FDA が提唱する
「AI モデルのリスクベース Credibility 評価」
フレームワークの概説

日本製薬工業協会　医薬品評価委員会

電子化情報部会 タスクフォース 4

2026 年 3 月 30 日

【免責事項】
本資料の記載内容は、現時点の情報に基づき記載しています。本資料を利用した結果生じた損害について、日本製薬工業協会は一切責任を負いません。

＜改定履歴＞

版数	発行日	改定理由
1.0	2026年3月30日	初版作成

表 1：用語の定義

用語	定義
AI Model（AI モデル）	データからパターン（入力と出力の関係性）を学習し、パラメータ（重み）を最適化する数学的アルゴリズム。より広範な IT システム（コアシステム）に組み込まれる「AI サブシステム」または「コンポーネント」として位置づけられる場合もある。
AI System（AI システム）	様々なレベルの自律性で動作する機械学習ベースのシステム。明示的または暗黙的な目的のために、受け取った入力から推論し、物理的または仮想的な環境に影響を与えうる予測、コンテンツ、推奨、決定等の出力を生成する。
Algorithm（アルゴリズム）	データからパターン（入力と出力の関係性）を学習し、問題を解決するための数学的な手順や計算式。AI モデルを生成するための「学習方法（例：ランダムフォレスト、ニューラルネットワーク）」そのものを指す。
Architecture（アーキテクチャ）	AI モデルの内部構造および設計仕様。ニューラルネットワークにおける層の数や結合方法、決定木の深さ等、モデルがどのように計算・推論を行うかを規定する構造的枠組みを指す。
Calibration（キャリブレーション）	AI の文脈において、訓練済みモデルが出力する値（Output）に対し、精度や再現性を向上させることを目的として行う微調整・補正のプロセス。なお、GxP の文脈における計測機器の「校正（標準機との比較・調整）」とは意味が異なる点に留意。
Credibility (信用度)	特定の利用状況（COU）において、AI モデルの性能が意図した目的に対して信頼できる度合い。従来のシステム開発における「バリデーション」とは異なり、AI モデルに固有の開発手法や不確実性を考慮した概念として、FDA が提唱している。

用語	定義
Credibility Assessment（Credibility 評価）	特定の利用状況（COU）において、AI モデルの Credibility を確立するために実施される一連の活動。モデルの開発プロセス、性能評価、不確実性の定量化、およびリスクへの影響度を体系的に評価することで、当該 COU に対するモデルの fit for use を判断する（広義）。本書では、狭義として、選抜されたモデルが意図した性能を満たすことを確認するテスト活動を指し、従来の CSV フレームワークにおけるテスト（または UAT）に相当するものとして用いる。
Context of Use (COU) / 利用状況	AI モデルの特定の役割と適用範囲を規定したもの。AI モデルが「何を」「どのように」行うのかを明示し、AI モデルのリスクと求められる Credibility のレベルを決定する上での基礎となる。
Data Drift (データドリフト)	時間の経過や環境の変化により、AI モデルの学習時に使用したデータと運用環境で入力されるデータの統計的特性が乖離していく現象、または入力と出力の関係性そのものが変化する現象（コンセプトドリフト）。Model Drift を伴う AI モデルに特有の性能低下の主要因となり得るため、継続的なモニタリングにより早期検知することが必要となる。
Data Integrity（データインテグリティ）	GxP 活動において取得、分析、保管および報告されるデータが信頼できることを確保するための要件の体系。ALCOA++原則に基づき、データの取り扱い、識別・アクセス管理、監査証跡、電子署名、およびセキュリティに関する要件を含む広範な事項を包含する。
Data Leakage (データリーク)	本来は評価専用であるべきテストデータが、意図せずモデルの開発やチューニングに使われてしまうこと。FDA AI ガイダンスおよび Annex 22 では、訓練・チューニング・テストの各データセット間の Data Independence を重視しており、Data Leakage はテストデータの独立性の欠如として重大なリスクと捉えられる。
Decision Consequence (DC) / 決定の結果	AI モデルの出力が誤っていた場合に、後続の規制上の意思決定、患者さんの安全、または製品品質に与える結果の重大性。モデルリスクを構成する 2 つの要素のうちの 1 つ。
Explainability (説明可能性)	AI モデル（特にブラックボックスモデル）の出力や判断根拠を、人間が理解・解釈できる度合い。GxP 上の重要な意思決定を支援する場合、なぜその結論に至ったかを説明できることは、信頼性を担保する上で極めて重要となる。
Feature（特徴量）	モデルを学習（訓練）する際に入力として用いられるデータ属性。表形式のデータでいえば、列（身長、体重、性別など）に該当する。また、その列に入っている値が特徴量値と呼ばれる。

用語	定義
Fit for Use (用途への適合性)	データが意図した目的に対して関連性（Relevant）があり、信頼できる（Reliable：正確、完全、追跡可能）状態であること。AI モデルの性能は学習データの質に大きく依存するため、データが Fit for Use であることが大前提となる。また、モデルの性能が意図する COU に対して十分かつ適切であると判断された状態も指す。
Grid Search （グリッドサーチ）	AI モデル（主に機械学習モデル）のハイパーパラメータの最適な組み合わせを見つけ出す手法のひとつ。事前に設定したハイパーパラメータ候補とその範囲の組み合わせを総当たりで試行し、最も性能の良い組み合わせを選択するプロセス。
Human in the loop (HITL)	AI システムの出力を人間（専門家等）が確認・検証し、最終的な意思決定に関与するプロセス。AI の誤りを補正し、ブラックボックス性等のリスクを低減するための重要な管理策。
Intended Use (使用目的、意図した用途)	システムやプロセスが「何のために使われるのか」という目的。GAMP®5 第 2 版におけるバリデーションの基本原則。AI システムの文脈では、これをさらに具体化した「Question of Interest (QoI)」や「Context of Use (COU)」として定義することが求められる。
Life Cycle Maintenance (ライフサイクルの維持)	AI モデルがその利用状況（COU）において「Fit for Use」であり続けることを保証するため、意図的または偶発的な変更を管理する一連の計画された活動。モデルの性能と適合性をライフサイクルを通じてモニタリング・維持すること。
Machine Learning (ML) / 機械学習	AI の技術体系の一分野であり、明示的なプログラミングによらず、計算プロセスを通じてモデルパラメータを最適化することで、データに基づいてタスクの性能を向上させるようにアルゴリズムを訓練するものをいう。
Model Influence （MI) / モデルの影響度	規制上の意思決定プロセス全体において、AI モデルの出力（証拠）が、他の証拠と比較して、どの程度その決定に影響を与えるかの大きさ。モデルリスクを構成する 2 つの要素のうちの 1 つ。
Model Risk (モデルリスク)	AI モデルの出力が誤った決定につながり、患者さんの安全や製品品質に有害な結果（Adverse Outcome）をもたらす可能性。FDA は「Model Influence（MI）」と「Decision Consequence（DC）」の組み合わせで評価することを提唱している。
Multiple Models （複数モデル）	AI モデルの開発段階（探索的開発）において、最適な性能を得るために並行して作成される複数のモデル候補。異なるアルゴリズム、アーキテクチャ、ハイパーパラメータの組み合わせによって生成され、比較検討の対象となるもの。

用語	定義
Overfitting（過学習）	AI モデルが開発データに過剰に適合してしまい、その結果、学習に使用していない未知のデータに対する予測性能（汎化性能）が低下してしまう状態。モデルの信頼性を損なう主要なリスクの一つ。
Question of Interest (QoI) / 関心のある質問	AI モデルを用いて解決したい、または回答を得たい、具体的かつ明確に定義された課題や問い。AI モデルの評価範囲とゴールを定める出発点となる。
Risk-Based Approach (リスクベースアプローチ)	患者さんの安全、製品の品質、データインテグリティへのリスクに基づいて、システムライフサイクル活動の厳密さと範囲を決定するアプローチ。GAMP®5 の基本原則であり、AI の保証においても中心的な考え方となる。
Selected Model（選抜されたモデル）	複数モデルの中から、チューニングデータを用いた評価により、COU（利用状況）を満たす最良の性能を持つと判断され、最終的なテスト（Credibility 評価）の対象として選ばれたモデル。 このモデルの「仕様」が COU を満たすものとして、最終的に採用される。
Static / Dynamic / Adaptive Model（静的/動的/適応型モデル)	**静的モデル**は、一度開発されるとパラメータが固定（凍結）され、運用中に新しいデータを取得した場合でもパラメータを変更しないモデル。モデルの再学習は、計画的な変更管理プロセスを経たバージョンアップの一環としてのみ行う。 **動的モデル**は、内部状態（Input 以外の変数）が時間とともに変化することを前提としたモデル。時系列や連続的な意思決定を扱う場面で用いられる。パラメータは固定でも、内部状態が推論のたびに更新される場合は動的モデルとみなす。 **適応型モデル**は、モデルの運用中に、新たなデータや環境の変化に応じてパラメータを自動的に更新するモデル。静的／動的モデルとも組み合わさり得る。 Annex 22 は、静的かつ決定論的な AI モデルのみを適用範囲として想定しており、運用中に継続的かつ自動的に学習・更新を行う動的／適応型モデルや、確率論的な出力を持つモデルは文書の適用範囲外とされている。本書では、動的とはモデルの内部状態が時間とともに変化することを指し、適応とは運用中にモデルのパラメータが更新されることを指す。これら動的／適応型モデルについて Annex 22 は、患者さんの安全、製品の品質、データインテグリティに直接影響するクリティカルな GMP アプリケーションでは使用すべきではない（should not be used）と明示して

用語	定義
	おり、その結果、クリティカルな GMP 用途で利用可能な AI モデルは実質的に静的かつ決定論的なモデルに限定される。
Test（テスト）	訓練済み AI モデルに対し、独立したテストデータを用いてその性能（汎化性能）を客観的に評価する活動。FDA の Credibility 評価フレームワークにおける「Step5」等の実活動を指す。
Test Data（テストデータ）	開発およびチューニングが完了した AI モデルの最終的な性能を、客観的に評価するために使用する、完全に独立したデータセット（「ホールドアウトデータ」とも呼ぶ）。
Training Data（訓練データ）	AI モデルを構築・学習させるために使用するデータセット。モデルの性能の基礎を決定づけるため、その品質と代表性が極めて重要となる。
Tuning Data / Validation Data（チューニング/検証データ）	訓練済みモデルの「評価を行い、最適なハイパーパラメータやモデルアーキテクチャを選定（探索）する」ために使用するデータセット。FDA のチューニングデータは、Annex 22 や GAMP®5 ではバリデーションデータ（Validation Dataset）と呼ぶ。なお、一般的な機械学習の文献でも「バリデーションセット（Validation set）」は同義で、モデル選択やハイパーパラメータ調整に使用するデータを指す。本書では、GxP 文脈での「バリデーション（システム検証活動）」との混同を防ぐため、FDA に倣い、このデータセットを「チューニングデータ」と統一して表記する。
Validation（バリデーション）	コンピュータ化システム全体が、予め定められた要件や仕様（意図した用途：Intended Use）を満たしていることを検証し、文書化する一連のプロセス。本書においては、AI モデル単体の性能評価（Credibility 評価）だけでなく、AI モデルを組み込んだシステム全体（コアシステム＋AI サブシステム）の適合性確認を指す上位概念として用いる。
Verification（検証）	仕様や要件に対して、作成されたもの（この場合は AI モデルやシステムコンポーネント）が正しく作られているかを確認する活動。本書においては、AI モデルの性能確認活動全般（評価、テスト、検証）を包括する用語として扱われるが、実務的な文脈では「テスト（Test）」と同義として扱う。

注：本用語集の定義は、FDA AI ガイダンス、Annex 22、GAMP® 5 等の国際ガイダンスの理解促進を優先し、AI に関する数理的・技術的な概念は、基本的かつ限定的な説明を採用している。各社の社内規程における定義を優先すること。

表 2：略語の一覧

略語	正式名称
AI	Artificial Intelligence
CAPA	Corrective and Preventive Action
COU	Context of Use
CSA	Computer Software Assurance
CSV	Computerized System Validation
DC	Decision Consequence
DWH	Data Warehouse
EMA	European Medicines Agency
EC	European Commission
FDA	Food and Drug Administration
GAMP®	Good Automated Manufacturing Practice
GCP	Good Clinical Practice
GMP	Good Manufacturing Practice
GVP	Good Vigilance Practice
GxP	Good x Practice
HITL	Human in the Loop
LLM	Large Language Model
MI	Model Influence
ML	Machine Learning
PoC	Proof of Concept
QMS	Quality Management System
QoI	Question of Interest
SaMD	Software as a Medical Device
SME	Subject Matter Expert

表 3：商標の一覧

商標または登録商標	商標権者
GAMP®	International Society for Pharmaceutical Engineering (ISPE)
Docker®	Docker, Inc.
TensorFlow®	Google LLC

本書に記載されているその他の会社名、製品名等は、各社の商標または登録商標である場合がある。

目次

1. はじめに

近年、AI 技術は目覚ましい進化を遂げており、その応用範囲は製薬業界においても急速に拡大している。創薬研究から臨床開発、製造、市販後安全性監視に至るまで、AI は医薬品のライフサイクル全体において革新的な価値をもたらす可能性を秘めている。一方で、世界的な医療ニーズの多様化と高度化、さらにはパンデミックのような予期せぬ危機への対応を背景に、より迅速かつ効率的な医薬品開発の実現が求められている。このような状況において、AI の活用は単なる選択肢にとどまらず、製薬企業が社会的責任を果たし、患者さんの QOL 向上に貢献するために必須の取り組みとなりつつある。

しかしながら、製薬業界における AI 導入には固有の課題が存在する。医薬品の開発・製造においては、患者さんの安全、製品の品質、そしてデータインテグリティの確保が絶対的な前提条件であり、これらは GCP、GMP、GVP といった厳格な規制要件によって担保されている。AI をこれらの GxP 領域に導入する場合、従来のコンピュータ化システムとは異なる性質――データ駆動型の学習プロセス、複雑な挙動、ブラックボックス性――を持つ AI に対して、どのように品質を保証し、信頼性を確立すればよいのか、その具体的な方法論は未だ確立されていない。

このような背景のもと、米国食品医薬品局（以下、FDA）は 2025 年 1 月に「Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry and Other Interested Parties」のドラフトガイダンス（以下、FDA AI ガイダンス）を発出し、AI モデルのリスクベース Credibility 評価フレームワークを提唱した。また、欧州委員会（European Commission、以下、EC）も 2025 年 7 月に「EU GMP Annex 22: Artificial Intelligence」（以下、Annex 22）のドラフトを公開し、より具体的な技術要件を示している。これらの規制動向は、製薬企業に対して AI 品質保証の新たなパラダイムへの対応を求めるものであり、業界全体としての理解促進と実践的な対応が急務となっている。

本書は、このような規制環境の変化を踏まえ、製薬企業において AI の導入・開発・運用に携わる実務者の方々、さらには AI 技術の専門家ではない方々も含めた幅広い関係者に対して、AI バリデーションの基本的な考え方と実践的な指針を提供することを目的として執筆された。本書は FDA が提唱する 7-Step のリスクベース Credibility 評価フレームワークを軸としながら、Annex 22、日本の経済産業省が発行した AI 事業者ガイドライン、そして国際製薬技術協会（“International Society for Pharmaceutical Engineering, Inc.”：以下、ISPE）が発行した「GAMP® 5 -A Risk-Based Approach to Compliant GxP Computerized Systems 2nd Edition（2022 年 7 月）」および「GAMP® 5 コンピュータ化システムの GxP 適合へのリスクベースアプローチ第 2 版（2025 年 4 月）」（以下、GAMP 5 第 2 版）といった国際的な標準との整合性を図り、AI バリデーションの全体像と各フェーズにおける具体的な品質保証活動を解説している。

本書の執筆にあたり、我々は「AI 技術の進化は日進月歩であり、規制要件も現在進行形で変化している」という現実を常に念頭に置いてきた。本書で参照している FDA のガイダンスおよび Annex22 はいずれもドラフト段階であり、今後の改訂や追加的なガイダンスの発出が予想される。したがって、本書は最終的な「答え」を提供するものではなく、むしろ製薬企業の皆様が自社の状況に応じて AI 品質保証のアプローチを構築する際の「思考のフレームワーク」として活用していただくことを意図している。重要なのは、「なぜこの AI を使うのか」「AI が間違った場合のリスクは何か」「そのリスクをどのように管理するのか」といったクリティカルシンキング（批判的思考）を実践し続けることである。

日本製薬工業協会の電子化情報部会タスクフォース 4 は、今後も規制動向の変化を注視し、業界内での知見の共有と議論を継続していく所存である。本解説書とケーススタディ集が、製薬企業の実務担当者の理解促進と実践的な AI 品質保証活動の一助となり、各社の Quality Management System （以下、QMS）の中で柔軟に活用されることで、医薬品開発のイノベーション加速と、最終的には患者さんへのより良い医療の提供に貢献できることを心より願っている。

2026 年 3 月

日本製薬工業協会　医薬品評価委員会
電子化情報部会　タスクフォース 4

2. 本書の趣旨と対象とする AI モデル

2.1 対象とする読者

本書は、AI 技術の進展並びにそれに伴う製薬規制当局の動向を整理し、製薬企業における AI システム活用の実務的指針を提供することを目的とする。

本書が想定する読者は、製薬企業において AI システムを利用する「AI 利用者」と、AI の導入・開発およびバリデーション活動（CSV 活動）に携わる「AI バリデーション実務者」である。前者に対しては AI システムの基本的理解を促進する概説的な解説を提供することを意図し、後者に対しては AI システム導入および運用に係る CSV 等の実務に資する手引きを示すことを目的とする。これにより、概説的内容と実務者向け手引きを一元的に包含する参照書となることを目指す。

FDA AI ガイダンスは本来、FDA の審査要件である「Credibility 評価」を示したものである。しかし、本書ではその枠を超え、医薬品ライフサイクル全体で AI 技術を活用するすべての人々に役立つ実践的な考え方を提言することも目的とする。

2.2 対象とする AI モデル

本書は、FDA AI ガイダンスが対象とする「規制上の意思決定支援ツール」に対しては当該ガイダンスの枠組みに準拠する一方、それ以外のユースケースにも同じ考え方を任意に拡張して適用することを意図している。本書で示す内容のうち、「FDA AI ガイダンスとしての明示的な要求」と「本書としての実務的な拡張・解釈」は明確に区別されるべきものであり、規制要件を上書きするものではない。

一般的に、AI 技術は以下のように分類される。

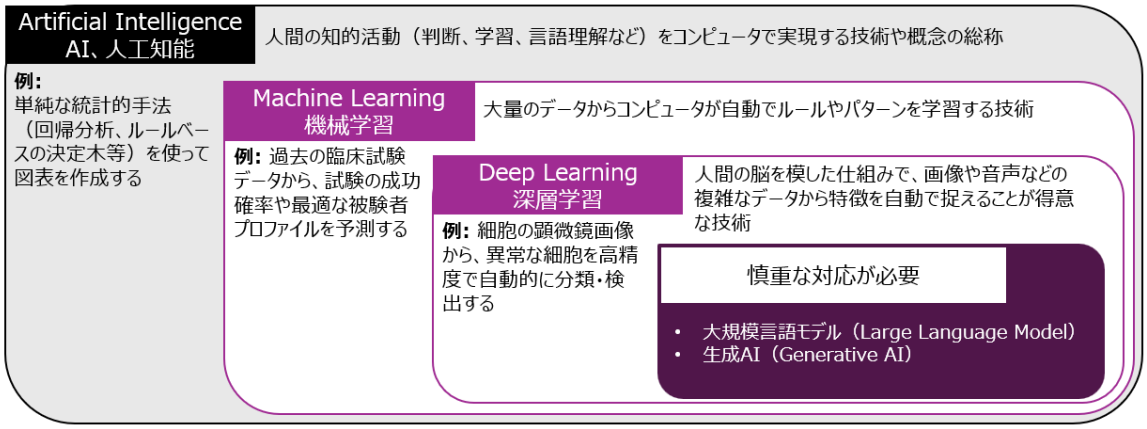

図 1：AI 分類

本書は、FDA AI ガイダンスが想定するデータからパラメータを学習する「機械学習（Machine Learning、以下、ML）モデル」を主な対象とし、以下も対象に含める。

- 機械学習（Machine Learning）モデル
- 深層学習（Deep Learning）は ML モデル（ML の一手法）
- 大規模言語モデル（Large Language Model、以下、LLM）
  - LLM は ML および深層学習に含まれるが、FDA AI ガイダンスは、主に従来型の機械学習モデル（予測、分類、検出等)を想定しており、LLM や生成 AI に特化した具体的な要件は含まれていない。そのため、LLM を規制上の意思決定の支援に用いる場合には、ブラックボックス性を考慮した上で、FDA AI ガイダンスの一般的なフレームワークを参考にしつつ、説明性・再現性・バージョン管理等の観点から慎重な検討を行い、当局との事前相談の機会をもつことが望ましい。
  - Annex22では、LLMや生成AIのような確率論的な出力を持つモデルについて、クリティカルなGMP分野での使用は、使用すべきではない（should not be used）と明示的に制限されている（6. 項参照）。

本書は、「データから学習する機械学習モデル」を総称して「AI システム」「AI モデル」と呼ぶ。この AI モデルを理解するために、仕組みと Output の観点から以下のように分類する。

- **仕組みの分類**
  - **静的モデル（static / locked）**
    モデルの運用中は学習済みパラメータが固定され、新しいデータを取得した場合でもパラメータを変更しないモデル。モデルの再学習は、計画的な変更管理プロセスを経たバージョンアップの一環としてのみ行う。
  - **動的モデル（dynamic）**
    モデルに「内部状態」が存在し、その状態が時間とともに変化することを前提とするモデル。時系列や連続的な意思決定を扱う場面で用いられる。パラメータは固定でも、内部状態が推論のたびに更新される場合は動的モデルとみなす。
  - **適応型モデル（adaptive）**
    モデルの運用中に、新たなデータや環境の変化に応じてパラメータを自動的に更新するモデル。静的／動的モデルとも組み合わさり得る。

- **Output の分類**
  - **決定論的モデル（deterministic）**
    同じ入力と同じ内部状態に対して、常に同じ出力を返すモデル。多くの AI モデルの推論プロセスは、この意味で決定論的に実装されている。

- **確率論的モデル（probabilistic / stochastic）**

  予測のばらつきを確率や分布として明示的に表現し、同じ入力・状態であっても出力が確率的に変動し得るモデルである。出力として確率や分布を返す場合も含まれる。例えば、確率論的モデルでは、Input データに応じて出力の確率分布を定め、その分布からサンプリングして Output を得る。

これらの仕組みと Output の分類は互いに独立し、例えば「静的かつ決定論的な AI モデル」「動的だが決定論的な AI モデル」「静的だが確率論的な AI モデル」「適応型かつ確率的な AI モデル」など、さまざまな組み合わせが存在する。

本書は、これらの組み合わせに関わらず「データからパラメータを学習する AI モデル」を広く対象とするが、実務的な有用性を考慮し「静的かつ決定論的モデル」を主な対象とする。一方で、動的・適応型 AI モデルは 6. 項における FDA AI ガイダンスと Annex 22 の比較の中で詳述する。

なお、AI モデル特有の議論に焦点を当てるため、以下の従来型手法は本書の対象外とする。ただし、高リスクな使用目的かつ規制上重要な判断に用いる場合には、本書の考え方を準用し得る。

- ルールベースのエキスパートシステム
- if-then ルール等のみから構成される単純な決定ロジック
- 従来から GxP 領域で広く用いられてきた単純な統計モデリング（例：説明可能な線形回帰モデル、事前に定義した閾値やスコアに基づく判定など）

3. AI システムの Credibility 評価フレームワークの基本原則と全体像（総論）

本項は、AI システムの Credibility 評価フレームワークの活動に初めて関わる担当者を対象に、なぜ AI に Credibility 評価フレームワークの活動が必要なのか、そしてその活動がどのような流れで進められるのか、その全体像と基本的な考え方は何か、を FDA AI ガイダンスに準拠し概説する。また、Annex 22、および、経産省の AI 事業者ガイドライン（以下、AI 事業者ガイドライン）も参照する。さらに、ISPE の GAMP 5 第 2 版のコンセプトとの整合性も考慮する。本項は、5. 項以降で詳述する具体的なプロセスを理解するための基礎となる。

