



A Unified Bayesian Framework for AI Uncertainty and Reliability

in Pharmaceutical Development

Anchored in ISO/IEC Guide 98 (GUM), ISO/IEC Guide 99 (VIM), ICH Q2(R2) / Q14 and USP <1220>

Bruno Boulanger & Nils Boulanger

ARLEND & SANAITIO, Belgium

AI-based Devices are Already Inside Pharmaceutical Development

So which framework qualifies it?

Measurement

Spectroscopic bioassays: NIR, Raman, MS
Image-based bioassays
AI-supported analytical procedures

Process and decision

Soft sensors in bioprocesses
Automated visual inspection
Batch disposition support

Framework	The concepts that already exist
FDA / EMA, January 2026	Principle 3: adhere to relevant legal, ethical, technical, scientific, cybersecurity and regulatory standards. The only family it names is GxP
ISO/IEC Guide 98-1:2024 (GUM-1)	Measurement is defined as an “experimental or computational process ”, on material or virtual objects
JCGM 106:2012 (Guide 98-4)	Conformance probability, decision rules, guard bands, consumer's and producer's risk
JCGM GUM-6:2020 (Guide 98-6)	Empirical, data-driven measurement models; parameter and model-choice uncertainty
ISO/IEC Guide 99 (VIM)	Target measurement uncertainty , rated operating condition, drift, metrological compatibility
ICH Q2(R2), Q14, USP <1220>, <1790>	ATP (Analytical Target Profile), established conditions, continued performance verification, probability of detection
ISO 15189, ISO/TS 20914, CLSI EP21/EP46	Mandatory MU , top-down estimation from routine QC, allowable total error limits
NASEM 2019	Reproducibility , replicability and generalizability defined, with an explicit VIM comparison

None of these frameworks needs a new concept for AI. Each already contains the slot the AI has to fill.



One Question, Six Vocabularies

Developers, quality units and regulators are describing the same object in six languages

**Personalized uncertainty
quantification (PUQ)**

machine learning

**Conformal
prediction**

machine learning

**Calibration
(ML sense)**

machine learning

**Uncertainty of
measurement**

metrology / GUM

**Total analytical
error (TAE)**

analytical & clinical chemistry

**Predictive values
PPV, NPV**

diagnostics

- Four professional communities: machine learning, metrology, analytical and clinical chemistry, diagnostics
- Overlapping but disconnected terminology
- No shared conceptual foundation, so acceptance criteria cannot be specified
- AI components cannot be qualified inside CMC and analytical lifecycle frameworks

One object underlies all six.



First, the False Friends

The same word, three different meanings

Word	ICH / USP sense	ISO / VIM sense	AI & Machine-learning sense
Calibration	Establishing the response-to-concentration relationship with reference standards: the calibration curve, or a multivariate calibration model	Relating values from measurement standards to instrument indications , then inverting that relation (VIM 2.39)	Predicted probabilities agree with observed event rates (Van Calster et al. 2025)
Accuracy	Closeness of the value found to an accepted reference value. This is trueness alone. Q2(R2) permits combining it with precision	Trueness AND precision together; a qualitative concept, never a number (VIM 2.13)	Proportion of correct classifications; a single number
Validation vs verification	Validation of a new procedure against the ATP, USP <1225>; verification that a compendial procedure suits the actual conditions of use, <1226>	Verification = specified requirements are met (VIM 2.44); validation = verification for an intended use (VIM 2.45)	Validation set = data used for tuning; test set = held out
Reproducibility	Precision between laboratories, from a collaborative study. Within-laboratory variation is intermediate precision	Precision under changed sites, operators and measuring systems; always a number (VIM 2.25)	Someone else can re-run the code and get the same number
Replicability	Not used as a term.	Not a VIM term.	NASEM 2019 defines it: new study, new data, same question

Accuracy is the major confusion between pharmacopoeial guidance and the other standards. In ICH and USP, accuracy means trueness. In the VIM it means trueness and precision combined. ISO/IEC TS 4213:2022, the ISO standard for ML classification performance, is built on the confusion-matrix family.

Before unifying the concepts, we have to stop the word confusion.

