21-Dec-2025

HPVM Technical White paper ＃2 (HiroPharmaConsulting® Co., Ltd)

HPVM（HiroPharma Validation Method）における AI 評価結果と人間系評価結果の統計的信頼性保証手法

1. 本資料の目的と位置づけ

本資料は、HPVM（HiroPharma Validation Method）を医薬品製造販売業者（MAH）が適用する際に選択可能な、**AI 評価結果の信頼性保証に関する三つの適用法域オプション**を整理し、それぞれの技術的考え方、統計的取扱い、合否判定ロジックおよび実務上の留意点を明確にすることを目的とする。

HPVM は、Pharmacovigilance（PV）業務において AI が生成する評価結果について、一定頻度で定期的に人間系評価と比較・点検を行い、MAH として合理的な説明責任を果たすための「統計的基本系手法」を提供するものである。

どの適用法域オプションを選択するかは、**各 MAH のリスク評価、運用方針、規制対応方針に基づき決定される事項**である。

2. 共通前提および用語定義

2.1 母集団（Used cases）

- 対象期間内に PV 業務で実際に処理された実症例データ群を指す。
- 対象期間、対象製品、対象データソース、除外条件（重複症例、情報不足症例等）は、評価計画書または SOP に明記する。

2.2 評価単位および評価要素

- 評価単位は「症例単位」とする。
- 各症例について、以下の 4 要素を評価する（例示）。
    1. 因果関係
    2. 重篤性
    3. 既知／未知
    4. 規制当局への報告要否

2.3 人間系評価

- MAH が定める SOP および教育を受けた評価者による症例評価を指す。
- 本資料では、AI が参照すべき基準評価（Reference）として位置づける。

2.4 合否判定および誤り分類

- 各評価要素の一致／不一致を集計し、症例単位で総合判定（OK／NG）を行う。
- 併せて、誤りの影響度に基づき以下の分類を行うことを推奨する。
    - Critical：規制報告・患者安全に重大な影響を与える誤り
    - Major：業務品質に影響するが即時規制影響は限定的な誤り
    - Minor：影響が軽微な誤り

3. 適用法域オプション（a）「簡易版」

3.1 目的と考え方

本オプションは、**定期的な監査・点検として AI の挙動を人間系評価との一致率で確認し、トレンド監視およびドリフト検知を行うこと**を目的とする。

本方式は、**母集団全体の性能を統計的に保証することを主目的とはせず**、運用上の合理的説明を可能とする点検手法である。

3.2 標本設計

- 母集団サイズを N とし、ランダムサンプリング数 n を以下で定義する。

$$n = \sqrt{N} + 1$$

- 例：N＝10,000 の場合、n＝101
- 乱数生成方法、抽出手順、抽出ログは記録として保存する。

3.3 合否判定方法

- 指標：AI 評価結果と人間系評価結果の総合一致率
- 計算式：

$$\text{一致率} = \text{一致（OK）症例数} \div n$$

- 合格基準（例）：一致率が 99%以上であること
- 併設推奨条件：Critical 誤りが 0 件であること

3.4 統計的取扱い

- 本方式では、二項検定等の統計的検定は必須としない。
- ただし、結果の表現は「当該サンプリングにおける観測結果」であることを明記する。

4. 適用法域オプション（b）「統計保証版」

4.1 目的と考え方

本オプションは、**AI 評価結果と人間系評価結果の一致率について、母集団における真の一致率が所定水準（例：99%）以上であることを統計的に保証する**ことを目的とする。

4.2 基本設計思想（二段構え）

本方式では、以下の二つの観点を組み合わせた合否判定を行う。

(1) 全体一致率の統計保証

- 成功＝一致（OK）、失敗＝不一致（NG）とする二項モデルを用いる。
- 一致率の**信頼下限**が所定水準（例：99%）以上であることを合格条件とする。
- 点推定値（単純な一致率）のみで判定しない点が重要である。

(2) Critical 誤りゼロ保証

- Critical 誤りの発生率について、
  観測件数および信頼上限を用いて、許容水準以下であることを確認する。
- 例：Critical 誤りが 0 件であり、その発生率の信頼上限が 1%以下であること。

