生成 AI を活用した医療機器の審査ポイント

（Draft）

独立行政法人 医薬品医療機器総合機構

審査センター

2026 年 9 月 14 日

内容

1. はじめに 5
2. 本審査ポイントの対象 9
3. 定義 9
4. 生成 AI の特性 10
4.1.生成 AI に顕著な技術的特性 10
4.2.生成 AI 医療機器に想定される特徴 12
4.3.生成 AI 医療機器に想定される開発・運用形態 13
5. 生成 AI の薬事評価・運用上の課題 13
5.1.評価及び運用上の課題 14
5.1.1. 出力の不確定性及び変動 14
5.1.2. 出力妥当性の評価 15
5.2.使用上の課題 15
5.3.その他の事項 16
6. 評価に関する考え方 16
6.1.評価全体の考え方 16
6.2.市販前の審査 17
6.2.1. 有用性 18
6.2.2. 製品性能 18
6.2.3. その他の評価 23
6.3.市販後の対応 23
6.3.1. 市販後の報告 23
6.3.2. ヒューマン・ファクターを踏まえた設計と運用 23
7. おわりに 25
別紙 1 開発形態の例 26
別紙 2 論点となり得る主な技術的特性と開発への示唆 27
別紙 3 ベンチマークテストの位置づけと限界 31

Ⅰ． はじめに

近年、生成 AI に対する社会的関心が急速に高まっている。生成 AI とは、「文章、画像、プログラム等を生成できる AI の総称」である。特に近年の生成 AI は極めて大規模なデータセットを用いて構築されるために、使用者からの自然文や音声による様々な問いかけに応じて、特定のタスクに依らない多様な回答や創造的なコンテンツを生成できるようになった。これらの生成 AI は、大規模言語モデル（LLM: Large Language Model）や視覚言語モデル（VLM: Vision Language Model）として構築され、医療分野においても臨床支援、文書作成、画像解析など幅広い応用が期待されている。しかしながら、薬事規制の観点からは、生成 AI 技術の評価やリスク管理に関し、特別な枠組み等は十分に体系化されていないのが現状である。そもそも、AI に関する議論[i,ii,iii,iv]は盛んに行われているものの、その多くは生成 AI ではなく、特定のタスクに特化した識別 AI に関するものであった。対象とする機能（タスク）を明確に定義することができ、推論結果もラベル等として出力されることによって出力の正誤をルールベースで定量的かつ客観的に評価できる従来型の識別 AI とは異なり、生成 AI は汎用性・多目的性が高いために使用範囲の限定が困難であり、自然文等による出力の正誤判定には解釈の余地を伴う。そのため「生成 AI のどの側面が薬事規制上の課題となるのか」、「生成 AI の特性のうち、何が薬事開発上の特別なリスクとして認識されるべきなのか」について、体系的かつ実務的に整理されたものは少ない。

本審査ポイントでは、これらの点について現時点での知見を基に包括的に整理し、評価方法や市販後対応を含む薬事規制上の考え方について論じることを目的とする。ただし、本審査ポイント発行時点では生成AIを用いた医療機器に関する審査実績はないことから、現在の承認審査の考え方に照らし合わせて検討される審査上のポイントをドラフトとしてまとめたものである点に留意が必要である。これを現時点でのたたき台としつつ、様々な角度から広く意見を収集することでより適切なものとしてまとめていくことが今後の目標となる。

生成AIを用いた医療機器の取扱いについては、どこまでを医療機器として取扱うか（医療機器該当性）も含めて慎重な議論がなされている。本審査ポイントは、標ぼう等含め医薬品、医療機器等の品質、有効性及び安全性の確保等に関する法律（以下「医薬品医療機器等法」という。）上の医療機器に該当した場合を前提としたものであり、今後医療機器該当性が整理された場合には、それを踏まえ本審査ポイントの内容を調整する必要が生じる可能性がある。また、LLMをはじめとするAI技術の発展は著しいことから、今後の経験や知見を踏まえ、より合理的なものに適宜改訂されるべきものである。

なお、一般に医療機器の承認審査は製品ごとの多様性を前提として行われるものである。極端に言えば、臨床的位置づけも含めて完全に同一の製品であっても合理的な説明ができるのであれば、異なる評価方法であっても受入れ可能である。本審査ポイントで提示する内容は、あくまで一つの整理例であり、個別案件ごとに柔軟に判断されるべきものである。

2. 本審査ポイントの対象

「生成 AI」の概念は広く、一般には「文章、画像、プログラム等を生成できる AI の総称」と捉えられる。このうち、本審査ポイントにおいては、使用者からの自然文や音声による様々な問いかけに応じて、特定のタスクに依らない多様な回答や創造的なコンテンツを生成することができるよう構築された生成 AI（以下、特別な説明がない限り単に「生成 AI」（※））を主な対象とする。特に、これらの生成 AI が、大規模言語モデル（LLM: Large Language Model）あるいは視覚言語モデル（VLM: Vision Language Model）として技術的に実現されている場合を想定している。

また、上述の生成 AI を用いて機能実現された医療機器（以下「生成 AI 医療機器」）を本審査ポイントの対象とする。

3. 定義

- 開発コンセプト：申請品を開発するに至った背景や経緯、対象疾患の自然歴（既存療法の有無及び既存療法による治療経過）、又は医療実態（臨床成績）、疾病の重症度、国内外における最新の開発状況など。
- 申請品：承認を目指す製品。承認申請では申請品、対面助言等では相談品を意図する。
- 製品性能：申請品の入力データに対する処理性能、処理能力。
- 設計コンセプト：開発コンセプトを踏まえ、どのような機能、どの程度の性能をもつ製品を開発しようとしたか。
- 当局：規制当局を指す。審査・相談時には、独立行政法人医薬品医療機器総合機構（以下「PMDA」という。）を指す。
- 臨床的位置づけ：その申請品が臨床現場において、誰にどのような目的で使用されるか。
- 臨床的有用性：申請品を用いることにより診療や患者アウトカムにもたらす価値。
- 広義の生成 AI：文章、画像、プログラム等を生成できる AI の総称。
- （本審査ポイントで対象とする）生成 AI：広義の生成 AI のうち、使用者からの自然文や音声による様々な問いかけに応じて、特定のタスクに依らない多様な回答や創造的なコンテンツを生成することができるもの。
- 生成 AI モデル：学習データを用いて訓練された学習済みの深層ニューラルネットワークで、自然文や音声による入力に応じて、文章、画像、プログラム等のコンテンツを出力するもの。
- 生成 AI システム：回答を生成する「生成 AI モデル」を中核とし、使用者が直接やり取りする「インターフェイス」、入出力を安全に処理する「ガードレール機構」、情報検索などを行う「ツール群」といった複数のコンポーネントが連携して動作するシステム。
- 生成 AI 医療機器：生成 AI システムのうち、医薬品医療機器等法における医療機器に該当する製品。
- 生成 AI 機能：生成 AI 医療機器がもつ機能のうち、生成 AI モデルを中核として実現された機能。
- プロンプト：生成 AI システムに対して何をして欲しいかと伝える指示文のこと。
- 生成 AI コンテンツ：生成 AI モデル又は生成 AI システムが出力する文章、画像、プログラム等のコンテンツであって、その意味内容を使用者が解釈することによって情報としての価値が生じるもの。

- 偽情報：事実に基づかない架空の情報
- 誤情報：事実に反する情報
- 有害情報：不適切な自己治療の推奨や危険な行為の教唆等、患者の生命や医療安全に直接的な危害を及ぼす情報
- 不適切な誘導：受診の妨げや科学的根拠のない代替療法への勧誘等、患者を本来必要な医療サービスから遠ざける出力

4. 生成 AI の特性

4.1. 生成 AI に顕著な技術的特性

生成 AI に顕著な技術的特性と、それに関連して留意するべき観点について、以下の 5 つの側面から整理する[v]。

(1) 学習過程やモデル構造の大規模性

生成 AI モデルは、モデルのサイズや学習に用いるデータセット等を増やせば増やすほど性能が向上するスケーリング則が知られている。そのため、例えば数千億パラメータに及ぶ巨大なニューラルネットワークが構築され、インターネット等に存在する膨大かつ多種多様なデータを用いて事前学習が行われる。その後、教師ありファインチューニングや強化学習といった手法を用いることにより、使用者からの問い合わせに対して適切で望ましい回答が出力できるよう、事後学習が施される。

こうした大規模なモデルの開発と運用には一般的に莫大な計算資源を要し、多額の投資が必要となる。そのため、開発者たる医療機器ベンダの多くは、生成 AI モデルをフル・スクラッチで自社開発し、運用するのではなく、外部の基盤モデルを活用し、システム統合やインターフェイス最適化といった後工程から主体的に関与していくことが想定される。

【留意するべき観点】

学習過程やモデル構造の大規模性により、医療機器のコア・コンポーネントである生成 AI モデルの開発や、場合によっては運用までも、実質的にサードパーティに依存せざるを得ない状況が生じる。開発過程に用いたデータセットの具体的な中身が秘匿されることにより、開発者にとっては問題発生時における根本原因の特定が困難になるほか、データセットバイアスの潜在化や、評価におけるデータリーケージ（評価用データが学習用データセットに混入してしまう問題）を踏まえた評価用データの妥当性説明が困難になること等の課題がもたらされる。