3.1 Credibility 評価フレームワークの活動対象となる AI モデル

最初に FDA AI ガイダンスの対象を特定する。図 2 は、FDA AI ガイダンスから対象の AI モデルを判定するために解釈、図示した。

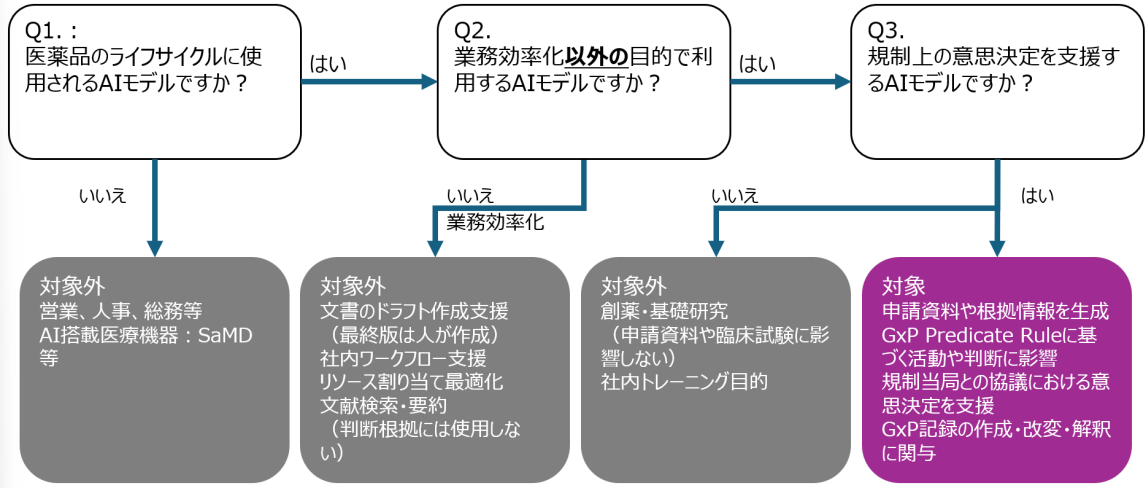

図 2：FDA AI ガイダンス適用判定フロー

FDA AI ガイダンスは、AI モデルの出力が医薬品のライフサイクルに関する FDA と製薬企業の両者、またはどちらかの規制上の判断を「支援する」目的で用いられる場合を対象とする。一方で、創薬や基礎研究段階にとどまる利用や、単なる業務効率化ツール等、規制判断や GxP 記録の信頼性に影響しない用途は対象外とされている。その判定のために、FDA AI ガイダンスに従い、下記の 3 つの質問からなるフローに落とし込んだ。

Q1：医薬品のライフサイクルに使用されるか

最初に、AI モデルが扱うデータが、医薬品の開発・製造・市販後安全対策等、医薬品ライフサイクルに利用するかどうかを確認する。営業、人事等医薬品と無関係なビジネス効率化ツールは、この段階でガイダンスの適用対象外と判断する。また、SaMD 等の医療機器は、FDA AI ガイダンスではなく医療機器関連ガイダンスの枠組みが適用されるため、対象外とする。

Q2：業務効率化以外の目的で利用されるか

AI モデルが GxP 文書のドラフト作成、社内ワークフロー支援、リソース割り当て等、業務効率化目的である場合は、対象外とする。

Q3：規制上の意思決定を支援するか

非臨床試験、臨床試験、製造、市販後安全管理等実際の医薬品ライフサイクルにおける FDA と製薬企業の両者、またはどちらかの規制上の判断を支援するために用いられるかを確認する。

この質問に「いいえ」となる場合、対象外となる。例えば、創薬・基礎研究、GxP 以外の社内トレーニング、文献調査等が該当する。

この質問に「はい」となる AI モデルが、FDA AI ガイダンスの対象となる「規制に関する意思決定支援ツール（Regulatory Decision Support Tool）」として、リスクベースの Credibility 評価やライフサイクル管理の対象となる。ただし、その厳密さは AI モデルが患者安全・製品品質に与えるリスクに応じて調整できる。 ただし、PQS（医薬品品質システム）業務を支援するモデルなど、直接的なリスクが低いモデルについては、5.3 項で述べるリスク評価に基づき、Credibility 評価および運用中の活動の厳密さを調整することができる。

上記「規制に関する意思決定支援ツール」を対象に、3.2 項より、AI モデルの Credibility 評価フレームワークに関する解説を行う。本解説は、ライフサイクルアプローチ、リスクベースアプローチ、及びデータインテグリティの確保を基本原則とする。

3.2 AI システムの Credibility 評価フレームワークの活動の必要性

医薬品開発や製造において AI システムを利用する場合、その判断結果が患者さんの安全や製品の品質、結果の信頼性に直接影響を及ぼす可能性がある。そのため、使用目的通りに、AI システムが安全かつ有効に機能することをエビデンスベースで保証する必要がある。これが AI システムにおける Credibility 評価フレームワークの活動の目的である。

CSV に代表されるようなバリデーションではなく、AI システムに特有の Credibility 評価フレームワークの活動が必要な背景には、AI モデル特有のリスクと従来のソフトウェアとの根本的な違いがある。

1) AI モデルがもたらす特有のリスク

AI モデルは、従来のソフトウェアにはない以下のようなリスクを内包する。

- **バイアス（偏り）のリスク：**AI モデルの開発に使うデータにバイアスがある場合、AI モデルが出力する結果にも偏りが生じ、特定の集団に対して不正確な結果を出力するリスクがある。
- **透明性の欠如によるリスク：**AI モデルがなぜその結論に至ったのか、その判断根拠を人間が完全に理解できない「ブラックボックス」状態に陥ることがある。予期せぬエラーが発生した際に、原因究明や対策が困難になる。
- **正確性が曖昧になるリスク：**AI モデルが出力した結果が正しいことを解釈、説明、または定量化することが困難になる。
- **性能低下のリスク：**運用を開始した後、実際のデータ（例：製造設備、患者背景）の統計的な分布が変化するデータドリフト（コバリエートシフト）や、入力と出力の関係性そのものの変化（コンセプトドリフト）が生じるリスクがある。本書ではこれらを総称してデータドリフトと表記するが、原因分析や対策検討の際には両者を区別して評価することが重要である。

2) 従来のソフトウェアとの違い

従来のソフトウェアは、人間が作成した明確な指示に従って、常に同じ結果を返すシステムである。一方、AI モデルは、大量のデータから法則性やパターンを自ら学習して判断ルールを構築する。このため、AI モデルの挙動は以下の特徴を持つ。

- **データへの強い依存性：**AI モデルの性能は、学習に用いるデータの質と量に依存し、事前に定義されたロジックに依存しない。不適切なデータで学習すれば、AI モデルの出力も不適切なものとなる。
- **複雑性と解釈困難性：**学習により獲得したパラメータ（重み）の組み合わせは膨大かつ複雑であり、人間がその全てを予測・把握することが困難な場合が多い（いわゆる「ブラックボックス」問題)。このため、結果だけでなく、学習プロセスの妥当性、学習していない実環境で初めて遭遇するパターンのデータへの頑健性も求められる。

この「データに依存し、自ら学習する」という性質が、従来の仕様どおりの動作を個々のテストケースで確認するだけのアプローチでは、品質を十分に保証できない理由である。このため、FDA AI ガイダンスでは、システムやデータに「Validation（バリデーション)」や「Reliability（信頼性)」を使っているが、AI モデルには「Credibility（信用度)」という新しい概念を取り入れていることからも、FDA は、AI モデルの特性を考慮したうえで、信用できる状態を継続的に確保するための一連の活動を求めていると解釈できる。

3.3 AI システムのライフサイクル

FDA AI ガイダンスが提唱する AI モデルの Credibility（信用度）を正しく理解するためには、その前提として、AI モデルを包含するシステム全体のライフサイクルを理解しておく必要がある。GAMP 5 第 2 版等が示すコンピュータ化システムのライフサイクルの中で、AI モデルの位置づけを理解することで、効果的な活動が可能となる。そのため、本項ではまずシステム構成とライフサイクルの全体像について解説する。

AI システムの構想から開発、運用、そして廃棄に至るまでの一連の活動を「ライフサイクル」として捉え、継続的に AI モデルの性能を評価するアプローチが求められる。

このライフサイクルを適切に評価するためには、まず「AI システム」の構成を理解する。「AI システム」は多くの場合、「コアシステム」、そのサブシステムとして機能する「AI サブシステム」、そしてその上で機能する「AI モデル」という、性質の異なる 3 つの要素で構成されている。

3.3.1 システム構成：コアシステムと AI モデルの関係

最初のステップとして、対象となる AI システムの構成を理解することが極めて重要である。AI システムの構成は多様であり、以下のようなパターンが例として考えられる。

- AI モデルがコアシステムに統合される構成
- AI モデルが独立したサブシステムとして連携する構成
- クラウドベースの AI API サービスを利用する構成

本書では、GAMP 5 第 2 版を元に、コアシステム（GxP システム）と AI モデルが明確に分離され、AI サブシステムを介して連携する構成を例として示す（図 3 参照）。実際のシステムアーキテクチャは、各社の IT 環境、リスク評価、および技術的制約に応じて選択する。

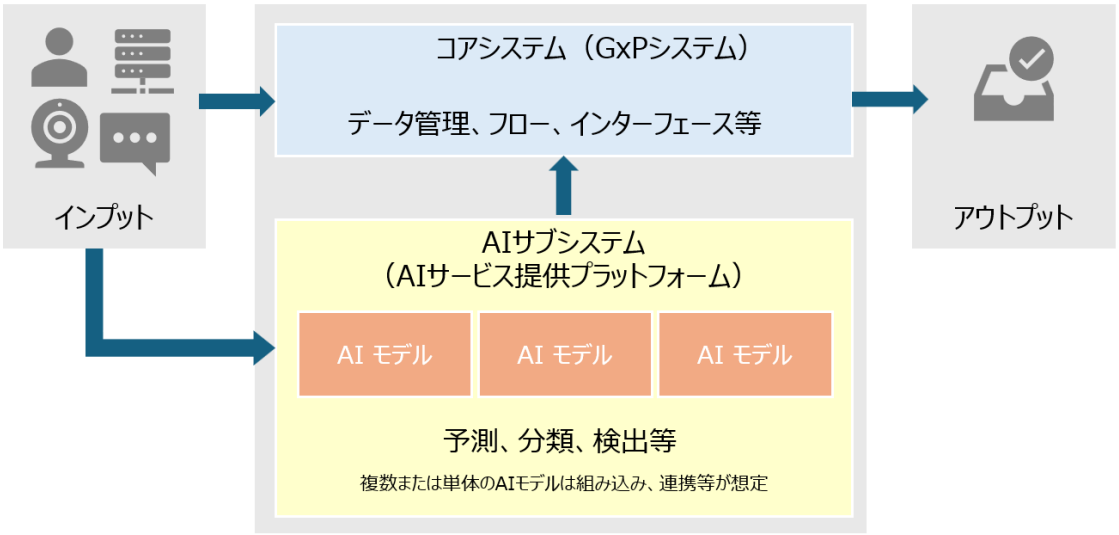

図 3：AI モデルを組み入れたシステムの構成例

- **コアシステム（GxP システム）：**ユーザが直接操作する、GxP 環境で稼働する主要なアプリケーションやプラットフォームを指すコアシステムである。例えば、製造実行システム（Manufacturing Execution System、MES）、臨床データ解析プラットフォーム等がこれにあたる。このコアシステムは、データの管理、業務ワークフロー、ユーザインターフェース、監査証跡といった機能を有する。従来の CSV の概念では、これが主たる「コンピュータ化システム」となる。
- **AI サブシステム（AI サービス提供プラットフォーム）：**単数、または複数の AI モデルを搭載し、コアシステムと連携するためのプラットフォームであり、搭載された AI モデルと合わせて、AI サブシステムという。コアシステムに AI モデルを組み込

む場合もあるが、本書では、コアシステムと AI モデルが AI サブシステムを通して連携するシステムを基本形として示す。

- **AI モデル：**予測、分類、検出といった特定の知的タスクを実行するために特化したコンポーネントである。通常、コアシステムから「呼び出される」形で機能する。AI モデルは、社内外のソースからデータを受け取って処理を行い、その結果をコアシステムに返す。例えば、「不良品／良品」の判定、将来のリスクスコア等が想定できる。

コアシステムと AI モデルの関係は、AI モデルが、コアシステムに統合された「プラグイン」や「サービス」のように機能すると考えると分かりやすい。つまり、コアシステムがバリデーションされた安定的な業務基盤を提供し、AI モデルが特定の高度な機能を提供する、という分業関係にある。AI モデルへのデータの流れは、コアシステムを経由する場合や、外部ソースから直接取り込む場合等多様なパターンが存在するが、いずれにせよシステム全体として機能することが重要である。

コアシステムが AI モデルとの間でデータを正しく送受信できること、そしてシステム全体として、本来の「使用目的」を確実に満たせることを検証する必要がある。

3.3.2 AI モデルのライフサイクルの 3 つのフェーズ

コアシステムと AI モデルでは、そのライフサイクルと管理アプローチが根本的に異なる。

- **コアシステムのライフサイクル：**従来のソフトウェア開発ライフサイクルに従う。そのバリデーションは、標準的な CSV に則り、業務と機能の要件、データインテグリティ、セキュリティ、監査証跡等に焦点を当てる。さらに、AI サブシステムとの接続、AI モデルの結果の出力およびエンドユーザが AI モデルの出力結果を理解したうえで、適切に規制上の判断を下せるプロセスであることを確認する。
- **AI サブシステムのライフサイクル：**基盤となるプラットフォームを提供する機能を有するため、バリデーションは、AI モデルが動作すること、AI モデルと接続することの検証、および構成管理に焦点を当てる。
- **AI モデルのライフサイクル：**基本的に動的かつ反復的である。実運用における性能低下（モデルドリフト）や、新しい訓練データの入手等に対応するため、利用環境に応じた再学習、更新、あるいはモデルの入れ替えが発生する。その Credibility 評価は、モデルの性能評価指標、データの品質に焦点を当てる。

AI システム全体のバリデーションでは、コアシステムと AI モデルがそれぞれ単独で正しく機能することに加え、「統合された状態で堅牢性を維持できること」を保証することが不可欠となる。つまり、コアシステムと AI モデルそれぞれに対してライフサイクルのバリデ

ーション活動があると想定する。ただし、AI モデルの開発は、従来のソフトウェア開発と本質的に異なる。AI モデル特有の開発方法の詳細は、4.2 項で述べる。

AI モデルのライフサイクルは、大きく構想フェーズ、プロジェクトフェーズ、運用フェーズの 3 つのフェーズで構成される。さらに、AI モデルではデータ管理が各フェーズの活動と関連付けられることから、AI システム全体のバリデーションの概要を図 4 に示した。

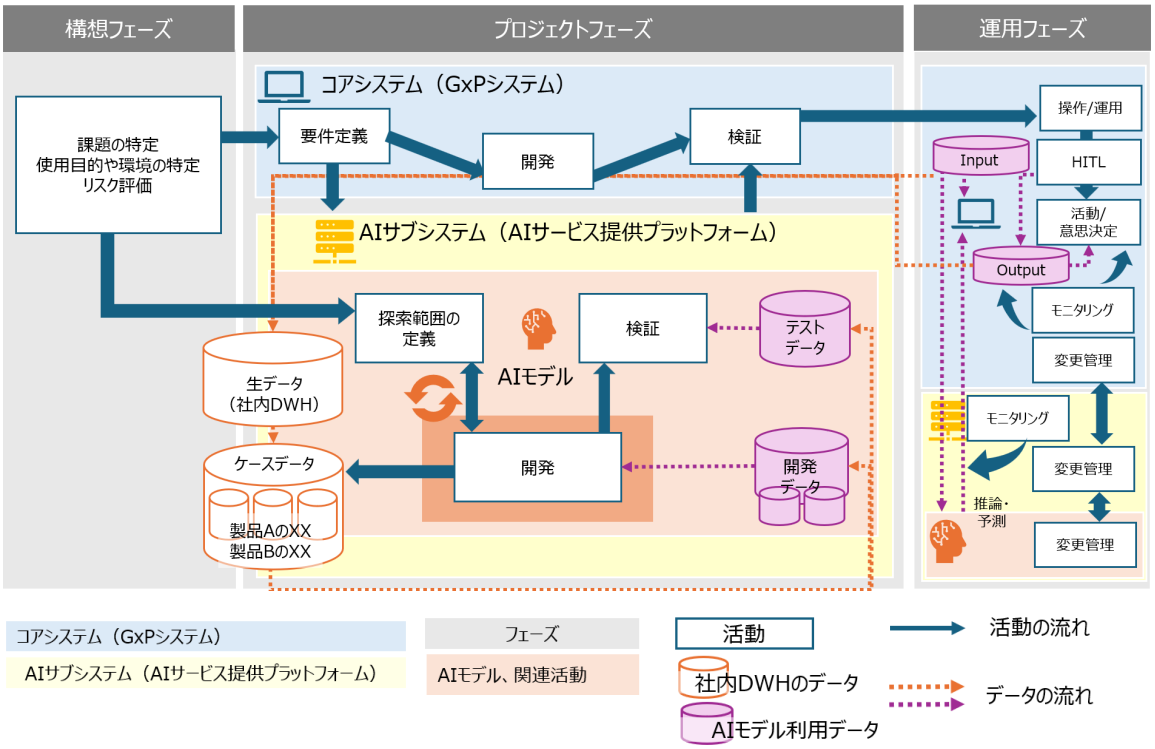

図 4：コアシステムと AI モデルのライフサイクル

注：Input と Output の入手、管理経路はコアシステムを含めた AI システム全体の構成によって異なる。以下に例を示す。

- コアシステムに Input、Output のデータベースが組み込まれる
- コアシステムとは別に、Input、Output はそれぞれ独立したデータベースをもつ
- Input は各種のデータを収集し、社内 DWH で統合、社内 DWH から入手する。Output も社内 DWH に格納される

GxP システムに AI サブシステムを組み込み、または、GxP システムが AI サブシステムを通して AI モデルと連携して Output を出力する場合の各フェーズの活動、データフローを説明する。以下、各フェーズの総論を述べる。

1)　構想フェーズ

AI バリデーションの出発点である。この段階では、「AI システムを使って何をしたいのか」という使用目的を明確にする。そのために、背景にある課題や使用する環境や制約条件等を定義していく。なお、現状の AI システムの結果は盲目的に信頼できるとは言えない場合もあるため、AI システムの結果にのみ依存するのではなく、専門家や関連情報を踏まえて、規制上の判断や次の活動に進むことも考慮する。本フェーズでは、既存のモデルやプロセスと比較しつつ、探索的に AI モデルの候補を開発する。そして、AI モデルの利用範囲や性能の概要をつかむ Proof of Concept（以下、PoC）も踏まえて、実現可能性を評価する。

さらに、AI システムを利用する業務の中での AI モデルからの Output が患者さんの安全、製品の有効性および品質に関する意思決定や判断に与える影響の大きさを分析し、潜在的なリスクを評価する。このフェーズでの定義と評価が、後続のすべての活動の質と方向性を決定づける。

2)　プロジェクトフェーズ

構想フェーズで定めた使用目的を達成するための、AI モデルの具体的な開発と検証を行う段階である。

まず、AI モデルの学習に用いるデータを準備する。次に、PoC での結果を参考にモデルアーキテクチャとハイパーパラメータ等の探索範囲を決定する。このプロセスでは、事前に詳細な実装仕様を確定するのではなく、指定した探索範囲の中で使用目的に最も適したモデルを選択する。そして、選抜されたモデルの性能評価を独立したデータセットで実施する。開発に使っていないデータを用いることで、モデルが実務において意図した性能を発揮できることを検証する（汎化性能の確認）。

この検証活動やユーザ受入試験（UAT、PQ 相当）に該当するテストのことを FDA AI ガイダンスでは「Credibility 評価」という。なお、本書では、「Credibility Assessment」を従来の CSV フレームワークにおけるテスト（または UAT）に相当するものとして位置づける。ただし、従来のソフトウェアテストとは異なり、Credibility Assessment は QoI・COU の妥当性、データ品質、訓練・チューニングプロセス、モデル性能評価指標など、AI モデルに固有のより広範な評価基準を包含する。この対応関係は、実務者が Credibility Assessment を GxP バリデーション業務に統合しやすくすることを目的として採用している。

AI モデルの開発で特徴的な活動のひとつは、開発やテストで使うデータを厳密に管理することである。これらのデータは、実際の環境、対象集団を代表していること、すなわち、実際に Input されるデータと高い類似性をもち、バイアスが少ないものが良い。また、準備したデータをどのように加工し、利用したのか、データの利用プロセスを説明で

きるように管理する。加工活動には、開発時の学習させるためのラベル付けとして、メタデータの付与、正解等の特徴の特定等の活動を含む。利用活動には、開発、検証等の目的に応じた分割や実際に特定のモデルやバージョンの開発や検証での利用を含む。

3) 運用フェーズ

検証した AI モデルを含むコアシステムを、実際の業務環境で利用する段階である。当初の構想フェーズで AI モデルの結果に加えて専門家の判断や関連情報を利用するプロセスを選択した場合、確実にそれらのプロセスが実行できる環境を整える。そのような人が積極的に最終的な意思決定に関与するプロセスを Human In The Loop（以下、HITL）という。

さらに、実運用が始まった後も、AI モデルの性能が維持されているかを継続的にモニタリングする必要がある。もし、時間の経過とともに環境変化や傾向が変化することによって、当初発揮していた性能が低下した場合、その原因を調査し、モデルの再学習や修正といった保守活動を行う。これにより、ライフサイクル全体を通じて AI システムの信頼性を維持する。

4. FDA ステップの全体像と機械学習のモデル開発手法

4.1 FDA ステップの全体像

本項では、FDA AI ガイダンスで提唱されたリスクベースの Credibility 評価フレームワーク（Risk-Based Credibility Assessment Framework）を基本骨格としながら、GAMP 5 第 2 版のライフサイクルアプローチを統合し、AI システムの構想から運用に至る包括的な Credibility 評価フレームワークの活動を 12-Step（Step 0 から Step 11）で解説する。

FDA AI ガイダンスの 7-Step のリスクベース Credibility 評価フレームワークは、主に AI モデルの Credibility 評価に焦点を当てているが、実際の GxP 環境で AI モデルを導入・運用するには、実現可能性の簡易的な検証である PoC（Step 0）、AI モデルとコアシステムとの統合（Step 8-9）およびコアシステムのリリースまでのバリデーション活動（Step10）、そして運用開始とライフサイクル維持（Step 11）といった一連の活動も不可欠である。本書では、FDA AI ガイダンスの 7-Step を拡張し、これらを総合して 12-Step の体系的なフレームワークを提示する。さらに、FDA AI ガイダンスのフレームワークと AI 事業者ガイドライン 別添（附属資料）の「図 2. AI の学習及び利用の流れの例」を統合し、データフローとの関連性を明確にした。図 5 として図 4 を詳細化し、12-Step の活動とデータの流れ、文書との関連性を示した。

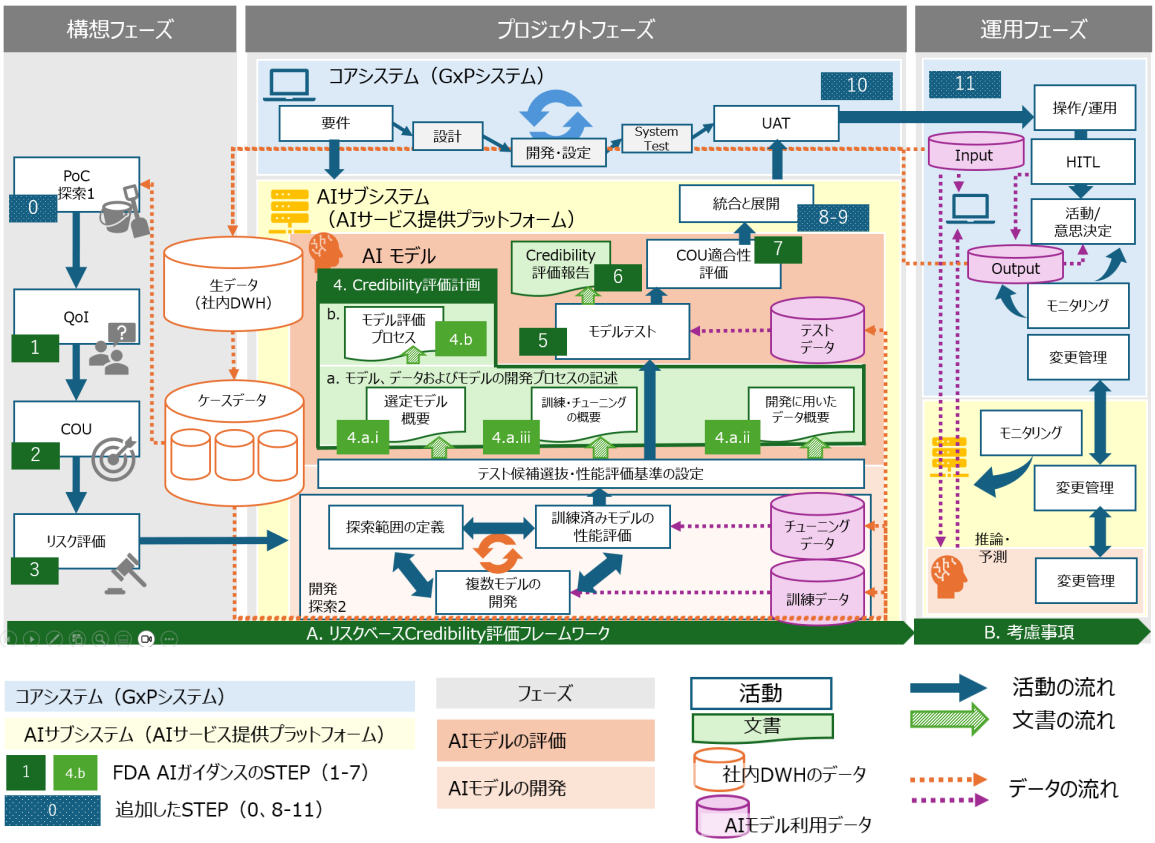

図 5：AI モデルのライフサイクルの活動

注：図 5 は FDA AI ガイダンスの 7-Step を 12-Step に拡張している。

図中の緑のボックスが、FDA AI ガイダンスの各 Step やセクションを示す。FDA AI ガイダンスは、リリースまでの構想フェーズ、プロジェクトフェーズの活動を「A. リスクベースの Credibility 評価フレームワーク」として 7-Step（Step 1-7）で示し、運用フェーズの活動を「B. 特別な考慮事項 (Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use)」で示している。

