

HPVM Technical White paper #2

(HiroPharmaConsulting® Co., Ltd)

Statistical Reliability Assurance Method for AI Evaluation Results and Human Evaluation Results in HPVM (HiroPharma Validation Method)

1. Purpose and positioning of this document

The purpose of this document is to list **three applicable jurisdiction options for reliability assurance of AI evaluation results** that MAHs (MAH) can select when applying the HPVM (HiroPharma Validation Method), and to clarify the technical concept, statistical treatment, acceptance/rejection logic, and practical considerations for each.

The HPVM provides a “basic statistical method” for MAHs to perform reasonable accountability by regularly comparing and checking evaluation results generated by AI with human evaluation at a certain frequency in pharmacovigilance (PV) activities.

The choice of applicable jurisdiction option is **determined based on the risk assessment, operational policy, and regulatory response policy of each MAH.**

2. Common assumptions and definitions of terms

2.1 Population (Used cases)

- This refers to the actual case data group processed in PV operations during the target period.
- The target period, target products, target data sources, and exclusion conditions (Multiple cases, cases with insufficient information, etc.) shall be specified in the evaluation plan or SOP.

2.2 Evaluation unit and evaluation elements

- The evaluation unit shall be case unit.
- For each case, the following 4 elements are evaluated (example):
 1. Causality
 2. Seriousness
 3. Expectedness/Labelled & Unlabelled
 4. Regulatory reporting required

2.3 Human evaluation

- This refers to SOPs specified by MAH and case evaluation by trained evaluators.
- In this document, it is positioned as a reference evaluation that AI should refer to.

2.4 Pass/fail judgment and error classification

- Concordance/discordance of each evaluation element is totaled, and an overall judgment (OK: Pass /NG: Fail) is made for each case.
 - In addition, the following classification is recommended based on the degree of impact of errors.
 - Critical: Errors that have a serious impact on regulatory reporting and patient safety
 - Major: Errors that affect work quality but have limited immediate regulatory impact
 - Minor: Errors that have a minor impact
-

3. Applicable jurisdiction Option (a) “Simplified version”

3.1 Purpose and concept

The purpose of this option is to **confirm the behavior of AI by the agreement rate with the human evaluation as a periodic audit and inspection, and to perform trend monitoring and drift detection.**

The **main purpose of this method is not to statistically guarantee the performance of the whole population**, and it is an inspection method that enables rational explanation in operation.

3.2 Sample design

- Let the population size be N , and the random sampling number n is defined as follows.

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

- Example: When $N=10,000$, $n=101$
- The random number generation method, extraction procedure, and extraction log are stored as records.

3.3 Pass/fail determination method

- Index: Overall concordance rate between AI evaluation results and human evaluation results
- Formula:
Concordance rate = number of concordant (OK: Pass) cases \div n
- Acceptance criteria (example): Concordance rate should be 99% or more
- Recommendation condition: 0 critical errors

3.4 Statistical handling

- This method does not require statistical tests such as binomial tests.
- However, it should be specified that the expression of results is "observed results in the sampling concerned."

4. Applicable jurisdiction option (b) "Statistical assurance version"

4.1 Purpose and Concept

The purpose of this option is to **statistically guarantee that the true agreement rate in the population is at least a predetermined level (e.g. 99%) regarding the agreement rate between the AI evaluation results and the human evaluation results.**

4.2 Basic design concept (two-stage approach)

In this method, pass/fail judgment is performed by combining the following two perspectives:

(1) Statistical guarantee of overall agreement rate

- A binomial model is used where success = agreement (OK: Pass) and failure = disagreement (NG: Fail).
- The acceptance condition is that the **lower confidence limit** of agreement rate is above a predetermined level (e.g., 99%).
- It is important that judgment is not made only by a point estimate (simple agreement rate).

(2) Zero Critical Error Guarantee

- Regarding the incidence of critical errors, Using the number of observations and the upper confidence limit, confirm that it is below the acceptable level.
- Example: The number of critical errors is 0 and the upper confidence limit of the occurrence rate is 1% or less.