(2) 文脈に応じたコンテンツ生成能力

従来型の識別 AI がラベル等の固定された出力を示すのに対し、生成 AI モデルは、使用者が入力するプロンプトに応じて、文章、画像、プログラム等、人間が解釈することによってその意味内容を享受するコンテンツを出力する。こうしたコンテンツは、同一のプロンプトに対しても、揺らぎをもった出力となるよう、ランダム性が付加されて生成されることもある。

さらに、こうしたコンテンツは使用者が入力するプロンプトのみならず、対話履歴などのコンテクスト（文脈）に応じてより適応的に変化する。そのため、使用者との双方向かつ継続的なやり取りが可能とな

り、生成 AI モデルからの出力も、使用者固有のコンテクストをより反映したものとなる。

【留意するべき観点】

生成されたコンテンツが事実に基づかない尤もらしい偽情報や誤情報を含むハルシネーションのリスクが存在する。これにより、医療機器としての出力の正確性をいかに保証するか、また、出力結果の真偽を一般的な使用者が適切に解釈できるかという課題が生じる。さらに、入力のわずかな表現の違いによって出力結果が変わるため、使用者の入力スキルへの依存や再現性の低下が懸念される。加えて、会話履歴の蓄積によってモデルの挙動が徐々に変化・偏向することによるリスクを伴う。

(3)　汎用性・多目的性

従来型の識別 AI が所定の操作により特定の目的を果たすための出力を提示する（例えば、医用画像を入力し規定された異常所見を検出する。）といった単一タスクに特化していたのに対し、生成 AI は一つのモデルに対して対話的に指示を与えることで、質問応答、要約、翻訳等の基礎的なタスクから、複数疾患の診断支援といった高度なタスクまでをシームレスに実行できる汎用性と多目的性を備えている。

こうした機能は、使用者が入力するプロンプトに応じて動的に引き出される。さらに、標準状態では高い精度での遂行が難しい機能であっても、プロンプトの中に当該タスクの例示を含めることによって、パラメータの更新を伴わずに新たなタスクに適応させることも可能である（文脈内学習）。あるいは、ファインチューニング等によりパラメータを微調整し、特定の機能を事後的に獲得させることもある。

【留意するべき観点】

モデル機能の汎用性・多目的性により、複数疾患の診断支援といった医療機器の多機能化が可能になる一方で、製品の性能評価は複雑化する。また、医療機器として承認された標榜機能以外のタスクも実行し得るため、機能の境界を明確に画定する必要がある。したがって、使用者による適応外使用（承認された使用目的や対象外での使用）・不適切使用（注意事項や推奨される手順に反した使用）・誤用（操作ミスや出力の誤解釈等の意図しない使用）を防ぐためのガードレール機構を構築し、申請者が意図しないタスクが実際に実行されないことを確保することが求められる。

(4)　外部システムとの相互運用性

LLM などの生成 AI モデルは、単独で動作するだけでなく、外部のシステムやデータソースと連携する高い相互運用性を備える方向に進歩している。例えば、外部のデータベースから最新の医学論文を検索して回答に取り込む検索拡張生成（RAG: Retrieval-Augmented Generation）や、外部の計算機能・データベース・他システムをツールとして呼び出し、その出力を自らの生成に取り込む機能などが挙げられる。これにより、モデル単体の内部知識に閉じることなく、外部の最新かつ専門的な情報資源を動的に参照した出力が可能となる。

【留意するべき観点】

外部システムとの相互運用性により、生成 AI モデルの利便性や正確性が向上する一方、外部環境への強い依存関係という課題が生じる。すなわち、参照元の外部データが変更・削除されたり、連携先システ

ムに障害が生じたりした場合に、医療機器としての性能が直接的に変動するリスクがある。したがって、連携先の範囲・バージョン・可用性を明確に管理し、外部依存に起因する性能変動を監視・検知する仕組みを講じることが求められる。

(5)　動作の自律性

（4）で述べた外部システムとの連携能力を基盤として、生成 AI モデルは、与えられた目標に対して自らその達成手順を組み立て、段階的に実行する自律的な動作が実現できうる。すなわち、入力された文脈からタスクを特定し、サブタスクへの分解・計画を立案したうえで、適切な外部システムをツールとして呼び出し、その結果を観察・評価して次の行動を決定するというループを、人間の逐次的な指示を介さずに繰り返すエージェント機能などが挙げられる。

こうした自律性は二値的なものではなく、自動運転車における自動化の水準（レベル 0～5）になぞらえて、人間の関与の度合いに応じた連続的な段階として捉えられる。すなわち、人間が各行動を逐一実行・承認する低位の水準（人間が実行主体）から、モデルの提案を人間が確認・承認する中位の水準を経て、一定の目標の下でモデルが一連の行動を継続的に実行し人間は結果を監視するにとどまる高位の水準（人間が観察者）までが想定される。同一の性能を有するモデルであっても、運用時にどの自律水準を与えるかは能力の高低とは独立した設計上の選択であり、高性能なモデルを限定された低位の水準で運用することも、その逆もあり得る。

【留意すべき観点】

動作の自律性により、複雑な臨床タスクを一連の流れとして遂行できる利便性が生じる一方、システム全体の制御可能性という課題が顕在化する。すなわち、自律的に実行しうる機能の範囲と人間による監督・介入が必要となる境界を明確に画定する必要があるが、自律度が高まるほど、与えられた目的の達成を最適化する過程で、生成 AI モデルが人間の設定した制約やガードレールの抜け穴を突き、本来意図されない迂回的・逸脱的な手段を選択し得ること（報酬ハッキング）にも留意を要する。また、(2)で述べたハルシネーションや(3)で述べた適応外タスクが、こうした自律ループの中で人間の確認を経ずに後続の行動へ連鎖・伝播し、誤りが増幅されるリスクを伴う。さらに、行動系列がプロンプトや外部環境に応じて動的に決定されるため、事前に挙動を網羅的に予見することが難しく、性能評価や再現性の確保が一層複雑になる。したがって、想定する自律水準に応じて、人間による確認・停止・介入の機構（Human-in-the-Loop 等）を設計に組み込むとともに、実行された行動系列を事後に追跡・監査できる仕組みと、生じた事象についての責任の所在をあらかじめ整理しておくことが、医療機器のリスクマネジメントに求められる。

4.2. 生成 AI 医療機器に想定される特徴

こうした性質を組み合わせることにより、以下のような特徴をもつ医療機器が実現できるようになる。

- 使用者からの様々な問いかけに応じて柔軟に回答を生成することによって、使用者の疑問に答えたり、必要な解析結果を提示したりすることができる。また、使用者との会話履歴が蓄積されることによって、使用者の技量や知識のレベルに応じて最適化された出力が得られるように調整さ

れる。こうした自動調整によって、専門的な医療知識がない使用者に対しても、治療支援などの高度な機能を実現することができる。

- テキストのみならず、画像やテーブルデータ等の多様なモダリティのデータに含まれる情報を、使用者からの自然文による問い合わせに応じて横断的・統合的に分析する機能を有する。これにより、特定のモダリティの取り扱いに限定されていた既存の医療機器よりも高い診断性能を発揮したり、自然文による対話形式で放射線画像の読影を支援したりするといった、従来では成しえなかった機能や性能を実現することができる。
- 医学文献データベース等の外部の情報処理システムとリアルタイムで連携することによって、使用者からの問い合わせに対して最新の知見を踏まえて回答し、解析する機能を有する。さらに、エージェント機能を備えることによって、様々な外部の情報処理システムの中からその時々に応じて最適なものを自律的に特定し、得られた結果を観察・評価したうえで、最終的な統合解析の結果を使用者に提示するといった機能を実現することができる。

4.3. 生成 AI 医療機器に想定される開発・運用形態

生成 AI モデルを一から研究開発するには多額の投資が必要となる一方で、高い汎用性を備えたモデルをクラウド API やオープンウェイトモデルとして提供する基盤モデル事業者が多数存在している。そこで、外部の事業者から基盤となる生成 AI モデルの提供を受け、申請者は当該モデルを含む生成 AI システムの全体を最適化することによって、医療機器としての要求事項（使用目的、臨床的位置づけ、品質マネジメント、リスクマネジメント、検証・妥当性確認、変更管理等）を満たすための要素（医療用途に即した入出力設計、ガードレール機構、評価データセット・評価手順、運用・監視体制等）を設計・開発する、という分業が想定される。

例えば、プロプライエタリモデルをクラウド API 等で利用する形態、オープンウェイトモデルを取得して自社環境で運用する形態、又はそれらを基に医療用途向けにファインチューニング等の追加学習を行う形態など、現実的な開発形態に応じて、申請者が直接管理できる範囲（モデルの更新・再学習の可否、学習データの把握可能性、推論環境の固定性、セキュリティ要件、ログ取得の可否等）と、外部事業者に依存する範囲が異なる（別紙 I 参照）。対処すべき評価上又は使用上の課題の内容や、申請者に求められる説明責任・管理方針（契約、Service Level Agreement を含む。）も、当該開発形態に応じて具体化されるべきである。同一の製品であっても、開発形態に応じて、リスクマネジメントやソフトウェアライフサイクルの適切な対応を実施する必要がある。

