2026 Field Guide to Scientific Operating Systems for Longitudinal Disease Research: Architectures, Workflows, Validation Frameworks, and Emerging Directions in ML-Enabled Biomedical Research

ARA S. KHACHATURIAN[1,2], NICOLAS P. PRONK[3], SUDHIR SIVAKUMARAN[4], JACQUES TOUCHON[5], MICHAEL W. LUTZ[6], JAKUB HORT[7], ORNIT CHIBA-FALEK[6], IRA R.J. HEBOLD HARALDSEN[8], STEPHEN Z. LEVINE[9], VÁCLAV SNÁŠEL[2], ZAVEN S. KHACHATURIAN[2,10], JEAN-MARIE BOUTEILLER[1,2,11]

Abstract / Executive Summary

Decades of biomedical investment have produced vast repositories of imaging, molecular, genomic, and longitudinal clinical data. Yet the field's capacity to translate this abundance into reliable predictions of individual disease trajectories remains deeply constrained. The principal obstacle is no longer one of data quantity; it is increasingly one of scientific architecture. Heterogeneous data streams remain fragmented across modalities and institutions, analytical workflows lack provenance and reproducibility, and most predictive models are evaluated in isolation rather than as components of a continuously learning scientific system.

The central scientific problem is therefore not simply the lack of data or algorithms. It is the absence of analytical frameworks capable of integrating and evaluating evidence across multiple biological scales, temporal horizons, and representational modalities while preserving scientific meaning.

This field guide examines how multimodal longitudinal disease modeling may require more than advances in AI models and algorithms. It may depend upon the deliberate construction of scientific operating systems: integrated systems spanning data, models, workflows, validation, governance, and regulatory alignment capable of ingesting heterogeneous data, preserving evidence provenance, supporting structured biological reasoning, and continuously refining predictions as new knowledge accrues. Rather than proposing a single implementation, this field guide surveys the emerging components of such a framework. The convergence of large multimodal foundation models, knowledge graph architectures, neuro-symbolic reasoning, and agentic AI systems now makes this aspiration technically plausible, but only if it is pursued as a coordinated systems-level effort rather than a collection of isolated modeling advances.

Six interdependent themes emerge repeatedly across the literature and current practice:

1. Data as scientific infrastructure. Dataset engineering, harmonization, annotation, and provenance tracking are not preparatory chores but first-class scientific activities.

2. Validation as a system-level property. Evaluation frameworks must incorporate temporal prediction, calibrated uncertainty, and real-world robustness rather than relying solely on conventional static performance metrics.

3. Modular and interoperable AI architectures. Foundation models, knowledge graphs, and agentic workflows should function as coordinated systems rather than isolated tools.

4. Workflow infrastructure as a scientific asset. AI-enabled pipelines should function as auditable, reproducible computational laboratory notebooks.

5. Precompetitive collaboration. Federated data access, shared benchmarks, and community-governed standards are essential for cumulative scientific progress across institutions.

6. Regulatory alignment by design. Regulatory frameworks, including FDA and EMA guidance, should shape traceability, lifecycle management, uncertainty quantification, and real-world validation from the outset.

The corresponding engineering challenge is the coordinated construction of the scientific infrastructure through which such frameworks become reliable, reproducible, and clinically actionable. Whether this opportunity is realized will depend not on any single breakthrough,

VM&E 2026;9:1-27
Published online October 21, 2026 http://dx.doi.org/10.14283/VME.2026.1

1. Brain Watch Coalition, Las Vegas, Nevada, USA
2. Center for Artificial Intelligence and Quantum Computing in System Brain Research, International Neurodegenerative Disorders Research Center, Praha, Czech Republic
3. HealthPartners Institute, Bloomington, Minnesota, USA
4. Lewy Body Dementia Association, Inc., Atlanta, Georgia, USA
5. Montpellier School of Medicine, University of Montpellier, Montpellier, France
6. Duke University School of Medicine, Durham, North Carolina, USA
7. Department of Neurology, Charles University, Second Medical Faculty and Motol and Homolka University Hospital, Praha, Czech Republic
8. Oslo University Hospital, Department of Neurology, Oslo, Norway
9. School of Public Health, University of Haifa, Haifa, Israel
10. President, Campaign to Prevent Alzheimer's Disease, Potomac, Maryland, USA
11. Institute for Technology and Medical Systems & Center for Neural Engineering, Alfred E. Mann Department of Biomedical Engineering, Viterbi School of Engineering & Keck School of Medicine, University of Southern California, Los Angeles, California, USA

Corresponding to:
Ara S. Khachaturian, 9812 Falls Road Suite 114-155, Potomac, MD 20854-3963,
Email: ara@pad2020.org;
http://www.brainwatchcoalition.org
http://www.pad2020.org

Received July 27, 2026
Accepted for publication August 28, 2026

but on the collective ability to build durable, interoperable, regulatory-aligned scientific systems through which predictive medicine becomes possible.

