ヒト生命情報統合研究を支える ICT 活用

ICT Application to Support Integrated Research of Human Life Information

福　田　健　太

要　約　日本ユニシスでは，医療・健康に関する情報基盤構築の実績をベースにして，ヒト生命情報統合研究を支える ICT 開発に取り組んでいる．本稿では，ヒト生命情報統合研究の概要，ICT の適用ポイント及び日本ユニシスの取り組みとして「医療・健康情報を統合する基盤のアーキテクチャ」と「医療テキスト情報の標準化」について紹介する．

Abstract　Nihon Unisys is promoting the ICT development based on business results of information platform construction related to medical care and health. This paper explains the outline of integrated research of human life information, the point of applying ICT, and the approaches of Nihon Unisys which are a platform architecture to integrate medical and health information and a standardization of medical text information.

1. は　じ　め　に

2013 年 7 月，日本学術会議の提言書「100 万人ゲノムコホート研究の実施に向けて」において，「ヒト生命情報統合研究」推進の提言がなされた．ゲノムコホート研究とは，固定した集団を一定期間追跡し，体質・生活習慣・環境と疾病発生との関連について遺伝子情報を含めて解析する観察的研究である．このゲノムコホート研究を 100 万人規模に統合したものが，ヒト生命情報統合研究であり，発症前の治療的介入による予防法の確立のために，多様な生体試料とデータを蓄積して，多次元かつ膨大な情報を最新の情報科学を用いて統合解析する[1]．提言書の中で，ヒト生命情報統合研究を推進する上での多くの課題が提起されているが，それらの解決のためには，様々な ICT を適切に用いることが求められている．

日本ユニシスでは，医療情報システムの開発，佐渡島や徳島県での地域医療連携基盤構築，滋賀県長浜市におけるコホート拠点での情報基盤構築を通じて，多くの技術やノウハウを蓄積してきた実績がある．それらをベースとして，今の日本が直面する高齢化時代における健康長寿社会の実現に貢献するために，ヒト生命情報統合研究の課題を解決する技術開発に取り組んでいる．

本稿では，2 章にてヒト生命情報統合研究の概要や ICT 適用ポイント，3 章と 4 章で日本ユニシスにおける具体的な取り組み，5 章で今後の展望について報告する．

2. ヒト生命情報統合研究について

本章では，ヒト生命情報統合研究の概要，目指す方向性，そして研究を支える ICT の適用ポイントについて説明する．

2.1　ヒト生命情報統合研究の概要

超高齢社会を迎えた日本において，活力のある健康長寿社会を実現するためには，病気の早期発見と発症前の治療的介入による予防法の確立が求められる．

今までの疾患研究の多くは，細胞や動物モデルを用いた解析により行われてきた．その主な理由は，ヒトへの侵襲[*1]が最低限にとどめられ，その情報を分析することに技術的な限界があったためである．また，ヒトから採取できる試料のほとんどが，血液や皮膚などの一部の組織に限られ，そこから得られる情報が限定的であった[2]．しかし，近年，生体試料からゲノム（遺伝子），オミックス（タンパク質，代謝物など）の情報を高い精度で定量的に測定できるようになった．また，医療・健康情報や生活環境情報についても，地域医療連携やライフログ[*2]機器の普及により収集するための基盤が整備され，今後ヒト生命情報統合研究を推進する動きが加速すると予想される[2]．

ヒト生命情報統合研究は，大規模な健常者[*3]集団を長期間追跡することが可能なゲノムコホート研究拠点を構築し，個人の生体試料やゲノム，オミックス，医療，健康，生活環境に関する多様なデータの蓄積と，最先端の測定・分析技術や統計学，計算科学を駆使した統合的な解析を行うことであり（図1），そのために多種多様なデータの統合と解析を支えるための技術開発が必要となる．