The Single Fundamental Object

Mind the notation! The two fields use it in opposite directions

- η the measurand: the true property we want (concentration, titer, potency, defect state, pH)
- z the indication or response the instrument produces (spectrum, image, chromatogram, pixel array)
- θ the model parameters
- D the training and validation data
- $g_0(\eta)$ the prior, eg the process distribution of the measurand

*NB: in machine learning the **indication** z is called x and the **measurand** η is called y . In the classical calibration curve it is the reverse: x is the concentration and y is the response, and the fit is inverted afterwards. Same two quantities, opposite letters.*

Multivariate calibration sits with machine learning: in PLS and PCR, X is the spectral response and y is the concentration.

A Discriminative fit

η was observed in training. The prior on η is implicit, equal to the label distribution in D .
The result is Posterior predictive of the measurand given the “indication”, ie the response.

$$p(\eta | z, D) = \int p(\eta | z, \theta) \cdot p(\theta | D) d\theta$$

B Calibration direction

The bracket is the posterior predictive of the indication.
The result is the posterior predictive for a latent unknown quantity.

$$p(\eta | z, D) \propto \left[\int p(z | \eta, \theta) \cdot p(\theta | D) d\theta \right] \cdot g_0(\eta)$$

- A needs one integral.
- B needs Bayes' theorem, a prior on the measurand, and a normalising constant. That prior is g_0 , the same process distribution JCGM 106:2012 Annex B.1 assigns from statistical process control (SPC)
- With a diffuse prior $g_0(\eta)$ over the working range, B reproduces the classical inverse-prediction distribution of a calibration curve

The deliverable is the distribution of the measurand given the indication or response.



Pharmacopoeial Validation

USP <1210> and ICH Q2(R2) compute this interval today, without calling it a posterior

- Integrates over the posterior $p(\theta | D)$: parameter uncertainty is carried, not discarded
- The complete statement of what the procedure knows about this sample, this patient, this batch
- Every reliability or performance metric is a functional of $p(\eta | z, D)$

The interval they already compute

USP <1210> recommends a two-sided tolerance interval as a single criterion validating accuracy and precision simultaneously. ICH Q2(R2) permits combined accuracy-and-precision criteria and uses prediction intervals.

Under a non-informative prior $g_0(\eta)$, the β -expectation tolerance interval and the credible interval of the distribution of the measurand are the same interval.

This is the **accuracy profile** as in SFSTP and the AFNOR / ISO guidance.

What the distribution adds

An interval answers yes or no: is it inside $\pm\lambda$?

The distribution answers with a probability: $P(|\eta - \mu| > \lambda)$, which is exactly $1 - p_c$ in JCGM 106:2012.

The standard deviation is the combined standard uncertainty u_c , so the half-width of the 95% interval is the expanded uncertainty U .

JCGM = Joint Committee for Guides in Metrology.

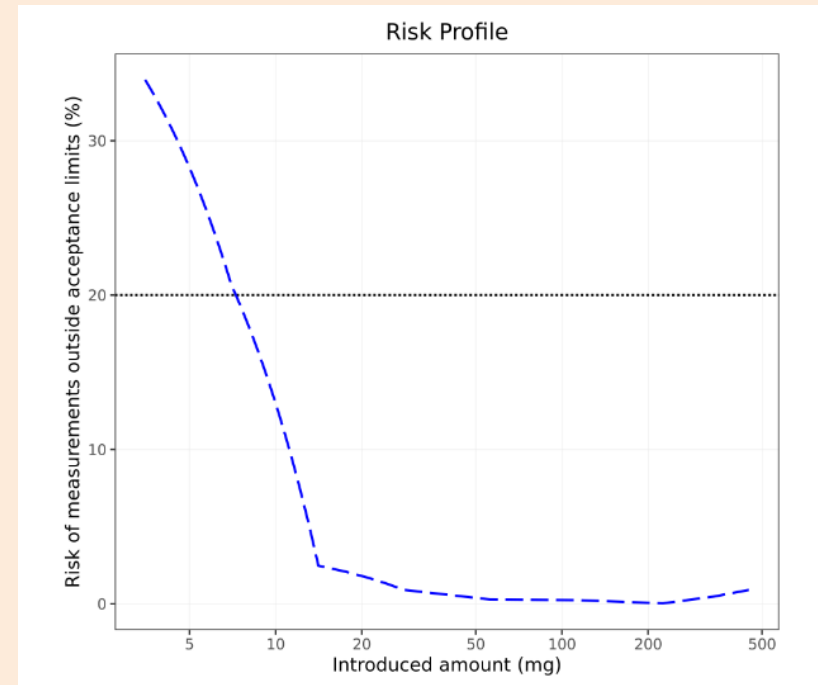
The interval is already computed. Only the probability is currently ignored in pharmaceutical world.

