイムアウトに起因するエラーは、PostgreSQL に直接接続して実施した場合には、起こりえない問題であった。今後の対策としては、クラウドシステム（GCP）で準備されているソフトウェアを使用するのではなく、限りなく直接接続に準ずるようなシステムを構築することがひとつの打開策になると考えられる。

最後に、期待される効果としては下記が想定できる。

- ・本研究で提案するシステムは、クラウド技術を援用することで研究者自身が作成したプログラムを、各医療機関に接続して実行、解析し、解析結果等の結果を得るシステムである。そのため、データベース言語や解析等の専門性に乏しい研究者に対しても利用を広げることで臨床研究の質、量ともに大幅に向上することが期待される。

- ・本研究では、複数の医療機関においてテストデータを用いて、データ受け渡しテストを試行し、データ漏洩が生じないシステムが構築できたことを確

認する。このため、今後他の機関に本研究の成果を導入する際にも問題が生じにくいことに加えて、患者データの漏洩対策として堅牢な検証が期待できる

- ・本研究では、利用者側（研究者側）で必要となるクラウド上での解析作業に必要となるソフトウェアを開発する際に、どのような問題を考慮する必要があるかという指針を示したと言える。

- ・また、本研究における検討成果は「臨中ネット」の取組みに組み込まれることを想定しているため、リアルワールドデータ利用の基盤整備の加速に資するものとなる。我が国でもさらに増えていくと予想されるリアルワールドデータを用いた臨床研究・治験等に対応することが期待される。

E. 結論

本研究では、複数機関においてクラウドシステム上でデータを解析するというわが国では類を見ない取り組みを実施した。開発したクラウドシステムは、各機関への接続方法やデータ抽出において様々な課題は見つかった。このため、今後複数機関でデータ共有、解析を実施するようなシステムを開発する際の端緒となることが期待できることに加えて、技術的な観点からも参考モデルになったと考えている。

F. 健康危険情報

特になし。

G. 研究発表

1. 論文発表

- 該当なし

2. 学会発表

- 該当なし

H. 知的財産権の出願・登録情報

特になし

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)

分担研究報告書

リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究

(大阪大学における検証)

研究分担者：岡田佳築 国立大学法人大阪大学 医学部附属病院 准教授

小西正三 国立大学法人大阪大学 医学部附属病院 特任助教（常勤）

研究要旨

わが国では AMED 事業において、臨床研究中核病院による RWD 利活用に向けた基盤構築の取組みである「臨中ネット」を進めており、各施設で共通の項目で構成された臨床研究のためのデータベース(臨中ネット共通 DB)を構築している。現状では、臨床研究を実施する研究者が自ら SQL を作成し、それを受け取ったデータ提供側の医療機関の担当者が自機関の臨中ネット共通 DB で SQL を実行することでデータ抽出を実施するが、データ抽出から受け渡しまで、研究者およびデータ抽出担当者にかかなりの負荷が生じることになる。本研究では、研究者がデータ提供元の各医療機関のデータベースに接続してデータ抽出が可能なデータ抽出システムを Google Cloud 上に構築し、効率的なデータ抽出および実際に本システムを用いる際の課題について検証することを目的としている。各施設からのデータ抽出については、システム内のソフトウェア(DTApp)を用いて抽出患者条件を設定して抽出する方法と、詳細条件を SQL で直接記載して抽出する方法を行ったが、前者ではデータ量が多い場合に抽出エラーになる場合があった。データ解析については、ARNI 内服開始前後の BNP・NT-proBNP の検査実施状況および検査結果の推移に関する解析を、解析に必要なデータの抽出をデータ抽出システムで行い解析は研究者の PC 上で行う方法と、解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う方法で行った。共に、データ抽出にかかる労力と時間を大幅に削減することができると考えられたが、データの解析までを Google Cloud 上でおこなう方法では、現在、医療機関ごとの解析のみが可能であり、今後、複数の医療機関からのデータを統合した解析の実施が可能とする点が、特に課題になると考えられた。

A. 研究目的

わが国では AMED 事業において、臨床研究中核病院による RWD 利活用に向けた基盤構築の取組みである「臨中ネット」を進めている。本事業においては、各施設で共通の項目で構成された臨床研究のためのデータベース（臨中ネット

