できるが、更新頻度等、運用上の都合により別マスタとした。これら3つのR-MIMは、レルム「日本(JP)」を付けバージョン番号01として定義した。また、既存のボキャブラリドメインを考慮しつつ、29の値集合を定義した。そのうち、既存のボキャブラリのみで対応可能なものは4つであった。図1に、本研究で定義した薬品マスタR-MIM図を示す。

4. 考察

本研究で開発したモデルは、神戸大学病院、麻生飯塚病院で運用されている薬品マスタに適応可能であり、薬品コードとしてローカルなコードと共にHOTコード[2]を採用しているため、今後、他施設の薬品マスタについても検討し、必要な情報を整理、追加していくことにより標準的な薬品マスタになりえると考える。また、一薬品一ファイルであるため、医薬品に付随する情報の一元的な管理が可能であり、医薬品の適切な使用や安全性の確保、流通・在庫管理業務の効率化に寄与すると考える。なお、開発した薬品マスタを、神戸大学病院で現在開発中の電子カルテシステムで実際に使用し、実用性の評価を行っていく予定である。

5. 課題

今回は既存のボキャブラリドメインに概念の存在しないコード化値についてはローカルなコード化体系を参照する値集合を定義し対応した。HL7v3のボキャブラリドメインでは既に保健医療分野における各領域の概念が体系的に整理され、コード化がなされている。しかし、そこで定義されているコード化値を日本におけるメッセージ定義の際にも使用できるとは限らない。日本特有の概念が存在するのも事実であり、今後、HL7との対応をとりつつ、日本でのコード化体系を整理していく必要があろう。

参考文献

[1] HL7 Version 3 Standard, Health Level Seven Inc., Available at http://www.hl7.org/, 2004.

[2] 標準医薬品名マスター, Available at http://www.medis.or.jp/4_hyojyun/download/index.html, 2004.

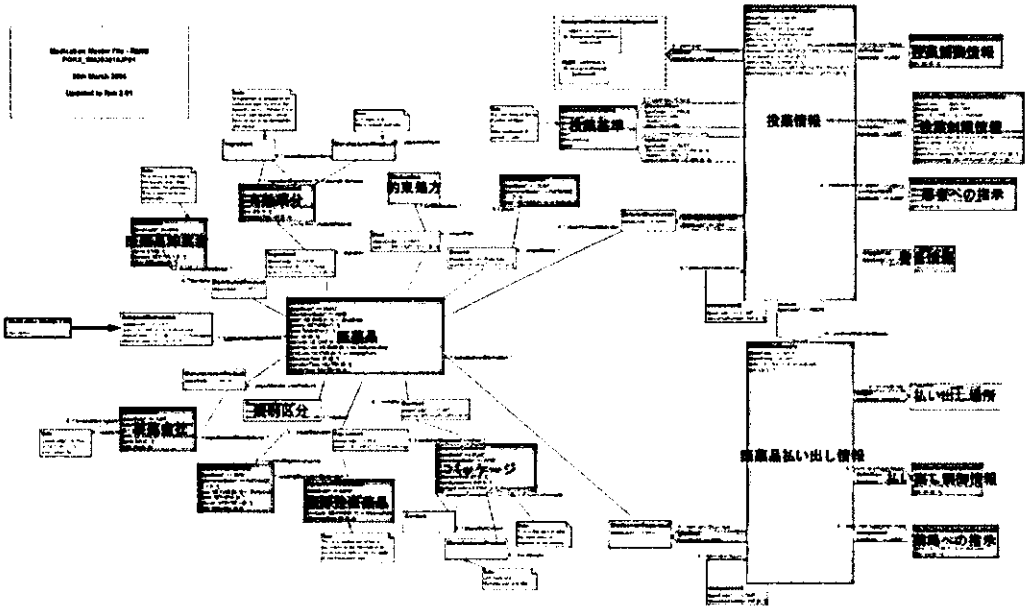

図1 Medication Master File R-MIM

電子化に向けた糖尿病カルテの分析

山本 さつき 前田 英一 星本 弘之 坂本 憲広
神戸大学 医学部附属病院 医療情報部

Analysis of The Progress Records of The Diabetic for Electronic Medical Records

Satsuki Yamamoto Eiichi Maeda Hiroyuki Hoshimoto Norihiro Sakamoto
Division of Medical Informatics, Kobe University Hospital

Abstract: To investigate an efficient way to the structural progress recording, we analyzed terms in the current paper-based medical records. Frequently used terms were picked up from each record of 100 diabetic subjects, and calculated the ratio of utilization of them in all patients. Most highly utilized term among all subjects was 'Plasma Glucose' (60%), however, majority of highly utilized terms were about general findings mainly in the objective findings record. In suffered subjects with the specific diabetic complications, terms about their complication were appeared in every day's records, however, their utilization ratio in all subjects were low. It is considered that suitable combinations of terms that are categorized by different characters in utility are useful for electronic progress recording.

Keywords: Electronic Medical Records, Standardization

1. はじめに

診療録の電子化において、経過記録用紙の記載は、その多様な記載様式のため標準化が進んでおらず、しばしば、自由文を中心とした記録となっている。経過記録用紙には、患者の主訴や、他覚的所見など診療時の症状のほか、医師の診断過程や診療計画案、治療評価などが記載されている。それらは、医療情報として重要な部分であり、経過記録を自由文として記載にすることは、電子診療録開発目的の一つでもある、記述の標準化や構造化、データマイニングへの利用性を損なうことになり、解決すべきテーマの一つである。我々は、糖尿病という罹患率が高く臨床像がよく知られた疾患をモデルとして、現在の紙診療録の記載を基に、経過記録用紙が実際にどのような形で記録されているかを検討し、項目の標準化について考察した。

2. 方法

平成15年度に、本院糖尿病内科に入院した患者の入院診療録を対象として、個々の診療録内で使用頻度の高い項目を抽出し、それらの項目が、全調査対象診療録間において普遍的に使用頻度の高い項目として用いられているかを検討した。まず、診療録ごと、項目ごとの使用頻度の分布を概観するために、最初に20症例分の診療録において、医師による経過記録用紙に記載されている全用語を項目立てし、項目ごとに、毎日の経過記録の記載の中での出現回数を計数して入院期間中の全記載日数に対する出現回数の割合を求めた。その結果を基に、単一診療録内で出現回数が20%以上記載のある項目を、「使用頻度の高い項目」と設定した。ついで、全調査対象診療録の経過記録から「使用頻度の高い項目」を抽出し、その項目が全調査対象診療録において、どの程度普遍的に高頻度として用いられているかを、「再現率」として表した。なお、一冊の同一診療録であっても、担当医の交代などで表記方法の変化があった場合は、別の診療録として取り扱ったため、最終的に、対象診療録数100冊、同記載医師数17名となった。それらを基に、経過記録に記載されている項目を、共通項目として電子診療録に展開することが、可能かどうかを検討した。

3. 結果

各診療録から使用頻度の高い項目として抽出された項目は、全対象診療録100冊において30項目であった。その中で再現率が3%以上であった項目および再現率を表1に示した。

