Federal Institute
for Drugs
and Medical Devices

Artificial Intelligence for Drug Development in Cardiovascular Medicine

Clemens Mittmann

AI definition

Disclaimer

The views expressed in this presentation are the personal views of the speaker and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency (EMA), its committees, working parties, or any of their members.

All examples are based on publicly available information and may not correspond to actual ongoing or past EMA procedures.

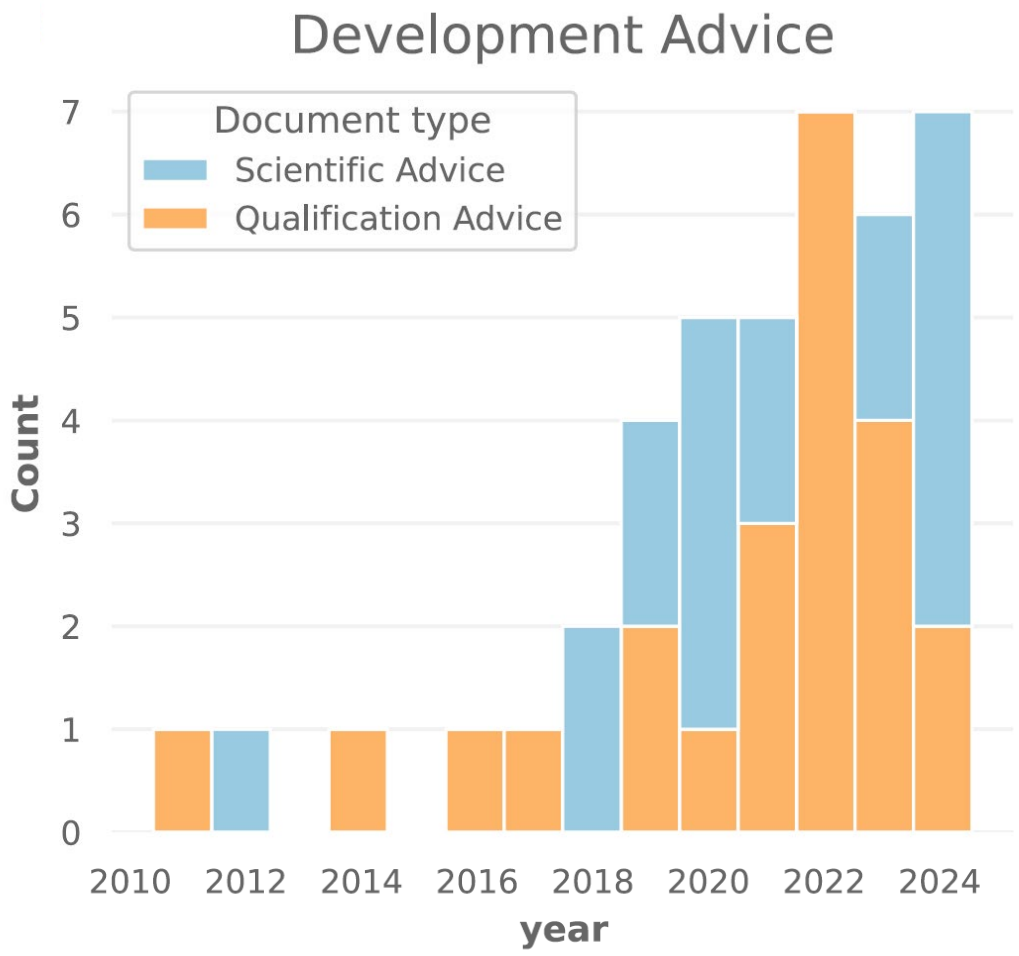
AI definition

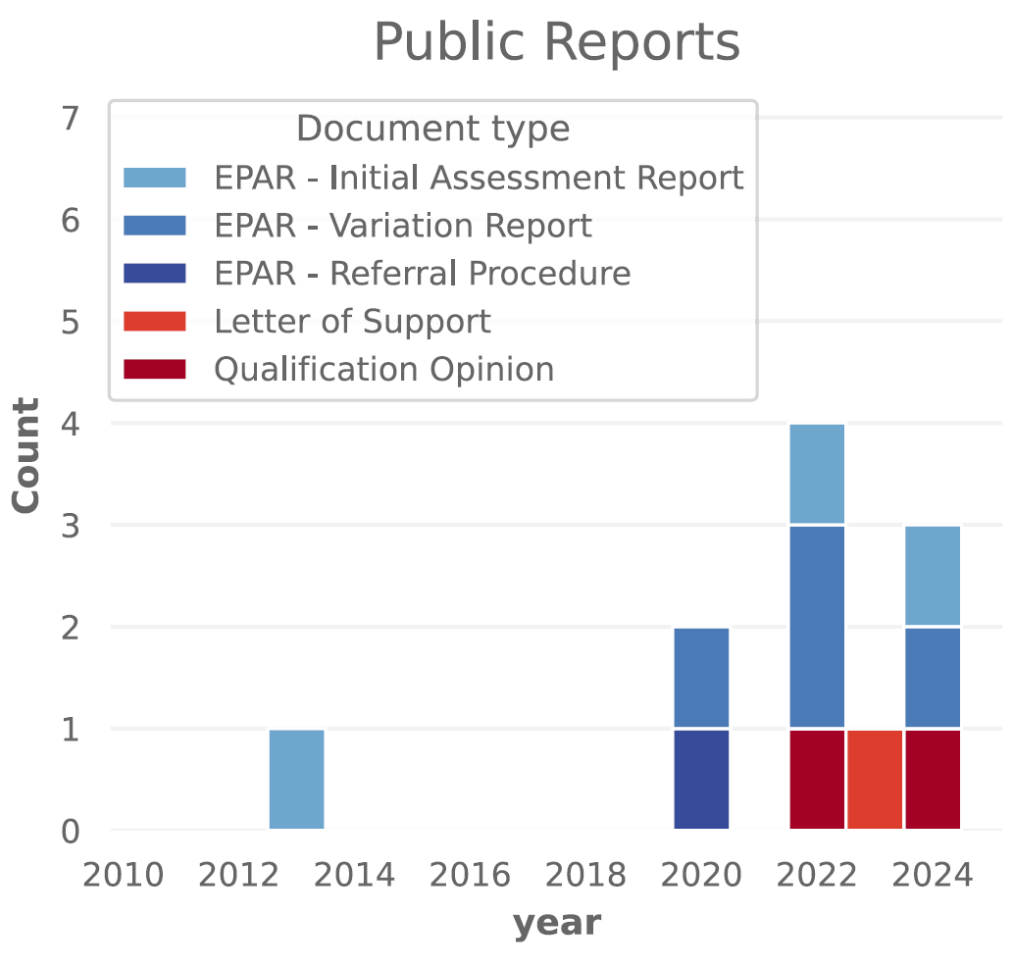
“An AI system is a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environment.“

In other words:
AI systems take in information, identify patterns, and produce outputs such as predictions, recommendations, or decisions that may influence real-world or digital outcomes.

(Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle, EMA/CHMP/CVMP/83833/2023)

