(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6956107号
(P6956107)
(45)発行日　令和3年10月27日(2021.10.27)　(24)登録日　令和3年10月6日(2021.10.6)

(51)Int.Cl.　　F I
G06F 16/25 (2019.01)　　G06F 16/25
G16H 10/60 (2018.01)　　G16H 10/60

請求項の数 24　(全 20 頁)

(21)出願番号　特願2018-553440 (P2018-553440)
(86)(22)出願日　平成29年4月19日 (2017.4.19)
(65)公表番号　特表2019-514128 (P2019-514128A)
(43)公表日　令和1年5月30日 (2019.5.30)
(86)国際出願番号　PCT/EP2017/059266
(87)国際公開番号　WO2017/182509
(87)国際公開日　平成29年10月26日 (2017.10.26)
審査請求日　令和2年4月9日 (2020.4.9)
(31)優先権主張番号　62/324,363
(32)優先日　平成28年4月19日 (2016.4.19)
(33)優先権主張国・地域又は機関
米国 (US)

(73)特許権者　590000248
コーニンクレッカ　フィリップス　エヌヴェ
KONINKLIJKE PHILIPS N.V.
オランダ国　５６５６　アーヘー　アインドーフェン　ハイテック　キャンパス　５２
(74)代理人　100122769
弁理士　笛田　秀仙
(74)代理人　100163809
弁理士　五十嵐　貴裕
(74)代理人　100171701
弁理士　浅村　敬一

最終頁に続く

(54)【発明の名称】明確な照合情報を持たない識別不能のヘルスケアデータベースの病院マッチング

(57)【特許請求の範囲】
【請求項１】
以下の動作を含むデータベース統合プロセスを実行することによって、Ｎ（Ｎは少なくとも３の値を有する正の整数）個の匿名化されたヘルスケアデータベースを統合するようにプログラムされた少なくとも１つの電子プロセッサを有し、
前記動作は、
Ｎ個の匿名化されたヘルスケアデータベースのうちの一組のデータベースｉ，ｊに関し、前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々が含まれる特徴のセットを識別するとともに、前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースの患者とマッチングする変換テーブルを生成する動作と、
Ｎ（Ｎ－１）／２個の変換テーブルを生成するために、前記Ｎ個の匿名化されたヘルスケアデータベースのデータベースの各一意の組について前記識別及び生成動作を繰り返す動作と、
を有し、
前記少なくとも１つの電子プロセッサは、前記Ｎ（Ｎ－１）／２個の変換テーブルを用いて前記Ｎ個の匿名化されたヘルスケアデータベースに含まれる１又は複数の匿名化された患者について患者データを検索する動作を含む患者データ検索プロセスを実行するように更にプログラムされる、匿名化されたヘルスケアデータソース装置。
【請求項２】
前記一組のデータベースｉ，ｊに関して前記特徴のセットを識別する動作が、特徴精度

メトリックが前記一組のデータベースｉ，ｊの各匿名化されたヘルスケアデータベースについての最小精度を満たす特徴を識別する動作を含む、請求項１記載の装置。
【請求項３】
前記Ｎ個の匿名化されたデータベースに含まれる前記患者データを検索する動作が、クエリ特徴に関し、
前記クエリ特徴が、前記Ｎ個の匿名化されたヘルスケアデータベースの１つにのみ含まれる場合、前記クエリ特徴を含む匿名化されたヘルスケアデータベースから前記クエリ特徴を検索する動作、及び、
前記クエリ特徴が、前記Ｎ個の匿名化されたヘルスケアデータベースの２つ以上に含まれる場合、前記クエリ特徴を含む各々の匿名化されたヘルスケアデータベースにおける前記クエリ特徴に関する特徴精度メトリックに基づいて、前記クエリ特徴を含む前記Ｎの匿名化されたヘルスケアデータベースのうちの２つ以上において前記クエリ特徴の値から前記クエリ特徴に関する検索された値を生成する動作、
を含む、請求項１又は２に記載の装置。
【請求項４】
前記変換テーブルを生成する動作が、ｍ（ｍは、前記一組のデータベースｉ，ｊにおいてマッチした患者の数）×２の変換テーブルを生成する動作を含む、請求項１乃至３のいずれか１項に記載の装置。
【請求項５】
前記データベース統合プロセスが、前記Ｎ（Ｎ－１）／２個の変換テーブルの間でマッチングする患者の整合性に基づいて、前記Ｎ（Ｎ－１）／２個の変換テーブルを改善する更なる動作を含む、請求項１乃至４のいずれか１項に記載の装置。
【請求項６】
前記改善する動作が、前記識別された特徴のセットを使用しない、請求項５記載の装置。
【請求項７】
前記データベース統合プロセスが、前記Ｎ個の匿名化されたヘルスケアデータベースのうちの少なくとも一組のデータベースに関し、各事象のタイムスタンプ間における時間間隔Δｔによって分離される一組のタイムスタンプされた事象によって規定される少なくとも１つの長さ方向の特徴を識別する動作、及び、
前記２つのデータベースｉ，ｊにおける患者に関する前記時間間隔Δｔの比較を含む前記長さ方向の特徴のマッチングに部分的に基づいて、前記一組のデータベースの患者にマッチングする前記変換テーブルを生成する動作、
を含む、請求項１乃至６のいずれか１項に記載の装置。
【請求項８】
前記長さ方向の特徴のマッチングに部分的に基づいて前記一組のデータベースの患者にマッチングする前記変換テーブルを生成する動作が、前記２つのデータベースｉ，ｊにおける患者に関する事象のタイムスタンプの比較を含まない、請求項７記載の装置。
【請求項９】
以下の動作を含むデータベース統合プロセスを実行することによって、ヘルスケアデータベースｉとヘルスケアデータベースｊとを統合するようにプログラムされた少なくとも１つの電子プロセッサを有し、
前記動作は、
一組のデータベースｉ，ｊに関し、事象のタイムスタンプ間の時間間隔Δｔによって分離される一組のタイムスタンプされた事象によって規定される少なくとも１つの長さ方向の特徴を含む前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々含まれる特徴のセットを識別するとともに、前記２つのデータベースｉ，ｊにおける患者に関する時間間隔Δｔの比較を含む前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースｉ，ｊの患者にマッチングする変換テーブルを生成する動作を含み、

前記少なくとも１つの電子プロセッサは、前記一組のデータベースｉ，ｊの患者とマッチングする前記変換テーブルを用いて、両方の匿名化されたヘルスケアデータベースｉ，ｊに含まれる１又は複数の匿名化された患者に関する患者データを検索する動作を含む患者データ検索プロセスを実行するように更にプログラムされる、匿名化されたヘルスケアデータソース装置。

【請求項１０】

患者類似度に基づいて前記一組のデータベースｉ，ｊの患者とマッチングする前記変換テーブルを生成する動作が、前記２つのデータベースｉ，ｊにおける患者に関する事象のタイムスタンプの比較を含まない、請求項９記載の装置。

【請求項１１】

前記特徴のセットを識別する動作が、前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに含まれる非長さ方向の特徴のセットを識別する動作を含み、各データベースｉ，ｊにおける各患者に関し、前記患者に関する前記非長さ方向の特徴のセットの値の連結を有する患者に関する統一識別子を生成する動作を含み、

前記変換テーブルを生成する動作が、前記２つのデータベースｉ，ｊにおける患者に関する前記統一識別子の比較を更に含む前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースｉ，ｊの患者とマッチングする前記変換テーブルを生成する動作を含む、請求項９又は１０に記載の装置。

【請求項１２】

前記特徴のセットを識別する動作が、前記特徴を抽出するための患者記録のテキストコンテンツについての自然言語処理を実行することによって、前記一組のデータベースｉ，ｊの少なくとも１つのデータベースにおける少なくとも１つの特徴を識別する動作を含む、請求項９乃至１１のいずれか１項に記載の装置。

【請求項１３】

前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々含まれる前記特徴のセットを識別する動作が、特徴精度メトリックが、前記匿名化されたヘルスケアデータベースｉと前記匿名化されたヘルスケアデータベースｊとの両方に関する最小精度を満たす特徴を識別する動作を含む、請求項９乃至１２のいずれか１項に記載の装置。

【請求項１４】

前記一組のデータベースｉ，ｊの患者にマッチングする前記変換テーブルを用いて両方の匿名化されたヘルスケアデータベースｉ，ｊに含まれる前記患者データを検索する動作が、クエリ特徴に関し、

前記クエリ特徴が、前記一組の匿名化されたヘルスケアデータベースｉ，ｊの１つのデータベースにのみ含まれる場合、前記クエリ特徴を含む前記匿名化されたヘルスケアデータベースから前記クエリ特徴を検索する動作、及び、