最も再現率の高い項目は「血糖値」であり、60%の診療録において、日々の経過記録中に20%以上出現する「使用頻度の高い項目」であった。ついで再現率の高い項目は、「BP」(血圧)、「BT」(体温)、「HR」(心拍数)といったバイタルサインに関する項目、「heart」「S1S2clear」「no murmur」(心音)、「N. V. S.」「no rale」(胸部聴診所見)、「abdomen」「soft & flat」「グル音」(腹部所見)といった一般的他覚的所見に関するものであった。

主訴に関する項目の再現率は、「no new complaint」(「著変なし」「no remarkable」「stable」)の17%を除いては、他覚所見に関するものと比べて全体に低く、3%以上の再現率を示した項目は「浮腫」「口渇」のみであった。「ふらつき」「浮遊感」「倦怠感」「冷や汗」などの低血糖症状に関する所見、「手足のしびれ」などの末梢神経障害に関する所見は、再現率は1%程度と高くはなかった。また、患者の訴えや感想等に関する記載は、同様の記載がほぼ毎日記述されており、単一の診療録に

おいては高頻度に出現したが、それらは、患者自身の言葉に基づく多様な表現により記述されていたため、異なる患者間での再現性は高くならなかった。
評価・立案の項目の記述に関する項目は、「BSコントロール」におけるインシュリン投与についての記載が再現率29%であった。なおインスリン投与の記載については、表記が様々であり、パターンがいくつか見られたが、今回の調査ではまとめて扱った。

4. 考察

今回の検討は、複雑な記載がなされる経過記録を構造化・標準化するための最初のステップとなる用語の項目立てが可能か否か、その効果はどの程度のものかを考えるために、糖尿病という比較的頻度が高い疾患を用いて行った。その結果、単一診療録において使用頻度の高い項目は、その診療録の中ではほぼ毎日記載されるなど繰り返し使用される傾向にあったが、全調査対象診療録を通してみた場合、再現率の高い項目は、主に他覚所見、なかでも疾患に特徴的ではない一般的な他覚的所見に集中する傾向が示された。これは糖尿病といえども一般的内科疾患でもあり、通常の内科診療で施行される他覚所見が高率であったことは不自然なものではなく、これら再現率の高い項目を電子診療録の標準的項目とすることは意義が高いものと考えられた。一方、糖尿病性合併症など、糖尿病に特徴的な所見に関する項目は、単一診療録においては極めて繰り返し使用されていたが、全対象診療録の中では再現率が高くなかった。これは、糖尿病合併症という糖尿病に特徴的なものであっても実際にはそれほど頻度の高いものではないためと思われた。従って、これらを全患者に適用する標準的な項目として扱う意義は必ずしも高いものではないが、必要な症例においては利用可能とすることは、意義が高いものと考えられた。さらに、「腹部エコー」「グルカゴン負荷」等の内科領域、糖尿病領域で頻繁に行われる検査項目は、ほぼ全ての対象診療録において一度は用いられる項目であったが、今回の調査では日々の経過記録を対象としたため表1には挙げられていない。しかし、これらは出現頻度こそ低いものの、診療過程においては、必ず使用される項目であり、日々の経過記録とは別のグループとしての項目化は必要と考えられた。このように、項目の用いられ方の性格に併せて、項目化の方法を工夫することにより、多くの記載が構造化・標準化できるものと思われる。今回の調査では、毎日診察が行われていたにも関わらず、経過記録には記載があまりない診療録も少なからず見られた。これらは、糖尿病の症状には日々大きな変化がなく、変化がなければ所見があっても経過記録として特に記述しないケースがあるためと思われたが、このようなケースであっても、項目を適切に設定することにより、日々記録される所見情報の量的質的向上が期待されるものと思われる。
一方、主訴や所見についての記述には微妙な表現も多い。例えば、「変わりない」という患者の言葉は、症状がないという意味と、前日との変化がなく、依然として神経障害などの持続的症状が続いている場合も含まれている。また患者によって表現の仕方も多様であり、その意味合いも微妙に異なっている。主訴のニュアンスをそのまま表現することもまた必要であると思われ、自由文入力も不可欠と考えられた。今後、本来の意義を損なうことなく診療録の電子化を進めて行くにあたり、これらをいかに効果的に組み合わせていくかが検討される。

表1 高再現率項目リスト

項目	再現率
血糖値	60 %
BP	35 %
BT	35 %
BSコントロール	29 %
HR (PR)	30 %
heart (no rale)	17 %
heart (no murmur)	14 %
abd (soft&flat)	13 %
PR	13 %
heart (S1S2 clear)	13 %
abd グル音	12 %
lung (N. V. S.)	12 %
no new complaint *	17 %
BW	7 %
浮腫	5 %
口渇	3 %

*「著変なし」「no remarkable」「stable」などを含む

PKIを用いた広域対応の臨床試験情報収集システムの構築

西腋 清行[1] 増田 剛[2] 坂本 憲広[1]
神戸大学医学部附属病院 医療情報部[1]
財団法人先端医療振興財団 臨床研究情報センター 遺伝子データベース研究部[2]

Construction of a wide area clinical trial information reporting system with PKI

Kiyoyuki Nishiwaki[1] Go Masuda[2] Norihiro Sakamoto[1]
Department of Medical Informatics, Kobe University Hospital[1]
Translational Research Infomatics Center,Foundation for Biomedical Research and Innovation[2]

Abstract: Paper based management of clinical trial data is inefficient, because the data is collected from nationwide. Therefore, construction of efficient data management system with online data collection would be cost effective. To assure the system's security, Public Key Infrastructure is suitable for the purpose.

However, a large sum of cost is needed to construct CA in the medical institutions in the whole country. Then, we construct a RootCA, and the certificate is issued from HPKI compliant SubCA in each medical institution. The digital signature can be verified, because RootCA is the same. And reductions in cost are possible.

In our practical experiment, HPKI compliant SubCA in Kobe University Hospital and TRI were used. SSL client authentication was introduced in the clinical trial data collection server, and digital signature was added to the clinical trial data. So security countermeasures were done.

It was possible to register and inspect the clinical trial data, it was also possible to verify the electric signature in the demonstration. However, there was a problem with operativeness deteriorates by SSL client attestation, addition of the digital signature, etc.

Therefore, in the future, it is important to construct system that users can use without the consideration of PKI.