本書では、FDA AI ガイダンスの 7-Step に加えて、 GAMP 5 第 2 版等のライフサイクル概念を踏まえて PoC の実施（Step 0）、AI サブシステムの構築と検証およびコアシステムとの統合（Step 8-9）、コアシステムリリースまでのバリデーション活動（Step10）、そして運用保守（Step 11）を含めた全 12-Step の構成とした。Step11 は、リリース直後のハイパーケアと、その後の定常運用を包含する運用フェーズとして位置づける。

本書にて独自に追加した Step を青（ドット）のボックスで示す。さらに、FDA が求める文書やコンテンツを緑の書類図形で示した。

各 Step の詳細な活動は、5. 項で述べる。

4.2 本書で想定する AI モデル

本書は、FDA AI ガイダンスに基づき、規制上の意思決定を支援する AI モデルの Credibility を保証するための実務的な手法を提示することを目的とする。したがって、製薬企業が社内の業務データを用いて、小規模から中規模の AI モデルを開発することを前提とする。また、規制上の意思決定を支援する AI モデルには高い説明可能性が求められるため、XGBoost、ランダムフォレスト、ニューラルネットワーク（以下、NN）等の説明可能性と予測性能のバランスをもつ二値分類モデルを想定する。

加えて、図 5 に示すように、AI モデルの開発は探索的（exploratory）かつ反復的（iterative）である。したがって、モデルアーキテクチャ（アルゴリズム）の選定やハイパーパラメータ（訓練の仕方や加減を決める設定）の調整だけでなく、データの特徴量（例えば、表形式データにおける身長、体重、性別などの列にあたるデータ属性）の選定や変換、欠損値処理、スケーリング等のデータ前処理・データエンジニアリングについても、探索の概念に含まれている。

また、本書では、指定したモデルアーキテクチャに対して、あらかじめ定めたハイパーパラメータ範囲の中でグリッドサーチ（grid search）等の一般的な探索的手法を用い、事前に定義した性能評価指標（例： AUC、感度、F1 スコア等）に基づき、訓練データおよびチューニングデータから複数の候補モデルを開発し、その中から最良のモデルを最終テスト候補として選抜する、という開発の流れを想定する。なお、本書の内容は、グリッドサーチ以外のハイパーパラメータ最適化手法にも適用可能である（ランダムサーチ、ベイズ最適化、等）。

用語の使い方としては、アルゴリズムやニューラルネットワークの構造そのものを「モデルのアーキテクチャ（設計）」と呼び、データによって学習され具体的なパラメータが与えられた数式を「モデル（完成品）」と呼ぶ。また、探索手法のひとつであるグリッドサーチでは、同一のモデルアーキテクチャに対して異なるハイパーパラメータ設定で複数のモデル（候補）を開発し、その中から最良のモデルを選抜することを想定する。

4.3 AI モデルの開発手法

FDA AI ガイダンスの一連のステップのうち、特に Step4（Credibility 評価計画の立案）から Step5（テスト）に至るまでの過程が、多くの CSV 実務者を悩ませている。なぜなら、従来のコンピュータ化システムで一般的に用いられてきた V モデル型の直線的なシステム開発手順とは異なり、AI モデルの開発は探索的かつ反復的なプロセスを前提にしているからである。したがって、この反復的な開発手法の理解が不十分な場合、FDA AI ガイダンスに準じた信頼性のある AI モデルを開発・運用することは困難である。そこで本項では、FDA AI ガイダンスのフレームワークの Step4、5 のテストを理解するために必要な基礎知識として、Step4、5 に至るまでの AI モデルの一般的な開発プロセスを整理し、それらの関係性を図 6 に示す。

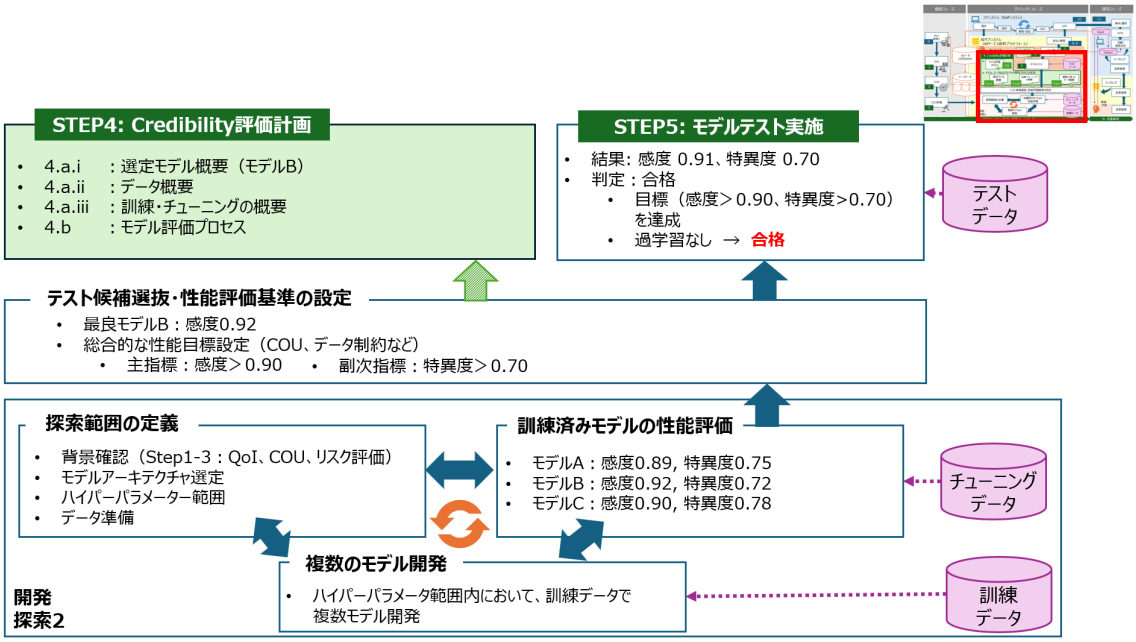

図 6：探索的開発とテスト（Step4、5）の関係

注：本図は図 5 の探索的開発（開発・探索 2）から Step5 の各活動を詳述したものである。

探索的モデル開発は、探索条件の範囲内で複数のモデルを作り、チューニングデータで各訓練済みモデルの性能を評価し、最良モデルを選抜するという流れとなる。もし COU を満たす性能に達していない場合は、モデルアーキテクチャの選定、ハイパーパラメータの見直しを含む探索範囲の定義から再度やり直す。この作業を反復的に繰り返し、COU を満たすモデルをテスト候補として選抜する。そして、その選抜されたモデルの概要（4.a.i）、開発に使ったデータの概要（4.a.ii）、モデルの訓練・チューニングの概要（4.a.iii）、モデルの評価プロセス（4.b）を Credibility 評価計画書に記載し、そのテストを実施する（Step 5）。

このような AI モデル開発のプロセスを理解する際の重要な原則は、以下の通りである。

- **探索的開発：**
  - 複数のモデルを並行して開発・比較する反復的プロセス
  - 訓練データで複数のモデルを開発し、それらの性能をチューニングデータで比較
  - 事前に設定したハイパーパラメータ候補とその範囲の組み合わせを網羅的に試行し複数のモデルを開発（グリッドサーチを想定）
  - 最良の候補を選定するための系統立てたな試行錯誤
- **データの分割と用途：**
  - **訓練データ：**複数のモデルを開発するために使用

- **チューニングデータ：** 訓練済みモデルを比較し最良モデルを選抜するために使用
- **テストデータ：** 選抜されたモデルの過学習を検証するために使用（汎化性能の確認）

- **従来のシステム開発手法との違い：**
  - AI モデルの開発プロセスは、従来の V モデルによる開発プロセスと本質的に異なる。具体的には、従来手法が「仕様を事前確定 → 開発 → 検証」という逐次的かつ直線的なプロセスであるのに対し、AI モデル開発は「探索範囲を定義 → 複数モデルを開発・評価 → 最良を選抜」という探索的かつ反復的プロセスであるためである。一例としては、ハイパーパラメータの範囲として「学習率」を 3 通り（0.001, 0.01, 0.1）、「正則化」も 3 通り（0.1、1.0、10.0）設定した場合、3 通り ×3 通りで 9 個のモデルを作成することになる。このように、開発の条件を変えながら複数のモデルを開発するプロセスが探索的と呼ばれる所以である。探索的開発の各活動例を図 7 に図示する。

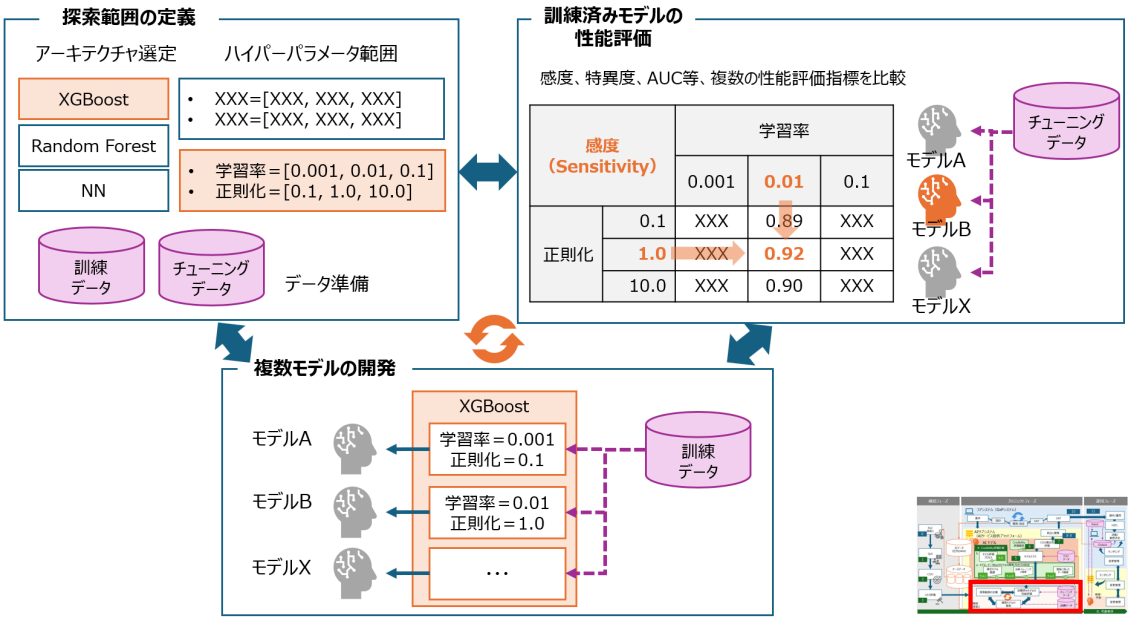

図 7：探索的開発の活動例

注：本図は図 5 の探索的開発（開発・探索 2）の活動を詳述したものである。

- **複数モデルを開発する背景：**

  AI モデルの性能は、モデルアーキテクチャ、ハイパーパラメータ、データ特性の複雑な相互作用で決まり、事前に予測できない。そのため、「複数の候補モデルを実際に開発して比較する」という探索的アプローチが科学的に合理的である。上記の探索的性質を考慮し、「探索範囲の定義 → 複数モデルの開発・評価 → 選抜 → テスト」というプロセスが適用される。

- **性能評価指標と合格基準：**

  選抜されたモデルは、過学習の可能性があるため、独立したデータで汎化性能を確認する必要がある。その選抜されたモデルのテスト（検証、評価）における、性能評価指標と合格基準は、COU に基づき以下の要素を総合的に考慮し、科学的かつ客観的に設定する。

  a) **COU 要求との整合性：**例）感度（Sensitivity）重視という方向性を満たすか
  b) **データ制約：**例）サンプルサイズ（N=1,200）、陽性率（15%）を考慮
  c) **達成可能性：**例）チューニングデータ最良候補の感度（Sensitivity）=0.92 達成 → 0.90 は現実的な性能指標候補
  d) **臨床的有用性：**例）感度（Sensitivity）＝0.90 でも臨床的に十分有用と専門家が判断
  e) **統計的信頼性：**例）信頼区間を考慮しても目標を満たせるか
  f) **現行プロセスとの比較：**現行プロセスのパフォーマンスと同等以上か

  上記を総合的に判断して、例えば、「主指標 感度（Sensitivity）≥0.90、副次指標 特異度（Specificity）≥0.70」を性能目標として決定する。

- **コアシステムと統合的な管理：**

  AI システム全体のバリデーションでは、AI モデルが単独で正しく機能することに加え、そのモデルの Output がコアシステムで正常に表示されて、「統合された状態で堅牢性を維持できること」を保証することが不可欠となる。つまり、コアシステムと AI モデルそれぞれに、ライフサイクルを通じたバリデーション活動が必要となる。

5. AI システムの Credibility 評価フレームワークの活動（各論）

本項では、各 Step の活動を詳細に述べる。なお、本項は、各 Step を FDA AI ガイダンスに基づく「概説」と、Annex 22 や関連規制、AI 事業者ガイドライン、GAMP 5 第 2 版等を参照した「考察」で構成した。

また、4.2 項で示したように、AI モデルの一般的な開発手順は、FDA AI ガイダンスで示す Step や文書化の構成順序とは異なる。必要に応じて、4.2 項のプロセスを参照し、各 Step の活動の順序と文書化の構成順序を関連づけることが必要である。

5.0 PoC の実施（Step 0: Proof of Concept）

AI システムを GxP 環境に導入する際、FDA AI ガイダンスに基づく Step 1-7 の活動を開始する前に、まず当該 AI システムの Credibility 評価フレームワークの適用性を判断し、プロジェクト全体の枠組みを確立する必要がある。これが Step 0 の目的である。この Step は、FDA AI ガイダンスには含まれていないが、GAMP 5 第 2 版のシステムライフサイクルアプ

ローチにおける「構想フェーズ」の考え方を参考に、AI システム導入の準備段階として本書で追加した Step である。

【考察】

Step 0 は形式的な準備作業ではなく、AI プロジェクトの成否を左右する戦略的な意思決定のフェーズである。適切な Step 0 の実施により、後続 Step での手戻りを大幅に削減できる。

Step 0 では、以下の主要活動を実施する：

1) 適用性の判断

FDA AI ガイダンスに従い、当該 AI モデルの適用性を判断する。主な判定基準は以下の通りであり、図 2 を参考に、全て満たす AI モデルに適用する：

Q1. 医薬品のライフサイクルに使用されるか

Q2. 業務効率化以外の目的で利用されるか

Q3. 規制上の意思決定を支援するか

全ての AI モデルの利用が FDA AI ガイダンスの適用対象となるわけではない。例えば、社内業務効率化ツールは対象外である。一方、臨床試験の結果に影響するデータやプロセスを制御する場合、製造工程の出荷判定に用いる場合、安全性評価や報告に用いる場合は対象となる可能性が高い。さらに、適用対象の AI モデルの中でも、医薬品の品質システム（PQS）の運用を支援する AI モデル等、直接的に患者の安全性や製品品質に関わらない場合、5.3 項の Step3（リスク評価）の評価結果に基づき、活動の厳密さを定める。

また、現状では、創薬・基礎研究段階の AI 利用は対象外と理解する。しかし、創薬・基礎研究段階の結果が規制申請資料に含まれる場合や、臨床試験デザインの根拠となる場合は対象となり得る。例を以下に示す：

- IND 申請に含まれる非臨床データの解析に使用される AI
- First-in-human 試験の用量設定根拠となる予測モデル
- 臨床試験のエンドポイント設定に影響するバイオマーカー解析

判断に迷う場合は、品質保証部門や薬事・薬制部門と協議し、必要に応じて規制当局に相談することが推奨される。適用性を誤ると、必要な品質保証を欠くリスク、または過剰な活動による AI モデルの開発遅延のリスクがある。

2) PoC の計画と結果

業務上の有効性、AI モデル構築の実現可能性や課題等の洗い出しを行う。実際に AI モデルを構築する PoC を通して、本開発に進める可能性があるのか、探索する。この探索には、以下項目を含む。

- 業務プロセス概要、背景情報
- 期待する性能、効果（既存プロセスとの比較、指標を含む）
- AI モデルアーキテクチャの候補
- 利用するデータ（データソース、特徴量、Input と Output のデータ経路等）
- サプライヤの要件（委託範囲、役割等）

データは AI モデルの開発を進めるために必要な重要な要素であるため、PoC の段階でデータ管理を実装しておく。図 5 では、二次利用可能なデータウェアハウス（DWH）にある社内データを使って開発、性能評価するプロセスを示した。

また、PoC 実施時にサプライヤに委託する範囲や求める能力等も明らかにする。

3) 本開発（Step1）への Go/NoGo 判断

PoC の結果から Step1 へ進むことの妥当性を評価する。

- 技術的な実現可能性
- 想定する開発（訓練・チューニング）データの妥当性
- AI モデルの性能評価

さらに、PoC から見出した要件、推奨事項、課題、リスク等を洗い出す。

5.1 QoI の定義（Step1: Define the Question of Interest）

【概説】

FDA AI ガイダンスは、Step1「Question of Interest（以下、QoI）の明確化」から始まる。Step1 の目的は「具体的に AI モデルに何を判断させたいのか、どんな課題を解決したいのか」を、誰にでも明確にわかる言葉で定義することである。

QoI は、後続の全ての Step（COU 定義、リスク評価、モデル設計、性能評価等）の基盤となる。曖昧な QoI は、不適切な AI モデル、不十分なテストデータ、そして最終的には信用度の低い AI システムにつながる。

GCP 分野、GMP 分野の QoI の具体例を示す。

GCP 分野の例（治験薬 A）

- 背景：
  - 治験薬 A には重篤な副作用のリスクがあるため、従来は、安全性を考慮し、被験者全員を 24 時間入院してもらい、モニタリングしていた。しかし、この手順は、医療機関、被験者の双方に大きな負担になっていた。過去のデータから、副作用のリスクが低い被験者もいることがわかっている。
- AI システムを使った活動や意思決定：
  - AI システムを使って事前にリスクが高い被験者と低い被験者を正確に判別したい
  - リスクの低い被験者には、入院の代わりに投与後の外来モニタリングで効率的で安全な治験を実施したい
- QoI：
  - どの被験者を低リスクと判断し、治験薬 A 投与後に入院を不要にできるか？

GMP 分野の例（注射剤 B）

- 背景：
  - 注射剤の製造ラインでは、バイアルへの薬液の充填量を目視検査している。目視検査は検査員の負荷が大きいこと、ヒューマンエラーのリスクがある。
- AI システムを使った活動や意思決定：
  - AI 搭載のカメラを使い、全製品の充填量を自動で高速かつ正確に検査し、不良品（充填量の過不足）を確実に、100%検出したい
- QoI：
  - 薬剤 B のバイアルは、定められた充填量を満たしているか？

QoI に答えるためには、AI の結果が間違っていた場合も補正可能なプロセスにする必要がある。AI の推測・予測結果だけに依存するのではなく、周辺の様々な関連情報や証拠源（以下、Evidentiary Sources）と組み合わせて、意思決定や次の活動を定めることも考慮する。Evidentiary Sources とは、例えば、in vitro 試験、in vivo 試験、臨床試験、または製造プロセスバリデーションから生成された種々のデータや人の判断結果、別のプロセスの結果等ある。Evidentiary Sources の特定は各 Step を経て深まる。Step1 では候補の洗い出しにとどまり、Step2 の COU 定義において具体化される。さらに Step3 では、AI の結果への影響を補正する要素として、リスク評価の中心的な役割を担う。

【考察】

QoI が曖昧な場合、後続の AI モデルの設計、学習に使うデータの選定、そして完成したモデルの評価基準が異なってしまう可能性があり、信用できる AI モデルを構築することはできない。QoI の明確化のプロセス概略図を図 8 に示す。

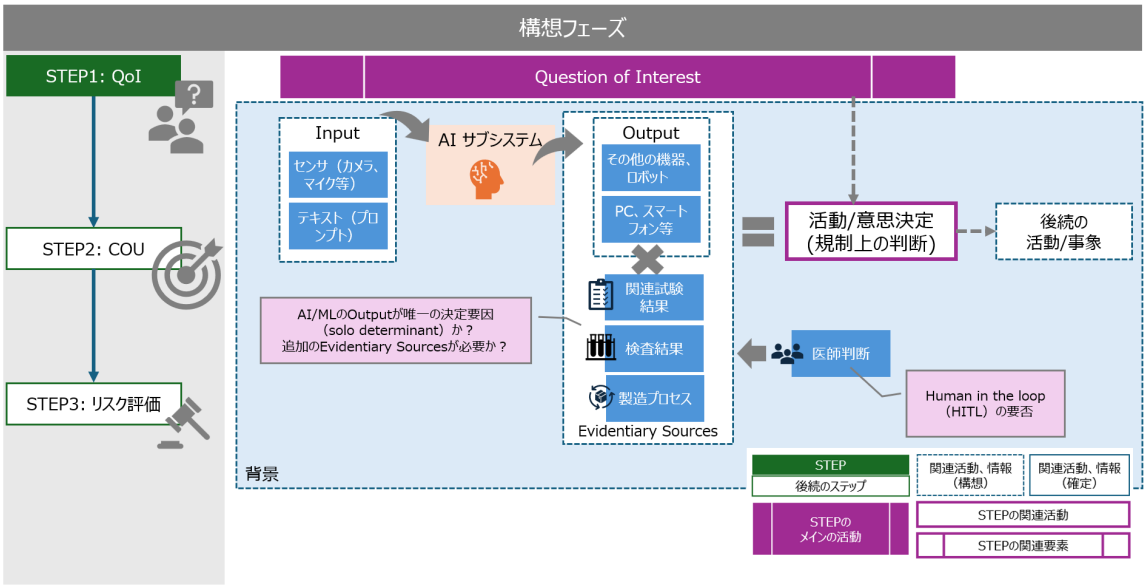

図 8：Step1 QoI の定義

QoI を定める Step は、「ビジネス課題の特定」と同様であるが、その解決方法に AI モデルを活用する点が異なる。AI システムを利用する業務プロセスの詳細化や AI システムの利用方法を深く検討することは Step1 ではなく、Step2 の COU の定義で行う。QoI の定義は、COU での詳細検討で必要となる要素の洗い出しという側面がある。

図 8 に示すように、GCP の事例では、「どの被験者を低リスクと判断し、治験薬 A 投与後に入院を不要にできるか？」を QoI と定義した。しかし、医師の判断を加えて、最終的な入院要否を決定するかもしれない。GMP の事例では、「薬剤 B のバイアルは、定められた充填量を満たしているか？」を QoI と特定した。一方で、目視検査と比較して出荷判定することも想定する。

このような「QoI に答えるための Evidentiary Sources」は、Step1 で洗い出し、Step2 に特定する。以下に Evidentiary Source の例を示す：