共通 DB）を構築している。臨中ネット共通 DB では、共通の項目でデータベースが構築されているため、データ抽出のための同じ SQL を用いて、各施設からデータ抽出が可能である。現状の計画では、臨床研究を実施する研究者が自ら SQL を作成し、それを受け取ったデータ提供側

の医療機関の担当者が自機関の臨中ネット共通DBでSQLを実行することでデータ抽出を実施するが、データ抽出から受け渡しまで、研究者および各施設のデータ抽出担当者にかかなりの負荷が生じることになる。上記の課題に対して、本研究ではGoogle Cloudが提供するモジュールを利用し、研究者がデータ提供元の各医療機関のデータベースに接続してデータ抽出が可能なデータ抽出システムを構築し、効率的なデータ抽出について検証するとともに、本システムを用いた解析を、システムの利用範囲を変更しながら試行的に実施し、実際に本システムを用いる際の課題について検証することを目的としている。

B. 研究方法

1. データ抽出システムを用いた各施設からのデータ抽出

まず、京都大学医学部附属病院を中心に構築したGoogle Cloud上のデータ抽出システムを、共同研究機関である、京都大学医学部附属病院、国立がん研究センター東病院、大阪大学医学部附属病院の臨中ネット共通DBに、各施設で接続するための環境構築を行った。本システムでは、Google Cloud上のデータ抽出・分析環境から、各施設の臨中ネット共通DBとODBD接続して、データ抽出が可能であるが、大阪大学医学部附属

病院では、臨中ネット共通DBが病院情報システムネットワーク内に設置されており、Google Cloudからの直接の接続・通信は、大阪大学医学部附属病院における情報セキュリティポリシーに抵触するため、大阪大学医学部附属病院の臨中ネット共通DBを複製したDBを、病院情報システムネットワーク外のセキュアな環境に構築し、データ抽出システムとの接続環境を構築した。

データ抽出に関しては、循環器疾患患者でアンジオテンシン受容体ネプリライシン阻害薬（ARNI）（商品名：エンレスト）が処方されている患者の、臨中ネット共通DBの対象症例IDリスト、検体検査結果テーブル、処方テーブルを抽出した。データ抽出システムを用いたデータ抽出については、システム内のソフトウェア（DTApp）を用いて、以下の2つの方法で行った。

- DTApp上で抽出患者条件を設定し、対象症例IDリスト、検体検査結果テーブル、処方テーブルを抽出（図1）
- SQLを直接記載し、対象症例IDリスト、検体検査結果テーブル、処方テーブルを個別に抽出（図2）

図1：DTApp画面（抽出患者条件を設定し抽出）

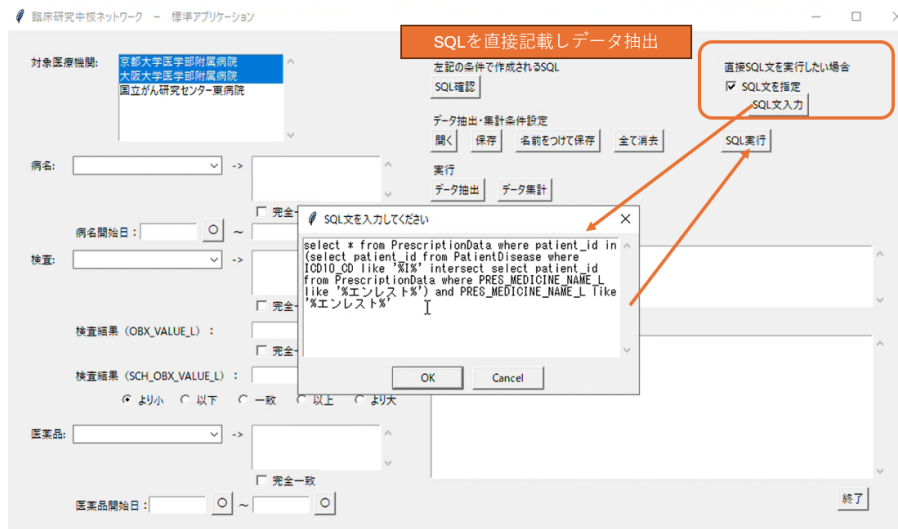


図 2 : DTApp 画面 (SQL を直接記載)