前記クエリ特徴が、前記一組の匿名化されたヘルスケアデータベースｉ，ｊの両方のデータベースに含まれる場合、前記クエリ特徴を含む各々匿名化されたヘルスケアデータベースにおける前記クエリ特徴に関する特徴精度メトリックに基づいて、前記一組の匿名化されたヘルスケアデータベースｉ，ｊにおける前記クエリ特徴の値から前記クエリ特徴に関する検索された値を生成する動作、

を含む、請求項９乃至１３のいずれか１項に記載の装置。

【請求項１５】

前記変換テーブルを生成する動作が、ｍ（ｍは、前記一組のデータベースｉ，ｊにおいてマッチした患者の数）×２の変換テーブルを生成する動作を含む、請求項９乃至１４のいずれか１項に記載の装置。

【請求項１６】

前記少なくとも１つの電子プロセッサが、Ｎ（Ｎ－１）／２個の変換テーブルを生成するために、前記Ｎ個の匿名化されたヘルスケアデータベースの各々一意の組のデータベースに関する識別及び生成動作を繰り返す更なる動作を含む前記データベース統合プロセスを実行することによって、前記匿名化されたヘルスケアデータベースｉ、前記匿名化され

たヘルスケアデータベースｊ、及び、少なくとも１つの追加的な匿名化されたヘルスケアデータベースを含むＮ個のデータベースを統合するようにプログラムされ、
前記少なくとも１つの電子プロセッサが、前記匿名化されたヘルスケアデータベースｉ，ｊの１つにおける患者の患者ＩＤを受信し、前記Ｎ（Ｎ－１）／２個の変換テーブルを用いて前記Ｎ個の匿名化されたヘルスケアデータベースに含まれる前記患者に関する患者データを検索する動作を含む前記患者データ検索プロセスを実行するように更にプログラムされる、請求項９乃至１５のいずれか１項に記載の装置。
【請求項１７】
Ｎ（Ｎは、少なくとも２の値を持つ正の整数）個の匿名化されたヘルスケアデータベースから匿名化された母集団イメージを再構成する匿名化された母集団イメージ再構成方法を実行するためのコンピュータによって読み取り可能及び実行可能な命令を格納する非一時的な記録媒体であって、前記匿名化された母集団イメージ再構成方法は、
前記Ｎ個の匿名化されたヘルスケアデータベースの一組のデータベースｉ，ｊに関し、前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々含まれる特徴のセットを識別するとともに、前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースの患者にマッチングする変換テーブルを生成する動作と、
前記Ｎ（Ｎ－１）／２個の変換テーブルによって統合される前記Ｎ個の匿名化されたヘルスケアデータベースのコンテンツを有する前記匿名化された母集団イメージを生成するために、前記Ｎ個の匿名化されたヘルスケアデータベースの各一意の組のデータベースに関して前記識別及び生成動作を繰り返す動作と、
を有する、非一時的な記録媒体。
【請求項１８】
前記格納された命令が、匿名化された母集団データクエリを受信するとともに、前記匿名化された母集団データクエリに応じて、前記Ｎ（Ｎ－１）／２個の変換テーブルを用いて前記匿名化された母集団イメージから患者データを検索する動作を含む匿名化された母集団イメージデータ検索方法を更に実行するためのコンピュータによって読み取り可能及び実行可能である、請求項１７記載の非一時的な記録媒体。
【請求項１９】
Ｎが、少なくとも３の値を持つ正の整数である、請求項１７又は１８に記載の非一時的な記録媒体。
【請求項２０】
前記変換テーブルを生成する動作が、ｍ（ｍは、前記一組のデータベースｉ，ｊにおいてマッチした患者の数）×２個の変換テーブルを生成する動作を含み、これにより、前記Ｎ（Ｎ－１）／２個の変換テーブルの各々は、ｍ×２の変換テーブルである、請求項１９記載の非一時的な記録媒体。
【請求項２１】
前記匿名化された母集団イメージ再構成方法が、前記Ｎ（Ｎ－１）／２個の変換テーブルの間でマッチングする患者の整合性に基づいて、前記Ｎ（Ｎ－１）／２個の変換テーブルを改善する更なる動作を含む、請求項１９又は２０に記載の非一時的な記録媒体。
【請求項２２】
前記改善する動作が、前記Ｎ（Ｎ－１）／２個の変換テーブルについて作用し、前記識別された特徴のセットを使用しない、請求項２１記載の非一時的な記録媒体。
【請求項２３】
前記匿名化された母集団イメージ再構成方法が、前記Ｎ個の匿名化されたヘルスケアデータベースの少なくとも一組のデータベースに関し、
事象のタイムスタンプ間の時間間隔Δｔによって分離される一組のタイムスタンプされる事象によって規定される少なくとも１つの長さ方向の特徴を識別する動作、及び、
前記２つのデータベースｉ，ｊにおける患者に関する前記時間間隔Δｔの比較を含む前記長さ方向の特徴のマッチングに部分的に基づいて、前記一組のデータベースの患者にマッチングする前記変換テーブルを生成する動作、

を含む、請求項１７乃至２２のいずれか１項に記載の非一時的な記録媒体。
【請求項２４】
前記長さ方向の特徴のマッチングに部分的に基づいて前記一組のデータベースの患者にマッチングする前記変換テーブルを生成する動作が、前記２つのデータベースｉ，ｊにおける患者に関する事象のタイムスタンプの比較を含まない、請求項２３記載の非一時的な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
以下は、概して、医療研究及び開発技術、ヘルスケアデータベースのキュレーション技術、ヘルスケアデータマイニング技術、及び、関連技術に関する。
【背景技術】
【０００２】
ヘルスケア研究開発の多くの領域は、医療患者のデータを含むヘルスケアデータベースを活用している。病院又は他の医療施設によって、及び／又は、心臓ケア機関（CCU）、集中治療室（CCU）、又は、緊急時アドミタンス部門などの個別の機関によって、病歴又は他の臨床データ、患者課金データ、病院ベッド占有率などの事項に関する管理記録が維持される。これらのデータベースには、（米国では）医療保険の相互運用性及び説明責任に関する法律（HIPAA）などの財政及び／又は医療プライバシー法の下で一般に秘密裏に維持されなければならない繊細な患者データが保存される。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
患者のプライバシーを維持しながら、患者データベースを臨床、病院管理、又は、他の目的のためのデータ分析に使用できるようにするために、患者識別情報（PII）を取り除くことによってデータベースを匿名化することが知られている。匿名化する必要のある情報には、患者名及び／又は医療識別番号（適切にランダムに割り当てられた番号などで置き換えられる）、住所などが含まれる。他の匿名化手段には、珍しい特徴の組み合わせによって識別可能な「まれな」患者の除去が含まれてもよい。例えば、特定の病気の102歳の患者は、その情報のみに基づいて識別され得るためである。
【０００４】
まれな患者に加えて、患者は、患者記録に記録された事象のタイムスタンプ情報に基づいて識別可能であるかもしれない。例えば、患者が特定の状態で特定の日付に病院に入院した場合、その情報は、可能性のある患者識別の数を少数に減らすのに十分であり得る。しかしながら、長さ方向の情報、即ち、事象の時間的順序及び種々の事象間の時間間隔は、医療データ解析において有用であることがある。例えば、入院と退院との間の時間間隔は、病院の効率及び／又は特定の治療の有効性を分析するために有用であり、又は、さらに重要であり得る。タイムスタンプを使用して匿名化された患者を特定するための潜在的可能性を減らし、医療データ分析の価値のある可能性のある長さ方向の情報を保持するために、一部の匿名化されたデータベースでは、タイムスタンプが、所与の患者の全てのタイムスタンプされた事象のための固定シフトを用いて、ランダムな量（一般に各患者ごとに異なる）だけシフトされる。タイムスタンプのランダムな固定時間シフトは、タイムスタンプによる患者の識別をより困難にするが、特に固定時間シフトの使用は、長さ方向の情報、即ち、事象間の時間間隔情報を保持する。
【課題を解決するための手段】
【０００５】
本開示の一態様では、匿名化されたヘルスケアデータソース装置が、以下の動作を含むデータベース統合プロセスを実行することによって、Ｎ（Ｎは少なくとも３の値を有する正の整数）個の匿名化されたヘルスケアデータベース１０を統合するようにプログラムされた少なくとも１つの電子プロセッサを有し、前記動作は、Ｎ個の匿名化されたヘルスケ

