ポスター
# ポスター2
# 情報セキュリティ・プライバシー
2017年11月21日(火) 14:15 ～ 15:15 L会場（ポスター会場2） (12F ホワイエ)

[2-L-1-PP2-4] ゲノム疫学研究における被検者重複登録回避のための実名登録システムの開発

山口 泉[1], 川口 喬久[1], 沖 俊吾[2], 内藤 千秋[3], 松田 文彦[1] （1.京都大学大学院医学研究科附属ゲノム医学センター, 2.日本ユニシス株式会社, 3.株式会社グリーン情報システムズ）

背景：疾患を対象とした多施設共同ゲノム疫学研究では匿名化した被検者情報を臨床情報や生体試料（ DNA検体など）と紐付けて解析拠点施設に集約し、解析を行う体制が一般的である。このような場合に特に希少難病では同一患者が複数医療機関を受診し、それぞれで同一の疫学研究などに登録されることがあるが、重複登録はゲノム解析などの研究資源の浪費や研究データの精度低下を招く。

目的：疫学研究における被検者の重複登録回避を目標として被検者実名情報の管理・照合・名寄せを行う実名情報管理システムを構築する。

方法：当施設のゲノム疫学情報管理基盤は検体提供元の施設で匿名化した被検者 IDとそれに紐付く臨床情報・検体情報を管理する。今回は実名情報の暗号化、匿名化、照合などの機能を備え、当方の現行情報基盤と連携可能な実名情報管理システムを開発した。開発の要点は以下の通り。①実名情報と臨床情報の分割管理、②実名情報の検索可能暗号化、③実名情報への厳格なアクセス制御とセキュリティ対策、④被検者情報照合機能

結果：①実名情報管理システムを独立サーバとし、現行情報基盤と物理的に離れた学内クラウドサービスに実装した。
②実名情報暗号化は業者に委託して検索可能分散暗号化技術を開発した。
③実名情報へのアクセス制御はユーザ権限設定とクライアント証明書により行い、リバースプロキシを経由してアクセスする体制とした。
④照合機能は (1) カナ氏名、(2) 生年月日、(3) 性別をキーとして重複の可能性がある被検者を提示し、個人情報管理者が重複の有無を判断して必要に応じて名寄せを行う仕組みとした。

考察：上記により実名情報を安全に管理し、必要に応じてそれを個人情報管理者にのみ提示し、名寄せによって重複登録が回避可能な仕組みが実現出来たと考える。本システムの実運用については疫学研究で実名情報を共同研究機関に渡す運用が一般的ではないため、①共同研究者の理解促進、②同意説明内容の工夫が必要で、現在対応中である。また、現在当施設で取り組んでいる希少難病レジストリシステムにも導入する予定である。

ゲノム疫学研究における被検者重複登録回避のための実名登録システムの開発

山口　泉[*1]、川口　喬久[*1]、
沖　俊吾[*2]、内藤　千秋[*3]、松田　文彦[*1]
*1 京都大学大学院医学研究科附属ゲノム医学センター、*2 日本ユニシス株式会社、
*3 株式会社グリーン情報システムズ

Development of a real name registration and verification system for epidemiologic studies in order to avoid overlapping of patient registration

Izumi Yamaguchi[*1], Takahisa Kawaguchi[*1],
Shungo Oki[*2], Chiaki Naitoh[*3], Fumihiko Matsuda[*1]
*1 Center for Genomic Medicine, Graduate School of Medicine, Kyoto University,
*2 Nihon Unisys, Ltd., *3 Green Information Systems, Ltd.

In multicenter genomic epidemiologic studies, clinical information and biological samples will usually be de-identified and sent to the genomic analysis facilities. In Japan, patients with rare diseases sometimes visit several specialized hospitals and thus an identical patient can be registered to the same disease study redundantly through several hospitals being de-identified differently. That will lead to a waste of research resources and a decrease in the precision of research data. In order to solve this issue, we have developed a real name management system for multicenter genomic epidemiologic studies, which stores personally identifiable information (e.g. name, gender, date of birth, address, phone number) of the study subjects in searchable encrypted form and can find the potentially identical study subjects using the stored personal information without decrypting. In this article, we report the background, the purpose, the process, and the results of the system development. Since registering personally identifiable information of the study subjects with the research management system of a collaborating research institute is not general in Japan, it is essential to promote the better understanding of both collaborators and study subjects about the purpose, benefit, and safety of the system. We are planning to utilize this system for our rare disease registry project in the future.

