医用画像の読影支援を目的としたコンピュータ診断支援プログラムの審査のポイント

独立行政法人 医薬品医療機器総合機構

2023 年 3 月 7 日

目次

医用画像の読影支援を目的としたコンピュータ診断支援プログラムの審査ポイント

はじめに

承認申請に際し、平成 26 年以降に承認を取得した医療機器プログラムから審査のポイントに関する情報を整理し、公表する。

- 本審査ポイントは、承認申請に際し、資料の作成の効率化及び審査の迅速化に資するため、規定する適用範囲に示す医療機器について、必要な評価項目等を示すものであること。
- 本審査ポイントは、現時点における科学的知見に基づき審査の考え方について示したものであり、今後の科学技術の進歩等に応じて随時見直され、改訂されるべきものであること。

1. 本文書の対象となる製品

「次世代医療機器評価指標の公表について」（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）別紙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))によれば、コンピュータ診断支援は以下のように定義されている。

CADe (Computer-Aided Detection)：画像上で病変の疑いのある部位をコンピュータが自動検出し、その位置をマーキングする機能を有する単体ソフトウェア又は当該ソフトウェアが組み込まれている装置。コンピュータにより医用画像データのみ又は医用画像データと検査データの両方を処理し、病変又は異常値の検出を支援する。

CADx (Computer-Aided Diagnosis)：病変の疑いのある部位の検出に加え、病変候補に関する良悪性鑑別や疾病の進行度等の定量的なデータを数値やグラフ等として出力する機能を有する単体ソフトウェア又は当該ソフトウェアが組み込まれている装置。診断結果の候補やリスク評価に関する情報等の提供等により診断支援を行うものを含む。

本文書では、医用画像で読影をする際の診断支援を行う CADe を対象に、承認申請に向けた審査のポイントについてまとめるものである。疾病の重症度の示唆や、検出したものの分類等は行わない（複数種類の所見を検出するものは対象とするが、検出した所見の種類を分類・分別して提示するものは対象としない）。また、Second reader 又は concurrent reader として用いられるものを対象とする。

- 本文書では、以下の意味で説明する。
  - 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。
  - 当局：規制当局を指す。審査・相談時には、PMDAを指す。

2. 申請品目の説明

2.1. 臨床的位置づけの整理

申請品に求められる評価パッケージの充足性や評価試験の妥当性を議論するためには、その申請品の仕様（入力データ、出力データ、性能に関する数値等）だけでなく、その申請品の臨床的位置づけを明確にして当局と共有することが重要である。臨床的位置づけとは、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ということである。

画像診断支援を意図した CADe においても、臨床的位置づけの整理は重要である。例えば、医用画像上の所見検出支援をする CADe に対して、次のようなバリエーションが想像できる（このバリエーションだけに留まるものではない）。

例）

- CADe の使用により読影医自身の診断成績を向上させることで、読影時の見落とし防止の支援を行う。
- 読影・検出が容易な所見を検出対象とせず、読影・検出が困難な所見のみを対象に検出支援をすることで、読影困難な所見に対する診断成績を向上させる。
- 二重読影体制が求められている状況において、CADe を 2 人目の読影医として扱うことにより、二重読影体制を解消する（このような開発が現実的に可能か、各種団体や医師法への配慮なども必要である）。

上のそれぞれにおいて、申請品が目指すべき性能や求められる評価の内容は異なる。臨床的位置づけを理解するにあたり、例えば、以下のような情報を整理するとよい。ただし、申請品の開発コンセプトに応じ、適宜説明を追加等すること。

- 対象とする疾患・患者等は何・誰か。
- 対象とする疾患・患者等に対する現状の診療において、どのような課題があるのか。
  - ガイドラインや文献等から、定量的に説明できる情報がある場合は、それらと合わせて説明することが望ましい。
- 申請品はその課題をどのような方法で解決するのか。
  - どのようなものを検出支援するか。
  - 誰（専門医、非専門医、コメディカル等）にどのように使用させるのか。
- 申請品を医療現場に導入することで既存の診療はどのように変わるのか（臨床的意義）。
  - 診療のフロー等がどのように変わるか。
  - 患者にとってどのような医学的メリットを享受しようとしているか。

既存の診療がどのように行われているか、またそこにどのように導入したいかという点は、性能の評価だけでなく申請品のリスク（主に、偽陽性・偽陰性のインパクト）を検討する上でも重要である。既存の診療の説明に対しては、各種診療ガイドライン等を引用しながら説明することが望ましい。

相談に際しては、以上を踏まえた承認を得たいとする使用目的の案も説明することが望ましい（※）。

※ あくまで相談時において相談者がどのような承認を得たいか理解するために記載を促すものである。最終的な使用目的は相談時には決定されず、審査を経て決定されることに留意されたい。

2.2. 設計のコンセプト

2.1 項で説明された臨床的位置づけに基づき、どのような機能・性能をもつ製品を開発しようとしたか（設計コンセプト）を整理する必要がある。言い換えれば、臨床的位置づけを達成するために、どのような機能が必要と考えたか、またその機能にはどの程度の性能が必要と考えたかを説明することともいえる。