アデータベースのうちの一組のデータベースｉ，ｊに関し、前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々が含まれる特徴のセットを識別するとともに、前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースの患者とマッチングする変換テーブルを生成する動作と、Ｎ（Ｎ－１）／２個の変換テーブルを生成するために、前記Ｎ個の匿名化されたヘルスケアデータベースのデータベースの各一意の組について前記識別及び生成動作を繰り返す動作と、を有する。上記少なくとも１つ少なくとも１つの電子プロセッサは、Ｎ（Ｎ－１）／２個の変換テーブルを用いて前記Ｎ個の匿名化されたヘルスケアデータベースに含まれる１又は複数の匿名化された患者について患者データを検索する動作を含む患者データ検索プロセスを実行するように更にプログラムされる。

【０００６】

本開示の他の態様では、以下の動作を含むデータベース統合プロセスを実行することによって、ヘルスケアデータベースｉとヘルスケアデータベースｊとを統合するようにプログラムされた少なくとも１つの電子プロセッサを有し、前記動作は、前記一組のデータベースｉ，ｊに関し、事象のタイムスタンプ間の時間間隔Δｔによって分離される一組のタイムスタンプされた事象によって規定される少なくとも１つの長さ方向の特徴を含む前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々含まれる特徴のセットを識別するとともに、前記２つのデータベースｉ，ｊにおける患者に関する時間間隔Δｔの比較を含む前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースｉ，ｊの患者にマッチングする変換テーブルを生成する動作を含む。上記少なくとも１つの電子プロセッサは、一組のデータベースｉ，ｊの患者とマッチングする前記変換テーブルを用いて、両方の匿名化されたヘルスケアデータベースｉ，ｊに含まれる１又は複数の匿名化された患者に関する患者データを検索する動作を含む患者データ検索プロセスを実行するように更にプログラムされる。

【０００７】

本開示の他の態様では、非一時的な記録媒体が、Ｎ（Ｎは、少なくとも２の値を持つ正の整数）個の匿名化されたヘルスケアデータベースから匿名化された母集団イメージを再構成する匿名化された母集団イメージ再構成方法を実行するためのコンピュータによって読み取り可能及び実行可能な命令を格納する。匿名化された母集団イメージ再構成方法は、前記Ｎ個の匿名化されたヘルスケアデータベースの一組のデータベースｉ，ｊに関し、前記一組のデータベースｉ，ｊの両方のデータベースｉ，ｊに各々含まれる特徴のセットを識別するとともに、前記特徴のセットによって測定される患者類似度に基づいて、前記一組のデータベースの患者にマッチングする変換テーブルを生成する動作を有する。また、Ｎ（Ｎ－１）／２個の変換テーブルによって統合される前記Ｎ個の匿名化されたヘルスケアデータベースのコンテンツを有する前記匿名化された母集団イメージを生成するために、前記Ｎ個の匿名化されたヘルスケアデータベースの各一意の組のデータベースに関する識別及び生成動作が繰り返される。

【０００８】

１つの利点は、２つ、３つ、４つ、又は、それ以上の匿名化されたヘルスケアデータベースの統合を提供し、ヘルスケアデータ分析タスクのためにデータベースに含まれる結合データを活用することにある。

【０００９】

別の利点は、１又は複数の匿名化されたヘルスケアデータベースが非構造化ヘルスケアデータベースである前述のものを提供することにある。

【００１０】

別の利点は、異なる匿名化されたヘルスケアデータベース内の匿名化された患者を照合する際に、長さ方向の情報、即ち事象間の時間間隔が活用される前述のものを提供することにある。

【００１１】

所与の実施形態は、本開示を読んで理解すると当業者に明らかになるであろうように、

前述の利点のうちの１つ、２つ、それ以上、又は、全てを提供することができ、及び／又は、他の利点を提供し得る。

【図面の簡単な説明】

【００１２】

本発明は、様々な構成要素及び構成要素の配置、並びに、様々なステップ及びステップの配置における形態をとることができる。図面は、好ましい実施形態を説明するためのものに過ぎず、本発明を限定するものとして解釈されるべきではない。

【図１】図１は、２以上の匿名化されたヘルスケアデータベースから統合された匿名化された患者データを活用する医療分析装置を概略的に図示している。

【図２】図２は、３以上の匿名化されたヘルスケアデータベースを統合するように構成された図１の装置によって実行されるデータベース統合プロセスの一実施形態を概略的に図示している。

【図３】図３は、異なる匿名化されたヘルスケアデータベースを統合する異なる特徴の選択の基準を模式的に示す表を示している。

【図４】図４は、図２のデータベース統合プロセスの実施形態の改善コンポーネントの動作を図式的に示している。

【図５】図５は、長さ方向の情報を利用する図１のデータベース統合プロセスの実施形態を図式的に示している。

【発明を実施するための形態】

【００１３】

匿名化されたヘルスケアデータベースの統合には多くの課題がある。様々な匿名化されたヘルスケアデータベースは、任意の２つのデータベース間では一部のみが重複するデータであり、範囲が大きく異なる場合がある。実際、この部分的な重複は、複数の匿名化されたヘルスケアデータベースを統合したいという、１つのデータベースに欠落している情報を他のデータベースのコンテンツで「埋める」という願望の重要な動機となっている。例えば、本明細書で使用されるように、「匿名化されたヘルスケアデータベース」は、総合的な電子医療記録（EMR）あるいは心血管系情報システム（CVIS）又は集中治療機関（ICU）情報システムなどのドメイン特有の医療データベースから抽出された匿名化されたデータベースなどの医療記録データベース、病院課金部門データベースから抽出された匿名化データベース、医療保険会社データベースから抽出された匿名化されたデータベース、病院入院部門データベースから抽出された匿名化されたデータベースなどであってもよい。CVISから抽出された匿名化されたデータベースは、心臓血管疾患の診断及び治療に関する医療記録を含むことが期待されるが、それらの診断／治療のための保険適用範囲に関する情報は含まれない可能性がある。対照的に、病院課金部門から抽出された匿名化されたデータベースは、保険償還情報を含むが、医療診断／治療データは含まないことが期待される。これらのデータベースを組み合わせることで、患者集団の全体像を提供することができる。しかしながら、統合の動機づけを提供する２つのデータベース間の限られたコンテンツの重複もまた、そのような統合を困難にする。

【００１４】

本明細書に開示される様々な実施形態において、複数（３以上）のヘルスケアデータベースの統合を活用することによって、これらの課題が克服される。これは、単一のプロセスでＮ個のデータベースの統合を行なうことへの動機付けとなる、全体的な重複の程度を大きくすることができる。しかしながら、逆説的に、統合を実行するためのより効率的で信頼性のある手法は、各組の変換テーブルを生成するために、最初に、一組の匿名化されたヘルスケアデータベースを統合し、その後、結果として生じるＮ（Ｎ－１）／２個の変換テーブルの間の患者マッチングの一貫性に基づいて、Ｎ（Ｎ－１）／２個の変換テーブルを改善することが開示される。この手法は、Ｎ個のデータベース間の特徴の重なりが小さくなる可能性があることを認識し、さらに、重複が存在する場合でさえ、幾つかのデータベースでは特定の特徴が信頼できない可能性がある。一組のデータベースを最初に統合する開示された手法を採用することによって、その一組の匿名化されたヘルスケアデータ

ベースに対してうまく選択されたそのようなペアワイズ統合ごとに、一連の特徴を選択することができる。次いで、複数の（Ｎ＞２）データベースによって提供される追加情報は、幾つかの実施形態では、特徴に依存しない後続の精緻化ステップにおいて活用される。

【００１５】

これに加えて、又は、これに代えて、本明細書に開示された実施形態では、これらの問題は、長さ方向の情報、即ち事象の時間的順序及び様々な事象間の時間間隔を利用することによって克服される。一般に、長さ方向の特徴は、事象のタイムスタンプ間の時間間隔Δｔによって分離された、匿名化されたヘルスケアデータベース内の単一の匿名化された患者に対する一組のタイムスタンプ付き事象によって定義される。このような長さ方向の特徴は、匿名化プロセスが各患者に関して全てのタイムスタンプのランダムではあるが固定のシフトを導入する匿名化されたヘルスケアデータベースにおいても明確に定義されている。これは、固定時間シフトは、事象間の時間間隔Δｔに影響を与えないためである。