Keywords: PKI

1. はじめに

全国多施設共同で実施される臨床試験のデータは、紙ベースで授受されてきたためにデータ管理の効率が悪く、インターネットを使用したオンラインシステムによる効率のよいデータ管理を行うシステムの構築が不可欠である。また、臨床試験データには個人情報が含まれるためユーザ認証・通信の暗号化と、データが誰によって作成されたものなのか確認する必要もあり、普及しつつあるPKI(公開鍵基盤) を利用する方法がある。

しかし全国に広がる医療機関から独自に電子証明書を発行・管理するに、RootCA設置、CP(証明書ポリシー)やCPS(運用規定)の独自策定などで多大な費用が必要となり、データを収集・利用する際には複数のRootCAを信頼する手間がある。

この問題を解決するにあたり、共通RootCAを設け、各施設にSubCAを設置し、RootCAの信頼、一箇所のCAから多施設のユーザへの証明書の発行の手間を省く方法がある。

そこで本研究ではMEDIS-DCのHPKI(ヘルスケアPKI) 準拠RootCAと、それをRootとするTRI((財)先端医療振興財団臨床研究情報センター)、神戸大学医学部附属病院のHPKI準拠SubCAを使用し、その有用性を検証した。

2. 方法

TRI、京都大学医学部附属病院、大阪大学医学部附属病院、株式会社エスアールエルの関係者にはTRIより電子証明書を発行し、神戸大学医学部附属病院の関係者には院内のSubCAより電子証明書を発行した。また臨床試験情報サーバはTRIに設置し、各施設のPCから模擬的な臨床試験データを入力してもらうことにより実験を行った。

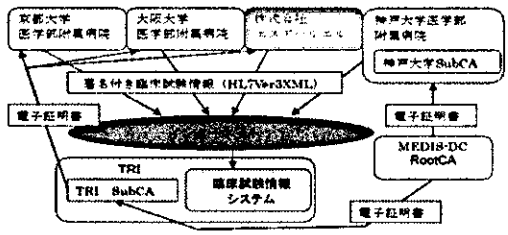

図1 システム構成図

2.1 臨床試験情報収集サーバシステム

収集サーバはMicrosoft Windows 2003 Serverを使用し、各施設の臨床試験データ登録・閲覧サイトはSharePoint Serviceにて構築した。セキュリティ対策としてWebサーバへの認証には、SSLクライアント認証を使用し、各施設のWebサイトへの認証はユーザ名・パスワードを用いた。SSLクライアント認証にて各施設のサイトへの認証もIIS (Internet Information Services)の証明書マッピング機能を使

用することで可能であるが、ユーザ数が多くなるとユーザと証明書のマッピングに時間を要するために見送った。

2.2 臨床試験情報収集クライアントシステム

臨床試験データを入力するクライアントアプリケーションにはMicrosoft InfoPath 2003を使用することにより、入力しやすいGUIと、HL7 Version3に準拠したXMLを作成しサーバへ保存する機能を提供した。

臨床試験データ登録時は、サーバに登録されているInfoPathフォームテンプレートをIE (Internet Explorer)上から開くことによりInfoPathが起動され、データを入力することが可能である。閲覧時はInfoPathにより作成されたXMLファイルをIE上から開き、InfoPathにて閲覧を行う。

作成されたXMLの真正性の確保と作成者・認証者の確認を行うため、サーバに保存する際には作成されたXMLに電子署名の付加が必須となるように構成した。さらに、サーバに保存されているXMLを閲覧する際は、電子署名の検証を行い、かつ電子署名に使用された証明書がHPKI準拠であるか検証を行うことにより、XMLの真正性と作成者・認証者の確認を行えるようにした。下記は証明書の検証時に行った確認事項である。

- 公開鍵証明書の発行CA署名検証
- 公開鍵証明書信頼チェインの確認
- 公開鍵証明書の証明書ポリシーを確認
- 公開鍵証明書有効期限の確認
- CRLの確認
- KeyUsage、BasicConstraintsの確認
- EE証明書のhcRole

3. 結果

同じ施設のユーザ間では、同じSubCAから証明書が発行されているため、他のユーザが登録した臨床データを閲覧する際の電子署名の検証が正常に行うことができるのはもっともであるが、他施設のユーザ間で異なるSubCAから発行された証明書を使用した電子署名でも、正常に検証可能であった。また、WebサーバによるSSLクライアント認証もRootCA、各SubCAを信頼しているため、各施設のユーザがSSLクライアント認証を行うことができ、権限の無い者による登録・閲覧は不可能となり、かつ通信の暗号化・改ざん防止が可能であった。

臨床データの登録・閲覧時に使用する証明書がCAにより失効された場合は、SSLクライアント認証は不可能となり、登録・閲覧時が不可能となった。また、すでに登録されている臨床試験データの電子署名の検証も失敗した。

4. 考察

臨床試験データに付加した電子署名の検証では、証明書が一度失効されると以前に登録していたデータも検証できなくなるため、中・長期的な臨床試験のデータを収集する際には問題が生じる。よって、運用方法または電子署名の付加方法を見直す必要がある。

また、実験終了後に行った電子署名の必要性(図1)と今後のネットワーク型電子カルテの利用(図2)についてのアンケートでは、実証実験に参加したすべての人が電子署名は必要であると回答した。しかし、今後のネットワーク型電子カルテの利用については、積極的な回答は少なく、その理由として自由回答ではセキュリティ上の問題と手間の増加があげられた。これは、エンドユーザの多くが最近多発する個人情報漏洩事件を問題視しているが、実務の効率性が落ちることも同様に問題視していると思われる。よって、RootCAを一つにすることにより、各施設にCAを構築する際のコストの問題は解決できるが、ユーザが満足するセキュリティの確保と、PKIを使用した場合の利便性の低下の相反する二つの大きな問題が浮き彫りとなった。

このことからも、今後はユーザが意識しなくともセキュリティの確保できるシステム構築も重要である。

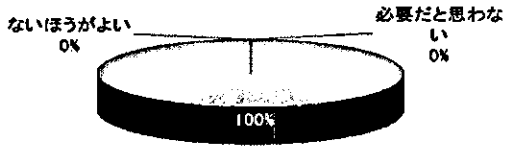

図2 電子署名は必要性があると思うか

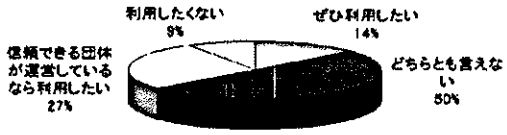

図3 今後、インターネット通信の「電子カルテ」システムを利用したいと思うか

保健医療情報標準化規格に基づく疾患関連遺伝子解析研究のためのデータ収集基盤

増田剛[1] 広井嘉栄[2] 坂本憲広[2]
財団法人先端医療振興財団臨床研究情報センター遺伝子データベース研究部[1]
神戸大学医学部附属病院医療情報部[2]

A Data Gathering Infrastructure Based on a Healthcare Data Exchange Standard for Candidate Genes Analysis

Gou Masuda[1] Kaei Hiroi[2] Norihiro Sakamoto[2]
Laboratory for Clinical Genome Informatics, Foundation for Biomedical Research and Innovation[1]
Department of Medical Informatics, Kobe University Hospital[2]

Abstract: Identification of disease related genes is one of the important research topics in post-genome era. Genome-wide association analyses with Single Nucleotide Polymorphisms (SNPs) are widely performed. In the analysis, it is pointed out that making use of detailed clinical information is essential in order to characterize and classify phenotypes. On the one hand, a lot of clinical data are collected electronically in healthcare institutions. However, it is difficult to analyze, compare and evaluate the data in a common measure because its format, code and normal value are different each other. A standardized method for clinical and genomic data is required. In this study, we develop a data gathering format for candidate gene analysis with clinical and genomic information based on HL7 Version 3 which is a standard protocol for healthcare information exchange. We then applied this format to the data set which is collected in the Japanese Millennium Genome Project. In addition that, we develop a piece of software library to develop any tools or applications based on the format. By using HL7 Version 3, the data gathering format becomes more complex than simple spreadsheet formats which are conventionally used. However, the software library makes it easy to handle such a complex format and to develop tools or applications based on the format.

