医療システムでの非構造化データ活用事例

Use Case of Unstructured Data in Health Care System

沖　　俊　吾

要　約　複数の医療機関の診療データや健診機関の健診データを数十年に渡り蓄積し，疫学研究用にデータを提供する疫学データベースシステムを構築した．本プロジェクトを遂行するにあたり，基盤機能として低コスト，大量データの格納（拡張性を含む），個人情報保護の要件を満たす必要があった．

低コストと大量データの格納の両立は，OSS（オープンソースソフトウェア）の Key-Value ストア型分散データベース製品である Cassandra を核に，並列分散処理基盤製品の Hadoop や全文検索エンジンの Solr を組み合わせて実現した．個人情報保護は，診療情報および健診情報と個人情報の分離（匿名化），およびアクセス経路の制限により確立した．

本稿では，疫学データベースシステムの構築を通して，医療データの活用事例を紹介する．

Abstract　In order to provide research data for the epidemiological study, the epidemiological database system was constructed, which stores medical data and health examination data gathered and accumulated in a number of medical organizations and health examination data from over several decades. Upon carrying out this project, the basic functionalities must meet the basic requirements for the low cost, the massive data store (including scalability), and the personal information protection.

Satisfying both of the low cost and the massive data store is realized through combining the full-text search engine Solr and the parallel distributed processing infrastructure product Hadoop using Cassandra the Key Value type distributed database product from OSS (Open Source Software) as a core. The personal information protection is realized through the separation of personal information from checkups and medical records (making anonymized data), as well as limited access passes.

This paper introduces the use case of the medical data through the epidemiological database system construction.

1. は　じ　め　に

糖尿病患者は，日本に 890 万人存在すると報告されている[1]．また，そのうちの約 4 割が治療を放置している（患者が通院していない状態）と言われている．佐賀県は全国に比べ，糖尿病患者の糖尿病性腎症進展率が高い[2]．糖尿病性腎症は合併症に至ると高額な透析治療が必要となり，医療資源と患者への負担となる．そのため佐賀県では，1 次予防[*1] による患者の早期発見と，2 次予防[*2] による医療費の削減および患者治療の標準化を目的とし，疫学[*3] データベースシステムを構築した．これは，複数の医療機関の診療データや健診機関の健診データを数十年に渡り蓄積し，科学的根拠のある予防医学の研究基盤の構築や，研究成果の臨床への還元に貢献するものである．

本稿では，総務省の 2010 年度「地域 ICT 利活用広域連携事業」における「佐賀県糖尿病医療連携推進事業に係る，システム開発及び運用保守業務委託事業」[3]入札案件プロジェクトを

通して，医療データの活用事例を紹介する．

2. 疫学データベースシステムに求められた要件

本プロジェクトの目的は，複数の医療機関の診療データや健診機関の健診データを数十年に渡り蓄積し，糖尿病予防の疫学研究にデータを提供する疫学データベースシステムを構築することである．総務省の公募案件であり，限られた予算でのシステム構築が求められた．

医療機関，健診機関は個人情報や診療記録など機微な情報を取扱っている．このため，文部科学省の「疫学に関する倫理指針」や厚生労働省の「医療情報システムの安全管理に関するガイドライン」など複数のガイドラインが規定されており，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．基盤機能の要件[*4]を表1に，アプリケーション機能の要件を表2に示す．アプリケーション機能は，基盤機能の上に構築されるため，要求事項を実現するには，基盤機能の要件を解決する必要があった．

表1　疫学データベースシステムに求められた基盤機能の要件[*4]

要求事項	詳細
低コスト	ハードウェア／ソフトウェア購入費用，初期構築費用，および保守費用が限られた予算の範囲内である．
大量データ格納	医療機関の診療データと健診機関の健診データを数十年単位に渡り蓄積し続ける（数百TB規模のデータを想定）．
拡張性	サーバはスケールアウトにより拡張可能とし，大量のデータに対してデータロードやデータ検索のパフォーマンスが劣化しない．
個人情報保護	個人情報から個人を識別することができる情報の全部または一部を取り除き，代わりにその人と関わりのない符号または番号を付す(匿名化)．