AI in EMA procedures

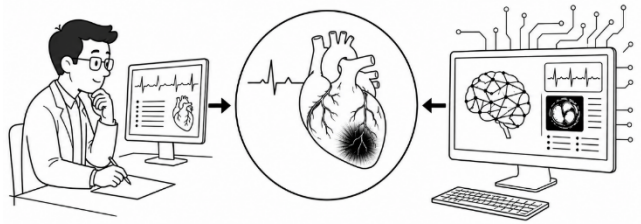




Nollen L-M et al., CLINICAL PHARMACOLOGY & THERAPEUTICS 2026; doi:10.1002/cpt.70300

Endpoint Adjudication in Cardiovascular Trials

- **Machine learning (ML) based** models
- Adaptive **hybrid AI model** combining large language models (LLMs) and ML
  (Vemulapalli et al. (ADAPT-CEC) Circulation 2026;153:1694–1706)
- **Rule based** approach based on LLM
- **Combined** orthogonal approach



"Human-in-the-loop" approach:

automated classifier is **confident**
→ **Accept result**
automated classifier is **uncertain**
→ Adjudication by **human experts (CEC) (triage)**

Triage for uncertain decisions
improved the results for all models.

- Clinical events dossiers
(455 CV deaths/332 CEC accepted, 659 MIs/363, 167 Strokes/125)
- Converted from PDF to plain text (OCR)

- → primary adjudication model: LLM, OpenAI o1-mini) (checks individual criteria, provides reasoning)

