大量かつ複雑な非構造化データを扱う解析基盤の仕組み

Architecture for a Data Analysis Base treating Massive and Complicated Unstructured Data

星　野　隆　之

要　約　2011 年より，ゲノムコホート研究のためのデータ解析基盤の設計およびプロトタイプシステム構築を行い，データ解析基盤を整理した．

ゲノムコホート研究は，集団を対象に追跡調査を行って，遺伝子と病気の発症などとの関係を探ることを目的とした研究で，診療情報，健康診断情報，ゲノム情報など，様々な情報を，長期間，収集・蓄積し，解析する．今回のデータ解析基盤は，このような研究活動を支援するもので，大量・非構造化データの扱い，個人情報の扱い，個人の識別などの課題を解決した．

ゲノムコホート研究におけるデータの扱いは，大量かつ複雑な非構造化データを扱う事例として参考になるものであり，この活動で得られたアーキテクチャは，疫学研究向けに留まらず，データを解析しその結果を事業に活用するための仕組みづくりに有効である．

Abstract　Since earlier 2011, we have designed the data analysis base and built its prototype system for the genomic cohort study, and then have organized the data analysis underlying architecture.

The genomic cohort study is a form of research activity, which intends to carry out the follow-up survey of specific groups in order to explore the relationships between genes and the onset of disease and such, to analyze a variety of information such as medical information, health examination information, including genomic information, which had been gathered and accumulated over the long term. The data analysis base raised here intends to support these research activities, by which the challenges such as the handling of bulk and non-structured data, the handling of personal information, and the personal identification are resolved.

The handling of data in the genome cohort study serves as a useful reference for a case study handling the massive and complicated unstructured data. The “architecture” gained through this activities would be helpful for not only epidemiological studies, but also a mechanism to analyze data and make use of the analyzed results in a related business.

1. は　じ　め　に

2010 年度より，疫学研究を題材として，大量かつ複雑な非構造化データを利活用するための情報基盤構築に取り組んでいる．2011 年度からは，京都大学医学研究科附属ゲノム医学センターとゲノムコホート研究のためのデータ解析基盤の設計およびプロトタイプシステム構築を開始した．

ゲノムコホート研究では，広範な源泉から多種のデータを取り込み，解析をする．そのデータは，大量であり構造化が難しい，いわゆるビッグデータであり，それを解析することは，「ビッグデータをどのように解析するか」という一般的な課題に対しても，解となり得る．また，

2011 年度に設計およびプロトタイプシステム構築を行ったデータ解析基盤は，一般のデータ解析基盤として適用可能なものであると考えている．

本稿では，先ずゲノムコホート研究の特性について整理し，データ解析基盤を構築する上での課題とその対応について検討する．そして，検討結果を反映したデータ解析基盤の仕組みを紹介する．

2. ゲノムコホート研究の特性

本章では，ゲノムコホート研究について説明する．ゲノムコホートの目的と，疫学研究で用いられるアプローチを整理し，ゲノムコホート研究におけるデータ解析基盤に求められる要件を抽出する．

2.1 ゲノムコホート研究とデータ解析基盤

ゲノムコホート研究は，地域住民などの集団を対象に追跡調査を行って遺伝子と病気の発症などとの関係を探ることを目的とした研究で，個別医療や予防医療を実現するために不可欠なものである．また，以下の点で，疫学という医学研究の一分野のみで担う研究活動ではなく，多分野の協力と連携で実現する学際的研究となっている[1]．

- 遺伝因子や分子レベルのバイオマーカーを加味した研究．
- 疾病リスク，薬剤応答性などは人種によって大きく異なるため，日本人を対象とした疫学研究が不可欠．
- ゲノム・バイオマーカー・疫学・臨床情報を融合し，長期的な集団の観察と膨大な情報の精緻かつ効率的な分析・解析による予防医学研究が必要．
- 分析・解析技術に加え，大規模データ利用の情報基盤やバイオインフォマティクスを用いた解析基盤の整備が必要．