- 既存の臨床データや文献情報
- 他の検査・測定結果
- 専門家（医師、薬剤師、検査員等）の判断
- 別の AI モデルやシステムの Output

QoI 定義時には、業務部門、データサイエンティスト、品質保証部門が協働し、これらの基準を満たすよう議論を重ねることが推奨される。

5.2 COU の定義（Step2: Define the Context of Use for the AI Model）

【概説】

Step2 は、AI モデルの COU、つまり、「AI の役割と範囲」を定義する活動である。Step1 で定義した QoI に答えるために、「AI モデルがどんな Output を出力し、これをどのように使うのか」を Evidentiary Sources と共に具体化していく。COU を定義する際、以下の 2 つの要素を明確にする。

- AI モデルの役割（Role）：
  - AI が「何をするのか」、AI の具体的なタスク
    例）リスクを予測する、不良品を検知する、画像を分類する等
- AI モデルの範囲（Scope）：
  - AI の Output の使用条件、その他 Evidentiary Sources の利用の有無
    例）AI の Output のみで決まるのか、人間の判断を補助する参考情報のひとつであるのか等

以下に 5.1 項の事例を基に、GCP 分野、GMP 分野の COU の具体例を示す。

表 4：COU の例

例	QoI	役割（Role）	範囲（Scope）
GCP 分野の例（治験薬 A）	どの被験者を低リスクと判断し、治験薬 A 投与後に入院を不要にできるか？	被験者のデータ（患者背景や臨床検査値等）に基づき、治験薬 A に関する副作用リスク、特に生命に関わるリスクを「高リスク」か「低リスク」に分類（層別化）すること	AI システムの Output を唯一の根拠として、被験者の入院モニタリング、または、外来モニタリングを決定すること
GMP 分野の例（注射剤 B）	薬剤 B のバイアルは、定められた充填量を満たしているか？	製造ラインを流れる全バイアルの画像を解析し、充填量が基準から逸脱しているバイアルを検知すること	AI システムは 100%検査のためのスクリーニングツールとしての役割を担う。しかし、最終的な製品 Lot の出荷可否は、従来通り担当者が抜き取りで行う独立した検証を根拠に判断すること

【考察】

COU を特定する目的は、AI モデルの役割と範囲を定義することである。Step1 で洗い出した背景情報（Input 情報、Output 情報、その他 Evidentiary Sources）を具体化し、AI モデルの COU を役割と範囲の 2 つの要素に分解し、定義する。Step2 の概念を図 9 に示した。

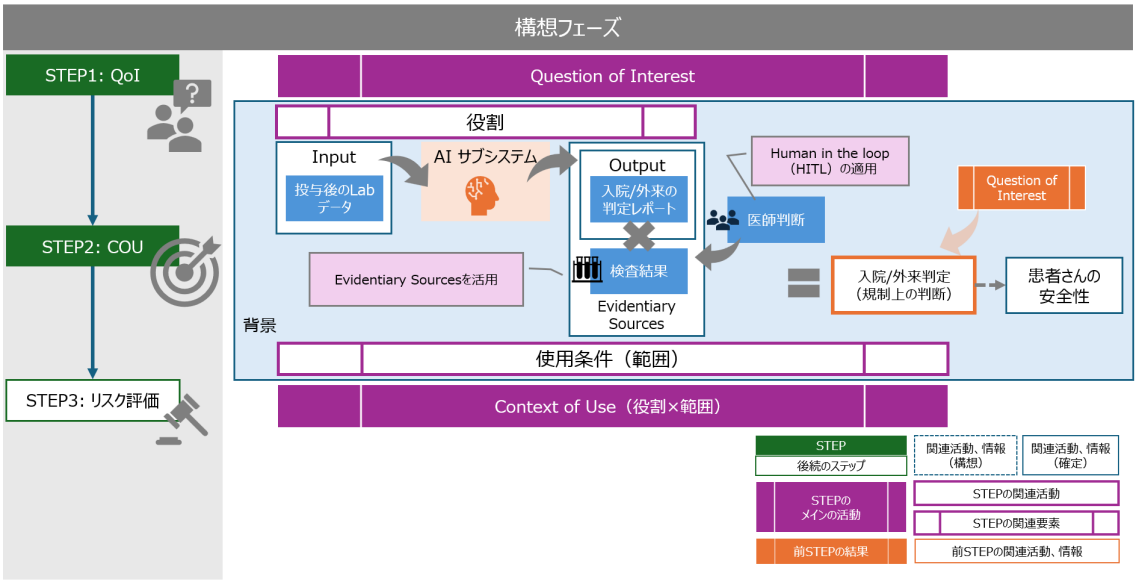

図 9：Step2 COU の定義

1) COU の定義の重要性

従来のコンピュータ化システムのバリデーションの Intended Use が AI システムの COU に当たる。しかし、従来の Intended Use の特定では、「コンピュータ化システム」の役割や範囲といった定義は、使用環境や要求する機能、シナリオの定義に留まっていると考える。しかし、GxP プロセスにおいては、AI システムの Output のみで完結するケースはごく限られており、多くの場合、その目的や使用条件を明確にしたうえで、人による確認等の追加的な検討が不可欠である。したがって、Step2 で求められる COU の定義は、AI システムへの依存度を適切に調整し、GxP 業務における意思決定を支援するうえで、重要な役割を果たすプロセスである。さらに、客観的なリスク評価の観点を明確にし、AI システムの性能や利用プロセスに関する説明責任を支える基盤として位置づけられる。

2) 役割と範囲の定義

COU は単に AI システムの Output を使って「何をするか」（役割）だけでなく、AI システムの Output とその他の Evidentiary Sources を明確にすることで、AI システムの使用条件（範囲）を定める。例えば、GCP 分野の例では、AI が被験者を「低リスク」と分類した場合、その Output のみに依存して入院の要否を決定するのか、医師による診断

結果を加味して決定するのかでは、AI システムの役割は大きく異なる。GMP 分野の例では、AI システムが充填不良を検知した場合でも、最終的な出荷判定には独立した抜き取り検査があるため、AI システムの役割は相対的に限定的となる。このように、COU における「範囲」（Scope）、すなわち AI の Output の使用条件や他の Evidentiary Sources との組み合わせ方は、結果として、AI モデルの役割（Role）の重さや位置づけに大きな影響を与える。

3) Evidentiary Sources の特定

Step1 で洗い出した AI システムの Output 以外の Evidentiary Sources を Step2 で特定する。AI システムの Output とこれらの Evidentiary Sources をどのように統合して意思決定を行うのかを明確にする。

4) HITL の位置づけ

HITL とは、AI システムの出力を専門家等が確認・検証し、最終的な意思決定に関与するプロセスである。COU の範囲を定義する際、特に重要なのが HITL の有無と、その関与の度合いである。Annex 22 は、AI システムの Output を人間の専門家が最終確認する HITL を組み込んだプロセスを推奨している。これにより AI システムの誤った Output が患者さんの安全、製品の品質および結果の信頼性に悪影響を及ぼすリスクを低減できる。ただし、HITL はどんな場合でもリスク低減の効果を発揮するわけではない。形式的な Output の確認作業になっている場合、確認する専門家が判断できる情報がない場合には、HITL の効果は期待できない。したがって、COU で HITL を定義する際は、専門家がどのような情報（Evidentiary Sources）に基づき、どのような判断基準で最終決定を行うのかを具体的に記述することが重要である。

5) COU の文書化（推奨）

この時点で COU を明文化することは必須ではないが、Step4 での文書化に必要な情報を準備できるため有意義である。COU は、以下の観点を含めることを推奨する：

- **AI モデルの役割**
  - AI モデルへの Input（データの種類、形式、取得方法）
  - AI モデルからの Output（予測値、分類結果、信頼性スコア等）
  - AI モデルからの Output の QoI、意思決定への寄与
- **AI モデルの範囲**
  - AI を利用した業務プロセス
  - 利用環境（ユーザ、場所、タイミング）
  - AI モデルからの Output の利用方法（自動判定、HITL での参考情報等）
  - その他 Evidentiary Sources とその利用/判断プロセス

- 利用上の制約や前提条件

さらに、COU の定義が AI プロジェクト全体の基盤となるため、その品質を担保することが極めて重要である。良い COU は以下の特徴を持つ：

- **具体性**：QoI の構成要素が明確である
- **測定可能性**：AI モデルの性能評価が可能な形で定義されている
- **臨床的/業務的妥当性**：実際の医療現場や製造現場のニーズに即している
- **実現可能性**：現実的なデータと AI モデル技術で対応可能

5.3 リスク評価（Step3: Assess the AI Model Risk）

【概説】

Step3 の目的は、Step2 で定めた COU に基づいて、AI モデルのリスクを評価することである。このリスクは、以下の 2 つの要素をかけ合わせて評価する。

- **Decision Consequence（以下、DC）：**
  AI の Output が間違っていた場合の結果の重大性
- **Model Influence（以下、MI）：**
  規制上の意思決定に対する AI から出力される Output の影響の度合い

AI モデルのリスクとは、AI モデルから出力された予測、分類、検出といった Output が誤った意思決定を促し、患者さんの安全、製品の品質、結果の信頼性に悪影響を及ぼす結果（Adverse Outcome）を導く程度である。

表 5 に 5.1 項の事例を基に、GCP 分野、GMP 分野の AI モデルのリスク評価の具体例を示す。

表 5：AI モデルのリスク評価

事例	Decision Consequence	Model Influence	AI モデルリスク
GCP 分野の例（治験薬 A）	**高：**患者さんの安全 Output が間違った（高リスク患者さんを低リスクと誤判定した）場合、患者さんは生命の危機に瀕する可能性があるため、結果は極めて重大	**高：**AI が唯一の判断材料 患者さんの入院／外来の判断を AI の Output だけに依存しているため、AI の影響度は最大。	**高×高＝高：** 最も厳格な管理と検証を要する

事例	Decision Consequence	Model Influence	AI モデルリスク
GMP 分野の例（注射剤 B）	**高：**投薬ミス 薬液の量が逸脱している場合、患者さんに正しい量の薬剤を投与できず、投薬ミスや健康被害につながる可能性があるため、結果は重大	**低：**別プロセスによる検証あり AI は全数検査を行うが、一次スクリーニングの位置づけ。最終的な出荷判定には、「検査員による抜き取り検査」という独立した検証プロセス（リスク管理策）が存在する。AI は唯一の判断材料ではないため、影響度は限定的	**高×低＝中：** 管理や検証のレベルを合理的に設定できる

【考察】

Step3 は、AI モデルの潜在的なリスクを評価する。「AI モデルのリスク」とは、AI モデルのアルゴリズム自体の欠陥や障害およびバグ、セキュリティリスクではない。「AI モデルの間違った推論結果に従って行動した結果、後続プロセスが患者さんの安全、製品品質、および結果の信頼性に対して、どの程度の悪影響を及ぼすのか」を評価することである。

GCP 分野の例では、AI システムの Output が入院要否を決定する唯一の根拠であるため、影響度は「高」と評価される。一方、GMP 分野の例では、AI システムは一次スクリーニングであり、独立した抜き取り検査があるため、影響度は「低」と評価される。

構想フェーズの全体像を Step1 から Step3 の活動を関連付けて図 10 に示す。FDA AI ガイダンスでは、AI システムの Output が間違っていた場合、規制上の活動や意思決定がどの程度の深刻な結果を及ぼすかを示す DC と規制上の活動や意思決定における AI システムの Output の影響度である MI の 2 軸から、AI モデルリスクを算出する。Step2 の COU がこれらのリスク評価の根拠となる。

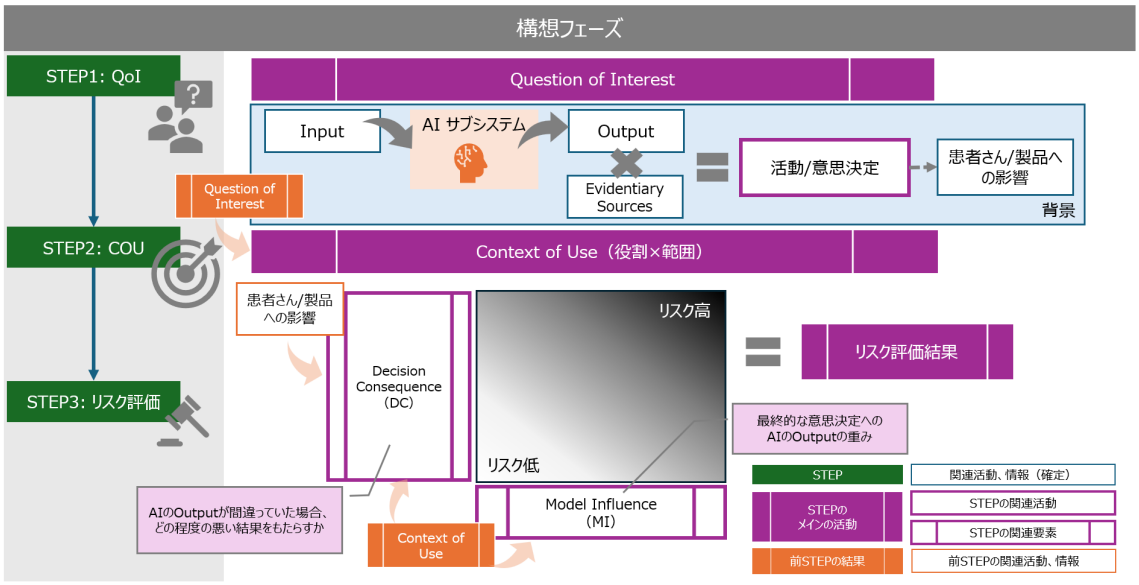

図 10：Step3 AI モデルのリスク評価

AI モデルリスクのレベルに応じて、次の Step の Credibility 評価（モデルのテスト）とその後の運用（モニタリング等）の厳密さや範囲を決定する。リスクに応じたレベルの活動は各社の SOP 等で定めることが必要である。

1) Decision Consequence の評価

AI システムの Output には、偽陽性、偽陰性、過不足等の間違いが発生し得る。DC のリスク評価時は、単に高低を評価するのではなく、COU を元に評価軸およびその結果を DC の根拠として明文化することで、DC の妥当性を示すことができ、COU が変更になった際の影響評価も容易になる。

2) Model Influence の評価

MI 評価では、Step2 で特定した「AI モデルの範囲」が重要な判断材料となる。以下の観点で評価する：

- AI の Output が唯一の意思決定根拠か、その他の Evidentiary Sources の一つか
- HITL の有無
- 他の独立した併用プロセスの有無

3) 動的・適応型 AI モデルのリスク評価

FDA AI ガイダンスのリスク評価では、AI モデルの自己学習の有無は、リスク評価の 2 軸（DC/MI）には含まれていない。しかし、5.7 項に述べるように、自己学習して環境に適応していく AI モデルを排除していない。一方で、Annex 22 においては、動的・適応型

AI モデルの GxP 使用を制限している。したがって、運用中に継続的かつ自動的に学習・更新する動的・適応型 AI モデルを利用する場合には、Adaptiveness 等の追加のリスク評価項目が必要と考える。

ただし、Annex 22 では動的・適応型 AI モデルは利用不可となっていることから、リスクはかなり高く評価すべきであり、現状ではハードルが高いこと、さらに、査察でもかなりレベルの高い Credibility 評価および関連する活動が求められる可能性がある。

5.4 Credibility 評価計画の立案 (Step4: Develop a Plan to Establish AI Model Credibility Within the Context of Use)

【概説】

Step4 では、AI モデルの Output の Credibility 評価のための計画立案を行う。

「AI モデルを特定の目的で利用する場合の当該 AI モデルの Credibility 評価の具体的な計画書」を策定する。後続の 4.a、4.b では、AI モデルの構築と Credibility 評価活動について、一般的な考慮事項と評価手法を述べる。

4.a: モデルとモデル開発プロセスの記述（結果）（Describe the model and the model development process）

i. モデルの概要（Describe the model）
ii. 開発に用いたデータ概要（Describe the data used to develop the model）
iii. モデルの訓練・チューニングの概要（Describe model training）

4.b: モデルのテスト（Describe the model evaluation process）

AI に関する技術は急速に進化しており、医薬品のライフサイクルにおける AI モデルの利用は今後ますます拡大していくことが予想される。したがって、AI モデルの Output の信頼性を確立するための活動は、一般的に COU とリスクに見合ったものが求められ、本書に記載した Credibility 評価計画の内容もテーラリングすることを前提にしている。

【考察】

「Credibility 評価計画」は探索的に開発された AI モデルのテスト計画のことである。

図 5 の Step4. Credibility 評価の計画を中心に簡略化して図 11 に示す。

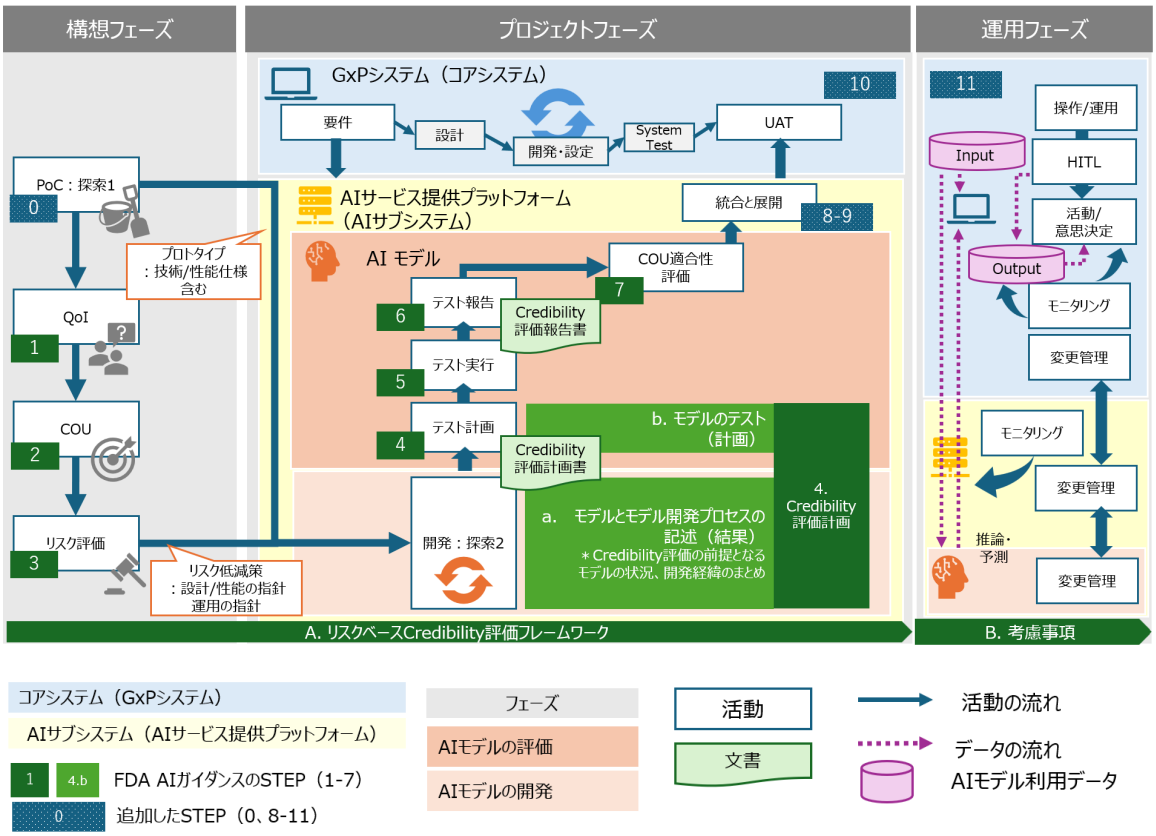

図 11：Step4 Credibility 評価の計画

Step0 の PoC の探索的な開発の技術的な情報および Step3 のリスク評価結果のリスク低減策をベースに探索的な開発を継続する。この「探索的な開発」が図 11 の「開発：探索 2」に該当する。AI モデルの探索的な開発が終了すると、複数のモデルの中で最良の性能を示すモデルアーキテクチャとハイパーパラメータの組み合わせが定まり、そのモデルが学習したパラメータ（重みとバイアス）が確定する。FDA AI ガイダンスのフレームでは、このハイパーパラメータ等、開発時の情報を「4.a モデルとモデルの開発プロセスの記述（結果）」としてまとめる。そして、この情報を基に、Credibility 評価である「b. モデルのテスト」プロセスを計画する。つまり、「モデルの Credibility 評価計画書」とは、探索的に開発され選抜されたテスト対象であるモデルの状態や経緯を概説し、その開発および選抜の妥当性と性能を検証するためのテスト計画書である。

この考え方は、従来のソフトウェア開発、CSV の概念とは大きく異なる。このため、FDA は敢えて、テストやバリデーションという言葉を使わず、「Credibility 評価」という新しい概念を使っていると考察する。あわせて、FDA AI ガイダンスでは、AI モデル導入全体を見据えた、いわゆる「バリデーション計画」についても、開発プロセスそのものについても言及されていない。この点も CSV との大きな相違点の一つである。

AI モデルの Credibility 評価は、その方法も技術も急速に進化しているため、画一的な方法では対応できない可能性がある。本書もドラフト段階の FDA AI ガイダンスを基にしているため、実際に AI モデルを開発する際には、最新の技術や規制を考慮して計画することが必要である。このため、FDA は、AI の利用範囲や評価方法を計画段階から規制当局と協議、合意することを強く推奨している。

FDA は、革新的な AI アプローチについて開発初期段階での対話を推奨している。特に前例のない新規 AI 技術の利用、高リスク COU、複雑な HITL プロセスを含む場合、早期対話により規制当局の期待値を把握し、Credibility 評価計画書を使って妥当性について事前に合意を得ることができる。これにより、FDA が求める事項が Credibility 評価計画書の内容となり、承認申請時の指摘リスクを大幅に低減できる。

本計画に求められる主な項目を図 12 にまとめた。

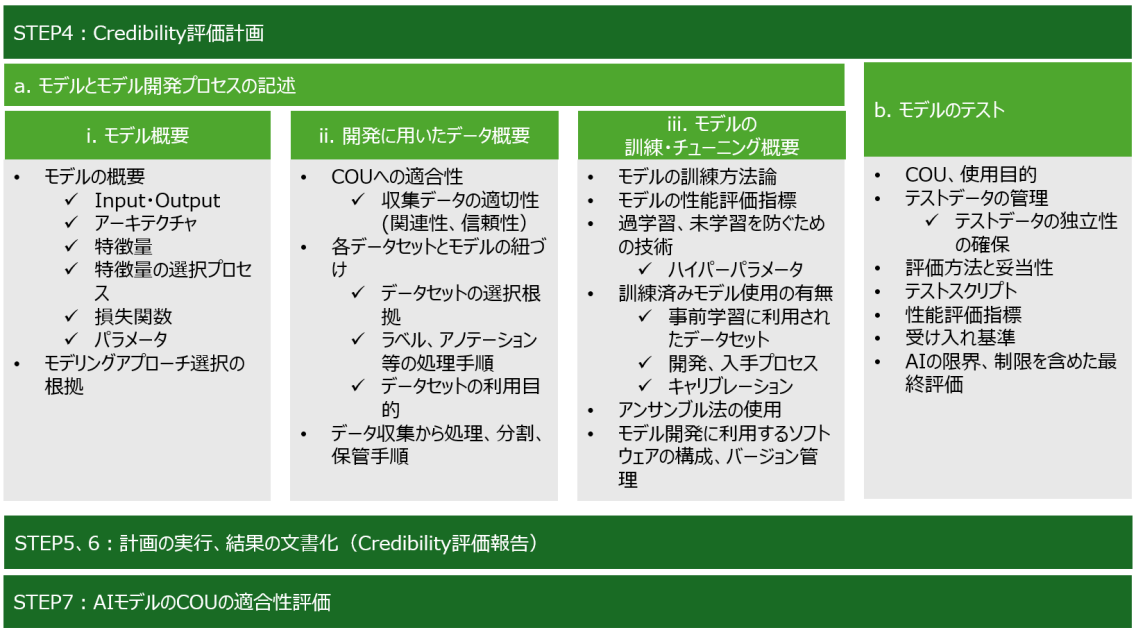

図 12：Step4 Credibility 評価計画の概要

注：本書における「4.a.i」等の記法は、FDA AI ガイダンス Section 4 のサブセクション（4.a.i, 4.a.ii 等）に対応する。本書の節番号とは異なる。

FDA は Computer Software Assurance（以下、CSA）のガイダンスに示すように、コンピュータ化システムにリスクと Intended Use に最適な信頼性保証活動のレベルや厳密さを求めている。同時に、その方法を選択した説明責任、透明性も求めている。CSA の具体的な評価や考え方は、日本製薬工業協会 医薬品評価委員会、電子化情報部会 2023 年度タスクフォース 4 が 2024 年 7 月に公開した「FDA CSA ドラフトガイダンスの概説と GxP 領域への適用の検討と考察（以下、ドラフト CSA の考察）」に詳述されている。同書においては、CSA