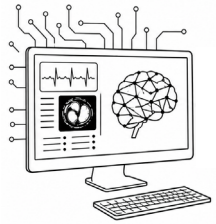
Plus

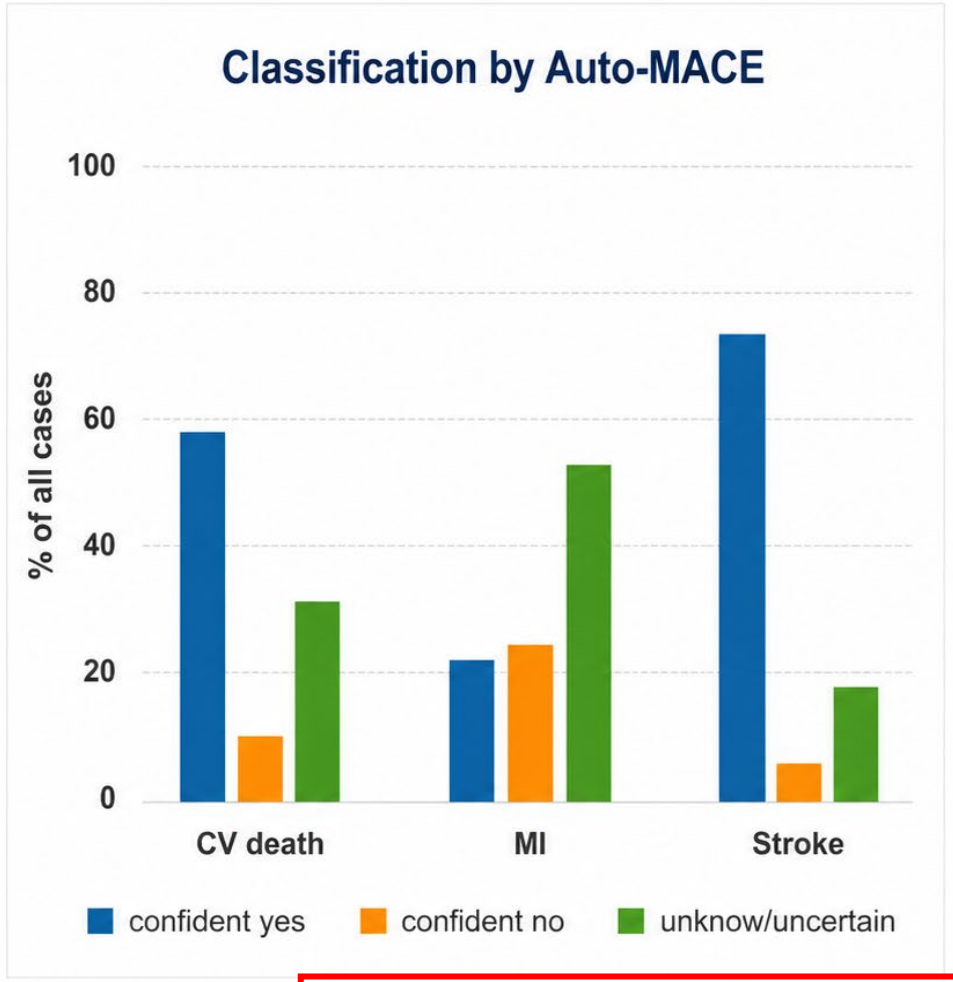
- → Clinical Longformer (trained on 6 clinical trials)

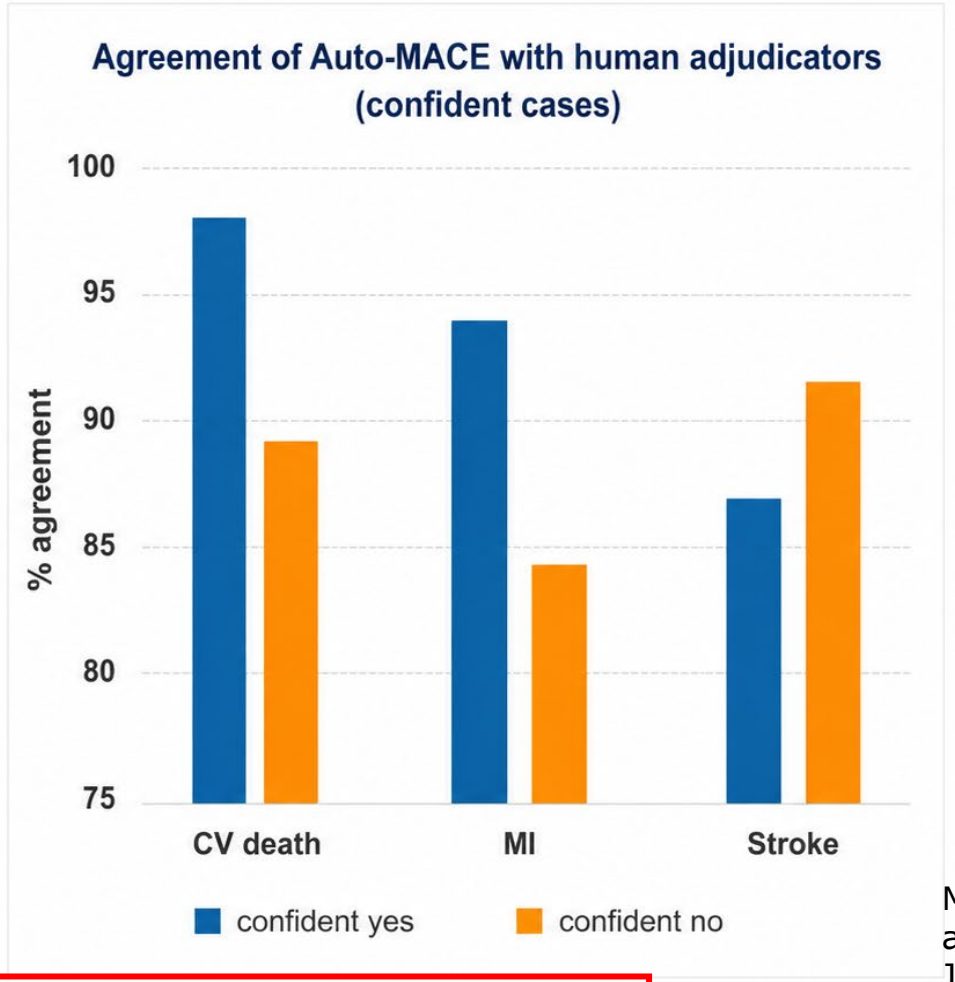
- → classification based on both methods together