Keywords: Clinical Genomics, Disease Related Gene Analysis, HL7 Version 3

1. はじめに

疾患の発症に関連した遺伝子を同定する疾患関連遺伝子解析研究はポストゲノム研究の重要な課題の一つである。現在、一塩基多型(SNP)を代表とする遺伝子多型を利用した全ゲノム探索による相関解析が広く行われている。しかし、糖尿病のような多因子疾患では、異なる病因の患者が同一の集団として扱われてしまうため偽陽性が起こりやすいという問題が指摘されており、この問題を回避するためには詳細な臨床情報を利用し表現型をさらに詳細に分類する必要がある[1]。一方で、臨床の現場では診療情報の電子化が進み、詳細な臨床情報が収集されつつある。しかしながら、これらを疾患関連遺伝子解析研究に利用する場合、データが各施設に固有の方法で記述されており、データ項目の定義や基準値、コードが統一されていないため、収集されたデータを共通に解析・比較・評価することが困難である。臨床情報と統合化された疾患関連遺伝子解析を行なうためには、標準化規格に基づいたデータの収集形式が必要である。さらに、このような収集形式がデータ収集のための情報基盤として利用されるためには、データ形式を簡便に扱うためのツールやライブラリが必要不可欠である。そこで本研究では、疾患関連遺伝子解析研究のためのデータ収集形式とその形式を簡便に扱うためのツールを構築することにより、疾患関連遺伝子研究のためのデータ収集基盤の構築を試みた。

2. 方法

保健医療分野における電子的な情報交換のための国際的な標準化規格の一つにHealth Level Seven (HL7) がある。特に、HL7の最新版であるHL7バージョン3(以下HL7V3)は、より表現力の高い情報モデルRIMを有しており、検査や処方データといった臨床情報だけでなく、アレルやSNPといった遺伝子情報をも統一的なモデルで表現することを可能とする。そこで本研究では、標準化規格としてHL7V3を用いた疾患関連遺伝子解析研究のためのデータ収集形式を開発する。また、著者らが開発しているHL7V3メッセージを扱うための基盤ライブラリ[2]を用いて、標準化規格の詳細を意識する必要なく、開発したデータ収集形式を扱うことを可能にするためのメッセージインターフェースを開発する。開発したデータ形式は、ミレニアムプロジェクト糖尿病疾患グループの収集データ[3]をモデルケースとして、その有効性を評価する。

3. 結果

3.1 データ収集形式

疾患関連遺伝子解析研究のためのデータ収集形式は、HL7バージョン3のメッセージ開発方法論に基づき、HL7RIMから導出されるHL7V3メッセージ型として定義した。臨床情報に関しては、臨床検査領域で定義されている検査結果メッセージを基礎として用いた。その結果、ミレニアムプロジェクトで収集される年齢、BMI、HbA1c、FPGといった、計44の項目をHL7V3の詳細化メッセージ情報モデル (R-MIM) として表現することができた。

一方、遺伝子情報に関して、現在HL7では、主に診療という観点からゲノム情報を扱うための標準化規格が検討されている。本研究では、その中で現在開発中である情報モデルを基礎として用いた。このモデルでは、SNPなどのゲノム情報は、観察という行為やその観察結果を表現するRIM Observationクラスを用いて、臨床検査結果と同様に統一的に表現される。このモデルを基礎として、疾患関連遺伝子解析研究で必要となる遺伝子やSNPの情報を、図1に示す通りHL7V3のR-MIMとして表現した。

3.2 データ収集基盤

標準化規格を用いた結果、従来までの単純なスプレッドシート形式と比較して、より複雑な形式となった。そのため、筆者らが研究を行なっているHL7V3メッセージングライブラリ[2]を利用して、規格の詳細を意識することなく簡便に扱うためのメッセージインターフェースを開発した。例えば、1つのSNPデータに対応するインターフェースISnpObservationを定義することで、SNPの識別子やタイピング結果を、HL7V3メッセージの詳細を知ることなくメッセージ中に設定・取得することができる。

4. 考察

遺伝子情報に関して、HL7で現在定義されている情報モデルは、主に診療の観点で定義されたモデルであり、個人に対して非常に多くのSNPを扱うユースケースが考慮されてはいない。しかしながら、疾患遺伝子解析研究の場合には、例えば、個人の10万のSNPデータを解析対象とするミレニアムプロジェクトの例のように、大量のSNPデータをメッセージとして表現する必要がある。今回モデルケースとして使用したデータは、1サンプルあたり58個のSNPデータを含み、これをHL7V3メッセージで表現すると約70KBのXML文書となった。単純に計算すると10万SNPでは約100MBのHL7V3メッセージとなり、1メッセージとして扱うことは実用的ではない。今後、複数のSNPデータを1つのObservationクラスの値として表現するようなモデルを考える必要がある。

5. おわりに

標準化規格を用いた結果、従来までの単純なスプレッドシート形式と比較して、より複雑な形式となった。しかしながら、標準化規格をサポートするライブラリとメッセージインターフェースを提供することで、規格の詳細を意識することなく扱うことが可能となった。今回明らかとなった大量のSNPデータの扱いに関する課題は、今後検討していく必要がある。

参考文献

[1] 坂本憲広: 臨床ゲノム情報学への誘い、医療情報学、23 (Suppl.)、138-139、2003.
[2] 増田剛、他: HL7バージョン3メッセージングライブラリの開発、医療情報学、23(Suppl.)、512-515、2003.
[3] 三宅一彰、他: Perspective: ミレニアムプロジェクトと今後の展望、Molecular Medicine、40 (9)、2003.

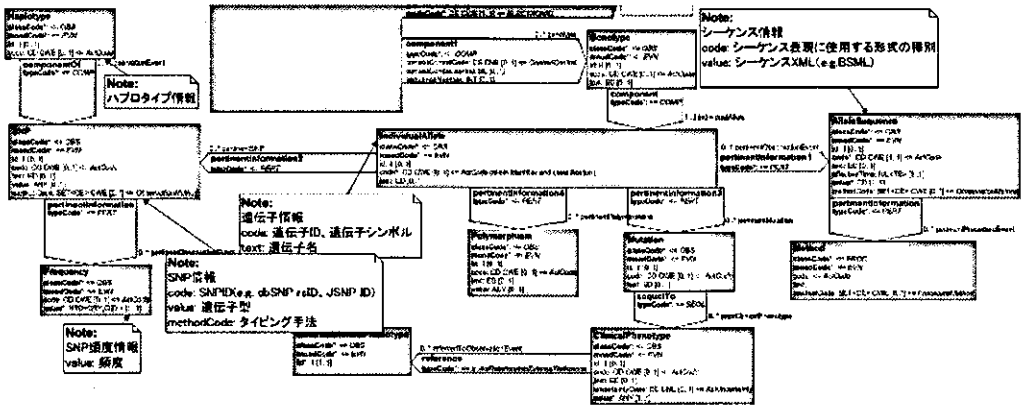

図1 開発した遺伝子情報のためのR-MIM (抜粋)

Towards the Construction of the Information Infrastructure for Genome Medicine

Norihiro SAKAMOTO

Department of Clinical Genome Informatics, Graduate School of Medicine, Kobe University

7-5-2, Kusunoki-cho, Chuo-ku, Kobe, JAPAN

nori@med.kobe-u.ac.jp

Abstract. One of the most expected applications of the human genome information is the development of genome medicine. Towards this goal, basic and applied biomedical laboratory researches and developments have been steadily in progress. However, clinical genome informatics and its applications in practice have been making slower progress compared to biomedical laboratory works. In order to facilitate the progress of clinical genome informatics and its powerful application to the development of genome medicine, the construction of the information infrastructure for genome medicine calls for urgent attention.

