データ利活用に基づく製薬研究開発モデルで勝ち抜くための6つの鍵

pwc

www.pwc.com/jp

目次

はじめに

これまでの製薬企業の研究開発モデルは、主にデータの一次利用に基づいたものであった。一次利用とはすなわち、あらかじめ定められた目的を満たすために取得されたデータを、本来の目的に限って利用するものである。例えば、ある薬剤候補の有効性や安全性に関する仮説を検証するためにGood Clinical Practice（GCP）に従った臨床試験を実施するが、そこで取得された患者データを薬事申請に使用することは、データの一次利用となる。それに対してデータの二次利用とは、当初の想定とは別の目的に取得データを活用することである。例えば、臨床試験で取得したデータを基礎研究で使用したり、病院で取得された患者の診療データを研究開発に応用したりするなど、あるデータを当初の目的外で利用することがデータの二次利用となる。

製薬企業では近年、リアル・ワールド・データ（RWD）や蓄積されている社内データなど、多様なデータを二次利用することでデータの価値を最大限引き出し、自社の競争原資とする動きが活発化し始めている*1。これは、戦略意思決定をデータ解析の結果や解釈に基づいて行おうとするデータドリブン（Data-driven）アプローチの全産業的な広がりや、既に当然となっている「科学的根拠に基づく医療（EBM, Evidence-based Medicine）」の実践とも呼応する。

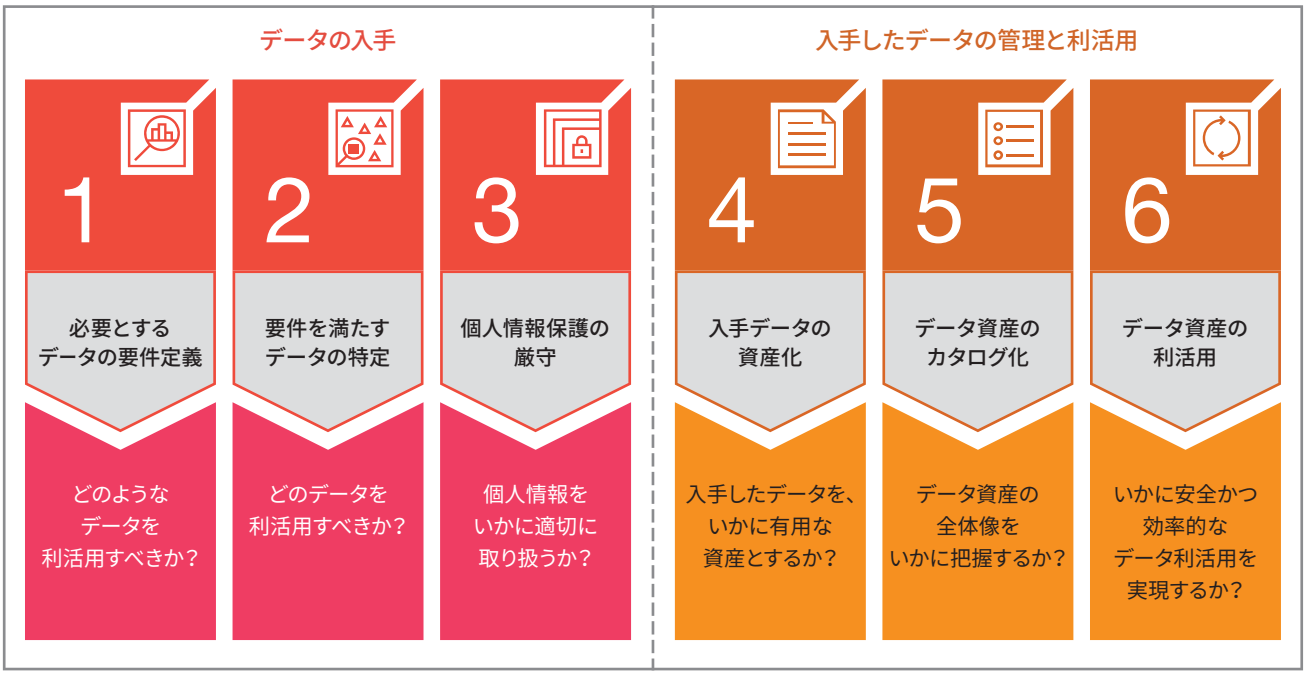
本稿では、このようにデータ利活用に基づく研究開発モデル（Data-driven pharma R&D）が製薬各社により推進されるなかでいかに自社が適切なモデル構築を先行できるか、そのための最も重要な6つの鍵（Six Keys）を紹介する。