画像診断支援を意図した CADe の設計コンセプトに基づき、2.2.1～2.2.3 項に関して整理すること。

2.2.1. 機能

申請品の具体的な機能を説明する必要がある。なお、承認申請時には、機能や仕様は具体化されている必要があるが、相談においては、相談を受けようとする時点の開発フェーズによっては、一部不確定な機能・仕様があることは致し方ない。この場合、それらの仕様が検討中であることが分かるように説明すること。

申請品が有する CAD 機能を記述する要素となる例を以下に示す。各申請品固有の機能・性能等に応じ、具体的に記載する必要があることに留意すること。

※「○○等をする機能」「○○を処理する機能」といった記述のみでは、申請品が有する機能の全貌や、どのようなアウトプットをする機能であるかが理解できない。網羅的かつ具体的な記載となるよう申請資料・相談資料の作成時には配慮されたい。

※ 申請・相談資料に基づき審査・相談が実施されるため、記載がない機能について審査員らは認知することができない。審査・相談中に、新たな機能が認知された場合、審査・相談の長期化や審査継続が困難となる可能性があるため、申請品が有する機能を網羅して審査員と共有するように準備をお願いしたい。

(1) インプット

- 解析対象となる被験者集団はどのような集団か。
- 撮影モダリティは何か。
- 撮影条件は何か。
- 撮影機種のベンダは何か。
- 画像処理（強調処理、フィルター処理等）が施された画像を用いるか/ raw data（再構成処理が必要な装置で処理前のデータ）を用いるか。
- 造影剤の有無

※ 申請品が適切な性能が発揮できるものとして指定するインプットの条件として設定されたい。また、試験時に評価対象として検討すべき条件となることに留意されたい。

(2) 検出対象

- どのような所見等の種類、性状を対象とするか。
- 医師が見つけにくい所見等、特殊な条件があればその内容は何か。
- どのような所見は検出しないのか。
- 従来の診療において、医師が解析対象画像から検出できる所見等か否か。

(3) 解析原理

- どのような解析原理か。
  - 演繹的に設計している場合は、どのような処理アルゴリズムか。
  - 機械学習を用いてモデルを実現している場合は、どのような学習アルゴリズムか、開発データ（データ収集施設、正解となるラベルの設定方法、どの程度の学習データ量か）は何か。（5 項も参照されたい。）
- 最終的な出力はどのように決定されるか。
  - どのような部分を統合したり、ポイントを決定したりして提示するか。
  - 判定閾値はいくつか。

(4) アウトプット

- 症例、画像単位にフラグを付与するか。又は、画像上の検出対象所見の候補にフラグを付与するのか。
  - 画像上のフラグ付与についても、点・バウンディングボックスやサークル・ヒートマップ等、どのようなフラグを付与するか。提示する寸法は不変か可変（所見サイズに応じて変わる等）か。
  - フラグは何個同時に表示させられるのか。表示個数に上限はないか。
  - 解析過程で算出される確信度（又はこれに相当するスケール等）を表示するのか。この場合、表示される確信度にどのような意義を持たせて情報提供するか。
  - 複数種類の所見を検出対象とする場合、これらの所見の種類を区別せず表示するか、区別して表示するか。

(5) その他

- CAD 機能以外にどのような機能があるか。その他の機能についても、上述の(1) – (4)を参考に、機能の詳細について説明すること。

2.2.2. 使用方法

申請品の具体的な使用方法を説明する必要がある。なお、承認申請時には、使用方法は具体化されている必要があるが、相談については、相談を受けようとする時点の開発フェーズによっては、一部不確定な使用方法があることは致し方ない。この場合、使用方法が検討中であることが分かるように記載すること。

申請品の使用方法を記述する要素となる例は以下の通りである。なお、各申請品固有の臨床的位置づけ等に応じ、具体的に記載する必要があることに留意すること。

(1) 使用者

- どのような医師等が使うのか。
  - 診療科や習熟度、資格の有無等についての指定がある場合はこれも説明すること。
  - また、使用を制限する医師等の条件があれば、説明すること。

※ 例えば、主たる使用者が「非専門医」であるとしても、「専門医」の使用を否定しない場合は、「専門医」は申請品の使用者として検討される必要があることに留意されたい。

(2) 使用方法

- Concurrent reader、second reader 等、実診療においてどのように申請品の結果を確認するか。
  - 特に second reader の場合、申請品が提示したフラグをすべて見直すのか、医師等は検出しなかったが申請品が検出したものだけを見直させるのか等。

※ 3.1 項で述べる読影試験を実施する場合、当該使用方法に基づき試験デザインが検討されることに留意されたい。また、試験デザインの妥当性を検討するにあたり使用方法をどのようにするか事前に検討する必要があることに留意されたい。

2.2.3. 性能

設計コンセプトに基づき、申請品の CAD 機能に対して、どのような画像に対してどの程度の性能を有しているべきと考えたかを説明すること。「どのような画像」は、一般的には 2.2.1 項(1)に対応する。「どの程度の性能」は、例えば以下のような内容となる。なお、これは単体性能試験の評価デザインと必ずしも一致しないことに留意されたい(申請パッケージ全体で評価されていればよい)。

