大規模ゲノム疫学研究の統合情報基盤の構築事例

Case Study of Integrated Information Infrastructure of Large-scale Genome Epidemiological Research

沖　　俊　吾

要　約　本稿では，長浜市役所と京都大学医学研究科附属ゲノム医学センターが連携して実施する「市民の健康づくりの推進」と「医学の発展への貢献」を掲げた全国初の 1 万人規模の大規模ゲノムコホート事業である「ながはま 0 次予防コホート事業」に関するシステム開発プロジェクトにおける大規模ゲノム疫学研究の統合情報基盤の構築事例を紹介する．

大規模ゲノム疫学研究の統合情報基盤は，京都大学医学研究科附属ゲノム医学センターにおける先制医療の実現に向けた疾患発症追跡のための研究基盤として開発し，2012 年より運用を開始した．本基盤は，将来の全国展開を視野に入れ，人的ミスがなくセキュアにデータ・試料を収集できるような汎用性の高いシステムを構築することが求められた．これらの要件を満たすために，二段階匿名化/秘密分散法，メタデータ管理，自動生成機能を開発した．

Abstract　In this paper, the case of building integrated information infrastructure of large-scale genome epidemiology research in system development project with regard to “Nagahama zeroji prevention cohort project” is introduced. The said project is conducted by Nagahama City Hall and Kyoto University Graduate School of Medicine, University Medical Genomics Center, and is the first countrywide large scale genome cohort project on the scale of ten thousand people that propounding the project concepts such as “Contribution to the development of medicine” and “promotion of health of citizens”.

Integrated information infrastructure of large-scale genome epidemiology research was developed as a research infrastructure for the disease onset tracking toward the realization of pre-emptive care in Kyoto University Graduate School of Medicine, University Medical Genomics Center and started its operation in 2012. With a view to nationwide deployment in near future, this infrastructure was requested to be constructed as a versatile system that can collect data sample securely without human error. In order to meet these requirements, we have developed a two-step anonymous/secret sharing scheme, metadata management, and automatic generation function.

1. は　じ　め　に

先進諸国の中でも少子高齢化が急速に進行する，いまや世界一の長寿国「日本」において，医療費の増大が国家の財政を圧迫している．内閣府によると，2013 年には高齢化率が 25.1％で 4 人に 1 人となり，2035 年に 33.4％で 3 人に 1 人となる．2060 年には 39.9％に達して，国民の約 2.5 人に 1 人が 65 歳以上の高齢者となる社会が到来すると推計されている．また，2009 年度の後期高齢者医療費は，約 12 兆 0,108 億円であり，国民医療費に占める割合は 33.4％となっている．今後も人口の高齢化や医療の高度化などに伴い，医療費が増大していくことが予想される．

近年，医療費の増大から予防医学への関心が高まっている．予防医学を推進することにより，病気の予兆を検知し，重篤な病気になる前に適切な処置・投薬を施すことが医療費を軽減させる優れた戦略である．大部分の病気は「遺伝要因」と「環境要因」の相互作用で発症するものであるので，この両者の情報を正確に分析し，どのような遺伝背景のもとでどのような病気を発症し易いか，またその発症に関わった生活習慣等をつぶさに長期に亘って観察することが不可欠である．このような医学的に極めて重要な「ゲノムコホート研究」[*1]が，内閣府総合科学技術会議によるアクションプランとして2011年から開始された．

本稿では，長浜市役所と京都大学医学研究科附属ゲノム医学センターが連携して実施する「市民の健康づくりの推進」と「医学の発展への貢献」を掲げた「ながはま0次予防コホート事業」を2章で説明し，当事業における大規模ゲノム疫学研究の統合情報基盤の構築事例を3章で紹介する．4章では今後の展開について述べる．