表2　疫学データベースシステムに求められたアプリケーション機能の要件

要求事項	詳細
検索	性別や傷病名等の値の完全一致検索と，検査結果や身長・体重等の上限値／下限値指定による範囲検索を行う．検索条件はユーザが指定できるフィールド指定検索とする．また，検査名，薬剤名，アレルギー等，複数の任意文字列による全文検索を行う．
名寄せ	医療機関，健診機関に跨った患者の特定を行う．名寄せにより医療機関と健診機関を跨り，かつ時系列に患者を追跡可能とする．

3. 疫学データベースシステムの構築事例

3.1 システム全体概要

佐賀大学地域医療支援センターのデータセンタに，疫学データベースシステム基盤をプライベートクラウドで構築した．疫学データベースシステムを構築するにあたり，

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） 疫学研究には個人情報は必要ないため，データを匿名化し検索可能とする
3） 診療データや健診データを統計解析するためのデータ抽出機能を設ける
4） 医療機関と健診機関のシステム間でデータのフォーマットやコードを標準化する
5） 診療データの連携フォーマットは今後の複数医療機関への拡大を考慮し，SS-MIX[*5]を使用する

これらの要求を満たすため「ゲートウェイ」「データ変換」「検索データベース」「個人情報管理データベース」「疫学データベース」の五つのコンポーネントで物理的に別サーバに分離す

るシステム構成とした（図1，表3）．これは目的やデータ所有者，物理的な配置場所を考慮した構成であり，個人情報保護に関連したガイドラインに規定されている匿名化を実現するためのアーキテクチャである．

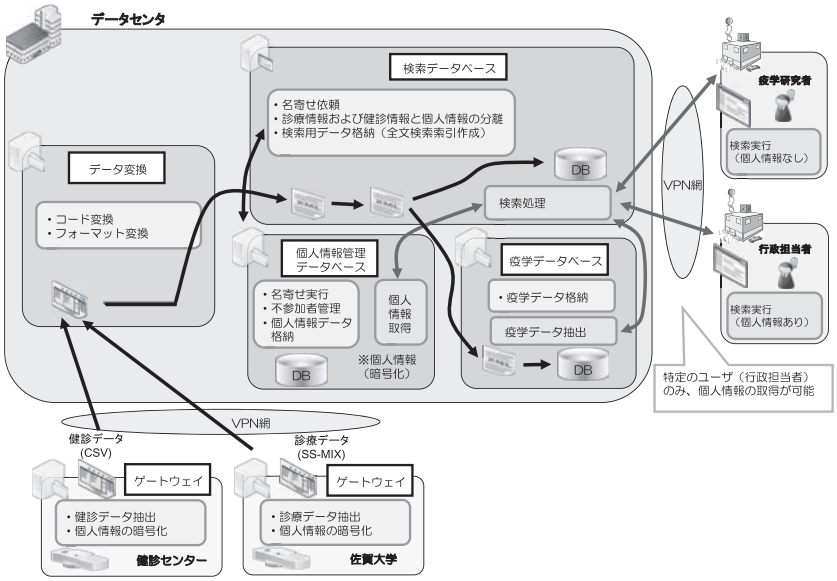

図1　システム全体図

表3　コンポーネントの役割

コンポーネント	役割
医療機関ゲートウェイ 健診機関ゲートウェイ	・医療機関や健診機関の情報システムから診療データや健診データを抽出 ・個人情報の暗号化
データ変換	・コード変換 ・フォーマット変換
検索データベース	・診療データや健診データから個人情報を分離（匿名化） ・匿名化した診療データと健診データに対する全文検索用データ（全文検索用インデックス）を作成 【データ検索処理】 ・フィールド指定検索用の検索条件設定 ・フィールド指定検索，全文検索
個人情報管理データベース	・患者の名寄せ（個人の特定），個人識別番号（UID）の払い出し ・個人情報の暗号化 ・個人情報の格納
疫学データベース	・匿名化した診療データ，健診データの格納