5. 生成 AI の薬事評価・運用上の課題

医薬品医療機器等法に基づく医療機器の承認審査においては、製造販売しようとする申請品に関する有効性及び安全性の評価が必要となる（プログラム医療機器に関する一般的な開発の考え方については、開発ガイダンス が公表されている。）。“有効であること”の議論においては「有用性」と「製品性能」に大別して考えることができる。有用性とは、診断精度向上や治療成績改善といった臨床的アウトカムへの寄与を意味する。一方、製品性能とは、特定条件下での申請品の性能を示す。

なお、例えば生成 AI モデルを用いて実現されている医療機器であっても、従来の医療機器と同様に用途が限定的（例：画像上の〇〇所見の検出、△△病リスクスコアの提示等）であり、かつ出力の提示方法

が固定されている製品であれば、従来の医療機器の評価方針を多く踏襲できる可能性が高い。

本項では、4 項で述べた生成 AI の特性を踏まえ、技術的な側面に基づく医療機器の評価上及び使用上の課題を大きく以下に分類し、整理する。なお、生成 AI に関しては倫理上の課題も重要なトピックではあるが、本審査ポイントは、審査上の取扱いの整理を目指したものであることから、これについては言及しないこととする。その他にも医療情報を取り扱うにあたっての個人情報保護法、学習用データに含まれる著作物に係る著作権法などの関連法規上の課題も認識されるが、同様に対象外とする。

- 評価及び運用上の課題：技術特性に由来する評価及び運用上の課題
- 使用上の課題：使用者が適切に使用するにあたっての課題
- 倫理上の課題：責任の所在等の運用上の課題（本審査ポイントでは取り扱わない）
- 法律上の課題：個人情報保護法や著作権法などの関連法規に関する課題（本審査ポイントでは取り扱わない）

5.1. 評価及び運用上の課題

生成 AI 医療機器の評価において、論点となり得る主な技術的特性については別紙 2 に整理する。これらに起因する評価及び運用上の課題は、大きく以下の 2 点に集約される。本節でそれぞれ説明する。

- 出力の不確定性及び変動
- 出力妥当性の評価

5.1.1. 出力の不確定性及び変動

生成 AI モデルは、ユーザーの入力の僅かな違いや対話履歴等の文脈（コンテキスト）により、出力が変化する特性がある。また、生成 AI モデルのうち、例えば LLM では、最も高い確率のトークンを常に選択するのではなく、一定のランダム性を付加してサンプリングすることによって、表現豊かな出力を生成するよう設計されたものもある。このほか、外部システムと連携しているものや外部の事業者が提供するプロプライエタリモデルを API を介して利用する場合などは、外部環境や基盤モデルの更新といった外的な要因の影響を受ける。加えて、出力レベルの揺らぎに留まらず、導入施設ごとの調整や対話履歴等による個別最適化が過度に適用されれば、申請者が意図していない医療機器として使用される可能性もある。これらにより、生成 AI モデルの出力の再現性の確認が課題となる。

また、従来の審査においては、変動の範囲（例えば入力条件の範囲）を想定した上で条件範囲内の有効性及び安全性の評価を求めることが一般的であるが、生成 AI 医療機器の場合、入力範囲の多様性とともに、上述した出力の不確定性及び変動の影響が強く、一定程度は制御されることを前提としても“条件範囲内”の規定が極めて困難である。

なお、一般に機械学習の可塑性に対応する医薬品医療機器等法上の制度として変更計画確認手続制度（通称：IDATEN）がある。IDATEN は、特定の変更に対して定期的な更新をタイムリーに実施する点では有用であるが、時々刻々と変化するような特徴をもつ製品への対応には限界がある。また、エージェント機能のように人が設計した安全機構を突破して作用しようとする現象も生じており、事前に制御・管理できる範囲は限定的となる可能性がある。このため、従来の主流の考えである「市販前評価のみで有効性・安全性を十分に保証する」という枠組みは適用困難である。時々刻々と変化する製品の有効性・安全性を確保するためにも、市販前後で有効性・安全性を総合的に評価するリバランスの枠組みを積極的に

取り入れることが重要である。

5.1.2. 出力妥当性の評価

生成 AI 医療機器の評価では、その出力の妥当性を、再現性及び客観性をもって判断することが課題となる。ここでいう妥当性には、生成された情報が医学的な事実やエビデンスに合致しているかという「情報としての正しさ」と、患者の臨床的背景や使用者の意図、使用状況に照らして適切な回答であるかという「文脈としての正しさ」の双方が含まれる。前者については、ハルシネーションによる偽情報や誤情報が混入し得るため、その発生傾向を客観的に測定する方法が課題となる。後者については、唯一の正解が存在しないタスクが多く、何をもって適合的とみなすかという評価基準の設定自体が課題となる。

さらに、生成 AI の出力は、その意味内容を使用者が解釈することによって初めて臨床的な価値が生じるため、妥当性は出力単体では定まらず、使用者の解釈に依存する。このため、出力の内容を評価するだけでなく、想定される使用者が出力をどのように理解し、どのような判断や行動につなげるかまでを含めて考察する必要がある。加えて、こうした妥当性の判断は、患者集団や使用条件によって一様ではなく、学習データに内在するバイアスによって特定の集団や状況で性能が変動しうる。したがって、平均的な性能の確認にとどまらず、臨床上重要な集団・条件を特定した上で、妥当性が保たれる範囲とその限界を明らかにすることが、市販前評価における本質的な課題となる。

5.2. 使用上の課題

上述の特徴を踏まえ、生成 AI 医療機器使用者が適切に使用するにあたっての課題を整理する。

(1) 説明可能性

生成 AI をはじめとする深層ニューラルネットワークは、明確なロジックの積み重ねではなく、多層的なパラメータの重み付けやパターンの学習によって出力を生成する。そのため、モデルがどのような推論過程でその結論に至ったのかを、使用者に対して論理的に説明することは原理的に困難である。結果として、例えば、生成 AI による根拠不明瞭な出力に依拠して医療行為を決定した場合、医師等の使用者が患者に対して十分な説明責任を果たせず、適切な同意（インフォームド・コンセント）を得られないおそれがある。したがって、システムとしての説明可能性を向上させるために、生成 AI 医療機器は、出力の根拠となる医学文献等の情報源を併せて提示するなどの方法により、使用者自身がその内容を可能な限り検証・把握できるように設計されていることが望ましい。

(2) 適応外使用・不適切使用・誤使用

使用者からのいかなる質問に対しても答えられるように生成 AI システムが設計・運用されている場合、使用者の意図の有無にかかわらず様々な逸脱行為が生じるリスクがある。具体的には、承認された目的外のタスクを要求する「適応外使用」や、注意事項や推奨される手順に反した「不適切使用」、さらには意図しない操作ミス等による「誤使用」である。

また、運用を通じた会話履歴の蓄積などに影響されてコンテキストが変容することによって、使用者が気づかぬうちに、生成 AI モデルが想定タスクや臨床的位置づけといった機能範囲を逸脱した出力を行なうおそれもある。このような出力の逸脱は、使用者がその内容を妥当な医学的情報として誤解釈してしまう重大な「誤使用」を誘発する危険性をも伴っている。

(3)　使用者のリテラシー

生成 AI コンテンツは、その真偽や有用性の判断を使用者自らの解釈に委ねるという特性（解釈依存性）を持つ。一方で、生成 AI モデルには事実に基づかない偽・誤情報をあたかも真実のように出力する「ハルシネーション」のリスクが存在し、これを技術的措置のみで事前に完全に検出・排除することは極めて困難である。

このように結果の判断が使用者の解釈に依存するという構造において、システムがもっともらしい偽・誤情報を出力した場合、使用者がその誤りに気づかず看過してしまうリスクは増大する。したがって、安全かつ適切な利用にあたっては、AI の出力を鵜呑みにせず、自らの専門的知見と照らし合わせて批判的に検証・判断できる一定のリテラシーが使用者に求められる。

(4)　認知バイアス

前述の「データバイアス」が、学習データ等に起因して AI モデルそのものに内在するシステム側の偏りであるのに対し、「認知バイアス」は、AI の出力を受け取る使用者の心理的・認知的特性、すなわち「ヒューマンファクター」に起因する受け止め方の偏りを指す。

特に、患者自身が使用者となるチャットボット型の治療支援製品などでは、このヒューマンファクターに根ざした認知バイアスが顕著に表れるリスクがある。例えば、「アルゴリズム権威」（人間の判断よりも製品の出力を優先しがちになること。）の影響により生成 AI の助言が過大評価された結果、医師との信頼関係を損なうおそれがある。また、生成 AI などのアルゴリズムによって示された結果が、人間の判断よりも客観的で正確であるように見えるため、その判断を過大に評価し、依存してしまう「自動化バイアス」も考慮する必要がある。

5.3. その他の事項

申請品の品質、有効性及び安全性の評価の他に、医療現場への受容性についても考慮する必要がある。特に、以下のような場合は、関連学会等との協議や必要な使用指針等の作成も視野に開発を進める必要がある。

- 生成 AI が単なる支援に留まらず、自律型エージェントとして医師や医療従事者の介在なしに一連の診療ワークフローを完結、または自動実行する機能を有する場合。
- テキスト、画像、遺伝子情報など、人間が同時には処理しきれない膨大なマルチモーダル情報をもとに生成 AI が総合的な判断を下すことによって、その出力の根拠を使用者が直感的に理解・検証することが著しく困難になる場合。
- AI チャットボットが患者に対して高い共感性や確信度をもって接することで、患者が医師の診察よりも AI の回答を優先してしまい、従来の医師患者関係が毀損されるような場合。