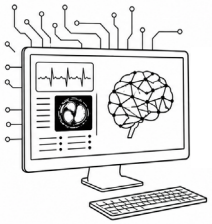
- (Marti-Castellote et al. (Auto-MACE) JACC 2026; 87: 1856 – 1866)

Endpoint adjudication







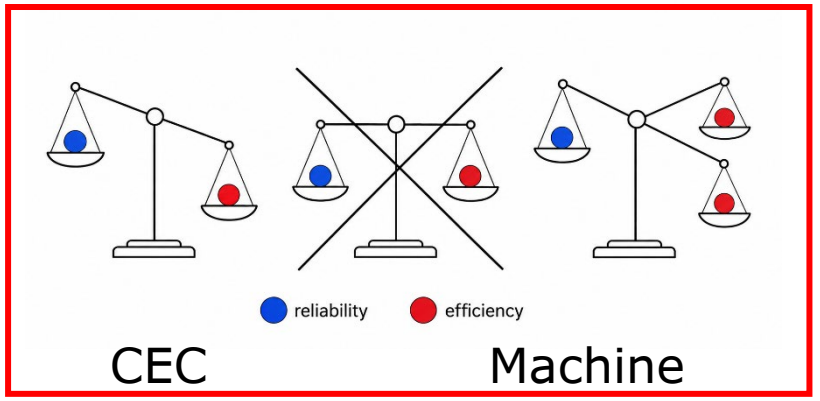


Main study result PARADISE-MI trial (HR for MACE (95% CI)

CEC :	0.90 (0.77-1.05)	(664 first events)
AI:	0.91 (0.78-1.07)	(600 first events)
AI with CEC:	0.92 (0.79-1.07)	

Marti-Castellote et al. (Auto-MACE) JACC 2026; 87: 1856 – 1866, data extracted from Tables 1, 3, and 4

Endpoint adjudication, Validation



Principles

- Qualification is **not** intended for a **general pipeline or framework**.
- A qualified method should be **broadly applicable**.
- **Efficiency** improvements are desirable without compromising **reliability**.

„It is the responsibility of the clinical trial sponsor, marketing authorisation applicant/holder or manufacturer to ensure that all algorithms, models, datasets, and data processing pipelines used are fit for purpose and are in line with legal, ethical, technical, scientific, and regulatory standards as described in EU legislation, GxP standards and current EMA guidelines.“

Model architecture and performance

Validation of each individual component.

Performance and Bias Monitoring (Lifecycle management)

- Case specific: Monitoring and quality controls, • Management of minor model updates,
- Risk management plan, • Predefined mitigation strategies for model, data, and concept drift.

Endpoint adjudication, Validation

Clinical validation

- **Endpoint** specific validation

- **Model performance is limited by the quality of the input data**
  - Detailed assessment of information that is available and used by model vs. human adjudicators
  - Temporal patterns and decision thresholds (ECG, cardiac biomarkers) are difficult to capture

- **The process should be clinically interpretable and understandable**
  - Is the information used by the model transparent and understandable?
  - Are the decision rules consistent with established adjudication criteria?
  - Prognostic baseline characteristics should generally not influence event classification.

- **Performance metrics alone do not replace qualitative assessment**
  - Human–machine and human–human discordance arise from different factors
  - Challenging: Myocarditis vs. MI, STEMI vs. NSTEMI, Patient history vs. new event, Multimorbidity
  - Only highly reliable classification may justify exemption from human review