の概念を本来のスコープである医療機器分野から拡張し、GxP 領域におけるコンピュータ化システムの信頼性保証活動にも適用している。また、そのレベルや厳密さを判断するための体系的なフレームワークも提示されている（図 13 参照）。現在、最終化された CSA ガイダンス（第二版）が発出されているが、図 13 で示しているドラフト CSA の概念に変更はなく維持されている。

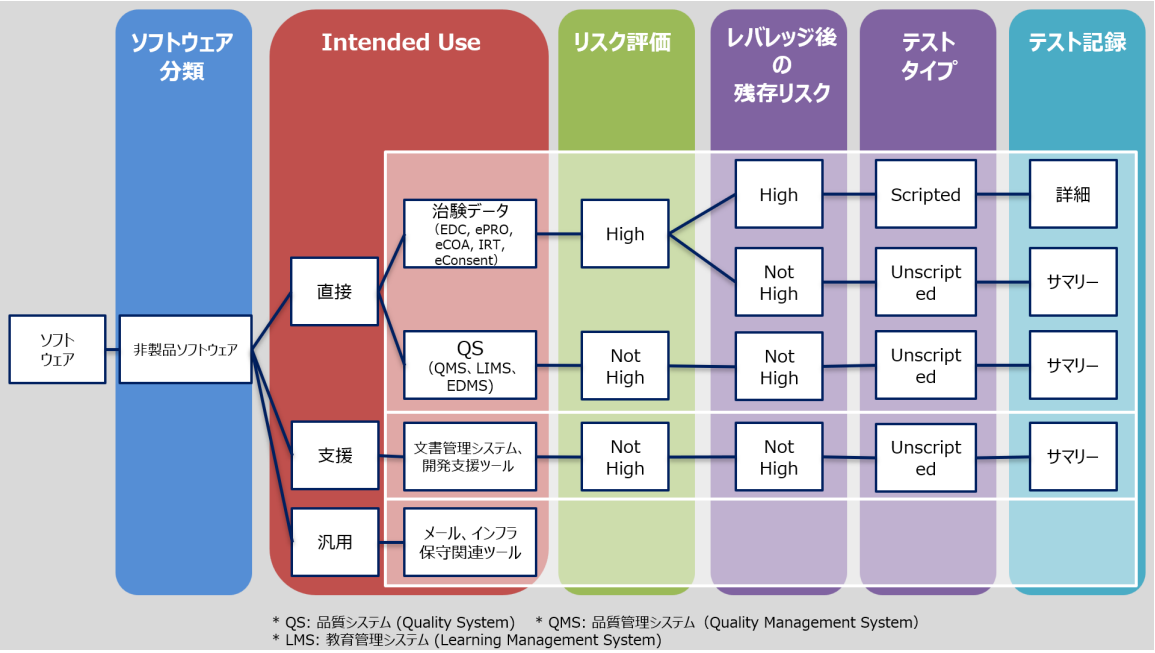

図 13：「GCP の CSA フロー」（「ドラフト CSA の考察」図 13 より引用）

これは、AI モデルの Credibility 評価でも同様であり、Step4 の AI モデルの Credibility 評価計画で各活動や選択した方法の根拠をそれぞれ説明すること、さらに、AI モデルの制限や限界についても述べることが求められる。これらは、主に COU に対する AI モデルの性能評価を中心として、関連事項を明文化する項目が列挙されている。

Annex 22 では、医薬品のライフサイクルに AI システムを用いる場合、その推論や予測が、既存のプロセス、専門家による判断と同等以上に信頼できることが求められる。特に、その AI システムの Output が患者さんの安全や製品の品質、結果の信頼性に直接影響を与える場合は、極めて高い信頼性が必要である。このため、現行のプロセス等に対する優位性を評価することも必要である。

5.4.1 モデル概要（Step4 a i. Describe the Model）

【概説】

開発されたモデル概要を Credibility 評価計画書にまとめる。

- モデルの概要
  - モデルの入力（Input）と出力（Output）
  - モデルのアーキテクチャ（CNN/畳み込みニューラルネットワーク等）
  - モデルの特徴量
  - 特徴量選択プロセスおよび、モデル設計と最適化のために使用された損失関数（該当する場合）
  - モデルのパラメータ
- モデリングアプローチの選択根拠：上記モデルおよび特徴量を選択する根拠

【考察】

AI システムの Output は、患者さんの安全、製品の品質に影響を及ぼす可能性があるため、AI モデルを「ブラックボックス」として扱うことはできない。したがって、製薬企業は「AI モデルの仕組みや開発の背景、コンセプト」を、第三者が理解できるレベルで説明する責任を負う。「モデルの概要」は、図 14 に示すように、探索的開発の結果を述べる。

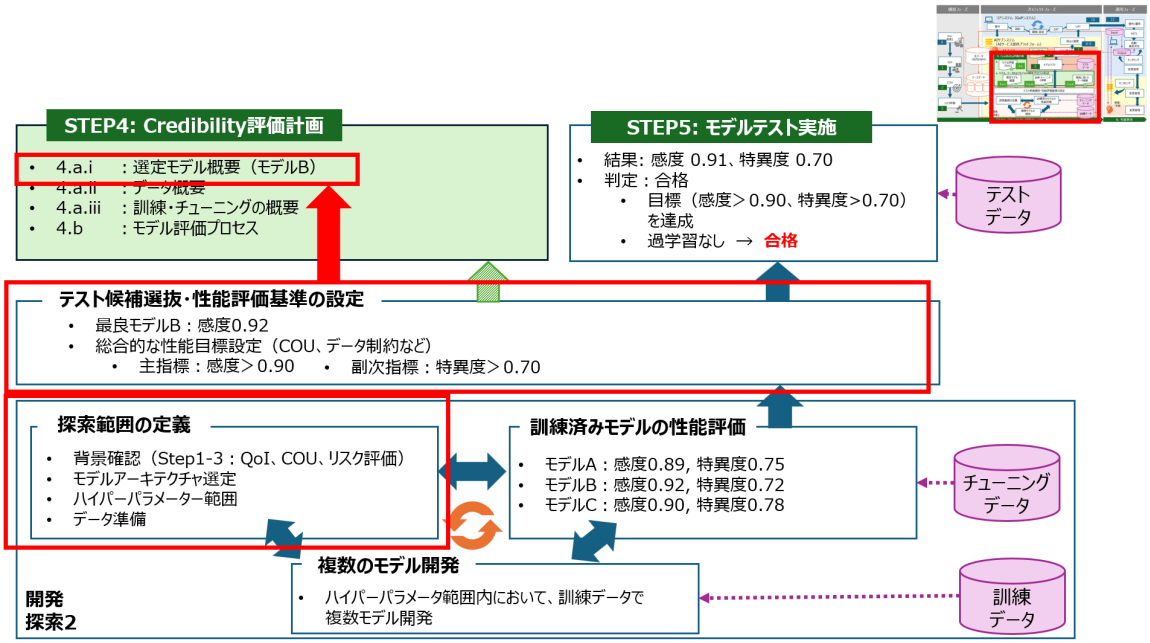

図 14：Step4 a i. AI モデルの概要

注：本図は図 5 の探索的開発（開発・探索 2）から Step5 の各活動を詳述したものである。

AI モデルの概要の詳細さは、AI モデルのリスクに依存する。高リスクモデルでは、上記の全項目の詳細な情報、および追加情報を要する可能性がある。一方、低リスクモデルでは、最小限の情報（Input、Output、アーキテクチャの概要等）で十分である場合がある。

以下に FDA が示した各項目について、述べる。

- **AI モデルの概要**
  - AI モデルの入力（Input）と出力（Output）：
    - 構想フェーズ（QoI、COU）で明確にした情報。Input、Output データの内容や媒体、経路を含む
    - Annex 22 では、予測や分類を行う AI モデルの場合、Output だけでなく、AI モデルの予測結果の確信度を示す「信頼性スコア」や閾値等付属情報を含めることを求めている。これらの情報は、AI モデルの予測値をどの程度活用して人間が意思決定できるのかを示し、最終判断をより安全なプロセスとすることができる
    - 信頼性スコアが極端に低い場合には、「予測不可」の Output を出力することも検討する
  - モデルのアーキテクチャ：
    - AI モデルの基本的な構造、種類等（CNN 等）
  - モデルの特徴量：
    - AI が Output 作成のために着目しているデータ内の具体的な観点、画像の場所、変数。実務的には、表形式データの列にあたるデータ属性と考えてよい。例えば、患者背景情報（年齢、性別、既往歴等）や検査値、投与情報など。どの属性を特徴量としてモデルに与えているかを明示することが重要。
  - 特徴量選択プロセスおよび、モデル設計と最適化のために使用された損失関数（該当する場合）：
    - 特徴量選択プロセス：数ある特徴量の候補から、最終的にそれらを選んだ理由、そのプロセス（例：既存の治験 A、治験 B より、対象疾患との相関が最も高い上位 20 個の特徴量を選択した）。
    - 損失関数：AI モデルの予測が、正解（真値）に対してどれだけ「望ましくないか」を定量的に評価するための関数。この損失関数によって計算される値（損失値）を最小化する内部パラメータの組み合わせを定義することが学習の目的である。損失値が小さいほど、モデルの予測が正解に近いことを意味する（例：見逃しや過剰検知等）
  - モデルのパラメータ（重みとバイアス）：
    - モデルが訓練データから学習により獲得する内部的な数値。例えば、ニューラルネットワークの各層の重み行列や、決定木における各ノードの分割に用いられる特徴量と閾値、XGBoost における各木の葉の重み等、学習によって定まる内部の規則や数値が含まれる。これらのパラメータは開発プロセスで自動的に最適化され、開発者が直接設定するものではない。

- **当該 AI モデリングアプローチの選択根拠：**
  - 対象とする課題・データ構造・リスク水準に照らして、なぜ当該モデリング手法を選択したかについての簡潔な理由付け。例えば、以下の観点を含める。
    - なぜこの問題設定に対してその手法を選んだのか
      例：線形モデルではなく非線形モデルとした理由
      例：ブラックボックス的な深層学習を採用した理由
    - なぜこのデータ構造に対して妥当と考えるのか
      例：時系列データであるため 時系列モデルが適していること
      例：構造化表データであるためツリーモデル等を選択したこと
    - なぜこのリスク・COU に対して妥当と考えるのか
      説明可能性・再現性・実装の成熟度・既存実績などを踏まえた選択理由

5.4.2 モデル開発に用いたデータ概要（Step4 a ii. Describe the data used to develop the model）

【概説】

AI モデルの開発時、一般的に訓練、チューニング、テストの 3 種類のデータに分割し、それぞれの目的にのみ利用する。訓練データは AI モデルの開発（モデルの重みづけ、結合、コンポーネントの定義を含む）時に利用する。チューニングは、AI モデルのテストの前に行われ、AI モデルの開発プロセスの一部である。

AI モデルの性能は、そのモデルの訓練およびチューニングに使用されるデータセットに大きく依存する。したがって、AI モデルの開発に使用されるデータは「用途に適して（fit for use)」いなければならない。これは、データが「関連性があり（relevant)」、かつ「信頼できる（reliable)」ことの両方を意味する。

- **関連性がある（relevant）：**主要なデータ要素と十分な数の代表的な被験者を含む、または、製造プロセスやオペレーションを代表する十分なデータである
- **信頼できる（reliable）：**正確で、完全で、追跡可能である

モデルのリスクに応じて、製薬企業は、開発データセット（訓練、チューニングデータセット）に関するデータ管理手順を定め、その開発データセットの特性を明らかにする。これらの手順は、データの潜在的な限界を特定し、特定の COU における AI モデルの利用可否を証明するための適切な Credibility 評価活動を特定するのに役立つ。

モデル開発に用いたデータの概要を Credibility 評価計画書にまとめる。

- 開発データセットの説明

- 開発データセットの説明（開発データセットがどのように訓練、チューニング、およびその他の追加サブセットに分割されたかを含む）
- 各データセットを使用してどのモデルを開発するのか

- 開発データの収集、処理、アノテーション、保管、管理手順、および、使用目的（訓練、チューニング）
  - 特定の開発データセットを選択した論理的根拠
  - ラベルやアノテーションの付与手順
- 開発データの COU への適合状況
- 開発データの関連性と信頼性
- 開発データの管理状況

【考察】

AI モデルを構築するためのデータ管理は AI システム開発特有の活動である。「どのような品質のデータを」「どのような管理プロセスを経て」「どのように利用したのか」という一連のデータ管理プロセス全体を通して、科学的根拠に基づいた透明性の確保とトレーサビリティの維持が求められる。「データ概要」は、図 15 に示すように、探索的開発で用いたデータについて述べる。

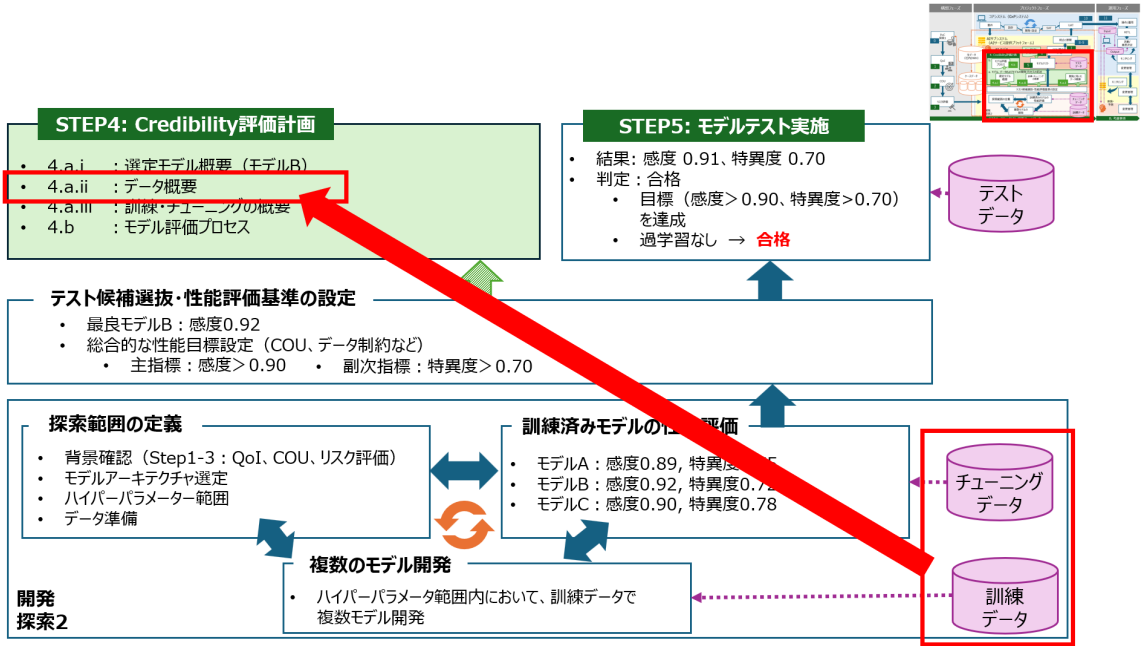

図 15：Step4 a ii. データ概要

注：本図は図 5 の探索的開発（開発・探索 2）から Step5 の各活動を詳述したものである。

事前に利用した訓練データおよびチューニングデータとテストデータはアクセス権を含めて厳格に分割、管理する。この厳密なデータの分割により、モデル評価の信頼度を担保するための透明性を示すことができる。

各データの種類と目的を表 6 にまとめた。

表 6：AI モデルに利用するデータの種類と目的

種類		**目的**
開発データ	訓練データ	AI モデルにパターンを開発（学習）させるためのデータ。モデルの基本的な性能（知識や能力）を形成する。
	チューニングデータ	訓練済みモデルの性能を評価し、テスト対象の AI モデルを選抜するためのデータ。
テストデータ		訓練とチューニングが完了したモデルの「真の実力（汎化性能）」を、最終的かつ客観的に評価するためのデータ。訓練やチューニングの過程では未使用の完全に独立したデータであることが基本的条件である。

注：FDA のチューニングデータは、Annex22 ではバリデーションデータと呼ぶ。また、一般的な機械学習の文献でも「バリデーションデータセット」はチューニングと同じ目的（モデル選定・ハイパーパラメータ調整）に使うデータを指すことが多い。本書では、GxP 文脈における「バリデーション」（システム検証活動）との混同を避けるため、FDA の表記に倣い「チューニングデータ」に統一している。

AI モデル開発と Credibility 評価（テスト）のためのデータの流れを図 16 に示す。

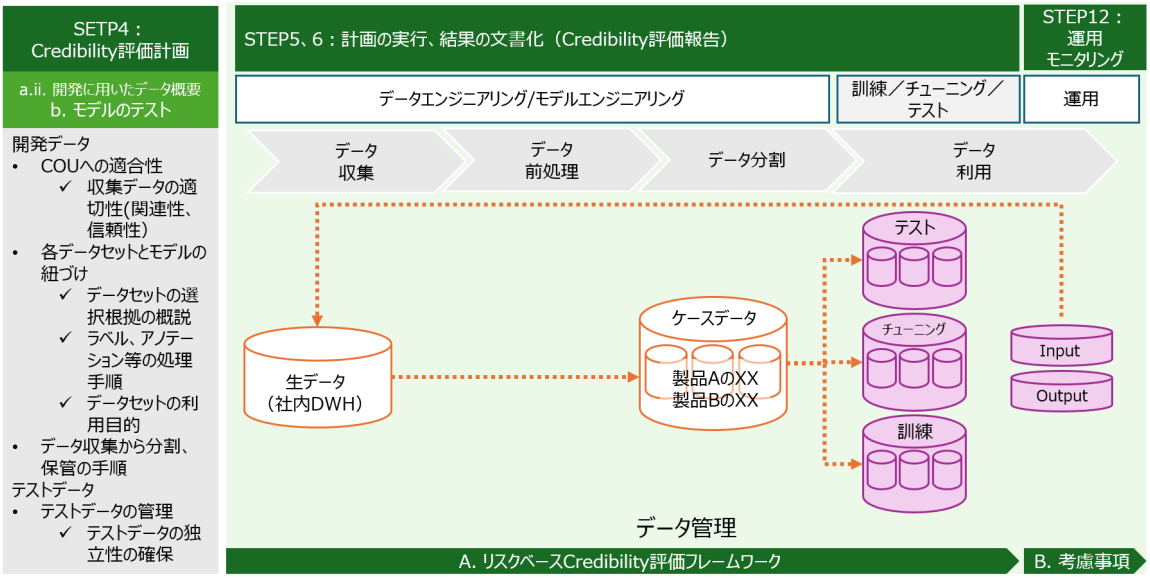

図 16：Step4 AI モデル開発、テストに使うデータ

注：Input と Output の入手、管理経路はコアシステムを含めた AI システム全体の構成によって異なる。以下に例を示す。

- コアシステムに Input、Output のデータベースが組み込まれる
- コアシステムとは別に、Input、Output はそれぞれ独立したデータベースをもつ
- Input は各種のデータを収集し、社内 DWH で統合、社内 DWH から入手する。Output も社内 DWH に格納される

AI モデルの開発およびテストに利用するデータはプロセス全体を通して、厳密に管理することが求められる。想定される主なデータ管理のプロセスを表 7 にまとめた。

表 7：データ管理のプロセス

プロセス	**目的と活動**	**主な要求事項**
データ収集 (Data Acquisition)	✓ **目的：** AI モデルの COU に合致した、生データを収集・選別する。 ✓ **活動：** データソースの特定、収集計画の策定、データ仕様の定義、収集したデータの初期品質確認。	✓ **収集元と根拠：** どの臨床試験、医療システム、製造バッチ等からデータを収集したか。また、対象データソースを選定した根拠を明確にすること。 ✓ **COU への適合性：** 収集したデータが「AI システムが実際に使用される環境（人種、性別、重症度等の分布）」を適切に代表していること（バイアスがないこと）を説明すること。 ✓ **データドリフト（環境変化、ギャップ）への考慮：** 過去のデータと実環境のデータの乖離リスクをどのように考慮したかを説明すること。
データ前処理 (Data Preprocessing)	✓ **目的：** 生データをモデルが利用できる形式に加工する。 ✓ **活動：** 欠損値処理、ノイズ除去、正規化、アノテーション（パターン別のラベリング）。アノテーション情報を基にケースデータ（パターン毎のデータセット）として、保管・管理する。	✓ **処理手順の文書化：** データを加工した手順の記録と、その再現性を保証すること。 ✓ **参照方法（Reference Method）：** アノテーションの基準となった方法（例：専門医の合議）の定義と、手順や基準が信頼できるものである根拠を説明すること。

プロセス	目的と活動	主な要求事項
		✓ **アノテーション品質：** アノテーション担当者の適格性、作業手順の SOP、品質管理（QC）プロセスを明文化すること。
ケースデータ（データセット）分割 (Dataset Strategy & Splitting)	✓ **目的：** データを目的別に分割する。 ✓ **活動：** 全データを訓練、チューニング、テスト用に分割する戦略の策定と実行。	✓ **独立性の確保：** 開発データとテストデータが完全に独立していることを保証する具体的な方法（例：時間的・場所的分離）を説明すること。 ✓ **分割の根拠：** その分割比率・分割方法を選んだ科学的根拠を説明すること。 ✓ **重複使用の正当化：** 開発データとテストデータの重複使用があった場合の詳細な説明と、テストへの影響を評価すること。
データガバナンスとトレーサビリティ (Data Governance & Traceability)	✓ **目的：** データのライフサイクル全体を通じて、データの完全性、セキュリティ、再現性を維持・管理する。 ✓ **活動：** データ保管、バージョン管理、アクセス制御、監査証跡の管理。	✓ **保管・管理手順：** データの保管場所、セキュリティ対策、アクセス権限の管理方法。 ✓ **バージョン管理：** データセットだけでなく、AI モデルの開発・テストに使用したソフトウェア、ライブラリ、ツールの正確なバージョン情報の記録と管理。 ✓ **品質保証：** コードの検証や計算の正確性を保証するための QA/QC 手順。データインテグリティ（ALCOA++ 原則）の担保。

AI モデルの性能は、モデルの開発からテストに使用されるデータセットに大きく依存する。したがって、「使用への適合性（fit for use）」を満たしたデータを収集することが求められる。データの「関連性（relevant）」を 5.2 項の例を基に述べる。

GCP 分野の例（治験薬 A）の「関連性（relevant)」

日本人の診断を補助する AI モデルを作る際、欧米人のデータしかない場合、データの関連性が低いと判断する。人種や文化によって疾患の傾向や身体的特徴が異なる可能性がある。

GMP 分野の例（注射剤 B）の「関連性（relevant)」

製造ラインの異常検知 AI モデルを作る際、正常な製品のデータだけでなく、想定される様々な種類の状況（欠け、汚れ、印字ミス等）のデータが「代表的」かつ「十分な数」蓄積済みである場合、関連性が高いと判断できる。

さらに、十分な「関連性（relevant)」と「信頼性（reliable)」をもったデータは同時に、モデルが実際に展開される環境で処理するデータを、統計的に代表するものでなければならない。過去と現在のデータ傾向が乖離する「データドリフト （Data Drift)」の可能性を常に考慮し、データセットが現在の COU を忠実に反映していることを説明する。

次に、各データを開発、テストに利用できるよう、前処理およびアノテーションを行い分類する。前処理のプロセス（変換、正規化、標準化、アノテーション等）は事前に計画し、使用目的を満たす範囲で行う。このため、データの過剰なクリーニングや除外は推奨されず、各活動の妥当性を説明することが求められる。なお、Annex 22 では、生成 AI によるデータの生成や拡張は推奨されていない。もし、生成 AI を利用したデータを開発やテストに用いる場合、当該データの正当性および妥当性も保証する。

アノテーションは、「正解データ」を作る作業であり、この品質が AI モデルの性能に大きく影響する。例えば、レントゲン写真から「がんの疑い」を判定する AI モデルの場合、「がん画像」の正解ラベル（アノテーション）を誰が付けたのかが重要な要素となる。例えば、「経験豊富な 2 名以上の専門医が協議してラベルを付けた」プロセスは、「研修医が 1 名で付けた」プロセスよりも信頼性が高いと評価される。

前処理完了後、各目的（訓練、チューニング、テスト）に応じて、収集したケースデータを分割する。データの分割は、AI モデルが訓練データだけを過剰に記憶してしまう「過学習（Over-fitting)」を防ぎ、汎用的な能力を評価するために不可欠な手順である。