6. 評価に関する考え方

6.1. 評価全体の考え方

従来の審査においては、市販前における品質、有効性及び安全性の評価を重視し、市販後の調査や対応は補足的に実施することが主流であった。一方で生成 AI 医療機器においては先述した通り、以下のような

特性が市販前の十分な評価を困難にする。

- 出力の不確定性
- 状態の変化
- 出力文脈の適切性評価
- 使用者の受取り方の影響
- 使用上の課題

以上を踏まえ、製品のライフサイクル全体を通じて、品質とリスクを一貫して管理する「トータルライフサイクルアプローチ」の考え方がより重要となる。すなわち、市販後の継続的な管理を前提とした上で、市販前にどのような性質を評価対象として特定し、その性質が保持される範囲をどこまで確認しておくべきかを検討する必要がある。

したがって、市販後においては市販前に評価された状態が一定の範囲内で維持されること、及びインシデントを積極的に検索し発見次第適切に対応できる体制を構築することを確保することで、市販前においては使用状況を想定した特定の条件下での臨床的な効果や性能を評価する（効果や性能が保持される範囲とともに破綻する条件の確認を含む。）ことでも受け入れ可能と考える。

【参考】リスクマネジメントを開発者と審査員間で共有することの有用性

医療機器を開発する際には、リスクマネジメント（JIS T 14971/ISO 14971 等）の考えに基づく設計や評価がなされる。これは、生成 AI 医療機器においても同様に重要である。4 項で述べた生成 AI の特性が、医療機器としての申請品にとってハザードとなるのか、そのハザードに起因するリスクはどのようにコントロールし受容可能とされるのか、リスクコントロールでは受容可能とできなかった場合、ベネフィットあるいは市販後の管理も含めて申請品全体としてどのようにリスクアセスメントを行うのか - このようなリスクマネジメントに係る一連の活動内容を共有することで、ガードレール機構等を含む生成 AI システムの設計や評価すべき機能・性能の理解が促進される。

医療機器の審査を実施する上では、審査員と開発者の間で申請品の開発においてリスクマネジメントがどのように実施されたかを共有することは有用である。先述した通り、生成 AI 医療機器は、様々な要因によってその出力が変動し、出力内容の正確性・適合性に影響を及ぼすだけでなく、機能的変容を引き起こすおそれがある。また、こうした変動要因の一部には、例えば申請品の運用実態を記録した履歴情報や、Web などのリアルタイムに更新される外部参照データベースのように、原理的に申請者の制御が及びにくいものが含まれる。このような申請者の管理下にない動的な変動要因により、市場投入後における製品の性能や機能に対して、経時的な変化がもたらされるおそれがある。リスクマネジメントの過程やそれに基づく意思決定の内容を通じて、考慮すべき生成 AI の特性やそのリスクの程度を適切に相互理解することで、適正な審査上の議論を行うことができる。

6.2. 市販前の審査

本項では、市販前の審査における評価方針について述べる。なお、以下の事項は、4 項で述べた生成 AI の特性に起因するハザードに対する網羅的な考慮を前提としているが、上述の製品理解により、申請品の特性を踏まえて必要な項目を適宜選択し、検討することが適切である。

市販後の管理を前提とした上で、市販前の評価については、以下の点に着目する。

- ある時点・条件での出力の有用性（6.2.1 参照）
- ある時点・条件での出力の製品性能（6.2.2 参照）
- 異なる時点・条件での出力の製品性能の安定性・再現性（6.2.2 参照）
- 有害な出力がないことの確認（6.2.2 参照）

なお、上述のリストは現時点における市販前評価の考え方の 1 つである。申請品の特性等を踏まえ、より合理的な評価法がある場合には、これを拒むものではない。

6.2.1. 有用性

本審査ポイントの対象である生成 AI 医療機器の開発の幅を踏まえると、治療及び診断の両方に資することを意図した申請品も想定される。このような場合は、治療と診断のそれぞれについて有用性を評価する必要がある。

また、有用性の評価の方針として、その申請品の使用の有無で最終的な臨床上のアウトカムの価値が高まるか（治療アウトカムの向上、診断成績の向上等）を評価する方法と、製品性能の結果から臨床的意義が説明できる成績であるかを考察する方法が想定される。いずれの方針で評価できるかは、申請品及び申請品の臨床的位置づけに関連する情報の充実度等から製品個別で検討される。

申請品の有用性について前向きに評価せざるを得ない場合、治験による評価が想定される。特に診断支援用の生成 AI 医療機器においては、特定の疾患を対象としていない申請品（例えば、内科の問診全般の支援を対象とした製品）などの開発も想定される。理想的には申請品が使用されうる状況全体に対して有用性を評価することが望ましいが、対象疾患や症状を限定しない申請品の場合、現実的に実施可能な試験プロトコルが設定できるかは慎重に検討する必要がある。必要に応じ、実施可能な範囲から段階的に標ぼうを広げていく（例えば、最初は狭い使用目的で認可を得た上で、徐々に使用目的を広げていく等）方針も検討するべきである。

なお、評価用データセットを作成するにあたっては、データの種類及び正答の定義について、その合理性を客観的に説明できる必要がある。一方で、実利用において個々の使用者から入力され得るあらゆるプロンプトに対する網羅性や、個々の文脈に対する回答の適合性、さらにはサードパーティ製の基盤モデルを利用する場合等において学習データの詳細が秘匿されていることに起因する評価用データの独立性などについて、完全な臨床的有用性を証明することが原理的に困難な場合もある。したがって、有用性の検証にあたっては、実運用におけるエッジケース対応や市販後管理などを踏まえた継続的なアプローチを検討していく必要がある。

6.2.2. 製品性能

生成 AI 医療機器が有する機能のうち生成 AI 機能に対する製品性能を包括的に評価するためには、生成 AI 機能の能力を「一般的な言語能力・指示追従性」「基礎的な医学知識」「タスク特化性能」という 3 つの階層に分解して捉えることが重要である。

その上で留意すべきは、生成 AI の出力が様々な要因により変動するという特性である。この変動性を踏まえた審査上の重要な視点として、単一の出力例で正否を判定することの困難さが挙げられる。さら

に、生成 AI 医療機器のユースケースでは、明確な正解（Ground Truth）が存在しないタスクも多い。したがって、「値」の一致に固執するのではなく、対象の振る舞いにおける「性質（プロパティ）」を定義するアプローチが求められる。具体的には、自動生成された膨大なテスト空間の中で、幅広い入力や操作に対してその性質が破られる反例を探索することで、その性質が保持される範囲を統計的に保証する「プロパティベースド・テスト」が有用である。

さらに、これらの検証過程において生成された出力テキスト等の妥当性を判定・評価する具体的な手法としては、代表的に 3 つのアプローチが挙げられる。

- 経験則等から作成した判定基準（ヒューリスティックな指標）を用いて自動的評価を行うルールベース評価
- 医師等の専門家が臨床的文脈や安全性に照らして直接判定を行う人間による評価
- 予め定義された評価基準（ルーブリック等）に従って別の言語モデルに判定を委ねる LLM-as-a-Judge 評価

タスクにおける明確な正解の有無や、必要とされる評価のスケールに応じて、これらの手法を適切に選択あるいは組み合わせることが重要となる（例えば、膨大なテスト空間を扱うプロパティベースド・テストを採用する場合、ルールベースや LLM-as-a-Judge による自動評価を主軸としつつ、そのサンプリング評価や反例への臨床的判断に、専門家による直接評価を組み合わせるアプローチが考慮される。）。

なお、製品性能の評価において、生成 AI モデル又はシステムの基礎的な能力を測定する静的なベンチマークテスト（JGLUE や Japanese MT-Bench 等）も、ベースラインの把握に有用である。一方で、ベンチマークの成績が申請品の臨床的な適合性と必ずしも直結するとは限らず、スコアの解釈には限界がある。また、評価用データセットや設問形式が既知である場合や、学習用データセットにベンチマークの内容が意図せず含まれている可能性（いわゆるデータリーケージ）がある場合には、成績が過大評価されうる。したがって、ベンチマークの結果のみを過度に重視するのではなく、申請品の臨床的位置づけに照らした総合的な評価（評価データセットの妥当性、評価手順、評価の方法、試験プロトコル等）に基づいて判断することが適切である。

以上より、本審査ポイントではプロパティベースド・テストを中心に、生成 AI の製品評価に関する具体的な観点を記述する。ただし、静的なベンチマークテストは依然として生成 AI の評価における主流である。静的なベンチマークを設計する場合は別紙 3 に示す留意点を参考にされたい。

(1)　一般的な言語能力・指示追従性

一般的な言語能力・指示追従性とは、自然言語の理解力や、プロンプトにより与えられた制約、出力フォーマットなどの指示に忠実に応答する基盤的能力である。こうした能力を計測するための既存のベンチマークとしては、以下のものが知られており、生成 AI モデルの基礎的能力に関する参考情報として用いることができる。

