

横断検索システムや企業向け相談窓口の整備等の体制構築

コホート研究のデータの標準データベースと検索 API の設計

研究分担者 荻島 創一 東北大学東北メディカル・メガバンク機構医療情報 ICT 部門 教授

研究要旨

次世代バイオデータ基盤の構築に向けて、令和元年度に実施した調査研究で取りまとめた「コホート・バイオバンクの産業利活用促進策」の具体化を図ることを目標とし、コホート・バイオバンクの産業利用促進のための調査研究を踏まえて、コホート横断検索システムを研究開発した。コホート研究により収集されるデータを収載する標準データベースを設計し、標準化された各コホート研究のデータベースを検索 API により横断して検索するシステムの研究開発を行った。また、解析・利活用基盤の構築に向けて、各コホート研究のデータベースへの API によるデータアクセスの研究開発を行った。

A. 研究目的

「バイオ戦略 2019～国内外から共感されるバイオコミュニティの形成に向けて～(令和元年6月11日 統合イノベーション戦略推進会議決定)」によれば、実現したい社会像として「医療とヘルスケアが連携した末永く社会参加できる社会」が掲げられ、具体的な取組として「バイオとデジタルの融合のためのデータ基盤の整備」が必要としたうえで、「大規模統合コホート・バイオバンクの構築」の中で、『健常人コホート等の実施主体が連携し、データを統合・強化する大規模健常人コホート・バイオバンクの構築』、『健常人コホート・バイオバンクについては、多様な分野において科学的エビデンスに基づいたサービスを提供できる環境を整備』することが明記された。

このような政策的位置付けの中で国立大学法人東北大学東北メディカル・メガバンク機構は、令和元年度の官民研究開発投資拡大プログラム(PRISM) バイオ技術領域で実施した調査研究(産

業界におけるコホート・バイオバンクの利活用のニーズ調査等)において、コホート・バイオバンクの産業利用促進策をとりまとめ、そのなかで、横断検索システム等のデータ基盤の整備が必要であるとした。

そこで、これらの産業利活用促進策の具体化を図るため、横断検索システムや企業向け相談窓口等の体制整備としてコホート・バイオバンクの産業利用促進のための調査研究を踏まえて、各コホート研究のデータのカatalogを横断して検索するシステムを研究開発した。また、解析・利活用基盤の構築に向けて、コホート研究のデータベースへの API によるデータアクセスの研究開発を行った。

B. 研究方法

1. コホート横断検索システムの研究開発

コホート・バイオバンクの産業利用促進策として、コホート横断検索システムを開発するため、株式会社ちとせ研究所による「コホート・バイオバンクの産

業利用促進のための調査研究」を通じてコホート横断検索システムに対する産業界のニーズを把握した。把握したニーズに基づき、コホート横断検索システムを概要設計した。

連携する PRISM バイオ技術領域「糖尿病個別化予防を加速するマイクロバイーム解析 AI の開発」及び「認知症に関与するマイクロバイーム・バイオマーカー解析」の施策の状況をヒアリングしたうえで、コホート研究で収集したデータを収載する標準データベースの設計を行った。

そのうえで、標準化された各コホート研究のデータベースを横断して検索する API を設計し、コホート横断検索システムの研究開発を行った。

2. コホート研究のデータベースへの API によるデータアクセスの研究開発

解析・利活用基盤の構築に向けて、コホート研究のデータベースへの API によるデータアクセスの研究開発を行った。横断検索システムのベースとなる統合データベースのデータ抽出機能の研究開発を行った。

(倫理面への配慮)

本研究はヒトゲノム・遺伝子解析、臨床研究、ヒトを対象とする医学系研究、動物実験等の実施はない。したがって倫理面の問題はないと判断した。

C. 研究結果

1. コホート横断検索システムの研究開発

コホート横断検索システムに対する産業界のニーズを把握し、コホート横断検索システムを概要設計した。PRISM バイオ技術領域の施策の状況を踏まえて、コホート研究で収集したデータを収載する標準データベースの設計を行った。標準化された各コホート研究のデータベースを横断して検索する API を設計し、コホート横断検索システムの研究開発を行った。

(1) コホート横断検索システムに対する産業界のニーズ

株式会社ちとせ研究所による「コホート・バイオバンクの産業利用促進のための調査研究」を通じてコホート横断検索システムに対する産業界のニーズを把握した。関心のあるデータ項目としては、睡眠やストレス・うつ、運動、飲酒、喫煙などの生活習慣、食習慣、罹患歴、検体検査値、マイクロバイーム情報、メタボローム情報、体重、BMI、体脂肪率などであることがわかった。このニーズを受けて、これらの項目を検索できるように設計することとした。