*1. https://www.pwc.ru/en/pharmaceutical/publications/assets/pwc-pharma-2020.pdf

データ利活用における6つの鍵

現在、多くの製薬企業ではデータ利活用推進の重要性が認識されており、各種データの統合と二次利用が計画、もしくは既に開始されている。ここでは、製薬企業がデータ利活用を進める上で考えるべき事柄と、多くの会社で見受けられる問題点を6つの項目に分けて示す（図表1）。ここでのデータ利活用の手順は、大きく2つのフェーズに分けられる。ひとつはデータ入手をするフェーズ、もうひとつは入手したデータを実際に管理・利活用するフェーズである。前半の「データの入手」のフェーズではまず、(1) 課題解決や目的達成のために必要とされるデータはどのようなものか、その要件について検討する。次に (2) 決まった要件に即したデータを実際に発見・特定し、そして (3) 特に臨床データについては、入手の際に個人情報保護への配慮を最大限行う。後半の「入手したデータの管理と利活用」のフェーズでは、(4) 入手した全てのデータについて、利活用が可能となるデータ資産に変換し、(5) それらデータ資産の有無と所在が今後とも容易に一覧・検索できるようにデータカタログを作成する。そして、(6) 運用における安全性と効率性をバランスよく兼ね備えた運用モデルを構築し、これにのっとってデータの利活用（二次利用）を適切・積極的に進めていくこととなる。以下では、これら「6つの鍵（Six Keys）」について、項目ごとに解説していく。

図表１：データ利活用に基づく製薬研究開発モデルで勝ち抜くための「6つの鍵（Six Keys）」



1 必要とするデータの要件定義

まずは自社にとって優先的に解くべき課題や検証するための仮説を確認し、
それを解くために必要となるデータの要件を定義する

製薬企業がデータの利活用をする上で、以下のような好ましくないケースが見受けられる。

- 取りあえずデータレイク（全ての構造化データと非構造化データを保存できる一元化されたリポジトリ）などの「ハコ」を作り、少しでも有用そうなデータを手あたり次第に保存してしまう
- データレイクに無秩序にデータをため込んでも、AIを導入することで自動的にデータを利用できる状態へと整理できるものと考えてしまう
- 活用方法や解析方法は後で考えるとして、まずはデータの集約だけに取り掛かってしまう

社内外の具体的なデータを見つけては、手あたり次第に集約しようとするボトムアップなアプローチは比較的楽であり、早くデータが集積・集約されていくため、成果が上がっているように見えやすい。しかし実際には効率が悪く、多くの作業が無駄に終わることが多いアプローチである。データレイクをいざ開けてみると、全く役に立たなかったり、もしくは使えそうなのに使えないデータばかりが詰まった「ぜい肉質のデータ資産」となっていたりすることが少なくないからである。

PwCでは、まずは各社における優先的な課題や検証したい仮説を具体的に確認した上で、必要となるデータの要件を確実に定めてからデータ発掘に進むという、「仮説提唱型のデータ利活用アプローチ」を推奨している（図表2）。

ここではまず、（A）自社の具体的な競争戦略を明確にする。例えば、オンコロジー（腫瘍学）領域の特定の疾患分野でマーケットリーダーになること、といった具体的な戦略を全社的に合意することが重要である。