- 診断をする専門医の診断成績と同等の成績で検出対象を検出できる。
- 診断をする非専門医の診断成績を超える成績で検出対象を検出できる。

2.3. 類似製品に関する情報

申請品の類似製品がある場合は、類似製品との比較をしながら申請品に関する情報を記載する。類似製品の比較方法については、「医療機器の製造販売承認申請書添付資料の作成に際し留意すべき事項について」(平成 27 年 1 月 20 日付薬食機参発 0120 第 9 号通知)等を参考にされたい。

3. 評価パッケージ

1 項及び 2 項で整理した臨床的位置づけや設計コンセプトを踏まえ、申請品の臨床的な有用性、臨床性能及び基本的な性能が評価できるよう評価パッケージを検討する必要がある。

以下に、多くの CADe に共通するハイレベルな概念的要求事項を示す。ただし、同じモダリティから同じ対象を検出する CAD であっても、どのような使用目的を持たせるか、誰がどのように使うか

といった臨床的位置づけや表示の方法等により、試験プロトコルの詳細は変化する。したがって、類似の前例品がどのような試験プロトコルで評価されているかを参考にすることは有益であるが、基本的には各申請品の内容に合わせて調整することが必要と考えるべきである。

<概念的要求事項> ※申請品の特性等に合わせて調整すること。

(1) 意図した入力データに対して申請品が解析した結果を用いることで、意図した使用者の診断成績が向上すること。(臨床的有用性)
(2) 意図した入力データに対して、申請品が臨床的に意義のある検出性能を有すること。(臨床性能)
(3) 臨床上許容できる時間内で処理が完了できること。(基本的な性能)
(4) その他の機能が、意図したとおりに動作すること。(基本的な性能)
(5) ソフトウェアライフサイクルが適切に管理されていること。

上述したように、これらの申請品の臨床的位置づけに応じて、各項目をどのように評価するか検討する。なお、基本的な考えとして、申請品が有する機能についてはいずれの機能についても一定の評価は必要である。ただし、どの機能を検証試験として評価するか、副次的な試験として実施するか、又は単に動作確認するに留めるかは、その機能の位置づけや、誤った出力に対する患者へのインパクトにより検討する必要がある。

※ ここで、「評価する」とは、試験を実施することのみを指すのではなく、実施された試験結果に基づく考察をもって検討した結果も含まれる。例えば、臨床性能の結果とともに、文献等の内容と合わせて考察することにより、試験を実施しなくても臨床的有用性に関する説明ができる場合、これは「臨床的有用性を評価した」といえる。

次項より、臨床的有用性及び臨床性能の評価の例について述べる。

3.1. 臨床的有用性を評価する試験

臨床的有用性を評価する主たる目的は、その申請品の開発コンセプトが達成できているかを評価することである。臨床的有用性を試験により直接評価する場合は、申請品が臨床導入された状況を模した上で、申請品の導入の価値や効果を直接的に評価する試験として実施する。申請品の医療機器としての価値や効果を直接試験により評価するため、承認申請に対しては非常に説得力のある結果が得られ得る。一方で、実臨床を模することの困難さや、試験に医師等の参加が必要となることから、実施に際しては一定のコストがかかる。次項で述べる臨床性能を評価する試験とともに、当該試験の実施の可否や要否については慎重に検討されたい。

申請品の臨床的有用性を評価する試験として実施される例として、申請品を用いた際の診断成績と申請品を用いずに診断した場合の診断成績を比較し、前者の優越性を評価する試験(以下「読影試験」という。)が挙げられる。この試験では、(理想的には)実臨床を再現した状況において

申請品の CAD 機能の使用が診断成績の向上に寄与することを検証的に評価することができる。

また、申請品の使用方法（例えば、concurrent reader、second reader、等）も試験デザインは影響を受ける。例えば、second reader の場合は、実臨床においても“CAD なしの読影後に CAD の結果を見ながら見直す”という流れで使用される。したがって、このような申請品の臨床的有用性を評価する試験においても、“(1) CAD なしの読影（通常読影）後に(2)CAD の結果を見ながら見直す（申請品併用読影）”（連続評定試験）として読影を実施し、(1)と(2)の診断成績の変化を比較することで、実臨床における申請品の有用性が評価できる。この際、(1)においては可能な限り十分な時間をかけて読影した結果となるように配慮することが望ましい（単に長時間の読影をした（又は複数回の読影をした）ために診断成績が向上した可能性を否定するため）。

一方で、concurrent reader の場合、初見から CAD の結果を参照して診断するものであることから、初見の症例に対してCADを併用することで診断成績が向上することを評価する必要がある。もし、second reader の例で例示した連続評定試験で実施した場合、CAD なし読影時の結果を覚えている状態で CAD ありでの読影をしてしまうため、実際の使用方法とは異なる状態における成績となる。この記憶によるバイアス（メモリーバイアス）の排除、又は影響を受けない試験デザインの設計や、メモリーバイアスがないことを事後的に評価する等の工夫が必要となる。このように、使用方法も考慮した試験デザイン、試験手順の設計が必要である。