ゲノムコホート研究では，1）参加者の募集と同意の取得，2）質問票の情報取得とデータ入力，3）生体試料の採取と保管，4）ゲノム解析を含む生体試料の測定，5）追跡調査とデータ入力，6）データの加工と統合化による解析用データセットの構築，7）研究目的に即したデータの検討と統合解析，8）研究成果の評価と社会への還元，という実施プロセス[2]に基づき，疾患と生活習慣，体質の関係を明らかにし，個人に適切な予防法，治療法の実施や創薬に結び付ける．

2011 年度に開始したデータ解析基盤の設計およびプロトタイプシステム構築は，このプロセスを情報基盤としてサポートするものであり，「ゲノムコホート研究の情報基盤の構築と公開」と「データベースの枠組みの提供と情報の連結」の実現を目指している．前者は，診療情報，健康診断情報（健診情報），生活習慣情報，環境情報，ゲノム・オミックス情報[*1]をデータベース化することで，集積した情報を，個人情報保護のもと，医学・生命科学研究者に提供するものである．また後者は，同様の研究を行う研究者に，即時活用可能な形でデータベースの枠組みを提供し，他の研究で蓄積されたデータを連結，共有することで，個別の研究で得られた情報の再利用ができる基盤を提供するものである．

2.2 ゲノムコホート研究に必要な情報

ゲノムコホート研究として集積が必要な情報の種類，情報を採取する際に考慮すべき点を整

理する．

診療情報については，医療機関のオーダリングシステムや電子カルテに，広範にわたり多くの情報が電子的に蓄積されているが，情報の規格化が不十分であり，医療機関毎に検査内容や病名などのコードが統一されていない，日本語文字列で記録されているケースが多い，という状況がある．このことは，研究利用を考えた場合に大きな課題となり，情報を入力する際にコードを統一することが望まれるが，現状では，コードの統一や日本語文字列からの情報抽出などはデータを取り込む側で対応している．

健診は，各種の検査で健康状態を評価するものであり，健診結果は，発症し医療機関での診察を受ける前の状態を知る情報として重要である．健診結果の電子データは医療保険者[*2]に集積されているが，特定健診情報がHL7-CDA形式と呼ばれるデータ形式となっている以外，一般の健診については規格化ができていない状態である．広範囲の疾病を対象とする可能性のあるゲノムコホート研究を前提とした場合，特定健診情報だけでなく一般の健診情報も取り込む必要がある．全ての健診情報の規格化が期待されるが，現状では，データを取り込む際，個別に対応している．

生活習慣情報，環境情報については，一般的には質問票を使ったアンケート調査という形で収集する．この調査では，生活習慣，本人が置かれている環境状況のほか，本人・家族の病歴・自覚症状など，診療情報や健診情報では得られない情報を得ることが可能である．しかしながら，その回答基準には個人差があり，回答の抜けや矛盾が発生するので，精度を保持するための対応が必要となる．具体的には，回答提出後の聞き取り，Webなどの電子的手段を用いて回答の抜けや矛盾回答をチェックする，などのフォローアップが有効である．

生体試料情報，ゲノム・オミックス情報については，研究内容に応じて，適宜，血液などの生体試料を採取し，検査機関によりデータ化されたものを取り込む．この時，検査会社や医療・測定機器メーカー間で検査方式や情報表現形式が違う可能性がある．この点についても規格化が期待されるが，現状では，検査方式や情報表現形式を確認し，個別に補正している．また，生体試料の採取に関しては，研究者から研究参加者に対し目的を明確にすることと，参加者本人の合意が必要となる．広く生体試料を採取しデータ化するには，そのための制度的な枠組みが必要となる．

2.3　情報の集積・統合

1章で述べたように，ゲノムコホート研究では，複数源泉からの多様な情報を集積し，長期間にわたる追跡調査を可能にするために，大量かつ非構造なデータを扱い，統合する．

ここでの「大量」とは，多数の項目を扱うこと，コホート研究では長期間にわたる情報が必要となること，ゲノム・オミックス情報など1単位あたりのデータ量が膨大なものを扱うことを指している．「非構造」とは，診療情報や健診情報において，個人の診療や健診の内容により記録する情報が異なるために，共通的なデータ構造の定義ができないケースや，医療画像など文字以外の情報も扱うケースがあるということである．