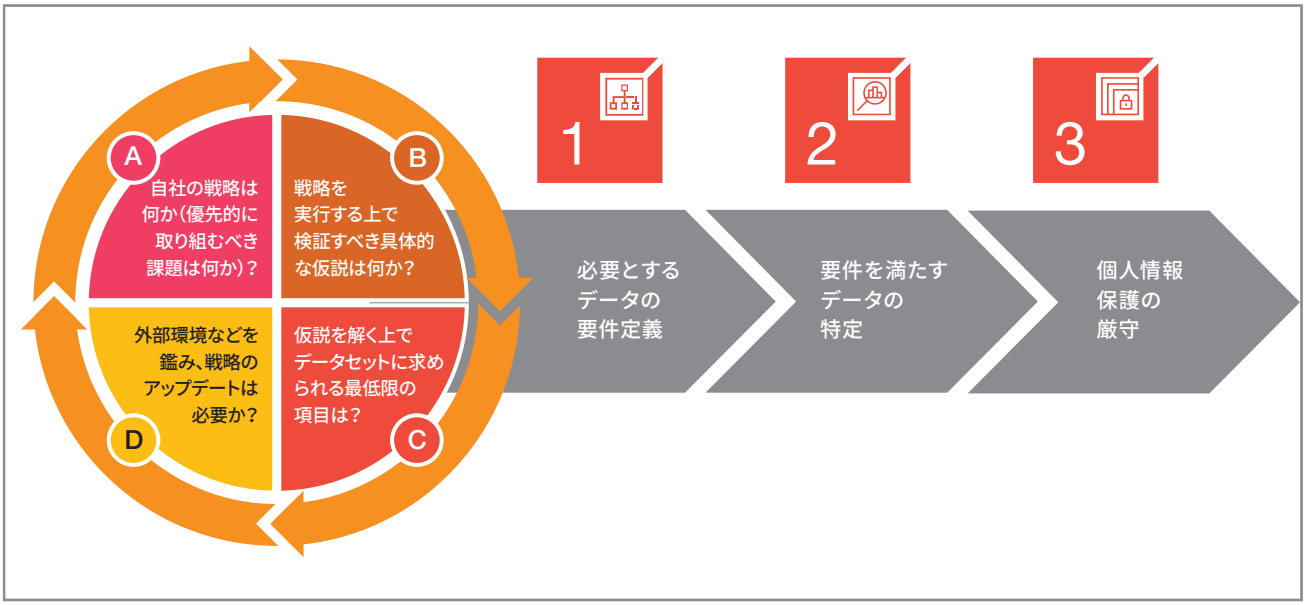
次に、（B）戦略を実行する上で検証すべき具体的な仮説を考える。例えば、自社のある医薬品を使用した患者のアウトカム（具体的な数値の改善、バイオマーカー指標の変化）をRWDで検証するといったサイエンスクエスチョンなど、ある程度具体的な指標にあたりをつけておくことが必要である。

そして、（C）データの要件を想定する段階では「仮説検証のためには少なくともこれらの情報が必要だ」といった必要最低限のデータ項目をリストアップする。

あれもこれも、とさまざまな項目をリストアップすると、理想のデータはいつまでたっても見つからない。一方、項目のリストアップが足りないと、データを集めても検証には不十分、もしくは使い物にならないことがある。データのイメージが定まったら、そのイメージに合うデータ「だけ」をデータプラットフォームに集約することで、はじめから利用目的にあったデータばかりが集められた「筋肉質のデータ資産」ができ上がる。

さらに、検証すべき仮説は時に応じて変化するものであると理解しておくことも重要である。（D）戦略の修正にひもづく検証仮説の更新と、そのためのデータのイメージというサイクルを高速に回し続けること、つまり「会社に必要な筋肉」が何であるかを考え続けることが、データ利活用の第一歩だと考えられる。

図表２：筋肉質のデータ資産を作る、「仮説提唱型のデータ利活用アプローチ」



2

要件を満たすデータの特定

必要となるデータの要件が定義できたら、それを満たす社内・社外データを発見・特定していく。ここでは、各データオーナーにデータリストを提示してもらう受け身だけでなく、明確な要件に基づいてデータを特定する「攻めの姿勢」が必要である

具体的なデータの要件が定まったら、実際にデータの特定に取り掛かる。ここでは大きく2つ、社内データ（Internal data）の特定と社外データ（External data）の特定とに分けて考えることにする。

