Appendix

臨床データ標準の新しい世界
—研究科学，テクノロジー，データソース—

抄録

20 年前以上前にデータサイエンティスト，統計学者，研究者が会合して臨床研究の成果を規制当局に提出するためのデータ標準規格の策定を試みたのが，CDISC（Clinical Data Interchange Standards Consortium）の嚆矢である．医薬品や医療機器の安全性を証明するためのデータは，発生源，データ収集，集計方法，内容（語彙）について厳密に管理される必要があることから，CDISC 標準の中でも非臨床試験と臨床試験むけに SEND，CDASH，SDTM，ADaM，ODM などの様々な規格が考案された．その結果，世界各国の規制当局に支持され，臨床研究データの世界的な標準規格の地位を占めるに至っている．日本では独立行政法人医薬品医療機器統合機構（Pharmaceuticals and Medical Devices Agency PMDA）が，米国ではアメリカ食品医薬品局（U.S. Food and Drug Administration FDA）が臨床試験のデータを CDISC 標準で提出することを義務つけている．さらに近年はコンピュータで自動的に CDISC 標準に準拠したプログラム，データを生成できるように CDISC Library が公開され，CDISC 360 プロジェクトの有志によって実証実験が進められている．そして，CDISC の対象範囲は一般的な臨床研究にも広がり，疾患領域別データ標準である Therapeutic Area User Guides （TAUGs）という拡張によって，腫瘍，血管疾患，神経疾患，感染症等における臨床研究もカバーされつつある．また，世界最大の Funding Agency である米国国立がん研究所（National Cancer Institute NCI）は研究者によって提出されたデータを共有するためのプラットフォームである Cancer Data Research Commons（CDRC）を構築し，そこに蓄積されるデータは CDISC 標準に準拠するように求めている．以上のように CDISC 標準はデータの品質管理，マネジメントを含む包括的な規格であるが故に，医学領域の各研究領域に浸透しつつあり，新しくかつ安全で有効な医療機器，治療方法の開発の迅速化に貢献している．リアルワールドデータ（Real World Data RWD）のデータも品質が低い傾向があるが，CDISC をマネジメントの中に組み込んでいくことで品質を引き上げていくことが期待される．また，CDISC は観察研究の利用も考慮している．

Ⅰ．はじめに

20 年以上前にデータサイエンティスト，統計学者，研究者のグループが非公式に会合を開き，臨床研究の成果の規制当局への提出と当局によるレビュープロセスの効率を向上させるために，共通のデータ標準セットの開発を始めた．この初期の取り組みから CDISC（臨床データ交換規格コンソーシアム: Clinical Data Interchange Standards Consortium）が形成された．CDISC は，臨床および非臨床研究のライフサイクル全体を記述する標準規格を構築するために，世界中にまたがった専門家のコミュニティを招集し，データのありかたを明確にする活動を継続している．

Ⅱ．データ標準とは何か?

データ標準に関する皮肉として，標準規格を記述するための「標準的な方法」が存在しない．確かに標準規格を定義する標準的な方法はない．標準規格とは一般に「メタデータ」，すなわちデータに関するデータを指す．つまり標準規格とは，誰が，何を，いつ，どこで，なぜ，どのように，データを収集したのかを確認しやすい方法で整理しようとする．臨床研究の主な関心対象は，調査中の疾患

または障害であり，またはそれに対する介入や，介入の頻度やその結果，評価であり，介入とその結果の評価，およびその他の病態 — 患者が既に経験していること，あるいは介入が始まる前，直後，介入中/介入後に経験したことなどがある．CDISC の実装者は CDISC 標準のことを「（データ）内容の標準規格」と呼ぶ．これは，CDISC 標準がデータの歪みを最小限に抑えながらデータの内容を標準化することを意味する．CDISC メタデータは，データが「どのように」収集されるべきであるかについて示しますが，どのようなデータを収集すべきか，またどのようなリサーチクエスチョンがなされるべきかについては言及していない．CDISC 標準には，メタデータを編成し，データセットを構造化する方法を標準化し，データ共有とシステム間の相互運用性をサポートするためのデータ交換規格が含まれる．