その他、申請品がどのような出力をするか（症例に対して解析結果を付す、医用画像に対し解析結果を付す、医用画像中の所見が疑われる個所に解析結果を付す等）によっても試験デザインが変わる。また、複数種類の所見を検出する場合、それらを区別して提示するか、しないのかという点も考慮する必要がある。（これに関連する留意事項は、4 項で述べる。）

また、臨床的有用性を評価する副たる目的として、使用医師が誘引されやすい偽陽性・偽陰性があるかを評価することにある。これにより、情報提供等の措置を講ずることでリスク低減することも期待できる。

試験を構成する各事項に対する留意事項については、3.2 項を参考にされたい。

<発展的な議論>

どのようなデザインの読影試験を実施するべきかは、申請品の開発コンセプトや設計コンセプトに依存する。例えば、申請品の CAD 機能が医用画像から何らかの所見候補を検出する機能だとしても、従来の読影による見落としの抑制、早期治療介入のための早期発見等、診断支援といってもその目的は多様である。この診断支援の目的が果たせるか否かを評価する試験デザイン（どのような症例を評価するか、どのような使用者を想定して評価するか、何を検出対象とするか、何を評価指標とするか、何が果たせれば意義があるといえるか等）は異なる。申請品の臨床的位置づけによっては、単に検出対象がより正しく検出できているかを評価するのではなく、検出した結果に基づき診療を行うことで患者アウトカムをより改善させられたかという試験が必要となる場合もある。申請品の開発コンセプトを整理した上で、その達成をどのように評価する試験デザインが想定できるか、何を評価項目とするか等の検討をする必要がある。

3.2. 臨床性能を評価する試験

臨床性能を評価する主たる目的は、その申請品が医用画像等の入力データに対して、どの程度正しく意図した出力ができるかを評価することである。臨床的有用性の評価で申請品の臨床上の価値や効果（申請品を使用することで診断成績を向上するか等）を評価していることに対し、臨床性能の評価では申請品の性能そのものを評価することになる。主に申請書の性能及び安全性に関する規格欄に記載される情報となる。

CAD 機能の評価において、臨床性能を評価する試験は、以下の2つの方針が想定される。

① 申請品の有効性及び安全性を、臨床的有用性に関する試験で評価する。臨床的有用性が確認された申請品の性能を規定するために臨床性能を試験で確認する。
② 申請品の有効性及び安全性を、臨床性能に関する試験で評価する。（臨床性能を評価する試験で臨床的有用性も評価できる試験を実施する。）

①の場合、臨床的有用性を仮説検証試験等の実施により評価し、この結果より申請品の有効性及び安全性を評価する。この場合、臨床性能に関する評価は、申請品の性能を規定する情報を得ることが主たる目的となる。一方で、臨床的有用性が説明できる臨床性能の評価試験が実施できる場合には、②の評価戦略を検討することができる。例えば、申請品の臨床的位置づけに関連する文献やガイドライン等の背景情報が豊富にあり、臨床的有用性が説明できる妥当なパフォーマンスゴール（以下「PG」という。）が設定できる場合、当該 PG に対する申請品の臨床性能の優越性や非劣性（申請品の位置づけにより適切な方法を選択する。）を検証的に評価することで、臨床的有用性と臨床性能を同一の試験で評価できる可能性がある。この場合 PG の妥当性は重要な論点となる。特に、申請品の臨床的位置づけに完全にマッチした既報やガイドライン等が存在しない可能性も往々にしてあるため、②の方針により適切な評価系が設計できるか慎重に検討されたい。

臨床性能を評価する副たる目的は、稀有な症例に対する申請品の検出性能を評価し、必要に応じ情報提供等の措置を講ずることである。臨床的有用性を評価する試験では、読影医等の参加により、実臨床を模した試験系が重視される。したがって、読影に用いるデータセットについても、症例バランスが実臨床と大きく異なる場合の試験結果への影響を考慮してデータセットを検討することが期待される。一方で、臨床性能を評価する試験のうち申請品単体の性能を評価する場合には、実臨床の症例とのギャップがあったとしても申請品がこれを加味しない設計であるならば症例バランスは気にならない。むしろ、多様なバリエーションに対する成績を検討し、より適切に使用されるような注意喚起等をするための情報収集が重要と考える。

3.3. その他の機能について

原則として、申請品が有する機能はすべて評価する必要がある。CAD 機能以外にも、臨床上意義のある機能や、誤った動作により患者へのリスクが及ぶ可能性がある機能については、その機能

に応じた臨床的有用性や臨床性能に関する試験が必要となる。一方で、データの入出力機能やデータ保存機能等の補助的な機能は、意図したとおりに動作ができることを評価することでよい。

4. 試験設計における留意点

本項では、試験デザインの詳細を設計する上で考慮すべき事項について述べる。

4.1. 試験検体

試験結果は、承認を受けたい検体（製品の version 等を含む。）の評価といえる必要がある。評価検体が最終製品と異なる場合は、評価検体と承認を受けたい検体の差分を明確にしたうえで、試験結果が外挿できる理由を明確にすること。