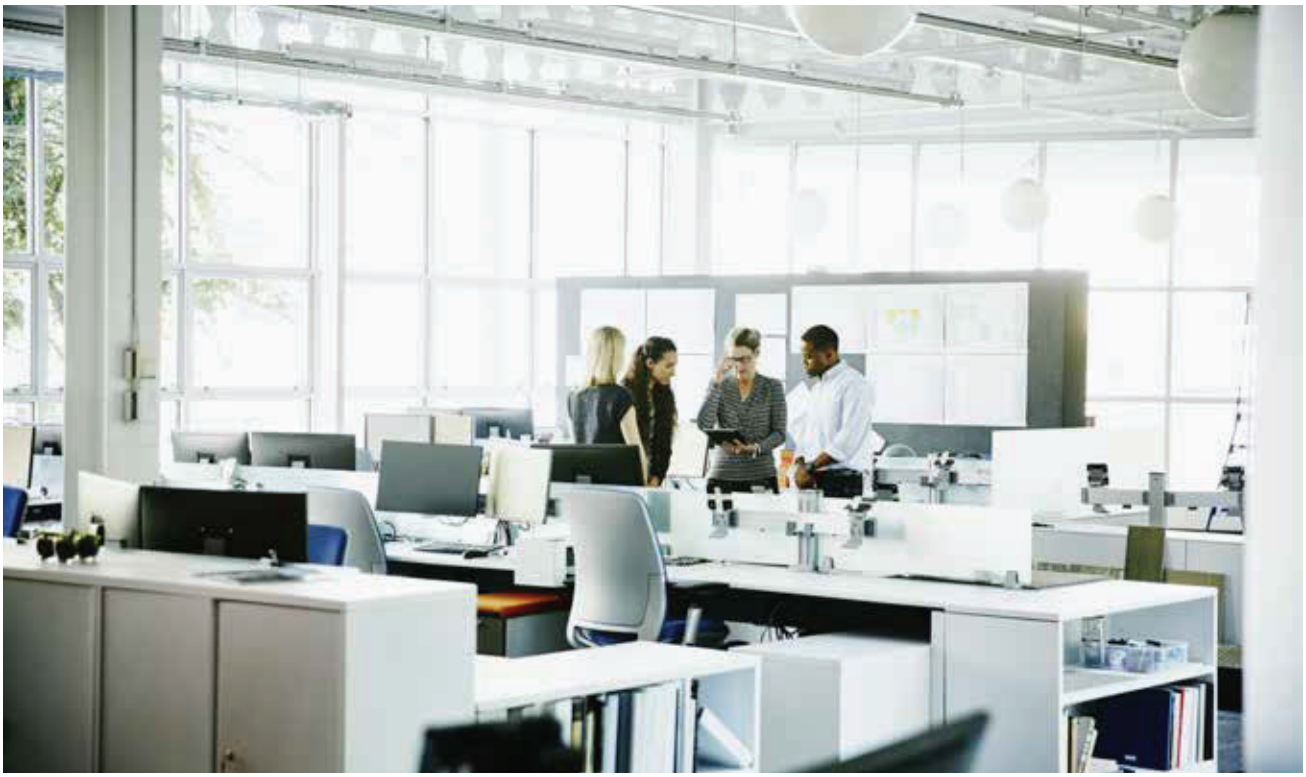
社外に比べ、社内データはある程度把握しやすいことは想像にたやすい。なぜならば、データに詳しい部門もしくは人間が少なくとも社内に存在するからである。ただし、社内のデータであっても、以下のような状況では多部門にわたるデータの全容をつかむのが難しくなる。

- 組織が縦割りになっており、協業や情報共有をする文化が根づいていない
- データオーナーは本業業務があるなかで、ビジネスリスクを負ってまで他部門にデータ共有する意義を感じない
- データユーザーが他部門のデータに対する理解に乏しく、明確な目的を提示できない

これらの課題に関しては、まずデータガバナンスを確立させることで、データの取り扱いにおけるリスクを減少させ、データオーナーの心理的ハードルを下げ、話し合いの土台に載せることが重要である。その上で、データ活用の目的を明確にし、患者、医師、自社へのメリットがあることを証明することで、データオーナーをプロジェクトに巻き込んでいくことが望ましい。これらデータガバナンスについては第6節「データ資産の利活用」でも詳述する。

一方で、社外データに関しては状況が異なる。さまざまなRWDが研究機関、医療機関、公共データベースなどに蓄積されているが、一部のデータベースを除いて「データカタログ」のような俯瞰的な一覧情報はなく、どのようなデータがどこの組織やネットワークに存在・保存されているのかを把握することは非常に難しい。特に、他社とは異なるユニークで価値があるデータを取得して、R&Dをはじめとする製薬企業活動に生かそうという場合には、市販されているデータではなく、独自のデータを発掘していくことが重要となる。

ここで求められるのは、積極的に他社・研究機関などとパートナーシップを組み、定義した要件に合うデータを積極的に探していく「攻めの姿勢」である。「仮説提唱型のデータ利活用アプローチ」により必要なデータの具体的なイメージを持てたら（要件が想定できたなら）、まず積極的に外部のパートナーシップ先を探し、相手方の持つデータを整理することが必要である。そして、その中で自社に有用なデータを特定（後ろ向きのデータ発掘）する。さらには、自社にとって有用なデータを共に作成していく（前向きのデータ創出）をする姿勢が必要となる。相手が持っているデータを俯瞰的に吟味するために、まずはデータカタログを出してもらい、その中から有用そうなデータを選びたいといった「受け身の姿勢」のために、いつまでも理想の相手が現れずに時間だけが経過してしまうという例はしばしば見られる。