2. データ抽出システムの利用範囲を変更した解析の実施

BNP や NT-proBNP は心不全のバイオマーカーとしてその測定が推奨されている。ARNI はネブリライシンの作用を阻害することで、BNP の分解を抑制するため、ARNI 導入時には内因性の BNP が一時的に上昇する。そのため、ARNI 導入時にはバイオマーカーとしての BNP 値の解釈に注意が必要である。一方、NT-proBNP には ARNI による影響はなく、今回の試行的な解析として、ARNI 内服開始前後の BNP および NT-proBNP の検査実施状況および検査結果の推移に関する解析を行った。具体的な解析内容は以下とした。

対象患者：

- ARNI が 90 日以上処方された患者

解析項目：

- ARNI 処方日前後 90 日以内の BNP/NT-proBNP の測定状況
- ARNI 処方日前後 90 日以内の BNP/NT-proBNP 測定値の推移

Google Cloud 上のデータ抽出システムには、
「1. データ抽出システムを用いた各施設から

のデータ抽出」で用いたソフトウェアである DTApp に加えて、データ収集・データ分析・結果配信をすべてクラウド上で行うことができる DataTrain が実装されている (図 3)。DataTrain はデータ収集 Train、データ分析 Train、結果配信 Train で構成され、それぞれ、以下の役割を担っている。

- データ収集 Train
 - データ収集 Train に設定した SQL 文を用いて、各施設の臨中ネット共通 DB からデータテーブルを取り出し、CSV ファイルとしてクラウドに保存。
- データ分析 Train
 - データ収集 Train が保存した CSV ファイルを取り出し、分析プログラムを実施し、分析結果をファイルとしてクラウドに保存。
 - 分析プログラムは、Python もしくは R のプログラムが利用可能。
- 結果配信 Train
 - データ分析 Train が保存した分析結果ファイルを、研究者がアクセス可能なクラウド領域に転送。

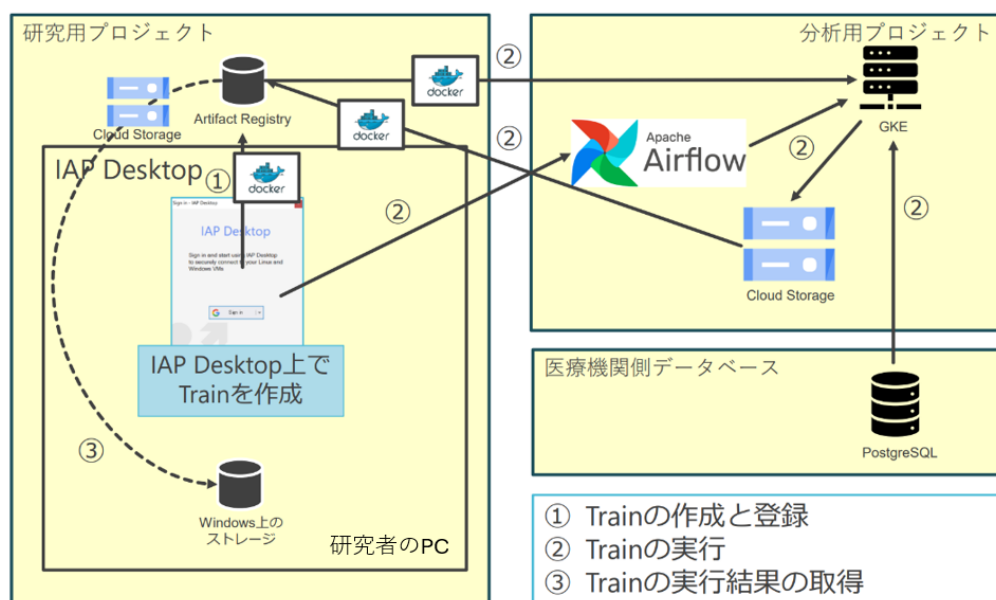


図 3 : DataTrain のイメージ図

上記のデータ抽出システムの機能を踏まえ、抽出システムの利用範囲としては、以下の 2 つのパターンを設定した。

- 解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。

【具体的な解析手順】

- DTApp を用いて、解析に必要なデータテーブルを抽出
- 抽出したデータテーブルから、研究者の PC 上で解析用データシート作成
- 解析用データシートを、統計解析ソフトを用いて解析（統計解析ソフトは、グラフィカルユーザーインターフェース（GUI）操作での解析が可能な SPSS とコマンド入力が必要なキャラクターインターフェース(CUI)である R を用いて同じ解析を行う。）