特に、検証試験に対しては、最終製品の出力の判定閾値も確定された製品が検証項目を達成できることを評価することになる点に留意すること。一般的な診断支援製品の開発において、ROC 等を用いて診断性能を確認し、感度・特異度が最も高くなる仕様や臨床的な位置づけから求められる感度・特異度が期待できる判定閾値を決定する方法はある。このような開発の方針は否定しないが、この時に得られた結果は検証された結果とはいえないので注意すること。

4.2. 人由来のデータの取扱い

一般に、承認申請に添付することを目的として実施される臨床試験は、GCP 省令に準拠して実施することが求められる。一方で、日常診療で得られた医用画像データ等を用いた試験の場合は、「追加的な侵襲・介入を伴わない既存の医用画像データ等を用いた診断用医療機器の性能評価試験の取扱いについて」（令和 3 年 9 月 29 日付薬生機審発 0929 第 1 号通知。以下「0929 通知」という。）に基づく取扱いが検討できる。

0929 通知 記 2. (1)の試験の場合、CAD では主に医用画像のみを用いた試験となる。医用画像のみから、申請品の臨床的有用性を評価するために必要な試験が設計できるか（例えば、試験における正解が医用画像のみから適切に定義できるか。）よく検討すること。これが困難な場合は、0929 通知(2)又は治験としての実施を検討する必要がある。

0929 通知 記 2. (2)の試験の場合、医用画像に紐づく診療データ（確定診断結果等）を用いることができる。その際、医用画像に紐づく診療データがある集団に対する評価となることについて、申請品の評価集団として妥当か注意する必要がある。例えば、医用画像に紐づく生検結果を試験上の正解として扱う場合、生検を行った集団に限定した評価になる。一方で、その申請品が生検に至らない患者に対しても用いられる場合、実臨床の集団と評価集団の差異が生じる。この差異が評価にどのような影響を与えるか、影響を与える場合はどのように対応し妥当な評価系を設計できるか検討する必要がある。

4.3. 評価データセットのバリエーション

申請品の臨床的位置づけ、申請品の性状及び試験の目的に応じ、試験に含めるべき評価デー

タセットのバリエーションについて検討する必要がある。申請品が診療上のどのようなフェーズにおけるどのような集団を対象としているか整理した上で、どのような評価データセットの収集を目指すのか明確にすること。また、試験の目的に応じ、実臨床における症例や有病率等のバランスについても考慮すること。併せて、非臨床的な要素に関するバリエーションについても検討すること。

なお、申請品の解析原理等から検討不要とできる説明ができれば、要素を削減することはできる。

これらを整理したうえで、目指す評価データセットを実現するために、どのような収集計画を実施するか説明すること。

4.4. 正解ラベル

試験における正解ラベルは、申請品目の臨床的位置づけを踏まえて、申請品目の評価が可能となるよう適切な方法により作成する必要がある。申請品の臨床的位置づけにより、どのような正解ラベルを定義するべきか（症例に対する所見の有無、医用画像上の所見の有無、医用画像中の所見位置等）、申請品に仕様と合わせて個別に検討する必要がある。正解ラベルの作成方法の妥当性は、試験結果の解釈並びに申請品目の品質、有効性及び安全性の評価可能性に影響を与える。したがって、試験実施前に正解ラベルの作成方法を明確化し、その妥当性について説明できるように準備する必要がある。

本指標の対象である CADe においては、入力データ中に存在し得る臨床上意義のある所見を検出することを目的とするものが多いと考えられる。従前の CADe では、熟練した医師であれば検出可能な所見に対する検出支援を目的とするもの多い。このような CADe に対しては、本邦の一般的な熟練医により入力データ中から検出されるものが正解ラベルとして定義されていることが説明できるよう作成方法を設計する必要がある。

例えば、正解ラベルを作成する熟練医（以下「正解判定医」という。）の解釈により正解ラベルを決定する場合、以下のような方法が考えられる。

- 属人性を排除するため 3 名以上による正解判定医により解釈をつける。解釈を付与する作業は独立して行う。
- 合議によるバイアスを排除するため、得られた解釈に対する多数決を用いて正解ラベルを決定する。

この際、正解判定医の妥当性（職種（医師、医師以外の医療従事者等）、専門性、経験年数等）、正解ラベル作成のプロセスで使用される選択除外基準等に起因するバイアスの影響も考慮すること。

なお、正解判定医は、読影試験に参加する医師とは独立していなければならない。

正解ラベルを作成するに際し、次のような参照情報を用いることも想定される。医用画像等のみで正解ラベルを付与できる場合もあるが、下に示すような参考情報を用いずに医学的に意義のあるラベルが付与できるかは慎重に検討する必要がある。また、レトロスペクティブな試験を実施する場合、患者に対する追加的な侵襲や介入がない場合であっても、非臨床試験として扱えない場合がある（0929 通知参照）。

- 別のモダリティ等といった他の機器や検査による診断結果
- 確定診断の結果
- フォローアップの画像診断情報

今後の開発によっては、他のモダリティや確定診断結果と合わせて開発することにより、熟練した医師でも入力データのみからでは検出が困難となる所見等を検出する製品が登場する可能性がある。このような製品については、正解ラベルの作成方法に留まらず、製品の臨床的位置づけ及び試験パッケージ全体から改めて検討する必要がある。