2.　ながはま0次予防コホート事業

2.1　ながはま0次健診

ながはま0次健診とは，滋賀県長浜市が市民の健康づくりの推進と医学発展の貢献を目的に，京都大学医学研究科附属ゲノム医学センターと共同で行っている全国初の1万人規模の大規模ゲノムコホート事業である．第1期は2007年から2010年まで実施したが，2012年より第2期の追跡調査が開始されている[1]．

図1は，ながはま0次健診の全体概要図である．ながはま0次健診は，通常の「特定健診」のほか，健康づくりのためのゲノムコホート研究に必要な140項目以上を追加した健診である．30～74歳の市民を対象に行われ，被験者は700項目以上に及ぶアンケートに回答し，詳細な健診を受け，血液などを試料として提供する．これらの情報と健診結果，試料は，個人情報を

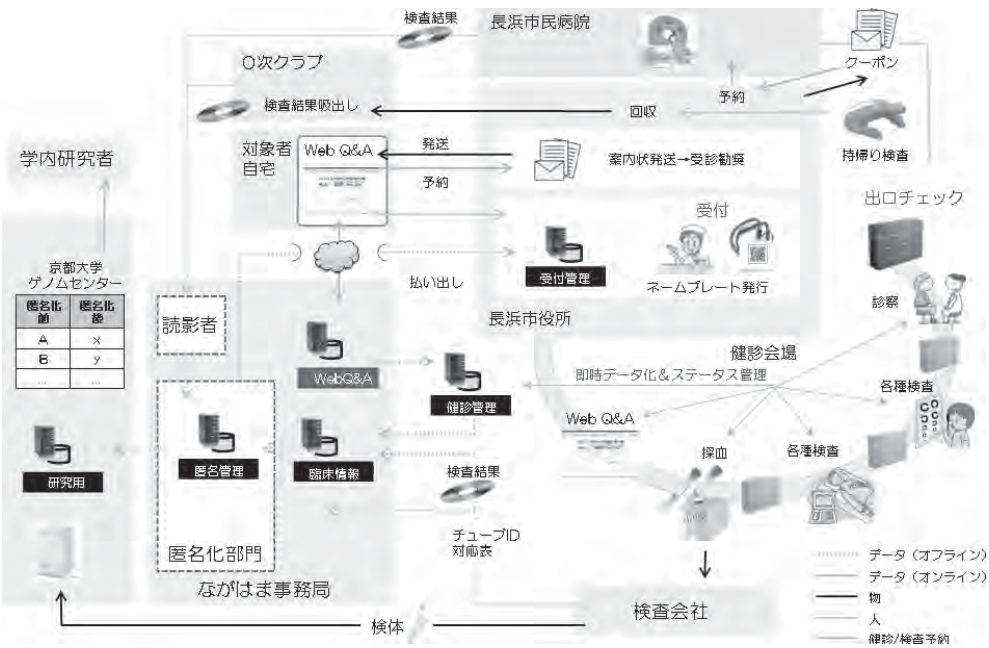

図1　ながはま0次健診　全体概要図

保護した上で京都大学の研究者によって分析され，生活習慣病の予防や早期発見，治療法の開発などに役立てられていく．

ながはま０次予防コホート事業では，産学官連携に加え，住民側からも自発的に「健康づくり０（ゼロ）次クラブ」が立ち上げられるなどの協力が得られた．大規模ゲノムコホート事業は，行政や大学側から頭ごなしに進めても意味がなく，住民に理解が得られないまま進めるとプロジェクト自体が潰れてしまいかねず，慎重に進めていかなければならない．そのためにも，時間をかけて住民に健康文化の重要性を啓発し，一体になってプロジェクトを進めていく枠組み作りが最も重要である．

ながはま０次健診には，受診する人達にもさまざまなメリットがある．例えば，大動脈波速度測定では血管年齢を知ることができ，特殊な血圧計でしか測れない中心血圧測定，肺気腫の早期発見に役立つ呼吸器機能測定，さらに，歯科医による歯周病検査など，一般の健康診断にはない高度な健診内容を無料で受けられる．