Ⅲ．CDISC 標準メタデータ

CDISC が想定している主要なユースケースは，バイオ医薬品，ライフサイエンス，医療機器産業によって規制当局に対して彼らの製品が安全で有効であることを提示するために作成されるデータセットを定義することであり，そのデータを表現するためのアーキテクチャの開発に何年もかけて漸次的に開発してきた．CDISC の主要な基盤規格には次のものが含まれる．

- SEND は，非臨床試験に関するデータ標準です．安全性データ，動物モデルおよび組織からのデータ，ヒト前臨床試験からのデータを整理し一貫性のある形で扱う
- CDASH はデータ収集に関する標準であり，収集されたデータを集計および分析や分析のモデルのセグメントに直接マッピングするために，調和した方法でデータ収集を標準化する
- SDTM はデータ交換に関する標準で，研究者がデータを整理，形式化，集計するのに使える．また，SDTM はデータの集約とデータベースへの格納をサポートする
- ADaM は，解析に関する標準で，トレーサビリティと再現性を最大化しながら，効率的に結果の生成を可能にする

上記の中核となる基盤規格は，以下の標準規格群によってサポートされている．

- TransCelerate BioPharma で開発された研究プロトコル表現モデル（Protocol Representation Model PRM）は，研究プロトコルの計画とデザインを標準化する．研究デザイン，適格基準や ClinicalTrials.gov，WHO，EudraCT レジストリからの要件などの研究に関する特性に焦点を当てている
- 統制用語（Controlled Terminology: CT）は，米国国立衛生研究所の米国国立がん研究所（NCI-EVS）と提携して開発された用語，概念，および変数に関する用語集である．CT には，一般的に使用される質問表，評価，および尺度（Questionnaires, Ratings and Scales: QRS）の項目の標準化も含まれる
- データ交換規格は，臨床研究のライフサイクル全体を通じて，さまざまな電子システム間でメタデータとデータの転送を容易にする
    - Define-XML は，CDISC SDTM および ADaM メタデータを機械読み取り可能な形式 [1] で格納するデータ交換標準であり，自動化を可能にし，データの理解と共有を容易にする
    - ODM は，研究データに関連するメタデータ（管理情報，参照情報，監査情報）と共に，臨床およびトランスレーショナルリサーチのデータを交換およびアーカイブするための，プラットフォームに依存しない表現形式である．ODM は，多くの電子的

データ収集システム（Electronic Data Capture: EDC）で症例報告書（Case Report Form: CRF）の内容を表すためにひろく利用されているため，CDISC 標準の中で最も広く使用されているものの 1 つである

広範にわたって一貫した CDISC 標準の使用を促進するために，CDISC は Critical Path Institute（C-Path），米国食品医薬品局（FDA），日本の独立行政法人医薬品医療機器統合機構（PMDA），NCI-EVS，及び様々なステークホルダーと協力して，疾患領域別ユーザーガイド（Therapeutic Area User Guides: TAUGs）の開発を開始した．TAUGs はその疾患や兆候について，最も一般的に利用される，特定のデータ収集に焦点をあて強調し，科学的および臨床的状況の両方の範囲で CDISC 標準の例を提供することによって，試験の標準化を活性化させるものである．現在では，がん，心血管疾患，神経疾患，感染症等の疾患カテゴリーをカバーする 30 以上の TAUG が存在する．CDISC TAUGs は，進化する科学と規制のニーズをサポートするために開発され続けるだろう．

