電子カルテ情報を用いた証拠性のある臨床研究手法に関する研究
(H27-医療-指定-016)
－研究用計測機器ユーザー認証等証拠性保全検討－
次世代シーケンサー等のデータソースおよび解析ソフトウエアの検討

研究分担者　澤　智博
帝京大学医療情報システム研究センター　教授

研究要旨: 医学研究における各種データ、特に次世代シーケンサーから出力されるデータ、解析ソフトウエア、ゲノムデータベースについて、証拠性保全を検討するための調査を実施した。次世代シーケンサー機器類、解析ソフトウエア類、ゲノムデータベースについて証拠性保全の視点でプロパティや機能を精査し、証拠性保全のために必要な要素について検討した。

A.研究目的

医学研究における各種データについて、電子カルテの証拠性保全の技術・運用の適用可能性を検討するため、次世代シーケンサーから出力されるデータ、解析ソフトウエア、データベースについて調査し、そのプロパティや機能について整理する。

B.研究方法

次世代シーケンサー機器、解析ソフトウエア、ゲノムデータベースに関するドキュメント類を精査する。

・次世代シーケンサー
- Illumina 社
  HiSeq, MiSeq
- Thermo Fisher Scientific (Life Technologies) 社
  Ion PGM, Ion Proton
- Pacific Biosciences 社
  PACBIO RS II

・次世代シーケンサーデータ種
BAM, TAB, ACE, FASTA, WIG, BED, VCF, MAF, GFF, CSV, TSV

・解析ソフトウエア
- Globus genomics
- Partek
- DNAnexus
- CLC Bio
- DNASTAR
- Maverix Biomics
- Seven Bridges
- Golden Helix

・ゲノムデータベース
- NCBI (National Center for Biotechnology Information)
- EMBL-EBI (European Bioinformatics Institute)
- DDBJ (DNA Data Bank of Japan)

倫理的配慮

平成 27 年度の分担研究においては、特に個人情報を取扱うなどの倫理的な課題は発生しなかった。

C.研究結果

・次世代シーケンサー

次世代シーケンサー（ＮＧＳ）は検体の分析器とそれを制御するコンピュータ（ハードウエアおよびソフトウエア）から構成される。ユーザー認証等の分析器に対する操作は制御コンピュータにて実施される。**図1**に NGS のソフトウエアコンポーネントを示した。

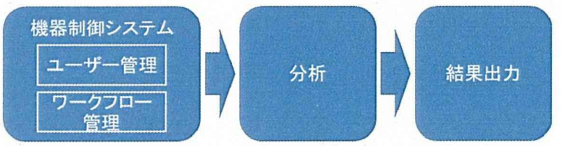

図1：NGS のソフトウエアコンポーネント

ユーザー認証等のユーザー管理、アクセス管理は図１における制御システム内のユーザー管理コンポーネントで行われる。

ユーザー管理コンポーネントでは、

・ユーザー名

・パスワード

・ユーザー種（アクセス権限）

が管理される。

・次世代シーケンサーデータ種

データファイルは、主にテキストファイルおよびバイナリファイルに分類される。代表的なファイル形式としては以下がある。

BAM: リードプレイスメントの配列アライメント/マップ形式のバイナリファイル

FASTA: シーケンスデータを記述したテキストファイル

・解析ソフトウエア

解析プロセスは一般的に以下の要素からなる。

- シーケンスファイルのインポート
- データの品質管理
- シーケンスリードのトリミング
- シーケンスデータのアライメント
- 重複リードの除去
- データ品質の再調整
- バリアントコントロール
- アノテーション付与

・ゲノムデータベース

ゲノムデータベースについて、ユーザー等のデータ作成や分析の実施に関する情報はメタデータオブジェクトに格納される。メタデータは XML 形式で記述される。

メタデータは以下の項目から構成される。

- Submission
- Study (BioProject)
- Sample (BioSample)
- Experiment
- Run
- Analysis

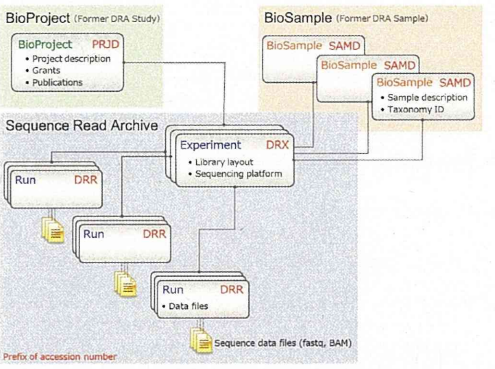

図2：メタデータモデルの例（DDBJ より）

- データバリデーション
データベースにデータを登録する際に、登録するデータが意図した内容で登録されているかを検証する方法としてＭＤ５ハッシュ関数が利用される。

D.考察

・次世代シーケンサーおよびシーケンスデータ

ＮＧＳにおけるユーザー認証等は、制御システムのソフトウエアにより管理される。ユーザーはパスワード管理され操作・閲覧などのアクセス権限が管理される。一方で、シーケンスデータは解析結果のみを含むファイルとなっており、ファイル内にはユーザー等に関する情報は含まれない。データ生成に関連する情報は、メタデータとしてデータと関連付けて管理される。従って、証拠性を担保するためにはデータファイルとメタデータとが不可分な状態でデータが取り扱われる必要があると考えられる。