2.2　システム開発の目的と方針

ながはま０次健診の実施にあたっては，長浜市役所と京都大学医学研究科附属ゲノム医学センターが定めた「ながはまルール」[2]を遵守しつつ，データマネージメント（データのチェック，データベース構築，匿名化等）における人的負担を軽減する仕組み作りが求められた．そのためには，臨床情報の収集と匿名化，生体試料の処理・保管と匿名化，情報のデータベース化を行う必要があった．

本プロジェクトでは，将来の全国展開を視野に入れ，人的ミスがなくセキュアにデータ・試料を収集できるような汎用性の高いシステムを構築することが求められた．ながはま０次予防コホート事業や希少難治性疾患関連研究事業など，事業により拠点やセキュリティポリシーが異なるが，各々の事業毎に個別にシステムを構築するのではなく，各々の事業に対応可能な汎用的なパッケージを構築することを目的としている．

3.　大規模ゲノム疫学研究の統合情報基盤の構築事例

3.1　システム要件

本プロジェクトでの要件を表１に示す．要求事項として特徴的であり，実現する上で考慮が必要であったのは，個人情報保護に関する高セキュリティの対応である．

表１　大規模ゲノム疫学研究の統合情報基盤の構築に求められた要件

項目	内容
情報基盤の構築と公開	「ながはま０次健診」の１万人の生活習慣・環境情報，臨床情報を標準化し，データベースを構築する．集積した情報を，個人情報を保護したまま，医学・生命科学研究者に提供する．
データベースの枠組みの提供と情報の連結	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，情報の再利用ができる基盤を提供する．
データ品質の保持	入力精度に個人差が発生する情報について，システムを利用することで回答の抜け漏れ防止，誤入力防止を行う．

医療機関，健診機関は患者情報や診察・診療記録など特にプライバシーに関わる機微な情報を取扱っている．このため，経済産業省，文部科学省および厚生労働省の「ヒトゲノム・遺伝子解析研究に関する倫理指針」等のガイドラインに加え，長浜市役所と京都大学医学研究科附属ゲノム医学センターが定めた「ながはまルール」が規定されており，本プロジェクトの要求仕様の範囲に留まらず，法令を遵守する必要性があった．

3.2　システム全体概要

京都大学医学研究科附属ゲノム医学センターのデータセンタに，大規模ゲノム疫学研究の統合情報基盤をプライベートクラウドで構築した．システムを構築するにあたり，以下の点に留意した．

1）　各機能に対して，個人情報の保持方法と保有管理責任を明確にする．
2）　システムのパッケージ化により複数拠点からの情報を容易に統合可能とする．
3）　参加医療機関が増える度に個別にシステム連携するのではなく，複数医療機関のデータを統合した EHR（Electronic Health Record）を診療情報のデータ源泉とする．

これらの要求を満たすため「マスタサーバ」「情報収集管理サーバ」「匿名化サーバ」「ゲノム連携サーバ」の 4 種類のサーバで物理的に分離するシステム構成とした．また，「ながはまルール」により，ながはま 0 次健診はスタンドアローン環境で実施する必要があった．そのため，京都大学医学研究科附属ゲノム医学センターのデータセンタと健診会場の健診サーバはオフラインでの連携としている．