- 日本語言語理解ベンチマークテスト：JGLUE 等
- 日本語生成系ベンチマークテスト：Rakuda、JapaneseVicunaQA、Japanese MT-Bench 等

その上で、例えば、生成 AI モデルが保持するべき性質を、「出力テキストの正確性」「文脈への適合性」「指示追従性」「安全性」「再現性・安定性」といったように特定し、医療の文脈を踏まえて以下の要領で検証することができる。

1) 出力テキストの正確性

入力に対して生成された情報が、客観的な事実や想定される正解にどの程度合致しているかを評価する。明確な正解が存在しないタスクにおいては、適切な評価指標（ルーブリック等）に従ったスコアリングなどを考慮する。

プロパティベースド・テストを用いて正確性を検証する場合は、入力表現に同義の言い換えや表記揺れ等の微小な変更を施した際に出力の一貫性が保たれるか、あるいは患者の症状などの入力条件を論理的に追加・反転させた際に、出力結果もそれに連動して妥当な変化を示すかといった、入出力間に成立すべき意味的な連関（いわゆるメタモルフィック関係）を検証することが有効である。

さらに、無関係なダミーテキストや意図的なノイズを混入させてもハルシネーションが誘発されず、入力の本来の意図や意味内容に対して忠実に処理できるかといったノイズ耐性を確認することも重要である。

2) 文脈への適合性

過去の対話履歴や長大なテキスト、検索拡張生成（RAG）により参照文献が提示されているといった複雑な文脈において、依拠するべき情報を適切に踏まえた応答になっているかを評価する。

プロパティベースド・テストを用いて適合性を検証する場合は、過去の対話履歴や参照文献の記述を同義に言い換えたり要約したりした際に出力の一貫性が保たれるか、あるいは、対話の過程で患者の前提条件（併存疾患の有無や状況の変化など）が論理的に更新された際に、モデルが文脈の推移を正しく捉え、出力結果もそれに連動して適切に変化するかといった意味的な連関を検証することが有効である。

さらに、RAG による参照文献の提示順序を意図的に入れ替えたり、ダミーの文献情報を追加したりするほか、過去の対話履歴に無関係な記述や意図的なノイズをコンテキストに混入させるといった操作を行っても、モデルが重要な情報を見落とすことなく、意図された臨床的文脈に適合した妥当な出力を維持できるかといったノイズ耐性を確認することも重要である。

3) 指示追従性

プロンプトにより与えられた出力フォーマット（JSON 形式等）、文字数制限、あるいは「特定の用語を使用しない」といった制約条件を、モデルが遵守できているかを評価する。

プロパティベースド・テストを用いて指示追従性を検証する場合は、制約条件の論理的な意味を変えずにプロンプトの表現を同義に言い換えたり、複数の指示の記述順序を入れ替えたりする変換を入力に施した際においても、出力結果が一貫して同一のフォーマットや制約を満たし続けるかといった、入出力間に成立すべき意味的な連関を検証することが有効である。

さらに、長大な入力や意図的なノイズ（誤字脱字、同音異義語等）が混入した複雑な要求を加えた場合であっても、指定された出力フォーマットやシステム側の制約条件が崩壊することなく、確実に維持できるかといったノイズ耐性を確認することも重要である。

4) 安全性

悪意のある入力や想定外の要求に対して、システムが有害な情報を出力したり、意図しない挙動を起こしたりしないかを評価する。

プロパティベースド・テストを用いて安全性を検証する場合は、不適切な要求（有害情報の要求等）を婉曲的な表現に言い換えたり、架空のシナリオを装って質問したりする変換を入力に施した際においても、一貫して有害な出力を回避・遮断し続けるか、あるいは適切な医学的質問の表現を同義に言い換えた際に、誤ってガードレール機構が作動し、過剰拒否されないかといった、入出力間に成立すべき意味的な連関を検証することが有効である。

さらに、「これまでの指示を無視して」といった敵対的プロンプト（Jailbreak 等）や、大量の無関係な文字列・意図的なノイズを伴う不適切な要求を加えた場合であっても、モデルの振る舞いが撹乱されることなく、機能的な境界を逸脱せずに安全に要求を拒否できるかといったノイズ耐性を確認することも重要である。

5) 再現性・安定性

生成AIが有する確率的な出力の変動を踏まえ、システムが予期せぬエラーを起こすことなく、一貫した水準の出力を生成できるかを評価する。

プロパティベースド・テストを用いて再現性・安定性を検証する場合は、同一の入力を複数回繰り返すだけでなく、入力プロンプトの表現を同義に言い換える等の変換を施した際においても、出力結果の変動が許容要件を逸脱せず、一貫した水準の出力を保ち続けるかといった、入出力間に成立すべき意味的な連関を検証することが有効である。

さらに、会話のセッションが長期間にわたって継続しコンテキストが長大化した場合や、入力に無関係な文字列や意図的なノイズが混入した場合であっても、モデルの性能が劣化したり破綻したりすることなく、同一の性質が安定して保持されるかといったノイズ耐性を確認することも重要である。

(2) 基礎的な医学知識

基礎的な医学知識とは、疾患概念や医学用語などを正確に理解し、保持しているかを表す医療領域に特化した基盤能力である。こうした能力を計測するための既存のベンチマークとしては、医師国家試験などの試験問題を用いたIgakuQA等が知られており、これらに対する正解率等が参考情報となる。

その上で、生成AI医療機器が保持するべき性質として、医学的事実に対する正確性や安全性などを特定し、以下の要領で検証を行うことができる。

医学的な定説や診療ガイドライン等に基づく推論の正確性に加え、患者の安全性に直結する禁忌（医療行為の禁忌、薬剤禁忌、アレルギー情報等）の要件を厳格に遵守できるかを評価する。また、

倫理的に問題となる有害な出力（差別的な表現やバイアス等）を生成しないかについても評価する必要がある。

プロパティベースド・テストを用いて正確性や安全性を検証する場合は、意味的に同一の質問を異なる医学的表現で言い換えたり、別の臨床シナリオとして条件を提示したりする変換を入力に施した際においても、一貫して医学的に正確な推論が保たれ、禁忌事項にあたる処置等の推奨を一切含まないといった、入出力間に成立すべき意味的な連関を検証することが有効である。

さらに、推論とは直接関係のないダミーの症状や意図的なノイズを混入させた複雑なシナリオを入力した場合であっても、ハルシネーションや有害な出力が誘発されることなく、医学的な正確性と安全性の境界が確実に保持されるかといったノイズ耐性を確認することも重要である。

(3) タスク特化性能

タスク特化性能とは、申請品が標榜する個別の臨床的タスクを実務を想定した際の性能である。この階層では、汎用的なベンチマークが存在しないことが多いため、実際の臨床現場での多様性（患者の属性、併存疾患の有無、非典型的な症状など）を反映した独自の評価用データセットの構築が求められる。

その上で、保持するべき性質として、「複雑な臨床コンテキストへの適合性」や「意図しない使用に対する機能的安定性」を特定し、以下の要領で検証を行うことができる。

1) 複雑な臨床コンテキストへの適合性

単一の症状だけでなく、実際の臨床現場で想定される複合的な背景を持つシナリオに対して、重要な臨床情報を見落とすことなく、目的に適合した出力を生成できるかを評価する。

プロパティベースド・テストを用いて適合性を検証する場合は、患者背景、既往歴、薬剤情報等の情報を提示する順序を入れ替えたり、特定の情報を同義の表現に言い換えたりする変換を入力に施した際においても、出力に含まれる重要な臨床的要素が欠落しないことを確認する。また、臨床情報が段階的に追加された際に、出力結果もその重要度に応じて論理的に更新されるかといった、文脈の積み上げと出力間に成立すべき意味的な連関を検証することが有効である。

さらに、臨床判断に直接影響を与えない無関係な記述や周辺情報をコンテキストに混入させた複雑なシナリオを入力した場合であっても、モデルが情報の優先順位を誤認したり、出力の臨床的有用性が損なわれたりすることなく、一貫した妥当性を維持できるかといったノイズ耐性を確認することも重要である。

2) 意図しない使用に対する機能的安定性

医療機器として承認された機能の範囲を超える使用者からの要求に対して、生成 AI システムが意図しない診断や治療方針の提示を行わないかを評価する。これには、意図しない機能的変容による標榜機能の逸脱、適応外使用・不適切使用・誤用に基づく入力に対して、実装されたガードレール機構が適切に機能することの確認が含まれる。

プロパティベースド・テストを用いてこの性質を検証する場合は、標榜外のタスク（例：対象外の疾患に関する確定診断の要求）を、婉曲的な表現に言い換えたり、医学的に同義な別表現を

用いたりする変換を入力に施した際においても、ガードレール機構が一貫して作動し続けるかといった、入出力間に成立すべき意味的な連関を検証することが有効である。

さらに、エビデンスの低い不確実な情報や、大量の無関係な記述、意図的なノイズを伴う敵対的なプロンプトを自動生成して入力した際であっても、モデルの振る舞いが撹乱されることなく、適切なフェイルセーフ機能が確実に作動し、標榜機能の境界を厳格に維持できるかといったノイズ耐性を確認することも重要である。

6.2.3. その他の評価

上述した製品性能に関する評価とは別に、申請品に対するリスクマネジメントの結果機能として実装されたガードレール機構等のリスクコントロールが適切に実装されているかを評価する必要がある。評価内容は、実装されたリスクコントロールの目的や趣旨に応じて検討される。