In this paper, we propose the architecture of the information infrastructure for genome medicine. It is required to provide three essential features: security mechanisms, comprehensive information models, and intelligent analyzers. To implement these features, we employed PKI as the security mechanisms and the HL7 Version 3 as the information models. To analyze information, knowledge discovery tools are expected to be implemented in addition to a lot of clinical and genetic statistical functions.

1 Introduction

The Human Genome Project was successfully completed in May, 2003. The project has produced the detailed map and the comprehensive dictionary concerning the human genome on a basis of about 3 billion nucleotide sequences. These results have changed the way researchers approach the life sciences and have pushed them to the post genomic era.

1.1 Clinical Genome Informatics

The major research topics in the post genomic era contain proteomics, system biology, genome-based drug discovery and other advanced life science areas. Among them, the research and development of genome medicine is one of the most expected applications of the human genome information. The research and development of genome medicine need both high throughput biological laboratory works and high performance computational power. Towards realizing genome medicine, basic and applied

biomedical laboratory researches and developments have been steadily in progress so far. However, informatics for genome medicine and its applications in practice have been making slower progress compared to biomedical laboratory works. Bioinformatics is the critical research field and the key information technology that supports genome medicine. Although bioinformatics seems to have been active in basic life science and genome based drug discovery among wide range of its scope, its activity in the application field of genome medicine seems to be lower. We have taken up and focused this research field of bioinformatics that supports the development of genome medicine and named it clinical genome informatics. In order to facilitate the progress of clinical genome informatics and its powerful application to the development of genome medicine, we consider that the construction of the information infrastructure for genome medicine calls for urgent attention. Towards the construction, we describe the architecture of the information infrastructure for genome medicine in this paper.

1.2 Research and Development Processes of Genome Medicine

As the first step of the research and development of genome medicine, genomic analyses for detecting disease susceptibility genes have been popular. The genomic analyses are mainly categorized into the following two approaches: the candidate gene approach and the whole genome scan approach. The whole genome scan approach for detecting disease susceptibility genes makes more use of the human genome information, particularly the single nucleotide polymorphisms (SNPs) and microsatellites (MSs), than the candidate gene approach. The whole genome scan approach is further categorized into the following subgroups based on the methodology of genetic statistics: the linkage analysis and the linkage disequilibrium analysis or the association study. In order to determine a disease susceptibility gene, the linkage analysis usually requires 10-200 samples or patients that are members of a few large families while the linkage disequilibrium analysis requires 200-2,000.

The main targets of genome medicine are so called common diseases that are very popular in today's society, for example, diabetes mellitus, hypertension, heart disease, cancer and so forth. The association analysis is more suitable for the genome analyses of such common disease because it is easier to collect a large number of samples of the common disease than to get a large family. The millennium genome project employed the approach based on the association study. The millennium genome project started in 2000 in order to detect susceptibility genes of the common disease. The project consists of the following 5 subgroups: Alzheimer's disease, asthma, cancer, diabetes mellitus (DM), and hypertension. The DM subgroups have been led by Prof. Masato Kasuga from Kobe University and we have been involved in the subgroup. Each of subgroups has collected and analyzed more than 200 patients (cases) and 200 normal controls on average. The DM subgroup collected 178 cases for the first screening and 752 cases and controls of the same number for the second and third screening. More than 15 medical schools have been cooperating to collect such a large number of cases and controls. This cooperation has been strongly

facilitated by using the Internet. We have set up a home page and shared analysis results. The accesses to the home page are secured through SSL.

Another example of such a large study was reported in the research paper on an affected sib pair study on DM by Prof. Hajime Nawata from Kyushu University [1]. They collected more than 100 cases and their siblings from more than 30 medical facilities. It is clear that the secure communication infrastructure through the Internet is indispensable to connect the participant medical facilities and researchers and share the information.

We need to handle various types of information in the research and development of genome medicine. Theses pieces of information include genome sequences, genomic structures such as promoters, exons and introns, amino acid sequences, protein structures and functions, genomic polymorphisms such as SNPs and MSs, metabolic pathways, signal transductions, and so forth. In the public biological databases, a part of them are usually managed in relational database systems. Such management is appropriate for the purposes of retrieving and browsing pieces of information. However, the efficient research and development of genome medicine requires the integrated management and analyses of the whole of those pieces of information. For this purpose, a relational model is not necessarily suitable but more powerful data model such as an object data model is preferable. Health Level Seven (HL7) is the international standards for the messaging and communications in the fields of healthcare and provides the reference information model (RIM). HL7 is now extending its scope to the filed of clinical genomics.

The common diseases are usually considered as a multifactorial disease. More than 10 genes in addition to environmental factors are involved in the development process of the disease. The extent of the involvement of each gene varies from case to case. According to such difference, a case shows different phenotypes or clinical symptoms. Therefore, efficient analyses of disease susceptibility genes require the combined analyses of the genome information and the clinical information. The genome information and the clinical information have been usually analyzed separately. In the genome analyses of disease susceptibility genes, the clinical information is used only when cases are diagnosed. However, in the genome analyses regarding the common diseases, patients of which present different phenotypes, any pieces of clinical information are required to be used in order to categorize the cases into subgroups. The combined information of the genome information and the clinical information also requires the combined method of clinical statistics, genetic statistics, and data mining. Such integrated intelligent approaches are critical for the efficient research and development of genome medicine.

1.3 Intelligent Database

To summarize the above requirements, the architecture of the information infrastructure for genome medicine shall possess the following features: security mechanisms, comprehensive information models,

and intelligent analyzers. We recognize the information infrastructure for genome medicine with these features as a natural extension of the intelligent database.

The intelligent database represents the state of the art in the evolution of database technology [2]. The intelligent database handles a variety of functions ranging from data modeling and management to information discovery, summarization, and presentation. The basic architecture of the intelligent database consists of three layers: the intelligent database engine, the object model, and a set of high-level tools that tend to vary somewhat according to the application [3]. The intelligent database handles the storage, retrieval and management of information. It also handles extraction of data from heterogeneous databases and the various communication and transaction protocols required by this activity. In terms of the intelligent database, the information framework for genome medicine is the clinical genome specific one featured by secure communication on the Internet. Therefore, the architecture of the information framework for genome medicine can be primarily a natural extension of the intelligent database.