Keywords: real name registration, subject identification, searchable encrypted database

1. 緒論

特定の疾患を対象として多施設共同で行うゲノム疫学研究に於いては匿名化された被検者情報を被検者識別のための情報として臨床情報や DNA 検体などの生体試料と紐付けて解析拠点となる施設に集約し、ゲノムシーケンスやゲノム解析を行うという体制を取ることが一般的である。

このような場合に特に希少難治性疾患においては同一の患者が紹介あるいは自発的に複数の医療機関を受診し、それぞれの機関を経由して同一のゲノム疫学研究や疾患レジストリに登録されるということが生じ得る。当施設は多施設共同ゲノム疫学研究に於いて上記の解析拠点の立場となることが大半で、異なる施設から届いた複数の検体や情報がゲノムシーケンスを行なった結果同一症例由来のものであると判明した事例を複数経験している。

そのような重複登録は①ゲノムシーケンスなどの研究資源の浪費、②研究データやレジストリデータの精度低下といった問題を招くため、可能な限り回避することが重要である。

そこで本研究では被検者の重複登録防止と名寄せによって上記の問題を軽減するとともに各共同研究機関の被検者情報管理の運用水準を一定以上に保つことを目標として、疫学研究において被検者の実名情報管理・匿名化などを行う実名情報管理システムを構築したため、報告する。

2. 目的

匿名化が前提である多施設共同疫学研究に於いて同一症例が重複して登録されることを回避するために被検者の実名情報を安全かつ限定的に管理し、匿名化及び必要に応じた照合や名寄せを行うことの出来る実名情報管理システムを構築する。

3. 方法

3.1 開発の概要

当センターは主として多施設ゲノム疫学研究におけるゲノム解析拠点として検体管理、臨床情報管理、DNA タイピングや全ゲノム解析を中心としたオミックス測定、統計解析といった業務を行っており、それらの業務を支援する目的で匿名化された被検者の検体情報及び臨床情報を管理するシステムをシステムベンダーに委託して開発し、運用している [1)]。今回の開発ではこのシステムに実名などのプライバシー情報の登録とそれを用いた被検者の名寄せを行なう機能を実装した。

3.2 開発のベースシステム

当施設に於いて従来から運用しているゲノム疫学研究用の情報管理基盤は実際の研究運用における組織体制に基づいて複数の仮想的な部門（マスター管理部門、臨床情報管理部門、解析部門など）をシステム内に作成し、それぞれの仮想部門の業務を担う複数のサーバによってシステムが構成されている。

このシステム構成の元に情報及び検体の提供元の施設に於いて匿名化された被検者 ID（一次匿名化 ID）とそれに紐付いた臨床情報・検体情報の登録、それらを解析部門用にさらなる匿名化（二次匿名化）を行なう機能などを備えて当施設

の疫学研究をサポートしている(図1)。

3.3 実名情報管理のための機能追加

上記のシステムをベースに実名情報の保存、匿名化、重複登録回避のための実名照合機能などを備えた仕組みを従来のシステムと独立した実名情報管理サーバを構築し、従来のシステムと連携して動作させる方針で開発した。実装の概要は以下の通りであり、①実名情報を取り扱う機能と②実名情報を適正に管理・防御する機能の二種類に大別される。

1. 実名情報を取り扱う機能
   a. 実名登録機能
      氏名、性別、生年月日などの情報を登録する
   b. 実名情報に基づく被検者名寄せ機能
      カナ名、性別、生年月日による照合と名寄せされた被検者のデータ統合機能
2. 実名情報を適正に管理・防御する機能
   a. データの分割管理
   b. 検索可能暗号化データベース
   c. 利用者権限による実名情報へのアクセスコントロール
   d. セキュリティ対策