また，「統合」とは，複数の源泉から集められた情報に関し同一人物を特定することである（一般に「名寄せ」と呼ばれる）．個人が複数の医療機関に跨って診療を受けた場合，源泉が複数の医療機関，医療保険者，検査機関などにわたり，それぞれの間で個人を特定する情報が統一されていないため，同一人物と認識されない問題への対処である．

統合については，EHR（Electric Health Record：電子健康記録）や PHR（Personal Health Record：個人健康記録）を源泉として利用するという考えもある．EHR や PHR は，情報を統合管理して医療機関間で共有し，個人に情報を提供する仕組みであり，実際にいくつかの地域で運用されている[3]．しかし，取り込まれている情報は，診療情報，健診情報が中心であり，不足する部分や情報の標準化に関しては，個別の取り込みと名寄せ処理など，取り込み側での対応が必要となる．

2. 4　個人情報の保護

ゲノムコホート研究に必要な情報は，個人情報の中でも機微（センシティブ）情報に該当するものであり，これらの情報を管理する際には，文部科学省の「疫学研究に関する倫理指針」[4]，厚生労働省の「医療情報システムの安全管理に関するガイドライン」[5]，文部科学省，厚生労働省，経済産業省による「ヒトゲノム・遺伝子研究に関する倫理指針」（以下「ゲノム指針」）[6]などの規定を遵守する必要がある．

「ゲノム指針」によると，個人情報とは，氏名，生年月日，その他の記述により，特定の個人を識別することができるものを指す．ゲノムコホート研究では，その個人情報から個人を識別する情報の全部または一部を取り除き，代わりに当該提供者とかかわりのない符号または番号を付して匿名化する．匿名化は研究部門以外で行い，かつ，匿名化情報と個人との対応表を厳密に管理する．

また，研究成果を自治体や医療機関にフィードバックし，健康指導や医療機関での個別医療や予防医療に活用するには，最終的に，個人を特定する必要があるため，匿名化の逆変換処理で，それを可能としている．

2. 5　情報の利用について

ゲノムコホート研究では，一般に研究計画（以下，プロトコール）を定義する．プロトコールには，データ解析に用いるデータセットと分析方法（モデルと手法）が記述されている．蓄積されている大量・非構造化データを研究に利用するためには，プロトコールに応じてデータセットを抽出する．即時利用可能な形でデータセットを提供するためには，新たな研究が開始される度に抽出プログラムを構築するのではなく，プロトコールの内容に従いデータセットを抽出するような方式をとるのが良い．研究者は，プロトコールの内容をメタデータとして登録しておき，メタデータの内容に応じてデータセットを抽出する．これにより，個別抽出プログラム構築の期間を省き，研究開始までの時間を短縮することができる．

また，同様な研究や，応用・発展させた研究に対しても，データセットを抽出するための情報がメタデータとして管理されていることにより，再利用が可能になる．

3.　ゲノムコホート研究のためのデータ解析基盤

ゲノムコホート研究において必要となるデータ解析基盤の仕組みを紹介する．ゲノムコホート研究で扱うデータに関わる課題は，1）大量かつ複雑な非構造化データの扱い，2）データの統合，3）個人情報の保護，4）データの精緻化・コードの統一化，である．これに対し，研究者側は，データに，1）プロトコールに応じたアウトカムや予測因子[*3]となるデータセットを取得できる，2）長期間にわたり，個人を追跡できる，3）個人情報を扱う上での制約に縛られ

ない研究ができる，4）研究に耐えられる精緻化，コードの統一化がされている，ことを要求する．

ゲノムコホート研究は，解析対象となるデータの発生源（以下「データソース」と記す）と，データ解析箇所（以下「ラボ」と記す），そしてデータソースとラボを連携するための基盤（以下「データバンク」と記す）という構成の上で実施される（図1）．データソースとラボの間に存在するギャップを解決/吸収するのがデータバンクの役割となる．