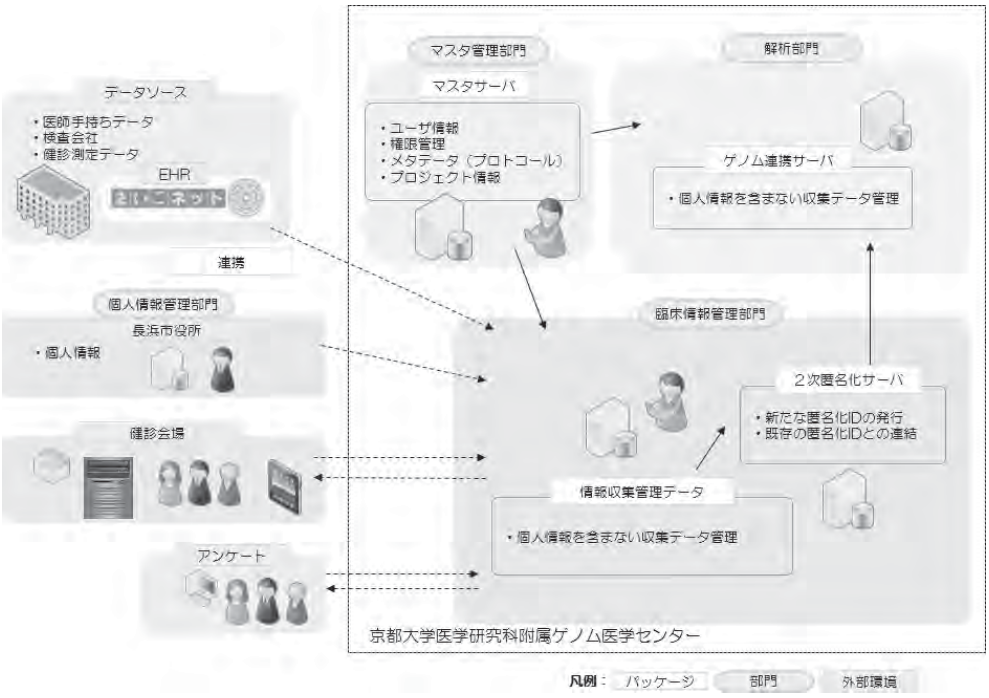

図 2　システム全体概要図

大規模ゲノム疫学研究の統合情報基盤のシステム全体概要図を図2に示す[3]．研究者が研究プロトコール（3.4節に後述）を定めて（マスタサーバ），臨床情報管理部門が匿名化する（匿名化サーバ）など法令やガイドラインを遵守しながら，食生活や生活習慣を追跡し（情報収集管理サーバ），解析部門でのゲノム・蛋白質や代謝物などの網羅的解析結果（ゲノム連携サーバ）と合わせて，健康や発病にどう結び付いたか調査する，という流れとなる（実線はオンライン連携，点線はオフライン連携）．

3.3　個人情報保護モデルの策定

大規模ゲノム疫学研究の統合情報基盤の構築は，個人情報保護が前提となる．「ながはまルール」では，二段階匿名化による個人情報保護が規定されており，個人データや試料が京都大学へ提供される際に個人情報を匿名化し，京都大学内で研究者へ提供される際に再度匿名化することで，個人情報を厳格に保護する必要があった．

本プロジェクトでは，個人情報保護に関連したガイドラインおよび「ながはまルール」に規定されている匿名化を実現するため下記の対応を行った．

1）　連結可能匿名化

ながはま0次健診は，被験者に健診結果をフィードバックする必要があるため連結可能匿名化を採用している．連結可能匿名化とは，必要な場合に個人を識別できるように対応表を残しておく方式を指す．連結可能匿名化を行ったうえで，研究機関において対応表を保持していない場合は，個人情報に該当しない情報となる．本プロジェクトでは，匿名化サーバに保存されている対応表を確認することでIDの可逆化を可能としている．

2）　再匿名化

本プロジェクトでは，異なる個人識別IDの匿名化処理を複数回実施する仕組みを構築した．二段階匿名化の概要を図3に示す．可逆変換（匿名後IDを匿名前IDに戻す変換処理）