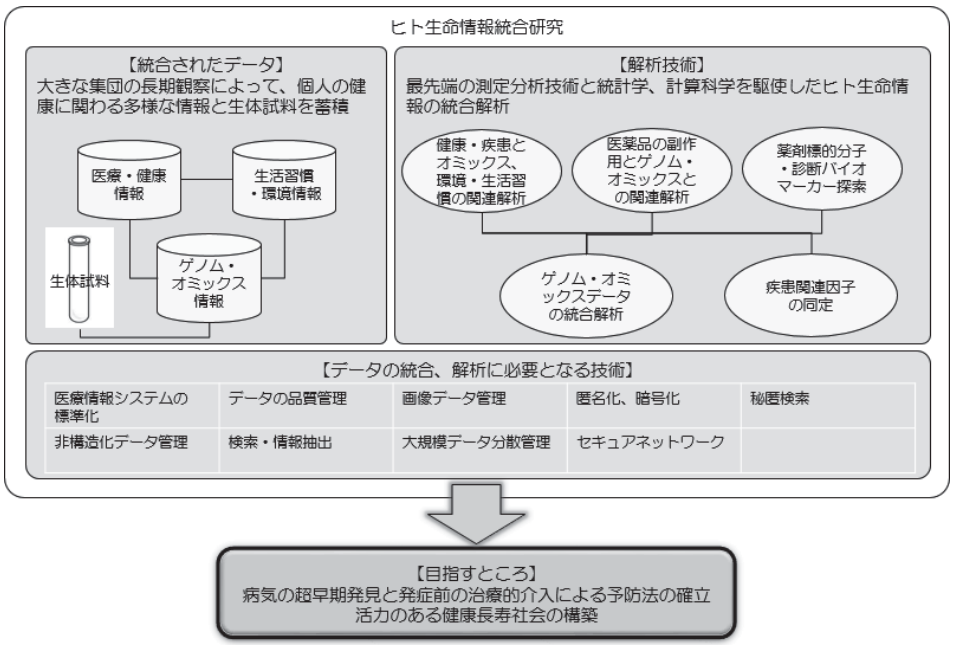

図1　ヒト生命情報統合研究の主要技術イメージ

2.2　ヒト生命情報統合研究の目指す方向性

日本においてゲノム解析を予定するコホート研究は約30件実施されている．対象者数が数百人の小規模研究から，数十万人の全国的な研究まであるが，地域を限定した数千人規模の研究が最も多い．日本における代表的な事例を図2に示す[3]．

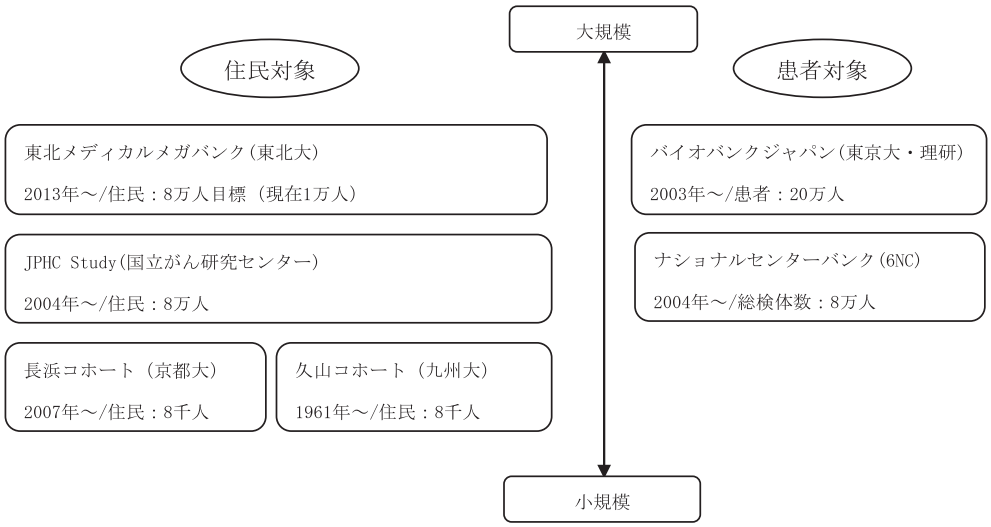

図2　日本における事例（2013年時点）

日本学術会議の提言書「100万人ゲノムコホート研究の実施に向けて」では，現在のわが国の疾患発症率をもとに多くの重要な疾患病因に迫ることが可能となる集団規模を100万人と設定している．しかし，対象者への説明や，適切に試料や情報を収集・管理する労力を考慮すると，1拠点で数十万人規模のコホートを構築することは容易ではない．また地域性や効率性，小中規模の研究拠点が多い日本の現状を考慮すると，10万人の事業参加者からなる実施拠点を10箇所程度，設置することが必要であると考えられている[1]．この各拠点から大規模ゲノムコホート研究拠点を統合的に構築するためには，多種多様で膨大な情報を網羅的かつ定期的に収集，整理，共有，統合化する基盤が必要であり，ICTを活用したインフラ（基盤）と運営・運用をうまく行うための仕組みを並行して整備していかなければならない．日本における100万人コホート拠点のイメージを図3に示す．