・解析ソフトウエア

ＮＧＳからの出力ファイルは、そのままの状態では結果を解釈することはできず、解析ソフトウエアを使用して目的に応じたデータ処理が実施される。このとき、シーケンスリードのトリミングに代表されるようにデータを加工することは不可欠であり、意図せぬデータ加工ミス、データの改ざん、解析プロセスとして正当なデータ加工を区別する必要があると考えられる。

・ゲノムデータベース

メタデータオブジェクトにはデータ生成や分析に携わったユーザーの情報を登録することが可能である。またデータファイルのバリデーション手段としてＭＤ５ハッシュ関数が活用される。一方で、データ登録の内容は登録者に任されておりここに意図せぬ登録ミスや不正なデータの登録の可能性があると考えられる。

E. 結論

次世代シーケンサー、データ、解析ソフトウエア、ゲノムデータベースについて調査した。ユーザー認証をはじめ、証拠性保全のために必要な要素について検討した。

F. 健康危険情報

平成 27 年度の本研究においては、生命、健康に重大な影響を及ぼすと考えられる新たな問題、情報は取り扱わなかった。

G. 研究発表

1. 論文発表

澤智博:周術期医療におけるビッグデータ活用とデータサイエンス，麻酔，64 増刊. S104-S112，2015.

澤智博:コンピュータはどこまで”医師”に近づいたか？，The Next Technology，日経ＢＰ社，146-151,2015. ISBN978-4-8222-7975-2.

澤智博:ＨＩＳ－既存システムの考察と今後あるべき姿を考える，月刊新医療. 42(11)67-70，2015.

澤智博.:人工知能時代を前に医師が考えるべきことは？大阪府保険医雑

誌, 2016(2)20-25, 2016.

2. 学会発表

澤智博: 周術期の臨床効果データベースと偶発症例調査事業, 第 35 回医療情報学連合大会, 医療情報学, 第 35 回医療情報学連合大会論文集, 35-Suppl., 88-89、11 月 3 日, 2015. 宜野湾市

澤智博:周術期医療におけるビッグデータ活用とデータサイエンス, 日本麻酔科学会第 62 回学術集会, 5 月 29 日, 2015. 神戸市

H. 知的財産権の出願・登録状況（予定も含む）

1.特許取得　　なし
2.実用新案登録　なし
3.その他　　なし

電子カルテ情報を用いた証拠性のある臨床研究手法に関する研究
(H27-医療-指定-016)
－測定、画像データ証拠性・安全運用環境検討－
遺伝子解析研究への証拠性の導入

研究分担者　作佐部　太也
藤田保健衛生大学医療科学部臨床工学科　　准教授

研究要旨：　目的：遺伝子解析研究における証拠性を確保する上でのワークフロー上のチェックポイントを見いだし、適切な方法を提案することを目的とする。**方法**：次世代シーケンサーを運用している研究現場においてワークフローについての聞き取り調査を行う。**結果**：ワークフローの概要は判明したが手作業が多く体系化したものではなく、情報システムによる支援が必要なことが明らかになった。**結論**：証拠性確保のためのチェックポイントを見出すためには、ワークフロー運用を支援する情報システムを開発することでワークフローの体系化を行う必要があることがわかった。

A.研究目的

今日において遺伝子解析研究における次世代シーケンサー（ＮＧＳ）は、基礎研究を超えて臨床研究のツールとなりつつある。しかし、基礎研究のツールとしての十分な運用経験の蓄積による成熟期間を経ていないため、データの管理方法などにおいて未成熟な状態であることが想定される。このことを別の視点からとらえると、証拠性に留意したデータ管理方法を広く普及させるには、むしろ望ましいタイミングであると考えられる。本分担研究において、医療機関の付属研究施設における実際のＮＧＳの運用やデータ解析処理の状況について調査分析し、証拠性確保のためのワークフロー上のチェックポイントを見いだし、また、適切な方法を提案することを目的とする。

B.研究方法

分担研究者が所属する藤田保健衛生大学に付属する総合医科学研究所の遺伝子解析関連の研究部門におけるＮＧＳの運用状況について聞き取りなどの調査を行い遺伝子解析の基本的なワークフローのパターンを明らかにする。次にワークフローを支援するようなデータ管理システムの開発し、研究所だけでなく大学病院の臨床研究者に利用してもらうことで、より一般的なワークフローのパターンを明らかにする。その上で、過誤によるデータに対する誤った操作が発生する可能性や悪意に基づいた改竄などの入る余地を分析し、適切なチェックポイントおよび証拠性を確保する方式を検討する。

倫理的配慮

平成27年度の分担研究においては、特に個人情報を取扱うなどの倫理的な課題は発

生しなかった。

C.研究結果

聞き取り調査によりデータ管理の体制などの情報が得られた。

まず人的体制については、計算機技術を独学で学んだ若手の生物系研究員が一人でデータ管理および基本的な解析処理を行っていた。このような状況は大規模な研究機関以外では一般的な状況であることもわかった。データの解析処理についてはＮＧＳから出力されるデータに対してサーバ規模の計算機上で各種処理コマンドを研究員が手動で動作させていた。