4. 結果

4.1 システムの実装

4.1.1 実名登録機能

被検者の氏名(漢字、カナ)、旧姓(漢字、カナ)、生年月日、性別、住所、電話番号などの他、保険証番号や将来策定予定の医療 ID も登録する仕組みとし、それら全てを暗号化してデータベースに登録する仕様とした。

4.1.2 実名情報に基づく被検者名寄せ機能

1. 名寄せ
   カナ名、生年月日、性別が一致する被検者を同一人物である可能性のある被検者として抽出する仕様とし、①被検者を新規に登録する際に同一人物である可能性のある被検者を呈示する機能と②研究プロジェクト単位に全体を検索して同一の被検者である可能性のある組み合わせを抽出する機能を実装した。
2. 同一被検者の統合
   名寄せ機能によって抽出された被検者の組み合わせについてシステムに登録されている他の情報を必要に応じて参照しながら個人情報管理者が確認・照合し、同一人物であると判断された場合に必要に応じて被検者 ID を統合する機能を実装した。後で統合が間違っていた場合に元に戻せるようにするため、統合した新しい被検者 ID を生成し、元の被検者 ID は無効化する仕様とした。
3. 関連データの統合
   異なる被検者 ID で同一人物が重複登録されていた場合に、元の被検者 ID で複数登録されているプライバシー情報を個人情報管理権限のある管理者が統合し、統合した新しい被検者 ID に引き継ぐことの出来る機能を実装した。統合は項目単位に管理者が登録情報を比較してどちらを引き継ぐか手動で選択可能な仕様とした。さらに ID 統合に関する情報は実名情報管理システムから臨床情報管理システムに引き継がれ、元の被検者 ID で複数登録されている臨床情報について同様のオペレーションによって臨床情報管理者が統合作業を行なえる仕組みを構築した。

4.1.3 データの分割管理

実名情報を管理するサーバと臨床情報を管理するサーバを論理的にも物理的にも異なるサーバとし、同一のサーバ内に実名情報と臨床情報の両者が同居しない構成とした。実名情報を管理するサーバは京都大学の情報環境機構が提供するクラウドサービスを使用した。

4.1.4 検索可能暗号化データベース

システムベンダーに検索可能暗号化技術の開発を依頼し、それを用いたデータベースを実名情報管理サーバに導入した。情報漏洩の経路が外部からの不正アクセスのみならず内部からの正当な手続きによる機密情報へのアクセスである場合があること、バックアップなどのオフラインメディアからの漏洩も想定する必要があることを考慮に入れ、以下のような特徴を持つシステムとした(図2、図3)。

1. データベースに収容された暗号化データが漏洩した場合の復元を以下の方法で困難なものとした。
   a. 秘匿化対象となるデータを暗号化する際にダミー情報を混入させる。復元時にはダミー情報の除去が必要となる。
   b. 暗号化情報を格納するテーブルに本システムが扱う秘匿化対象情報の全てを格納することにより目的とする情報の特定を困難にする。
   c. 暗号化した情報を複数のテーブルに分散して格納する。格納テーブルの順番は複数のパターンからランダムで決定する。
   d. プライバシー情報を被検者毎に収容するテーブルには上記の暗号化情報格納テーブル内の該当情報へのリンク情報を暗号化して収容する。
2. データ暗号化を行なう際の暗号化鍵として機密情報を収容するデータベースサーバとアクセス元となる PC のネットワークカード情報を使用するため、データベースを別環境にコピーしても情報が復元出来ない仕様とした。これによりバックアップメディアが漏洩した場合に正規環境以外での復元は防止される。
3. 暗号化対象情報はテキストデータとし、一文字単位で暗号化することにより部分一致検索が可能な仕組みを実現した。

4.1.5 利用者権限による実名情報へのアクセスコントロール

本システムではシステムの利用するユーザの職種や研究プロジェクトにおける立場などを反映したロールの概念を導入し、ロール単位で実行可能な操作の制御を行なっている。

今回の実装では「個人情報管理者」というロールを設定し、この権限を付与されたユーザのみが被検者の実名情報を登録及び閲覧出来る仕組みとした。

4.1.6 セキュリティ対策

実名情報を収容したサーバへの不正アクセス対策として以下の対策を行なった。

1. 個人情報管理者がアクセスする Web システムと通常のユーザがアクセスする Web システムを別立てとした。
2. 個人情報管理者用の Web システムへのアクセスには一般的なパスワード認証の他にクライアント証明書によるクライアント認証を必須とした。