4.5. 正答について

試験における正答は試験の成否に影響することから、客観的な判定ができるよう正答の定義を事前に定めておく必要がある。また、試験の目的や申請品の臨床的位置づけ、申請品の出力方法も考慮し、試験における正答の定義する必要がある。

CAD の出力仕様においては、(1)解析症例に対して判定結果を付与するもの、(2)解析画像(動画)に対して結果を付与するもの、(3)解析画像上の検出対象に対して結果を付与するもの等の出力パターンが想定される。(1)及び(2)に対しては、解析対象又は解析画像の正解ラベルと申請品の出力結果の一致を正答とすることが考えられる。一方で(3)については、申請品が提示する位置が正解ラベルの位置と適切な位置関係の下で一致していることも評価できるよう、正答の定義を検討する必要がある。

また、申請品の出力仕様をさらに詳細に分類すると、(3−1)検出対象の領域をトリミングして提示するもの、(3−2)検出対象全体が包含されるバウンディングボックスが提示されるもの、(3−3)検出対象領域の重心や中央を点又は大きさが一定の円等で提示するもの、(3−4)CAD が解析した確信度に基づくヒートマップを重畳して提示するもの等、さらに多様な仕様が想定される。各申請品の仕様等に基づき個別に検討されるべきではあるが、正解ラベルと出力仕様の関係に基づき以下の点を参考に正答の定義を検討されたい。

表 1　評価対象の関係と正答の定義設定における留意点

評価対象の関係 （正解ラベルと申請品の出力）	正答の定義の設定における留意点
面 対 面	一般には、Dice 係数、IoU、Simpson 係数等を用いて定義される。申請品の臨床的位置づけに基づき、臨床上の位置情報の正しさの重要性を考慮しながら、意義のある面の重なりの程度であることが説明できる必要がある。
面 対 点（※）	一般には、点が面内に包含されることで定義される。申請品にとって過度に有利な評価とならないか留意する必要がある。
点（※） 対 点（※）	一般には、点間の距離により定義される。申請品の臨床的位置づけを考慮した際に、意義が説明できる距離であることが説明できる必要がある。

※ 出力仕様に対しては大きさが一定の円等も含む。

同様に、臨床的有用性を評価する試験では、読影医の判定結果と正解ラベルとの一致の程度を評価することが想定される。試験の目的や読影医の判定結果の提示方法も考慮し、表 1 を参考に正答の定義の妥当性を説明すること。

4.6. 評価項目

一般に、試験の主要な目的に直結した最も適切かつ説得力ある証拠を与え得る変数を主要評価項目とし、主要な目的に関連した補足的な測定値又は副次目的に関連した効果の測定値を与え得る変数を副次評価項目として設定する。申請品の位置づけによっては、感度と特異度のいずれも重要と考えられる製品もある。このときは、複合エンドポイントとして設定することも検討する必要がある。

なお、副次評価項目として設定したものであっても、臨床的な有用性に疑義を生じるような結果がある場合は、医療機器の承認を取得する上で大きな問題となる可能性がある。申請品の全体的な成績を考慮して、臨床上の位置づけにおける有用性や安全性が確保されるか、必要な情報提供等の措置は十分か等、検討する必要がある。

また、何を主要評価項目として検証試験を実施するかについては、申請品の仕様も考慮する必要がある。例えば、複数の所見を検出し、かつ検出したものを分類して提示する製品について考える。検出しかつ分類された提示内容を参考に最終的な分類結果の患者へのインパクトが高い場合（例えば、緊急時の使用が想定されかつその分類結果に依存し患者の処置が変わる位置づけで使用されるもの等）、各分類結果が適切に検出できることを検証することについても慎重に検討する必要がある。

4.6.1. 有効性を評価する指標

申請品の臨床的位置づけと試験の目的から、評価したい事項が説明できる指標を設定する必要がある。CAD 機能の評価に用いられる指標の例を以下に示す。なお、申請品の評価としての妥当性が説明できれば、下に提示した以外の指標であってもよい。

(1) AUC、FOM

診断支援性能の評価に用いられる指標として、ROC を用いた AUC による評価が良く知られている。ただし、検出したものの位置情報の正しさも併せて評価する場合は AUC では評価できないため、FROC を用いた FOM による評価を行うことが多い。

AUC や FOM は、二つの診断法の診断性能を比較する際には有用である。例えば、読影試験において CADを用いない医師群とCADを用いた医師群の読影成績を比較する場合が想定される。ただし、申請品の AUC や FOM のみで申請品自体の性能を評価する場合には、臨床上有用といえる性能を有しているかを説明することが困難となる場合が多い。この場合は、(2)に示す感度・特異度等から説明することがよい。

また、二つの診断法の診断性能を比較する場合であっても、2 つのカーブが交差する場合等には、解釈が困難となる可能性がある。カーブの傾向から診断性能の特性を考察することはできるが、2 つの診断法の優劣を比較するという観点では、解釈が難しくなることがある。事前に実施される予備試験等の結果も踏まえながら、AUC や FOM で評価が可能であるか検討する必要がある。