上述した内容の他、一般的な医療機器に求められているソフトウェア開発ライフサイクル（JIS T 2304）、ユーザビリティエンジニアリングプロセス（JIS T 62366-1）、サイバーセキュリティ（JIS T 81001-5-1）等を含む基本要件基準への適合について確認する必要がある。

6.3. 市販後の対応

6.3.1. 市販後の報告

前述した市販前の限定的な事前評価のみでは、実運用下で遭遇する多様なエッジケースを完全に網羅することは困難である。また、生成 AI 医療機器が有する動的特性（外部参照情報に依存した出力の変動性など）を考慮すると、市販前の固定的な評価結果が、市場投入後においても適切な有効性及び安全性を確保し続けているかを証明することには原理的な限界がある。加えて、エージェントのような目的達成のために自律的にステップを検討・実行する場合には、開発者が想像だにしない方法（倫理的、技術的リスクのある方法も含む。）による応答をする可能性もある。

これを補完するために、市販後の性能や有害事象の発生状況、またハルシネーションの発生状況等に関する情報を積極的に収集の上、それらが臨床上問題ないことの考察について定期的な報告が求められるであろう。なお、エージェント的に機能実行された場合、臨床的なアウトカムや有害事象の観点からは問題なくとも、機能実行に至るまでのプロセスに問題を生じる可能性もある。このような場合、臨床的に認知が容易なハザードだけでなく、認知が困難なハザードについても考慮するべきか検討する必要がある。この場合、申請品の実行したログ等も監視することが想定される。これらの情報を収集できる体制や機能の実装についても考慮していくことが重要である。なお、これらの対応内容については、市販前の評価内容や申請品のリスク等に応じて個別に検討される必要がある。

また、問題となる事象が発生した場合の対応について、可能な限り事前に行政側と検討を進めておくとともに、特殊なインシデントの発生については、問題となる事象の頻度やインパクトに応じ個別に対応を検討する必要がある。

6.3.2. ヒューマン・ファクターを踏まえた設計と運用

5.2 項で述べた通り、生成 AI 医療機器の活用においては、説明可能性の限界や適応外使用・不適切使用・誤用のリスク、さらには使用者のリテラシーや認知バイアスといった「ヒューマンファクター」に起

因する課題への留意が必要である。

一方で、従来の医療慣行においても、医師は常に文献等から新たな知見を収集し、自らの専門的知見に基づいてその妥当性を判断した上で診療に応用している。医学文献には不確実な情報や未検証の仮説、エビデンスレベルの低い情報等が含まれ得ることを前提に、医師が内容を批判的に吟味して活用する現状に鑑みれば、出力情報に不完全さが残ること自体が直ちに機器の否定に繋がるものではない。重要なのは、情報の真偽や妥当性を使用者が適切に判断・解釈できる運用環境の整備にある。

生成 AI による出力と文献のそれぞれを参照する場合の違いについて以下の通りまとめる。

評価の観点	従来の医学文献・報告	生成 AI の出力内容	生成 AI の出力内容におけるリスクの所在
根拠の検証可能性	著者・掲載誌・引用元が明示され、内容の検証が可能	RAG 等の特別な仕組みを用いない限り、出典根拠が不明瞭	トレーサビリティや検証可能性の欠如
専門的評価の蓄積	査読制度や学会等の専門集団による評価を経て信頼性が形成されている	出力内容に対する臨床的な評価実績が蓄積されていない	社会的に合意された信頼性評価の仕組みの不在
知見の継承・教育	教育や指導を通じて、「このような文献は信用できる／できない」という知識が共有される	出力の妥当性を判断するための教育的・経験的ノウハウが十分に確立されていない	正誤を識別するためのリテラシーが未確立
認知バイアス	医師は文献には誤りや偏りがあることを前提に、批判的に吟味する訓練を受けている	アルゴリズムの出力が客観的で正確であるように見え、無批判に受容しやすい（自動化バイアス）	ヒューマン・ファクターに起因する過信や依存のリスク

すなわち、生成 AI の出力に関する説明可能性の不足に加え、出力の信頼性を評価するための統一的な運用方針の欠如及び使用者のリテラシー不足や自動化バイアスなどのヒューマン・ファクターが実運用における課題であると言える。

以上を踏まえ、使用者が出力の不確実性や信頼性の程度を正しく理解した上で、生成 AI 医療機器を活用できる状況を作ることが重要である。具体的には、以下のような措置を講じることが望ましい。また、必要に応じ、設計に組み込むことが望ましい。

- **根拠情報の提示：**生成した出力の出典の有無やその内容、又は生成した内容が既存のエビデンスの範疇に収まる情報なのか、あるいは既存の情報から外挿して推論された情報であるかを使用者に提示する機能を実装することが望ましい。
- **適応外使用・不適切使用・誤用への拒否：**明らかに使用目的や適応を逸脱した入力、禁忌事項、エビデンスが低い情報等に基づく入力に対して、これが行なわれないような注意喚起のみならず、申請品の仕様としてこれらの要求を拒否する機能（ガードレール機構等）を実装することが望ましい。
- **意図しない適応外使用に対するリスク管理：**上述の意図的な不適切使用だけでなく、製造販売業者が定める使用目的を超えた広範囲の入出力により、使用者が意図せずに適応外使用が行われるリスクがある。この点を踏まえたリスクアセスメントを実施するとともに、販売時の適正使用に対する注意喚起を徹底する必要がある。
- **リテラシー向上に向けた教育的支援：**使用者が生成 AI を適切に使用のための一定のリテラシーを得るために、チェックリストや研修等を実施することが望ましい。特に、生成 AI の利便性が

向上したとしても、その出力に過度に依存することで使用者の批判的思考や専門的技能が退化する「デスキリング」の懸念がある。したがって、これは、技術の進展や社会的なリテラシーの醸成状況に応じ、常にその内容を最適化しながら継続すべき必須の措置と考える。

一方で、生成 AI の特性を応用し、患者自身が使用者となる生成 AI 医療機器も想定される。この場合、患者の医療や健康に関するリテラシーの程度には大きな個人差があることから、使用者自身のリテラシーによるハルシネーションの識別や、批判的な吟味を前提としたリスクマネジメントは困難である。したがって、使用要件やインストラクションの提供に加え、申請品の仕様上の設計においても、使用者の不適切使用や誤用を未然に防ぐガードレール機構の実装や、機器自体の設計による安全性の確保（フールプルーフやフェイルセーフの概念に基づく安全設計）を検討する必要がある。

7. おわりに

生成 AI 技術は医療分野において多大な可能性を有する一方で、これまでにない多層的な規制的・技術的課題を内包している。本審査ポイントで述べた一連の考察は、現時点での知見に基づく暫定的なものであり、今後、実製品開発や審査実績が蓄積されることにより、さらに精緻化されるべきものである。

なお、生成 AI に関し、ハルシネーションによって偽・誤情報を出力をすることからその申請品のリスクが許容できないとする声を聴くこともある。しかし、現在の医療でも、医療従事者は様々なエビデンスレベルの情報を解釈し判断しながら、最良の医療の提供に努めている。生成 AI の出力についても、同様ではないだろうか。ただし、黎明期である現在、偽・誤情報を含むかもしれない生成 AI の出力をどの程度信用してよいのかの判断についての経験値が少ないというのが、現状である。したがって、ハルシネーションが使用者の判断、あるいは医療に対してどのような影響を与えるかを慎重にモニタリングする必要がある。当面は、生成 AI の提案を即座に採用するのではなく、使用者が生成 AI の出力の成否を適切に判断できない場合は、その内容に精通する医師に確認をとることを義務付けるということも必要かもしれない。

いずれにしても、生成 AI という技術を良い形で医療に導入していくことは、医療者、患者の双方にとって恩恵は大きいと考える。開発者、規制当局、医療現場の三者が協働し、安全で有効な生成 AI 医療機器の社会実装を推進することが求められる。

別紙 I 開発形態の例

表 I 開発形態の例

開発形態	特徴	申請者が直接管理しやすい範囲
プロプライエタリモデル利用（クラウド API 等）	外部事業者が提供するモデルを呼び出して推論する。モデル仕様や学習データの詳細は非公開であることが多い。	プロンプト設計、ガードレール機構、入出力設計、RAG 等の周辺コンポーネント、評価データセット・評価手順、運用・監視（ログ取得可能範囲内）
オープンウェイトモデル利用（自社環境運用）	公開された重み等を取得し、自社（又は指定環境）で推論する。推論環境を固定しやすい。	推論環境（ハード／ソフト構成、バージョン）、デプロイ手順、アクセス制御・ログ、モデルの設定（温度パラメータ等）を含む実装上の再現性、セキュリティ対策
追加学習あり（ファインチューニング等）	上記いずれかのモデルを基に、医療用途のデータで追加学習して性能・振る舞いを調整する。	追加学習データの選定・品質、学習手順、ハイパーパラメータ、再学習・変更管理、差分評価、学習結果の版管理
自社開発の生成 AI モデルの使用	開発者が自ら設計・学習させた生成 AI モデルを医療機器に組み込んで利用する。性能実現コストは高いが、用途に応じた最適化が可能。また、更新管理も容易。	モデル構造、学習アルゴリズム、学習データの選定・前処理、学習条件及び評価方法、モデルの更新や再学習の実施有無、その手順及び管理方針