- 解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う。

【具体的な解析手順】

- 解析に必要なデータテーブルの抽出用 SQL をデータ収集 Train に設定
- 分析プログラムを R のプログラムとして作成し、データ分析 Train に設定
- DataTrain を実行
- 結果配信 Train により転送された分析結果ファイルをクラウド領域から研究者の PC へダウンロード

C. 研究結果

1. データ抽出システムを用いた各施設からのデータ抽出

【DTApp 上で抽出患者条件を設定し、対象症例 ID リスト、検体検査結果テーブル、処方テーブルを抽出】

DTApp 上で、病名の条件を循環器疾患（ICD10 コードが「I」ではじまる）および医薬品の条件を ARNI（薬品名が「エンレスト」）として、データ抽出を行った。DTApp の仕様では、この操作にて、臨中ネット共通 DB 上の、循環器疾患の病名登録があり、かつ、ARNI の処方がある患者の対象症例 ID リストと、検体検査結果テーブル、処方テーブル内のすべてのデータが抽出される。

大学医学部附属病院、京都大学医学部附属病院、国立がん研究センター東病院から抽出された対象症例 ID リストの ID 件数、検体検査結果テーブルのデータ件数、処方テーブルのデータ件数を表 1 に示す。

表 1：抽出されたデータ件数（DTApp 上で抽出患者条件設定）

	阪大病院	京大病院	国がん東病院
対象症例 ID リスト ID 件数	433	853	65
検体検査結果テーブルデータ件数	975,404	3,231,326	69,353
処方テーブルデータ件数	218,564	340,671	5,339

検体検査結果テーブルについては、施設によっては対象のデータ数が非常に大きくなり、抽出時のリソースが不足し、抽出エラーとなるケースが見受けられた。そのような場合は、医薬品開始日を限定し、数回に分けてデータを抽出することで、データの抽出が可能であった。

【SQL を直接記載し、対象症例 ID リスト、検体検査結果テーブル、処方テーブルを抽出】

SQL 文を直接記載する方法では、より詳細な抽出条件で、対象症例 ID リスト、検体検査結果テーブル、処方テーブルをそれぞれ抽出可能であるため、データ件数の多かった、大阪大学医学部附属病院および京都大学医学部附属病院の DB から、次の「2. データ抽出システムの利用範囲を変更した解析の実施」の解析テーマで必要な範囲のデータに絞り抽出を行った。具体的には、検体検査結果テーブルは BNP および NT-proBNP のデータのみ、処方テーブルは ARNI のデータのみ、以下の SQL 文を用いて抽出を行った。

- 対象症例 ID リストの SQL
 - `select patient_id from PatientDisease where ICD10_CD like '%I%' intersect select patient_id from PrescriptionData where PRES_MEDICINE_NAME_L like '%エンレスト%'`
- 検体検査結果テーブルの SQL
 - `select * from ObservationResult where patient_id in (select patient_id from PatientDisease where ICD10_CD like '%I%' intersect select patient_id from PrescriptionData where PRES_MEDICINE_NAME_L like '%エンレスト%') and OBX_NAME_L in (select OBX_NAME_L from ObservationResult where OBX_NAME_L like '%BNP%' or OBX_NAME_L like '%NT-proBNP%')`
- 処方テーブルの SQL
 - `select * from PrescriptionData where patient_id in (select patient_id from PatientDisease where ICD10_CD like '%I%' intersect select patient_id from PrescriptionData where PRES_MEDICINE_NAME_L like '%エンレスト%')`

抽出されたデータ件数は、表 2 に示す通り大幅に減少し、抽出時のリソースが不足等による抽出エラーを認めることなく抽出が可能であった。

表 2：抽出されたデータ件数（SQL を直接記載）

	阪大病院	京大病院
対象症例 ID リスト ID 件数	433	853
検体検査結果テーブルデータ件数	7,082	15,320
処方テーブルデータ件数	2,990	8,953

2. データ抽出システムの利用範囲を変更した解析の実施

【解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。】

解析に必要なデータの抽出は、件数の多い大阪大学医学部附属病院および京都大学医学部附属病院から、「1. データ抽出システムを用いた各施設からのデータ抽出」の SQL 文を直接記載する方法で行った条件で、ARNI が処方されている症例の検体検査結果テーブル（BNP および NT-proBNP のデータのみ）と、処方テーブル（ARNI のみ）を抽出した。