3.2　低コストと大量データ格納の両立

疫学データベースシステムが大規模化すると，ハードウェア/ソフトウェアの購入費用およ

び保守費用が増大する．また，データが大規模化することで性能（データロード時およびデータ検索時性能）の低下が想定され，データベース管理者によるチューニング作業の運用管理コストが発生する．システム拡張に伴う性能の劣化は，大規模改修やシステムの再構築を伴う場合がある．これらの課題を解決するため，各データベースサーバのオペレーティングシステム（OS），およびミドルウェアには以下に挙げるOSS製品を採用し，導入コストを抑えた．使用ソフトウェアのうちデータベースに関連するものについて，その役割を表4に示す．

- サーバOS（共通）　：　CentOS（Linux）
- WEBサーバ（共通）　：　Apache Tomcat
- 疫学データベース　：　Hadoop，Cassandra（3ノード構成）
- 検索データベース　：　Solr，Fess
- 個人情報管理データベース　：　MySQL

表4　使用ソフトウェアの役割

OSS	役割
Hadoop	診療データや健診データなどの大量データを疫学データベース（Cassandra）の複数サーバへ並列に分散配置する実行処理基盤である．診療データや健診データのバックアップとしての機能も併せ持つ．
Cassandra	診療データや健診データなどの大量データを複数サーバに分散して格納する分散データベースである．
Solr	診療データや健診データを，検査名，薬剤名などのキーワードで検索するための全文検索エンジンである．
Fess	Solrと連動した全文検索用ポータルである．ユーザインタフェースはGoogle風であり，マニュアルレスで利用が可能である．
MySQL	患者の個人情報を管理するリレーショナル型データベースである．

3.3　スケールアウト可能なデータアーキテクチャ

複数の医療機関の診療データや健診機関の健診データを蓄積し続けるため（数百TB規模のデータを想定），データの増加に伴い検索処理やデータロードのパフォーマンスが劣化しないようにする必要があった．そのため，データベースはスケールアウト方式のKVSであるCassandraを採用した．Cassandraは複数のサーバが連携して一つのシステムとして稼働するため，参加医療機関が増え，データが増える場合には，データセンタの疫学データベースサーバを増やすことにより対応する．追加するサーバはコモディティサーバで済むため，コストを抑えることができる．また，データベース構造は各サーバが並列で管理するため，単一サーバに障害が発生してもサービスを継続でき，サーバの追加もダウンタイムなしで可能である．冗長性の確保として，データのレプリケーション数を設定することにより，データは複数のサーバに自動でレプリケーションされる．

Cassandraは検索処理をサーバに跨り並列実行するため，大量データに対する検索性能が劣化しない[4]．また，データロードのパフォーマンス対策として，診療データや健診データを解析した中間形式ファイルをHDFS（Hadoopの分散ファイルシステム）に格納し，MapReduce（Hadoopの並列分散処理フレームワーク）を使用して並列処理でCassandraにデータロードすることで，処理時間の短縮を図った（図2）．次節で述べる，Cassandraに対する転置索引作成処理でも，MapReduceを使用した．