3. リバースプロキシサーバを導入し、実名情報管理サーバに外部から直接アクセス出来ないサーバ構成とした。

4.2 研究倫理審査

疫学研究に於いては被検者と直接接点を持つ研究機関に於いて個人情報管理者が匿名を行い、実名との対応表をその施設で保管し、他の研究分担施設や共同研究機関に対してその情報は提供しないことが通例であるため、当施設が解析拠点を務めるゲノム疫学研究に本システムを組み入れるに当たって倫理審査申請では以下について項目毎に克明に記載した詳細資料を作成して提出し、実名情報管理システムを管理運用する立場にある当施設では承認された。

1. 「人を対象とする医学系研究に関する倫理指針」の「安全管理」への対応状況
2. 「ヒトゲノム・遺伝子解析研究に関する倫理指針」の「試料・情報の取扱い等」「個人情報の保護」への対応状況
3. 「医療・介護関係事業者における個人情報の適切な取り扱いのためのガイドライン」の「III 医療・介護関係事業者の義務等 3.個人情報の適正な取得、個人データ内容の正確性の確保、 4. 安全管理措置、従業者の監督および委託先の監督」への対応状況
4. 「医療情報システムの安全管理に関するガイドライン第 4.2 版」への対応状況

5. 考察

5.1 システムの設計・開発について

一般に多施設共同の疫学研究では被検者情報の管理という観点で異なる権限の様々な立場の研究者が実務に携わることとなり、実名情報を取り扱うことが出来るのは研究指針に於いて個人情報管理者と命名された役割の研究者とされており、他の立場の研究者は実名情報を取り扱ったり匿名化 ID との対応情報を知ったりしてはならないことになっている。

そのため、①被検者と直接接点を持つ研究機関に於いて匿名化を行い、実名との対応表をその施設で保管する、②複数の研究機関を取りまとめる研究事務局や当施設のような解析拠点では匿名化 ID を用いて情報や試料を取扱い、被検者の実名を知ることはない、という運用が一般的である。

以上の実情を踏まえ、本システムでは実名情報の登録・管理が出来る利用者の権限を限定し、その条件を満たさない利用者には実名情報が全く見えないように設計した（ユーザ権限によるアクセスコントロール）。また、クライアント証明書によるクライアント認証を採用することによってシステムへの接続環境の制限を行なうとともに（システムへの接続制限）、リバースプロキシサーバを導入して実名情報を収容したサーバに外部から直接アクセス出来ないネットワーク構成とした（システムへの不正アクセス防止）。

それにも関わらずシステムに何らかの方法で侵入されるなどしてデータベースからデータを直接引き出される事態になった場合も上述の手順で暗号化された形でデータが収容されており、正規の手順以外での復号は事実上不可能である。さらにシステムのバックアップメディアが万が一盗難に遭った場合でも上述の仕組みにより正規システム以外でそれを元にして復号を行なうことは出来ない仕組みになっている（不正侵入及びデータ漏洩時の対策）。

他方、名寄せについては全国民を一意に識別することの出来る番号体系で医療に使用可能なものが存在しない現状では氏名、性別、生年月日など複数の情報を元にして推測する以外に有効な方法がなく、システムによる自動名寄せを行なうことは現状では不可能であると判断して本システムでは同一人物である可能性のある既登録の被検者を見つけ出して個人情報管理担当者に呈示し、個人識別の目的でシステムに登録された情報を元に人が判断する仕組みとした。

結婚などで姓が変わっているケースにも対応できるようにするために姓は判定材料とせず、氏名については名の部分のみを用いる仕様とした。この仕様では被検者数が多い場合に無効な候補者が多数抽出される恐れが高いが、被検者数の規模がそれほど大きくない疫学研究や元々患者数が多くない稀少難病の領域では大きな支障はないと推測している。

以上によりゲノム疫学研究に於いて実名情報を安全に管理し、必要に応じてそれを個人情報管理者に対してのみ提示し、名寄せを行って重複登録が回避可能な仕組みが実現出来たと考える。