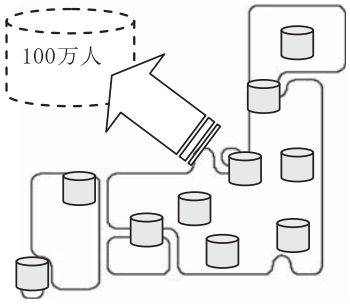

図3　日本における100万人規模の拠点構築イメージ

2.3　ヒト生命情報統合研究におけるICT適用ポイント

100万人規模のコホート拠点構築を目指す上で，統合化の対象となるデータは，生体試料から得られるゲノム・オミックス情報，罹患・投薬履歴などの医療・健康情報および生活習慣・環境情報が挙げられる．それぞれの情報について，データ源，データ例，そして統合化を目的

とした上でICTを適用するポイントについて，表1に示す．

表1　ヒト生命情報とICTを適用する意義について

	データ源	データ例	ICT適用ポイント
ゲノム・オミックス情報	生体試料 etc	・ゲノム情報 ・代謝物発現量 ・タンパク質発現量	測定・分析技術に依存する領域ではあるが，標準的技術が確立されれば，データモデル化や統合管理においてICTの適用が有効
医療・健康情報	医療機関 保険者 etc	・罹患情報 ・投薬情報 ・検査値情報 ・健診情報 ・介護情報	地域医療連携やコホート研究基盤構築の実績とノウハウをベースとした技術展開が可能であるためICTの適用が有効
生活習慣・環境情報	質問表 ライフログ センサー etc	・運動・活動量 ・食事情報 ・睡眠情報 ・地域特性情報	運動・活動・睡眠情報についてはウェアラブルデバイスを用いて高頻度かつデジタルに収集可能．食習慣などは質問表を用いた主観回答の収集にとどまっており，情報精度向上のためにICTの適用が有効

次章以降，医療・健康情報を中心とした技術開発への取り組みを説明する．3章ではテキストデータや画像データのような非構造化データを含む医療・健康情報を格納するための基盤技術について，4章では非構造化データから必要な情報を抽出し，データマイニングなどの二次利用が可能な形式へ変換する技術開発の取り組みを紹介する．

3.　医療・健康情報に関する収集蓄積基盤について

日本ユニシスでは，佐渡島における地域医療連携システム構築や京都大学における疫学研究基盤構築により培った技術やノウハウを活かして，ヒト生命情報統合研究を目的とした医療・健康情報に関する収集蓄積基盤の検討を重ねてきた．基盤要件として以下のことが挙げられる．

・大容量かつ多様な構造化されていない情報（画像，図形，テキスト）を格納可能
・新たなデータ項目追加に対応可能な高い拡張性
（解析技術や機器の進歩により新たなデータ項目が発生する可能性があるため）
・個人情報への高いセキュリティの確保
（医療情報は他の個人情報と比較して秘匿性が高いため）
・データ利活用時において，統合蓄積されたデータ群から抽出する際に高い精度で範囲検索や部分一致検索を行うための抽出機能

上記の要件を考慮した医療・健康情報統合基盤アーキテクチャを図4に示す．基盤設計におけるポイントは以下の通りである．

・個人情報等の構造化データと非構造化データ部分を切り分けて格納
・構造化データ部分は，リレーショナルデータベースに格納し，秘密分散処理機能を加えることによりセキュリティを担保
・非構造化データ部分は，データ形式の多様性や拡張性の要件に対応するためNoSQLに格納