また、利用条件として産業界が二次利用可能であることが事前にわかることが重要であるというニーズがあることがわかり、横断検索システムにおいて産業界が二次利用可能であることを検索できるように設計することとした。

さらに、代表的なコホートやバイオバンクのデータのカatalogもひとつのプラットフォームで閲覧できるようにしたいというニーズがあることがわかった。

本研究では、PRISM バイオ技術領域の本施策と「糖尿病個別化予防を加速するマイクロバイーム解析 AI の開発」及び「認知症に関与するマイクロバイーム・バイオマーカー解析」の施策のコホートデータ横断検索システムを構築することとしているが、将来的には、多目的コホート研究(JPHC Study, JPHC-NEXT Study)、日本多施設共同コホート研究(J-MICC Study)などの代表的なコホート、AMED ゲノム医療実現推進プラットフォーム事業(ゲノム研究プラットフォーム利活用システム)によるバイオバンク・ジャパン、ナショナルセンター・バイオバンクネットワークなどの代表的な12のバイオバンクを横断した検索システムを構築することが求められており、AMED のバイオバンク横断検索システムと相互運用性を担保したシステムを設計することとした。

(2) コホート横断検索システムの概要設計

把握したニーズに基づき、コホート横断検索シス

テムの概要設計を行った。

コホート横断検索システムは、横断検索を実行するハブと各コホート研究のデータベースのノードにより構成することとした。ユーザがフロントエンドで検索条件を組み立て、検索クエリをハブ API に渡すと、ハブ API は各ノード API に横断的に検索リクエストを送信する。各コホートのノード API は検索リクエストを処理し、ハブ API に検索結果を返す。ハブ API は各ノード API から返された検索結果をとりまとめて、ユーザに対してフロントエンドで検索結果を表示する方式とした。

この方式は AMED ゲノム医療実現推進プラットフォーム事業(ゲノム研究プラットフォーム利活用システム)によるバイオバンク横断検索システムと同様の方式であり、これにより相互運用性を担保したシステムとすることができる。

(3) コホートデータの標準データベースの設計

把握したニーズに基づき、連携する PRISM バイオ技術領域「糖尿病個別化予防を加速するマイクロバイオーム解析 AI の開発」及び「認知症に関与するマイクロバイオーム・バイオマーカー解析」の施策の状況をヒアリングしたうえで、コホート研究で収集したデータを収載する標準データベースの設計を行った。

コホート横断検索システムに対する産業界のニーズによると、関心のあるデータ項目としては、睡眠やストレス・うつ、運動、飲酒、喫煙などの生活習慣、食習慣、罹患歴、検体検査値、マイクロバイオーム情報、メタボローム情報、体重、BMI、体脂肪率などであったため、生活習慣・食習慣は環境曝露イベント、罹患歴は協力者の既往症・併存症、マイクロバイオーム情報、メタボローム情報は解析情報、体重、BMI、体脂肪率などは検査イベントとして取り扱えるようにデータ構造を定義した(図 1)。

また、利用条件として産業界が二次利用可能であることが事前にわかることが重要であるというニーズについては、このことを同意条件として取り扱えるようにデータ構造を定義した(図 1)。

このデータ構造に基づいて、コホート研究のデータを収載する標準データベースの設計を行った。

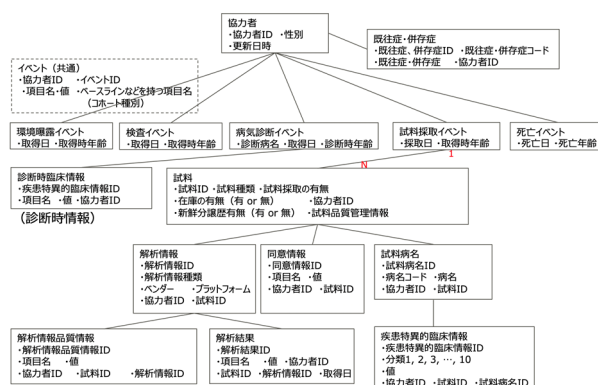


図 1 コホート研究で収集したデータの標準化されたデータ構造

(4) 分散データベースの横断検索 API の設計とコホート横断検索システムの研究開発

標準化された各コホート研究の分散データベースを横断して検索する下記の API を設計した。

- HistoryFetch : ログインしている利用者の検索履歴を返却する。
- ResultFetch : 検索状況や各バイオバンクの状況を返す。
- MasterFetch : 検索条件のマスター情報と利用者情報を返す。
- SearchAccept : 検索の実行を受け付ける。