Scope and Intended Use

This document is intended as a periodically updated field guide for researchers, clinicians, engineers, and policy stakeholders working at the intersection of machine learning and longitudinal disease research. Its purpose is to survey emerging architectures, methods, infrastructure requirements, validation frameworks, and implementation challenges across a rapidly evolving field.

The document does not propose a single preferred architecture, technology stack, regulatory pathway, or commercial model. References to scientific operating systems, multimodal AI systems, foundation models, knowledge graphs, agentic workflows, and related concepts are intended to describe and organize the current landscape rather than establish standards of practice or make claims regarding superiority, readiness, or future adoption. Readers are encouraged to view this publication as a living resource that will evolve as technologies, evidence, and operational experience mature.

Introduction: The Need for ML-Driven Predictive Disease Modeling

The biomedical community now possesses decades of heterogeneous data spanning imaging, molecular biomarkers, genomics, pathology, digital measures, and longitudinal clinical trajectories (1-4). Yet despite this abundance of information, the field still struggles to identify early signals that reliably predict long-term disease progression (5-8). This gap is not simply a consequence of insufficient data accumulation; it reflects persistent challenges in integrating heterogeneous observations across modalities, scales, and time (9-12).

As disease datasets become increasingly multimodal, longitudinal, and distributed, the challenge shifts from model development alone to the management of the broader scientific lifecycle in which models are created, validated, deployed, and updated.

This shift reflects a deeper scientific challenge. The central problem is no longer simply acquiring more data or developing more sophisticated algorithms. It is the absence of analytical frameworks capable of integrating and evaluating evidence across multiple biological scales, temporal horizons, and representational modalities while preserving scientific meaning. Disease models increasingly draw upon observations that differ in biological interpretation, temporal resolution, evidentiary strength, and clinical context. The challenge is not simply to combine these heterogeneous sources. Instead, it is to organize them into coherent scientific representations that can support reliable prediction, biological interpretation, and continuous learning.

Artificial intelligence offers an important opportunity, but not merely as a better prediction engine. Its greater promise lies in enabling scientific operating systems for longitudinal disease modeling: an integrated, governed infrastructure layer that connects multimodal data ingestion, harmonization, provenance tracking, model development, knowledge representation, workflow orchestration, validation, regulatory documentation, and post-deployment monitoring into a continuous and reproducible lifecycle for developing, validating, and updating disease trajectory models (13-16).

For the purposes of this field guide, a Scientific Operating System (SOS) may be defined as an integrated framework that coordinates data, knowledge, models, workflows, validation, governance, and regulatory processes into a continuously learning scientific lifecycle. Rather than functioning as a single software platform, a scientific operating system serves as an organizational and computational architecture through which scientific evidence is assembled, evaluated, updated, and translated into actionable knowledge.

Such systems can be conceptualized as a set of interdependent layers, each corresponding to a distinct scientific function:

Rather than treating these components as separate activities, such systems organize them into a coordinated framework in which data, models, and scientific reasoning evolve together over time. This perspective shifts the focus from isolated predictive models to the infrastructure required to make those models more reliable, interpretable, reproducible, and continuously improvable.

Layer	Purpose	Concrete artifacts
Data layer	Multimodal ingestion, harmonization, versioning	Data dictionaries, data contracts, cohort definitions, and missingness maps
Knowledge layer	Biological and clinical semantics	Ontologies, knowledge graphs, causal assumptions, evidence links
Modeling layer	Predictive and representational models	Foundation models, temporal models, survival models, graph models
Workflow layer	Research automation and auditability	Pipelines, agent logs, notebooks, provenance graphs
Validation layer	Scientific and regulatory credibility	Calibration reports, external validation, subgroup analysis, uncertainty estimates
Governance layer	Control, compliance, access, accountability	Model cards, datasheets, audit logs, change-control plans, COU documentation