AIM-NASH: Qualification opinion

AI Support for Central Pathologist in MASH Clinical Trials

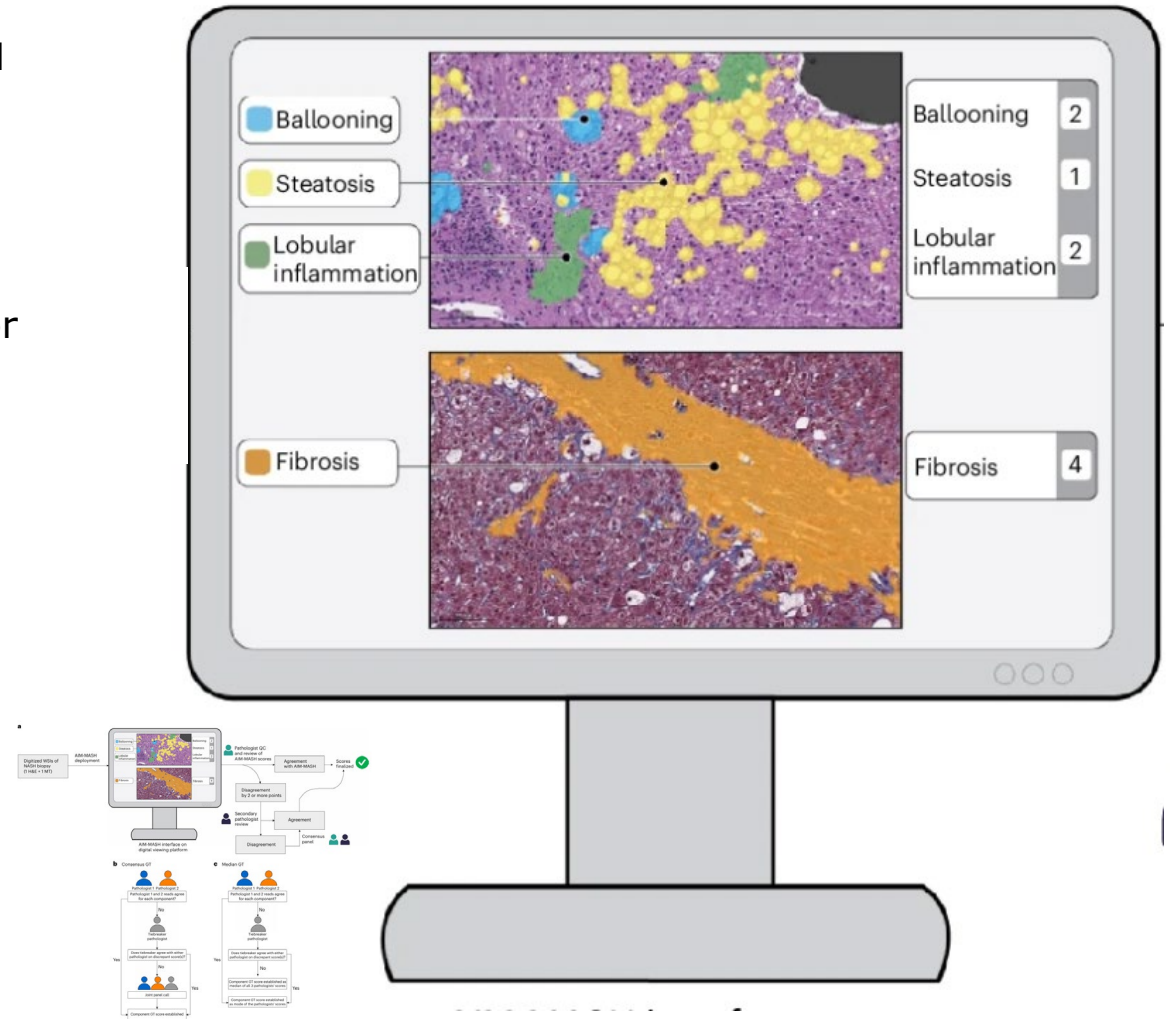
- Enrolment/inclusion of patients into clinical phase 2 and phase 3 trials in MASH (**Metabolic-Dysfunction-Associated - Steatohepatitis**)
- Evaluation of the study outcomes (primary or secondary)

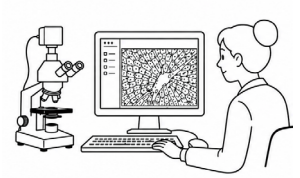
AI-based Determination of disease activity biomarker in biopsies

- NAS component scores (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage

Role of Pathologist (human-in-the loop)

- Confirms sample evaluability
- accepts/rejects AI assessment of NAS components and fibrosis stage
- determines presence of any additional findings