Ⅳ．臨床データ標準のボランティア開発

CDISC 標準は，ボランティアで参加している専門家のグローバルコミュニティによって開発され続けている．CDISC は実際にボランティアによって設立された．これらのボランティアには，組織の管理職，データサイエンティスト，統計学者，コンピュータプログラマー，標準規格に関する専門家などが含まれている．一部のボランティアは CDISC プロジェクトで自分の時間を割いて参加しており，多くのボランティアはチームベースのプロジェクトに取り組むために彼らの雇用者のサポートを受けている．ボランティアのチームは，規格のスコープ定義，活動の立ち上げ，規格の更新・開発を管理し，標準化の対象となる新しい領域（CDISC ファーマコゲノミクス（ゲノム薬理学）や遺伝学の標準，PGx など）を特定するのに貢献する．ボランティアによる標準規格の開発作業は，一般的な CDISC モデルに方法と準拠の両方に合致するように編成されている．オープンな開発プロセスの勘所には，ボランティアやスタッフを含む専門のデータモデルの設計者で構成されるグローバルガバナンスグループによる品質レビューが組み込まれる．CDISC 標準開発は，最高規格責任者と標準規格開発に関する 2 名の責任者によって率いられ，小規模ではあるが強力なスタッフであるコンテンツ専門家とプロジェクトマネージャによってサポートされている．

CDISC のボランティアは，世界の臨床データ標準化に関わるコミュニティにとって重要な他分野に大きく貢献している．ボランティアは日本語や中国語等を含む英語以外の言語への CDISC 標準の正確かつ専門的な翻訳を保証しサポートする．また，CDISC Interchange と呼ばれる地域会議や，世界中に散らばっているユーザーグループの開発を支援している．CDISC のボランティアは，標準を利用する人々が標準規格を理解し，標準化から最大の恩恵を引き出すことを支援するための教材も開発している．全ボランティアの取締役会は，経営戦略を設定し，非営利団体の最高経営責任者（CEO）を選出する．

Ⅴ．グローバルスタンダード

CDISC 標準は臨床研究の事実上のグローバルスタンダードとなっている．その理由は，CDISC 標準が医薬品，診断技術，医療機器の開発のスポンサーである事業者と，安全かつ効率的であると実証された医薬品，診断技術，医療機器のみを承認することを通して公衆衛生を保護する規制当局との間で，不可欠なやり取りを促進する為に広く活用されている規格だからである．CDISC 標準を扱っている規制当局には以下がある．

- 日本の PDMA は，2016 年 10 月から 3 年間の移行期間を設けて CDISC 標準によるデータ提出を義務付けている．2020 年 3 月に移行期間を無事に終えることとなり，ほとんどの臨床試験のデータは現在 CDISC 標準で提出されている
- 米国 FDA は 2016 年 12 月から CDISC 標準による提出を義務付けている（2017 年 12 月以降の新薬の新規出願（IND）から）
- 中国国家医療製品庁（China National Medical Products Agency : NMPA，旧 China FDA）は CDISC 標準を強く推奨している．2018 年現在，NMPA に提出された 70%以上の患者試験データが CDISC 標準を利用している
- 欧州医薬品庁（European Medicines Agency :EMA）は，患者レベルの試験データを必要としない．それにもかかわらず，EMA と各 EU 加盟国を代表する医薬品機関の代表は，グローバルなデータ標準規格の利用を明示的に推奨しており，CDISC を臨床研究の為の標準規格として特に指名している．EMA は基本的に，すでに標準規格が存在している場合は，新しいデータ標準規格の開発に反対している

Innovative Medicines Institute や米国国立がん研究所（NCI）などの政府系の研究資金提供者も CDISC 標準を使うように推奨している．最近，NCI は開始から CDISC 標準を利用したクラウドベースのデータ共有，分析，保存，データの再利用のサービスを提供する Cancer Data Research Commons を立ち上げた．その NCI は現在，がん研究の世界最大の資金提供者である．また，ビル・アンド・メリンダ・ゲイツ財団，レオナ・M・アンド・ハリー・B ヘルムズリー・チャリティー・トラストなどの生物医学研究の民間資金提供者は，関心のある分野での CDISC 標準の開発を支援が増えており，臨床研究計画において標準も活用してきている．

