

Human liver models as regulatory NAMs: what it takes to move from biological realism to decision-ready evidence

Human 3D liver models can reveal toxicities that simplified systems miss, but regulatory utility depends on context of use, reproducibility, technical characterization and a credible link between model output and the decision being made.

PharmaToxAI · 2026-09-14

Human liver models are becoming substantially more sophisticated. Hepatocyte spheroids, multicellular organoids and microphysiological systems can now reproduce aspects of metabolism, cell-cell signaling and tissue-level injury that are difficult to capture in conventional two-dimensional cultures. At the same time, regulatory policy is changing: New Approach Methodologies (NAMs) are no longer discussed only as experimental alternatives to animal studies, but increasingly as data sources that may contribute directly to drug-development decisions.

The difficult part is connecting those two developments.

Biological complexity does not automatically produce regulatory credibility. A model may contain several human cell types, generate transcriptomic data and reproduce a plausible injury mechanism, yet still be poorly suited to a regulatory question if its context of use is vague, its performance cannot be reproduced, or its outputs have not been related to a decision threshold. Conversely, a narrower assay can be highly useful when its purpose, operating conditions and uncertainty are well characterized.

This distinction has become particularly important since the U.S. FDA issued its 2025 roadmap for reducing animal testing and, in March 2026, a draft guidance on general considerations for NAMs in drug development. The emerging regulatory problem is therefore less about whether a model is “advanced” and more about whether the evidence it produces is fit for purpose.

The regulatory shift is now operational

The FDA's *General Considerations for the Use of New Approach Methodologies in Drug Development*, issued as draft guidance in March 2026, provides a useful indication of how the agency is approaching this transition. The document does not endorse a particular organoid, chip or computational architecture. Instead, it establishes a validation framework centered on the intended use of the method, its human biological relevance, technical characterization and its ability to support the proposed regulatory decision.

That is an important constraint. NAM validation is not a generic certification exercise. The evidence needed to use a liver model for early compound ranking is not necessarily sufficient to use the same model to replace a specific nonclinical safety study in a regulatory submission.

FDA experience reported by Yao and colleagues from CDER's Office of New Drugs reinforces this point. NAMs have already appeared across regulatory interactions and submissions, but their roles are heterogeneous: mechanistic support, safety assessment, pharmacology, contextual evidence and, in some cases, alternatives to conventional tests. Regulatory acceptance therefore depends on what claim is being made from the data rather than on the technological label attached to the assay.

The FDA's April 2026 report on the first year of its roadmap makes the direction explicit. The agency established NAM coordination infrastructure, expanded pathways for qualification and issued guidance intended to reduce unnecessary animal studies. This is meaningful regulatory movement, but it should not be interpreted as blanket acceptance of non-animal methods. The burden shifts toward demonstrating that a proposed method answers the relevant question with sufficient reliability.

Why the liver is a demanding test case

Drug-induced liver injury (DILI) remains difficult to predict because "hepatotoxicity" is not a single biological process. Direct cellular injury, reactive metabolite formation, mitochondrial dysfunction, cholestasis, immune-mediated effects and indirect toxicity can produce overlapping clinical phenotypes. Some liabilities are concentration-dependent and relatively accessible to hepatocyte-based assays. Others depend on multicellular interactions, immune signaling, genetic susceptibility or exposure patterns that are difficult to reproduce experimentally.

This heterogeneity creates a basic modeling problem. Increasing model complexity can add biologically relevant mechanisms, but every additional component also introduces sources of variation. Cell provenance, differentiation state, extracellular matrix composition, culture duration, media, oxygenation and assay endpoints can all alter the observed phenotype.

A useful liver NAM therefore needs two forms of validity that are related but not identical.

Biological validity asks whether the system reproduces mechanisms relevant to human liver injury.

Decision validity asks whether its output can reliably change a defined development or regulatory decision.

The second is harder to demonstrate and is the one most likely to determine practical adoption.

A 2026 example: multicellular organoids reveal toxicity invisible to simpler systems

A recent study by Sun and colleagues in *Nature Communications* illustrates both the promise and the remaining validation problem. The authors constructed human embryonic stem cell-derived three-dimensional multi-lineage hepatic organoids containing five major hepatic cell populations: hepatocytes, cholangiocytes, endothelial cells, hepatic stellate cells and Kupffer cells.

The system was evaluated with a panel of hepatotoxic compounds and was designed specifically to capture indirect hepatotoxicity. In the reported dataset, the multi-lineage organoids identified 28 of 34 hepatotoxic drugs, corresponding to 82.4% sensitivity, with 75% specificity. Importantly, the study did more than report classification performance. It identified a toxicity mechanism for imipramine that depended on communication between hepatic stellate cells and hepatocytes.