Annex 22 も参考にし、データの独立性を確保する具体的な分割方法の事例を以下に示す。

- **時間的な分割：**特定の時点を区切り、それ以前のデータで訓練し、以降のデータでテストする（例：2020 年までのデータで開発、2021 年のデータでテスト)。
- **場所的な分割：**異なる施設や地理的拠点を基準にして、訓練・チューニングとテストに分割する（例：A 病院のデータで開発、B 病院のデータでテスト)。
- **物理的な分割：**異なるバッチやロットを基準にして、開発とテストに分割する。
- **特性的な分割：**各ケース固有の特性（例：製品の種類、疾患の重篤度）から一定量をランダムに開発とテストに分割する。

AI モデルの信頼性の大きな要素であるデータの独立性は、独立性の確保に対する論理的根拠を明確に説明し、透明性を維持することが求められる。

開発データとテストデータの独立性は、モデル評価の信頼性を担保するための重要な原則である。ただし、以下のような場合には、科学的に妥当な方法論として、データの一部重複使用や特殊な分割方法が選択されることがある：

- **交差検証（Cross-validation）：**データを複数のフォールドに分割し、各フォールドを順番にテストセットとして使用する方法
- **層化サンプリング（Stratified sampling）：**希少事象を含む全てのサブグループが開発・テストに適切に分布するよう制御する方法
- **Leave-one-out validation：**症例数が極めて限定的な場合の方法

これらの方法を採用する場合は、以下を文書化する：

- 当該方法を選択した科学的根拠
- データ特性（症例数、希少性等）
- 方法の詳細（分割方法、フォールド数等）
- 性能評価への影響（過学習や過大評価等へのリスク）
  - 過学習：同一被験者由来のデータを開発・テストデータの双方に利用することによる過学習のリスク
  - 過大評価：モデル選択（チューニング）に使用した検証結果を最終性能として扱ってしまうことによる性能の過大評価のリスク
- リスク管理策（保守的な閾値設定、追加検証等）

本項では、「データの管理」について、開発データを中心に述べた。続く 5.4.3 項、5.4.4 項では、これらのデータをどのように利用するのか、訓練、チューニング、テストの活動を説明する。

5.4.3 AI モデルの訓練、チューニング概要（Step4 a iii. Describe the Model training）

【概説】

モデルの開発結果概要を Credibility 評価計画書にまとめる。

- モデルの開発手法
  - 開発方法論（例：教師あり学習、教師なし学習）
  - モデルを評価するために使用された性能評価指標。すべての性能評価値は、信頼区間とともに提供される。以下に性能評価指標の例を示す。
    - Receiver Operating Characteristic（ROC）

- ✧ リコールまたは感度、特異度、陽性/陰性適中率（PPV/NPV）、真陽性/偽陽性および真陰性/偽陰性の数（例：混同行列における）
- ✧ 陽性/陰性診断尤度比（PLR/NLR）、適合率、F1 スコア

➢ 過学習または未学習を防ぐために採用された技術（例：正則化技術）

➢ ハイパーパラメータ（例：学習率、正則化係数、バッチサイズ）を適切に設定する必要がある。

➢ 損失関数はハイパーパラメータではなく、モデルアーキテクチャの設計段階で選択する構造的要素（例：クロスエントロピー損失、Mean Squared Error）であり、本書では区別して記載する。

- 事前学習済みモデル（または複数の事前学習済みモデル）の使用の有無
  - ➢ 事前学習に使用されたデータセットと、その事前学習済みモデルの開発、入手方法
- アンサンブル手法の使用
- AI モデルのキャリブレーション
  - ➢ 訓練済みモデルの Output に対する、精度および/または再現性の向上を目的とした微調整
- コンピュータソフトウェア（開発環境を構成するツールやパッケージを含む）の品質保証および管理手順と、バージョン履歴

【考察】

本ドラフトガイダンスの本セクションでは、最終的に評価テストへ送るモデルの仕様などを記載することが求められている一方で、そのモデルをどのような手順で開発すべきかについては具体的な言及がない。これが、Step4 a iii.「訓練・チューニングの概要」の理解を難しくしていると考える。そこで本項では、一般的な機械学習の手法であるグリッドサーチによる開発を例に用いて、探索的かつ反復的なモデル開発の理解促進を目指す（図 17、図 18 参照）。なお、グリッドサーチを例とする本項の内容は、ランダムサーチやベイズ最適化など他の探索的なハイパーパラメータ最適化手法にも適用できる。

AI モデルの探索的な開発活動とは、複数のモデル候補を並行して作成・比較し、最適解を見出していく活動である。この開発活動と結果を「訓練・チューニングの概要」として「Credibility 評価計画書」にまとめる必要がある。

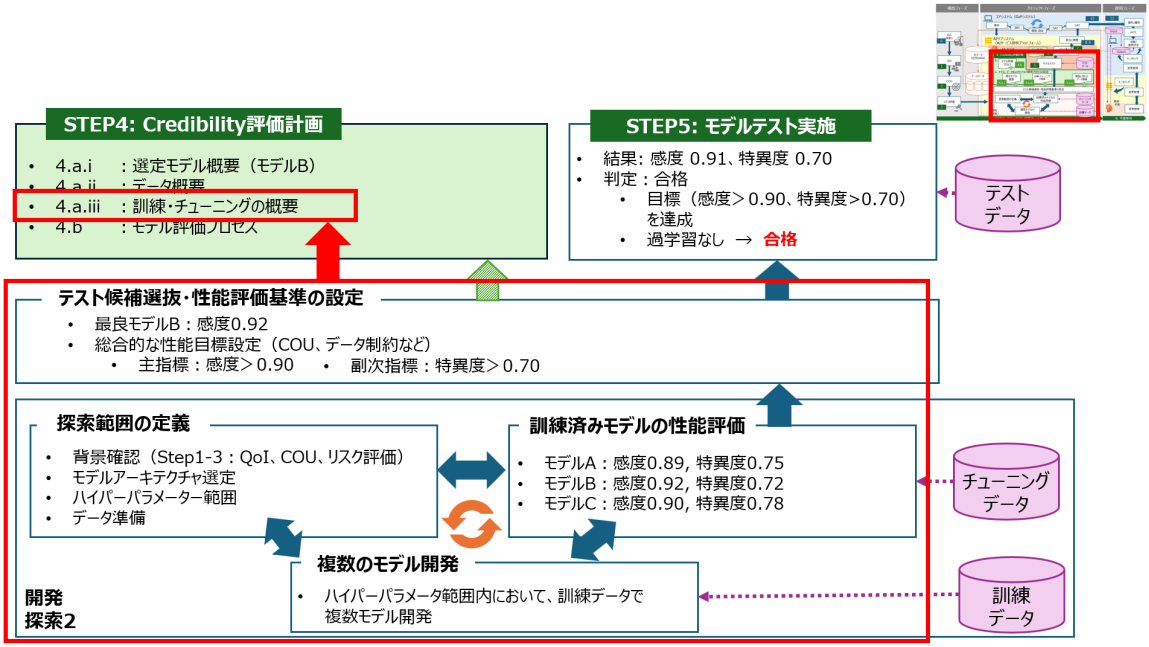

図 17：Step 4 a iii. 訓練・チューニングの概要

注：本図は図 5 の探索的開発（開発・探索 2）から Step5 の各活動を詳述したものである。

図 17 が示す通り、探索的かつ反復的な開発とは、定義された探索範囲の中で訓練データを用いて複数のモデルを開発し、そのモデルの性能をチューニングデータで評価することを反復する作業である。そして、その中で最も性能が良かったモデルをテスト候補として選抜する。

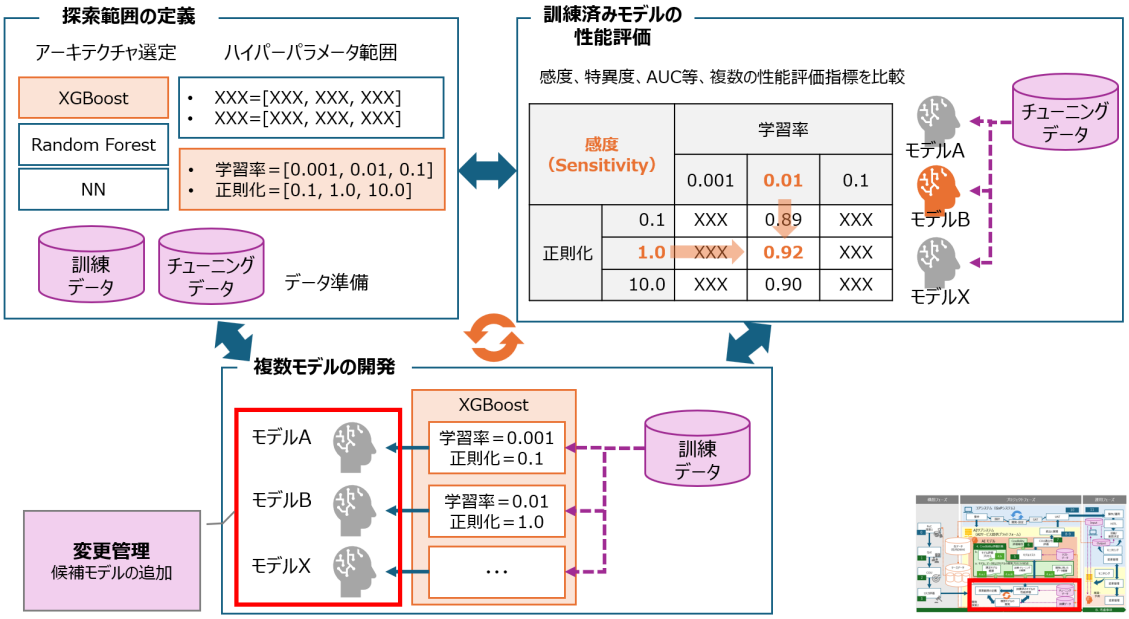

図 18：探索的・反復的な訓練およびチューニングの詳細イメージ

注：本図は図 5 の探索的開発（開発・探索 2）の活動を詳述したものである。

図 18 は、グリッドサーチによるハイパーパラメータ最適化を想定し、学習率や正則化係数の範囲と刻み幅をあらかじめ定義し、その組み合わせごとにモデルを訓練・評価する探索的な開発プロセスの一例を示している。この探索的プロセスにおける各試行は、従来の変更管理上の「変更」ではなく「候補モデルの追加」として扱われる点も、AI モデル開発に特有の考え方である。

以下に図 11 に示した Step4 に対応する変更管理の考え方を示す：

a. モデルとモデル開発プロセスの記述

- 複数のモデルパターンを並行的に構築・評価
- 各パターンは独立したモデルの記録
- 従来型の「変更」ではなく「モデルの追加」として管理
- 最終的に最適な性能を示したパターンを選定

b. モデルのテスト

- 選定されたモデルの仕様を確定・文書化
- 以降の変更は従来型の変更管理プロセスを適用
- 再学習、パラメータ調整等は変更管理の対象

このように、探索的開発では、単一のシステムに対するバージョンアップのような管理を適用せず、「探索結果をモデル記録として管理する」というアプローチを採用する。

優れた訓練データを利用しても、開発方法に瑕疵がある場合、信用できる推論や予測は得られない。AI モデルの性能を評価するためには統計やデータサイエンスの知識が不可欠であるため、統計の専門家やデータサイエンティストと協議して性能評価指標を選択、設定する。開発時に目標とした性能評価を満たした AI モデルの開発経緯を Credibility 評価計画書にまとめる。本項では、AI モデルの開発方法とその品質保証について基本的な内容を述べる。

1) モデル開発方法論（Model Learning Methodology）

モデルの性能と信用度を根幹から支える開発プロセスは、アプローチの選定から性能評価、汎化性能の確保に至るまで、一貫した科学的根拠に基づいて設計する。モデルの開発方法は、COU と利用可能なデータによって決定する。主要な開発方法は「教師あり学習」と「教師なし学習」が挙げられる。

- **教師あり学習（Supervised Learning）:** 正解ラベルが付与されたデータセットを用いて、Input と Output の関係性を学習する手法である。例えば、診断支援や不良品検知等、明確な正解が存在する COU で採用される。
- **教師なし学習（Unsupervised Learning）:** 正解ラベルのないデータから、データ内に潜在する構造やパターン（クラスタリング等）をモデル自身が発見する手法である。例えば、カルテの医師所見といった非構造化データから被験者候補を特定する、異常値や外れ値の登録状況および CRF の情報から患者背景データのクラスタリングにより、類似した病態サブグループを発見する等、明確な正解が存在しない COU で採用される。

2) 性能評価指標（Performance Metrics）

モデルの性能は、単一の指標（例：正解率）のみで評価することは不十分である。特に、偽陰性（例：疾患 A の見逃し）と偽陽性（例：健常者を疾患 A と誤診）のリスクが非対称な課題においては、多角的な評価が不可欠となる。この性能評価指標は、開発とテストの各段階で基準を設ける。

- **混同行列（Confusion Matrix）:** 性能評価の基礎となる数値で、モデルの予測結果と実際の正解の数を対比させた表である。これにより、真陽性、真陰性、偽陽性、偽陰性の数を客観的に把握できる。

表 8：混同行列表

		予測	
		陽性	**陰性**
実際の状況	**陽性**	真陽性 （True Positive：TP）	偽陰性 （False Negative：FN）
	陰性	偽陽性 （False Positive：FP）	真陰性 （True Negative：TN）

- **主要な性能評価指標：**混同行列に基づき、感度/再現率（Sensitivity/Recall）、特異度（Specificity）、適合率（Precision）、F1 スコア等の指標を算出する。これらの指標はトレードオフの関係にあることが多く、モデルの使用目的に応じて、どの指標を重視するかの根拠を明示する必要がある。

表 9：性能評価指標例表

性能評価指標	算出式	目的
感度、再現率（Sensitivity / Recall）	$\frac{TP}{TP+FN}$	取りこぼし無く Positive なデータを正しく Positive と予測できていることを示す指標。
特異度（Specificity）	$\frac{TN}{TN+FP}$	取りこぼし無く Negative なデータを正しく Negative と予測できていることを示す指標。
適合率（Precision）	$\frac{TP}{TP+FP}$	Positive と予測したとき、実際に Positive だった割合。この値が高い場合、取りこぼす Positive が多く発生する可能性がある。
F1 スコア（F1 Score）	$\frac{2\times（適合率\times感度）}{（適合率+感度）}$	適合率と感度の調和平均。双方が高い場合、1 に近くなり、どちらかが低い場合、0 に近づく。

- **信頼区間（Confidence Intervals）：**算出した性能評価指標が、統計的にどの程度のばらつきがあるかを示す区間である。点推定値（例：感度 95%）だけでなく信頼区間を併記することで、AI モデルの Output の総合的な評価結果に対する頑健性を示すことができる。

1) モデルの最適化と性能評価指標の確立

a) 過学習、未学習の防止 (Techniques to prevent overfitting / underfitting)

AI モデルは、学習したデータだけでなく未知のデータに対しても安定した性能を発揮する「汎化性能」を持つ必要がある。汎化性能により、実環境でも耐え得る品質と性能を発揮する AI を開発することができる。そのための技術として、モデルの複雑さにペナルティを課す「正則化（Regularization）」等が用いられる。開発においては、過学習防止策を講じたかを具体的に記述することが求められる。

さらに、学習率や正則化係数等、モデルの開発プロセスを制御するハイパーパラメータは、モデル性能を大きく左右する。これらのハイパーパラメータを系統的に探索・チューニングし、最終的に採用した値の選定根拠を文書化する必要がある。

b) 性能評価指標とその合格基準の設定プロセス

性能評価指標とその合格基準は、COU に基づき以下のプロセスで決定する。

- COU 要求の明確化（Step1-3 で実施済み）
  - Step1-3 で定義した COU に基づき、性能要求の「優先順位」を明確にする。例えば、「偽陰性（見逃し）の最小化が最優先（感度重視）」「サブグループ間での公平な性能担保」「予測の不確実性の定量化」等である。この優先順位が、性能評価指標選定の基準となる。
- 性能評価指標候補の検討
  - COU 要求を満たし得る複数の性能評価指標候補を検討する。例えば、Sensitivity（感度）、Specificity（特異度）、F1 Score、AUC（Area Under the Curve）、Balanced Accuracy 等である。各指標が COU の要求とどのように対応するかを明確にする。
- データ特性の確認
  - 実際の開発データの特性（陽性率、サンプルサイズ、データ不均衡の程度、サブグループ分布等）を把握する。これらの特性は、統計的に信頼できる評価を行うための性能評価指標の選択に影響する。
- 探索的評価と最適化
  - 複数のモデルアーキテクチャ、データ前処理方法、性能評価指標の組み合わせを試行し、COU を最もよく満たす組み合わせを特定する。この過程では、以下を記録する。
    - 検討した性能評価指標候補（例：Sensitivity 単独、F1 Score、Sensitivity + Specificity 組み合わせ等）
    - 各性能評価指標が COU 要求を満たす程度
    - データ制約下での達成可能性

- ✧ 最終的な性能評価指標（組み合わせを含む）を選定した根拠
- ➢ 指標・閾値の確定と文書化最終的に選定した性能評価指標と合格基準について、以下を文書化する。
  - ✧ 選定した指標（例：主指標 Sensitivity ≥ 95%、副次指標 Specificity ≥ 70%）
  - ✧ 採用した性能評価指標が COU に最適と評価した理由（COU との論理的整合性）
  - ✧ 他の候補を採用しなかった理由
  - ✧ 閾値設定の根拠（臨床的許容限界、データ特性、統計的妥当性等）
  - ✧ 信頼区間の考慮
  - ✧ 統計専門家またはデータサイエンティストによるレビュー結果

c) 重要な原則

このアプローチは、機械学習の探索的性質を許容しつつ、COU への忠実性と科学的厳格性を両立させるものである。重要な点は、「恣意的に指標を選ぶ」のではなく、「COU 要求と特定のデータ制約下で最適な指標を論理的に導き出し、そのプロセスを透明化、文書化する」ことである。

特に以下の場合は、性能評価指標選定の妥当性について規制当局との事前協議を強く推奨する。

- リスク評価結果＝「高」（DC：「高」×MI：「高」）の COU
- 新規の AI 手法
- 複雑な HITL プロセスを含む場合
- データセットのサブグループ間で性能が大きく異なる場合

2) 訓練済みモデル（Pre-trained Model）の利用

自社によるモデル開発ではなく、既存の AI モデルを活用することは効率的であるが、当該 AI モデルを利用できる根拠を明確にすることが求められる。訓練済みモデルを基盤とし、特定のタスク向けに自社データで追加学習（ファインチューニング）を行う手法は広く採用されている。しかし、GxP 領域で訓練済みモデルを利用する場合、当該訓練済みモデルに利用しているデータおよびバージョン管理を含めたリスク評価が必要である。

訓練済みモデルを利用する場合、以下のような情報を明確にすることで、AI モデルの透明性を維持することができると考えるが、FDA との協議を要する。

- 利用した事前訓練済みモデルの名称とバージョン
- モデルの入手元（公式レジストリ等）
- 元のモデルが開発に用いたデータセットの特性とバイアスに関する情報

3) アンサンブル法（Ensemble Methods）

Output の品質を向上させるひとつの手法として、複数の異なるモデルを組み合わせ、各モデルの予測・推論を基に、総合的な判断を下すことで単一モデルよりも高い精度と頑健性を目指すアンサンブル法がある。これを採用した場合、どのようなモデルを、どのような方法（例：多数決、平均化）で組み合わせたかを明確にする。

4) 最終調整と品質保証

開発が完了したモデルは、実用に向けて最終的な調整と、開発プロセス全体の品質保証が求められる。

- **キャリブレーション：**モデルが出力する Output に対する個別の「信頼性スコア」（例：90%の確率で陽性）を実現できるように補正や微調整するプロセスである。Output に対する適切なキャリブレーション（補正）を行うことで、モデルが提示する結果への信頼性確率が向上し、臨床現場等、利用環境により即した Output の出力が可能になる。
- **ソフトウェアの品質保証と再現性：**AI モデルの信頼性は、そのプロセスの再現性に大きく依存する。したがって、開発に使用した環境やツールを記録し、管理する。
  - **開発環境の記録：**プログラミング言語（例: Python）、主要なライブラリ、および関連ツール自体の構成とバージョンを記録する。
  - **バージョン管理：**ソースコード、設定ファイル、実行手順の変更履歴を追跡するため、バージョン管理システムを利用し、その管理体制を示す。また、乱数シード（Random Seed）の固定と記録もモデル再現性の担保に不可欠である。使用したフレームワーク（例：Python、TensorFlow®等）における乱数シードの設定値を、他のパラメータとあわせて開発記録に残すことが望ましい

5.4.4 AI モデルのテスト（Step4 b. Describe the model evaluation process）

【概説】

本項は、訓練済みモデルのテストについて概説する。AI モデルのテストは、COU に対してモデルの性能が十分であるかを、テストデータを用いて評価する活動である。

テストデータは開発データから独立し、開発中に利用してはならない。テストデータは開発後に AI モデルの性能を評価するために使用される。開発データと同様に、これらのデータも「用途に適して（fit for use)」いることが求められる。

Credibility 評価計画書に含めるテスト関連の情報は以下がある。

- テストデータの収集、処理、アノテーション、保管、管理手順、および、AI モデルの使用目的
  - 開発データとテストデータの独立性
    - データの独立性は、異なる臨床試験や医療システムからのデータ、あるいは異なるバッチや製品を用いて取得されたデータを使用することで達成され得る。
  - 開発とテスト間でデータの重複使用があった場合は、それらのデータの使用方法と、その使用が適切であった正当性
  - 該当する場合、テストデータを作成するために使用された参照方法（reference method）と参照方法の性能の要約
- テストデータの COU への適合状況
  - 予測モデルが過去の開発データを用いて開発された場合、その開発データが実環境で遭遇するデータと異なっていれば、AI モデルは COU で期待通りの性能を発揮しない可能性がある（データドリフト）
- テストデータによるモデルの予測と実測値の一致性
- 選択したモデル評価方法の論理的根拠
  - その評価方法が、使用されたモデリング手法および COU に適用可能であることを説明すること。「HITL」を伴う COU の場合、評価方法はモデル単体の性能だけでなく、人間と AI をひとつのチームと捉えての性能を考慮していること
- 当該「テスト方法や性能評価指標」を選択、計画した正当性を明確に記述する。さらに、その評価方法が、開発時の「モデル構築手法」や「使用目的（COU）」に対して、適切であることを説明する。
  - Receiver Operating Characteristic（ROC）
  - リコールまたは感度、特異度、陽性/陰性適中率（PPV/NPV）、真陽性/偽陽性および真陰性/偽陰性の数（例：混同行列における）
  - 陽性/陰性診断尤度比（PLR/NLR）、適合率、F1 スコア
  - モデル予測の不確実性と信頼度レベルが推定されたプロセス
- 潜在的なバイアスを含む、モデリングアプローチの限界
- コード検証のための品質保証および管理手順
  - エラーや異常の解決を含む（例：コードにエラーがないこと、計算が正確であること）。

【考察】

従来の CSV では仕様を事前に確定してから開発を進めるが、AI モデルの開発は探索的である。目標とする性能基準を満たしたモデルが、最終的な仕様となる。開発プロセスが探索

的であればあるほど、最終評価には高い客観性が求められる。FDA AI ガイダンスと Annex 22 がテストデータの完全独立性を強調する理由はここにある。

AI モデル開発ライフサイクルの最終関門は、完全に開発が完了したモデルの客観的な性能評価、すなわちテストである。このプロセスは、モデルが学習で獲得した知識が、未知のデータに対しても有効に機能する「汎化性能」を有しているかを検証するものである。これは、訓練、チューニングを終えたモデルが、実使用環境で、意図した性能を安定して発揮できるかを評価する、科学的に最も重要なプロセスである。図 19 に示すように、「Credibility 評価計画書」は本活動を計画するための文書である。