5.2 システムの運用について

名寄せを目的として実名情報を他施設のシステムに預ける運用は匿名化を必要としない地域医療連携システム [2] や多施設電子カルテ相互参照システム [3] においては報告があるが、疫学研究の現場に於いては実名情報を共同研究機関が構築・管理するシステムに預ける運用自体が一般的でないため、被検者からインフォームドコンセントを取得する立場にある共同研究者で当惑する者も少なくないのが現状である。そのため、被検者のみならず同意取得を行なう研究者自身も実名情報を登録する必要性、利点、安全性を十分に理解することが不可欠であり、(1) 共同研究者を対象とした説明を通じた共同研究者の理解促進、(2) 被検者の同意取得のための説明同意文書の記載と説明方法の工夫等が必要となる。現在はこれらに取り組んでいる段階である。

一部の被検者のみで実名登録が実現しても名寄せを十分に達成することは出来ないため、実名登録が全被検者について行えることが理想だが、個人情報保護法の厳格化の影響もあり、共同研究機関における倫理審査の敷居がさらに高くなっているのが実情である。

本システムは現在当施設で取り組んでいる稀少難病レジストリ事業に将来活用する予定で上記の準備を進めている。疾患レジストリにおいても被検者の重複登録はデータの質に関わる大きな問題であり、本システムのような仕組みが一定の効果をもたらすと考えている。

6. 結語

疫学研究や疾患レジストリにおける被検者の重複登録による研究資源の浪費やデータの品質低下を回避するため、被検者の実名情報を安全に登録・管理しそれを用いて名寄せを行なう仕組みを構築した。共同研究機関における倫理審査や被検者への説明同意など実名登録の運用実現の敷居が高いため現在はそのための準備を進めている。被検者を一意に特定出来る有効な仕組みが現状では存在しないこともあって名寄せの精度は高くないが、被検者規模がそれほど大きくない疫学研究や疾患レジストリにおいて一定の効果が期待出来ると考える。

参考文献

1) 山口 泉, 川口喬久他. 多施設共同ゲノム疫学研究のための臨床情報管理基盤の構築. 医療情報学 2014; 34(Suppl.): 888-891.

2) 中村直毅，中山雅晴他．地域医療連携システムにおける他施設
による名寄せの試み．医療情報学 2016；36(Suppl.): 633-635.
3) 近藤博史，寺本 圭他．名寄せ管理サーバを中心にした多施設
電子カルテ相互参照システムの開発と運用．医療情報学 2012;
32(Suppl.): 1118-1121.