【００１６】

図１によれば、Ｎ個の匿名化されたヘルスケアデータベース１０は、「データベース１」、「データベース２」、・・・、「データベースＮ」とそれぞれ付されている。一般に、Ｎは、少なくとも２である、幾つかの実施形態では少なくとも３である、正の整数である。下限であるＮ＝２が、幾つかの実施形態で想定される。匿名化されたヘルスケアデータベース１０は、幾つかの実施形態では１００万人以上の患者エントリを有するような大規模データベースを匿名化するために、好ましくは自動化された（例えば、特定のクラス又はタイプのデータを除去するようにプログラムされたコンピュータによりコンピュータ実装された）適切な匿名化プロセス（図示せず）によって生成される。オプションで、匿名化は、例えば、特定の稀少な患者を除去するため、又は、他の異常な状況に対処するために、何らかの手動処理を含むこともできる。Ｎ個の匿名化されたデータベースを生成するために使用される匿名化プロセスは、一般に異なり、及び／又は、同じ情報を匿名化してもしなくてもよい。各匿名化プロセスは、好ましくは、患者名、患者住所、社会保障番号などの患者を直ちに識別することができる個人識別情報（PII）、及び、病院の名前、郵便番号などの他の情報と組み合わせてPIIとなる可能性のある情報を匿名化する。情報が他の情報と組み合わせてPIIとなり得る場合、その組み合わせの一部のみを匿名化することで十分である。例えば、郵便番号、性別、及び、生年月日の組み合わせは、個人的に識別することができるが、郵便番号情報のみを匿名化することによって、受け入れ可能な患者の匿名性を達成することができる。匿名化プロセスは、任意に、ある特定の患者を識別することができるまれな情報、例えば、例えば、90歳などのある特定の最大年齢を超える任意の年齢、及び／又は、一般的な診断のリストの中にない診断を除去することもできる。

【００１７】

一般に、特定のデータの匿名化は、データを削除する（修正する）か、又は、データをプレースホルダで置き換えることによって行なうことができる。後者は、匿名化ではあるが、その特定のタイプの情報との相関が望ましく保持される場合に好適である。例えば、医療機関（例えば、病院又はケア機関）のエントリは、データベースに対して内部的に一貫したプレースホルダによって置き換えられ得る。これらのプレースホルダは、所与のデータベース内で内部的に一貫しているが、基本的にデータベース間でランダムに変化する。例えば、データベース１では、病院「Blackacre総合病院」は、常にプレースホルダ（例えば、「8243」）で置き換えられ得る。一方、「Whiteacre地域医療センタ」は、常にプレースホルダ「1238」で置き換えられ得る。この例では、データベース１の医療機関「Blackacre総合病院」の全てのインスタンスが、（同じ）プレースホルダ医療機関「8243」に置き換えられ、データベース１の医療機関「Whiteacre地域医療センタ」の全てのインスタンスが、（同じ）プレースホルダ医療機関「1238」で置き換えられる。一方、データベース２の例を続けると、データベース２の医療機関「Blackacre総合病院」の各インスタンスは、（匿名化されたデータベース１におけるBlackacreに対して使用されたプレースホルダ「8243」とは異なる）同一のプレースホルダ医療機関「EADF」によって置き換

えられることができ、「Whiteacre地域医療センタ」の各インスタンスは、（匿名化されたデータベース１におけるWhiteacreに対して使用されたプレースホルダ「1238」とは異なる）同一のプレースホルダ医療機関「JSDF」によって置き換えられることができる。匿名化されたデータベース内で内部的に一致する医療機関プレースホルダによる医療機関の上記匿名化により、データベース上で動作するヘルスケアデータ分析プロセスが、患者の匿名性を維持しながら特定の医療機関との相関を識別することが可能になる。例えば、Blackacreが平均的な病院よりも心臓移植の成功率が統計的に有意に高い場合、これは、匿名化された病院「8243」で行なわれた心臓移植の統計的に有意に高い成功率として、データベース１（心臓移植結果データを格納していると仮定）に表示される。

【００１８】

一方、一部の情報は、修正、つまり削除によって匿名化される場合がある。例えば、住宅アドレス情報は、完全に修正されることができ、これは識別性が高く、典型的なヘルスケアデータ分析プロセスでは、住宅アドレスとの有用な相関が期待できないためである。変形実施形態では、住宅アドレス相関が健康管理データ分析プロセスの有用な入力であると予想される場合、住所の匿名化は、各住宅アドレスをより地理的エリア（例えば、この都市が許容可能なレベルの匿名性を保証するのに十分な人口を有している場合には、住宅都市）で置き換えることによって実行されてもよい。人口が十分に少ない住宅都市又は郡は、個人的に識別できる「まれな」データを保持しないように完全に修正され、又は、居住国などの適切なより広い地理的単位で置き換えられてもよい。

【００１９】

匿名化されたヘルスケアデータベース１０は、一般に、例えばリレーショナルデータベースフォーマット又は他の構造化データベースフォーマットのような幾つかの構造化フォーマットで、スプレッドシート、検索可能な列区切りのリッチテキストファイルなどのようにそれぞれフォーマットされることが期待される。しかし、幾つかの実施形態では、データベース１０のうちの１又は複数は、例えば患者に書かれたテキストレポートを格納するような非構造化データベースであってもよく、又は、限定された構造（例えば、患者名や人口統計情報などの情報に構造化されていないテキストレポートが続く構造化見出しなど）を有してもよい。そのような場合、自然言語処理（NLP）を使用して、データベースコンテンツの構造化された表現（例えば、テキスト文書の単語の表現）を抽出することができる。

【００２０】

図１に示されるように、ヘルスケアデータ分析装置は、コンピュータ１４（又は、より一般的には電子プロセッサ１４）上で実装される匿名化されたヘルスケアデータソース装置１２を含み、これは、例えば、ネットワークベースのサーバコンピュータ、クラウドコンピューティングリソース、サーバクラスタなどであってもよい。コンピュータ１４は、データベース統合プロセス１６及び患者データ検索プロセス１８を実行するようにプログラムされ、後者は、Ｎ（Ｎ－１）／２個の変換テーブル２０のセットを使用する。本明細書の例示的な実施形態では、各変換テーブルは、Ｎ個のデータベース１０のうちの一組のデータベースに関するｍ×２変換テーブルである。一般性を失うことなく、一組のデータベースをそれぞれデータベースｉとデータベースｊとし、集合的に一組のデータベース（ｉ，ｊ）を形成する。各変換テーブルは、データベース統合プロセス１６によって一組のデータベース（ｉ，ｊ）にマッチングされたｍ人の患者の行（又は、代替的に列）と、匿名化されたデータベースｉにおける匿名化された患者ＩＤと、匿名化されたデータベースｊにおける匿名化された患者ＩＤとをリスト化している２つの列（又は、代替的に行）と、を有するｍ×２のテーブルである。Ｎ＝２の場合、単一の組のデータベース（ｉ，ｊ）が存在する。Ｎ＞２の場合、Ｎ（Ｎ－１）／２個の一意なデータベースの組（ｉ，ｊ）が存在する。これは、ｎのセットからｋ個の要素をとる組み合わせ数の組み合わせ式を使用して取得され得る。

【数１】

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$$

現在のケースでは、組が図示されているため、ｋ＝２であり、セットは、Ｎ個の匿名化されたヘルスケアデータベース１０であるので、ｎ＝Ｎである。従って、組み合わせは、Ｎ（Ｎ－１）／２まで減少する。一般に、Ｎ＞２である場合、データベースの組（ｉ，ｊ）毎に一致する患者数ｍは異なるかもしれないが、データベースの組間の患者の重複は、３以上の匿名化されたヘルスケアデータベースの有用な統合に対して期待される。

【００２１】

Ｎ（Ｎ－１）／２個の変換テーブル２０は単一のテーブルとして具体化されることが考えられる。例えば、単一のｍ×［Ｎ（Ｎ－１）］テーブルを形成するために、各々次元ｍ×２のＮ（Ｎ－１）／２個テーブルを連結したものである。この場合、全てのＮ（Ｎ－１）／２個の構成要素であるｍ×２変換テーブルがマッチされる患者数ｍを有すると仮定する。そうでない場合、パディングを使用して「欠落した」匿名化患者を説明することができる。例えば、データベース１の患者４９がデータベース３に一致がない場合、組（ｉ，ｊ）=（１，３）の構成要素であるｍ×２変換テーブルは、<ヌル（null）>又はゼロ又は他のプレースホルダによって適切に埋められる。

【００２２】

また、コンピュータ１４は、匿名化された患者データをＮ（Ｎ－１）／２個の変換テーブル２０を用いてＮ個の匿名化されたヘルスケアデータベース１０から検索するための患者データ検索プロセス１８を実行するようにプログラムされる。例えば、クエリは、データベース１において使用される匿名化された患者ＩＤによって識別される所与の患者に関するクエリ特徴の値を取得するため、患者データ検索プロセス１８の対象となり得る。この患者ＩＤは、データベース１からクエリ特徴の値を検索するために直接的に使用され得るが、データベースｊ＝２，・・・，Ｎの各々に関し、データベースの組（１，ｊ）のための適切な変換テーブルが、データベースｊからクエリ特徴値を検索するために、データベースｊにおいて患者ＩＤをマッチングすべく使用される。