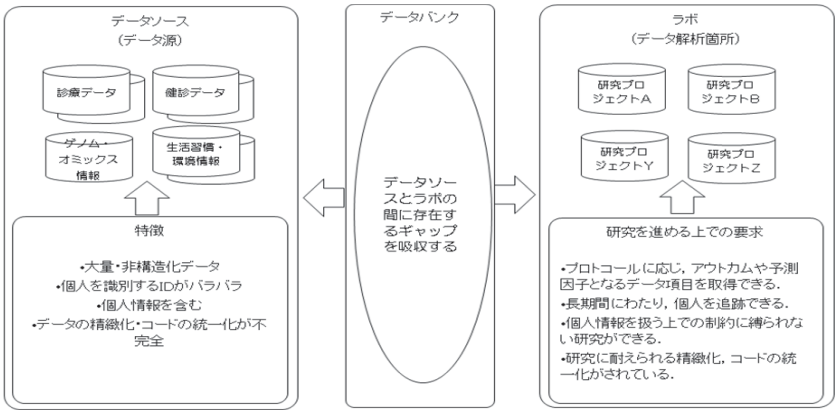

図1　ゲノムコホート研究基盤の基本構造

3.1　大量・非構造化データの扱い，データの精緻化・コードの統一化

データソースについては，複数の医療機関に跨った診療情報/健診情報を，数十年単位で蓄積し続けるため，数百テラバイト規模にのぼる大量データを想定した対応が必要である．また，データの増加に伴い，データ処理効率（データのロードおよび検索時間など）が劣化しないような配慮も必要である．この点については，クラウドコンピューティング基盤などで使われている大規模分散データ処理技術を適用し，まずはプロトタイプシステムとして小規模な構成でスタートし，必要に応じてスケールアウトする，という方針で進めるのが適切である．

非構造化データについては，データの収集/蓄積を目的とする機能と，プロトコールに基づいたデータ解析に対しデータを提供する機能に分けて検討する（図2）．

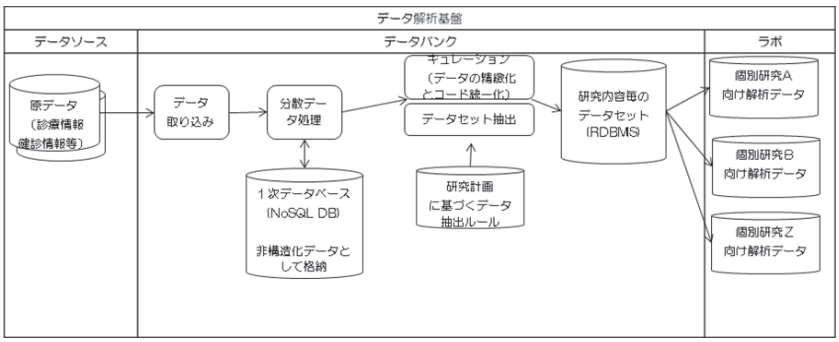

図2　大量・非構造化データの格納とデータ精緻化・コード統一化

データの収集/蓄積では，データ源からのデータをそのままの形で取り込むことにし，各々の研究が必要とする（可能性がある）データを格納しておく．格納には，RDBMS ではなく NoSQL データベースを用い，非構造化データに耐えられるようにする．

大量・非構造化データの管理については，2010 年度に徳島大学医学部・歯学部付属病院の病院情報センターと共同で行った EHR の構築[7]，同じく 2010 年度に佐賀大学と共同で行った疫学データベースの構築[8]において，大規模分散データ処理技術を適用し，技術を実証しており，今回のデータ解析基盤でも同様の技術を採用する．

データ解析のためにデータセットを抽出する際は，既にプロトコールが定義され，扱うデータ項目も明確になっていると判断できる．この段階では，分析モデルが存在し．データ解析の対象はデータセットとして定義できるはずである．このため，提供用のデータ格納には RDBMS を用いて分析モデルと対応付け，分析ツールなどからのアクセスを容易にすることを考える．プロトコールに従ったデータセット抽出は，そのルールをメタデータとして設定し，その設定に従って抽出するようにする．また，研究者にデータを提供するにあたって，キュレーション処理（コードの統一などの標準化，欠損値や矛盾値などの確認と補正，データの精緻化）も行う．