Imipramine engaged TRKB in stellate cells and altered exosomal signaling. Stellate-cell-derived exosomes enriched in miR-34a-3p were transferred to hepatocytes, where suppression of XIAP was associated with caspase-3 activation and apoptosis. The injury was therefore not simply a direct effect of the drug on hepatocytes. Simpler hepatocyte-only systems would be structurally incapable of reproducing the full mechanism because the initiating cell population and intercellular signal are absent.

This is exactly the type of result that makes multicellular human models attractive for predictive toxicology: additional biological organization can reveal a liability that is invisible when the assay is restricted to the presumed target cell.

But the same study also shows why biological sophistication cannot be equated with regulatory readiness. A mechanistically persuasive result for one compound does not establish prospective predictive performance across the chemical space relevant to drug development. The reported sensitivity and specificity are encouraging, but they arise from a finite compound set and a specific experimental implementation. Broader

external validation, inter-laboratory reproducibility and predefined decision rules would be required before treating those values as general properties of the platform.

The model also lacks adaptive immune components, which the authors identify as a limitation for low-frequency immune-mediated DILI. This matters because a negative result cannot be interpreted independently of what biology the model contains. Applicability domain is therefore not only a computational concept. Experimental NAMs have domains too.

Experimental NAMs need an applicability domain

QSAR practitioners are accustomed to asking whether a query compound lies inside the chemical space represented by a training set. A similar discipline is useful for organoids and microphysiological systems.

For a liver model, the domain includes at least four dimensions.

First, there is **biological coverage**: which cell types, pathways and injury mechanisms are represented? A hepatocyte spheroid may be appropriate for metabolism-dependent direct toxicity while being poorly suited to an immune-mediated liability.

Second, there is **exposure coverage**: can the system reproduce the relevant concentration range, duration and metabolite profile? Nominal medium concentration is not equivalent to human tissue exposure, particularly for highly bound, unstable or extensively metabolized compounds.

Third, there is **technical coverage**: over what range of culture conditions, batches and operators does the assay remain reproducible?

Fourth, there is **decision coverage**: what conclusion is the assay actually validated to support? Hazard identification, rank ordering, mechanistic investigation and replacement of an in vivo study are different contexts of use.

A credible report should make these boundaries visible rather than hide them behind a single toxicity score.

Complexity should be earned by the question

There is a tendency in translational model development to treat increased biological complexity as an intrinsic improvement. That assumption should be tested rather than accepted.

A five-cell-type organoid may reproduce intercellular mechanisms that a monoculture cannot, but it also requires more extensive characterization. If the regulatory question concerns a well-understood hepatocyte-autonomous mechanism, the additional

complexity may contribute little while increasing variance and cost. If the question concerns indirect injury, inflammatory signaling or cholestatic responses, the same complexity may be essential.

This leads to a practical design principle: the model architecture should be justified by the failure modes of simpler systems.

The strongest case for a complex NAM is not that it looks more like an organ. It is that a specific feature of the model measurably improves prediction or mechanistic discrimination for a defined class of problems.

That improvement should ideally be demonstrated against comparators. For example, a multicellular liver model can be evaluated against 2D hepatocytes, hepatocyte spheroids and available clinical annotations using the same blinded compound panel. The relevant outcome is then incremental information: which liabilities become detectable, which false positives disappear, and which uncertainty remains unresolved?

What validation should look like

The FDA's 2026 draft framework points toward a validation strategy that is closer to analytical method development than to a one-time biological demonstration.

A decision-ready NAM should have a predefined context of use. The endpoint should be explicit, and the relationship between the measured signal and the regulatory interpretation should be documented. Technical performance should include repeatability, reproducibility, dynamic range, controls, sources of variability and failure criteria. Biological relevance should be supported by evidence that the system contains the pathways needed for the proposed use.

Prospective and blinded evaluation are particularly important. Retrospective compound selection can unintentionally favor well-characterized mechanisms or compounds used during model development. A stronger test locks the protocol and analysis before evaluating an external set of compounds.

Inter-laboratory transfer is another critical step. A platform that performs well only in the laboratory that developed it may still be scientifically valuable, but its use in distributed drug-development workflows will be limited. Standard operating procedures, reference compounds, acceptance criteria and quantitative quality-control metrics are therefore part of the technology, not administrative additions.

Uncertainty also needs to be reported at the level of individual predictions. Binary labels such as "hepatotoxic" and "non-hepatotoxic" conceal whether a result is close to a decision threshold, inconsistent across replicates or generated under conditions where the model has limited biological coverage.

Integration with computational toxicology

Human 3D models and computational methods are often presented as competing alternatives to animal testing. In practice, their strongest use may be complementary.

Computational models can prioritize compounds, identify structural alerts, estimate exposure and suggest mechanisms. PBPK models can translate in vitro concentrations into plausible human exposure scenarios. Transcriptomic or imaging models can extract high-dimensional phenotypes from organoid experiments. Experimental NAM data can then provide human-relevant evidence for mechanisms or liabilities that are difficult to infer from structure alone.