【００２３】

しかしながら、一般的に、クエリ特徴は、Ｎ個全てのデータベースに含まれない可能性がある。クエリ特徴が、Ｎ個の匿名化されたヘルスケアデータベースのうちの１つのみに含まれる場合、クエリ特徴は、クエリ特徴を含む（単一の）匿名化されたヘルスケアデータベースから検索される。一方、クエリ特徴が、Ｎ個の匿名化されたヘルスケアデータベースのうちの２以上に含まれる場合、検索された値は、クエリ特徴に関して、クエリ特徴を含む、Ｎ個の匿名化されたヘルスケアデータベースのうちの２以上のデータベースにおけるクエリ特徴の値から生成される。これは、例えば、クエリ特徴を含む各々の匿名化されたヘルスケアデータベースにおけるクエリ特徴に関する特徴精度メトリックを用いて実行され得る。例えば、クエリが、患者４９に関する一次診断を要求するとともに、データベース１，２，３が一次診断フィールドをそれぞれ含む場合、（適切なｍ×２変換テーブルを用いたデータベース２，３に関する匿名化された患者ＩＤ４９の変換後に）患者４９の一次診断に関する３つの値が提供される。データベース１，３が、一次診断に関して９７％の精度を持つ一方、データベース２が、この特徴に関して極めて低い精度（例えば７１％）を持つことが知られている場合、検索される値は、最も精度が高い可能性のあるデータベース１，３から取得された一次診断として生成される。異なるデータベースが所与のクエリ特徴に関して異なる値を格納する場合、その特徴に関する最も高い精度メトリックを有するＮ個のデータベース１０のデータベースの値を取る、又は、最も一般的な値を取る（例えば、６個のデータベースがその特徴に関する値を持ち、そのうち５個のデータベースが一致する場合、６個のデータベースのうちの５個のデータベースにおいて現れる値が選択され得る）、又は、それらの値の平均値（あるいは、その特徴の精度メトリック

が最も高いデータベースの一部のサブセットの値の平均値、又は、識別可能な異常値を削除した後の平均値）を有する数値をとるなど、様々な手法を使用して、検索値が生成される。

【００２４】

患者データ検索プロセス１８によって検索及び処理されるクエリは、クエリの目的に依存して変化し得る。例えば、３０歳～５０歳の全ての男性患者について一次診断を得ることが望ましい場合がある。この場合、クエリは、年齢及び性別による適切なフィルタリングの後で、一次診断のセットの要求として定式化（それぞれ異なる診断に関する列挙）され得る。この場合のクエリ結果は、各要素（診断、カウント）が診断を示す文字列と、その診断を有する患者数（年齢／性別フィルタリング後）を格納するデータの組｛（診断、カウント）｝であってもよい。Ｎ個のデータベース１０がリレーショナルデータベースである場合、患者データ検索プロセス１８は、SQLクエリを受け取るSQL（Structured Query Language）クエリエンジンとして実装され得る。

【００２５】

引き続き図１を参照すると、ヘルスケアデータ分析装置は、コンピュータ２４（又は、より一般的には電子プロセッサ２４）上に実装されたヘルスケアデータ分析ツール２２を更に含み、これは、例えば、ネットワークベースのサーバコンピュータ、クラウドコンピューティングリソース、サーバクラスタ、デスクトップコンピュータ（図示のように）などであってもよい。コンピュータ２４は、例示的なキーボード２８、マウス又は他のポインティングデバイス３０、ディスプレイ２６のタッチセンシティブオーバレイなどの１又は複数のディスプレイコンポーネント／装置２６並びに１又は複数のユーザ入力コンポーネント／装置を含むか、それらと動作可能に接続される。ヘルスケアデータ分析ツール２２は、（一例として）特定の医療処置に関する保険適用範囲の評価、医療処置の生存率決定、患者に最も一般的に提供される医療ケアの種類との人口統計学的相関についての評価などの様々なヘルスケア分析を実行する。適切な実施形態では、ユーザは、ユーザ入力デバイス２８，３０を操作して、実行される分析のタイプを構成する。例えば、ヘルスケアデータ分析ツール２２は、匿名化されたヘルスケアデータソース装置１２の患者データ検索プロセス１８を介して匿名化されたデータベース１０から適切なデータを取り出し、そのデータについて選択された分析分析を実行し、その結果は、ディスプレイコンポーネント２６上に、図形表現などとして提示される。例えば、処置の保険適用範囲を日付間隔で格納されたヒストグラムとしてプロットするか、又は、異なる保険会社に対応するスライスを含む処置の保険適用範囲を示す円グラフとしてプロットするか、又は、地理的位置の関数として生存率をプロットするなどである。

【００２６】

例示的な匿名化されたヘルスケアデータソース装置１２が、コンピュータ１４上に実装されるように図１に示されている。一方、ヘルスケアデータ分析ツール２２が、異なるコンピュータ２４上に実装されるように図１に示されている。しかしながら、他の実施形態では、匿名化されたヘルスケアデータソース装置とヘルスケアデータ分析ツールとが、単一のコンピュータ上に実装されてもよい。また、他のハードウェアセグメンテーショントポロジも考えられる。例えば、データベース統合プロセス１６及び患者データ検索プロセス１８は、異なるコンピュータ上で実装され得る。さらに、本明細書に記載のヘルスケアデータ解析装置の開示された機能は、開示された機能を実行するために電子プロセッサ１４，２４によって読み取り可能且つ実行可能な命令を記憶する非一時的な記憶媒体として実施され得ることが理解されよう。非一時的な記憶媒体は、例えば、ハードディスクドライブ又は他の磁気記憶媒体、光ディスク又は他の光学記憶媒体、フラッシュメモリ、読み出し専用メモリ（ROM）、あるいは、他の電子記憶媒体、それらの様々な組み合わせなどを有していてもよい。

【００２７】

図２を参照して、Ｎ＞２の場合のデータベース１０に対するデータベース統合プロセス１６の一実施形態を説明する。この実施形態では、Ｎは少なくとも３であり、より一般的

には、Ｎは３以上の任意の正の整数であり得る。動作４０において、Ｎ個のデータベース１０から、一組（第１の組）の匿名化されたヘルスケアデータベース（ｉ，ｊ）が選択される。一手法では、ｉ及びｊの値は、最初はそれぞれ１と２とに設定され、ｉとｊとの全ての組の組み合わせが選択されるまで各次の反復して変更される。ここで、１＜ｉ＜Ｎ、且つ、１＜ｊ＜Ｎである（インデックス１，・・・，Ｎを用いて、Ｎ個のデータベース１０の構成データベースを表している）。データベースの組（ｉ，ｊ）は２つの異なるデータベースを統合するので、その組は、ｉ＝ｊである全ての縮退ケースを除外する。

【００２８】

以下、選択されたデータベース（ｉ，ｊ）における患者マッチングについての一例が説明される。動作４２において、一致／除外基準を適用して、マッチさせるデータベース部分を選択する。データベースｉ及びデータベースｊからの患者記録を照合するために、おそらく関連性のある２つのデータベースのサブセットが抽出される。例えば、データベースｉが医療外科及び熱傷ICU患者のデータのみをカバーする場合、データベースｊから、入院時に医療外科及び熱傷ICU病棟に入院した患者のサブセットが抽出される（即ち、含まれる）一方、データベースｉと重複しない他のエリアからのデータは除外される。除外された／包含されたデータは、特定のデータベースの組（ｉ，ｊ）についての重複によって決定され、異なる組に対して異なることがあることに留意されたい。

【００２９】

動作４４において、特徴のセットが、データベース組（ｉ，ｊ）を統合する際の使用のために識別される。ここで、データベースｉ及びデータベースｊを確実に統合するための一意でない識別の特徴のセットが選択される。選択された特徴は、それぞれ、データベース組（ｉ，ｊ）のデータベースｉとｊとの両方に含まれる。さらに、選択された特徴は、信頼性に関する利用可能な情報に基づいて任意に選択される。例えば、データベースの１つが患者の性別記録に関して比較的不正確であるが、データベースｉとデータベースｊとの両方が体重記録の点で正確であることが分かっている場合、体重を特徴として適切に選択し、性別は、特徴として適切に選択されない。