4.3 標本数設計の考え方

- 標本数は、求める信頼水準、許容する不一致件数および Critical 誤り許容率に応じて設

計する。

- √N＋1 方式よりも多い標本数（数百件規模）が必要となる場合がある。

4.4 合否判定の表現例

- 「全体一致率の信頼下限が 99%以上であり、かつ Critical 誤り率の信頼上限が許容基準以下であるため、AI 評価機能は所定の信頼性基準を満たすと判断した。」

---

5. 適用法域オプション（c）「（神）評価含版（Ground Truth 導入）」

5.1 目的と考え方

本オプションは、**人間系評価自体にも誤りが含まれる可能性を前提とし、運用上の真（Operational Ground Truth）を構築した上で AI 評価の信頼性を評価する**ことを目的とする。

5.2 Ground Truth の構築方法（推奨）

1. 複数の人間評価者による独立評価を実施する。
2. 評価結果が不一致となった症例について、アジュディケータ（上級評価者、医学的責任者等）が最終判定を行う。
3. この最終判定を Ground Truth として確定する。

5.3 合否判定指標

- AI 評価結果と Ground Truth の一致率（要素別および総合）
- 人間系評価結果と Ground Truth の一致率（参考指標）
- Critical 誤りの発生状況（Ground Truth 基準）

5.4 記録および成果物

- 評価計画書
- 症例別評価記録（AI、人間評価者、アジュディケータ、最終判定理由）
- 集計レポート（統計指標、前回比較、CAPA 要否）
- 教育記録および監査証跡

---

6. 三つのオプションの使い分けに関する考察

- オプション（a）：運用負荷が低く、定期点検およびトレンド監視に適する。
- オプション（b）：導入時、重大変更時、規制当局説明時における統計的裏付けとして有効。
- オプション（c）：人間系の誤りも含めた最も厳密な評価が可能であり、議論が生じた場合の説明力が高い。

---

7. まとめ

HPVM は、**AI 評価機能の信頼性を一律の方法で縛るものではなく、MAH が自らの責任において適切な適用法域を選択できる枠組み**を提供するものである。本資料に示した三つのオプションは、その合理的選択と説明を支援するための補足技術的整理である。

規制当局・監査向け 想定 Q&A

HPVM（HiroPharma Validation Method）AI 評価結果の信頼性保証に関する考え方

Q1. HPVM とは何を目的とした手法ですか？

A1.

HPVM は、Pharmacovigilance 業務において AI が生成する症例評価結果について、**一定頻度で定期的に人間系評価と比較・点検を行い、MAH として合理的に信頼性を説明するための基本系手法**です。

AI そのものを自動的に「正解」とみなすものではなく、**人間系評価を基準とした監視・保証の枠組み**を提供します。

Q2. HPVM は AI の「バリデーション」を目的とした手法ですか？

A2.

いいえ。HPVM は**開発時の一回限りのバリデーション手法ではありません。**

本番運用後において、**AI 評価結果が継続的に人間系評価と整合しているかを確認するための運用時信頼性保証手法**です。

これは従来の CSV では十分にカバーできなかった領域を補完する考え方です。

Q3. AI 評価の正解・不正解はどのように定義していますか？

A3.

HPVM では、原則として MAH が定めた SOP に基づく人間系評価を参照基準（Reference）とします。

したがって評価指標は「AI と人間系評価の一致／不一致」であり、**AI が AI 自身を評価することはありません。**

Q4. 人間系評価も誤る可能性があるのではありませんか？

A4.

はい、その可能性はあります。

HPVM では、この点を否定せず、**MAH の選択により以下の 3 段階の適用法域を用意**しています。

- 人間系評価を基準とする簡易運用
- 統計的保証を付加した運用
- 人間系評価の誤りも織り込む Ground Truth（アジュディケーション）導入運用

どのレベルまで求めるかは、**MAH のリスク評価および運用方針に基づいて決定**されます。

Q5. なぜ √N＋1 というサンプリング方法を採用しているのですか？

A5.

√N＋1 は、**監査・点検として合理的かつ実務的なランダムサンプリング数**として製造現場で用いられてきた一般的に合理性があるとされてる手法です。

HPVM では、**頻度高く定期的に信頼性保証を取ることを前提とした基本手法**に、この手法を推奨方

式例として記載しています。
これは「統計保証」を補完するものであり、**目的に応じて統計保証版へ拡張可能**な設計としています。

Q6. 101 件中 100 件一致＝99%であれば、統計的に 99%保証されたと言えますか？

A6.