解析結果のデータを用いて複数の研究者が探索的な研究を行っているが、研究プロジェクトのワークフローの管理は、表計算ソフトウェア、SNS、電子メールなどのプリミティブな情報システムを組み合わせて行っていた。データ管理については、全てのデータはファイル形態でＮＡＳに蓄積され、研究員が開発した簡易な Web アプリケーションで管理を行っていた。

データ処理のワークフローや研究プロジェクトのワークフローの管理についてはほとんどが手作業となっており、情報システム化が望まれていることがわかった。このような状況は一般的なものと考えられ、実際、ＮＧＳを開発・販売しているメーカや大手の研究 機関などが解析ワークフローおよびデータの管理を行うクラウド・サービス を提供している。しかし、患者のデータを扱う臨床研究においては、患者からの承諾の取得から研究倫理審査の社会的プロセスの整備だけでなく、そもそも学外へのヒトゲノム情報の転送の可否についての運営的な方針の整備が必要となる。実際、本学においてはそのような整備は十分ではないため、クラウド・サービスを利用することはできず、他の医療機関においても同様であると推測される。

また、検査センターなどのように定型業務としてＮＧＳをランさせているわけではなく、その後の解析処理を含め、研究プロジェクトなどに応じてプロセスをその都度変更するため、自動化やアウトソーシング自体が難しい。

一方、情報処理技術的な観点からみると、データ量についてはＮＧＳから出力となる FASTQ 形式のファイルが 1GB 程度、その後の解析マップ処理などの出力となる BAM 形式のファイルが 5GB 程度。リファレンスに対する変異の抽出した結果の VCF 形式のファイルが数 100MB 程度と比較的サイズの大きいファイルが扱われてはいるが、発生するデータ数は多くはなく、今日の MDCT による放射線診断画像と比べれば処理が困難になるほどの量ではない。

D.考察

聞き取り前の想定通り、データ管理の方式は洗練されたものではなく改善の余地があった。何より手動でコマンドを実行することが多く、動作の記録や再現性の確保などに不安があった。よって、解析処理のワークフローを支援する情報システムが必要と考えられる。

一方、研究プロジェクトの管理として、対象患者の選択から検体の採取、管理、

数理的な解析や探索的な分析、研究の指揮や管理などのワークフローについても未整備な状況であり、やはり研究のワークフローを支援する情報システムが必要と考えられる。

今後は聞き取り調査の段階から協同での作業の段階へと作業を進め、ワークフローの詳細な洗い出しと、そのワークフローの運用を支援するプロトタイプ・システムの開発を行う。情報システムとしては二種類必要となり、一つはＮＧＳの出力から臨床研究者が探索的分析を行うことができるデータに処理する計算機上でのプロセスを支援するシステム、もう一つはより大きな視点から研究プロジェクトの中での探索的分析のワークフローをデータ中心に管理するシステムが考えられる。

このような半自動化したシステムを開発し運用することで、詳細な動作記録を自動的に採取することができるため、証拠性確保のためのチェックポイントを見出すための各種の情報が得られることが期待される。なお、開発するワークフロー支援システムについては諸事情が許せばオープンソースとして配布することを計画しており、他の研究機関で試用してもらうことで、より本学固有の事情に左右されない、一般的な情報が収集できることも期待される。また、ワークフロー支援システムがＮＧＳを用いた臨床研究の効率化、活性化に寄与することも期待される。

E. 結論

実際の研究機関におけるＮＧＳおよびデータの運用状況について調査より、改善の余地があることが明らかにった。

証拠性の確保の前段階としてデータ処理や研究プロジェクト運営についてのワークフローの詳細な分析と、ワークフローの運用を支援する情報システムの導入が必要であることが明かになった。

F. 健康危険情報

平成 27 年度の本研究においては、生命、健康に重大な影響を及ぼすと考えられる新たな問題、情報は取り扱わなかった。

G. 研究発表

1. 論文発表

なし

2. 学会発表

なし

H. 知的財産権の出願・登録状況（予定も含む）

1. 特許取得　　なし
2. 実用新案登録　　なし
3. その他　　なし

電子カルテ情報を用いた証拠性のある臨床研究手法に関する研究
(H27-医療-指定-016)
－研究用計測機器ユーザ認証等証拠性保全検討－

研究分担者　渡辺　浩
国立長寿医療研究センター 治験・臨床研究推進センター 室長

研究要旨:国立長寿医療研究センターの治験臨床研究推進センターの医療情報室では研究者の研究支援基盤構築プロジェクトとして「電子カルテ領域の臨床データを匿名化モジュールを介して施設基盤レベルでデータを匿名化する仕組み」を試験的に導入している。今回の木村班の研究に関しても有用なソリューションであるためこれを報告する。

A.研究目的

今回の班研究では、臨床研究用のデータを如何に研究者に安全に活用させるかコンセプトになっている

B.研究方法

長寿医療研究センターは、病院と研究所から成り立っており、病院情報システムはクローズドなネットワーク、研究は外部との接続も可能な情報系のネットワークとして独立して存在している。2012 年にバイオバンクシステムが導入された際には匿名化されつつ連結可能なデータソースのバンキングが必要になった。バイオバンクは研究機関内の部門であり、検体を管理する検体管理モジュールは、研究棟内の情報系ネットワークに存在しており、一方病院から送られた検体を処理するには電子カルテ上の患者属性、病名、検査結果、処方内容など、各種情報をサンプルのカタログとして必要であったが、直接アクセスをさせることはポリシー上相応しくないと考えた。そこで、病院系と情報系ネットワークの間にいずれにもアクセス可能な特殊な DMZ ネットワークを置き、ここに患者 ID とバイオバンク協力者 ID の変換や番号照合を行う匿名化システムを導入した。同時にこの機能を匿名化モジュールとして、本幹ステムから分離独立させた。