3 個人情報保護の厳守

患者データは原則として個人情報であり、その利活用には個人情報保護の配慮が必要となる。これに対し、同意を取得した上での匿名化、もしくは匿名加工による非個人情報への変更という、2つのアプローチがある

利活用したいデータセットを社内外で特定しても、これをすぐに入手できるわけではない。特に、該当するデータが患者の個人情報である場合には、個人情報保護法（これは各国や地域で法律が異なる）にのっとった処理が必要となる。通常、データ提供者（ヘルスケア分野の多くの場合には「患者」）の同意文書により、データの活用方法は特定の用途に限定されている。臨床開発データであれば、特定の医薬品候補の有効性や安全性の検証のため、また診療情報データの場合であれば、患者への最適な医療提供のため、といった具合である。この場合、例えば臨床開発で取得したデータを患者の同意なく、営業部門の社員に共有したり、営業活動目的で使用したりすることはできない。

個人情報データの二次利用を可能にするためには、一般に次の2つのアプローチがある（図表3）。ひとつは①個人情報のまま使用する方法であり、もうひとつは②非個人情報に変換して使用する方法である。個人情報保護法により、個人情報を使用するには情報提供者（データ提供者）から同意を得る必要があるとされている。例えば、臨床開発で取得したデータを個人情報のまま基礎研究部門で使用する場合には、臨床開発でデータ取得する際に、基礎研究でも使用される可能性があることについて患者から同意を取得する必要がある。同意を取得した場合は、臨床試験で取得した患者データ（患者の個人情報）を基礎研究で使用することが可能であり、多くの場合は患者名を伏せて、データユーザーにデータが受け渡される。尚、患者名を削除または変更すること（一般的には、特定の個人を識別することができる記述などの全てまたは一部を削除＜置換含む＞すること）を匿名化というが、この場合データ提供元（ここでは、臨床試験の実施担当など）において特定の個人を容易に照合でき得るため「個人情報がデータユーザーに渡されている」と見なされてしまう。

一方で、匿名加工とは「特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報であって、当該個人情報を復元することができないようにしたもの」と定義される（個人情報保護法第2条第9項）。匿名加工情報は特殊な加工をすることによって、データ利用者のみならず、データ提供側であっても個人を特定できないように不可逆的な処理をしたデータである。このように匿名加工を実施することで、データを非個人情報に変換して二次利用することも可能である（図表3の②）。匿名加工を経たデータ、すなわち匿名加工情報は個人情報とは見なされないため、個人情報保護法が適用されず、よってあらかじめ患者同意を取得していないデータでも法律上では使用することが可能となる。しかし、一般的には①のアプローチの方がデータを使用する際の自由度が高い。従って、データ取得時に二次利用の可能性を考慮し、できるだけ広いスコープで患者同意を取得しておくことが望ましい。

図表3：データを二次利用するための2つのオプション（匿名化と匿名加工）

取り得るオプション	データの取得・利用の流れ
① 個人情報のまま使う 患者データを匿名化してユーザーに提供（ユーザーは匿名化個人情報を使用）	1. 二次利用を含んだ文言で患者同意を取得（Translational research など） 2. ユーザーへ渡る前に、患者データを匿名化（例えば患者名などを伏せるなど） 3. ユーザーが患者同意の範囲内で、匿名化された患者データ（匿名化個人情報）を使用
② 非個人情報として使う 患者データに匿名加工を実施してユーザーへ提供（ユーザーは非個人情報である匿名加工情報を使用）	1. ユーザーへ渡る前に、患者データを匿名加工（複数の患者データをグルーピングし、統計情報に変換するなど）。この時点でデータは患者の個人情報ではなくなり、個人情報保護法の規制を受けなくなる 2. ユーザーが自由にデータを使用 3. 匿名加工情報の利活用に関する義務*2の履行