Agree/Disagree

Agree/Disagree

Agree/Disagree

Agree/Disagree

Disagreement by ≥ 2 points → Second pathologist

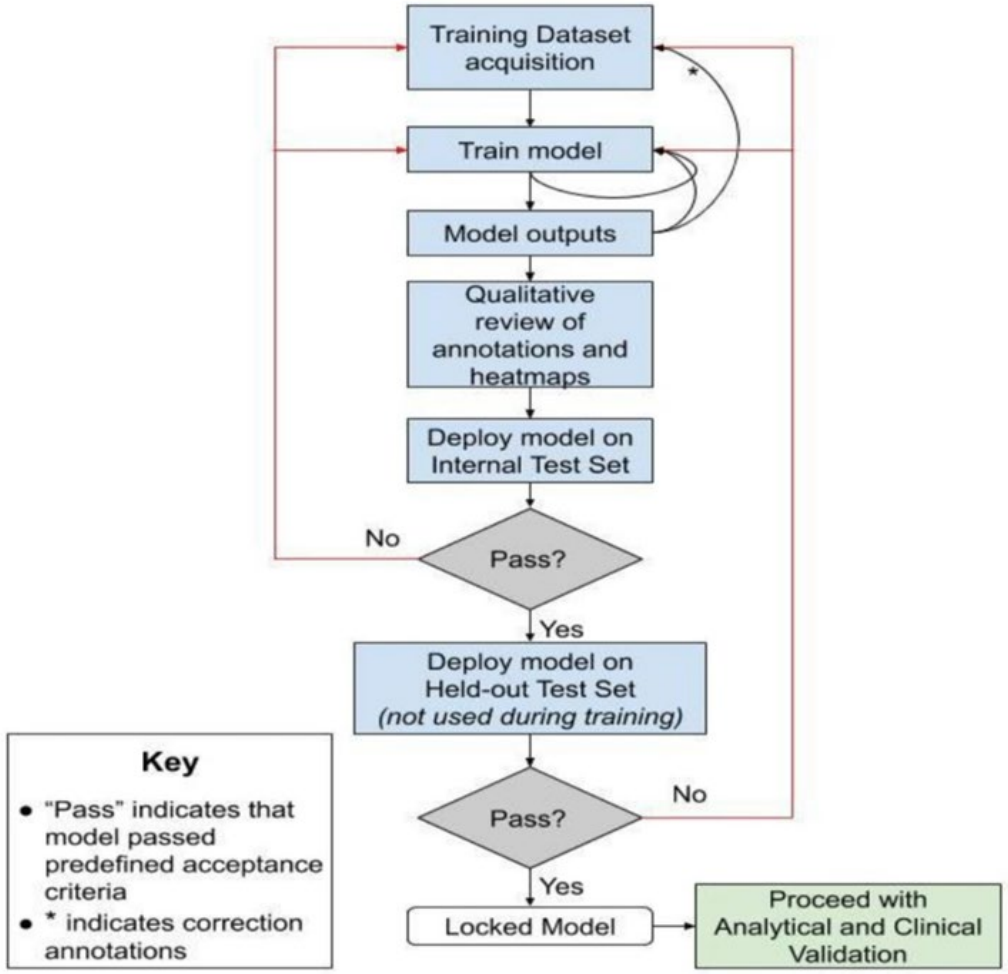
AIM-NASH: Qualification opinion for Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology in Liver Biopsies to Determine Disease Activity in NASH/MASH Clinical Trials EMADOC-1700519818-1761332 Corr.1

Pulaski H. et al., Nature Medicine 2025; 31: 315 –322

EMA CV 02.07 2026

AIM-NASH: Qualification opinion

Figure 2: AIM-NASH Iterative Model Development



Training datasets:
- NAS Feature (H&E): n = 6227
- Fibrosis Stage (trichrome): n = 6215

Different models developed:
Artefact detection, H&E Tissue Model, Trichrome Tissue, Trichrome Large Septa, NAS component scoring and fibrosis staging.

Lifecycle-management
- Monitoring plan in place
- No „variation“ procedure for the qualification opinion is foreseen. Minor changes should be documented and reported.

AIM-NASH: Qualification opinion for Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology in Liver Biopsies to Determine Disease Activity in NASH/MASH Clinical Trials
EMADOC-1700519818-1761332 Corr.1. page 12

AIM-NASH: Qualification opinion

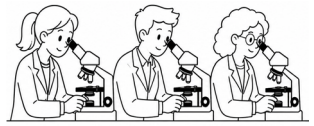
Table 22: Primary endpoint results for each NASH histologic component

Feature	Modality	N	WK (95% CI)	Difference (95% CI)	p-value for NI	p-value for Superiority
Steatosis	AI-assisted vs GT	1467	0.677 (0.652, 0.703)	0.003 (-0.026, 0.038)	<0.0001	0.3455
	IMR vs GT	1481	0.674 (0.651, 0.694)			
Lobular inflammation	AI-assisted vs GT	1465	0.419 (0.361, 0.46)	0.123 (0.069, 0.173)	<0.0001	<0.0001
	IMR vs GT	1478	0.297 (0.265, 0.329)			
Hepatocellular ballooning	AI-assisted vs GT	1465	0.563 (0.519, 0.601)	0.15 (0.104, 0.194)	<0.0001	<0.0001
	IMR vs GT	1476	0.414 (0.385, 0.442)			
Fibrosis	AI-assisted vs GT	1429	0.653 (0.627, 0.676)	0.008 (-0.026,	<0.0001	0.42
	IMR vs GT	1453	0.645 (0.622, 0.665)	0.039)		

Clinical validation

AIM-NASH compared with individual pathologist reads (IMR):