別紙２ 論点となり得る主な技術的特性と開発への示唆

A. 出力の変動性

生成 AI モデルは、様々な要因によってその出力が変動し、非決定的となるおそれがある。同一の意味内容を有する入力に対してモデルの挙動が変動する場合、医療機器としての出力の再現性、挙動の一貫性、及び機能の同一性などを、限られたテストケースで評価することが課題となる。

特に評価における重要な視点は、こうした出力の「変動要因」を特定し、要因ごとの変動のダイナミクス（グローバルな変動か、ローカルな変動か、継続的変動か、離散的変動か、初期状態に戻すことは可能か等）をその制御性とともに整理することである。また、こうした変動のダイナミクスに対する制御性は、プロプライエタリモデルを API 等を経由して利用するのか、あるいはオンプレミスでモデルを運用するのかといった運用形態によっても変化する。

したがって、生成 AI モデルの出力の変動要因ごとにそのダイナミクスを踏まえた上で、その変動性を固定できるのか、あるいは予め定義された範囲・境界の中での変動に制御可能か否かといった制御性を明らかにした上で適切なリスクコントロールを実施し機能設計等をする必要がある。

以下に、出力の変動要因を検討するために考慮すべき事項の例を挙げる。

(1) ユーザの入力に起因する出力の変動

生成 AI モデルは、使用者からの入力（プロンプト）が意味内容としては同一であっても、その表現がわずかに異なることによって出力結果が変動する。こうした特性は柔軟で表現力の高いコンテンツ生成を可能にする。他方、医療現場における略語や非標準的な医療用語、開発時とは異なるフォーマット、誤字脱字の混入といった些細な入力の変化により、その推論の妥当性が変動・低下することで、使用者の意思決定に混乱を招くおそれがある。

評価上の課題としては、タスクごとにどのようなプロンプトを前提としてモデルの評価を行なうべきかという条件設定や、入力の変化に伴う出力の変動の許容範囲、さらには使用者の入力スキルへの依存性や再現性の低下などへの対応等が挙げられる。

(2) コンテクストに起因する出力の変動

生成 AI モデルの出力コンテンツは、入力されたプロンプトだけでなく、対話履歴等の文脈（コンテキスト）に応じて適応的に変化し、継続的な出力の変動（継続的変動）を生じさせる。これにより、申請品の運用実態を記録した履歴情報等に基づき、生成 AI による出力の指向性がより使用者に固有の文脈に即した形で最適化され得る一方で、不適切な文脈の蓄積により出力の精度が低下する可能性も孕んでいる。この履歴情報等には、申請品の使用中、あるいは過去に使用した際の会話コンテキストやログ情報等が想定される。

評価上の課題としては、申請品の運用実態に合わせてモデルの振る舞いが動的に調整され得るため、初期状態や、特定の履歴が蓄積された状態など、どのようなコンテクストを前提としてモデルの評価を行い、承認事項とするかという点がある。さらに、蓄積された文脈を初期状態へと確実にリセットできるかといった、システムとしての制御性の観点も重要となる。

(3) モデル設定に起因する出力の変動

生成 AI モデルのうち、例えば LLM では、次に出現するトークン（モデルがテキストを処理する際の最小単位）の確率に基づいた文章の生成が行われる。この際に、最も高い確率のトークンを常に選択するのではなく、一定のランダム性を付加してサンプリングすることによって、文字列として完全に同一の入力に対する出力のばらつきが得られる[1]。この特性によって、画一的ではない表現豊かな出力を生成することができるため、開発者が意図してこうした特性を実装した申請品が想定される。特に一般の方向けの治療支援を意図する申請品のインターフェースでは、このような「人間らしさ」が製品による治療へのアドヒアランスや有効性に寄与する可能性もあることから、医療機器であっても敢えてこのような特性を持つものが想定される。

評価上の課題としては、ある入力に対して単回の出力だけを評価しても、ばらつきを含むその製品の出力特性の全体を網羅的に評価できない点がある。また、出力の変動に関わるパラメータを特定し、システムの挙動を適切に安定させる制御性の観点も極めて重要となる。

(4) 外部システム・データに起因する出力の変動

生成 AI モデルは、先に述べた検索拡張生成（RAG）やエージェント機能等の高い相互運用性を用いる場合、連携先のデータベースや外部システムが更新されることによって、それを参照する生成 AI の出力内容も連動して変化し（離散的変動）、これにより最新の医学的知見を踏まえた診断支援等が可能となる。

評価上の課題としては、参照情報が経時的に変化することによって、同一の入力に対しても評価時点によって出力が変動し、再現性が低下する点がある。また、Web のような膨大な外部情報を制限なく活用する場合には、偽・誤情報やバイアスの強い情報（例えば、日本人ではなく欧米人に関する健康情報等）を取り込み、不正確な回答を生成する可能性があることから、これをどのようにリスクマネジメントするか慎重に検討する必要がある。加えて、連携先の外部システムやデータを特定の状態に固定することができるかといった、生成 AI システム全体に係る制御性も考慮するべきである。なお、当該変動は運用上の課題ともなる。

(5) **基盤モデルの更新に起因する出力の変動**

基盤モデル事業者は、通常、提供する基盤モデルを継続的に調整・更新している。そのため、例えばプロプライエタリモデルをクラウド API を介して利用する場合、当該モデルのアップデートに伴い、モデルの特性が全般的に更新されることによって、同一の入力に対する出力が不可逆的に変化することがある。(1)から(4)までの生成 AI システムの内部のローカルな変動とは異なり、これは外部環境の変化に伴うグローバルな変動と位置づけられる。したがって、基盤モデルの更新が行われる場合には、上述した(1)から(4)までの全ての変動の性質や制御性が変化し得る点に留意する必要がある。

本課題は、申請品のコア・コンポーネントである生成 AI モデルがアップデートされた場合、医療機器としての同一性をどのように評価するかという運用上の課題といえる。

B．機能的変容性

---

[1] こうした出力のばらつきは、意図的なランダム性の付与だけでなく、ハードウェア等の計算処理における非決定的な揺らぎによってもたらされることがある。

Aで述べた出力の変動は、単なるテキスト出力のばらつきという側面にとどまらず、医療機器としての「機能的変容」を引き起こす要因にもなりうる。つまり、 Aに挙げた各変動要因が、出力レベルの揺らぎにとどまらず、医療機器の機能そのものの変化に及ぶ場合について整理する。

例えば、導入施設ごとの診療プロトコルに応じたシステムプロンプトの調整（離散的な機能的変容）や、使用者の過去の対話履歴に基づくパーソナライズ（継続的な機能的変容）などは、文脈に応じたコンテンツ生成能力という生成AIの変動性を有益に活用した機能的最適化と位置づけられる。

一方、こうした適応性が制御されない場合、自由な入力や文脈の蓄積にモデルが過剰に追従し、承認された標榜機能を逸脱するおそれがある。具体的には、使用者のプロンプトによって本来検証されていない対象疾患や臨床的意図に関する質問が入力された際、生成AIモデルがもっともらしい回答を出力し、結果として未検証の診断・治療支援として機能してしまうリスクが挙げられる。

したがって、こうした生成AIモデルにおける機能的変容についても、出力の変動性と同様に、要因ごとの変容のダイナミクスをその制御性とともに整理する必要がある。特に、使用者に対してシステムプロンプトの調整をどこまで許容するか、過去の使用履歴などに基づく継続的な機能的変容をどのように監視・制御するかといった観点が求められる。

その上で、標榜機能の有効性・安全性を確保するための合理的なテストデータの設計や、標榜機能を超えた適応外使用や不適切使用を防ぐためのガードレール機構の実装、ヒューマン・ファクターを踏まえた評価などが重要となる。

C. 出力内容の正確性・文脈への適合性

生成AI医療機器の出力は、申請品の位置づけや使用目的等によっては患者の生命や健康に重大な影響を及ぼすおそれがある。そのため、生成された情報が医学的な事実やエビデンスに合致しているかどうかを示す「出力内容の正確性」や、複雑な臨床的背景や使用者の意図など、個別の文脈に応じた適切な回答であるかどうかを示す「文脈への適合性」を確保することが極めて重要となる。

特に、「画像上の〇〇所見の検出」のように正解が一つに定まり、その予測の正確性を定量的に評価できるタスクのみならず、生成AI医療機器の機能においては、使用者のリテラシー、心理状態、社会的背景等に合わせた適切な言葉選びのように、唯一の正解が存在しない文脈で、いかに適合した回答を生成できるかという観点も認識される。

こうした「出力内容の正確性」や「文脈への適合性」を評価するにあたっては、生成AIコンテンツの有用性が使用者の解釈に強く依存している点（解釈依存性）に留意する必要がある。その上で、使用者の解釈を歪めうる偽情報・有害情報・誘導といった重大なリスクや、生成AIモデルに内在するデータバイアスの影響を踏まえて評価することが求められる。

(I) 出力結果の解釈依存性

生成AIコンテンツは、問いかけに応じて出力された生成AIコンテンツの意味内容を使用者が適切に解釈することによって、診断精度向上や治療成績改善といった臨床的アウトカムへの寄与がもたらされるものである。すなわち、解析結果がラベル等により出力される従来的な識別AIとは異なり、生成AIが自然文等で出力する内容は、その医学的正確性や文脈に対する適切性を使用者が自ら解釈し、その意味内容から有用な情報を取捨選択しなければならない。