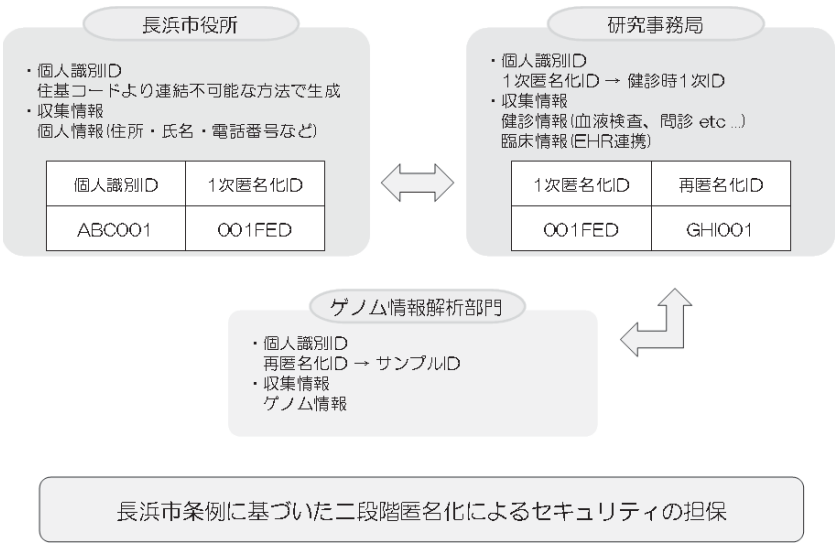

図3　二段階匿名化の概要

を行う場合は，匿名化を行ったルートを逆に辿らないと元に戻せない．このため，匿名化処理を行うすべての匿名前後 ID を知らなければ，データベース上の個人識別 ID から，個人情報を導き出すことができない．

3） 秘密分散法

匿名化サーバに保存されている対応表について，データベースのセキュリティ機能を強化するため，AES による暗号化に加え，データ分割機能によってデータベースのデータを分割管理することにより，データ漏洩に対する秘匿性を保障する仕組みを構築した．データ分割では，分割されたデータを異なる複数テーブルに格納し，さらに各テーブルも異なるサーバ上の複数ディスクに配置する．データの復元には，分割されたテーブル上のデータを結合する．このため，一つのテーブル上の断片化されたデータからのみでは，元のデータを復元することは不可能であることを保障する．秘密分散法のイメージを図 4 に示す．

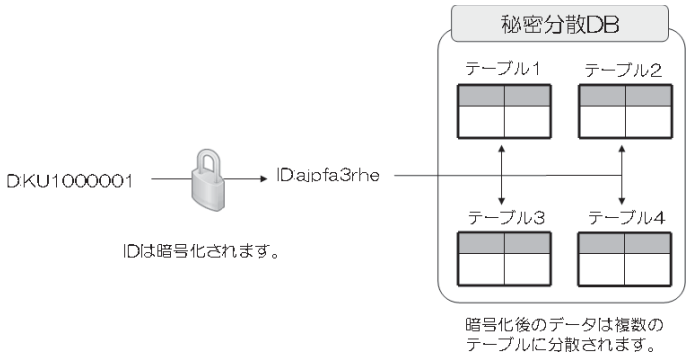

図 4　秘密分散法イメージ

4） 匿名化サーバの閉塞と閉塞解除

匿名化サーバは夜間の一定時間しか開放せず，日中は特別な権限を持つユーザが閉塞解除を実施しない限り接続することができない仕組みを構築した．

3. 4　メタデータ管理

ゲノムコホート研究では，一般に研究を実施するにあたっての「研究計画」を定義する．この研究計画のことをプロトコールと呼ぶ．プロトコールには，解析に利用するデータ項目のセット，分析方法（モデルと手法）が定義されている．蓄積されたデータからプロトコールに応じたデータセットを抽出し，研究に利用する．