Pharmacopoeial Validation and ISO reconciliation

ICH-Q2(R2) Interval and TAE



Risk $\Rightarrow P(| \eta - \mu | > \lambda | z, D)$



With the distribution $p(\eta | z, D)$ you can compute the interval AND the probability $P(| \eta - \mu | \leq \lambda | z, D) = \int p(\eta | z, D) d\eta \text{ over } C = p_c$.

The Same Interval, Read Three Ways

ICH Q2(R2), total analytical error and JCGM 106 are three readings of one quantity

$$P(|\eta - \mu| \leq \lambda | z, D) = \int_C p(\eta | z, D) d\eta = p_c$$

ICH Q2(R2), USP <1210>

Fix the probability at β and read off the interval.

The prediction or β -expectation tolerance interval, compared with $\pm\lambda$.

Answers yes or no: is the interval inside the acceptance limits?

Total analytical error

CLSI EP21 and CLSI EP46.

The same spread, aggregated over the domain, with bias added linearly, and compared with an allowable total error limit.

A **procedure-level** construction, not a synonym for **measurement** uncertainty.

JCGM 106:2012, cl. 7.1

Fix the tolerance at λ and read off the probability p_c , **the conformance probability**.

Its complement $1 - p_c$ is the specific consumer's risk for this result.

Same integral, same interval, one extra number.

μ the target or reference value the result is compared with

λ the tolerance around it

$C = [\mu - \lambda, \mu + \lambda]$ the acceptance interval

$p_c = P(\eta \in C | z, D)$ the conformance probability

β the minimum acceptable value of p_c

The accuracy profile becomes a risk profile: the same interval, with a probability attached, varying across the range.



The 2024 GUM Part 1 Already Admits Computation

ISO/IEC Guide 98-1:2024, identical to JCGM GUM-1:2023, replaces the 2009 introduction. The GUM proper, 98-3 / JCGM 100:2008, is unchanged.

What changed in the definition

Measurement is now described as an experimental **or computational process** producing an estimate plus an uncertainty statement, intended to support decision making.

The property measured may belong to a material or virtual object, or to a process or event.

Why it matters here

A trained model is a computational process acting on a virtual object.
An AI-based device therefore falls inside the scope of the GUM as currently written.

Part of the GUM suite	What it gives an AI-based device
98-3 / JCGM 100:2008	Law of propagation of uncertainty; combined and expanded uncertainty, coverage factor
98-3 S1 / JCGM 101:2008	Propagation of distributions by Monte Carlo, which is what posterior sampling already does
98-4 / JCGM 106:2012	Conformance probability, acceptance limits, guard bands, consumer's and producer's risk
98-6 / JCGM GUM-6:2020	Empirical models: parameter uncertainty (cl. 7), model-choice uncertainty (cl. 8.5), errors in inputs and outputs (cl. 9-10)

JCGM = Joint Committee for Guides in Metrology.

GUM-6 already covers empirical models fitted to data. In principle it applies to any AI model today.



ICH Q2(R2) and Q14 Already Cover Multivariate Models

Adopted November 2023, effective June 2024. The AI-relevant clauses are already written.

ICH clause or concept	What it says	Read it as
Q2(R2) scope	Validation extended to multivariate calibration models, bioassays and non-chromatographic procedures	An AI calibration model is explicitly in scope
Q2(R2) two phases	Model development is calibration plus internal testing, then validation on independent samples	The train / test split, written into an ICH guideline
Q2(R2) combined criteria	Accuracy and precision may be assessed together, with prediction intervals among the tools	The tolerance interval, in ICH wording
Q2(R2) transfer and co-validation	Partial or full revalidation, or comparative analysis, when moving laboratory; multi-site data can satisfy transfer	Re-verification at a new site, and a hierarchical multi-site model, both already permitted
Q14 MODR, plus Q12	Method operable design region (MODR) from enhanced development; established conditions and pre-agreed change management	The region where the ATP holds, and the update protocol