2013年度にこのアーキテクチャをベースにしたプロトタイプシステムを作成した．2014年

度は実用化に向けたブラッシュアップを行っており，2015 年度にかけて技術実証を行う予定である．

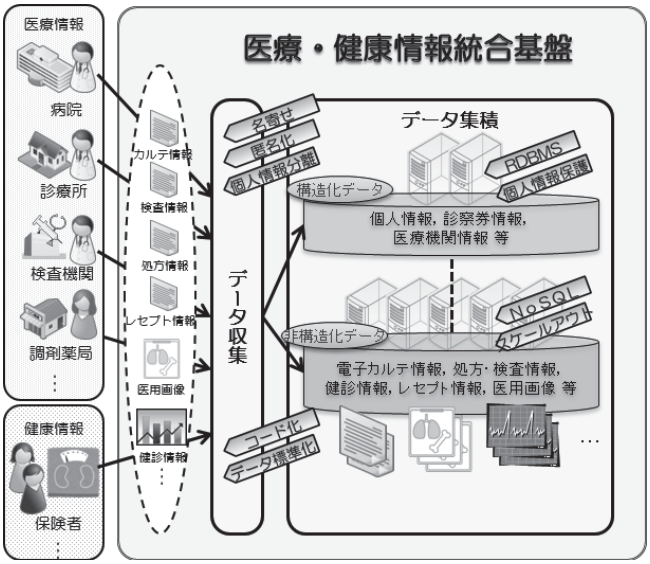

図 4　医療・健康情報統合基盤

4.　医療情報標準化技術の取り組みについて

近年，医事会計システムや電子カルテシステムの普及により，院内における医療情報のデータ化が可能となった．また，電子カルテや検査情報管理システムは，データ出力規格が統一されておらず，多施設間におけるデータ収集や統合には困難が伴うことが多かったが，標準化されたデータ形式（SS-MIX2 形式など）で入出力する機能拡張により，多くの地域間医療ネットワークが構築されてきた．

一方で，ヒト生命情報統合研究を推進する上で，電子カルテ情報に含まれている罹患情報に関連するテキスト情報，画像情報，波形情報など従来の RDBMS では取り扱うことの難しい情報を利活用する仕組みが必要である．例えば，重要な医療情報として，病名が挙げられる．病名は，医事会計システム側から取得可能なレセプト病名があるが，より精度の高い病名を取得するには，電子カルテに記載されている情報から抽出する必要がある．

本章では，非構造化データであるテキスト情報に注目して，重要な情報を抽出し，マイニング可能な形式に変換するための技術開発に関する取り組みについて紹介する．

4. 1　医療テキスト情報の標準化技術へのアプローチ

医療情報においてテキスト形式の情報を多く含んでいるのは電子カルテである．電子カルテに記載されている病名や医薬品名などの情報を標準的に抽出することが必要である．それらを支援する仕組みとして，電子カルテシステムには，医師によるカルテデータ入力時の標準化を補助するためのテンプレートが用意されているケースが多い．しかし，それらを活用した入力方法は医師の感覚になじまないことが多く，限定的な使用にとどまっている．データマイニングを見据えた標準化処理の負荷を考慮した場合，本来，データ入力と同時に標準化されること

が望ましいが，患者の状態や医師の考察が，それぞれの医師の言葉でカルテ内に記載されている．我々は，図5のようにそれらの医療テキストから情報を機械判読に適したデータ形式で抽出し，標準コードへ正規化するための言語処理アプローチを考えている[4][5]．

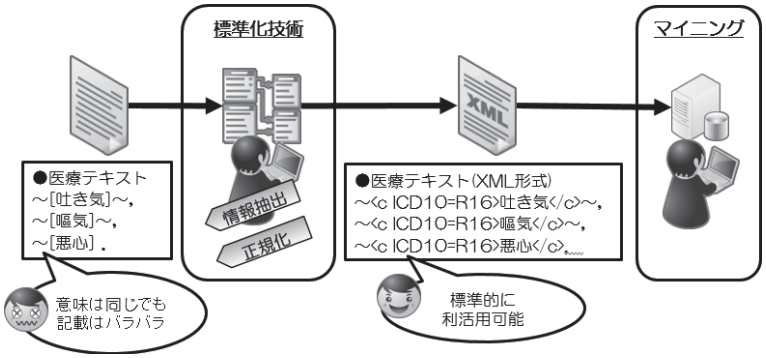

図5　医療テキスト処理のイメージ

しかし，医療テキストにおける言語処理技術研究には以下のような課題がある．

- 秘匿性の高い個人情報を多く含んでいるため，医療テキストを大学や企業の言語研究者が共有できる仕組みがない．また，病院側にはデータを解析するための人材リソースがない．
- 症状，病名や医薬品名の標準化コードをメタデータとして付与した文書が存在しない．
- 医療テキストにおける症状や病名をメタデータとして付与する統一方針がない．