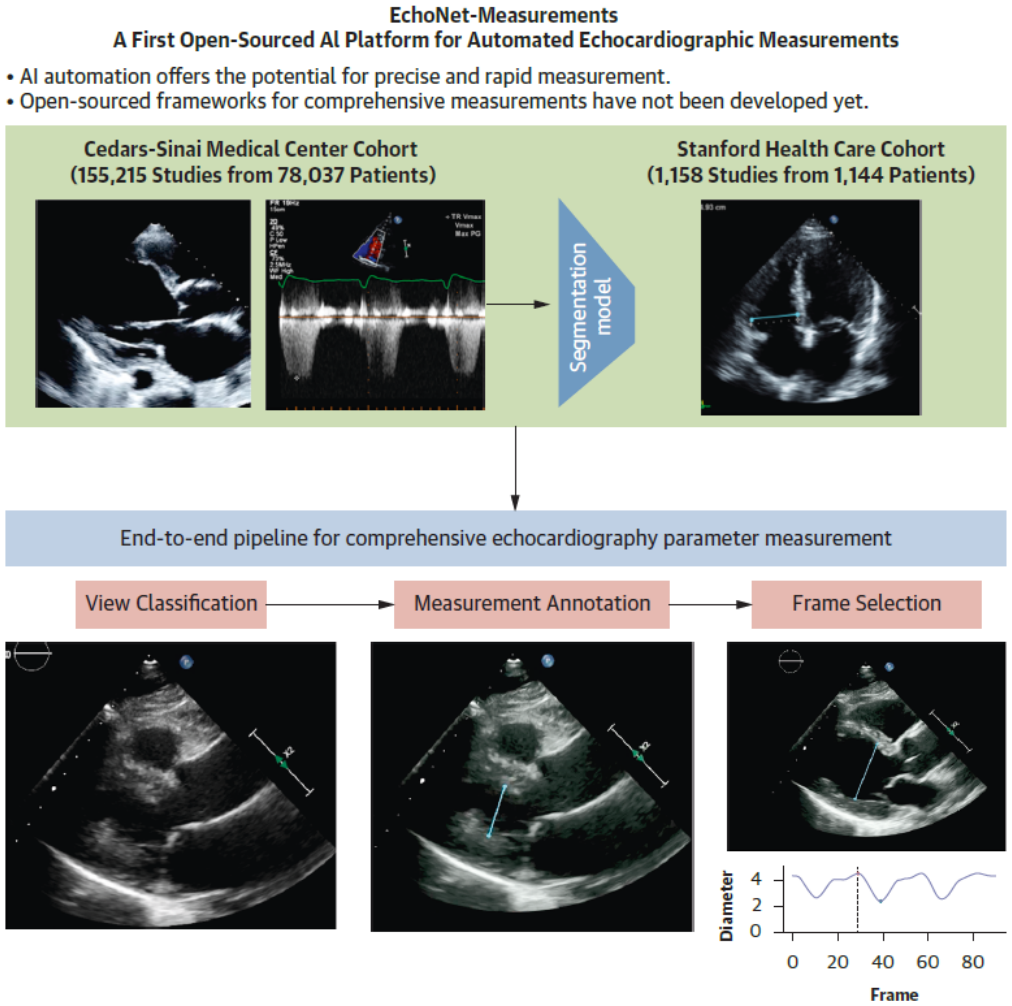
- **Non-inferior** for the features of fibrosis staging and steatosis grading.
- **Superior** for the assessment of lobular inflammation and hepatocellular ballooning.



GT: Ground truth
WK: Weighted Kappa (agreement measure)
NI: Non inferiority

AIM-NASH: Qualification opinion for Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology in Liver Biopsies to Determine Disease Activity in NASH/MASH Clinical Trials
EMADOC-1700519818-1761332 Corr.1.

Automated measurement of echocardiography parameters



High accuracy in automated echocardiographic measurements compared with clinicians.

Sahashi Y, et al. JACC. 2025;86(13):964-978.

- Trained on data from Cedars-Sinai Medical Center (CSMC).

- Outputs compared with sonographer measurements using data from CSMC and an external data set (Stanford Health Care, SHC).

→ High accuracy demonstrated on held-out datasets.
→ Performance consistent across patient subgroups.

Use in drug development?