2 Method

We designed the architecture of the information framework for genome medicine based on the concept of the extended intelligent database. Here, we describe the approaches of design and implementation for the three outstanding features: the security mechanisms, the comprehensive information models and the intelligent analyzers.

2.1 Security Mechanisms

For making use of the genome information and clinical information across healthcare institutes, standardization of data formats and vocabularies of representing the information are essential. The detail regarding this issue is described in the next subsection. In addition to them, digital signature is a critical key technology for making the information to be exchanged secure and trustworthy [4]. Especially for the genome information, digital signature is much more important because of its reusability. Efficient exchanges of authorized information with a digital signature in healthcare information networks require a construction of a public key infrastructure (PKI).

The necessity of PKI in the healthcare domain has been internationally discussed and recognized [5]. However, there have been few reports on the implementation and practical use of PKI for healthcare. A plan to use PKI for out-patients' prescription is under way in Korea [6]. Japan is ahead in a point of view of implementation of PKI because some of the EPR projects have already implemented a PKI and been operating a CA for different purposes [7]. One of the purposes of using a PKI is user authentication. A second is an access control and privilege management. A third is a digital signature on clinical information. A composite usage of them can be a forth purpose. We employed a PKI for the security

mechanism for the construction of the information infrastructure for genome medicine because it provides these various security solutions.

2.1.1 Community of the PKI

The community of a PKI is composed of the following entities: the root CA (Certification Authority), the sub CAs, and end entities. The end entities are in turn categorized into subscribers and verifiers.

- Root CA: The root CA authorizes the sub CAs by means of issuing PKCs (Public Key Certificates) for digital signature to them. It also authorizes itself as the root CA by a self-issued PKC.
- Sub CAs: The sub CAs authorize end entities by means of issuing end entity PKCs for digital signature. In this study, there are no intermediate sub CAs that issue PKCs to subjects that are CAs.
- Subscribers: The subscribers are end entities that are the subjects of PKCs, hold the PKCs and sign pieces of genome information and clinical information by using them. In terms of healthcare, they are referring doctors that send clinical information to a referred doctor.
- Verifiers: The verifiers are end entities that make use of PKCs and verify digital signatures. In terms of healthcare, they are research leaders that receive clinical information and genome information signed by a participant researcher of different medical institutes.

2.1.2 Policies and Statements

In order to manage the hierarchical PKI, coordinated security policies are essential. We defined the security policies to manage both the root CA and the sub CAs as two kinds of documents. These documents are referred to as Certificate Policy (CP) and Certification Practices Statement (CPS) [8].

A policy is a set of rules established to govern a certain aspect of organizational behavior. The CP addresses the components of a PKI in total. It describes the goals, responsibilities, and overall requirements for the protection of the CAs, PKCs and their supporting components. It is a high-level document that describes a security policy for issuing PKCs and maintaining certificate status information. This security policy describes the operation of the CA, as well as the users' responsibilities for the requesting, using, and handling of PKCs and keys.

Compared to the CP, the CPS is a highly detailed document that describes how a CA implements a specific CP. The CPS identifies the CP and specifies the mechanisms and procedures that are used to achieve the security policy. Each CPS applies to a single CA. The CPS may be considered the overall operations manual for the CA.

The policy information is indicated in the three policy extensions of the PKC: certificate policies, policy mapping, and policy constraints. Now, we have prepared the secure mechanism based on the PKI. Next we describe the comprehensive model that defines the forms of information to be communicated.

2.2 Comprehensive Information Models

The architecture of the intelligent database proposes the adoption of the object model to handle complex data. In the area of healthcare, HL7 is the international standards and have entirely adapted the object oriented technologies for the latest version, HL7 Version 3. Therefore, the architecture of the information structure for genome medicine is expected to follow HL7 Version 3. However, HL7 Version 3 has been strongly focusing on the clinical information. The HL7 information model for the genome information is not completed. Hence, in this paper we devise and propose the comprehensive information models that integrate the genome information and the clinical information by extending HL7 Version 3.

2.3 Intelligent Analyzers

The genome information and the clinical information require their specific analysis methods respectively. Namely, they are genetic statistics and clinical statistics. For efficient analyses of both the genome information and the clinical information, the analyzing system should provide the various kinds of statistical analysis functions.

2.3.1 Genetic Statistical Functions

The following functions are primary genetic statistical functions that the intelligent analyzers are expected to provide.

- Hardy-Weinberg Equilibrium
- Case-Control study
- Linkage Disequilibrium Analysis
- Haplotype Inference
- others

2.3.2 Clinical Statistical Functions

The following functions are primary clinical statistical functions that the intelligent analyzers are expected to provide.

- Paired t-test
- Student t-test
- Welch t-test]
- Mann-Whitney U-test
- others

In addition to these traditional statistical analysis methods, a sort of knowledge discovery approaches or data mining approaches are known to be powerful for efficient analyses of the integrated genome and

clinical information. Among the data mining approaches, we employed machine learning approach and back propagation neural network approach.

3 Results

Towards the construction of the information infrastructure for genome medicine, we implemented some software components for PKI and HL7. Then we developed a prototype system based on the architecture that was described in the above and evaluated it.

3.1 PKI Components

3.1.1 Configuration of the root CA

The root CA was installed in The Medical Information System Developing Center (MEDIS-DC). The X.500 distinguished name (DN) for it was set to 'c=JP, o=MEDIS-DC, cn=MD-HPKI-01-MEDIS -TopCA-for-CAs-and-TSAs'. The DN is a structured data type that supports a hierarchical naming system. The most common X.500 naming attributes are 'c' (country), 'o' (organization), 'ou' (organization), and 'cn' (common name).

The primary functions of the root CA are 1) issue of the certificate signing certificate to the root CA (the root CA signing certificate), 2) issue of the certificate signing certificates to the sub CAs (the sub CA signing certificates), 3) revocation of the sub CA signing certificates, 4) distribution of the authority revocation list (ARL) that lists the revoked sub CA signing certificates, and 5) distribution of the certificate revocation list (CRL) that lists the end entity signing certificates that are revoked by each sub CA. The certificate validity period of the root CA signing certificate is set to 8 years. That of the sub CA signing certificates is also set to 8 years. Distribution of ARLs and CRLs are made via HTTP protocol. The distribution points of ARL and CRL are http://repository.medis.or.jp/crl/root.crl and http: //repository.medis.or.jp/crl/, respectively.

The system architecture of the root CA is illustrated in Figure 1. Since it shall be operated in top level secure way, the CA system that generates PKCs and ARL is installed on the private network that is isolated from either the intranet of MEDIS-DC or the Internet. The web server for distributing ARLs and CRLs is installed on the demilitarized zone (DMZ).