規制当局が CDISC 標準を選ぶ理由として下記の様に CDISC に説明している．

- 標準化により，レビューワーの品質と適時性が向上する
- CDISC 標準は，レビューワーがデータを確認し，スポンサーの主張を理解するのをサポートするように適切に設計されている
- 多くの多国籍製薬会社，CRO，およびこれらの企業にサービスを提供するテクノロジー企業は，すでに CDISC 標準を利用している．CDISC 標準をエンタープライズ・アーキテクチャに組み込んでいる．CDISC 標準を活用するようスタッフメンバーを育成し，CDISC 標準は業界の事実上の国際標準となっている
- CDISC の組織が長期間存続し安定しているため，CDISC は標準のバージョン管理や更新を含め，長期間に渉って標準情報モデルと標準規格を安定したものにしている
- CDISC 標準から構築された可視化ツールや検証ツールなど，グローバル市場で利用できる
- CDISC は，統計的なレビューワーや医学的なレビューワーにトレーニングを提供する

CDISC 標準は成熟しており，世界的に認められ，活用されている．CDISC 標準は，臨床試験データの世界標準となっている．

VI. データ標準を使用する新しい方法

CDISC 標準は安定した規格であるが，同時に発展中でもある．新しいデータソース，新しいタイプの科学研究，および新しい技術の出現は，標準規格を更新し，新しいバージョンを開発することが必

要となる．2017 年，CDISC は野心的なプロジェクトを立ち上げた．それは，CDISC 標準について単一の機械読み取り可能なインスタンスを生成することである．CDISC Library という名前のメタデータリポジトリは，2019 年 4 月に立ち上げられた．CDISC Library は，「リソース記述フレームワーク（RDF: Resource Definition Framework）」を介してグラフデータベースとして表現される新しいセマンティック技術基盤を使用して構築された．CDISC Library は，リンクされたデータと REST API を使用して，標準ベースのプロセスの自動化に役立つ CDISC 標準メタデータを eSystems に提供する．CDISC Library 公開時には，過去 5 年間に開発された全ての基盤規格と統制用語が含まれる 100 万以上のリソースが含まれていた．グラフデータベースを使用すると，リソースを概念レベルでリンクできるため，CDISC Library 内に 600 万を超えるリンケージ（例：RDF トリプル）がある．CDISC 標準は，今やすべての実装に対して単一かつ厳密なソースを持つようになった [2)]．

この画期的な標準メタデータリポジトリは，堅牢な基盤であり，新しい革新的なアプリケーションやツールをサポートするテクノロジープラットフォームである．CDISC は，CDISC 標準に準拠した実装を容易にし，かつ一貫性のあるものとして実現するために，標準に関連した活動の自動化を含むフロントエンドのユーザーインターフェイス/エクスペリエンス層の構築に役立つツール開発者と協力し始めている．

四半期ごとに基本的な標準規格，TAUGs，QRS 補遺，追加の統制用語についてのバージョン更新，新規追加を含む新しいコンテンツが CDISC Library に追加される．

2019 年 1 月に CDISC は CDISC Library の新機能を基にしたオートメーション・パイロット・プロジェクトである CDISC 360 を開始した．CDISC 360 の目標は，特定の分野における臨床研究評価において，CDISC 標準の完全で機械処理可能な自動化を最初から最後（終始一貫した自動化）まで実証することにある．このパイロットプロジェクトには，ヨーロッパ，日本，北米，中国の 26 の加盟企業から派遣された人達が参加している．成功した場合は，CDISC 360 は多次元標準を構築するためのテンプレートを提供する．CDISC の目標は，まず機械読み取り可能な標準規格を構築し，次いで人間向けの標準規格を策定することである．この新しい標準規格には，標準化と視覚化を可能にするために必要なリンクされたメタデータが含まれる．

- 電子症例報告書（CRFs）
- 表，図，およびリスト（TFLs）
- CDISC 標準の一貫した実装をサポートするその他の研究成果物
- HL7 FHIR などの他の規格への標準化された機械読み取り可能なマッピング

CDISC 360 は，短期的な集中プロジェクトに分かれた野心的な取り組みである．CDISC 360 プロジェクトに関する最新情報は，CDISC のウェブサイトから入手できる．

Ⅶ. 新しいデータソース

今後数年間で，新しいデータソースが臨床研究の世界に入ってくるだろう．ほとんどの標準規格開発組織と同様に，CDISC の取り組みはもともと保守的なものである．科学技術が定着した後は対応する標準規格を効率的に開発することができる．それにもかかわらず，CDISC は，CDISC 標準が今後数年間で発展し続ける可能性が高いいくつかの分野を予測している．