本プロジェクトでは，プロトコールをメタデータとして定義し，情報を収集する仕組みを構築した．これにより，下記の 3 点のメリットが生まれる．

1） 個別抽出プログラムの構築期間を省き，研究開始までの期間が短縮可能
2） 即時利用可能な形でデータセットの抽出が可能
3） 定義の再利用が可能

メタデータ（metadata）とは，データに関する情報を記述したデータであり，data about data と英語で表現されることもある．これは，実際にデータベースに格納される生データに対して考えられるもので，データベースの構造と内容に関する属性などの情報を意味している．ゲノムコホート研究におけるメタデータとは，例えば「生年月日」という項目は日付型データ，「性別」という項目は，1 バイト長の文字列型データでありコードとして「1：男性」「2：女性」を使用するというような情報を意味する．このため，メタデータを把握することにより，データベースの構造や内容を把握することができるようになる．一般にこれらのメタデータは個別のアプリケーションに定義されているケースが多い．しかしながら，ゲノムコホート研究のようにデータの構造が研究対象疾患ごとに変化する（例えば，糖尿病と膠原病では研究で必要な診療データが異なる）という状況に柔軟に対応して複数のデータベースを統括・管理するという目的のためには，メタデータ自体をデータベースとして管理して各疾患研究データベース間で共有しておき，その中から必要なデータ定義を選択して疾患ごとのデータベースを作成する方法を用いることが効果的である．さらに，メタデータとして管理する情報の中に入力画面のレイアウト情報を含める仕組みを構築することにより，効率化すなわち作業負担と管理コストの抑制を実現することができる．また，事前にデータの形式を表 2 に示した「データ型」として定めておくことで，解析を行う際に扱いやすいデータを収集することができる．

表 2　データ型

データ型	内容
continuous	連続値（数値）を扱う
binomial	2 値から選択を行うデータを扱う
catord	3 値以上で順序の存在するカテゴリから単数選択を行うデータを扱う
cats	3 値以上で順序の存在しないカテゴリから単数選択を行うデータを扱う
catm	3 値以上で順序の存在しないカテゴリから複数選択を行うデータを扱う
date	日付に関するデータを扱う
time	時刻に関するデータを扱う
string	文字列を扱う
biallelic	遺伝子多型を扱う

3. 5　入力画面の自動作成

本プロジェクトは，全国展開を視野に入れた汎用性の高いパッケージの開発が目的である．入力画面については，ゲノムコホート研究で用いるデータ項目のセットを決めれば，Web フォームが動的に作成される仕組みが求められた．

入力フォームを開発するにあたり，問診業務で要求された 1 画面に複数のデータ項目を表示するケースと，アンケート調査業務で要求された 1 画面に単項目を表示するケース（1 問 1 答形式）の二つのパターンを開発した．

1)　1 画面に複数のデータ項目を表示するケース

臨床情報入力画面を図 5 に示す．体裁の微調整（表示順，見出し，表示方法（ラジオボタン（縦/横），セレクトボックス，チェックボックス（縦/横））も Web 画面から設定可能

である．ラジオボタンやセレクトボックス等のオブジェクトは，データ型に応じた選択が可能である．

図 5　臨床情報入力画面イメージ

2）1 画面に単項目を表示するケース

設問の回答画面は，表 3 の全 14 種類の中からデータ型に応じた画面パターンが選択可能である．また，アンケート画面は iPad での表示を考慮したデザインとしている．アンケート入力画面を図 6 に示す．

表 3　アンケートの画面パターン

数値選択	ラジオボタン（1 列）	時刻選択型（24H）
Visual analogue scale	ラジオボタン（2 列）	時刻選択型（AM/PM）
単一ボタン選択	チェックボックス（1 列）	テキスト入力
複数ボタン選択	チェックボックス（2 列）	非表示
日付選択型（和暦）	日付選択型（西暦）	