$P(|\eta - \mu| \leq \lambda | z, D) = p_c$, the conformance probability of JCGM 106:2012, cl. 7.1

**Q2(R2) avoid uncertainty concept, only use interval and acceptance criteria.
JCGM 106:2012 named that probability 14 years ago, SFSTP 23 years ago.**

USP Already Has the Three Operations

Validate, transfer, verify, plus a documented rule for when re-verification may be waived

<1225> Validation

Validation of Compendial Procedures.
Establish the performance characteristics:
accuracy, precision, specificity, range,
robustness.

Initial qualification of the model against the ATP

<1224> Transfer

Transfer of Analytical Procedures.
Comparative testing, co-validation between
laboratories, partial revalidation, or a justified
transfer waiver.

Deploying the model to a new site or CDMO

<1226> Verification

Verification of Compendial Procedures.
Confirm the procedure is suitable under the
actual conditions of use, without full
revalidation. No acceptance criteria clearly
defined, but inherited from <1225>.

Local re-verification at the receiving site

The transfer waiver is the pharmaceutical PCCP: a rule agreed in advance for when a change does not require re-verification.

*PCCP: FDA 2024, Marketing Submission Recommendations for a **Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions**.*

Validation asks whether the procedure can meet the ATP. Verification asks whether it does so here.
Transfer is the operation in between. An AI-based device needs all three.



Reproducibility, Replicability, Generalizability

NASEM 2019 settled the terminology, and compares its terms with the VIM definitions directly

The question actually being asked	NASEM 2019	VIM / ISO	ICH and USP	For an AI system
Same data, same code, same result?	Reproducibility (computational)	Not a metrology concept	Data integrity, raw-data traceability	Fixed seeds, containerised inference
Same lab, repeated or slightly varied?	Not a NASEM category; this is precision	Repeatability (2.21); intermediate precision (2.23)	Repeatability and intermediate precision across day, analyst, equipment: Q2(R2), <1225>	Run-to-run at fixed input; then library, hardware or version change
New data, same question?	Replicability	Reproducibility, as a precision number (2.25)	Reproducibility between laboratories; co-validation	External validation at a new site
A new laboratory takes the procedure over?	Replicability	Metrological compatibility (2.47)	Transfer <1224>, verification <1226>	Site deployment plus local re-verification
Other contexts or populations?	Generalizability	Outside the rated operating condition (4.9)	Outside the MODR or the validated range	Transportability, domain shift

Your AI supplier and your QC unit may mean opposite things by the word reproducible.

Reliability is not a function of model complexity. It is a function of where the training data live relative to where the procedure is used !!



Six Vocabularies Reconciled

Each concept is a functional of $p(\eta | z, D)$.

Today's concept	Functional of $p(\eta z, D)$
Trueness, precision (ICH Q2, GUM)	Location and spread
Uncertainty of measurement	Credible interval of the distribution, which the GUM calls a coverage interval
β -expectation tolerance interval (USP <1210>)	The same interval, under a non-informative prior
Total analytical error (TAE)	The same spread, aggregated over the domain, with bias added linearly
Personalized uncertainty quantification (PUQ)	The predictive distribution for one sample
Predictive values, PPV and NPV	Specific consumer's and producer's risk (JCGM 106:2012)

The six vocabularies become one vocabulary



Mapped onto the ICH Q14 and USP <1220> Lifecycle

Three regulatory vocabularies, one column of mathematics

ICH Q14 / Q2(R2) / USP <1220>	GUM / VIM anchor	Bayesian counterpart
Analytical Target Profile	Target measurement uncertainty (VIM 2.34); conformance probability (98-4)	Probability statement
Performance characteristics	Trueness, precision, selectivity, range (VIM 2.14, 2.15, 4.13, 4.7)	Location, spread, coverage, calibration
Working, validated range	Measuring interval and rated operating condition (VIM 4.7, 4.9)	Region where the ATP probability holds: $\{ z : C_m(z) \geq \text{target} \}$
Reportable value	Measurement result, which may be reported as a PDF (VIM 2.9)	A summary of $p(\eta z, D)$, with the replication strategy
Decision rule (USP <1220>, Specification and Decision Rules)	Acceptance limits and guard bands (98-4, cl. 8)	Accept when $p_c = P(\eta \in C z, D) \geq \beta$; the guard band follows, rather than being set separately
Transfer to a new site, USP <1224>, <1226>	Metrological compatibility (VIM 2.47); proficiency testing (ISO/IEC 17043)	Site posterior, tested for compatibility
Continued performance verification	ISO 15189:2022, ISO/TS 20914 top-down MU from routine QC	Sequential posterior update