上記で抽出したテーブルから、対象患者を ARNI が 90 日以上処方された患者を対象に、解析項目である、ARNI 処方日前後 90 日以内の BNP/NT-proBNP の測定状況と ARNI 処方日前後 90 日以内の BNP/NT-proBNP 測定値（複数回測定の場合は平均値）の推移を解析するためのデータテーブルを R のプログラムを用いて作成し、解析用のデータテーブルを用いて、統計ソフトの SPSS および R のプログラムを用いて解析を実施した。

表 3 に ARNI が 90 日以上処方された患者の、ARNI 処方日前後 90 日以内の BNP/NT-proBNP の測定状況を示す。（数値は、SPSS による解析、R による解析ともに同じ数値である。）

表 3：BNP/NT-proBNP の測定状況

	阪大病院	京大病院
N	161	129
pattern (%)		
fromBNP_toBNP	30 (18.6)	31 (24.0)
fromBNP_toNT-proBNP	11 (6.8)	9 (7.0)
fromNT-proBNP_toBNP	0 (0)	6 (4.7)
fromNT-proBNP_toNT-proBNP	61 (37.9)	17 (13.2)
no_post	10 (6.2)	15 (11.6)
no_pre	49 (30.4)	51 (39.5)

ARNI が 90 日以上処方された患者は、大阪大学医学部附属病院で 161 症例（症例の処方開始日の範囲は 2020 年 9 月 25 日から 2022 年 12 月 26 日）で、京都大学医学部附属病院で 129 症例（症例の処方開始日の範囲は 2021 年 1 月 19 日から 2025 年 2 月 25 日）であった。処方日前 90 日以内および処方日後 90 日以内の両方に、BNP もしくは NT-proBNP の測定があった症例は、大阪大学医学部附属病院では 102 症例であり、京都大学医学部附属病院では 63 症例であった。このうち ARNI 開始前に BNP を測定していた症例は大阪大学医学部附属病院で 41 症例であり、ARNI 開始後に NT-proBNP に変更した症例（fromBNP_toNT-proBNP）は 11 症例、引き続き BNP を測定した症例（fromBNP_toBNP）は 30 症例であった。京都大学医学部附属病院で ARNI 開始前に BNP を測定していた症例は、40 症例で、NT-

proBNP に変更した症例は 9 症例、引き続き BNP を測定した症例は 31 症例であった。ARNI 開始前に BNP を測定していた多くの症例は、そのまま BNP が測定されていた。

ARNI 処方日前後 90 日以内の BNP/NT-proBNP 測定値の推移については、処方日前後に BNP の測定がある症例および NT-proBNP の測定がある症例を対象として解析を行った。解析については、平均値と標準偏差を算出し、箱ひげ図でグラフ表示を行い、前後 90 日以内の測定値の推移については、対応のある t 検定にて検定（有意水準：0.05）を行った。結果を図 4 に示しているが、大阪大学医学部附属病院、京都大学医学部附属病院ともに ARNI 開始後、NT-proBNP は有意に低下を示していたが、BNP は有意な低下は認めていなかった。