*2. 詳細については、個人情報保護委員会のページを参照　https://www.ppc.go.jp/personalinfo/

4 入手データの資産化

データを適切に二次利用するには、データを「資産化」するための処理が必要である。
入手した全てのデータついて、今後とも利活用可能なデータ資産への変換を行う

要件を満たすデータが社内外から入手できたら、これを利活用するための処理を施す（これを「資産化」という）。データの一次利用者や取得者（臨床開発データであれば、臨床開発部門の担当者）には内容が明らかなデータであっても、二次利用を意図するユーザーには理解ができなかったり、もしくは誤って使用されてしまったりすることが起こり得る。従って、二次利用者が活用しやすいようにデータを処理することで初めて、データは社内での利活用に耐え得る資産となる。この資産化の工程には、データの記述方法の変更やデータのミス・漏れの修正といった簡単なもの（一般的にデータクレンジングと呼ばれる）から、データフォーマットを変換するといった複雑な処理（PwCではこれをデータエンリッチメントと呼ぶ）など、さまざまな作業が含まれる。特に、データエンリッチメントの手法はデータタイプによってそれぞれであり、例えばテキストデータ（例えば、ゲノムシーケンシングなどのバイオマーカーデータ）においては、以下の観点で手法を考えていくことが多い。

データエンリッチメント手法を検討する際に考慮すべき項目の例

1. データエンリッチメント対象のデータセットの構造が分かりやすいか
2. それぞれのメタデータの意味合いは何か
3. 真に重要なメタデータと、特に重要でないメタデータが、それぞれ認識できるか
4. 複数のデータで横断的に検索をかけた際に、全データ集団の中から目的にあったものを見つけることができるように、重要なメタデータが全て記載されているか
5. （資産化する段階で元のデータフォーマットからより画一的なフォーマットに変換することが求められる場合）一義的にデータを相互変換することが可能か

図表4では、データ資産化の例を示す。まず、元々マトリックス型のデータフォーマット（Matrix Format）（図表4の左部）を、ロングテーブル・フォーマット（Long Format）に変更している（①）。マトリックス型のフォーマットでは、全データテーブルの一部を抜き出し利用することが比較的難しいが、ロングテーブル・フォーマットであれば、特定の遺伝子といったデータのサブセットを検索することが容易になるなどのメリットがある。次に、追加的に情報を記載し、同じ患者の他のデータセットと関連付ける（②）。このように、元のデータセットには記載されていない情報を付け足すことで、同じ患者から得られた各データをひもづけることができる。最後に、データタイプをそろえる作業も重要である（③）。元データでは、テーブルに数字とテキストデータが混ざっているが、LOD（Limit of Detection）、つまり検出不能は実質、値がゼロであると同じであるため、数字「0」に置き換えている。これらのように、全社的に標準化されたルールを早期に規定することで、高品質なデータ資産を築き上げることが可能となる。

図表4：データ資産化において実施する処理の例

元データ（マトリックスフォーマット）　（RPM）*3

遺伝子発現	サンプル1	サンプル2	サンプル3
遺伝子A	1	20	5
遺伝子B	40	LOD	6
遺伝子C	5	9	45

二次利用可能な資産に変換
データフォーマットを変更（①）
他データセットと関連させるために情報を追記（②）
記載の不具合を修正（③）

元データ（ロングテーブル・フォーマット①）

患者番号	組織名	サンプル1	遺伝子	発現量（RPM）*3
XXX	XXX	サンプル1	遺伝子A	1
XXX	XXX	サンプル1	遺伝子B	40
XXX	XXX	サンプル1	遺伝子C	5
XXX	XXX	サンプル2	遺伝子A	20
XXX	XXX	サンプル2	遺伝子B	0 ③
XXX	XXX	サンプル2	遺伝子C	9
XXX	XXX	サンプル3	遺伝子A	5
XXX	XXX	サンプル3	遺伝子B	6
XXX	XXX	サンプル3	遺伝子C	45

②

*3. Reads per million mapped reads

データフォーマットに関しては、遺伝子変異の表記に特化したVCF（Variant Call Format）や、C-DISCのSDTMデータなど、広く受け入れられているデータフォーマットも出現してきている。しかし少なくとも現状において、特に研究が集中するオンコロジーなどの領域においては、ゲノムシーケンシング（Genome Sequencing）の普及により取得されるデータの量は急増しているにもかかわらず、「One size fits all」となるような業界標準のデータフォーマットは存在していない。よって、自社の研究・製品ポートフォリオに応じて、自社で定義したデータエンリッチメントを経て、データの資産化を進めていくアプローチが必要だと考えられる*4。

*4. “Integrated Omics: Tools, Advances, and Future Approaches”

5

データ資産のカタログ化

資産化されたデータの価値が最大化され、今後とも効率よく利活用されるためには、その有無や所在が容易に一覧・検索されなければならない。そのために、重要なメタデータを整理したデータカタログを作成する