This creates a more useful workflow than simply replacing one assay with another:

chemical structure and formulation → exposure modeling → computational hazard hypotheses → targeted human NAM experiments → mechanistic and phenotypic readouts → uncertainty assessment → weight-of-evidence decision

The regulatory advantage is traceability. Each component answers a narrower question and can be challenged independently. A PBPK assumption can be revised without invalidating the organoid assay; a structural alert can be contradicted by experimental evidence; a negative organoid result can be qualified by the model's biological domain.

For PharmaToxAI, this is the relevant architecture for predictive toxicology. The objective is not to generate a larger number of predictions. It is to assemble evidence in a way that preserves provenance, model boundaries and uncertainty.

OECD guidance points in the same direction

The regulatory movement is not confined to drug development. The OECD's third edition of its *Guidance on Grouping of Chemicals*, published in December 2025, explicitly incorporates NAMs including adverse outcome pathways, omics, high-throughput screening and (Q)SAR into grouping and read-across strategies.

The document is particularly relevant because it treats mechanistic and computational information as evidence for substantiating similarity and characterizing uncertainty. That logic is transferable to pharmaceutical safety assessment: a NAM result becomes more useful when it contributes to a structured argument about why compounds, mechanisms or exposures should be considered comparable.

The OECD framework also reinforces an important point about evidence integration. No single method needs to reproduce an entire organism to be useful. A method needs to contribute reliable information to a defined inference.

What should be reported from a regulatory-oriented liver NAM

For a human liver model intended to support development decisions, a useful report should contain enough information to reconstruct both the experiment and the inference. At minimum, this means documenting cell source and differentiation state, culture architecture, relevant functional characterization, exposure conditions, analytical endpoints, controls, replicate structure, predefined acceptance criteria and the method used to convert raw measurements into a conclusion.

The report should also state what the model does not represent. Missing adaptive immunity, absent systemic metabolism, limited chronic exposure, immature cellular phenotypes or restricted donor diversity are not generic caveats; they determine which negative results can be trusted.

Where computational components are included, their versions, input data, applicability domains and uncertainty estimates should be preserved. If a machine-learning classifier is trained on imaging or omics outputs, validation must separate biological replicates appropriately. Randomly splitting correlated wells or images from the same experimental batch between training and test sets can produce excellent metrics while measuring batch recognition rather than generalization.

This is one area where predictive toxicology can borrow directly from mature QSAR practice: external validation and domain definition matter more than an impressive internal cross-validation score.

A realistic near-term role

Human liver NAMs are unlikely to replace every conventional nonclinical study through a single universal platform. The more plausible near-term trajectory is modular adoption.

Some methods will be used to eliminate studies that add little information for specific drug classes. Others will refine dose selection, investigate mechanisms, resolve species discrepancies or provide human-specific evidence alongside conventional studies. As confidence accumulates, narrowly defined contexts of use can expand.

The FDA's recent guidance is significant because it makes the validation problem explicit. Developers now have a clearer target: define the question, establish human relevance, characterize the method technically and demonstrate that its performance is adequate for that question.

For organoids and other complex 3D models, this may ultimately be more consequential than increasing biological sophistication alone. A reproducible model with a narrow, defensible claim can enter a decision process. A highly elaborate model without a defined claim remains a research system.

The next phase of predictive toxicology will therefore depend less on whether a platform can mimic an organ in general and more on whether it can generate evidence that changes a decision for the right reason.

References

1. U.S. Food and Drug Administration. **General Considerations for the Use of New Approach Methodologies in Drug Development.** Draft Guidance for Industry. March 2026. Docket FDA-2025-D-6131. [FDA guidance](#).
2. U.S. Food and Drug Administration. **FDA Achieves Year 1 Goals in Reducing Animal Testing in Drug Development.** April 20, 2026. [FDA](#).
3. Yao J, Peretz J, Bebenek I, et al. **FDA/CDER/OND Experience With New Approach Methodologies (NAMs).** *International Journal of Toxicology*. 2026;45(2):136–156. doi:10.1177/10915818251384270. [Article](#).
4. Sun L, Zhang Y, Niu Y, et al. **Multi-lineage hepatic organoids reveal toxic exosome mediated indirect hepatotoxicity.** *Nature Communications*. 2026;17:2926. doi:10.1038/s41467-026-69548-0. [Article](#).
5. OECD. **Guidance on Grouping of Chemicals, Third Edition.** OECD Series on Testing and Assessment, No. 418. OECD Publishing, Paris; 2025. doi:10.1787/b254a158-en. [OECD](#).

Online edition: <https://pharmatoxai.ar/whitepapers/human-liver-models-nams-regulatory-readiness/>