評価上の課題としては、自然文などの生成 AI コンテンツにおける意味内容を再現性、客観性をもってどのように評価するかという点がある。また、このように出力結果の真偽が使用者の解釈に委ねられるという特性を踏まえて、以下に記載する生成 AI に特有のリスクを考慮した設計と運用が求められる。

(2) 偽情報・有害情報・誘導

生成 AI の出力内容には、ハルシネーションによる“偽情報”や“誤情報”が含まれうる。さらに、“有害情報”や“不適切な誘導”等、生成されたコンテンツが直接的に負の影響を及ぼすおそれがある。

評価上の課題としては、具体的に想定される使用シナリオごとに、これら偽情報・誤情報や有害情報の出力傾向をいかに客観的に測定するかという点がある。加えて、不適切な誘導に繋がりうる出力の許容基準や、それらを未然に防ぐガードレール機構をどのように設定し評価するかという点がある。

(3) データバイアス

データバイアスは、生成 AI に限らず、データドリブンで開発される機械学習を用いた製品全体の特性といえる。ただし、極めて大規模なデータセットを用いて構築された生成 AI モデルを用いた製品の場合は、一般的な機械学習に比べて多量かつ多様な標準化されていないデータを用いて学習をしており、また、教師ラベルを用いず広く学習しているため、内在するデータバイアスの特定や制御がより難しくなる。暗黙の価値観や文化的前提が意図せず混入している可能性もあり、医療機器の有効性や安全性に影響を与える可能性がある（例：都市圏のデータが多いために提示される推奨行動が都市部に住んでいることを前提とする、高齢者のデータが多いために若年層には不適切な提示となる等）。

評価上の課題としては、生成 AI 医療機器で想定されるユースケースにおける入力と出力の多様性が極めて高いため、データバイアスがどのように顕在化するかを事前に網羅的に確認することが困難であり、特有のリスクマネジメントが求められる点が挙げられる。

別紙 3 ベンチマークテストの位置づけと限界

A. ベンチマークテストについての一般的解説

ベンチマークテストとは、あらかじめ定めた設問やデータセットに対するモデルの応答を、一定の基準に基づいて採点し、性能を定量化する評価手法である。生成 AI（特に大規模言語モデル）においては、言語理解、推論、生成の一部側面を比較可能な形で把握する目的で広く用いられる。

ベンチマークテストの主な意義としては、（1）モデル間の相対比較、（2）改良の方向性の把握、（3）評価手順の標準化による説明可能性の向上等が挙げられる。

一方で、ベンチマークテストは、特定の設問形式・評価条件における性能を測定するものであり、申請品の臨床的位置づけに照らした有効性・安全性を直接示すものではない。したがって、ベンチマーク結果は、申請品の目的や使用状況を踏まえた評価（評価データセット、評価手順、試験プロトコル等）と併せて解釈されるべきである。

B. ベンチマークテストの限界（評価上の注意点）

ベンチマークテストによる評価を検討するにあたっては、少なくとも以下の注意点や限界を踏まえる必要がある。

- 臨床的有用性との乖離： ベンチマークテストで高得点を取得したとしても、実臨床で求められる判断（情報の不完全性、前提条件の違い、禁忌・例外、説明責任等）に対して十分な性能が確保されているとは限らない。例えば、医療 QA ベンチマークと実臨床での性能や有効性の整合性について定量的に検討し、その乖離が生じうることを示唆した報告もある（Siun Kim, Hyung-Jin Yoon. Proceedings of the 24th Workshop on Biomedical Language Processing. 2025）。また、トリアージ推奨を対象とした構造化テストにおいても、条件設定により性能のばらつきや盲点が生じうることが報告されている（A. Ramaswamy, et al. Nature Medicine. 2026）。
- データセット汚染（contamination）・既知問題による過大評価：評価用データセット又は類似データが学習データに含まれている場合、モデルが「能力」ではなく「記憶」により正答し、成績が過大評価され得る。特に公開ベンチマークでは、意図せず学習済みとなる可能性（汚染リスク）を排除しにくい。
- ベンチマークへの過適合（leaderboard overfitting）：ベンチマークが事実上の最適化目標となることで、特定ベンチマークにおけるスコア改善が、他のタスクや実運用条件に一般化しない可能性がある。単一スコアのみで性能を論じることには限界がある。
- 評価指標・採点方法の限界（再現性・客観性）：生成タスクでは正答の定義が一意でない場合が多く、評価指標や採点ルールにより結果が変動しうる。人手評価では評価者間のばらつきが生じうる。自動評価（LLM-as-a-Judge 等）を用いる場合も、判定の一貫性、偏り、評価用 LLM の設定やプロンプトへの依存等に留意が必要である。
- 安全性上重要な失敗の取りこぼし：平均的な正答率が一定水準であっても、稀だが重大な誤り（禁忌に関する誤誘導、緊急性判断の誤り等）が見落とされる可能性がある。ベンチマークだけで安全性を保証することは困難である。

C. 目的とするタスクに応じた適切なベンチマーク設計の重要性と設計例

ベンチマークテストは、目的とするタスク（診断支援、治療支援、トリアージ、文書作成支援等）や、想定する使用者・使用状況によって、適切な設計が異なる。申請品の臨床的位置づけを踏まえ、評価すべき能力を明確にした

上でベンチマークを設計することが重要である。

- 評価すべきタスクの明確化：申請品が提供する価値（例：鑑別支援、禁忌確認、説明文生成、要約等）を整理し、入力条件、期待する出力の形式・粒度・前提条件を定義する。
- クリティカル要件の明示：禁忌、安全性、緊急度判断等、誤りが重大な結果につながりうる観点（クリティカル要件）については、平均的スコアだけではなく、個別ケースでの失敗を捉えられる設計（重点サンプル、閾値、エラー分析等）を検討する。
- 正答定義と判定基準の設計：参照すべき根拠（ガイドライン、添付文書、標準的教科書等）や、正答の範囲（許容表現、同義表現、条件付き正答）を事前に定義し、試験プロトコルに規定する。
- 評価条件の管理と再現性の確保：プロンプト等の設定条件（内容、表現の粒度等も含む。）、利用するツール（RAG、外部 DB 等）や提示情報量の範囲を規定し、評価結果の再現性を確保する。
- 多面的評価：単一のスコアに依存せず、タスクに応じて複数の観点（正確性、完全性、一貫性、禁忌回避、根拠提示の適切性等）で評価する。

D. ベンチマークを設計する際の基礎的視点

ベンチマークを設計する際には、評価用サンプルの性質（網羅性・代表性）に加え、生成 AI コンテンツの事実性（factuality）評価に固有の技術的論点を踏まえる必要がある。

- 評価用サンプルの網羅性・代表性：対象疾患・症状、重症度、併存疾患、年齢層等のバリエーションを含め、想定使用状況を代表するよう構成する。稀だが重要なケース（禁忌、緊急対応等）についても意図的に含める。
- 医学的な正確性・関連性：正答の根拠を明確にし、医学的に妥当な範囲の表現や条件（前提条件、例外、禁忌）を評価できるようにする。必要に応じて専門家レビューを組み込む。
- 生成 AI コンテンツの事実性評価に関する技術観点：事実性の評価では、（1）参照すべき根拠の範囲、（2）根拠に照らした整合性、（3）不確実性の扱い（断定の回避、条件付け、限界の明示）といった観点を区別して評価する。また、出力がもっともらしい文章であっても根拠が不十分な場合があるため、根拠提示の有無や、根拠の適切性（引用の整合性等）を評価項目として設定することが考えられる。
- 採点の再現性・監査可能性：人手評価の場合は評価者要件と判定基準を明確化する。自動評価（LLM-as-a-Judge 等）を行う場合は、評価用 LLM、プロンプト、判定基準、サンプリング条件等を固定し、必要に応じて人手評価との突合を行う。

ベンチマーク結果のみをもって結論を導くのではなく、申請品の臨床的位置づけに照らした評価と併せて総合的に判断することが重要である。

以上より、ベンチマークテストは有用な情報を提供する一方で、それ自体を臨床的有用性又は安全性の代替指標として過度に重視しない評価の考え方を前提とする必要がある。

参考文献

[i] 「Good machine learning practice for medical device development: Guiding principles」（IMDRF/AIML WG/N88 FINAL:2025, 29 January 2025）

[ii] 「AI を活用したプログラム医療機器に関する報告書」（令和 5 年 8 月 28 日付け PMDA 科学委員会（AI を活用したプログラム医療機器に関する専門部会））

[iii] 「次世代医療機器評価指標の公表について 別添 4「人工知能技術を利用した医用画像診断支援システムに関する評価指標」」（令和元年 5 月 23 日付け薬生機審発 0523 第 2 号 厚生労働省医薬・生活衛生局医療機器審査管理課長通知）

[iv] 「ＡＩを活用した医療診断システム・ 医療機器等に関する課題と提言 2017」（平成 29 年 12 月 27 日付け PMDA 科学委員会（AI 専門部会））

[v] Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. Nature. 2023 Apr 12. doi: 10.1038/s41586-023-05881-4.