［別添 3］

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す作業に際し、研究者とデータ提供元の担当者に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ってくると考えられる。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。

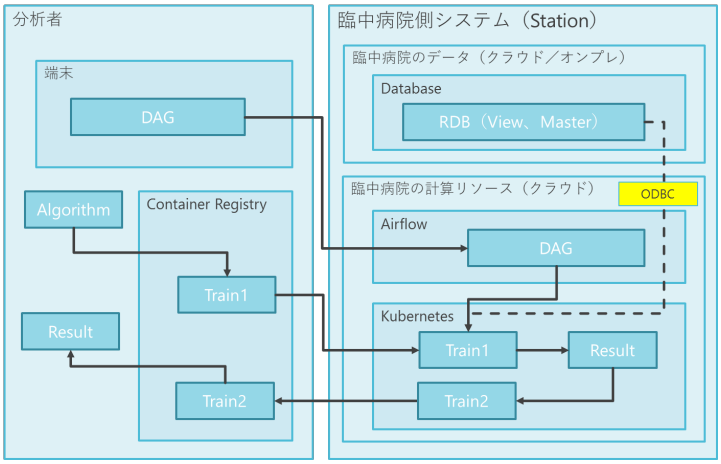

図 1 データ抽出環境の概要

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、検索結果をウイ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、利用者の多用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。

クラウドし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、そしれ、国立がん研究センター東病院ではMicrosof 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