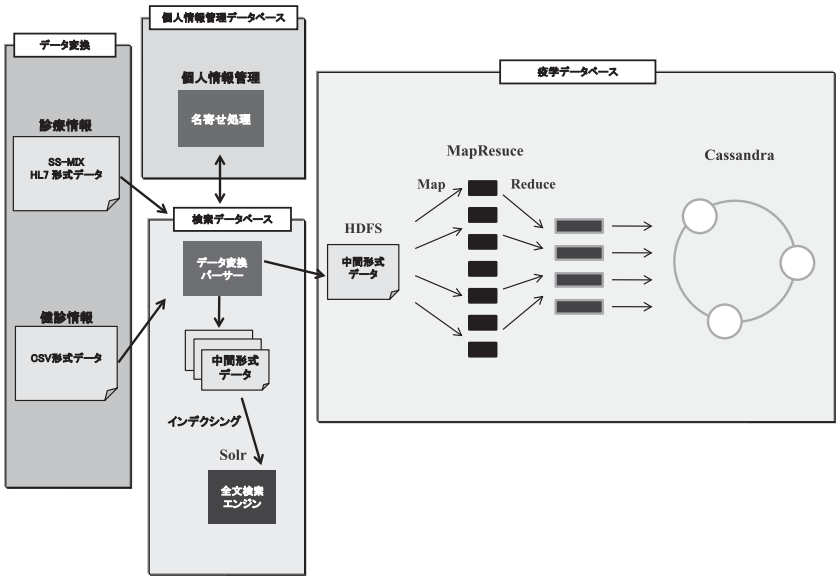

図 2　データロード処理手順

3. 4　非構造化データの格納

診療データや健診データは，患者や健診者に対して時系列にデータが発生する追加型のデータであり，検査項目や傷病名等は省令改正により項目が増加する．処方データは様々な種類の薬剤があり，同じ薬剤でも患者によって用法が異なる．血液検査（検体検査）結果は，検査種別毎に単位が異なり，検査値は数値や文字列の場合があるため，二次元の表形式に収まりきらない非構造化データである．こういった特性を持つデータを，Cassandra の「スキーマレス」「Key-Value ストア型」の特長を生かして，以下のようにデータベースに格納した．

診療情報や健診情報のファクトデータは，一つの大きなカラムファミリー（RDBMS のテーブルに相当．以降「ファクトテーブル」と呼ぶ）に格納した．患者に対して時系列で分析するという利用シーンが主であるため，ファクトテーブルのキーを「日付_UID」とし，カラム（RDBMS の列名相当ではなくデータ値を保持する構造体）には処方や検査結果等の診療情報を格納した．レコード毎にカラム名を自由に設定する「スキーマレス」の特長を生かし，傷病名コード値や検査コード値をカラム名の接頭辞に設定した．

RDBMS における検索は，SQL 文を発行することで検索条件に任意項目を指定することができるが，Cassandra における検索は，検索条件にキーを指定する必要がある．このため，範囲検索の手段として，対象とする検索項目（カラム）に対して転置索引用のカラムファミリーを作成した．具体的には，検査結果値の範囲検索用として，キーに検査結果値，カラムに「日付_UID」（ファクトテーブルのキー）を持つカラムファミリーを作成した．転置索引用のカラムファミリーに対して範囲検索し，「日付_UID」（ファクトテーブルのキー）を抽出後，ファクトテーブルからデータを抽出する．転置索引用のカラムファミリーを用いて検索が実行される様子を図 3 に示す．