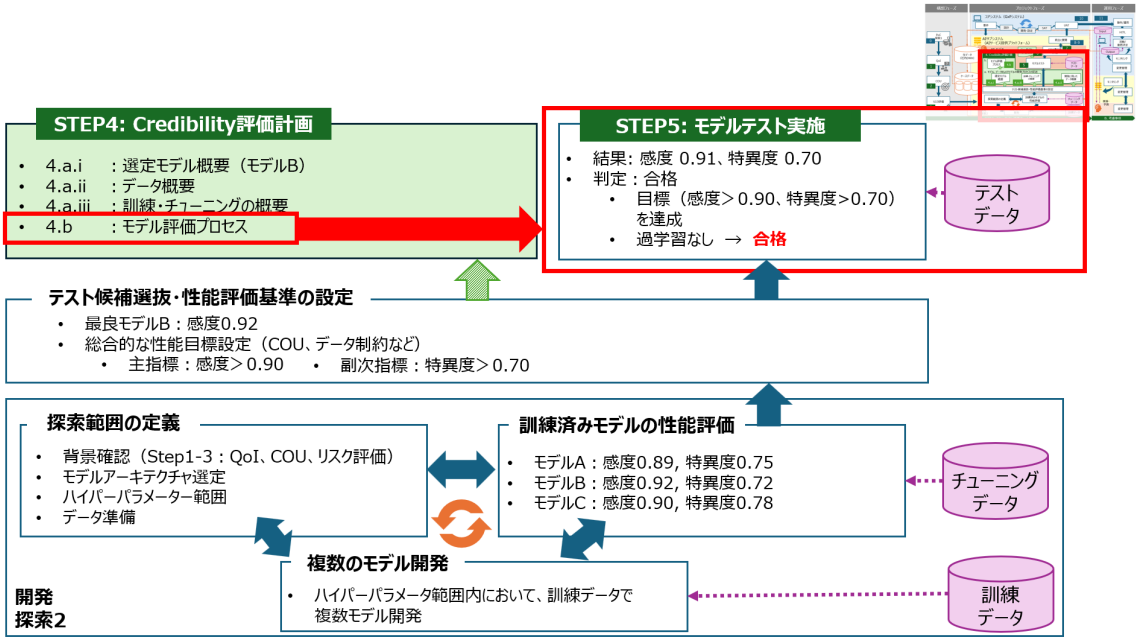

図 19：Step 4 b. モデル評価プロセス

注：本図は図 5 の探索的開発（開発・探索 2）から Step5 の各活動を詳述したものである。

AI モデルに対するテストも通常の GxP システム（コアシステム）の開発同様、テスト開始前までに、テスト計画である Credibility 評価計画書を承認する。テストでは、以下のような観点を含めることを想定する。

- COU、使用目的
- テストデータの管理
  - テストデータの独立性
  - テスト担当者の独立性
  - テストデータの参照を含む監査証跡の保持
  - テストデータの利用方法
- 評価方法と妥当性

- 人間の関与を含めた AI の利用プロセスとテスト範囲
- AI の予測結果の不確実性、確信度の推測プロセス
- AI の特徴量の提示

- テストスクリプト
- 性能評価指標
  - 汎化性能の評価
  - 利用する性能評価指標の妥当性
  - 計算方法
- 受け入れ基準
- 最終評価
  - AI の限界、制限等

本項では、AI モデルの最終評価における Credibility 評価の要点、特に下記 3 点を詳述する。

1) テストデータの要件と管理
2) 評価方法の妥当性
3) 性能評価指標

1) テストデータセットの要件と管理

5.4.2 項でも述べたように、評価の信頼性は、使用するテストデータの品質と管理体制に大きく依存する。テストでは「テストデータの独立性」を中心に説明する。

- **データの独立性確保：**FDA AI ガイダンス、Annex 22 共通でデータ自体の独立性確保を要求している。
  - テストデータはモデルの訓練・チューニングデータと完全分離（データリークの防止）
  - 時間的、地理的、物理的な分割により独立性を確保する。
  - データベースでのアクセス制御と監査証跡
- **担当者の独立性確保：**Annex 22 では、データにアクセスする担当者の独立性も求めている。開発担当者とテスト担当者を完全に分け、データへのアクセス権およびその監査証跡での実証できることが理想である。しかし、組織の規模やリソースによっては担当者を完全に分けることが難しい場合もある。このような場合でも、テストデータへのアクセスを記録・承認制とすること、ハイパーパラメータ変更やモデル再開発時に第三者レビュー（4-eyes principle）を行うこと等の、開発・テスト活動の時系列記録を保持すること等により、実質的な独立性と透明性を確保することが考えられる。

- **データ重複使用の取り扱い：**
  - 開発とテストデータに重複があれば、それはモデルの真の実力を測るテストではなく、開発内容の「答え合わせ」に過ぎず、過学習（Over-fitting）を見過ごすリスクがある。この重複には、テストで「Fail」と評価した後に、このテストデータを訓練やチューニングに利用することも含まれる。
  - 利用できるデータ数が限定されている場合は、開発データとテストデータの重複使用が避けられないこともあり得る。交差検証（Cross-validation）等、モデルの訓練・チューニングにデータを重複して利用する場合は、5.4.2 項で示したように、「当該方法を選択した科学的根拠」等を文書化する。
- **テストデータの利用履歴管理**
  - Annex 22 では、どのデータがいつ（開発パターン情報等の AI モデルの特定情報）、何回テストに利用されたのか、を記録し、テストの妥当性を示す。テストデータの利用状況には、テストデータを「利用しなかった」状況も含む。予め準備していたテストデータを利用しなかった場合は、テストの逸脱として扱い、結果への影響を評価する。

2) 評価手法の妥当性

評価手法は、モデルの特性と実際に利用される環境、COU に適合したものでなければならない。

- **説明可能性の活用**
  - AI モデルを評価する場合、特徴量を可視化（SHAP 値、LIME、ヒートマップ等）し、AI モデルが何を根拠に結果を予測したのかをレビューすることは、モデルが関連性のある適切な特徴量に基づいて、かつリスクに基づいて決定を下していることを保証するエビデンスになる。特徴量の重要度や寄与度が解釈可能（interpretable）、または、算出可能である場合は、AI モデルの説明可能性のために活用する。
- **HITL を含む総合的評価**
  - COU が「HITL」を想定している場合、評価方法はモデル単独の性能だけではなく、人間と AI システムを総合的に評価する。Annex 22 では、少なくとも、AI システムを導入する場合、AI システムを利用していないプロセスに対し、性能が同等以上であることが求められる。つまり、AI モデルによって置き換えられるプロセスの性能を比較できることが必要である。
  - 診断支援 AI システムのように、最終判断を医師が下すシステムの場合、AI モデル単体の性能評価だけでなく、AI システムの支援によって医師のパフォーマンス（診断精度、作業時間等）変化を評価することが不可欠である。AI システムを利用することにより、人間のミスを誘発するリスクもあるため、人

間のプロセスを含めて、総合的な運用プロセスや AI システムの利用制限を考慮する。

- 本評価はコアシステムのユーザ受入試験（UAT、PQ 相当）で実施することも許容されると考える。

- **信頼性スコアと不確実性の定量化**
  - モデル予測の「信頼性スコア」を定量的に示すことは、Output を用いるユーザが適切に解釈し、過信を防ぐ上で極めて有益である。予測結果の自信の度合いである信頼性スコアも付記した（例：信頼性スコア 99%の陽性、信頼性スコア 50%の陽性）AI システムの Output を参照する場合、予測結果を基に最終判断を下す際の医師の意識は全く異なったものになる。
  - AI が予測するために必要な情報を得るための閾値等の制限がある場合、適切な条件下で Output を出力できるように、適切な閾値設定を設ける。信頼性スコアが極端に低い場合には、信頼性の低い予測を行うのではなく、「予測不可」の Output を出力することも検討すべきである。

3) 性能評価指標

モデルの性能は、COU におけるリスクを考慮した複数の指標を用いて評価する。性能評価指標と受け入れ基準を予め計画し、これらの指標等を満たすことで、モデルの性能が COU に適していること、受け入れ可否を評価できる。COU が複数のサブグループや製品カテゴリにまたがる場合（例：入院と外来、製品 A と製品 B）には、個別のケースごとに受け入れ基準を設定する。

また、高リスクの COU では、総合的な性能評価指標に加えて、主要な患者サブグループや製造条件ごとの性能評価指標と信頼区間を確認し、特定の集団や条件で性能が有意に劣らないか（バイアスや局所的な弱点がないか）を評価することが推奨される。

以下に AI モデルのテストの主な性能評価のための観点を述べる。

- **汎化性能の評価：**モデルが使用環境に適合し、「汎化性能が高い」こと、すなわち、モデルが新しいデータに対して十分な性能を持つことを評価する
  - 過学習、未学習の可能性を検出することが含まれる
  - 「特定の患者群では性能が低下する可能性がある」等、モデルの既知の弱点、苦手な状況、潜在的なバイアスを開示し、透明性を確保する
- **性能評価指標：**混同行列、および、感度、特異度、精度、F1 スコア等、本 COU に適した性能評価指標を提示する（5.4.3 項参照）
  - サンプルサイズの具体的な算出方法は、COU、モデル種別、性能評価指標によって異なる。実務上は、訓練データで得られた事象の頻度やイベント数（例：陽性症例数、不良品数）を参考にしつつ、「想定される性能値」と「許

容できる誤差幅・信頼水準」を統計担当者と協議しながら、テストデータに必要な症例数／サンプル数の目安を見積もることが望ましい。

- **信頼区間：**全ての性能評価値には、統計的なばらつきを示す信頼区間を併記し、結果の信頼性を担保する（5.4.3 項参照）

その他、Credibility 評価の観点から、ツール等を評価に用いた場合には、その信頼性も求められる。

- **評価コードの品質保証：**評価プロセスの計算に使用したプログラムコード自体の品質保証も必須である。コードレビューやユニットテスト等を通じて、コードにバグがなく、計算が正確であることをどのように検証したか、その手順を明確に説明する必要がある。これは、評価結果そのものの信頼性を保証する。

特に LLM や生成 AI を規制上の意思決定支援に用いる場合には、代表的なプロンプトや Input に対する Output について、専門家による系統的なレビューと、プロンプトおよびモデルバージョンを固定したうえでの再現性の確認を通じて、当該 COU における信頼性を検証することが重要である。

5.5 AI モデルの Credibility 評価計画の実行と結果の文書化 （Step5: Execute the Plan, Step6: Document the Results of the Credibility Assessment Plan and Discuss Deviations From the Plan）

【概説】

Step5 は Step4 の Credibility 評価計画の実行、Step6 は実行結果の文書化である。Step6 では、Step1～4 の計画の結果、逸脱をまとめて、報告書にまとめる。この Credibility 評価報告書は、COU に適した AI モデルを開発した保証を示す文書である。Credibility 評価報告書は FDA から求められることがある。

【考察】

「Credibility 評価計画（5.4 項）」の策定後、AI モデルのプロジェクトフェーズの最終段階として、計画の実行と結果の文書化を行う（図 11 参照）。Step5、6 の 2 つの Step は、AI モデルが意図した目的（COU）において信頼できるものであることを、客観的な証拠をもって示すための活動である。FDA AI ガイダンスでは、これらの活動に着手する前に、規制当局と計画内容や報告方法について協議し、期待値をすり合わせておくことの重要性を強調している。

1) Credibility 評価計画の実行

計画書に記載された「b. モデルのテスト（データ管理、モデル評価、性能評価等）」に従って実行する。Credibility 評価計画からの逸脱が発生した場合は、その内容と理由をすべて記録しておく。

- **規制当局との協議**

  FDA AI ガイダンスでは、計画を実行する前に規制当局と協議することが、双方にとって有益であると示唆している。事前の協議を通じて、以下のような利点が得られる。

  - **期待値の整合:** 計画した評価活動の妥当性（モデルのリスクや COU に見合っているか）について規制当局と期待値を合わせることができる。これにより、後工程での手戻りや、承認申請時の指摘リスクを大幅に低減することが可能となる。
  - **潜在的課題の特定:** 企業側だけでは予見しきれない潜在的な課題を特定し、それらにどのように対処すべきかについて、規制当局と早期に議論・検討することができる。

2) 実行結果の文書化と逸脱の考察

文書化の目的は、計画書に従って実行した活動結果を「Credibility 評価報告書」として体系的に文書化することである。この報告書は、AI モデルの信頼性に関する一連の活動の成果物である。

本報告書には、主に以下の内容を含めることが求められる。

- **評価結果の要約：**Step5 で得られた全ての結果を客観的に記述する。
- **計画からの逸脱に関する考察：**計画と異なる手順や手法を用いた場合、その事実、理由、およびその変更が評価結果に与える影響を明確に考察し、記述する。逸脱を隠さず、その妥当性を論理的に説明することが信頼性確保の観点から重要である。
- **関連情報の参照：**リスク評価と計画時に定義した QoI、COU、リスク評価の結果の要約を含め、評価の前提条件を明確にする。

作成した Credibility 評価報告書の規制当局への提出は、AI モデルのリスクや申請の種類によって異なる。例えば、規制申請や対面助言等のパッケージの一部として正式に提出する場合もあれば、査察時に提示できるよう社内で保管する場合も考えられる。この提出戦略についても、AI モデルの Credibility 評価の実行前に規制当局と合意しておくことを推奨する。

5.6 AI モデルの COU 適合性評価（Step7: Determine the Adequacy of the AI Model for the Context of Use）

【概説】

Credibility 評価計画に基づき、AI モデルの適合性を評価した結果、当該モデルが特定の COU に対して適合すると評価できれば、運用を開始する。しかし、特定の COU に対して適合しないと評価することもあり得る。当該モデルの信頼性が不十分であると判断した場合であっても、リスクに応じて、複数の対応策が検討できる。

FDA AI ガイダンスは 5 つの代替案を提示している。

(1) AI モデルからの Output と併せて、追加の Evidentiary Sources を組み込むことにより、モデルの影響度を格下げする
(2) Credibility 評価活動の厳格性を高めるか、または追加の開発データを加えることによってモデルの出力を改善する
(3) リスクを軽減するために適切な管理策を確立する
(4) モデリングのアプローチを変更する
(5) AI モデルの出力の信頼性が COU に対して不十分であると判断し、その結果、モデルの COU を却下、または反復的なやり方で修正する。

【考察】

当該モデルの信頼性が不十分であると判断した場合の 5 つの対応策の例を示す。

表 10：AI モデルの信頼性が不足している場合の対応策

対応策	例
(1) 追加の Evidentiary Sources の組み込み	AI システム予測に依存するのではなく、AI システム出力結果を参考値にする
(2) Credibility 評価活動の厳格性を上げる、または追加の開発データを追加	Credibility 評価の厳格さを高めるため、単一から複数の性能評価指標を追加し、エビデンスの質を上げる。または、モデルの汎化性を向上するため、類似試験のデータに加え、患者レジストリ等のデータを追加して再学習を行う
(3) リスクを軽減するために適切な管理策の確立	AI モデルはそのまま利用するが、運用プロセス上、モニタリング数を増やす等、性能を人間の追加プロセスで補完する
(4) モデリングのアプローチを変更	根本的に異なる技術や設計思想でモデルを再構築する
(5) モデルの COU を却下、または反復的なやり方で修正	本 AI モデルの利用の中止、または COU の範囲を限定的な利用に狭める

FDA は、AI モデルが常に完璧な結果を出力できるとは前提にしていない。このため、AI システムの導入は、一度失敗すれば終了ではなく、継続的な開発ルートを残している。AI モデルの限界を正しく理解し、信頼性が不十分だった場合には、どのように、患者さんの安全や製品の品質、結果の信頼性を保証するのか、そのリスクを管理したうえで、再度開発に取り組む、または、運用に反映させることが重要である。

5.7 AI サブシステムの統合、展開 (Step8: AI Model Implementation, Step9: Model Integration and Deployment)

【考察】

本項の活動は、FDA AI ガイダンスにはないが、本書独自に GAMP 5 第 2 版をベースに追加した Step である。AI モデルを AI サブシステムに構築し、AI モデルと GxP システムを接続する。ソースシステムから出力される Input データを AI モデルがそのまま利用できない場合、AI サブシステムの機能で処理することもある。

AI サブシステムのメイン機能はデータの受け渡しであるため、ネットワークの接続やデータの授受を含めた構成設定とその検証が主な活動と想定される。

AI サブシステム構築・運用においては、実務的に以下のような考慮点が含まれる。

- インターフェース仕様の確定
  - Input データのフォーマット、範囲、欠損値の扱い
  - Output データのフォーマット、信頼性スコアの有無
  - エラーハンドリングの方法
- AI サブシステムの構成管理
  - モデルバージョン管理（モデルファイル、重み、設定）
  - ライブラリ・依存関係の固定
  - 推論環境の再現性確保
    - 推論環境（AI モデルの実行環境）を安定して再現するための技術として、主に、「コンテナ」と呼ばれる仕組みが用いられる。例えば、Docker®などのコンテナ技術を用いることで、推論に必要な OS、ライブラリ、ランタイム環境等をひとつのコンテナイメージとしてパッケージ化できる。これにより、異なるサーバーやクラウド環境に展開した場合でも、同一のコンテナイメージを用いることで、実質的に同一の推論環境を再現することが可能となる。
    - GxP 分野においては、トレーサビリティ（追跡可能性）およびデータインテグリティの観点からも、推論環境を一意に特定できることは非常に重要である。コンテナイメージをバージョン管理し、特定の AI モデル出力と紐

づけて記録することで、「どの環境でどのバージョンのモデルが実行されたか」を後から検証可能となり、監査対応や再解析時の説明性を高めることができる。

- 性能要件の確認
  - 推論速度（レイテンシ）の要件
  - スループット（同時処理数）
  - リソース使用量（CPU、GPU、メモリ）
- セキュリティとアクセス制御
  - AI モデルへの不正アクセス防止
  - Input データの機密性保護
  - 監査証跡の記録
- フェイルセーフ機能
  - モデル推論失敗時の代替処理
  - 信頼性スコアが閾値を下回る場合のエスカレーション
  - 異常 Input 検知時の警告

GAMP 5 第 2 版では、従来型の V モデルからリスクベースのテストアプローチへの転換が強調されており、IQ/OQ/PQ といった用語自体は必須ではない。しかし、現場の CSV 実務では依然として IQ/OQ/PQ という区分が広く用いられているため、本書では理解のしやすさを優先し、以下のように整理する。AI サブシステムのインストール確認が IQ、機能・接続性の確認が OQ、コアシステムと統合したエンドツーエンドの業務プロセス確認が PQ（ユーザ受入試験：UAT）に相当する活動として位置づける。ただし、各社の CSV 手順に応じてテスト活動の名称と範囲を定義する。

AI サブシステムに対する GAMP 5 第 2 版でのカテゴリ分類は、その取得形態に依存する。カスタム開発した AI サブシステムは Category 5（カスタムソフトウェア）に該当することが多い。商用プラットフォームを主にパラメータや構成要素の設定によって構成して利用する AI サブシステムは Category 4（構成可能ソフトウェア）、設定要素が比較的限定された商用 AI サービスを利用する場合は Category 3（構成不要ソフトウェア）に相当する可能性がある。AI サブシステムの Category 分類は、AI プラットフォームの種類と構成方法、ならびにモデルリスクの評価結果を踏まえ、各社 QMS で定めることが望ましい。

5.8 コアシステムのバリデーション（Step10: Validation of the Core System）

【考察】

本項も FDA AI ガイダンスにはないが、GAMP 5 第 2 版をベースに追加した Step である。Input の収集と Output の提供を担うコアシステムである GxP システムの CSV 活動が該

当する。ただし、AI モデル、AI サブシステムを統合したコンピュータ化システム全体に対する Intended Use への適合性を検証する。基本的な信頼性保証のアプローチは各社で定めている CSV、CSA の手順に則る。

AI システムに関する活動としては、コアシステム単体のユーザ受入試験（UAT、PQ 相当）だけでなく、Input 情報の提供から Output の活用までの一連のプロセスを検証する。特に以下に留意する。

- Input に対して期待した Output が出力されること
- Output を活用できるプロセスが適切であること

AI サブシステムと連携する GxP コアシステムの CSV においては、実務的に以下のような考慮点が含まれる。

- データパイプラインの信頼性
  - Input データ収集の自動化プロセス
  - データ前処理の再現性
  - データ品質チェック機能
- AI サブシステムとの統合テスト
  - 正常系: 想定される Input 範囲での E2E テスト
  - 異常系: 想定外 Input（欠損、範囲外、異常値）の処理
  - 境界値: モデルが不確実な領域での HITL 動作確認
- 監査証跡
  - どのモデルバージョンで推論したか
  - Input データとモデル推論結果の紐付け
  - HITL による承認・却下の記録

5.9 運用フェーズ（Step11: Operation and Maintenance）

【概説】

Step11 は、PoC を含む Step0 から Step10 に続く最終ステップとして、AI モデルの本番運用フェーズ全体を指す。本書では、この Step11 の中に、リリース直後の集中的な立ち上げ期間である「ハイパーケア」と、その後の「定常運用（ライフサイクルの維持)」の 2 つのフェーズが含まれるものとして整理する。

それを踏まえて、FDA AI ガイダンスの「B. Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use」を概説する。

FDA AI ガイダンスにおける「ライフサイクル維持」とは、AI モデルが医薬品ライフサイクルを通じてその利用状況に対し「用途に適した」状態を保つための一連の計画的な管理活動である。AI モデルは、データ駆動型であり、再開発により性能が変化する特性を持つ。

特に、運用中に新しいデータで継続的に学習する Adaptive AI モデルは、その Output が一定の性能を継続的に維持することが不可欠となる。

これを実現するため、モデルの性能評価指標をリスクベースで継続的にモニタリングし、偶発的・意図的な変更を医薬品品質システム（PQS）内で管理する。例えば、製造プロセスの変更が AI モデルの性能に与える影響等を評価し、その影響度に応じてモデルの再開発や再テストを実施する。性能に影響を及ぼす重要な変更は、規制要件に従って規制当局へ報告する必要がある。

ライフサイクル維持に関する詳細な計画（モニタリング頻度や再テストのトリガー等）は、PQS の一部として文書化し、レビュー可能な状態に保つべきである。さらに、ICH Q12 の Established Conditions（確立された条件）といったツールを活用することも推奨される。モデル関連要素を Established Conditions として定義し、その変更管理計画を販売承認申請に含めることで、どのような変更が事前承認不要となるかについて、あらかじめ規制当局の見解を得ることが可能となる。これにより、企業は規制上の予見性を高め、効率的にモデルの維持管理を進めることができる。

【考察】

FDA AI ガイダンスでは、運用フェーズに関しては「B. Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use」として考慮事項が述べられている。本考慮事項を含めて、本書では、Step11 にリリース直後の混乱期をフォローするハイパーケアと定常運用（ライフサイクルの維持）を統合した一つの運用フェーズとして位置づけた。本項では、これらの Step の活動に関する「特別な考慮事項」について、その規制的な背景、内在するリスク、管理手法を述べる。

AI モデルの信頼性は、AI のライフサイクルを通して保証することで担保される点では、従来の CSV 活動と同じである。しかし、従来の静的なコンピュータ化システムと比べて、AI モデルの取り扱う範囲が拡張している。

- **時間的な拡張：**AI の利用期間の長期化
- **データの拡張：**固定データ等の静的なデータから、日々の動的なデータを扱うことに伴う対象データの拡張
- **役割の拡張：**報告書を一度作成するだけの役割から、常時モニタリングし、オペレータの判断を継続的に支援する役割への拡張

これらの要求の拡張に伴い、AI の Credibility（信用度）を確立する活動も動的なマネジメントへのパラダイムシフトが求められている。

AI モデル、特に製造プロセスで継続的に使用されるモデルは、その開発基盤となったデータと、実運用で利用されるデータとの間に生じるギャップにより、時間経過と共に性能が

低下する「モデルドリフト」のリスクを内包する。この性能低下は、例えば外観検査における不良品の見逃し（偽陰性）といった形で顕在化し、製品品質および患者さんの安全、結果の信頼性に対する直接的なハザードになり得る。したがって、COU への適合性を 7-Step で評価するだけでなく、モデルの性能を継続的にモニタリング・維持するライフサイクルアプローチが必須となる。