倫理的配慮

今回の分担研究上、特に個人情報を取扱うなどの倫理的な課題は発生しなかった。

C.研究結果

今回の仕組みを導入することにより機微な情報を含む本モジュールのセキュリティをより強固な物にすることが可能になった。また、センター内の他システムで匿名化処理が必要な場合、本モジュールを流用することによりセンター内の匿名処理を一元管理できるようになった。実際の動きとして、当センターでは電子カルテシステム導入時より SS-MIX 標準化ストレージが導入されていたが、2014 年に治験臨

床研究推進センター医療情報室の研究者支援基盤構築プロジェクトとして　病院ネットワークにある標準化ストレージを匿名化モジュールを利用して、ストレージ全体の匿名化処理をし「匿名化ストレージ」として情報系ネットワーク内に導入した。（**図 1** 参照）

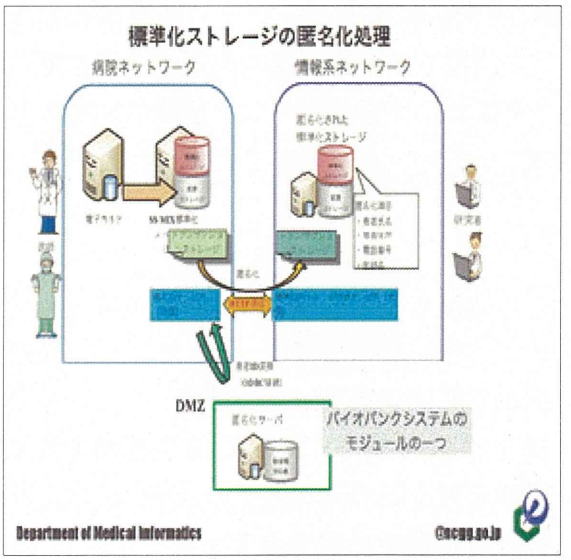

図 1
匿名化モジュール概念図

D.考察

本システムは、現在試用を始めたばかりではあるが、今後、不容易な電子カルテアクセスが減ることや臨床情報を元にした研究が活性化する事が期待される。

E. 結論

電子カルテデータの匿名化処理をインフラレベルで構築することには臨床研究推進の支援として期待が持てる。

F. 健康危険情報

今回の分担研究では、生命、健康に重大な影響を及ぼすと考えられる新たな問題、情報は取り扱わなかった。

G. 研究発表

1. 論文発表

なし

2. 学会発表

渡辺浩：モジュール単位開発のメリットを活かした研究者支援基盤システム構築の報告，第 35 回医療情報学連合大会，医療情報学，第 35 回医療情報学連合大会論文集，35-Suppl.，418-419，11 月 2 日，2015．宜野湾市

H．知的財産権の出願・登録状況

1. 特許取得　　なし
2. 実用新案登録　　なし
3. その他　　なし

Ⅲ.研究成果の刊行に関する一覧表

研究成果の刊行に関する一覧表

雑　誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
澤智博	周術期医療におけるビッグデータ活用とデータサイエンス	麻酔	64増刊	S104-S112	2015
澤智博	コンピュータはどこまで"医師"に近づいたか？	The Next Technology.	ISBN978-4-8222-7975-2	146-151	2015
澤智博	HIS－既存システムの考察と今後あるべき姿を考える	月刊新医療	42(11)	67-70	2015
澤智博	人工知能時代を前に医師が考えるべきことは？	大阪府保険医雑誌	2016(2)	20-25	2016

Ⅳ.研究成果の刊行物・別刷

【論文発表】

1. 澤　智博:

周術期医療におけるビッグデータ活用とデータサイエンス, 麻酔, 64 増刊. S104-S112, 2015.

周術期医療におけるビッグデータ活用とデータサイエンス

澤　智博

麻　　酔
第 64 巻 増 刊 別 刷
克誠堂出版株式会社

Jr, Henry MC, Parnia S. Cerebral oximetry levels during CPR are associated with return of spontaneous circulation following cardiac arrest : an observational study. Emerg Med J 2015 ; 32 : 353-6.
56) Neuloh G, Schramm J. Monitoring of motor evoked potentials compared with somatosensory evoked potentials and microvascular Doppler ultrasonography in cerebral aneurysm surgery. J Neurosurg 2004 ; 100 : 389-99.
57) Allen JS, Tranel D, Bruss J, Damasio H. Correlations between regional brain volumes and memory performance in anoxia. J Clin Exp Neuropsychol 2006 ; 28 : 457-76.
58) Heckmann JG, Lang CJ, Pfau M, Neundörfer B. Electrocerebral silence with preserved but reduced cortical brain perfusion. Eur J Emerg Med 2003 ; 10 : 241-3.
59) Ajisaka H. Early electroencephalographic findings in patients with anoxic encephalopathy after cardiopulmonary arrest and successful resusitation. J Clin Neurosci 2004 ; 11 : 616-8.
60) Doppenberg EM, Zauner A, Bullock R, Ward JD, Fatouros PP, Young HF. Correlations between brain tissue oxygen tension, carbon dioxide tension, pH, and cerebral blood flow—a better way of monitoring the severely injured brain? Surg Neurol 1998 ; 49 : 650-4.
61) Ungerstedt U, Rostami E. Microdialysis in neurointensive care. Curr Pharm Des 2004 ; 10 : 2145-52.
62) Pleines UE, Morganti-Kossmann MC, Rancan M, Joller H, Trentz O, Kossmann T, et al. S-100 beta reflects the extent of injury and outcome, whereas neuronal specific enolase is a better indicator of neuroinflammation in patients with severe traumatic brain injury. J Neurotrauma 2001 ; 18 : 491-8.