4.3 Concept of sample size design

- The sample size should be designed according to the required confidence level, the acceptable number of discrepancies, and the critical error tolerance rate.
- A larger sample size (hundreds of samples) than the $\sqrt{N}+1$ method may be required.

4.4 Expression example of pass/fail judgment

- "Since the lower confidence limit of the overall agreement rate is 99% or more

standard; upper confidence limit of the critical error rate is below the acceptable standard, the AI evaluation function was judged to satisfy the prescribed reliability standard."

5. Applicable jurisdiction option (c) "(God) evaluation inclusion (introduction of ground truth)"

5.1 Objectives and concepts

The purpose of this option is to **evaluate the reliability of AI evaluation after constructing operational ground truth on the premise that human evaluation itself may contain errors.**

5.2 Method of constructing ground truth (recommended)

1. Independent evaluations by multiple human evaluators are conducted.
2. The adjudicator (Senior evaluators, medical officers, etc.) makes a final judgment for cases with discordant evaluation results.
3. This final judgment is confirmed as the ground truth.

5.3 Pass/Fail Judgment Index

- Concordance rate between AI evaluation results and ground truth (by element and overall)
- Concordance rate between human evaluation results and ground truth (reference index)
- Critical Occurrence of errors (ground truth criteria)

5.4 Records and deliverables

- Evaluation plan
 - Case-by-case evaluation records (AI, human evaluator, adjudicator, final decision reason)
 - Summary reports (Statistical indicators, previous comparisons, CAPA required or not)
 - Education records and audit trails
-

6. Considerations for the use of the three options

- Option (a): Low operational load, suitable for periodic inspections and trend monitoring.
 - Option (b): Useful for statistical support during implementation, major changes, and regulatory explanations.
 - Option (c): Enables the most rigorous evaluation, including human errors, and provides greater explanatory power in the event of controversy.
-

7. Summary

HPVM does not constrain the reliability of the AI assessment function in a uniform manner but provides a framework for MAHs to select appropriate jurisdictions on their own responsibility. The three options presented in this document are supplementary technical arrangements to support rational selection and explanation.

Anticipated Q & A for regulators and auditors

HPVM (HiroPharma Validation Method) Approach to reliability assurance of AI assessment results

Q1. What is the purpose of HPVM?

A1.

HPVM is a basic system method for rationally explaining the reliability of case assessment results generated by AI in pharmacovigilance operations by **comparing and checking them with human assessment at a certain frequency and periodically as MAH.**

AI itself is not automatically regarded as “correct” but provides a **framework for monitoring and assurance based on human assessment.**

Q2. Is HPVM a method for AI "validation"?

A2.

No. HPVM is **not a one-time validation method during development.**

It is an **operational reliability assurance method to confirm whether the AI evaluation results are continuously consistent with the human evaluation** after the actual operation.

This idea complements areas that were not sufficiently covered by conventional CSV.

Q3. How do you define correct and incorrect AI evaluation?

A3.

In principle, HPVM uses the human evaluation based on the SOP established by MAH as the reference standard.

Therefore, the evaluation index is “agreement/disagreement between AI and human evaluation,” and **AI does not evaluate itself.**

Q4. Isn't there a possibility that human evaluation may also be wrong?

A4.

Yes, there is a possibility.

HPVM does not deny this point and **provides the following three levels of application jurisdictions depending on the choice of MAH.**

- Simplified operation based on human evaluation
- Operation with statistical assurance
- Operation with ground truth (adjudication) incorporating errors in human evaluation

The level to be determined is **determined based on the risk assessment and operation policy of MAH.**

Q5. Why is the $\sqrt{N}+1$ sampling method used?

A5.