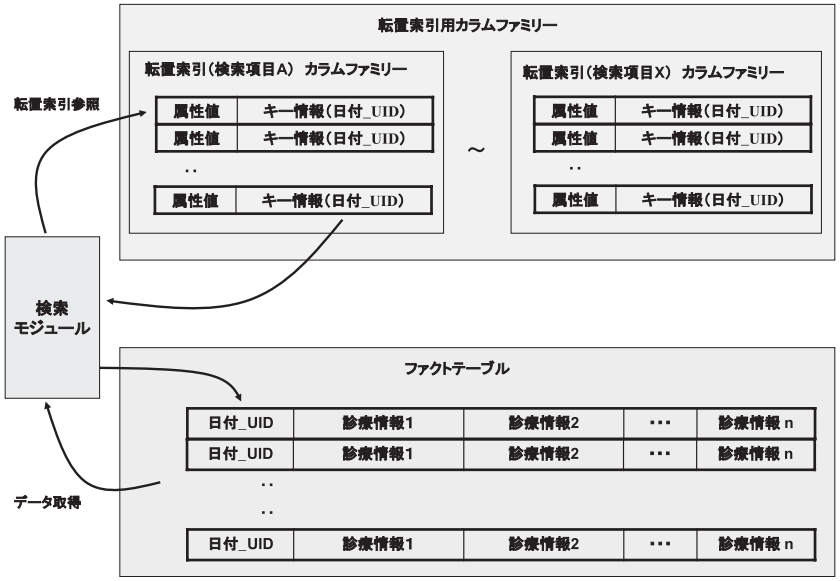

図 3　転置索引用のカラムファミリーを使った検索

3. 5　非構造化データの検索

検査結果には基準値があり，数値に対して上限値/下限値を指定する範囲検索や，性別や傷病名等の値で完全一致検索を行うという利用シーンがある．範囲検索と完全一致検索は「フィールド名」で指定されたフィールドに対する「検索語」による検索である．また，検査の所見情報による検索を含め，複数の任意文字列を検索項目の指定なしで検索するという利用シーンがある．これら二つの利用シーンに対して柔軟かつ高速な検索が求められたため，フィールド指定検索と全文検索の二つの機能を実装した．以下に説明するとともに，非構造化データ検索の概要図を図 4 に示す．

1）　フィールド指定検索

ⅰ）利用者が画面から検索項目を設定する仕組み

検索対象項目の追加や，転置索引カラムファミリーを作成する設定画面および検索画面を実装した．転置索引用カラムファミリーの作成処理は，該当 Web サービスのリクエストを検索データベースから疫学データベースへ発行することにより実現した．

ⅱ）論理演算の実装

RDBMS に対する検索は，SQL 文を発行することで検索条件に任意項目を指定することができるが，Cassandra に対する検索は，検索条件にキーを指定し，バリュー（値）を取得する方式である．データ構造が単純なため，データの取り出し時間が短くて済むが，キーは一意である必要があるので，部分一致による検索や，論理演算（SQL における条件句の AND，OR）や表結合（JOIN）ができない．そのため，本プロジェクトでは独自クエリ言語を用意し，検索条件文を含んだ Web サービスのリクエストを検索データベースから疫学データベースへ発行することで，条件検索を実現した．これにより，検索データベースから疫学データベースを見た場合，RDBMS の SQL のように検索リクエストを発行することができる．

2）　全文検索

Cassandra ではキーを検索条件項目とする必要があるため，任意カラムに対するあいまい検索や部分一致による検索はできない．本課題を解決するため，本プロジェクトでは，検索データベースの全文検索機能に全文検索エンジンである Solr を採用した．ファクトテーブルのキーである日付と UID を基に診療情報，健診情報のドキュメントを作成し，文字列検索用のインデクシングを行っている．Solr 機能で文字列を検索し，検索結果より日付と UID を抽出後，Cassandra のファクトテーブルのデータを抽出する処理を開発した．本方式により，検索処理と抽出処理をシームレスに連携している．

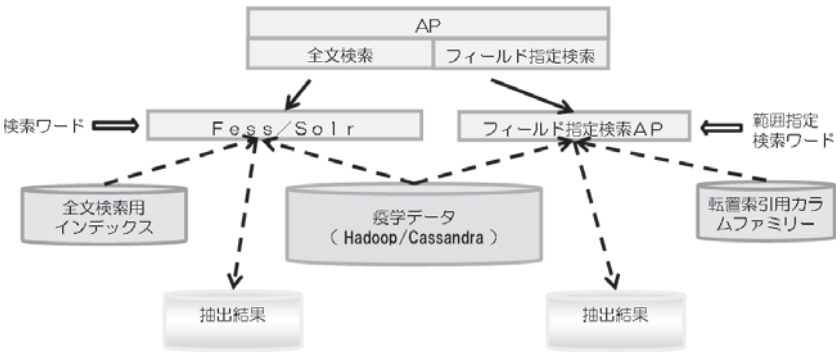

図 4　非構造化データ検索

3. 6　個人情報匿名処理

個人情報や診療記録などの漏洩を防ぐため，通信の暗号化（IPSec-VPN）に加えて，個人情報の匿名化を行った．診療情報および健診情報と個人情報を分離し，個人情報は疫学データベースには保有せず[*6]，個人情報管理データベースに暗号化して保有する仕組みとした．しかしながら，行政機関は糖尿病予備軍等の病院受診率の実態を調査し，未受診者を特定して受診