The AI is qualified with guidance that already exists.



The Ten FDA and EMA Principles, and Where They Land

Guiding Principles of Good AI Practice in Drug Development. FDA CDER and CBER with EMA, January 2026. Titles as published.

Principle	Where this talk answers it
1 Human-centric by design	A values question. Out of scope here
2 Risk-based approach	Model risk as a number: conformance probability, specific and global risk
3 Adherence to standards	The standards it does not name: GUM 98-1, 98-4, 98-6, VIM, ICH, USP
4 Clear context of use	Target measurement uncertainty (VIM 2.34) and rated operating condition (4.9)
5 Multidisciplinary expertise	Six vocabularies, and the false friends between them

Principle	Where this talk answers it
6 Data governance and documentation	Traceability of the labels; computational reproducibility
7 Model design and development practices	Generalizability and robustness as the precision hierarchy
8 Risk-based performance assessment	Metrics appropriate to the context of use: proper scores, net benefit, specific risk
9 Life cycle management	Data drift: g_0 and SPC; sequential re-verification
10 Clear, essential information	Report the conformance probability and the capability

Principle 3 asks for adherence to relevant standards.

Thank you

And one question we cannot answer yet

Principle 8 asks for a risk-based performance assessment of the complete system, including human-AI interactions.

Nobody has joined the two into one risk statement.

Specific risk extends naturally to a two-stage human-plus-machine rule. That work is not done, and we would like to do it with people during the panel discussion.



Trusted by major pharma for nearly two decades.

bruno.boulanger@arlenda.com



References

Backup slide

■ Metrology

- ISO/IEC Guide 98-1:2024 (JCGM GUM-1:2023), Introduction
- JCGM 100:2008, law of propagation of uncertainty
- JCGM 101:2008, propagation of distributions by Monte Carlo
- ISO/IEC Guide 98-4 (JCGM 106:2012), conformity assessment
- JCGM GUM-6:2020, developing and using measurement models
- ISO/IEC Guide 99:2007 (VIM 3); VIM 4 in preparation
- ISO 5725; ISO/IEC 17025:2017; ISO/IEC 17043
- ISO 33406:2024, reference materials with qualitative properties

■ Pharmaceutical and clinical

- FDA CDER/CBER with EMA, Guiding Principles of Good AI Practice in Drug Development, January 2026
- FDA 2024, PCCP for AI-Enabled Device Software Functions
- FDA CDER, AI in Drug Manufacturing, discussion paper, 2023
- ICH Q2(R2), Validation of Analytical Procedures, 2023
- ICH Q14, Analytical Procedure Development, 2023; ICH Q12
- USP <1220>, <1224>, <1225>, <1226>, <1010>, <1210>
- USP <790> and <1790>, visual inspection of injections
- ISO 15189:2022; ISO/TS 20914:2019
- CLSI EP21, 3rd ed. 2025; CLSI EP46, 1st ed. 2025

■ AI, statistics and evidence

- NASEM 2019, Reproducibility and Replicability in Science, doi:10.17226/25303
- Plant & Hanisch, Reproducibility in Science: A Metrology Perspective
- Barba 2018, terminology usage across disciplines
- Hubert et al., accuracy profile and β -expectation tolerance intervals (SFSTP)
- Van Calster et al. 2025, STRATOS topic group 6
- Bilson, Cox, Pustogvar & Thompson 2025, arXiv:2504.03359
- Nordin et al. 2018, IFCC-IUPAC nominal property vocabulary
- Eurachem / CITAC 2021, qualitative chemical analysis
- ISO/IEC 42001:2023; ISO/IEC TS 4213:2022, under revision as DIS 4213