資産化したデータを最大限活用する上で必要不可欠なのは、これをデータベース化して、しかるべきユーザーがアクセスできるようにすることである。どのようなデータが自社にあるかを手軽に検索できれば、前述の「仮説提唱型のデータ利活用アプローチ」に基づいて、仮説検証の計画を策定することもより容易になる。社内で資産化されたデータから、自分の計画や仮説の何が検証でき、何が足りないのかを判断し、必要に応じて追加でデータを探していく、といったアクションを検討・実行できるためである。

上記のようにデータを検索するには、さまざまなアプローチがある。テキストデータであれば、その中身も含めて検索ができるSQL機能を含んだデータプラットフォームを導入することもできるが、画像データや音声など、テキストデータ以外のものに関しては、データの中身をテキストベースで検索することは難しい。よって、重要そうなデータ属性（ある程度の患者背景や治療歴、アウトカムなど）をテキストのメタデータとして付加し、カタログ化しておかないと、そのデータは有益な財産とはならず、データの海に埋もれてしまう。

また、そもそも何をキーワードとして検索したらいいか分からない、もしくは仮説検証のためでなく、ただ自社に蓄積されたデータ資産の全体像を把握したい、といったことも考えられる。そのために、データをどこかのプラットフォームに保存したら、データカタログに登録しておくべきである。データカタログの粒度やデータの分類法は会社それぞれであるが、適切なカタログのあり方は会社の状況によって変化するため、定期的にこれを見直し、古いデータは削除するか、アーカイブしておくなど、取り扱い方法をあらかじめ決めておくことも推奨される。