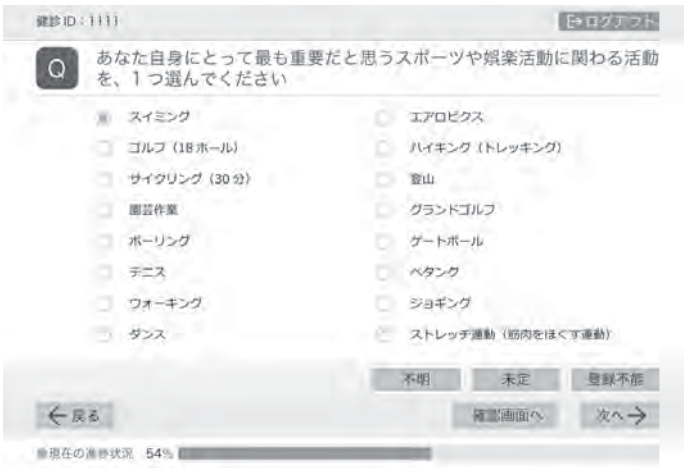

図 6　アンケート入力画面イメージ

アンケート調査業務についての全体概要を図7に示す．アンケートは設問数が700項目以上と多かったため，自宅にインターネット環境がある被験者は事前回答してもらう運用とした．自宅で回答済の被験者は，会場では回答する必要はなく，途中まで回答した被験者は続きから再開できる．その他，工夫した点として，アンケートの回答画面で音声を再生することができる仕組みを構築した．MP3形式の音声ファイルを一括で登録し，回答画面ではプレイヤーを利用して，アップロードした音声を再生することが可能である．

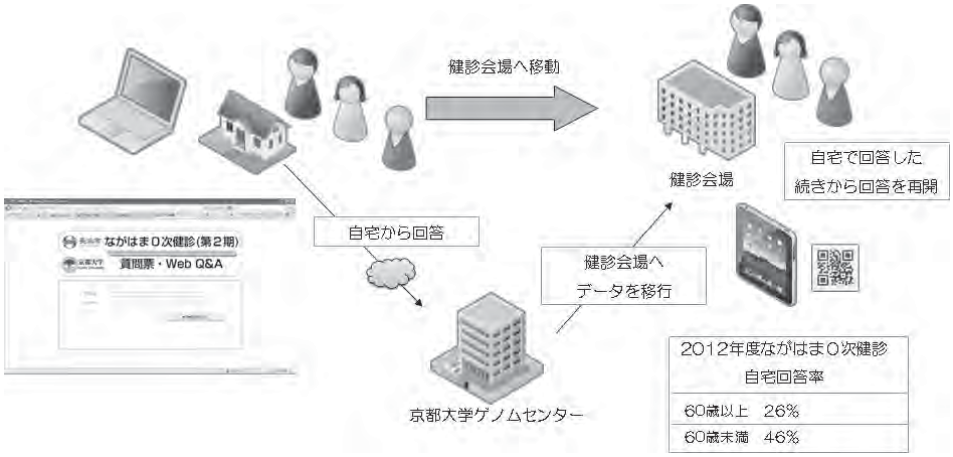

図7　アンケート調査業務全体概要

3.6　データ品質の保持

個人差が発生する情報について，システムを利用することで回答の抜け漏れや誤入力が防止できる．本プロジェクトでは，収集する情報間の論理矛盾の解消として，「欠損値・矛盾値などの確認と補正」を目的に，表4に示した論理矛盾チェック機能を開発した．

表4　論理矛盾チェック機能

種類	内容	例
単項目チェック	最小・最大値，有効桁数，日付・時刻のフォーマット，項目数の値チェック	身長（250cm以下）
回答権の制約	複数の項目間での論理矛盾チェック	男性は出産に関連する質問の回答権がない
値の制約	複数の項目間で値が取りうる範囲の値チェック	0 < 喫煙期間 <= 現年齢－喫煙開始年齢

回答権および値制約については，メタデータのデータ項目IDと演算子を用い，論理式で記述させる仕組みとした．論理式は，複合演算・基本的な集合演算が可能である．論理矛盾チェック機能を設けることにより，プログラムレスでビジネスロジック作成を実現した．

論理矛盾チェック機能は，論理式をBNF記法にて定義することによって単純化している．これによりJavaCCを利用することができ，論理式の妥当性を検証することを可能としている．このような手法を活用することは独自に論理式を解析することと比較し，パフォーマンス面でも助力となっている．