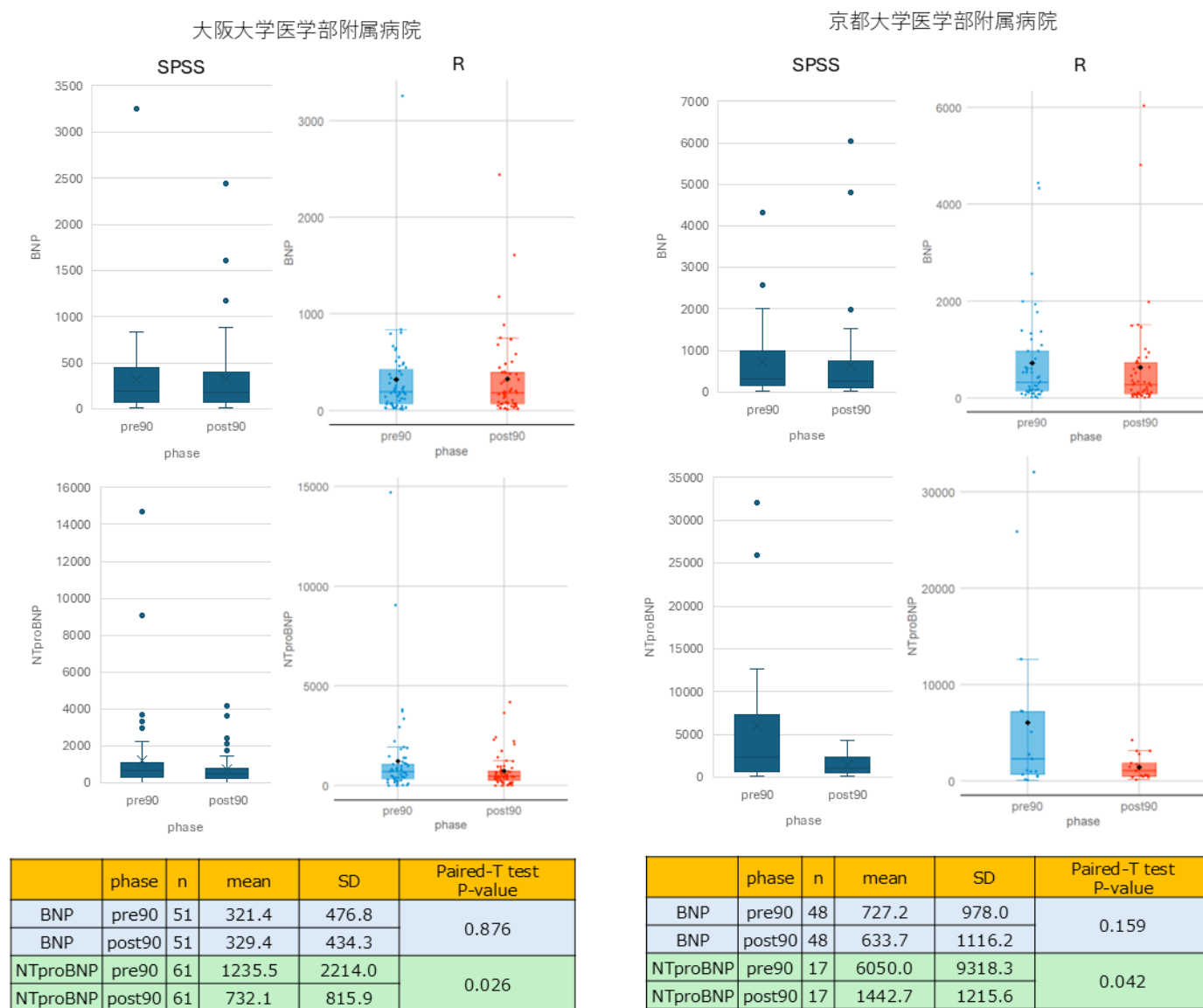


図 4 : ARNI 投与前後の BNP/NT-proBNP 値

【解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う。】

データ収集 Train に設定する、解析に必要なデータテーブルの抽出用 SQL は、「1. データ抽出システムを用いた各施設からのデータ抽出をデータ収集 Train に設定」の「SQL を直接記載し、対象症例 ID リスト、検体検査結果テーブル、処方テーブルを抽出」の部分で用いた、SQL を用いた。

データ分析 Train に設定する分析プログラムは、本セクションの前半の「解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。」部分で作成・使用した、解析用データテーブルの作成および統計解析用の R のプログラムを流用した。プログラムの流用については、R のプログラム上のファイル読み込みおよび書き出しのファイルパスを指定する部分を、Google Cloud のデータ抽出システムの構成に合わせたパスに修正することで流用可能であった。

上記にて設定を行った DataTrain を実行した後に、結果配信 Train により転送された分析結果ファイルをクラウド領域から研究者の PC へダウンロードして、分析結果を確認した所、本セクションの前半の「解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う。」部分で R を用いて行った、表 3・図 1 と同じ、解析結果と箱ひげ図が、分析結果としてダウンロードされていた。

D. 考察

1. データ抽出システムを用いた各施設からのデータ抽出

データ抽出については、DTApp 上で条件を指定することでデータ抽出が可能であったが、検体検査結果テーブル、処方テーブルについては、対象症例のすべてのデータが抽出されるため、葬礼数によっては、非常に膨大なデータ量になり、データ抽出がエラーとなり抽出できないケースが認められた。そのため、必要な項目を絞った形で抽出が可能な、SQL を直接記載しての抽出のほうが、解析に向けたデータ抽出においては、データ抽出がスムーズにいくものと考えられた。