$\sqrt{N}+1$ is a generally reasonable method that has been used at manufacturing sites as a **reasonable and practical random sampling number for audits and inspections.**

HPVM describes this method as an example of the recommended method in the **basic method that assumes that reliability assurance is performed frequently and regularly.**

This method complements the statistical assurance and is designed to be **expandable to the statistical assurance version according to the purpose.**

Q6. If 100 out of 101 matches =99%, can it be said that 99% is statistically guaranteed?

A6.

No.

A result of 100 out of 101 indicates that the **observed concordance rate in the sample was**

99%, but it **does not statistically guarantee that the overall population concordance rate is 99% or higher.**

To provide statistical guarantees, a separate design using confidence intervals and tests (statistical guarantee version) is used.

Q7. What kind of guarantees are provided by the statistical guaranteed version?

A7.

The statistical guaranteed version evaluates on the following two levels:

1. For the **overall agreement rate**, the lower limit of the confidence interval must be greater than or equal to a predetermined level (for example, 99%).
2. For the **critical error**, the upper confidence limit of the occurrence rate must be less than or equal to an acceptable level.

This provides **reasonable statistical guarantees for population performance** rather than mere point estimates.

Q8. Why are critical errors evaluated separately?

A8.

In pharmacovigilance, not all errors carry the same risk.

In particular, **errors in seriousness and reporting requirements can have a significant impact on regulatory responses and patient safety.**

Therefore, HPVM adopts a risk-based approach to **managing critical errors independently of the overall agreement rate.**

Q9. What is the ground truth?

A9.

The ground truth in HPVM is not an absolute right answer that exists naturally, but a "final assessment" that is operationally rationally defined. It is typically constructed through independent assessments by multiple human evaluators and adjudication of discrepancies.

Q10. With the introduction of ground truth, can I claim that AI is better than humans?

A10.

The purpose of HPVM is **not to claim that AI is better than humans.**

The purpose is to demonstrate **that AI evaluation results are operating stably and reproducibly at an acceptable level for MAH.** If necessary, comparison results with human evaluations may be provided for reference.

Q11. Does HPVM depend on specific AI technologies or vendors?

A11.

No.

HPVM is a neutral methodology that **does not depend on a specific algorithm, AI model, or PV system vendor.**

The evaluation target is the output result of AI, regardless of its internal structure or implementation method.

Q12. How often do you expect HPVM to be applied?

A12.

The frequency of application is **determined by MAH based on its own risk assessment.** In general:

- Periodically (Examples: quarterly, semi-annually, annually)
- After a significant change (AI model update, SOP revision, etc.)

If the HPVM results fail, what happens?

Q13. What happens if HPVM results fail?

A13.

HPVM is a monitoring method that leads not only to pass/fail judgment but also to **corrective and preventive actions (CAPA)**.

In the case of failure,

- trend analysis of errors
 - Analysis of factors causing divergence from human evaluation
 - Review of AI settings, operation, education, and procedures and reevaluate.
-

Q14. As a regulator, what can HPVM help me see?

A14.

HPVM helps us see the following:

- MAH does **not leave AI assessment results as a black box**
- AI assessment results are **compared and monitored with human assessments**
- There are **records and corrective processes that can be explained** when problems occur

This is an important element from the perspective of responsibility and quality management in GVP.

Q15. Will HPVM be able to respond to future AI regulatory trends?

A15.

Yes.

HPVM is a methodology based on the principle of “**continuous monitoring and accountability by human systems**” without relying on specific regulatory language.

Therefore, it **can be applied while maintaining its basic structure** even when regulations and guidance on AI evolve.

Reference Material Statistical Explanation (Supplementary Reference Material)

We present statistical methods and indicators for evaluating the reliability of AI based on a sample scenario (Comparison and validation of AI and human evaluation results using 101 cases randomly sampled $\sqrt{N}+1$ from a total of 10,000 cases).

Here, we use a method to calculate the “correct answer rate (agreement rate)” of AI **and its “confidence interval”** when human (doctor/PV expert) evaluation is “correct”.