Table 1. Representative interdependent layers of scientific operating systems

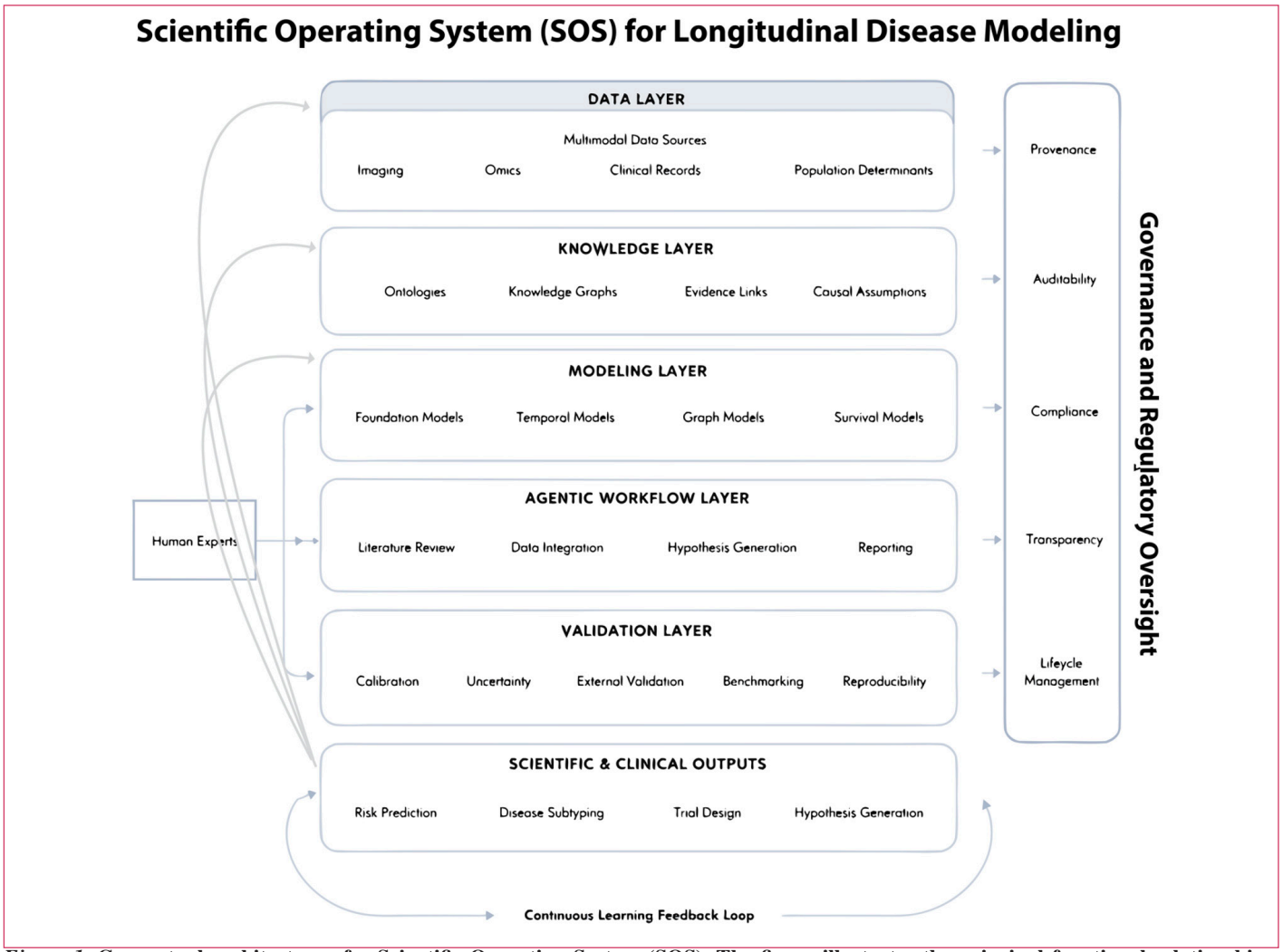


Figure 1. **Conceptual architecture of a Scientific Operating System (SOS). The figure illustrates the principal functional relationships among the major components of a Scientific Operating System. It is intended as an architectural framework rather than an operational workflow or data-processing pipeline. The connections represent three broad categories of interaction: (1) governance and oversight, (2) human supervision and intervention, and (3) continuous learning and refinement. The architecture illustrates how these functions interact within a continuously learning scientific ecosystem without prescribing a specific implementation or sequence of operations**

One motivating scientific challenge is determining whether signals observed across multiple biological scales and time horizons can be integrated to improve understanding and prediction of long-term disease trajectories (17-20). Neurodegenerative disorders such as Alzheimer's disease provide a particularly well-studied example of this challenge, but the underlying problem extends to many chronic progressive conditions in which pathology may evolve for years or decades before clinical recognition (21-24).