3. 2　個人情報の扱いと名寄せ

個人情報の扱いについては，「ゲノム指針」を遵守した方式を考える．具体的には，医療機関などの源泉から情報を収集する際に個人情報を分離し，蓄積する時には個人情報を一切含めない方式をとる．個人を識別する識別子を匿名化し，情報の流出が発生した場合でも，個人を特定できないようにする．

「ゲノム指針」にあるように，匿名化を実施する組織は他の組織とは独立させ，その組織には実情報（診療情報，健診情報など）を保持させない．このことにより，中間組織からの情報流出を防ぐ．また，データソースからラボに至るまでに，二段階の匿名化を行うことで，データ解析結果から個人を特定されるリスクを低減する．さらに，個別研究毎に異なる匿名化を行い，複数の研究結果を照合することで個人が特定されてしまうリスクを低減する．

ゲノムコホート研究では，研究成果を自治体での健康指導や医療機関での個別医療・予防医療にフィードバックするが，このためには，匿名化の逆処理が必要である．逆処理では，匿名化でたどった経路を逆順に進み，匿名化と矛盾しない安全性を確保する．

実際のゲノムコホート研究での匿名化運用事例としては，「ながはま 0 次コホート事業」[9]があり，約 1 万人のゲノム情報を採取し，コホート研究，自治体の健康増進事業へのフィードバックを行っている．

一方，個人の特定は名寄せ処理により実現するが，その方式としては，

1）　既知の属性情報を組み合わせる方法
2）　複数の識別子を統合する辞書を用意する方法
3）　共通番号を使用する方法

が考えられる．1)の方法は，個人情報を構成する属性情報（被保険者証の記号・番号，生年月日，カタカナ氏名など）を組み合わせて生成する方法である．この方法は，既知の情報を用いるため，名寄せにあたって新たな情報収集や手続きを追加する必要がなく，導入はしやすい．しかしながら，転職などで医療保険を切り替えた場合や，結婚などで氏名が変わった場合など

に個人を追跡しきれない可能性があり，広域で長期間の追跡調査を行うような国家レベルでのコホート研究には適用が難しい．

2)の方法は，病院などの施設で個人を識別している番号情報を，研究の参加者がそれぞれ登録し，辞書化して，施設別の識別番号を統合用番号に変換することで，個人を特定する方法である．この方法は，1)に比べ精度を高めることが可能であるが，施設別の識別番号の登録は，研究参加者の意思によるもので，登録率の高さが精度に影響する．

3)の方法は，個人を識別する共通番号を使用する方法である．研究参加者に対し，参加時に共通番号を発行する．研究参加者は，診療・健診を受ける際，生体試料の採取に応じる際などに，この共通番号を提示する．各施設では，データに共通番号を付加して，データバンク側に引き渡す．この方法は，他の方法よりも精度を保つことが可能である．しかし，広域にわたってこの方法を運用する場合，共通番号を発行する機関の位置づけや，個人情報の扱いなど，制度・組織面からの検討が必要となる．地域において，この方式を運用している事例としては「まいこネット」[3]がある．

今回のデータ解析基盤プロトタイプでは，3)の方法と 2)の方法を組み合わせている．具体的には，EHR から入手するデータについては 3)の方法，EHR 以外から入手するデータについては 2)の方法を用いている．

3.3　ゲノムコホート研究向けデータ解析基盤

本章で述べた内容を整理すると，データ解析基盤は，1）データソース，2）個人情報管理，3）一次匿名化，4）データバンク，5）二次匿名化，6）ラボ，という要素に分類することができる．この要素間におけるデータ授受の流れを図 3 に示す．