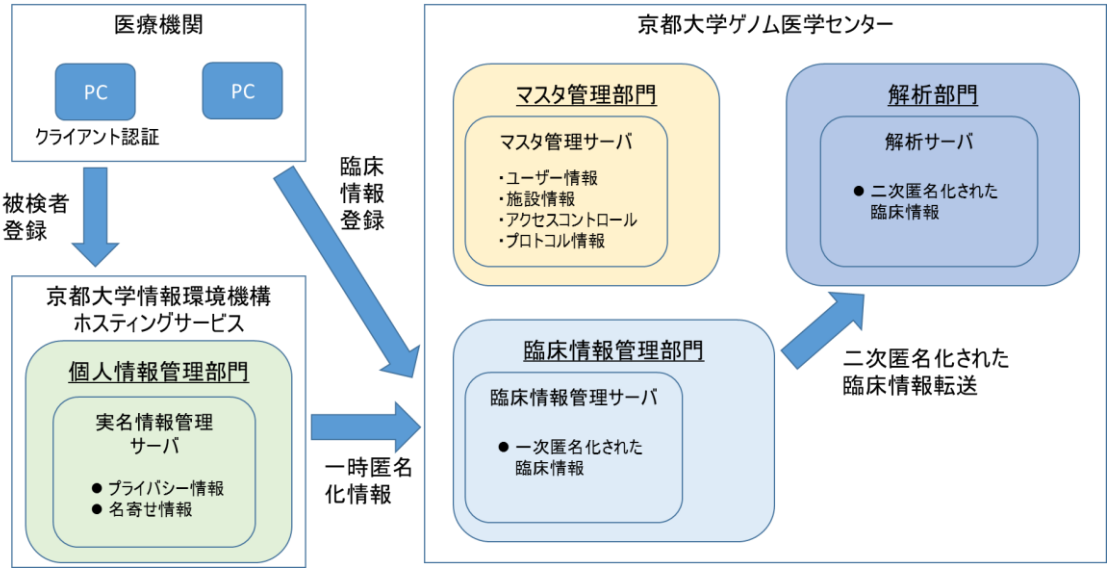

図1 システムの全体構成図

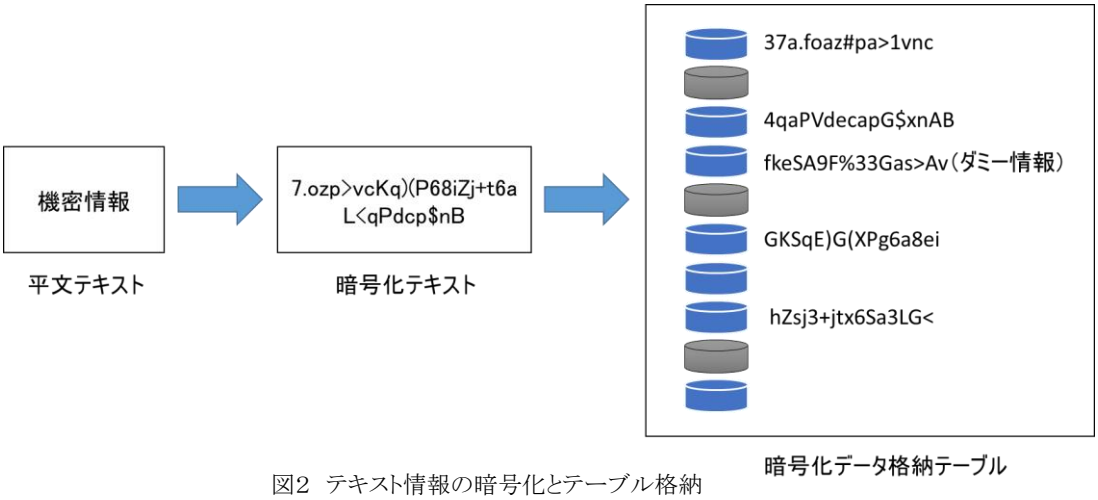

図2 テキスト情報の暗号化とテーブル格納

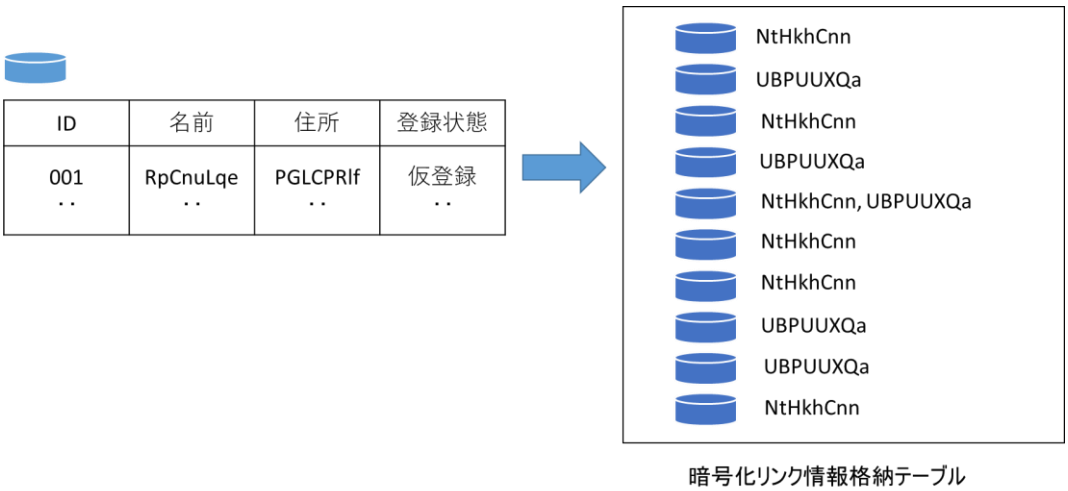

図3 暗号化テキスト情報への暗号化リンク情報