The Scientific Problem: Multimodal Longitudinal Disease Modeling Across Multiple Biological Scales

Heterogeneity of Biomedical Data Modalities

A central challenge in disease modeling is that relevant data are heterogeneous not only in format, but in biological meaning, scale, and evidentiary role (3, 4, 9, 25). This heterogeneity spans both individual-centered measurements and population-level determinants and reflects different levels of biological and contextual (26-28).

Individual-Centered Multimodal Data

At the individual level, disease is observed through a diverse set of modalities, including imaging data (e.g., MRI and PET), molecular biomarkers, genomics and transcriptomics, digital biomarkers derived from wearables, and longitudinal clinical records (4, 6, 29, 30). Each modality provides a partial and often indirect view of the underlying biological state (31, 32). Imaging may reflect structural or functional changes, molecular markers may capture biochemical processes, and clinical data encode downstream manifestations of disease (33, 34).

These data are heterogeneous not only in format, but in resolution, noise characteristics, and temporal sampling (35-37). More importantly, they do not align naturally: signals from one modality may precede, lag, or contradict signals from another (8, 33). An important scientific challenge is therefore not simply to combine these data sources, but to determine how they relate to one another over time—how signals interact, constrain interpretation, and collectively define evolving disease states (44, 45).

Population-Level and Contextual Determinants

In parallel, disease processes are shaped by factors that operate at the population level, including environmental exposures, socioeconomic conditions, behavioral patterns, public policy, culture, and broader social determinants of health (27, 28, 38, 39). These variables are often not tied to individual-level biological measurements in a straightforward way, yet they exert significant influence on disease risk, progression, and heterogeneity (39).

The many interactions among the multiple factors that make up social and structural determinants are complex, interdependent, interactive, dynamic, and multilevel, a fact sometimes obscured when defining these terms separately. Hence, multilevel frameworks may be required to integrate observations across individual and population levels, as well as across temporal and generational scales (26, 38, 40).

Unlike individual-centered data, these determinants are typically coarse-grained, longitudinally diffuse, and context-dependent (41). They introduce additional layers of variability that may modulate or even dominate biological signals in certain settings. Incorporating them into disease models requires frameworks that can link individual trajectories to population-level structure without collapsing one into the other (12, 42).

Implications for Modeling

The coexistence of individual-level biological data and population-level contextual drivers creates a fundamentally multiscale modeling problem (9, 25, 43, 44). Effective disease models must integrate these layers in a way that preserves their distinct meanings while enabling interaction across scales (12, 44). This requires not only technical integration, but conceptual clarity about the role each type of data plays in explaining and predicting disease trajectories.

Temporal Complexity of Disease

Disease processes unfold across multiple, interacting timescales, ranging from rapid physiological fluctuations to molecular changes that emerge over months or years, to clinical trajectories that develop over decades (23, 24, 45, 46). These temporal layers are not independent. Instead, they are coupled: short-term dynamics may accumulate into longer-term change, while slow-moving processes may constrain or

modulate faster signals (32, 47).

At the level of individual-centered data, different modalities operate on distinct temporal regimes (35, 37). Molecular and cellular changes may precede detectable imaging abnormalities, which in turn may precede clinical symptoms (23, 24). Digital biomarkers may capture high-frequency fluctuations that reflect early instability, while clinical diagnoses often represent late-stage summaries of underlying processes (29).

Temporal complexity is further amplified when population-level and contextual factors are considered (39). Environmental exposures, behavioral patterns, and social determinants often exert their influence cumulatively over long periods (39). These effects may interact with individual-level biology in nonlinear ways(46, 47).

An important implication is that longitudinal disease modeling involves more than the aggregation of time-stamped observations (12, 46). It requires models that explicitly represent temporal relationships across scales, including lag structures, feedback loops, and phase transitions between disease states (30, 47).

Temporal integration is therefore not an auxiliary feature of disease modeling; it is one of its defining scientific requirements. Models that fail to account for multiscale temporal structure risk conflating cause and consequence, missing early predictive signals, and generating trajectories that lack biological and clinical validity.