そのため，他分野の言語処理と比較して進歩が遅いことが指摘されており，その課題を解決するために，2011年に京都大学や国立情報学研究所の研究者を中心として，メタデータを付与した研究利用可能な日本語医療コーパス[*4]を構築し，情報抽出に関する解析タスクを主催するコミュニティ（MedNLP）が発足した．本コミュニティにより，医療コーパスにリーチできなかった研究者や企業が，自らの解析技術を適用し，課題を共有することにより，言語処理技術のさらなる発展を目指している．

タスクに用いる医療コーパスは，医師によって書かれた擬似的な病歴要約を用意し，テキスト内に個人情報と症状などのメタデータを付与した．図6にアノテーション済みコーパス例を示す．

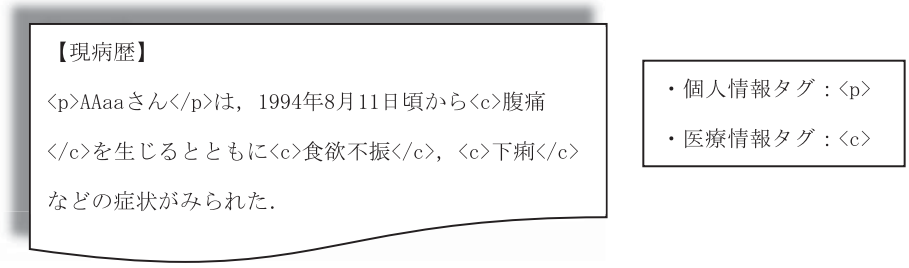

図6　アノテーション済みコーパス例

このアノテーション済コーパスや公開されている医療辞書データを活用した標準化モデル構築，モデル評価，そして変換したデータを活用したマイニングを研究のスコープとして位置づけている（図 7）.

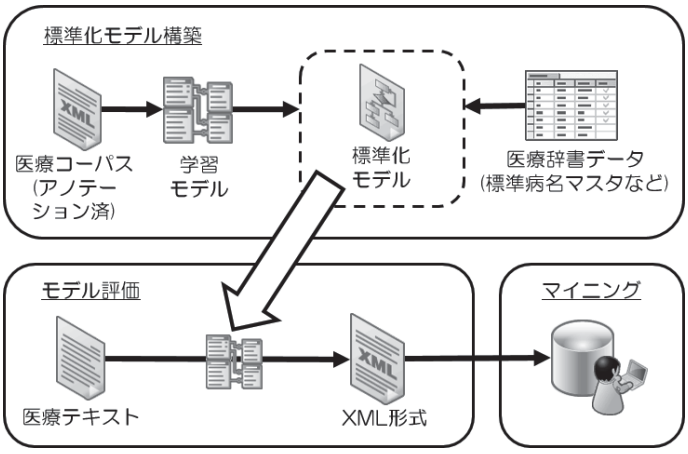

図 7　スコープイメージ

日本ユニシスは，2013 年度下期より MedNLP に参画し，医療テキストを標準的に利活用するための技術検討に取り組んできた．次節にて，その取り組みについて具体的に紹介する．

4. 2　MedNLP での取り組み

MedNLP では，模擬医療コーパスを活用した言語処理のシェアドタスクを開催している．シェアドタスクとは，複数のグループで同一の実験材料を共有し，手法を評価することである．それにより，実験材料の入手という障壁を取り払い，各研究者同士で手法ごとに解析手法の評価や議論が可能となる．医療分野の言語処理タスクは，英語圏においては，いくつか開催されているが，日本語でのシェアドタスクは，MedNLP が唯一のものである．

2013 年度は，医療コーパス内（2,244 文）とアノテーションガイドラインを公開し，個人情報（年齢，日時，病院名，場所，個人名，性別）と症状・病名を抽出するタスクを実施した．結果として，外部データを取り込んだルールベースによる抽出手法や機械学習の手法を駆使することにより，高い精度で抽出することが可能であった[6].