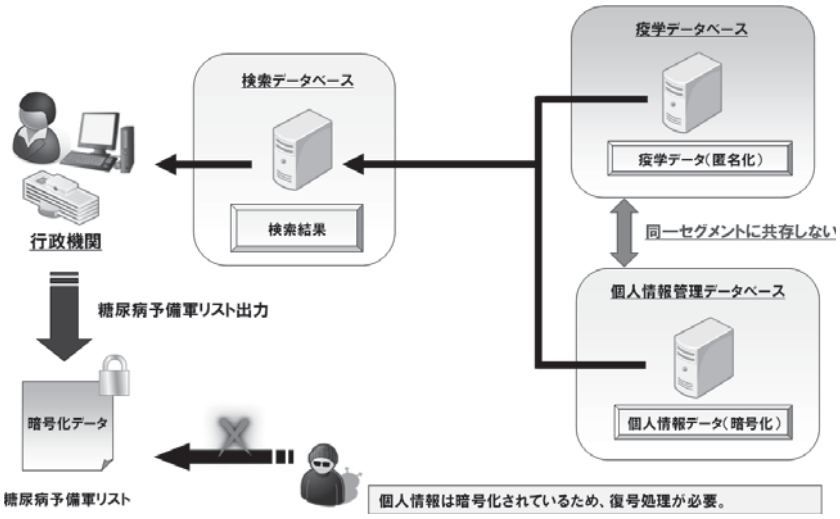

図 5　健診情報の可逆化

をフォローする必要があったため，疫学データと個人情報データを，個人識別番号（UID）で連結可能とした．すなわち疫学データと個人データを連結した暗号化データを行政機関が管理しているキーにより復号し，可逆化することが可能な連結可能匿名化方式とした．健診情報の可逆化の概念図を図5に示す．

4. 今後の予定

2011年度の疫学データベースシステムの展開予定として，京都大学医学研究科付属ゲノム医学センターと共同研究で，メタデータ管理を重要テーマとしたシステムの構築がある．複数の医療機関と連携する場合，異なる電子カルテシステムより診療データを抽出して標準化する必要が生じる．また，複数の疾患を対象とする場合，共通的に取り込むべき診療データの明確化などの検討，検証が必要である．第一段階として，一つの疾患を題材に，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．2012年度以降は，1疾患/1医療機関を，1疾患/複数医療機関，複数疾患/1医療機関というモデルに拡大し，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．

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．疫学データベース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「0：男性」「1：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，疫学データベース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できる．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と診療データマッピング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．

5. お わ り に

複数のOSS製品を組み合わせて新しい価値を生み出し，データ統合システムを短期間に構築した．分散している大量データの統合と個人情報の保護を両立し，データの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．また，OSSを活用してクラウド基盤上にこのような仕組みを実現したいという要求は高まると予想される．

＊1　疾患の発生を未然に防ぐ行為．生活習慣の改善や予防接種など．
＊2　疾患を早期に発見・処置する行為．健康診断や早期治療など．
＊3　地域や集団内で．疾患や健康に関する事象の発生の原因や変動するさまを明らかにする研究．
＊4　具体的な数値については伏せている．

＊5　SS-MIX（Standardized Structured Medical 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．
＊6　生年月日は，疫学研究で重要な情報だが個人情報であるため，偽装して格納した．

参考文献　［1］　平成19年国民健康・栄養調査結果の概要について，報道発表資料，厚生労働省，2008年12月，http://www.mhlw.go.jp/houdou/2008/12/h1225-5.html
［2］　佐賀県保健医療計画，佐賀県，2009年3月，http://www.pref.saga.lg.jp/web/var/rev0/0023/4494/keikaku1.pdf
［3］　佐賀県糖尿病医療連携推進事業に係るシステム開発及び運用保守業務委託契約に関する仕様書，佐賀県，2010年12月
［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2012年1月30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．医療情報技師．