招請講演

周術期医療におけるビッグデータ活用とデータサイエンス

澤　智博*

キーワード：周術期医療，ビッグデータ，データサイエンス

はじめに

本稿では，“ビッグデータ”や“データサイエンス”というキーワードを通して，麻酔科医とは何者であるのか，麻酔科学とはどのような領域であるのかを考えていきたい。

1 ビッグデータ？　データサイエンス？　データサイエンティスト？

“ビッグデータ”という語は，2010 年ごろからよく使用されるようになったと考えられているが，わが国で“ビッグデータ”が注目されるようになった契機の一つは2011 年3月の東日本大震災である。震災時の人々の動きを大量のデータを使用して，空間上に経時的に可視化することができ“ビッグデータ”は注目を集めることとなった。その後も“ビッグデータ”は国内，そして世界的にもホットトピックスとして扱われている。“ビッグデータ”に対する考え方のポイントは，データの大きさが“ビッグ”であるかどうかという視点においてより，ビッグデータの持つ性質やそれを扱うのに必要な技術の開発に注目が集まっているところにある。

“データサイエンス”は，コンピュータの黎明期である 1960 年代にはすでに存在していたといわれている。したがって，“データサイエンス”はコンピュータとともに発展してきたと考えられる。“データサイエンス”を端的に表現すると，データを，統計学的に，数学的に，プログラミングによって処理することで“data products”，つまりデータの形をした価値ある産物を生み出すことにある。“data products”がもっとも活用されている分野の一つは広告であり，“The best minds of my generation are thinking about how to make people click ads.”といわれているほどである[1]。

データサイエンスとともに注目を集めることになったのは“データサイエンティスト”である。特に，2012 年 10 月のハーバードビジネスレビュー誌で“Data Scientist：The Sexiest Job of the 21st Century”[2]という記事がもとになり，世界の注目が高まった。データを駆使して現状を把握し，企業や組織の方向性を指し示す。データサイエンティストは，現代においてもっとも必要とされる職業である，という雰囲気が高まった。わが国も例外ではなく，いや，むしろ，かなり熱を帯びてデータサイエンティストを“外部”から採用する必要性が論じられ，データサイエンスに近い領域にいた統計学者やコンピュータサイエンティストは我先にデータサイエンティストを名乗り始めたのである。

2 データサイエンティストは期待外れ？

ところが，ビッグデータやデータサイエンスへの期待が変わらず高い状態であるのに対して，データサイエンティストという語への注目は低下してきている。数字やコンピュータだけを扱ってきた人がある日突然にこれまで働いたこともない領域にヘッドハンティングされても役に立たな

* 帝京大学医学部麻酔科学講座・帝京大学医療情報システム研究センター

麻酔 2015；64 増刊：S 104—S 112.

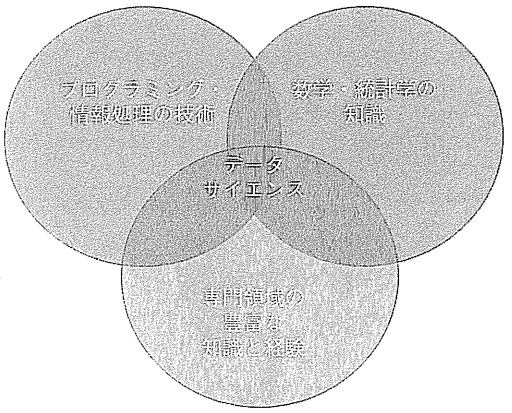

図1　データサイエンスの3要素
〔Drew Conwayのベン図より改変引用（Schutt R, O'Neil C. Doing data science. O'Reilly Media；1版（2013/10/9）〕

かったのであろうか。そして，数字だけをこね回して現場の実感と乖離した予測を提示しても現場にはピンとこなかったということであろうか。むろん，期待どおりのパフォーマンスを発揮するデータサイエンティストは存在するのだが，そうではないデータサイエンティストはなぜ発生してしまったのであろうか。図1は，データサイエンスに必要な3要素，つまり，プログラミング・情報処理技術，数学・統計学の知識，専門領域の豊富な知識と経験の関係について図示したものである[1]。ここで注目すべきは，専門領域の豊富な知識と経験が必要とされる点である。専門領域によっては，プログラミング技術の習得や数学・統計学の獲得よりも，その領域での知識と経験を培うことのほうがはるかに大変なことがある。それでは，医学はどうであろうか。麻酔科学はいかがであろうか。プログラミング技術と数学・統計学の知識のあるデータサイエンティストがやってきて，ただちに活躍してくれそうだろうか。われわれ麻酔科医が抱えている課題を解決してくれるであろうか。