2014 年度は，医療コーパスを追加することで，より精度の高い抽出手法の確立や抽出した症状・病名を ICD-10 コード（疾病分類コード）[*5] に変換することを目的として，タスクを実施している．2014 年 7 月時点のステータスとして，抽出や正規化手法の技術評価を実施しており，2014 年 9 月中に評価結果の取り纏めが完了する見込みである．2015 年度は，マイニングをメインテーマとして，医療テキスト情報から抽出・変換したデータと他の医療データを組み合わせて，有益な分析ができるかをシェアドタスクとして検証する．

4. 3　標準化するための処理技術

MedNLP において，医療テキストをマイニング目的として標準化するために，図 8 のように複数段階の処理を実施してきた．

まず初めに必要となるのは，患者や関係者の個人情報を抽出し匿名化する処理である．次に，研究利用する際に重要となる用語と属性情報を抽出する処理である．具体的に重要となる情報には症状や病名などがあり，属性情報としては発症有無が挙げられる．最後に，抽出した用語を標準コードや数値などへ正規化する処理である．特に症状や病名は，医師による自由入力の場合，表記が統一されていないため，標準的な病名コードに正規化することで表記ゆれを統一する必要がある．そして，これらの処理を行うことにより，初めて医療テキストから患者の罹患情報を標準的に抽出することができる．

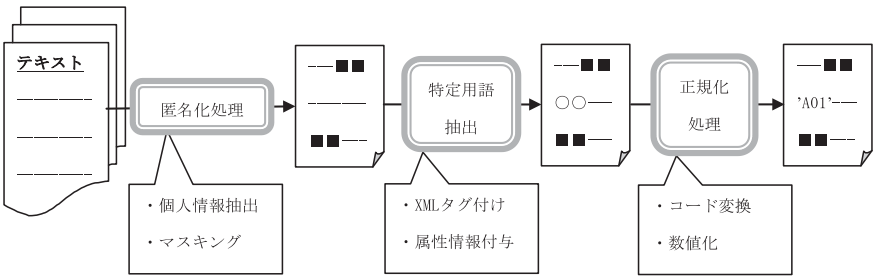

図 8　医療テキストの標準化処理イメージ

2014 年度は，特定用語抽出・正規化処理技術について継続的に検証していくが，将来的には 3 章で述べた医療・健康情報統合基盤の機能として実装することを検討している．医療テキストのような非構造化データを利活用可能な形式で統合することにより，精度が高く，かつ有効なデータマイニングが可能になる．

また，具体的な利活用シーンとして，ゲノムコホート拠点において，参加者の医療・健康に関するテキスト情報から確定病名や症状など研究に必要な情報を抽出・正規化する部分に適用することにより，高い精度の網羅的な解析や研究の推進が期待される．

5.　今後の展望

今後の展望として，医療テキスト情報の標準化技術の適用先について述べる．

2000 年のヒトゲノム全配列の解読以降，30 億塩基対からなる遺伝暗号がどのような意味を持っているか解析が進められてきた．これらの解析を進めていく中で，個人間の塩基配列は殆ど同じであるが，数百箇所のうち一箇所ほど差異があることが示された．この差異のある箇所を一塩基多型（SNP）といい，全遺伝配列の中で 1000 万箇所ほど存在し，病気のなりやすさや薬効の個人差に関与していると考えられた．そして，数十万箇所の SNP を説明変数，疾患発症を目的変数として網羅的に解析する手法として，ゲノムワイド関連研究（GWAS）が推進され，数百件もの表現型に関連した SNPs の同定で 1300 件以上の報告があげられた．

一方で，GWAS の課題として，数十万もの説明変数から統計的な有意差を示すことが困難であるケースも多く，それを補完するためにフェノムワイド関連研究（PheWAS）の必要性が提言されている[7]．PheWAS は，GWAS と全く逆の解析手順であり，表現型を説明変数として，遺伝型との関連を明らかにする研究である．遺伝型は数百万種類であるのに対して，表現型は数千種類であるため，統計的に有意差を検出しやすい．GWAS と PheWAS の研究イメージについて図 9 に示す．