- 電子健康記録（EHR）データをソース化 [3)] し，利用をスケールアップ．CDISC は HL7 FHIR および TransCelerate Biopharma と提携しており，EHR のデータセットを組み込んだ臨床試

験の実施の改善に注力している．CDISC のスタッフは，現在の EHR データソースに内在する課題と限界を認識しているが，同時に潜在的な洞察の可能性も認識している

- 機械学習（ML）やディープラーニング（DL）を含む人工知能（AI）が臨床研究の状況を変える可能性があることを示唆しているが，これらのアプリケーションは，言うは易く行うは難しである．単により多くのデータが必要というに留まらず，完全かつ膨大な量のデータを要求すること，そして現状はデータ品質が不十分であり，改善にコストがかかるという警告に対して傾聴すべきである．CDISC のスタッフは，上記に述べた問題意識に関して，CDISC Library と CDISC 360 プロジェクトが，AI 及び ML の可能性を引き出すために重要であると考えている
- ゲノムデータと個別化医療は進化し続けている．CDISC SDTM PGx チームは，この分野で CDISC 標準の更新に取り組んでいる
- レジストリは CDISC への関心を高めている．CDISC ブルーリボン委員会は，CDISC がステークホルダーと協力して，レジストリが CDISC 標準で設計されるようなツールを構築する必要があることを示唆している

Ⅷ．観察研究

歴史的に，CDISC 標準は主に規制当局への医療製品の市場投入の承認をサポートする臨床試験データの提出に使用されてきた．しかし，最近では，疾患領域別ユーザーガイド（TAUGs）開発による CDISC 標準の適用範囲拡大と CDISC の存在の認知向上により，医学研究の他の分野におけるデータ標準規格としての価値を見いだされることにつながっている．TAUGs で主に説明されている CDISC 標準の既存の生物医学的なコンセプト内容は，研究の種類に関わらず，観察研究で収集されたデータの限定的な比較から検討された類似の概念とよく似ている．CDISC 標準は，データ標準化の恩恵を受けることを望む学術研究者によって利用されている．これらの研究者は，迅速かつ低コストにエビデンスの生成，堅牢な分析，研究データの共有，研究の再現性を確保できる手段として標準規格を位置づけている．CDISC 標準は，Data Sphere プロジェクトや Vivli などのグローバルなデータ共有プラットフォームで活用されている．CDISC のツールは，REDCap などの学術的利用に焦点をあてたプラットフォームでも利用されている．CDISC TAUGs から提供されている，トレーニング製品および派生成果物は，他の目的や低・中所得国でのデータ収集を支援するためにも開発された．

CDISC 標準を活用しようとする学術研究者が直面する課題は多数ある．観察研究は，研究目標，研究デザイン，被験者集団，臨床環境，規制/研究に関する監視要件，データ収集およびデータ管理に方法論等において，無作為化比較対照試験とは異なる取り組みである．これらの多くの違いから示された課題によって，観察研究に CDISC 標準を採用することへの障壁が認識されている．無作為化比較対照試験とは異なり，観察研究は介入を伴わず，健康上のアウトカムに影響を与えるために試験責任医師から働きかけることもない．学術・政府の研究機関で収集された場合，観測データは大体において高品質な傾向がある．これらの研究は研究プロトコルに従って実施されており，観察研究モニタリング委員会（Observational Study Monitoring Board）による監視の対象となる．無作為化比較対照試験と同様に，観察研究は採用される研究デザインに様々な形態があり，一般的にケースコントロール，コホート，または横断研究に分類される．無作為化比較対照試験の目的は，介入の安全性および/または有効性を決定することである．対照的に，観察研究は疾患の結果に関連しているであろう，潜在的なリスク因子を探索することを目指している．ランダム化が欠如しているため，観察研究はバイアスを起こしやすくなる．そのため，分析中にバイアスを制御するために潜在的な交絡因子を収集する必要がある．1 つの化合物に関する試験の分析は，標準化することは比較的簡単にできる．それは，数週