いいえ。
101 件中 100 件一致という結果は、**当該サンプルにおける観測一致率が 99%であった**ことを示しますが、**母集団全体の一致率が 99%以上であることを統計的に保証するものではありません。**
統計保証を行う場合は、信頼区間や検定を用いた別途の設計（統計保証版）を採用します。

Q7. 統計保証版では、どのような保証を行うのですか？

A7.

統計保証版では、以下の二段構えで評価します。

1. **全体一致率**について、信頼区間の下限が所定水準（例：99%）以上であること
2. **Critical 誤り**について、発生率の信頼上限が許容水準以下であること

これにより、単なる点推定ではなく、**母集団性能に対する合理的な統計保証**を行います。

Q8. なぜ Critical 誤りを別枠で評価するのですか？

A8.

Pharmacovigilance 業務では、すべての誤りが同じリスクを持つわけではありません。
特に、**重篤性や報告要否の誤りは規制対応や患者安全に重大な影響**を与える可能性があります。
そのため HPVM では、**全体一致率とは独立して Critical 誤りを管理**するリスクベースの考え方を採用しています。

Q9. Ground Truth（神評価）とは何ですか？

A9.

HPVM における Ground Truth とは、自然に存在する絶対的な正解ではなく、運用上合理的に定義された「最終評価」を指します。通常は、複数の人間評価者による独立評価と、不一致時のアジュディケーション（裁定）によって構築されます。

Q10. Ground Truth を導入すると、AI は人間より優れていると主張できますか？

A10.

HPVM の目的は、**AI が人間より優れていることを主張することではありません。**
AI 評価結果が、**MAH として許容可能な水準で安定的かつ再現性をもって運用されていること**を説明することが目的です。必要に応じて、人間系評価との比較結果を参考情報として提示することはあります。

Q11. HPVM は特定の AI 技術やベンダーに依存しますか？

A11.

いいえ。

HPVM は、**特定のアルゴリズム、AI モデル、PV システムベンダーに依存しない**中立的な方法論です。
評価対象は「AI の出力結果」であり、その内部構造や実装方式は問いません。

Q12. HPVM の適用頻度はどの程度を想定していますか？

A12.

適用頻度は、**MAH が自らのリスク評価に基づいて決定**します。一般的には、

- 定期的（例：四半期、半年、年次）
- 重大変更後（AI モデル更新、SOP 改訂等）

に実施されることを想定しています。

Q13. HPVM の結果が不合格となった場合、どのように対応しますか？

A13.

HPVM は合否判定だけでなく、**是正・予防措置（CAPA）につなげるための監視手法**です。
不合格の場合には、

- 誤りの傾向分析
- 人間系評価との乖離要因分析
- AI 設定・運用・教育・手順の見直し

等を行い、再評価します。

Q14. 規制当局として、HPVM により何が確認できますか？

A14.

HPVM により、以下の点が確認可能となります。

- MAH が AI 評価結果を**ブラックボックスとして放置していない**こと
- AI 評価結果が**人間系評価と比較・監視されている**こと
- 問題発生時に**説明可能な記録と是正プロセス**が存在すること

これは、GVP における責任体制・品質管理の観点から重要な要素です。

Q15. HPVM は将来の AI 規制動向にも対応可能ですか？

A15.

はい。
HPVM は、**特定の規制文言に依存せず、「人間系による継続的監視と説明責任」という原則に基づく方法論**です。
そのため、AI に関する規制・ガイダンスが進展した場合でも、**基本構造を維持したまま適用可能**です。

「参考資料」 統計的な解説（補足参考資料）

サンプルシナリオ（全 10,000 例から√N＋1 ランダムサンプリングした 101 例を用いた、AI と人間系の評価結果の比較・検証）に基づき、AI の信頼性を評価するための統計的手法と指標を提示す。
ここでは、人間（医師・PV 専門家）の評価を「正（正解）」とした場合の、**AI の「正解率（一致率）」とその「信頼区間」**を算出する方式としている。