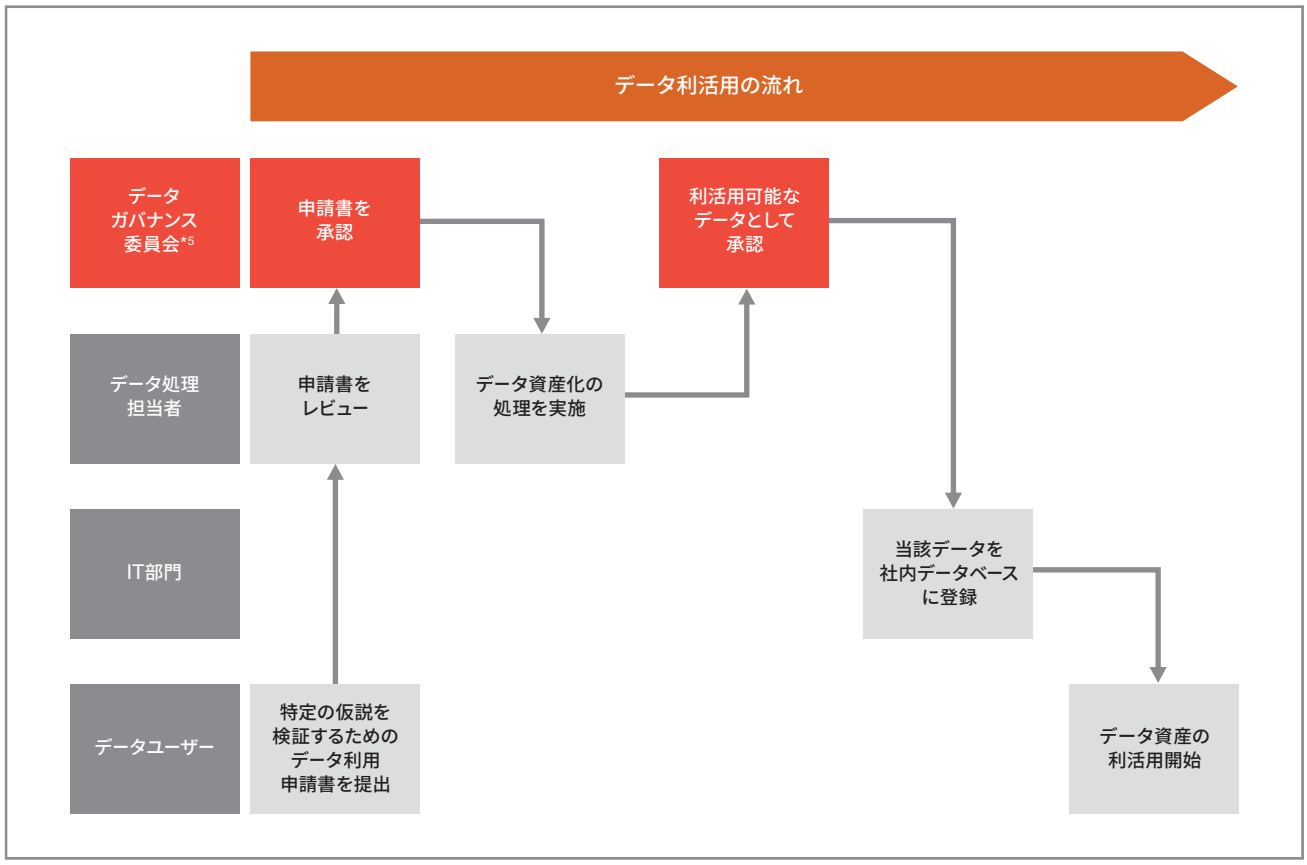
6 データ資産の利活用

データ資産の利活用を適切・積極的に進めるには、安全性と効率性をバランスよく兼ね備えた運用モデルを整備する必要がある。明文化されたデータガバナンスが、その基盤となる

最後に、データカタログからデータ特定し、活用する際には、あらかじめ定められたデータガバナンスと特定のオペレーションフローに沿って進めることが重要である。ここでは社内の部門Aが保持しているデータを別部門Bが利活用するケースを一例として紹介する。例えば、臨床開発部門の責任者がデータ管理者となっている臨床開発データを、基礎研究部門が使用する場合などがこれにあたる。この場合、新たなオペレーションフローの確立が必要となる。まず、データユーザーが明確な仮説検証計画に基づくデータアクセスの申請書を作成する。そして、データ管理者や法務・コンプライアンス部門など、ビジネスリスクに関して責任を負う者がこれをレビューし、データアクセスなどの重要なアクションに対して意思決定をするという流れである。意思決定の内容によっては、専門的な知識を提供する者を加えたチームから成る「データガバナンス委員会」を形成し、意思決定を実行していくことが多い（図表5）。

もちろん実際の運用においては、部門内での二次利用は部門長の承認、より複雑な場合は複数のプロセスを組み合わせるなど、最適なオペレーションフローは会社の組織構造や組織の機能によって異なるが、設計する上で極めて重要なことは、これを不必要に複雑にしないことである。意思決定のポイントが多く、データアクセスを申請してからアクセスできるようになるまで何カ月もかかるといった状況では、ユーザーにとって大きな負担となり、会社全体のデータ利活用が遅々として進まない。よって、安全かつ迅速なオペレーションフローを議論していき、会社のニーズに応じて変更していくことが求められる。

図表5：データ利活用における意思決定のオペレーションフローの例



*5. データガバナンス委員会はデータオーナーやデータ利活用に対してのリスクの責任を負う者が意思決定を行う会議体である。ただし、データに対する専門家などが同席し、意見を提供する場合が多い

おわりに

以上、製薬企業のR&Dが、よりデータ利活用に基づく（Data-driven）研究開発を推進していく上で、競争に打ち勝つためのアプローチについて私たちの考えを述べた。6つの鍵のどこに、どのような課題があるかは、会社によって異なるため、自社の課題を洗い出した上でこれに対処することは極めて重要である。PwCコンサルティングでは、DoubleJump Health™をはじめとするさまざまなソリューションをもとに、製薬企業のデータ戦略・デジタル戦略の策定と実行を支援している。例えばDoubleJump Health™は、業界のベストプラクティス、クライアント独特のビジネス状況およびPwCの豊富な実績に基づき、6つの鍵によってクライアントのケイパビリティを格段に強化することを目的としている。PwCが持つ業界屈指のノウハウと経験をもって、製薬企業のビジネスに貢献することが、PwCの使命だと考えている。

著者

クリストファー アルバーニ
Christopher Albani
PwCコンサルティング合同会社
パートナー

船渡 甲太郎（医師）
Kotaro Funato, M.D., Ph.D.
PwCコンサルティング合同会社
マネージングディレクター

宋 云柯（理学博士）
Yunke Song, Ph.D.
PwCコンサルティング合同会社
シニアマネージャー

お問い合わせ先

PwC Japanグループ
https://www.pwc.com/jp/ja/contact.html

www.pwc.com/jp

電子版はこちらからダウンロードできます。www.pwc.com/jp/ja/knowledge/thoughtleadership.html
発刊年月：2020年12月　管理番号：I202007-08