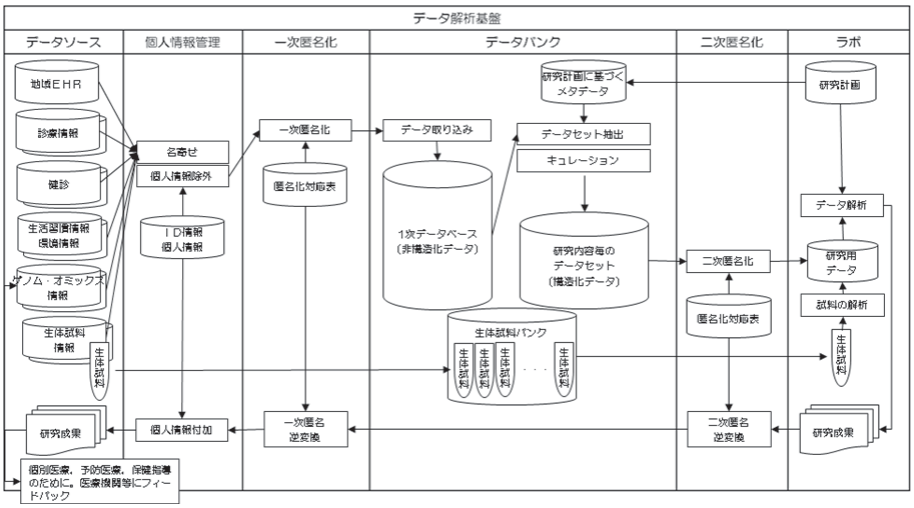

図 3　疫学研究向けデータ解析基盤のイメージ

「データソース」には複数の施設（医療機関，保険者，検査機関など）があり，ゲノムコホート研究に必要な情報（診療情報，健診情報，ゲノム・オミックス情報，生活習慣・環境情報）の源泉として機能する．

「個人情報管理」は，複数のデータソースから送付された情報に対し，名寄せを行う．また，情報漏えい時のリスクを軽減しラボでのデータの扱いを容易にするため，個人情報を排除する．

「一次匿名化」では，不測の事態でデータの流出が発生した場合でも，個人の特定を不可能にするため，個人を識別する識別子を別名に変換する．この機能は，データソース，個人情報管理，データバンク，二次匿名化，ラボとは独立した機関で運営され，匿名化対応表の秘匿性を守る．

「データバンク」では，一次匿名化されたデータを取り込み，そのままの形（非構造化データ）で一次データベースに格納する．さらに，ラボでの利用に向けて，研究計画に基づいたコードの統一やデータの補正・確認（キュレーション）を行い，各研究向けのデータセットを作成する．

「二次匿名化」は，ラボにおける個人の特定を更に不可能にするため，一次匿名化された識別子を別名に変換する．この機能は，データソース，個人情報管理，一次匿名化，データバンク，ラボとは独立した機関で運営され，匿名化対応表の秘匿性を守る．

「ラボ」では，データや生体試料を取り込み，研究計画に基づいた研究活動（データ解析など）を実施する．ラボでのデータ解析の成果は，二次匿名化機関，一次匿名化機関の順に識別子を逆変換し，保健指導や個別医療，予防医療のために医療保険者や医療機関にフィードバックしたり，他の研究のデータソースとして活用する．

生体試料（血液など）については，試料自体は物理的にデータバンクを経由してラボに引き渡すが，生体試料情報（試料の属性）は，他の情報と同様に，個人情報管理，一次匿名化，データバンク，二次匿名化を経てラボに送る．

4. 今後の展開

医療情報については，厚生労働省での規格化[10]や，HL7 などの規約[11]の整備により，効率的な収集が可能になることが期待できる．

EHR，PHR については，地域での試行に加え，国家レベルでの検討，整備に，今後期待したい．これにより，国家レベルでのコホート研究の情報源として利用が可能になり，情報の取得や名寄せの効率が向上する．

生活習慣や環境情報のアンケート形式での情報収集については，欠損回答や矛盾回答をフォローするために，Web を用いたアンケートの実施が検討されており，アンケート手法の新技術を適用するなどの動きと共に，新たな試みが期待できる．また，生活習慣に関してはアンケート形式に限らず，センサーなどを用いてディジタルに情報を取得することも試みられており[12]，新たなデータソースとなり得る．環境情報についても，今後，放射線の影響などを因子として取り込む可能性があり，地理情報をデータ源とする研究の展開が予想される．