間または数ヶ月の期間にわたって，介入群はベースラインから変化があったか？またコントロール群と比較して変化したか？といった内容である．観察研究は，何十年間もののコホートを追跡し，ライフスタイル，行動様式，環境，社会経済的要因が健康上のアウトカムにどの程度寄与するかを検討するだろう．現状の CDISC モデルでは，こういった観測研究特有の事情を表すために追加の構成要素が必要になる．

第二の課題は，学術研究のための研究プロトコルを標準化することは非常に困難であるということである．研究活動の性質上，学術研究者は新しい知識を構築している．簡単な研究のためのプロトコルであっても，細部にわたって細かくみていく必要がある．現状の標準規格レベルでは，この微妙な違いを標準に反映することは非常に困難であり，CDISC 標準だけでなく，どのデータ標準でも任意の粒度でプロトコルも標準化することが強く求められるようになるだろう．

臨床研究にとどまらず，EHR，請求や支払い，患者レジストリ，モバイルデバイスなどの「リアルワールド」のデータソースから観察データが生成される場合もある．これらのデータは，研究を支援する目的で収集されていないため，一般的に研究で収集されたデータよりも不完全で低品質である可能性がある．

CDISC のスタッフは，スタッフの規模と予算上の制約により，大規模に成長しつつある CDISC 標準を利用している研究者のグローバルコミュニティとの交流が限られていた．しかし，CDISC のスタッフは，データの標準化を改善し，標準化のメリットを研究者の業務に適用していく大きな機会があると考えている．将来の取り組みが期待される領域は次のとおりである．

- 学術研究者が CDISC 標準を最大限に活用する方法を理解する
- 学術研究者をサポートする用途にあわせた CDISC 標準を開発する（ただし，スポンサー・規制当局のユースケースには適用されない）
- データ標準化に対する既存の障壁を克服するための具体的な戦略を策定する
- データ共有のためのより優れた基盤を生成し，プラットフォームに依存しないデータ共有をサポートする
- CDISC Library 上に構築されたオープンソースツールやその他の手頃な価格のツールを開発する
- 研究者によるデータ標準化に関する良い取り組みを見いだし，ジャーナルへの投稿を奨励する

IX. 結論

CDISC 標準は，臨床研究のライフサイクルを表すために考案された，複数かつ相互に運用上の配慮がなされた標準規格群より構成されているものである．CDISC 標準は，研究事業者が利用できる新しいデータソースや新しいテクノロジーを活用するために進化を続けている．この進化は，標準規格の開発をより効率的にし，標準規格自体がより豊かで有用になる．CDISC では，これらの変化がデータの内容を明確にし，切実に治療を求めておられる患者に対して，新しくかつ安全で有効な治療法の開発と承認を迅速化するのに貢献するものであると確信している．

脚注

[1)] 機械読み取り可能（machine-readable）という表現は，データの電子化の高度な段階を表現している．例えば，これまではソフトウェアやデータの構造を説明する仕様書などは自然文章で書かれており，利用者はその文章を読

んで自らの手を動かして作り上げていく必要があった．「機械読み取り可能」な形式で作成されたものはコンピュータが自ら読み取ることが可能であり，自動的にプログラムやデータを生成できるようなかたちになっている．

[2] CDISC 標準は複数の標準規格から構成されており，仕様も膨大なものである．膨大な仕様を適切に読み解き，首尾一貫かつ正確な実装を人手で作るのは困難になってきている．そこで，CDISC の諸標準を機械読み取り可能な形で記述し，コンピュータで CDISC 標準に準拠したツール等を自動的に生成できるようにした．これまで沢山の資料があったが，CDISC Library におさめられたリソースを紐解くことで，CDISC 標準に関する情報が一元的に取り出せるようになった．このことを著者は「CDISC 標準は，今やすべての実装に対して単一かつ厳密なソースを持つようになった」と表現している．

[3] 「ソース」≒入手源．「データソース」は「データの入手源」．