(2) 感度・特異度

感度・特異度は、正解ラベルとしての陽性・陰性のそれぞれを適切に判定できた割合を表す指標である。従来の診断法の感度・特異度が文献等により示されている場合には、これと比較することで申請品の臨床的な有用性を直接的に説明することが比較的容易な指標である。

感度・特異度で評価する場合は、その試験において何を真陽性・真陰性と定義したか、どのような状態を正答と定義したか等により算出された値の意味が異なる。加えて、解析単位（症例単位、所見単位等）も算出された値の意味を変える。したがって、試験における感度・特異度の定義は明確にしておく必要がある。また、従来の診断法等と比較する場合は、比較対象となる感度・特異度の定義が申請品を評価しようとする感度・特異度の定義と一致しているか、比較可能であるか等についても検討すること。

(3) 正診率、陽性的中率・陰性的中率

正診率は、評価集団全体に対して真陽性・真陰性を適切に判定できた割合を示す指標である。また、陽性的中率・陰性的中率は、評価検体の陽性判定・陰性判定のそれぞれが適切に判定できた割合を示す指標である。これらの指標は、評価集団の陽性率により影響を受けることに留意する必要がある。他の診断法等と比較する場合は、比較対象との陽性率の差異がないか確認するべきである。また、これらの指標により申請品の性能を規定する場合も、解析対象の陽性率に関する情

報と合わせて規定する必要がある。加えて、感度・特異度と同様に、これらの指標の定義（解析単位を含む。）は明確にする必要がある。

表 2 感度・特異度、正診率、陽性的中率・陰性的中率

		正解ラベル	
		陽性	陰性
申請品の解析結果	陽性	真陽性（TP）	偽陽性（FP）
	陰性	偽陰性（FN）	真陰性（TN）

感度 ： $\frac{TP}{TP+FN}$

特異度 ： $\frac{TN}{FP+TN}$

正診率 ： $\frac{TP+TN}{TP+FP+TN+FN}$

陽性的中率 ： $\frac{TP}{TP+FP}$

陰性的中率 ： $\frac{TN}{TN+FN}$

4.6.2. 暗転と好転

読影試験では、CAD を用いずに診断した結果と、CAD を用いて診断した結果を比較し、CAD を使用することの有用性を評価する。FROC 等により全体として診断成績が向上しているかは評価できるが、一般には真陽性の検出増加とともに偽陽性も増加する（感度は向上するが特異度が低下する）。CAD を用いることで判断が改善した事例（好転）と、判断が悪化した事例（暗転）の内容について考察し、特に暗転についてはその傾向を検討の上、注意喚起等の情報提供の要否についても検討する必要がある。

4.6.3. サブグループ解析

申請品の性能や限界を使用する医師等に正しく理解して使用してもらうための情報提供ができるように、適宜サブグループ解析を実施すること。以下にサブグループ解析を行う切り口の例を示す。申請品の特性に応じ適宜取捨選択、追加又は調整を検討すること。

- 撮影モダリティのベンダー別
- 画像処理条件、撮影条件別
- 患者背景別
- 所見のサイズ、性状等別
- 読影試験における読影医等の診療科、経験年数別

4.7. 読影試験における読影医

読影試験は、実臨床を可能な限り再現したうえで、申請品を導入することの価値を評価することが目的である。したがって、参加する読影医（申請品によって、読影試験の参加者は医師とは限らない。）も、申請品の使用対象者が代表されるように選定されることが望ましい。選定に際しては、診療科、経験・技量、専門医の有無等を検討すること。

申請品の開発コンセプトによっては、診断に不慣れな医師の診断成績を向上させることで、診断成績を均てん化させることを目指す製品もある。この場合、その申請品の主たる使用者は診断に不慣れな医師となる。そのような製品であっても、経験豊富な医師の使用を制限しないことがある。この場合の読影試験の対象は、経験豊富な医師を含む使用し得る読影医群を検討すべきである。一方で、（医師の経験による制限はしないものの）経験豊富な医師の診断成績が十分高く診断支援が実質不要であることから、申請品が経験豊富な医師に使用されることがほとんど想定されない場合には、読影試験の対象医師を限定した群に対する評価をピボタルとすることも検討できる。この場合であっても、経験豊富な医師が申請品を使用した場合に診断結果を悪化させることがない等について考察し、適切な措置を講ずること。

4.8. その他

試験結果（4.6.2、4.6.3 項を中心に参照されたい。）を踏まえて、申請品の性能や限界等を正しく使用者に理解してもらうための情報提供の要否について検討すること。例えば以下のような内容が想定される。

- 対象外所見に対する偽陽性・偽陰性
- 他の臓器との重なりがある部分の偽陽性・偽陰性

5. 機械学習を用いた製品に対する追加的留意事項

昨今、深層学習等の機械学習を用いた医療機器（以下「MLMD」という。）の開発は盛んであり、MLMD たる CADe の開発も主流になりつつある。機械学習を用いて開発された判別器は、入力に対して出力が非線形に変化する特性（入力の変化に対して出力結果が大きく変わり得る性質）があり、複雑なパターン認識問題を解くための重要な性質を有している。一方で、その特性ゆえに、未知データに対する振る舞いが予想困難である場合も多く、予想もしない誤りが発生する可能性があること、過学習により性能が低下すること等の特徴もある。また、機械学習、特に深層学習は、その特性上、ニューラルネットワークによる判断の過程が解釈困難であるため、通常の医療機器にてその性能確保のために承認事項として規定することが求められる原理（実装する検出アルゴリズム等）や設計仕様等のみをもって、アウトプットの品質が確保されていることを説明することは難しい。