「症例評価サンプルシナリオ」

1) ある個別有害事象報告の Intake & Triage 評価を実施した場合、「AI での評価結果：[A]」、と「医師、PV 専門化の人間系の評価結果：[B]」がある。
2) この場合、リアル実症例の母数、10,000 例から、ランダムサンプリングした√N＋1 方式でピックアップした Used Case をテスト症例とした、2 つの評価結果が、全部で１０１組ある。
3) [A]:1 から [A]:101」と「[B]:1 から [B]101」を比較し、一致しているのは「OK」とする。
4) [A]:1 から [A]:101 と [B]:1 から [B]101 を比較し、一致していないものは「NG」とする。
5) この場合、人間系評価結果を「神評価：正」として取り扱う。
6) この１０１組の２項比較結果から、AI 評価が、どの程度の信頼性・信用性があるか、を統計的な手法で「信頼限界」と「信頼性評価指数：数値で表現」を提示する。
7) 統計的な手法名とその解説を提示する。

以下は具体的な指標と計算手法の解説：

1. 信頼性評価指数：数値での表現

AI の性能を評価する最も基本的な指標は、人間と判断が一致した割合（正解率）である。

- **指標名**：**正解率（Accuracy）** または **一致率（Agreement Rate）**
- **計算式**： 正解率 $(\hat{p}) = \frac{\text{OK の数 (一致した数)}}{101}$

例えば、101 例中 95 例で一致（OK）した場合、正解率は $95 \div 101 \approx 0.941$つまり **94.1%** となる。これが「点推定値」としての信頼性評価指数である。

---

2. 信頼限界（信頼区間）の算出

101 例というサンプル（標本）の結果から、母集団全体（10,000 例）における AI の真の実力を推測するために、「信頼区間」を求める。通常は **95%信頼区間**を用いる。

- **指標名**：**母比率の 95%信頼区間（95% Confidence Interval for Population Proportion）**
- **計算式（ウォルド法による近似）**： 信頼限界 $= \hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - $\hat{p}$：上記の正解率（OK 数 ÷101）
  - n：サンプルサイズ（101）

- 1.96：信頼係数（95%の確率に対応する標準正規分布の値）

この計算により、「下限値 ～ 上限値」という幅（レンジ）が算出されます。「AI の真の精度は、95%の確率でこの範囲内に収まる」ということを意味する。

具体的な計算例（仮に OK が 95 例だった場合）

- $\hat{p} = 0.941$
- $n = 101$
- 標準誤差 $= \sqrt{\frac{0.941\times(1-0.941)}{101}} \approx \sqrt{\frac{0.0555}{101}} \approx 0.0234$
- 誤差範囲 $= 1.96 \times 0.0234 \approx 0.046(4.6\%)$

結果：

- 信頼性評価指数：**94.1%**
- 信頼限界（95%信頼区間）：**89.5% ～ 98.7%**
  - $(0.941 - 0.046 \sim 0.941 + 0.046)$

---

3. 使用した統計的手法名とその解説

この分析で使用する統計手法は「母比率の区間推定（Interval Estimation of Population Proportion）」である。

解説

1) **背景**：今回のケースは、結果が「OK（一致）」か「NG（不一致）」かの 2 値（二項分布）になるデータです。全数（10,000例）を調査するのは大変なため、√N(10,000) + 1=101 例をサンプリングしてテストを行っている。
2) **目的**：手元にある「101 例の結果（標本比率）」から、バックグラウンドにある「10,000 例全体での AI の真の正解率（母比率）」がどの程度になるかを推測する。
3) **手法の選定理由**：
   - **正規分布近似（ウォルド法）**：サンプルサイズ $N = 101$は統計的に十分な大きさ（大標本）とみなせるため、二項分布を正規分布に近似させて計算するこの手法が一般的かつ実務的である。
   - **信頼水準 95%**：統計的検定において標準的に用いられる基準となる。「同じ調査を 100 回行えば、95 回はこの範囲に真の値が含まれる」という精度を示す。

補足：より厳密な手法が必要な場合

もし「OK」の数が極端に少ない（または極端に多くて 100%に近い）場合は、近似計算の精度が落ちるため、「ウィルソン・スコア区間（Wilson Score Interval）」や「クロッパー・ピアソン法（Clopper-Pearson interval）」という手法を用いることが推奨されるが、P Vビジネス上の一般的な精度検証であれば、上記の標準的な計算式で十分許容される。

以上