3 データ，テクノロジー，サイエンス

ここでは，データ，テクノロジー，サイエンスの3つの視点から，周術期医療におけるデータをどのように取り扱うことができるかを考えていきたい。“データ”の項では，医療を取り巻くデータの変化について，ビッグデータにも触れながらデータの量や性質を中心に概説する。“テクノロジー”の項では，ビッグデータをはじめとする新たな種類のデータを扱う技術の紹介や，データベース解析に必要な考え方について解説する。“サイエンス”の項では，テクノロジーで紹介した市販のハードウエアやソフトウエアとして購入できる枠組みを超えて，コンピュータサイエンスやインフォマティクスといった科学分野において研究されている知見について紹介する。

1）データ

医療に関するデータには，検体検査の結果のような“数値”であったり，心電図に代表される“波形”，放射線画像のような“画像”，さらに画像は，心エコーや内視鏡のような“カラー動画”，あるいは自動麻酔記録のような毎分・毎秒に生成されるバイタル測定値の時系列データもあり，非常に多種多様であることが特徴である。

データ自体は，人類が文字を扱うようになって以来，またはコンピュータが符号を処理できるようになって以来，存在している。ここ数年で変化してきたのは，データの量が増加してきたことと，種類が多様になってきたことである。ビッグデータを論ずる際の“ビッグ”について，データの量として2020年ごろには35-40ゼタバイト（ZB）になるといわれている。ゼタバイトとは，われわれが通常使用するハードディスクの容量単位であるテラバイト（TB）の1,000倍（ペタバイト：PB）の1,000倍（エクサバイト：EB）の1,000倍である。また，ビッグデータを特徴づける4つのVが提示されている。そのVとは，volume（量），variety（種類），velocity（速度），value（価値）である。これまでに扱うことのなかったデータ量と，これまでに扱うことのなかったデータ種を，これまでより短時間でデータ処理できるようになると，これまでになかった価値を生み出せる可能性があると解釈できる[3]。

医療データにおいて量の増加という視点で注目されるのは，カラー動画である。従来から存在する電子カルテシステムには，検体検査結果に代表

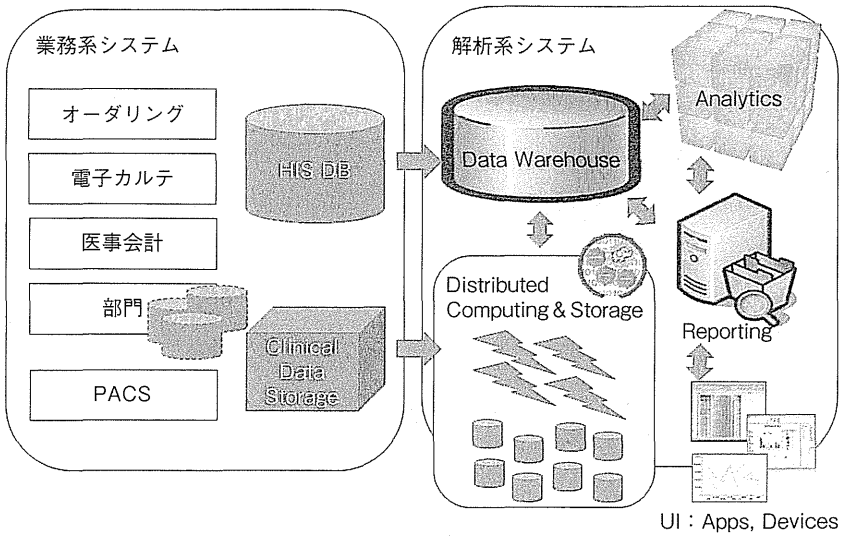

図 2　データ処理に適した病院情報システムのアーキテクチャ例

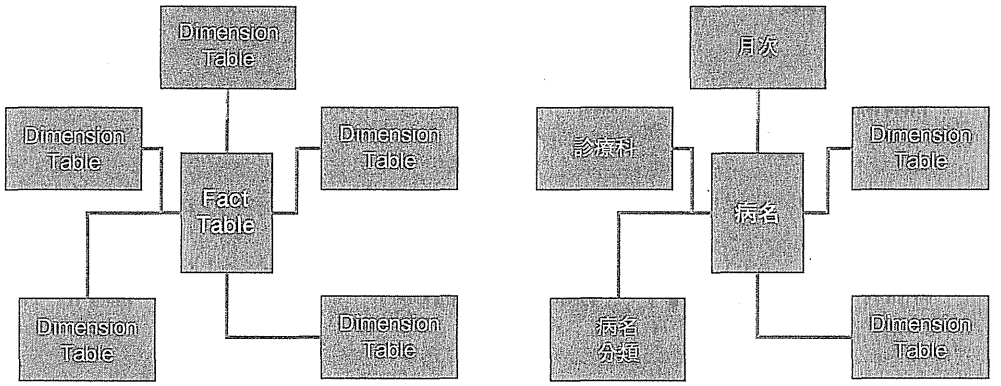

図 3　スタースキーマの例

左はスタースキーマの一般的な構造を示す。解析対象となる fact table を中心にして複数の dimension table を関連づけ多次元のデータモデルを表現する。右は，病名データを解析する際のスタースキーマの例である。病名テーブルを中心に，月次，診療科，病名分類といった切り口でデータ解析を可能にする。