SQL を直接記載しての抽出では、SQL を研究者が作成する必要があるが、DTApp では、DTApp 上で条件を指定した際に DTApp 上で作成される SQL を

確認することが可能であり、この SQL を土台として、必要な SQL を作成することで、比較的 SQL 作成は行いやすいものと考えられた。

2. データ抽出システムの利用範囲を変更した解析の実施

最初に行った、解析に必要なデータの抽出をデータ抽出システムで行い、解析は研究者の PC 上で行う方法については、研究者が一般的に行う解析手順になると考えられる。また、今回の解析項目を問題なく評価することが可能であった。

従来の臨中ネット共通 DB における解析手順と比べると、同一の SQL を用いて、クラウド上で必要なデータ抽出を行うことができるため、データ抽出にかかる労力および時間は、大幅に軽減されているものと考えられる。また、データ抽出後は、グラフィカルユーザーインターフェース (GUI) 操作での解析が可能な SPSS とコマンド入力が必要なキャラクターユーザーインターフェース (CUI) である R を用いて同様の結果が出ており、普段研究者が利用している統計解析ソフトによる解析につなげることが可能であると考えられる。

ただ、データ抽出については、検体検査結果テーブルや処方テーブルなどのデータテーブルは、別々に抽出されるため、これらのテーブルから解析用のデータテーブルを作成する必要がある。この解析用データテーブルの作成については、今回 R のプログラムを用いて、スムーズに作成することが可能であったが、R や Python のプログラムを普段利用していない研究者が解析用データテーブルを作成する場合には、この部分には労力がかかる可能性があると考えられた。

後半に行った、解析に必要なデータの抽出、データの解析までを Google Cloud 上のデータ抽出システム上で行う方法については、普段の解析で R や Python でプログラムを作成して解析を行っている場合には、比較的容易に行うことが可能であった。そのため、本システムを用いて、解析までをすべてクラウド上で実施することが可能であり、データ解析の労力は、大幅に軽減できるものと考えられた。しかしながら、現在のシステムの仕様では、医療機関ごとの解析しか実施できず、複数の医療機関からのデータを統合したデータを用いた解析の実施については、データ抽出後に研究者が自身の PC 上で統合したデータを作成して解析を行う必要がある。今後、

本システムを用いた臨床研究を推進していくためには、複数の医療機関からのデータを統合した解析の実施が行うことができるよう、システムを更新していくことが必要であると考えられた。

E. 結論

Google Cloud 上のデータ抽出システムを用いることで、複数医療機関の臨中ネット共通 DB から、効率的にデータ抽出およびデータ解析が可能であった。より効率的な実施には、システムに実装されている DTAApp や DataTrain などの機能を、データ収集・解析の目的に応じて、活用することが重要であり、今後、本システムを用いた臨床研究を推進していくためには、複数の医療機関からのデータを統合した解析の実施が行うことができるよう、システムを更新

していくことが、特に課題になると考えられた。

F. 健康危険情報

なし

G. 研究発表

1. 論文発表

なし

2. 学会発表

なし

H. 知的財産権の出願・登録状況

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし

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)

分担研究報告書

リアルワールドデータの二次利用加速にむけた多施設データ処理方式の導入の試行研究

(国立がん研究センター東病院における検証)

研究分担者：青柳吉博 国立がん研究センター東病院 臨床研究支援部門 臨床研究推進部 システム管理室

A. 研究目的

わが国では AMED 事業において、臨床研究中核病院による RWD 利活用に向けた基盤構築の取り組みである「臨中ネット」を進めている。2024 年 3 月時点の「臨中ネット」では、各施設に臨中ネット共通 DB を構築し、研究者が SQL を作成し、各施設のデータ提供者に SQL を配布することで、施設ごとにデータ抽出を実施し、データ収集を行う。この方法には、研究者が SQL を作成するという高いハードル、各施設のデータ提供にかかる工数の負荷、安全にデータ授受を行う仕組みの欠如といった問題がある。これらの問題点を解決するために、臨床研究中核病院である京都大学、大阪大学、国立がん研究センター東病院の 3 施設で、研究者が SQL を利用せずとも安全にデータ抽出ができる仕組みについて施行を行い、将来的に全ての臨床研究中核病院での導入可能性についても検討していく。