したがって、現時点においては、構築されたネットワークの内容を詳細に精査するのではなく、臨床的な位置づけに基づきインプットに対して適切なアウトプットが得られているかを確認することに重点を置き評価を確認している。すなわち、機械学習を用いた開発の有無に限らず、“医療機器”と

しての適切性を評価している。以上の考えに基づき、申請品の有効性及び安全性を示すためには、上述した通り、医学的かつ統計学的に妥当な方法となるよう配慮し、申請品の性能等を検証することが求められる。

そのうえで、機械学習を用いて開発された CADe に追加的に考慮すべき事項についてまとめる。

5.1. 評価データセットに対する留意事項

多くの申請品は、本邦の任意の施設で使用されることを前提に開発され、また承認を得ることを想定している。したがって、本邦の任意の施設で、申請品の品質、有効性及び安全性が確保されることを評価する必要がある。すなわち、試験成績の一般化可能性が審査上の論点となる。

以下では、MLMD である申請品の CAD 機能の評価について、当該 CAD 機能が有するバイアスへの配慮と、帰納的な設計であるために考慮すべき試験バリエーションへの配慮について述べる。

5.1.1. 学習データとの関係

MLMD である申請品の CAD 機能の解析結果に対する試験成績は、評価データセットがもつバイアスとモデルが有するバイアスの影響を受けている。評価データが有するバイアスの処理は、申請品が MLMD であるか否かによらず、通常の医療機器の評価として考慮されるべきである（一般にはデータ収集施設、選択除外基準により管理される。）。

一方で、MLMD の場合は、モデルが有するバイアスについても配慮する必要がある。

例えば開発データと評価データが同一施設で収集されている場合、その施設を受診する患者、撮影機種、撮影法、診療の方針・体制等、施設が持つバイアスの影響を強く受けた試験成績となっている。したがって、他の施設のデータで同様の試験を実施した際の結果と乖離する可能性がある。すなわち、試験結果の一般化可能性の考察が困難となる可能性がある。

別の例として、開発データの教師と評価データの正解ラベルの作成法に作成医の主観的な要素が多く含まれる場合、当該作成医の判断を模倣できることに対する評価としては成立するが、一般に医師が検出すべきとする対象を適切に検出できる評価となっていない可能性がある。これも、試験結果の一般化可能性の考察が困難となる可能性がある。

このように、同一の臨床的位置づけである CAD であっても、開発データとの関係により評価データとしての適切性に対する判断が異なる可能性がある。したがって、表 3 などを参考に、開発データと評価データの関係については整理して説明する必要がある。

表 3 開発データと評価データの関係の整理の例

	学習データ	評価データ
データ収集場所	○○大学病院 ○○病院	
患者背景	健診患者/ I 次スクリーニングで要検査となった患者	
撮影機種、撮像条件		
教師データ/正解ラベルの作成方法	陽性：日本の○○専門医○名が画像情報のみを用いて読影し○と思われる部位を○と定義した。 陰性：・・・	
学習データとの関係		学習データを一部含むデータセット / 学習データと同一施設で収集された別のデータセット / 学習データとは異なる施設で収集された別のデータセット 等

バイアスの影響について定量的に考察することは困難であることから、バイアスの影響を受けにくいと考えられる評価データセットを収集し評価することが望ましい。例えば以下のような事項を整理する。

- 学習データと評価データの収集施設を重複させない。
- 教師データと正解ラベル作成者を重複させない。

なお、これらの措置が難しい場合は、これにより生じ得るバイアスの有無や影響を考察の上、試験結果の一般化可能性を説明すること。

5.1.2. バリエーションに対する配慮

申請品が開発者により演繹的に機能設計されている場合、その申請品が影響を受けない因子を特定すること（又はその妥当性を説明すること）が比較的容易であることから、評価データに含める必要がない要素についても比較的容易に説明できる。一方で、MLMD の場合、帰納的に機能設計されていることから、その申請品が影響を受けない因子を特定すること（又はその妥当性を説明すること）が困難である。したがって、申請品に入力され得る多様な因子を想定し評価データセットのバリエーションに含める必要がある。以下に、因子の例を記載する。

（1）臨床的な因子の例

- 対象疾患の重症度
- 検出対象の種類
- 検出対象の数
- 検出対象の解剖学的な位置
- 患者の年齢、性別

(2) 非臨床的な因子の例
- 撮像機種
- 撮像条件
- 撮像パラメータ

なお、開発時の学習データセットの多様性から、申請品が影響を受けない因子について考察されることがある（例えば、多様な機種から取得したデータで学習しているため、機種差の影響がない等の考察）が、一般には学習データに含まれている説明だけをもって、それの因子に対する解析性能の安定性が確保されていることを説明することはできない。原則としては、上の因子の例を参考に、多様なバリエーションが含まれるよう評価データの収集計画を検討されたい。

以上