される文字や数値のデータが保存されてきた。また，放射線画像の多くは静止画として PACS（picture archiving and communication systems）に保存されてきた。近年は，循環器系検査などに伴うカラーエコー画像や手術映像といった種類の動画が出現し，病院情報システムにおいて他のデータ種と同様に系統的に管理する際に，データ量が課題となってきている。

医療におけるデータの種類の多様化という視点では，ゲノム関連のデータが挙げられる[4]。これまでの医療では，各種検査によって臓器や組織のレベルでの異常を検出してきた。DNA 配列の測定や遺伝子発現の計測が安価に高速に実施できるようになり，臨床においてもこれらのデータ活用，それに伴うデータ管理の必要性が高まってきている。このようなゲノムデータを活用した医療は，precision medicine と呼ばれ，診療のあり方はもとより，病院情報システムのあり方についても

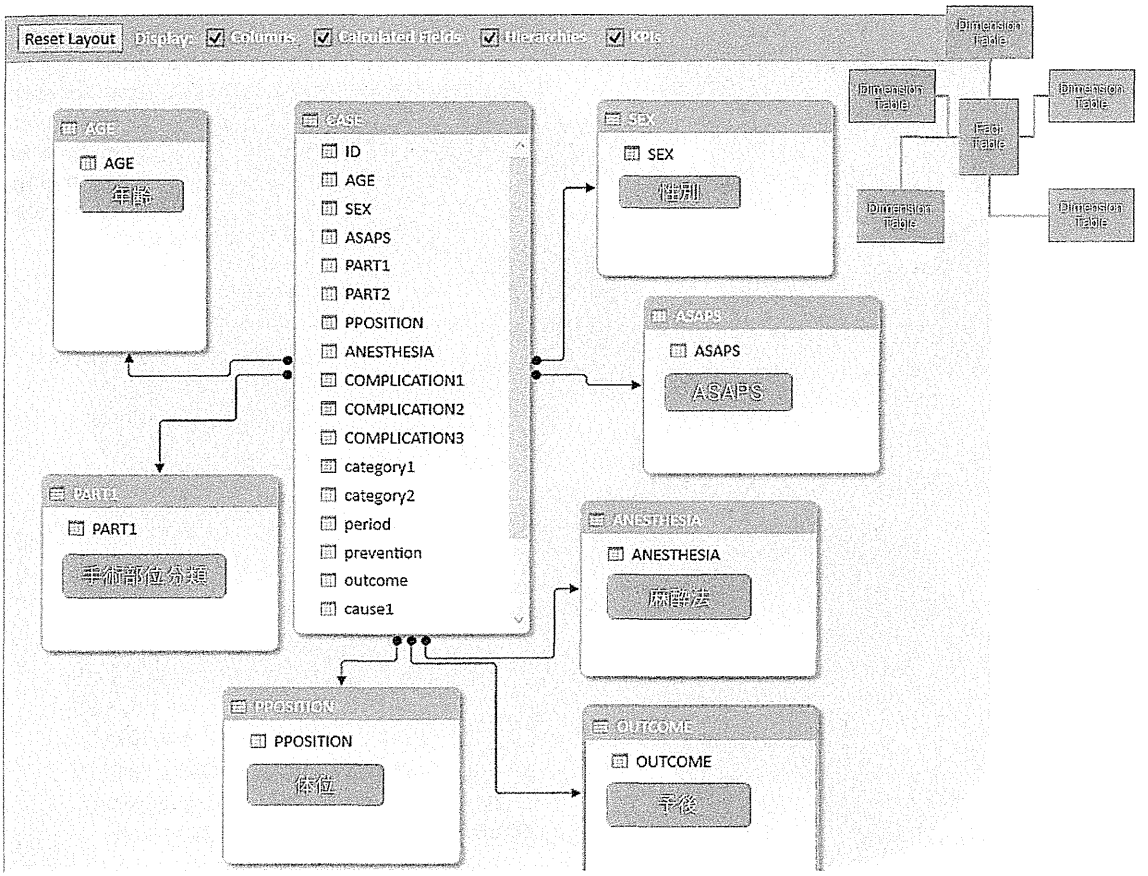

図4　JSAPIMSデータを解析する目的でスタースキーマを構成した例

影響を与えている。

細胞レベルでのデータ活用が可能になる一方で，ライフログのような個人の生活上で発生するデータ，あるいは社会や環境のデータが医療で活用され始めている[4]。このようなデータは，patient-generated health data[6]とも呼ばれ，PHR（personal health records）などの活用とともにconsumer ehealthの推進力として期待されている。米国では，medical home，あるいはpatient-centered medical home（PCMH）として，さらに周術期医療においてはperioperative surgical home[7]として積極的な取り組みがなされている。

2）テクノロジー

このように量が増加し，種類が多様化する医療のデータをどのように処理すべきか。その方法と手段について議論したい。ここで“テクノロジー”として紹介する内容は，基本的な機能がハードウエアやソフトウエアの製品として提供されているものである。したがって，どのような製品であるか，どのように使用するかを知ることで活用の幅が大きく広がるであろう。