【００３０】

図３を簡単に参照すると、所与のデータベースの組（ｉ，ｊ）を統合するために選択される特徴のセットは、一般に、特定のデータベースｉ及びｊに依存することに留意されたい。例えば、図３は、３つの匿名化されたヘルスケアデータベースＸ、Ｙ、及び、Ｚの特徴の表を示しており、各データベースの各特徴の確度をパーセンテージとして表にしている。図３に示す表の最後の３行は、各特徴を、指示されたデータベースの組み合わせｉ－ｊの特徴の集合として選択すべきかどうかを示している。例えば、図３は、データベースＸ及びＹが、人種、死亡率、滞在の長さ、年齢、及び、体重の記録において正確であることを示しているので、これらの５つの特徴は、データベースＸ及びＹの照合のために選択される。同様に、人種、滞在期間、年齢、一次診断、体重などの特徴セットが、データベースＸとデータベースＺとを統合するために適切に選択される。また、性別、人種、滞在期間、年齢、体重などの特徴セットが、データベースＹとデータベースＺとを統合するために適切に選択される。図３の例では、精度パーセンテージは、特徴精度メトリックを形成し、サンプリング（患者の代表的なサンプリングを選択し、サンプルの特徴精度を検証すること）に基づいて、又は、明らかに誤った特徴値（例えば、年齢＝０又は年齢＝２００）に基づいて、又は、誤った特徴値（「エラー」として各々の欠落した特徴値をとる）に基づいて、生成され得る。

【００３１】

再び図２によれば、動作４６において、動作４４で選択された特徴のセットが、データベースｉ及びｊ内の患者をマッチングするために使用される。様々な手法を用いることができる。単純な手法では、特徴のセットの特徴の使用可能な値の閾値分数（又は数）が一致する場合、データベースｉとデータベースｊとのそれぞれにおける２人の患者の間で一致が見つかる。オプションで、マッチングは、データベースに誤って記録された特徴値を有する可能性、特徴の選択性などの要因に基づいて、異なる特徴に異なる重みを適用する

ことができる。本質的に、データベースｉの各患者は、動作４４で選択された特徴のセットの値を記憶する要素を持つ特徴ベクトルによって表され、同様に、データベースｊの各患者は、動作４４で選択された特徴のセットの値を記憶する要素を持つ特徴ベクトルによって表される。これらの値の幾つかは空白であってもよい（例えば、ベクトルは<null>又は他のプレースホルダを格納する）。２つのそのような特徴ベクトルの類似性を計算するための任意の手法を使用して、患者を比較し、２つのデータベース内の同様の患者を識別することができる。例えば、特徴の数がＦである場合、適切な類似性計算は、次式によって与えられる２つの特徴ベクトルp_iとp_jとの間の差であってもよい。

【数２】

$$D(p_i, p_j) = \frac{1}{F}\sum_{f=1}^{F} w_f\left(p_i(f) - p_j(f)\right)^2$$

ここで、p_i及びp_jは、データベースｉにおいて比較される患者とデータベースｊにおいて比較される患者とをそれぞれ表す特徴ベクトルであり、p_i（ｆ）は、患者p_iに関するf^{th}番目の特徴の値を表し、同様に、p_j（ｆ）は、患者p_jに関するf^{th}番目の特徴の値を表す。パラメータw_fは、様々な特徴ｆ＝１，・・・，Ｆの相対的重要性を示すとともに、（必要に応じて）異なる特徴タイプを共通の単位に変換してその和を計算可能にするために選択される特徴重み及び／又は単位変換係数である。この式では、Ｄ（p_i，p_j）の値が小さいほど、より似ている患者を示し、結果、Ｄ（p_i，p_j）が幾らかの閾値よりも小さい場合に、２人の患者はマッチし得る。任意の欠落した特徴は、様々な方法で処理され得る。例えば、Ｄ（p_i，p_j）を形成する和からそれらを単純に省いたり（及び、然るべく１／Ｆスケーリングしたり）、又は、欠落している特徴ｆの場合、p_i（ｆ）－p_j（ｆ）の幾らかの初期値を割り当てたりする。上記は単なる例示であり、実質的に他の比較形式を使用して、それぞれのデータベースｉ及びｊにおける一致する患者を識別することができることを理解されたい。

【００３２】

動作４８では、動作４６で識別されたデータベース間の患者の一致が、データベース対（ｉ，ｊ）の患者ＩＤ変換表に集計される。例えば、この表は、次のようなｍ×２の表であり得る。

表１：データベース組（ｉ，ｊ）に関する患者ＩＤ変換表

【表１】

患者ＩＤ（データベースi）	患者ＩＤ(データベースj)
1	43
2	17
3	＜ヌル＞
4	98
5	2
5	3
6	6
…	
96	9
＜ヌル＞	6
＜ヌル＞	9
＜ヌル＞	23

ここで、この例では、データベースｉの患者ＩＤ＝３はデータベースｊに一致せず、同様に、患者ＩＤ＝６、ＩＤ＝９、及び、ＩＤ＝２３はデータベースｉに一致しないことに留意されたい。表１の例示的な例は、データベースｉの患者ＩＤによってソートされているが、データベースｊの患者ＩＤによるソートを実行することは自明である。そうすれば、表をより効率的に読み取ることができる（例えば、図１の患者データ検索プロセス１８によって受信されるクエリがデータベースｊ内の患者ＩＤによって索引付けされる）。

【００３３】

幾つかの実施形態では、患者のマッチングは排他的ではないことに留意されたい。これは、データベースｉの患者ＩＤ＝５がデータベースｊの患者２とデータベースｊの患者３との両方と一致する表１に示されている。このオプションの非排他性は、患者のマッチングにおける不確実性を捕捉することを可能にする。医療データ分析アプリケーションの場合、上記の非排他的マッチングは、そのような不確実なマッチングの数が比較的少ない場合には必ずしも問題ではなく、そのような場合、このように複数のマッチングを許容すると統計的な全体的な精度が向上する。表１に示されるデータベース（ｉ，ｊ）の例示的な変換テーブルでは、ストレージは、データベースｉの患者ＩＤ又はデータベースｊの患者ＩＤのうちのいずれかのテーブルのソートを容易にする利点を有する、データベースｉの患者ＩＤ５の重複エントリによるものである。

【００３４】

決定動作５０では、一意のデータベース組（ｉ，ｊ）毎に患者ＩＤ変換テーブルを生成するために、統合されているＮ個のデータベース１０のセット内の各一意の一組のデータベース（ｉ，ｊ）について、処理を繰り返す。従って、このループは、Ｎ個のデータベースから得られるＮ（Ｎ－１）／２個の一意のデータベース組に対するＮ（Ｎ－１）／２個の変換テーブルを生成するためにＮ（Ｎ－１）／２回実行される。例えば、Ｎ＝３の場合、組（１，２）、組（１，３）、組（２，３）の３つの繰り返しがある。他の例示的な例として、Ｎ＝５の場合、（１，２）、（１，３）、（１，４）、（１，５）、（２，３）、（２，４）、（２，５）、（３，４）、（３，５）、（４，５）の１０個の繰り返しがある。決定動作５０により実施されるループは、例えば、ネストされたループｉ＝１ to（Ｎ－１）；ｊ＝ｉ＋１ to Ｎ（ここで、ｊは、内側のループである）によって実施され得る。