Case Evaluation Sample Scenario

- 1) When Intake & Triage evaluation of an individual adverse event report is performed, there are “AI evaluation result: [A]” and “Physician, human evaluation of PV specialization: [B].”
- 2) In this case, there are a total of 101 pairs of 2 evaluation results in which used cases picked up by the $\sqrt{N}+1$ method of random sampling from a population of 10,000 real actual cases are used as test cases.
- 3) [A]: 1 to [A]: 101 "and" [B]: 1 to [B] 101 "
- 4) [A]: 1 to [A]: 101 and [B]: 1 to [B] 101.
- 5) In this case, the human evaluation result is treated as "God evaluation: positive."
- 6) Based on the results of these 101 pairs of 2 term comparisons, the reliability and trustworthiness of the AI evaluation are presented using statistical methods as "confidence limits" and "reliability evaluation index: numerical expression."
- 7) The names of statistical methods and their explanations are presented.

The following are specific indicators and explanations of calculation methods:

1. Reliability evaluation index: numerical expression

The most basic indicator for evaluating AI performance is the percentage of agreement between human and judgment (accuracy rate).

- **Indicator name: Accuracy or Agreement Rate**
- **Formula:** % correct (\hat{p}) = $\frac{\text{"Number of OK's (number of matches)"}}{101}$

For example, if 95 out of 101 examples agree (OK: PASS), the accuracy rate is **94.1%**. $95 \div 101 \approx 0.941$. This is the reliability evaluation index as a point estimate.

2. Calculation of confidence limits (confidence intervals)

A confidence interval is calculated to infer the true performance of an AI in the entire population (10,000 examples) from the results of a sample of 101 examples. Normally, a **95% confidence interval** is used.

- **Indicator: 95% confidence interval for population ratio (95% Confidence Interval for Population Proportion)**
- **Formula (approximate by the Wald method):** Confidence limits = $\hat{p} \pm$

$$1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- \hat{p} : Percentage of correct answers above (101 OK: PASS)÷
- n : Sample size (101)
- 1.96: Confidence coefficient (standard normal distribution value corresponding to 95% probability)

This calculation produces a range from the lower limit to the upper limit. This means "True AI

accuracy falls within this range 95% of the time."

Example of calculation (assuming 95 subjects were OK: PASS)

- $\hat{p} = 0.941$
- $n = 101$
- Standard error = $\sqrt{\frac{0.941 \times (1 - 0.941)}{101}} \approx \sqrt{\frac{0.0555}{101}} \approx 0.0234$
- Error bars = $1.96 \times 0.0234 \approx 0.046$ (4.6%)

Results:

- Reliability rating index: **94.1%**
 - Confidence limits (95% confidence interval): **89.5% to 98.7%**
 - $0.941 - 0.046 \sim 0.941 + 0.046$
-

3. Name of statistical method used and its explanation

The statistical method used in this analysis is "Interval Estimation of Population Ratio (Interval Estimation of Population Proportion)."

Explanation

- 1) **Background:** In this case, the data has a binary (binomial) distribution of OK: PASS or NG. Since it is too much to investigate all 10,000 cases, we sampled VN (10,000) + 1=101 cases for testing.
- 2) **Purpose:** To estimate the true accuracy rate (population ratio) of AI across all 10,000 cases based on the available results (sample ratio) of 101 cases.
- 3) **Reason for selecting the method:**
 - **Normal distribution approximation (Wald method):** Since the sample size can be considered statistically large (large sample), this method, which approximates the binomial distribution to the normal distribution, is common and practical. $N = 101$
 - **Confidence level 95%:** This is the standard criterion used in statistical tests. It indicates the accuracy of "In 100 of the same surveys, 95 will have a true value in this range."

Note: When a more rigorous method is required

If the number of "OK: PASS" is extremely small (or extremely large, close to 100%), the accuracy of the approximation calculation is reduced, and it is recommended to use the techniques called "Wilson Score Interval" or "Clopper-Pearson interval". However, the above standard calculation formula is acceptable for the accuracy verification in general PV business.

(end)