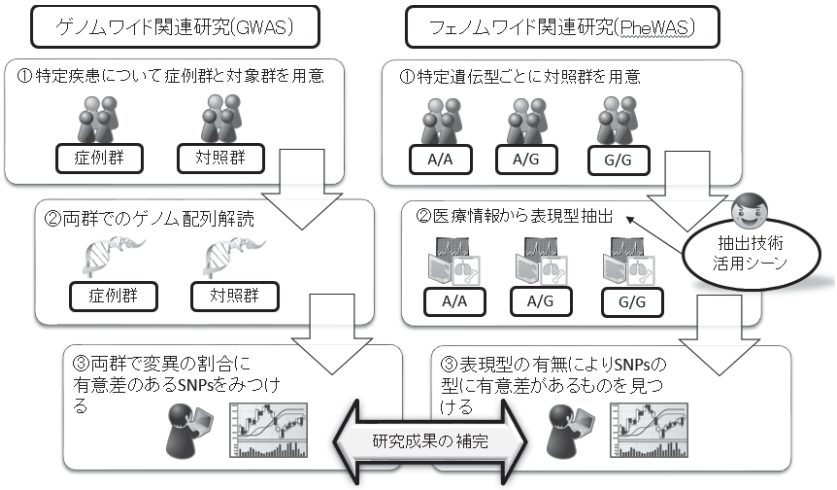

図 9　GWAS と PheWAS の研究イメージ

PheWAS を推進する上で必要となるのが，医療情報に含まれる情報から表現型となる症状や病名を抽出する技術である．情報抽出の観点で考えると，解析機器より抽出される遺伝型情報と比較して，表現型情報の抽出コストは大きい．アメリカの研究事例として，電子カルテの情報から表現型を ICD-9 コードとして自動的に出力する技術開発が行われている[7]．今後，日本でも，日本語を含む多言語の医療情報から表現型を抽出するための技術開発ニーズがよりいっそう高まることが予想される．MedNLP における研究開発を通じて，この分野における技術適用に貢献していきたい．

6. お わ り に

ヒト生命情報統合研究を支える技術として「医療・健康情報統合基盤のアーキテクチャ」と「医療テキスト情報の標準化」について紹介したが，統合化された多種多様なデータを利活用する上で特に重要となるのは，非構造化データの情報抽出や正規化，検索，及び管理技術である．それらの情報と構造化情報を組み合わせることにより，情報網羅性の高い解析ができるようになる．今後，多様な非構造化データを有効に取り扱うために，テキストマイニング，自然言語処理，機械学習などの技術を取り入れた仕組みを開発し，医療・健康情報統合基盤の機能として組み入れていきたい．

最後に，本稿執筆にあたり調査・研究活動にご協力・ご指導いただいた皆様に深く御礼申し上げます．

＊ 1　医学において生体の内部環境の恒常性を乱す可能性がある刺激全般のことである．投薬・注射・手術などの医療行為や外傷・骨折・感染症などが含まれる．
＊ 2　ヒトの活動を映像，音声，位置情報などの経時的データとして記録することである．
＊ 3　心身に病気や障害のない健康な人のことである．
＊ 4　言語学において，自然言語処理研究に用いるために，文章を構造化し大規模に集積したものを指す．一般的に構造化では言語的なメタ情報（品詞，統語構造など）が付与される．
＊ 5　WHO による病名分類のためのコードであり，アルファベット 1 桁と数字 3 桁からなる．

参考文献
［1］ 日本学術会議提言書，「100万人ゲノムコホート研究の実施に向けて」
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t176-1.pdf
［2］ 日本学術会議提言書，「ヒト生命情報統合研究の拠点構築」
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t155-1.pdf
［3］「個別化医療・創薬のための大規模ゲノムコホート研究の最新動向」，株式会社シード・プランニング，シーエムシー出版，2014年4月
［4］ 森田瑞樹，狩野芳伸，大熊智子，宮部真衣，荒牧英治，「医療分野の言語処理研究の環境整備に向けて」，言語処理学会，2013年3月，
http://www.anlp.jp/proceedings/annual_meeting/2013/pdf_dir/Y1-2.pdf
［5］ 森田瑞樹，狩野芳伸，大熊智子，荒牧英治「NTCIR MedNLP-2：医療分野の言語処理」，動向情報の要約と可視化に関するワークショップ（MuST），2013年10月，
http://must.c.u-tokyo.ac.jp/sigam/sigam05/sigam0512.pdf
［6］ 荒牧英治，大熊智子，「NTCIR MedNLP：本邦初の医療分野の言語処理コンテスト」，第33回医療情報学連合大会，2013年10月
［7］ Scott J. Hebbring，「The challenges, advantages and future of phenome-wide association studies」，British Society for immunology，2013年10月

参考文献のURLは2014年7月25日現在での存在を確認

執筆者紹介 福 田 健 太（Kenta Fukuda）
2006年日本ユニシス（株）入社．SEとして金融システムの構築を経験した後，社内公募制度を利用して2012年に総合技術研究所に異動．以来，研究員として医療・ヘルスケア分野の新たな情報基盤構築に向けた技術や新サービスの開発に取り組んでいる．