【００３５】
Ｎ（Ｎ－１）／２回のループ反復の出力は、Ｎ個のデータベース１０のうちのＮ（Ｎ－１）／２個の一意のデータベース組に対するＮ（Ｎ－１）／２個の変換テーブルとなる。幾つかの実施形態では、これは、患者データ検索プロセス１８によって使用されるＮ（Ｎ－１）／２個の変換テーブル２０（それぞれｍ×２の次元）を提供する最終出力である。しかし、この時点でデータベース統合プロセス１２が終了する場合、複数（３つ以上）のヘルスケアデータベース（即ち、Ｎ≧３）からの情報は、個々のｍ×２ペアワイズ変換テーブルを改善するために効果的に活用されない。
【００３６】
引き続き図２を参照するとともに図４を更に参照すると、例示的な実施形態では、Ｎ（Ｎ－１）／２個の変換テーブルが構築された後に改善動作５２が実行される。これは、Ｎ（Ｎ－１）／２個の変換テーブル間の患者マッチングの整合性に基づいて、Ｎ（Ｎ－１）／２個の変換テーブルを改善する。例示的な実施形態では、改善動作５２は、動作４４の反復で識別された特徴のセットを使用しない。むしろ、改善動作５２は、Ｎ（Ｎ－１）／２個の変換テーブル間で予想される整合性を考慮することによって、図４に図式的に示すように実行される。図４の例では、各円は、匿名化された患者ＩＤ（例えば、「患者１」は匿名化ID＝１でラベル付けされている）及びデータベース（この例ではＸ、Ｙ、又は、Ｚ）によりラベル付けされた単一の匿名化された患者を表している。異なるデータベースの患者を結ぶ実線又は破線は、動作４２，４４，４６，４８のペアワイズマッチングによって見出される可能性のあるマッチングを示す。この例では、データベースＸの患者１は、Ｘ－Ｙ変換テーブルに基づいて、データベースＹの患者２２にリンクされる。一貫性を維持するために、データベースＸの患者１とデータベースＹの患者２２との両方をデータベースＺの同じ患者にリンクする必要がある。しかし、組ｉ＝Ｘ、ｊ＝Ｚに対するペアワイズマッチングプロセスでは、データベースＸの患者１は、データベースＺの患者７２及び患者３１との両方にマッチングされる（このような非排他的マッチングは、表１の例におけるデータベースｉの患者５の例について上述のように、許容され得る）。組ｉ＝Ｙ、ｊ＝Ｚに対するペアワイズマッチングプロセスでは、データベースＹの患者２２は、データベースＺの患者７２及び患者１４の両方に一致した。自己整合性を維持するために、データベースＸの患者１とデータベースＹの患者２２とがデータベースＺの患者７２と一致しなければならず、他の可能性のある一致は不一致である。従って、改善動作５２では、データベースＸの患者１とデータベースＺの患者３１との間の一致がＸＺ変換テーブルから除去され、同様にデータベースＹの患者２２とデータベースＺの患者１４との間のマッチングがＹＺ変換テーブルから除去される。
【００３７】
他の実施形態では、このような整合性解析は、反復ループ４０，４２，４４，４６，４８，５０の間に実行することができる。このアプローチは、既に作成された組毎の変換テーブルを活用することによって、後のループ反復を実行する処理時間を短縮する。例えば、Ｘ、Ｙ、Ｚのインデックスを付けられたデータベースを有するとともに、Ｘ－Ｙ、Ｘ－Ｚ、Ｙ－Ｚ変換テーブルを作成するために反復ループ４０，４２，４４，４６，４８，５０がこの順序で実行されるケースを考える。Ｘ－Ｙ及びＸ－Ｚ変換テーブルの作成後、データベースＸの患者１０はデータベースＹの患者１１にリンクされ、データベースＸの患者１０はデータベースＺの患者１５にもリンクされることが分かる。次に、Ｙ－Ｚ変換テーブルを作成する最後の反復の間に、Ｙ－Ｚ変換テーブルと既に作成されたＸ－Ｙ及びＸ－Ｚ変換テーブルとの一貫性（整合性）を保証するために、データベースＹの患者１１をデータベースＺの患者１５にリンクする必要があることが既に知られている。
【００３８】
これに加えて、又は、これに代えて、本明細書に開示された幾つかの実施形態では、患者のマッチングを改善するために、長さ方向の情報が活用される。一般に、長さ方向の特徴は、事象のタイムスタンプ間の時間間隔Δｔによって分離された、匿名化されたヘルスケアデータベース内の単一の匿名化された患者に対するタイムスタンプ付き事象の組によ

って定義される。このような長さ方向の特徴は、匿名化された医療データベースにおいても明確に定義されており、厳密な時間シフトは事象間の時間間隔Δｔに影響を与えないので、匿名化プロセスは各患者の全てのタイムスタンプのランダムではあるが厳密なシフトを導入する。

【００３９】

図５を参照すると、長さ方向の特徴の一例が記載されている。長さ方向の特徴は、時間間隔Δｔで区切られたｅ型の事象と、ｆ型の事象とによって定義される。図５の例では、データベースＸの患者ｍは、時間間隔Δｔで分離された事象タイプｆの事象の発生に続く事象タイプｅの事象の発生を有する。同様に、データベースＹの患者ｎは、同じ時間間隔Δｔで分離された事象タイプｆの事象の発生に続く事象タイプｅの事象の発生を有する。対照的に、データベースＺの患者ｐは、事象タイプｅの事象とそれに続く事象タイプｆの事象を有するが、タイプｅ及びｆの事象間の時間間隔は時間間隔Δｔよりもはるかに長い。従って、時間間隔Δｔで区切られた事象シーケンスｅ→ｆの時間的特徴に基づいて、データベースＸの患者ｍは、データベースＹの患者ｎと一致するが、データベースＺの患者ｐとは一致しない。このような長さ方向の特徴を一致させる際に、例えば、タイムスタンプの入力の可能性のあるエラーを考慮して、異なるデータベースの患者のためにΔｔにおける或る程度の変動を許容することが考えられる。

【００４０】

より複雑な長さ方向の特徴、例えば、第１の時間間隔Δt_1で分離される事象ｇ→ｅを有する事象タイプｇ→ｅ→ｆと、第２の時間間隔Δt_2で分離される事象ｅ→ｆとが想定される。他の考えられる長さ方向の特徴では、Δｔにおける許容可能な変動は、タイプｅ→ｆの事象がそれらの間の時間間隔に関係なく（Δｔにおける許容可能な変動で規定されるある制限内で）発生した場合に、実質的に長さ方向の特徴がマッチするように十分大きくてもよい。

【００４１】

例示的な長さ方向の特徴は、２つのデータベース（ｉ，ｊ）の患者の事象のタイムスタンプを比較するのではなく、事象間の時間間隔Δｔを使用する。前述したように、事象の絶対的なタイムスタンプに頼るのではなく、事象間の時間間隔に依存するこの手法は、患者のタイムラインが匿名化プロセスの一部としてランダムにシフトされた可能性に対して堅牢である。

【００４２】

幾つかの実施形態では、長さ方向の特徴は、動作４４で識別され、動作４６（図２参照）で使用される特徴のセットの他の特徴と同様に処理される。しかしながら、この手法は、長さ方向の特徴の識別及び処理が計算上複雑であり得るので、不要な計算上の複雑さを招く可能性がある。例えば、平均患者がＥ＝３０事象を有する場合、ｅ→ｆ形式の長さ方向の特徴を識別するために必要なペアワイズ事象比較の数は、Ｅ（Ｅ－１）／２＝４３５個の事象組である。一方、長さ方向の特徴のかなり高い特異性は、それらが合致する患者に対して非常に差別的であり得ることを意味する。従って、幾つかの実施形態では、患者のマッチング動作４６は、長さ方向の特徴に依存することなく最初に実行され、長さ方向の特徴は、困難なマッチング（例えば、非長さ方向の特徴のみが使用される場合の、データベースＹの複数の患者と一致するデータベースＸの患者）に対してのみ計算及び活用され得る。

【００４３】

幾つかの実施形態では、各患者の普遍的な患者ＩＤ（又はＵＩＤ）を使用して、非長さ方向の特徴マッチングが実行される（又は、部分的に実行される）。ＵＩＤは、患者の性別、人種、年齢、及び、体重などの共通の特徴の集合の連結として構築される。例えば、患者のＵＩＤ1518170は、以下の特徴を使用して生成される。男性又は性別１（1518170の最初の桁）、ネイティブアメリカン又は人種５（1518170の２桁目）、１８才（1518170の３桁目及び４桁目）、体重１７０ポンド（1518170の５桁目～７桁目）。従って、患者のために新しい記録（医学報告又は請求記録）が生成されるたびに、ＵＩＤが患者記録に割

り当てられる。ＵＩＤは特徴ベースなので、異なる匿名化されたデータベースで同じにする必要がある。オプションで、幾つかの許容値が受け入れられる。例えば、年齢に対して±1年の許容しきい値を使用した場合、データベースIIの80歳は、データベースIの79歳～81歳と同じとみなされる。特徴マッチングのためのそのようなＵＩＤ手法は、患者とのマッチングに使用される特徴のセットの全ての特徴に対して使用されてもよく、あるいは、より小さな特徴のサブセットが連結されてＵＩＤを形成してもよい。ここで、ＵＩＤを形成している特徴のセットは、全てのＮ個のデータベース１０に対して共通である。この後者の手法は、ＵＩＤが一度計算され、図２の（ｉ，ｊ）ループの各反復に対して再使用されることを有利に可能にし、計算効率を高めることができる。この手法では、次の３段階マッチングプロセスが想定される。（１）ＵＩＤに基づくマッチング、（２）困難な場合には、ＵＩＤに含まれない追加の非長さ方向の特徴に基づくマッチング、最後に、（３）より困難な場合には、長さ方向の特徴を使用したマッチング。