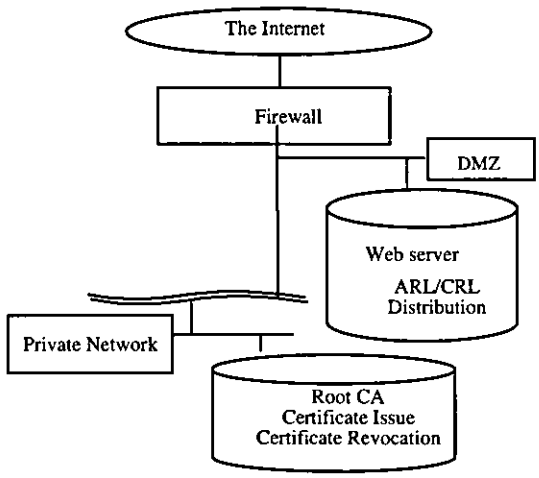

Fig. 1. The system architecture of the root CA installed in MEDIS-DC

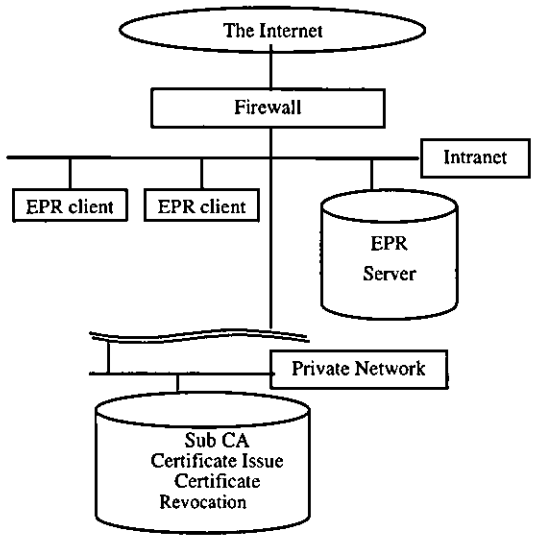

Fig. 2. The system architecture of the sub CA installed in Kobe University Hospital.

3.1.2 Configuration of the sub CA

The sub CAs were installed on the private networks in Kobe University Hospital and Kobe Translational Research Informatics Center. For example, The DN for Kobe University Hospital was set to 'c=JP, o=Kobe University Hospital, ou=Kobe University Hospital, cn=MD-HPKI-01-KUH-CA-for-non -Repudiation'. The DN of the end entities were, for example, identified by a DN like 'c=JP, o=Kobe University Hospital, ou= regulated health professional, cd=Real Name' where 'Real Name' is a doctor's real name.

The primary functions of the sub CA are 1) issue of the clinical document signing certificates to end entities (the end entity signing certificate), 2) revocation of the end entity signing certificates, 5) registration of the CRL to the CRL distribution point on the web server of the root CA. The certificate validity period of the end entity signing certificate is set to 4 years that is a half of that of the sub CA signing certificates.

The system architecture of the sub CA is illustrated in Figure 2. The CA system that generates PKCs and CRLs is installed on the private network that is isolated from either the intranet of each EPR project where EPR servers and EPR clients are running or the Internet.

3.2 HL7 Components

The clinical documents including individual genome information are composed of two components. The first component is clinical information and genome information itself. We designed the format of those information of a patient according to the Japanese Set of Identifiers for Medical Record Information Exchange (J-MIX) [14] and HL7 Version 3. The items and their XML elements are listed in Table 1. We exchanged the clinical information and the genome information as an XML document in conformity to J-MIX and HL7 HMD.

The second component of the clinical documents is a digital signature and its related information. Digital signatures are usually transferred with the signer's PKC that is used for verifying the digital signature and the CA PKCs that are necessary for building the certification path to validate the signer's PKC.

There are various choices regarding the format of the digital signatures and signed document formats. An XML signature is one of the most possible choices because we use J-MIX and HL7 Version 3 in an XML format. However, the implementation of the XML format is not straightforward. For example, it requires a lot of preprocessing of XML documents such as canonicalization. PKCS #7 is another possible choice. It is the de facto standard specification for protecting information with digital signature.

The basic PKCS #7 message format has two fields: the content type and the content. The content types defined by PKCS #7 are data, signedData, envelopedData, signedAndEnvelopedData, digestedData, and encryptedData. Since PKCS #7 is a basic building block for cryptographic applications, such as the S/MIME v2 electronic mail security protocol, there are already a lot of libraries or modules available on may platforms. Since this means the ease of implementation, we chose PKCS #7 as the signature format in this study.

A part of the XML scheme for the clinical information and genome information described in HL7 Version 3 format with a digital signature is illustrated in Table 2.

Table 1. J-MIX items that are used for the description of clinical information and genome information

Item code	XML element name	Data type
MD0010050	Patient.WholeName	String
MD0010110	Patient.Birthday	Date
MD0010120	Patient.Sex	Category
MD0010150	Patient.WholeAddress	String
MD0020180	Referral.Date	Date
MD0020220	Referring.Provider. Name	String
MD0020410	Referring.Physician. WholeName	String
MD0020480	ReferredTo.Provider. Name	String
MD0020670	ReferredTo.Physician. WholeName	String
MD0020730	ReferralNote	Text

3.3 Prototype System

We developed a prototype system by using the above components and evaluated it in Kobe University Hospital, Kyoto University Hospital, Osaka University Hospital and Kobe Translational Research Informatics Center. The system was developed using with Microsoft Internet Information Servers, Microsoft InfoPath, and Microsoft .Net C#. The overall system architecture was illustrated in Figure 3.

About 60 users participated in using the prototype system. First they applied for registration through the registration web page to the sub CA either in Kobe University Hospital or Translational Research Informatics Center depending on their affiliation. The sub CA verified the registration and issued a private key and the corresponding public key certificate stored in a USB token. The users generated pieces of translational research information that contained SNP information in some cases in an HL7 Version 3 format and signed it with the USB token. The signed documents were uploaded through the Internet by using SSL to the Translational Research Information System that was maintained in Translational Research Informatics Center.

The system was running successfully and the system architecture we proposed in this paper was evaluated to be appropriate.

Table 2. A part of XML scheme for a clinical document described in HL7 Version 3 format with a digitall signature