B. 研究方法

国立がん研究センター東病院では、研究対象のデータを Microsoft 社が提供するクラウド基盤である、Azure 上に格納した。環境構築に関しては、富士通社提供の Healthy Living Platform サービス（以下、HLPF という。）を利用した。分析環境は他機関と同様に Google クラウド上に構築した。研究対象データは電子カルテ系のネットワークにある電子カルテ DWH からのデータを HLPF 上の臨中ネット共通 DB へ手動で転送した。HLPF 上の臨中ネット共通 DB から Google クラウド上の臨中ネット共通 DB へは ODBC 接続することでデータ抽出を可能とした。

なお、ODBC 利用環境を構築するために、以下の設定を行った。

- HLPF 内の臨中ネット共通 DB 環境と Google クラウドの間にバーチャル FW を設置する。
- HLPF 内の臨中ネット共通 DB と Google クラウドとの通信に VPN を利用する。
- 臨中ネット共通 DB と Google クラウド解析環境を ODBC 接続する。

今年度は、京都大学が実施する臨床研究において、電子カルテ DWH から実臨床データを HLPF 上の臨中ネット共通 DB へ転送し、京都大学および大阪大学のデータと共にリアルワールドデータを用いた解析が行えるか否かを確認した。研究者は国立がん研究センター東病院にある端末から、Google クラウド上の計算リソース（Airflow、Kubernetes モジュール）に接続して、データ抽出のためのユースケースを策定し、データ抽出および解析を行った。

また、令和 6 年度はさらに、各施設で利用されるクラウド環境を安全に接続するためのセキュリティ対策について調査を行った。調査方法は Azure において利用されているクラウド間接続に関するベストプラクティスを中心に要件の概要を公開情報から収集した。調査は HLPF を提供する富士通社の支援のものと行った。

C. 研究結果

国立がん研究センター東病院では、HLPF 上に格納されている臨床データに対して Google クラウ

ド環境から ODBC 接続でデータ抽出が可能であることを確認した。本研究を通じて、実患者におけるデータの提供・取得が可能であることが明らかになった。また、クラウド間の接続におけるセキュリティ対策の調査では、トークンを用いたクラウド型情報交換技術仕様の確認やその実装状況を確認することができた。

D. 考察

データ分析では、複数の施設の大量データ抽出によるデータベースの負荷等、RWD ならではの課題が発生すると考えている。国立がん研究センター東病院では、京都大学が実施する臨床研究を通じて、京都大学および大阪大学のデータを国立がん研究センター東病院のデータを統合して解析することができた。この結果は、今後、臨床研究中核病院で行われている臨中ネットのような分散型の研究用データ統合プロジェクトにおいて有用であると考えられた。また、クラウド間接続におけるセキュリティ対策調査の結果、クラウド型情報交換技術仕様を適切に各施設環境に取り入れていくことで、より安全に各施設のデータを共有できる環境ができる可能性が示唆された。この結果は、電子カルテ情報共有サービスで利用される環境と ARO 等の研究機関で利用されるクラウドサービス間を連携する際のセキュリティ対策として発展することが可能ではないかと考えられた。

本研究における課題は各施設を増やした際にデータベースへの負荷が十分に検討できなかったことである。参加施設を増やすことにより、より大量のデータを研究に利用できることが可能となるが、その際に各施設の環境に与える負荷については十分把握することができなかった。国立がん研究センター東病院では、データベースをクラウド上に設置しており、負荷の度合いによってフレキシブルに性能を上げることができるが、その分だけ運用費用等への影響がある。今後は負荷に対する対応などを検討していくことにより大規模な研究を実施することが可能になると考えられた。

E. 結論

臨床研究中核病院である京都大学、大阪大学、国立がん研究センター東病院の3施設で、実患者を利用した分析ができることが確認できた。本研究の成果を基に多規模研究を行う際の負荷等の課題を検討することでリアルワールドデータの二次利用加速にむけた多施設データ処理環境を構築する可能性が示唆された。

F. 健康危険情報

該当なし。

G. 研究発表

該当なし。

H. 知的財産権の出願・登録情報

該当なし。

[別添 5]

研究成果の刊行に関する一覧表

論文

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
該当なし					