【００４４】

所与の実施形態では、開示された態様の様々な組み合わせを採用することができることが理解されよう。例えば、２つのデータベースの統合（Ｎ＝２）と複数のデータベースの統合（Ｎ≧３）との両方に、長さ方向の特徴マッチングが使用され得る。自然言語処理（ＮＬＰ）を使用して、Ｎ＝２且つＮ≧３の統合タスクのために、非構造化又は半構造化データベースから一連の特徴を生成できる。

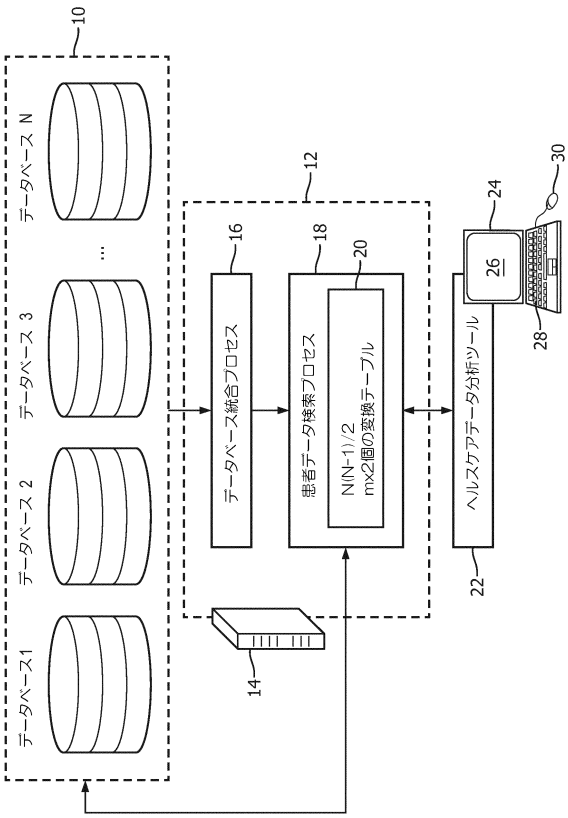
【００４５】

図１の開示されたヘルスケアデータ分析装置を見るための別の手法では、Ｎ個の匿名化されたヘルスケアデータベース１０を統合するプロセスは、Ｎ個の匿名化されたヘルスケアデータベース１０から匿名化された母集団イメージを再構築する匿名化母集団イメージ再構成方法と見ることができる。この代替図では、再構成された匿名化された母集団イメージは、Ｎ（Ｎ－１）／２個の変換テーブル２０によって統合されたＮ個の匿名化されたヘルスケアデータベース１０の内容を含む。この代替的な視点では、匿名化された母集団イメージ再構成法は、Ｎ個の匿名化されたヘルスケアデータベース１０の形式の母集団イメージデータを、Ｎ（Ｎ－１）／２個の変換テーブル２０によって統合されたＮ個の匿名化されたヘルスケアデータベース１０の内容を有する匿名化された母集団イメージに再構築（又は変換）する。

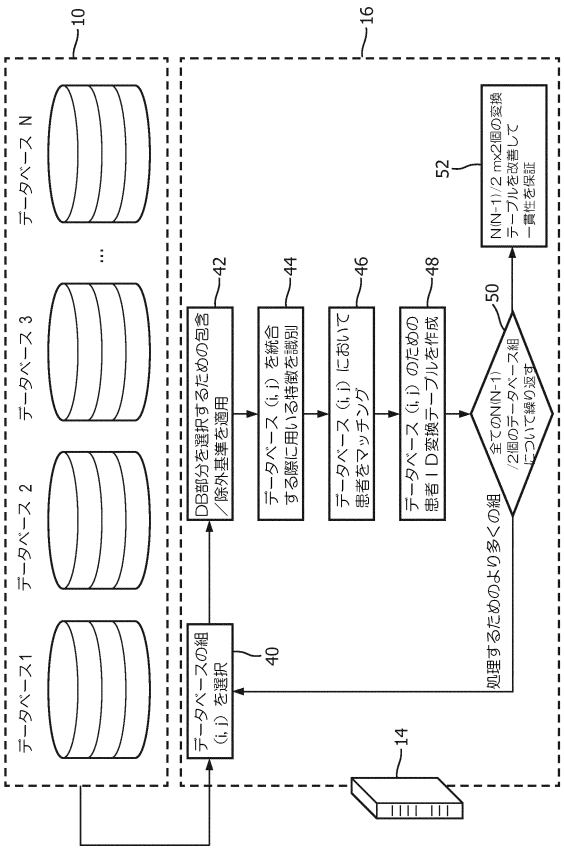
【００４６】

本発明を好ましい実施形態を参照して説明した。前述の詳細な説明を読んで理解すると、他への修正や変更がなされ得る。本発明は、添付の特許請求の範囲又はその均等の範囲内に入る限りにおいて、そのような改変及び変更の全てを含むと解釈されることが意図される。

【図１】



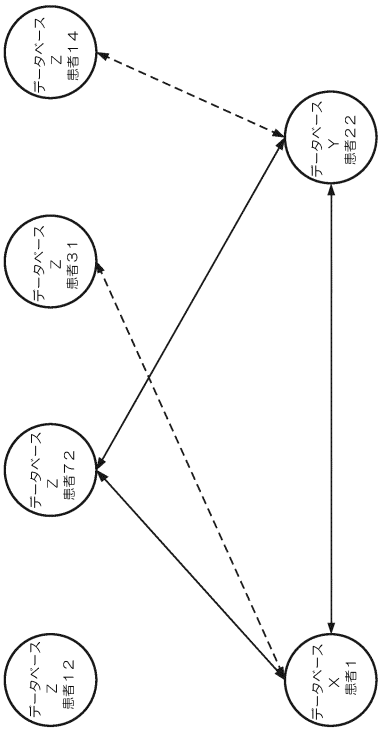
【図２】



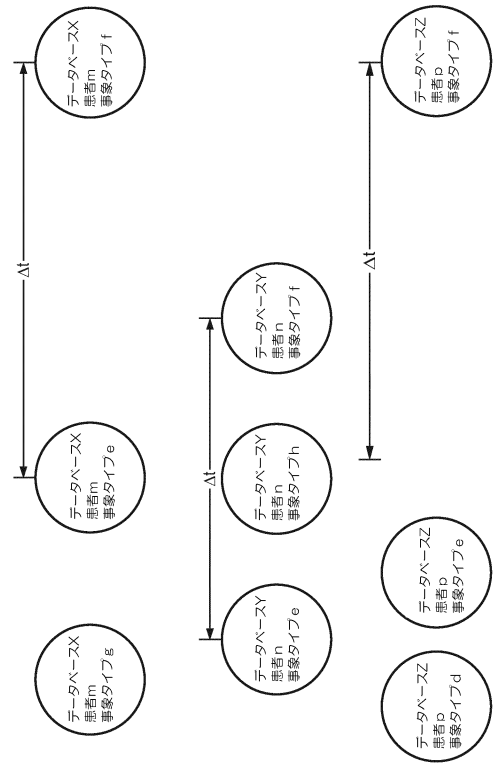
【図３】

データベース	性別	人種	死亡率	在院日数	年齢	一次診断	体重
X	80%	99%	97%	95%	97%	97%	99%
Y	97%	97%	98%	94%	99%	71%	98%
Z	99%	96%	70%	98%	95%	97%	97%
X-Yの特徴？	いいえ	はい	はい	はい	はい	いいえ	はい
X-Zの特徴？	いいえ	はい	いいえ	はい	はい	はい	はい
Y-Zの特徴？	はい	はい	いいえ	はい	はい	いいえ	はい

【図４】



【図５】


データベースX
患者m
事象タイプf
Δt
データベースX
患者m
事象タイプe
データベースX
患者m
事象タイプg
データベースY
患者n
事象タイプf
Δt
データベースY
患者n
事象タイプh
データベースY
患者n
事象タイプe
データベースZ
患者o
事象タイプf
Δt
データベースZ
患者o
事象タイプe
データベースZ
患者o
事象タイプd

フロントページの続き

(72)発明者　シャリフィ　セデュー　レザ
オランダ国　５６５６　アーエー　アインドーフェン　ハイ　テック　キャンパス　５
(72)発明者　エルフォルト　ダニエル　ロベルト
オランダ国　５６５６　アーエー　アインドーフェン　ハイ　テック　キャンパス　５
(72)発明者　トゥルエン　ロエル
オランダ国　５６５６　アーエー　アインドーフェン　ハイ　テック　キャンパス　５

審査官　後藤　彰

(56)参考文献　特開２０１６－０３８７８０（ＪＰ，Ａ）
欧州特許出願公開第０２８７９０６９（ＥＰ，Ａ２）
米国特許出願公開第２００２／００７３１３８（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００－１６／９５８
Ｇ１６Ｈ　１０／００－８０／００