電子カルテ，あるいは病院情報システムのデータ分析は，どのようなシステムで実施すべきか。過去には電子カルテシステムのデータベースに保存されているデータに対して，処理プログラムを直接接続させる方式がよいのか，あるいは別途に解析用のデータベースを構築し，その解析用データベースを操作する方式がよいのか，意見が分かれることがあった。今日においては，図2に示すように，業務系システムである電子カルテシステムとは別に，データ処理や解析を目的としたデータベース（データウエアハウス）を構築するのが

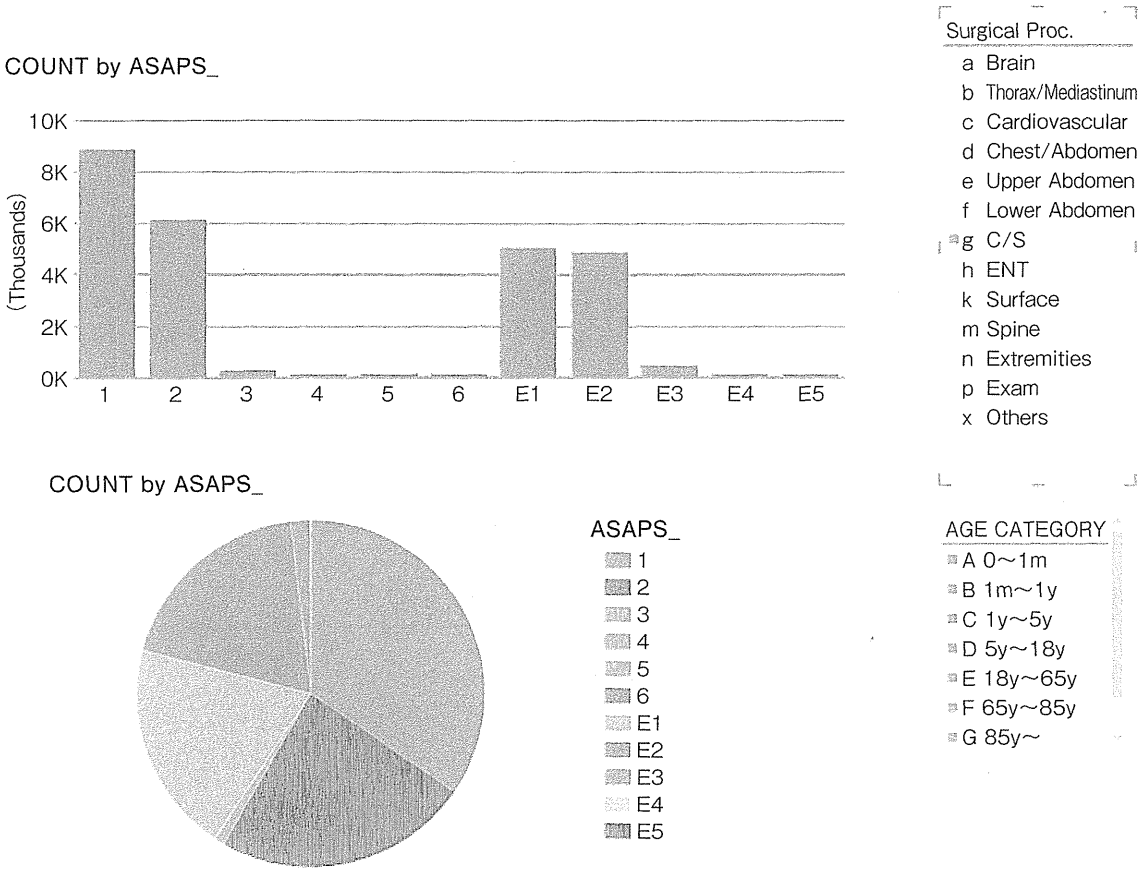

図 5　スタースキーマデータベースに BI 製品を適用し ASAPS の分類を可視化した例

主流となっている。

このようなデータ解析を目的として構築したシステムを活用するために，アナリティクス（analytics）という分野が発達してきている。アナリティクスでは，統計学やプログラミングを駆使してデータを加工することにより，意味あるパターンの発見を目指す。また，データ処理により発見したパターンを提示するためにデータの可視化を有効に行う方法や，ツール類も開発されている。特に企業向けに一連の解析技術や製品群は，BI（business intelligence）と称して提供されている。ここでは，アナリティクスで活用されるデータベース技術や BI における可視化技術を通して，周術期データへの適用可能性について検討する。

（a）スタースキーマ（star schema）

リレーショナルデータベースにおいて，一般的にデータは，項目である“列”と各データを格納する“行”から成る“表”として実装され，複雑なデータは複数の表とそれら表の関係により構築される。電子カルテのデータも例外ではなく，多くのシステムではこのようなリレーショナルデータベースを採用している。リレーショナルデータベースにおいて，表の関係や各表での項目の構造を定義したものをスキーマ（schema）と呼ぶ。スタースキーマ（star schema）は，このようなデータ構造の定義方法の一つであり，特にデータ分析において採用される構造である。図 3 にスタースキーマの例を示した。図 3 の左は，スタースキーマの一般的な構造を示している。スタースキーマは，解析対象となる fact table と称する表（データ）を中心として，複数の解析の切り口として dimension table を準備し fact table に関連づける。