- Fit for purpose validation depending on patient risk and regulatory impact

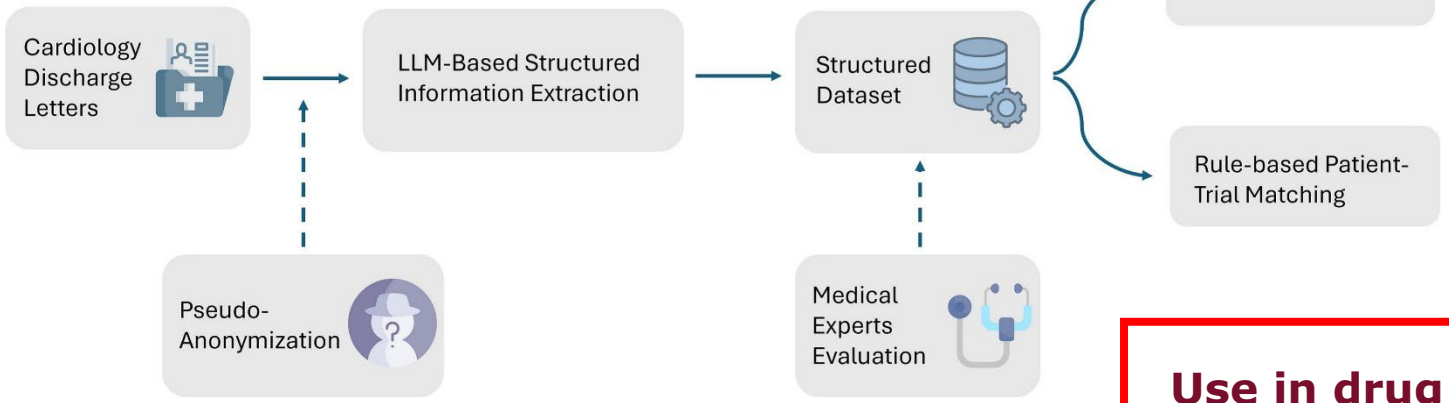
EMA Qualification procedure

- Not mandatory, but recommended for high-impact use cases, such as
  - Endpoints relevant to benefit-risk assessment
  - Patient safety (e.g. dosing in HOCM)
  - Novel AI derived efficacy or safety parameters

EMA CV 02.07
2026

Patient selection for clinical studies

Pescol (2026)



Testing for eligibility criteria for Inclisiran and Alirocumab studies (n=100):

False negatives below 3%: 13/14 criteria
False positives below 3%: 12/14 criteria

Average potential error in criteria verification: below 10%

Use in drug development?
AI/ML systems may support patient selection based on disease characteristics or other clinical parameters.

Key considerations

- Selection bias, enrichment bias and generalizability to the target population
- Transparency and explainability
- Human oversight

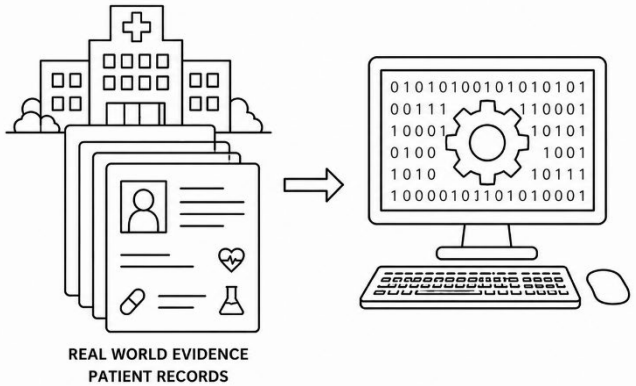
Pescol F et al.; International Journal of Medical Informatics 2026; 217: 106509

Use of AI when analysing Real-World Data

Effectiveness and drug repurposing: SGLT2 Inhibitors (use case 4)

Evaluate the comparative effectiveness of SGLT2 inhibitors by applying a target trial emulation (TTE) approach to RWD, using artificial intelligence and machine learning algorithms.

The targeted maximum likelihood estimation is employed for TTE. Using a doubly robust estimation approach, machine learning and AI methods such as deep learning are applied and compared.



Some preliminary regulatory thoughts

1. Assess **quality of database**

Fit for purpose assessment according to

- Reflection paper on use of real-world data in non-interventional studies (EMA/99865/2025)
- Data Quality Framework for EU medicines regulation, EMA/326985/2023

2. Assess **target trial emulation** methodology
3. Assess validation and performance of the **AI model**

Real4Reg

https://www.real4reg.eu/

AI and GxP aspects

- Has the intended use been clearly defined?
- Are the data sources controlled and documented?
- Has the model been validated and is validation documented?
- Is model performance continuously monitored?
- Are changes managed through a controlled change management process?
- Is an audit trail in place?
- Is there human oversight and approval of critical decisions?
- Is the documentation complete and up to date?

High regulatory impact or high patient risk in a study:
For systems used outside the qualified context or not previously qualified, additional information on model architecture, development, validation, training data, and data processing should be available for regulatory assessment (e.g., CTA or MAA). GCP inspections take qualification status into account and follow a risk-based approach.

- Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle EMA/CHMP/CVMP/83833/2023
- ICH E6 guideline for good clinical practice (GCP)

Summary

The development and use of AI in drug development should align with EMA and FDA guiding principles.

- Human-centric by design
- Risk-based approach
- Adherence to standards
- Clear context of use
- Multidisciplinary expertise
- Data governance and documentation
- Model design and development practices
- Risk-based performance assessment
- Life cycle management
- Clear, essential information

• Responsible integration of AI into drug development can accelerate innovation and improve the efficiency of drug development while maintaining high standards of safety and efficacy.

Guiding principles of good AI practice in drug development, EMA/FDA (2026)