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <xs:schema targetNamespace="urn:hl7-org:v3"
elementFormDefault="qualified"
xmlns:fo="http://www.w3.org/1999/XSL/Format"
xmlns:msg="urn:hl7-org:v3/mif" xmlns:hl7="urn:hl7-org:v3"
xmlns:voc="urn:hl7-org:v3/voc" xmlns="urn:hl7-org:v3"
xmlns:my="http://schemas.microsoft.com/office/infopath/2003/myXSD/200
4-02-20T09:07:01" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:include schemaLocation="datatypes.xsd" />
  <xs:include schemaLocation="voc.xsd" />
  <xs:import
namespace="http://schemas.microsoft.com/office/infopath/2003/myXSD/20
04-02-20T09:07:01" schemaLocation="my.xsd" />
  <xs:element name="OutcomeResearchReport"
type="UUDD_MT990100.OutcomeResearchReport" />
- <xs:group name="UUDD_MT990100">
- <xs:sequence>
  <xs:element name="OutcomeResearchReport"
type="UUDD_MT990100.OutcomeResearchReport" />
  </xs:sequence>
  </xs:group>
- <xs:complexType name="UUDD_MT990100.OutcomeResearchReport">
- <xs:sequence>
  <xs:element name="id" type="II" />
  <xs:element name="recordTarget" type="UUDD_MT990100.RecordTarget" />
  <xs:element name="component" type="UUDD_MT990100.Component"
minOccurs="0" maxOccurs="unbounded" />
  <xs:element name="signature" type="my:SignatureType" />
  </xs:sequence>
  <xs:attribute name="classCode" type="ActClass" />
  <xs:attribute name="moodCode" type="ActMood" />
  </xs:complexType>
- <xs:complexType name="UUDD_MT990100.RecordTarget">
- <xs:sequence>
  <xs:element name="patient" type="UUDD_MT990100.Patient" />
  </xs:sequence>
  <xs:attribute name="typeCode" type="ParticipationType" />
  </xs:complexType>
```

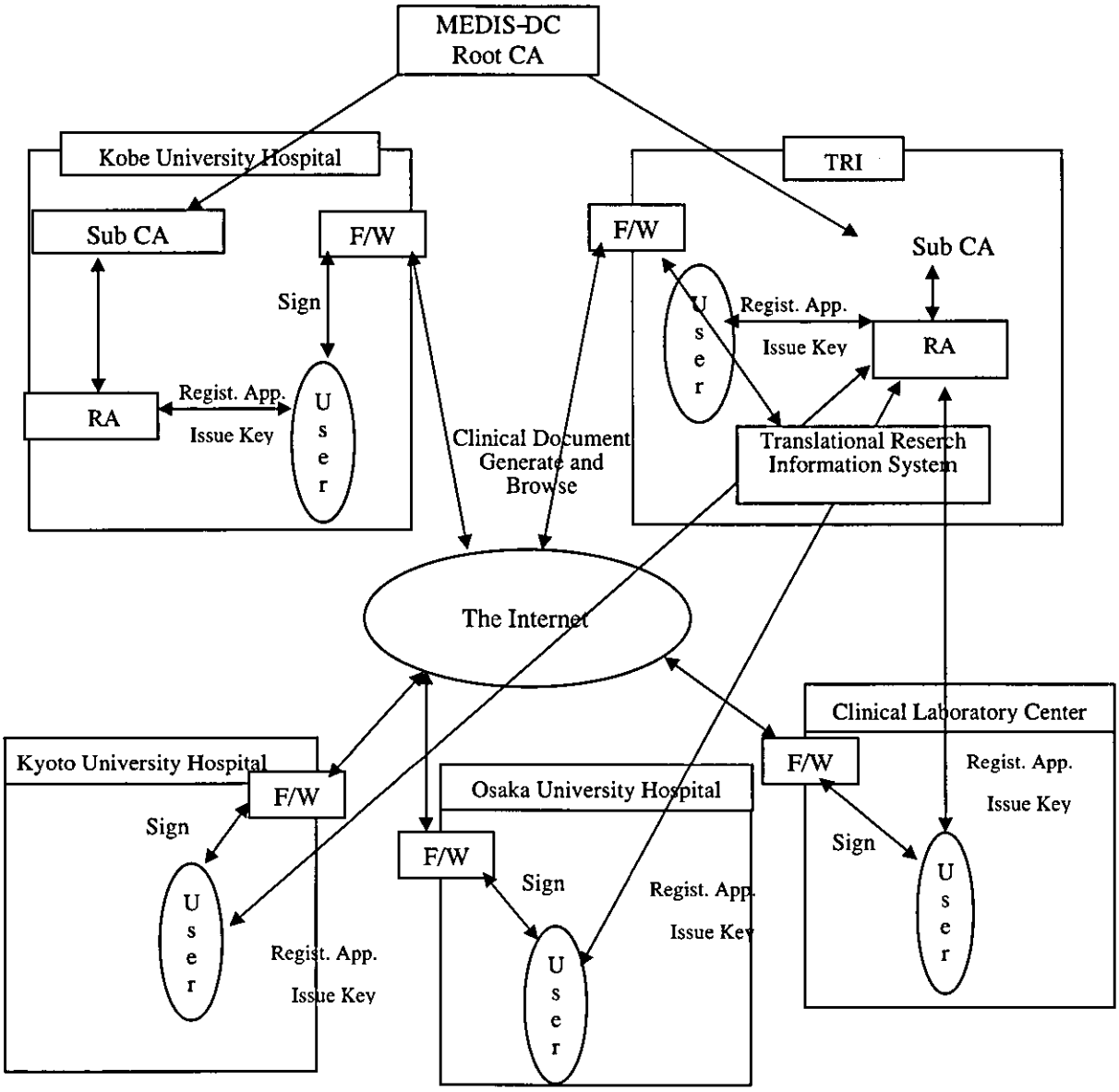

Fig. 3. The overall system architecture of the prototype system. The users in Kobe University Hospital posses the PKC that was issued by the sub CA installed in Kobe University Hospital while the other users possess the PKC that was issued by the sub CA installed in TRI.

4 Discussion

During the evaluation process of the prototype system, we had no technical problems and the system was running without any troubles. The most difficult part of the process was in the steps of registration of users and issuing their public key certificates because these steps required a lot of collaboration of the users who didn't know much about PKI. We needed to ask them to bring forward their resident's card for

the personal identification and their medical license for conformation of their qualification, This procedure put a burden on the users and it took about seven day for half of them to get a complete set of the papers. In order to make these steps much speedier and easier, social infrastructure for digital application should be developed. Especially, the development of nation wide PKI that is specific for the healthcare field and assert healthcare professionals is a pressing issue.

The cost of issuing the public key certificate is another serious problem. One PKC stored in a USB token cost about 24,000 yen or 220 USD. It is obviously higher compared to current seals and handwritten signature. .Though this problem about the cost is the one that is expected to be solved as PKI becomes popular and the number of the PKC in use increases, it is another pressing issue.

We succeeded in the development of the comprehensive information models for clinical information and genome information on a basis of HL7 Version 3. However, the more comprehensive the information models are getting, the more complicated the implementations becomes. In the development of the prototype system, the implementation of the clinical document of HL7 Version 7 XML format took the longest time (man month) in the whole implementation processes. A development of a set of software tools that supports the HL7 Version 3 is an urgent business for the construction of the information infrastructure for genome medicine.

We pointed out the importance of the intelligent analyzers that provide a lot of statistical functions for the information infrastructure for genome medicine and its essential components. However, its implementations are still in progress and were not included in the prototype system. In order to make the prototype system much more practical, we need to work hard for its implementations.

5 Conclusion

The Internet is providing the powerful foundation for the construction of the information infrastructure for genome medicine because the Internet is now stable, trustworthy and fast. The Public Key Infrastructure is adding another good feature, namely security, to the Internet. PKI is expected to be more popular through the Internet society in the near future, which will provide more sound foundation for the construction of the information infrastructure for genome medicine.

Though PKI is one of the key technologies that is critical for all of those who want to use the Internet in safety, other methods that are specific for the development of genome medicine should be devised by the research community of clinical genome informatics. Among those methods, comprehensive information models and intelligent analyzers are most essential. In this paper, we proposed an implementation of the information models based on the HL7 Version 3 and demonstrated the integrated architecture of the models and the PKI. The architecture was evaluated to work well and we consider that the architecture gives the basis for the construction of the information framework for genome medicine.