3. 7　マスター情報一括入出力

Excelにメタデータ，入力画面のレイアウト情報，入力値の制約（論理矛盾）の情報を記載し，一括で登録する仕組みを構築した．登録後は内容の確定と被験者の登録を行い次第，プログラムレスで健診をはじめることができる．また，アップロード時にエラーチェックを行い，不正な値が登録されることを防ぐ．

マスター情報一括入出力機能を設けることにより，マスター情報更新作業の入力工数を大幅に下げると同時に，管理者によるマスター定期メンテナンス作業の負荷も軽減することを実現した．また，アップロードしたExcelファイルをダウンロードすることで，同様の研究を行う他の研究者に，即時活用可能な形でデータベースの枠組みを提供可能としている．マスター情報一括入出力機能概要を図8に示す．

図8　マスター情報一括入出力機能概要

4.　今後の展開

医療情報の収集については，現状ではMML（Medical Markup Language）[*2]が対象であるが，SS-MIX（Standardized Structured Medical record Information eXchange）[*3]やレセプト情報に対応することにより，効率的な収集が可能になると期待できる[4]．

第一段階として，長浜の拠点病院を対象にし，連携データの洗い出し，連携方式の実装を行い，妥当性を検証する．EHR（Electronic Health Record）と連携をし，複数疾患/複数医療機関というモデルに到達させることを最終目標と考えている．

5.　お　わ　り　に

本稿で述べたような，データ利活用を図る仕組みは，医療・疫学研究の場面に限らず，広く適用できると考えている．本稿で述べた「大規模ゲノム疫学研究の統合情報基盤の構築」は，2012年11月にながはま0次健診で本稼働を迎えたが，2014年度も継続して開発中であり，有益なデータ出力，情報入力効率の向上，情報の誤入力防止に向けて機能強化を図っている．

最後に，本稿の事例となった大規模疫学研究の統合情報基盤のシステム構築を行った方々，および本稿執筆にあたりご指導を賜りました方々に厚くお礼を申し上げます．

＊1　ゲノムコホート研究とは，健診者に対して長期に亘って追跡し，ゲノム情報（遺伝子情報），医学的な情報，環境・生活習慣の情報を発症前から収集し，発症する疾病，治療の効果とその反応を将来にわたって分析すること．
＊2　MML（Medical Markup Language）．MedXML コンソーシアムで開発された異なる医療機関（電子カルテシステム）の間で，診療データを正しく交換するために考えられた規格．
＊3　SS-MIX（Standardized Structured Medical record Information eXchange）．厚生労働省電子的診療情報交換推進事業で定義された標準的電子カルテ情報交換フォーマット．

参考文献

［1］ ながはま 0 次予防コホート事業，NPO 法人　健康づくり 0 次クラブ，http://zeroji-club.com/0ji_kohot.html
［2］ ながはま 0 次予防コホート事業，NPO 法人　健康づくり 0 次クラブ，http://zeroji-club.com/0ji_kohot.html#Content5
［3］ 松田文彦，大規模ゲノム疫学研究の統合情報基盤の構築，JST バイオサイエンスデータベースセンター，「基盤技術開発プログラム」および「統合化推進プログラム」平成 24 年度 進捗報告会，2013 年 1 月 21 日 http://biosciencedbc.jp/gadget/rdprog_over/H24-t10_matsuda.pdf
［4］ 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，31 巻 2 号，2012 年
※上記参考文献の URL は 2014 年 9 月 12 日時点での存在を確認．

執筆者紹介

沖　　俊　吾（Shungo Oki）
2006 年(株)ハルクより日本ユニシス(株)に転籍．サービスインダストリ事業部ヘルスケアビジネスの UniCare 担当 SE として医療情報システム開発，適用に従事．2010 年より疫学データベースシステム開発に従事し，2011 年から総合技術研究所先端技術ラボに所属．2012 年から公共システム本部ヘルスケアサービス部に所属．医療情報技師．