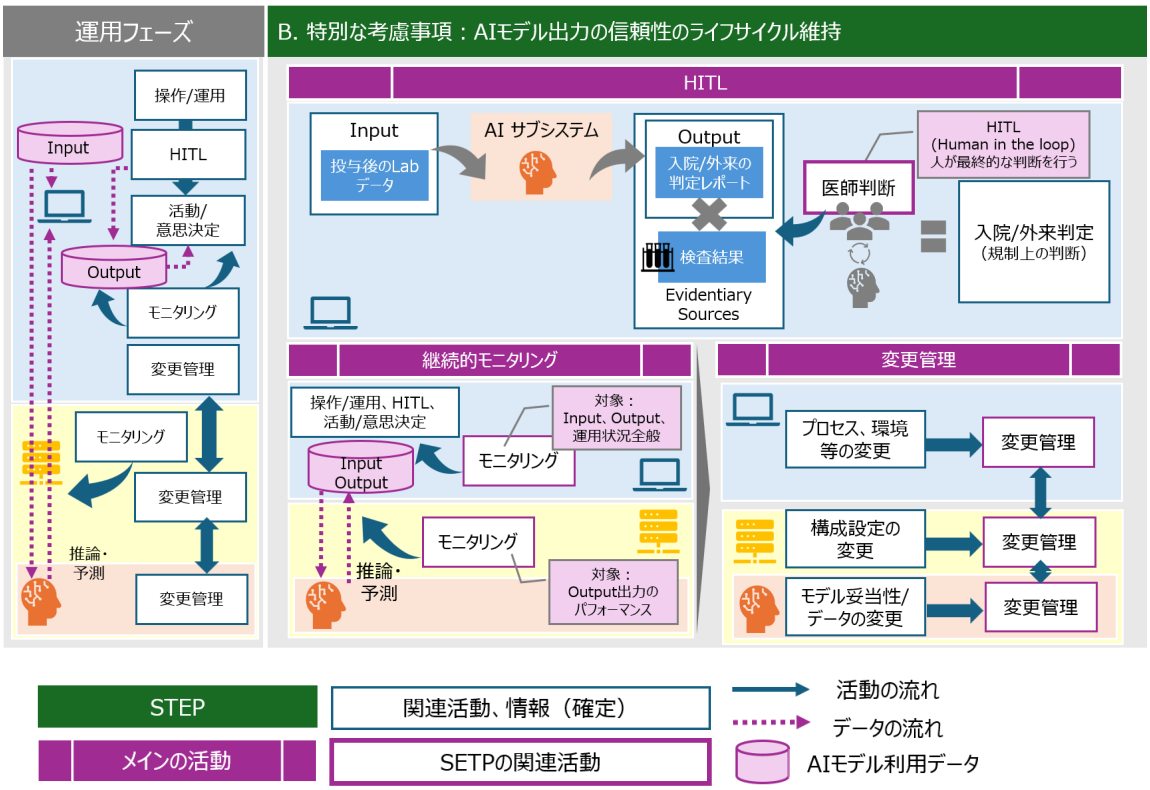

図 20：運用フェーズの特別な考慮事項

以下に、主な管理手法について、詳述する。

- HITL
  - HITL の目的は、AI モデルの Output を使った意思決定のリスクを下げることである。HITL の代表的な活動は、以下の 2 点が挙げられる。
    - **AI モデルの Output を人間がレビューする活動**：プロセスの重要度とモデルのテストレベルに応じて、AI モデルからの全ての、または予め設けた基準に該当する Output をレビューすることを想定する。
    - **AI モデルとは別の並行した活動の結果を利用する活動**：5.2 項 表 4 の GMP 事例にあるように、「担当者が抜き取りで行う独立した検証」を行うような事例が該当する。

- AI モデルの Output を利用して、人間が最終判断を行う HITL の場合、かつ、HITL を COU に組み込み、モデルのテストが軽減されていた場合、このプロセスの記録を保持する。
- ハイパーケア時に、HITL が適切に運用できない等のリスクが検出された場合は、運用プロセスを見直すことも考慮する。

- 継続的モニタリング
  - Output がモデルの性能評価指標の範囲内にあること、入力データがモデルの開発時の訓練データの範囲内にあること（Input Sample Space の確認）を示す、そして実運用で利用する環境や条件を網羅する体系的なモニタリング計画を策定し、実行する。これにより、モデルドリフトの早期検知と、それに対する是正措置・予防措置 (CAPA) を開始できる。
  - Input データは、当初の COU の範囲内であり、性能評価指標の範囲内であることをモニタリングするための性能評価指標を定める。代表的な統計的モニタリング手法として、入力特徴量の分布変化を定量化する PSI（Population Stability Index）、出力スコア分布の時系列比較（KS 検定等）、混同行列の推移追跡等がある。いずれの手法を採用する場合も、アラートを発するための閾値とその根拠、および閾値超過時のエスカレーション手順（例：調査開始、再学習の検討、運用一時停止等）をモニタリング計画に事前に定めておくことが重要である。
  - ハイパーケア時においては、ユーザからのフィードバックも重要なモニタリング項目となる。例えば、画面上の表示が分かりにくい、特定の値の判断に迷うといったユーザの声を確認することは、画面表示の変更やトレーニング等により AI システムの適切かつ効果的な運用に有益である。
  - また、AI モデルの運用フェーズでは、性能評価指標やデータドリフト指標等技術的なモニタリング結果を個別に確認するだけでなく、ビジネスユーザ、IT、データサイエンティスト、QA や CSV 担当者等、関係部門が定期的に情報を共有する会議体を設けることが望ましい。2026 年 1 月に FDA と EMA が連名で発出した「Guiding Principles of Good AI Practice in Drug Development（以下、AI Principles）」が示す「Multidisciplinary expertise」の考え方に沿い、モデル出力の解釈、業務プロセスへの影響、運用上の課題、および今後の改善方針について、ステークホルダー間で継続的に議論・合意形成することで、AI モデルの Credibility と GxP 適合性をライフサイクルを通じて維持しやすくなる。

- 変更管理
  - モデルの再学習、性能に影響を与えうるプロセスや IT インフラの変更等、モデルの妥当性に影響を及ぼす全ての事象は、確立された変更管理手順の下で評価、承認、文書化する。

- モデル、システム、またはそれが使用されるプロセスへのいかなる変更（モデルが入力として使用する物理的オブジェクトへの変更を含む）も、モデルの再テストの要否を評価する。テストされたモデルは、リリース前に構成管理を行う。不正な変更を検出するための管理手法を導入する。なお、Annex 22 では、現状、AI の自律的な変更を許容していない。

● モデルカード（エンドユーザへの情報開示）

- HITL では、人間の役割が重要であるが、エンドユーザが AI の特徴を理解して利用することが大切である。FDA は、2025 年 1 月に発出した「Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations」（FDA AI 医療機器ガイダンス）で、AI の Output を利用するエンドユーザ向けに「どのような AI モデルなのか」を分かりやすく「モデルカード」として開示することを推奨している。この考え方は、医薬品の GxP 領域でも活用できると考える。「モデルカード」は、AI モデルを妄信することなく、安全かつ効果的に利用するために、AI モデルの概要、性能、バイアス、そして「してはいけないこと」や「AI モデルの限界」を理解してもらうことを目的としている。「モデルカード」には、一例として以下のような情報が提示されている。これらを参考に、マニュアルやトレーニングに AI モデルの情報を提供し、エンドユーザへの AI モデルの理解を促すことを推奨する。
  - **利用目的：**この AI モデル（またはシステム）の想定利用者は誰か？想定している利用目的は何か？
  - **性能：**テストの結果、どれくらいの精度（感度、特異度等）があるか？
  - **限界とバイアス：**どのような状況や対象では性能が低下するか？（例：「特定の非常に稀な症例では見逃す可能性がある」「アジア人集団での検証は十分ではない」）
  - **使い方：**どのような入力データを使えばよいか？結果はどのように解釈し、意思決定や行動すればよいか？
  - **リスク：**AI モデル、データ、結果に関連するリスクは何か？（例：個人情報の取得、サイバーセキュリティ）
  - **変更履歴：**バージョンとその履歴、変更内容は何か？（例：対象年齢を 18 歳以上から 22 歳以上に変更した。利用できる 12 誘導心電図が A、B から A、B、C、D に増えた）

さらに FDA AI ガイダンスは、ICH Q12 ガイドラインの「確立された条件（Established Conditions）」や「承認後の変更管理計画（Post Approval Change Management Plan：PACMP）」の概念を AI モデル管理に適用する可能性を示唆している。これは、モデルの許容可能な運用範囲と変更管理手順をあらかじめ定義し、規制当局との間で合意形成を図る先

進的なアプローチである。この戦略的活用により、製薬企業は規制上の予見可能性を高め、承認後の変更に伴う負担を軽減しつつ、技術革新を迅速に導入することが可能となる。

総じて、規制当局は、AI モデルを単なる IT ツールとしてではなく、患者さんの安全、製品の品質、結果の信頼性に直接的な影響を与える重要なプロセス構成要素として位置づけている。コンプライアンス準拠とプロジェクト成功の鍵は、AI モデルのライフサイクル管理を、既存の医薬品のライフサイクルにおけるプロセスにシームレスに統合し、継続的なモニタリングと厳格な変更管理を通じて、その信頼性を恒久的に実証し続けることにある。

なお、廃棄フェーズの詳細は本書の範囲外であるが、GxP 規制の観点から、AI モデルのアーカイブポリシー・訓練データの保持期間・廃棄手順を QMS 上で定めることは不可欠である。各社の既存 SOP（コンピュータ化システム廃棄手順）を AI モデルおよび AI サブシステムに適用・拡張することを推奨する。

6. 主要規制要件の比較とギャップ分析

ここまで FDA AI ガイダンスを中心に AI モデルのライフサイクルの構想から運用フェーズを概説し、Annex 22 も含めて考察してきた。

FDA AI ガイダンスと Annex 22 が提示する要求は、FDA AI ガイダンスがプロセス、Annex 22 がデータと、それぞれ主眼を置く点が異なる。Annex 22 の要求がデータガバナンスに深く踏み込み、一見してより厳格に映るのは、GDPR（General Data Protection Regulation）や AI Act といった欧州の強力な法的枠組みを遵守することが前提となっているためである。

実務において、FDA AI ガイダンスが提唱する 7 つの Step およびリスクベースフレームワークは、従来のバリデーションの考え方や CSA とも親和性が高く、開発ライフサイクル全体のプロセスの「骨格」を理解する上で有用である。一方で、Annex 22 のガイダンスは、テストデータの独立性の確保といった、より具体的かつ実践的な活動レベルでの要求事項を理解する上で極めて参考になる。したがって、両者は対立する概念ではなく、相互に補完し合う関係と捉え、双方を深く理解し自社の活動へ統合していくことが肝要である。

6.1 FDA AI ガイダンスと Annex 22 の比較

ここで、FDA AI ガイダンスと Annex 22 の規制のギャップを比較する。

表 11：FDA AI ガイダンスと Annex 22 のギャップと考察

項目	FDA AI ガイダンス	Annex 22	ギャップと考察
全体フレームワーク	リスクベースの Credibility 評価フレームワーク(7-Step)を提唱。概念的。	Annex 11 を補完する具体的な技術的要件を提示。より具体的。	FDA AI ガイダンスは「どう考えるか」、Annex 22 は「何をすべきか」に重点。両者を

項目	FDA AI ガイダンス	Annex 22	ギャップと考察
			組み合わせることで網羅的なアプローチが可能。
適用範囲	医薬品ライフサイクル全般の広範な AI 利用を想定。	静的（Static）かつ、同一入力に対して常に同一出力を返す決定論的（Deterministic）な AI モデルのみをクリティカルな GMP 用途で許容し、運用中に継続的かつ自動で学習・更新を行う動的／適応型モデルはクリティカルな GMP 用途では明示的に制限している。	現時点では Annex 22 のほうが適用範囲は狭い。
データ	Fit for Use（目的に適合したデータ）の重要性を強調。	訓練・チューニングデータから独立したテストデータ（Test Data）の準備と、その技術的・手順的な分離を要求している。テストデータの開発・チューニングへの流用（Data Leakage）を防ぐため、データ独立性（Data Independence）の確保を重視している。	Annex 22 の要求はより具体的で厳しい。テストデータの管理は最重要課題となる。
モデル評価	信頼区間（Confidence intervals）*を含むモデルの性能評価と限界の記述を要求。	説明可能性(Explainability)や信頼性スコア(Confidence Score)**といった、個々の AI の Output の評価軸を具体的に要求。	FDA AI ガイダンスは仕組みとしての信頼性を要求し、Annex 22 は個々の結果の信頼性も要求。共にブラックボックスのままでは承認が難しい可能性を示唆。
運用	ライフサイクルを通じたメンテナンスの重要性を指摘。	性能モニタリングや入力データモニタリング等、具体的なモニタリング項目を要求。	運用フェーズでの継続的な検証活動が共通で必須となる。

項目	FDA AI ガイダンス	Annex 22	ギャップと考察
データインテグリティ	既存の GxP ガイダンスに基づき、監査証跡、アクセス制御、記録の完全性・真正性の確保等、コンピュータ化システム全般に対するデータインテグリティ要件を適用。	Annex 22 および関連文書において、訓練データ・テストデータ、AI モデルの出力の完全性と追跡性を明示的に要求している。データセットのバージョン管理や変更履歴管理に加え、ハッシュ値等によるデータ固定（改ざん防止）のような技術的手段も含め、AI ライフサイクル全体でのデータインテグリティ確保を強調している。	FDA AI ガイダンスは既存の GxP 要件の適用という「枠組み（General）」重視であるのに対し、Annex 22 は AI 特有のデータ処理に対する「技術的詳細（Specific）」まで踏み込んで要求している点に明確な差分がある。

*信頼区間（Confidence intervals）：AI モデルとしての性能評価指標

** 信頼性スコア（Confidence Score）：AI モデルが出力する個々の Output の確信度

Annex 22 は、AI のうち静的かつ決定論的なモデルのみを適用範囲として想定しており、運用中に継続的かつ自動的に学習・更新を行う動的／適応型モデルや、確率論的な出力を持つモデルは文書の適用範囲外とされている。これらのモデルについて Annex 22 は、クリティカルな GMP アプリケーションでは使用すべきではない（should not be used）と明示しており、その結果、クリティカルな GMP 用途における動的／適応型／確率論的モデルの利用は明示的に制限される。したがって、Annex 22 のもとでは、クリティカルな GMP 用途で利用可能な AI モデルは、実質的に静的かつ決定論的なモデルに限定される。

一方、FDA AI ガイダンスは、AI モデルのライフサイクルを通じて、「Credibility（信用度）」を維持するためのリスクベースフレームワークを提示しており、Annex 22 のように 運用中に継続的かつ自動的に学習・更新する動的・適応型 AI モデルを明示的に禁止してはいない。むしろ、変更管理やドリフト監視を含めたライフサイクルの維持を前提としつつ、モデルの更新・再学習も想定し得るスタンスと解釈されるが、そのような AI モデルを規制判断に用いる場合には「早期の FDA との協議」が強く推奨されている。

本書では、こうした FDA AI ガイダンスと Annex 22 のスタンスの差異を踏まえつつ、グローバル展開における基本方針として、以下の考え方を推奨する。

- クリティカルな GMP 用途では、Annex 22 の要件を満たす「静的かつ決定論的なモデル」（運用中に自動学習を行わず、同一入力に対して常に同一出力を返すモデル）を共通ベースラインとして採用することを原則とする。すなわち、米国においても、多

くのケースで規制判断に直結する用途では、運用中に自動学習を行わない静的モデルを用い、AI モデルの変更はオフラインでの再学習、Credibility 評価・変更管理を経て新バージョンとして反映することを基本とする。

- AI モデルの分類（静的、動的・適応型）を問わず、運用中の性能劣化への対応としての変更管理の重要性は FDA AI ガイダンス、Annex 22 双方で強く強調されている。さらに、米国であっても、動的・適応型 AI モデルを規制上の意思決定に用いることは、自動的に「許容される」と解釈すべきではない。このため、COU とモデルリスクを明確化した上で、医療機器分野の AI における Predetermined Change Control Plan（以下、PCCP）を参考に「事前定義された変更管理計画」を準備することが望ましい。すなわち、どの範囲のモデル変更（再学習・アルゴリズム改良等）を想定しているのか、その場合にどのような検証・影響評価・リリースプロセスを経るのかを、ライフサイクル維持計画としてあらかじめ文書化することを、本書ではベストプラクティスとして推奨する。
- その上で、当該計画および COU・モデルリスクの内容について、個別案件ごとに早期の規制当局との相談（事前エンゲージメント）を行い、当局側の期待水準と整合した形で動的・適応型 AI モデルの適用可否や運用条件を確認することを推奨する。これにより、「静的・決定論的モデルを下限としつつ、米国では事前合意された枠内で限定的な適応性を認める枠組みを検討することが可能となる」という実務的な落としどころを検討しやすくなる。
- 非クリティカル用途（業務効率化や補助的分析等）で動的・適応型 AI モデルを用いる場合も、「クリティカルな GMP プロセスからは論理的・技術的に分離された用途」であることを QMS 上明確に位置付け、Annex 22 および各地域ガイダンスとの整合性を確認することが望ましい。

さらに、FDA と EMA の共通認識として、専門家との協力体制が挙げられる。本書では、「データサイエンティスト等と協働して」等、各分野の専門家（Subject Matter Expert、以下、SME）や担当者と協議しつつ、AI モデルの開発を進めることを繰り返し述べた。AI Principles では、「5. Multidisciplinary expertise」「10. Clear, essential information」が含まれている。AI Principles は、「各領域の専門家の活用」と異なる背景、分野の SME や担当者が協働し、合意を得ていくことが重要であることから、「各担当者にとって明確に意味を理解できる言葉を使うこと」を提唱している。AI モデルの開発では、既存の体制では当たり前に使っていた言葉が通用しない、担当者間で背景理解が不足していることもあるため、コミュニケーションにより留意することが大変重要である。

6.2 日本の規制動向

現時点では、厚生労働省および医薬品医療機器総合機構（Pharmaceuticals and Medical Device Agency、以下、PMDA）から、医薬品開発における AI モデルの Credibility 評価に該当するガイダンスは発出されていない。しかし、本書でも参照している経済産業省「AI 事業者ガイドライン」は、国内の汎用的な AI ガバナンス指針である。また、PMDA との事前相談により、各 AI システムの管理方法や利用範囲を合意していくことにより、FDA AI ガイダンスや Annex 22 を参考とした実践的な対応が可能になる。日本の製薬企業は、FDA AI ガイダンスおよび Annex 22 の要求事項を満たすことで、将来的な国内規制にも対応可能な体制を構築することが推奨される。

参考までに、国内の関連通知を以下に示す。

- 「人工知能（AI）を用いた診断、治療等の支援を行うプログラムの利用と医師法第 17 条の規定との関係について」（厚生労働省、2018 年 12 月）：
  - AI による診療支援においても、最終的な判断の主体と責任は医師にあり、当該行為は医師法第 17 条の医業として位置づけられることを明確化した通知。
- 「医療デジタルデータの AI 研究開発等への利活用に係るガイドライン」（厚生労働省、2024 年 3 月）：
  - 医療機関に蓄積されたデータを、仮名加工情報として民間企業の AI 製品開発等へ円滑に利活用するための法的根拠と、具体的な加工・運用手順を体系化したガイドライン。

7. 本書の適用

本書は、AI モデルの Credibility 評価フレームワークの活動に関する規制と実務の基本原則を理解する上で有用であるが、その適用には重要な限界があることを認識する必要がある。主に参考にした FDA AI ガイダンス、Annex 22 は 2025 年に発出されたドラフトであり、日進月歩で進化する技術や規制の動向を完全に反映したものではない。例えば、本書は現行のドラフトガイダンスに基づき廃棄フェーズの詳細な解説は範囲外としているが（5.9 項参照）、GxP のライフサイクル管理の観点からは、将来的に AI システムや関連データの廃棄・アーカイブ要件も重要な論点となる可能性がある。

特にプロジェクトフェーズにおいて使用される LLM や生成 AI などの事前学習済みモデルに関しては、FDA AI ガイダンスで示されている「事前学習済みモデルの評価方法」は現時点では概念的な段階であり、実務的なガイドラインとしての適用範囲は限定的である。このため、今後はより具体的かつ実務的な評価手法の確立が課題となる。

LLM 等の事前学習済みモデルを提供するテクノロジー企業は、モデルの性能、限界、安全性評価について Model Card や Technical Report として公開しているものの、GxP 環境での利用に必要な以下の情報については、十分な開示が得られない可能性がある。

- 訓練データの詳細な構成（ソース、サンプリング方法）
- データ前処理の具体的手順
- アノテーション品質管理プロセス
- 開発時のハイパーパラメータ詳細

このような情報の不透明性に加え、従来の QMS 監査を中心としたサプライヤ評価手法は、訓練データ管理の観点が十分に組み込まれていないことが多く、業界全体として新たなアプローチの確立が求められる。この状況は、FDA AI ガイダンス、Annex 22 が AI モデルを自社開発することを前提としていることから、製薬企業に「自社の COU に即した AI モデルそのものの Credibility 評価フレームワークの活動」という重い責任を課す。

最も懸念すべきは、検証の難しさやその必要性を認識せず、「補助的なツールだから」と安易に判断し、GxP の重要な管理策である「クリティカルシンキング（批判的思考）」を放棄することである。クリティカルシンキングとは、AI モデルの Output を無条件に受け入れず、「なぜ AI システムの Output を信じられるのか？」「AI システムの Output が誤っていた場合のリスクは？」と能動的に問い、自身の専門知識と照らし合わせる知的なプロセスである。このクリティカルシンキングの省略は論理的な判断の形骸化に直結する。

したがって、本書を単なるチェックリストとせず、クリティカルシンキングを実践するための思考のフレームワークとして活用すべきである。最も堅牢な Credibility 評価フレームワークの活動とは、文書やシステムだけでなく、担当者一人ひとりが実践するクリティカルシンキングそのものであることを忘れてはならない。

8. まとめ

本書は、規制当局の動向を踏まえ、AI 技術の非専門家を含む幅広い関係者が AI システムの品質保証を理解するための一助として作成した。本書で示した品質保証の考え方が羅針盤となり、AI という強力なツールがもたらす大きな可能性を、患者さんの安全を最優先にしつつ、着実に現実のものにしていくことを期待する。今後は、GCP 領域における施設選定の最適化や、PV 領域での有害事象判定・シグナル検出の高度化等、日本製薬工業協会内の各プロジェクトとも連携し、本書の議論をより具体的な実践へと昇華させていくことが望まれる。

本活動の過程では、より実践的なシナリオを設定し、具体的にどのような特徴量の選定や性能評価手法、各 Step における成果物イメージ等のシミュレーションも試みた。しかし、AI 技術特有のメリットを最大化する COU の設定には多角的な検討が必要であり、今回はこれらの詳細化を今後の課題として残すこととした。また、事前学習済みモデルの利用規約等の制限から、従来の QMS 監査を中心としたサプライヤ評価の実施は困難であると考えられる。このため、公開文書、第三者認証等の活用が評価の主軸になっていくことが予想される

が、実践的な観点や方法は確立されていない。したがって、業界全体での継続的な議論と知見の集積が不可欠である。

今後、各領域の SME と共に、本書の実践編として、ケーススタディの検討や新たなサプライヤ評価手法の確立等、AI 技術を適正に活用できる環境整備を継続したい。そして、規制当局と協調して製薬企業に変革をもたらすイノベーションを促進することで、信頼性のある AI システムを GxP プロセスに統合し、真に患者さんの QOL 向上に寄与していきたい。

9. 参考資料

- Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry and Other Interested Parties（FDA、2025 年 1 月、ドラフト）
- EU GMP Annex 22: Artificial Intelligence（EC、2025 年 7 月、ドラフト）
- AI 事業者ガイドライン（経済産業省、2025 年 3 月、第 1.1 版）
- GAMP® 5 -A Risk-Based Approach to Compliant GxP Computerized Systems（ISPE、2022 年 7 月、2nd Edition）
- GAMP® 5　コンピュータ化システムの GxP 適合へのリスクベースアプローチ（ISPE、2025 年 4 月、第 2 版）
- Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations（FDA、2025 年 1 月、ドラフト）
- Computer Software Assurance for Production and Quality System Software Guidance for Industry and Food and Drug Administration Staff（FDA、2026 年 2 月）
- Guiding Principles of Good AI Practice in Drug Development（FDA、EMA、2026 年 1 月）
- ICH Q12　Technical and regulatory considerations for pharmaceutical product lifecycle management - Scientific guideline（ICH、2019 年 11 月）
- 人工知能（AI）を用いた診断、治療等の支援を行うプログラムの利用と医師法第 17 条の規定との関係について（厚生労働省、2018 年 12 月）
- 医療デジタルデータの AI 研究開発等への利活用に係るガイドライン（厚生労働省、2024 年 3 月）
- 「FDA CSA ドラフトガイダンスの概説と GxP 領域への適用の検討と考察」（日本製薬工業協会 医薬品評価委員会、電子化情報部会 2023 年度タスクフォース 4、2024 年 7 月）

執筆者

電子化情報部会

部会長	佐久間　直樹	帝人ファーマ株式会社
副部会長	井上　学	MSD 株式会社
	染谷　美紀	ファイザーR&D 合同会社
	渡辺　博司	第一三共株式会社

電子化情報部会タスクフォース４サブタスクフォースメンバー（順不同）

TF リーダー	渡辺　博司	第一三共株式会社
SFT リーダー	三宅　哲郎	バイエル薬品株式会社
	中尾　進	エーザイ株式会社
	新開　浩平	キッセイ薬品工業株式会社
	中村　優希	大正製薬株式会社
	中島　洋子	中外製薬株式会社
	渡辺　博司	第一三共株式会社