これらの実現は，国家レベルのゲノムコホート研究向けデータ解析基盤の実現に通じ，ビッグデータの解析基盤として，実証された技術を提供することができると認識している．

5. お わ り に

これまで，ゲノムコホート研究向けのデータ解析基盤について述べてきたが，広範な情報源より大量かつ複雑な非構造化データを収集・格納し，個人情報の保護に配慮しつつ，計画的なデータ解析を行うことは，一般のデータ解析にもあてはまると考えている．今回紹介したデー

タ解析基盤の仕組みが，ビックデータを扱うデータ解析基盤の参考になれば幸いである．

最後に，今回紹介したデータ解析基盤の研究開発プロジェクトに助言・支援をくださった方々にお礼を申し上げる．

＊1　生体のもつ全ての遺伝子情報，および生命分子（タンパク質，RNA）の網羅的情報．
＊2　健康保険に関する事務・責務を業務として行う者（市町村など）．
＊3　二つの変数（因子）の関係において，一方が他方より時間などの尺度で先行していると判断できる場合，前者を予測因子，後者をアウトカムと呼ぶ．

参考文献

［1］ 松田文彦，「トーゴーの日シンポジウム 2011　大規模ゲノムコホート研究の統合情報基盤の構築」，2011 年 10 月，http://events.biosciencedbc.jp/images/togo2011/07.pdf p4, 5
［2］ 田島和雄，「ゲノムコホート研究の基礎知識」，メディカル・サイエンス・ダイジェスト 10 月臨時増刊号，ニューサイエンス社，第 37 巻第 12 号通巻 487 号，2011 年 10 月
［3］ まいこネット–京都地域連携医療推進協議会，http://www.e-maiko.net/
［4］ 疫学研究に関する倫理指針，文部科学省・厚生労働省，2002 年 6 月，http://www.lifescience.mext.go.jp/files/pdf/37_139.pdf
［5］ 医療情報システムの安全管理に関するガイドライン，厚生労働省，2010 年 2 月，http://www.mhlw.go.jp/shingi/2010/02/s0202-4.html
［6］ ヒトゲノム・遺伝子解析研究に関する倫理指針，文部科学省，厚生労働省，経済産業省，2001 年 3 月，http://www.mhlw.go.jp/topics/bukyoku/seisaku/kojin/dl/161228genomu.pdf
［7］ 森川富昭，玉木悠，田木真和，青木雅美，井内伸一，中山陽太郎，「医療情報の二次利用に向けた医療クラウドデータベース設計」，医療情報学，日本医療情報学会，第 31 巻第 2 号，2012 年
［8］ 沖俊吾，「医療システムでの非構造化データ活用事例」，ユニシス技報，日本ユニシス，Vol.31 No.4 通巻 111 号，2012 年 3 月
［9］ ながはま 0 次予防コホート事業，長浜市役所，2008 年 10 月 http://www.city.nagahama.shiga.jp/index.cfm/9,3709,19,158,html
［10］ 厚生労働省，「厚生労働省において保健医療情報分野の標準規格として認めるべき規格について」，2010 年 2 月，http://www.mhlw.go.jp/shingi/2010/01/dl/s0125-12a.pdf
［11］ 日本 HL7 協会，http://www.hl7.jp/index.html
［12］ 新堀聡，「予防医療を実現する生体情報収集基盤」，ユニシス技報，日本ユニシス，Vol.30 No.4 通巻 107 号，2011 年 2 月
［13］ L. Gordis，木原正博，木原雅子，加地正行，「疫学」，メディカルサイエンス・インターナショナル，2010 年 5 月
［14］ S. B. Hulley，S. R. Commings，W. S. Browner，D. G. grady，T. B. Newan，木原正博，木原雅子，「医学的研究のデザイン第 3 版」，メディカルサイエンス・インターナショナル，2009 年 2 月

※上記参考文献の URL は 2012 年 2 月 13 日現在での存在を確認．

執筆者紹介　星　野　隆　之（Takayuki Hoshino）

1985 年日本ユニシス（株）入社．人材育成部門，データ利用技術部門，電力事業部門で情報系のシステム開発プロジェクトに従事．2006 年より R&D 部門において，データエンジニアリング関連の研究開発プロジェクトに従事．現在，総合技術研究所先端技術ラボに所属．PMAJ 公認 PM レジスタード．