検索については、時系列の検索を可能とした。そのうえで、コホート横断検索システムを研究開発し、フロントエンドについても実装し、東北大学東北メディカル・メガバンク機構が運用する AMED スーパーコンピュータの Web 公開区画に導入した。

① ログイン

認証は Auth0 のサービスを利用した(図 2)。



図 2 ログイン画面

Auth0 は柔軟で拡張性が高く、あらゆるタイプのアプリケーションへ簡単に迅速に組み込むことが可能な認証サービスである。認証基盤を独自にもつことは、セキュリティに対する脅威が多様化するなか現実的ではない。Auth0 はソーシャルログイン、多要素認証などさまざまな認証機能を持ち、必要に応じてそのセキュリティを高めることができる。



図 3 ユーザ登録画面

上記の画面のようにユーザ登録を行い、ユーザは利用を開始することができる(図 3)。将来的に、ユーザ認証のみならず、所属機関の認証を行うため、認証基盤を researchmap と連携させる予定であり、researchmap の関連情報として researchmap の各研究者の permalink と KAKEN 研究者番号を optional で入力できるようにした。

② 検索

本研究課題の東北メディカル・メガバンク計画、連携する PRISM バイオ技術領域「糖尿病個別化予防を加速するマイクロバイオーム解析 AI の開発」及び「認知症に関与するマイクロバイオーム・バイオマーカー解析」のコホートを横断して検索できるようにシステムを開発した(図 4)。



図 4 ログイン後のトップページ

なお、現状ではテストデータを収載しており、令和3年度以降に各コホートのカタログデータを収載する予定である。

左側の検索クエリのパネルにて、環境曝露イベント、協力者の既往症・併存症、マイクロバイオーム情報、メタボローム情報の有無等の解析情報、検査イベントについて時系列で検索可能なものとした(図 5)。

検索結果は協力者単位で表形式により表示され、列単位でソートやフィルタができるようにした。



図5 検索クエリと検索結果

これにより利用者は研究に必要なデータの有無を確認することができる。ユーザは検索履歴を保存することができ、研究に利用したいデータがあった場合には別途保存して(図6)、保存した協力者のIDリストを各コホートの企業向け相談窓口へ連絡することができる。



図6 保存された協力者

2. コホート研究のデータベースへのAPIによるデータアクセスの研究開発

解析・利活用基盤の構築に向けて、コホート研

究のデータベースへのAPIによるデータアクセスの研究開発として横断検索システムのベースとなる統合データベースのデータ抽出機能の研究開発を行った。

認証、検索、個票検索、CSVダウンロードなどのAPIの研究開発を行った。

D. 考察

コホート横断検索システムに対する産業界のニーズを把握し、概要設計した。PRISM バイオ技術領域の施策の状況を踏まえて、コホート研究で収集したデータを収載する標準データベースの設計を行った。協力者について環境曝露、検査、病気診断、試料採取、死亡のイベントを定義し、病気診断のイベントについては詳細な疾患特異的臨床情報を定義可能とした。また、試料採取イベントについては試料の情報を定義可能とし、解析情報は試料に由来する情報として定義可能とした。データ項目はISO、ICD-10、試料についてはSPRECなどの国際的な標準や事実上の標準に従うように標準化を行った。

標準化された各コホート研究のデータベースを横断して検索するAPIを設計し、コホート横断検索システムの研究開発を行った。この際、コホートのデータであることから時系列の環境曝露イベント、検査イベント、試料採取イベントとそれに紐づく試料及び由来する解析情報を検索できるようにするために時系列検索のアルゴリズムを開発し、十分な検索性能が保証されるようにテストデータによる検証を重ねた。

本研究により研究開発したコホート横断検索システムを令和3年度に実証し、令和4年度で実装及び運用することにより、コホート・バイオバンクの産業利用が促進し、多様な分野において科学的エビデンスに基づいたサービスを提供できる環境を整備し、エビデンスに基づいたデータ駆動型のヘルスケア産業を創出するとともに、データを統合・強化する大規模健康常人口コホート・バイオバンクの構築を通じて、国民のQOL向上に資する疫学研究の発展に寄与することが期待される。

E. 結論

次世代バイオデータ基盤の構築に向けて、コホート・バイオバンクの産業利用促進のための調査研究を踏まえて各コホート研究のデータのカatalogを横断して検索するシステムを研究開発した。また、解析・利活用基盤の構築に向けて、各コホート研究のデータベースのデータへの API によるデータアクセスの研究開発を行った。令和3年度からは、PRISM バイオ技術領域「糖尿病個別化予防を加速するマイクロバイーム解析 AI の開発」及び「認知症に関与するマイクロバイーム・バイオマーカー解析」と連携して、コホート横断検索システムの実証に取り組む。

F. 健康危機情報 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