The Predictive Challenge

The predictive challenge in longitudinal disease modeling arises from the need to integrate heterogeneous signals across both biological scales and temporal regimes in order to generate clinically meaningful forecasts (12, 48). It is not a single problem, but a set of interdependent objectives that collectively define the utility of a disease model.

First, many disease-modeling efforts seek to enable earlier detection of risk, identifying signals that precede overt clinical manifestation by years or even decades. Achieving this requires linking subtle, often noisy patterns in individual-centered data—such as molecular changes, imaging features, or digital biomarkers—to long-term outcomes. These early signals are typically embedded within complex temporal dynamics and must be distinguished from background variation, aging-related change, or transient fluctuations (49, 50).

Second, models must predict divergent disease trajectories over time, reflecting interactions between biological mechanisms and environmental factors. Individuals with similar baseline profiles may follow markedly different paths, reflecting interactions between biological mechanisms, environmental exposures, and contextual factors. Capturing this heterogeneity requires models that can represent how multiscale influences evolve and interact across time, rather than assuming a single, average progression pattern (8).

Third, disease-modeling frameworks are often used to explore biologically meaningful subtypes associated with distinct trajectories and underlying mechanisms. These subtypes are not merely static clusters, but often correspond to distinct trajectories shaped by underlying mechanisms and modulated by population-level determinants (8). As such, subtype identification must be integrated with temporal modeling, allowing categories to emerge from patterns of progression rather than being imposed a priori.

These three objectives—early detection, trajectory prediction, and subtype identification—are deeply interconnected and require models that represent disease as a dynamic process rather than a static classification problem (18). Early signals may only be interpretable in the context of the trajectories they predict, while meaningful subtypes may only become apparent through longitudinal patterns (51). A model that excels in one dimension but neglects the other risks being scientifically incomplete or clinically limited.

The overarching challenge, therefore, is to develop models that can link early, multiscale signals to future disease pathways in a temporally coherent manner. Such models seek to move beyond static prediction toward representations of disease as a dynamic process. In principle, these representations may support improved forecasting and generate hypotheses regarding disease progression, heterogeneity, resilience, and potential intervention opportunities.

Emerging AI Architectures for Multimodal Modeling

No single AI paradigm is sufficient for multimodal longitudinal disease modeling. Deep learning systems are powerful pattern recognizers but are often opaque and weakly constrained by

biological knowledge. Symbolic systems offer interpretability but lack statistical scalability. Agentic systems can accelerate workflows but introduce risks related to trust and validation. Progress is increasingly likely to depend not on identifying a single dominant architecture, but on combining complementary capabilities within integrated and interoperable systems.

Large Multimodal Foundation Models

Large multimodal foundation models are designed to learn from diverse data types, including imaging, clinical tabular data, omics, and unstructured text such as electronic health records and scientific literature, and extract shared latent representations. Their central promise lies in their ability to learn shared latent representations that capture subtle, nonlinear relationships across modalities without requiring each interaction to be explicitly specified in advance (52, 53). These models retain the capacity to detect subtle, nonlinear patterns that would be difficult to uncover through conventional analytical pipelines, although their effectiveness depends strongly on the quality, scale, and representativeness of the underlying data (54, 55).

These systems rely on transformer architectures and attention mechanisms to model cross-modal relationships. Different fusion strategies can be employed, ranging from early fusion of raw inputs to late fusion of modality-specific embeddings, as well as cross-attention mechanisms that explicitly model interactions between modalities. Temporal extensions, including time-aware transformers and sequence models operating over patient trajectories, are essential for capturing the longitudinal structure of disease (56, 57).

The strengths of these models include their ability to capture complex nonlinear relationships, scale effectively with increasing data and computational resources, and support transfer learning across datasets. They are particularly well suited for end-to-end representation learning. However, these advantages are counterbalanced by important limitations, including their reliance on large, harmonized datasets, limited interpretability, weak incorporation of prior biological knowledge, and sensitivity to biases in the training data (58, 59).

In the context of disease modeling, one potential contribution of these approaches is the identification of latent representations that may capture patterns not directly observable within any single modality. These representations then may support the prediction of future trajectories, the identification of early predictive signatures, and the discovery of transitional states between health, risk, and overt disease. Their value is often greatest when coupled with systems that can interpret and validate the patterns they uncover.

Knowledge Graph and Symbolic AI Approaches

Knowledge graph and symbolic approaches address the complementary challenge of representing biomedical knowledge in forms that are explicit, interpretable, and amenable to reasoning. Knowledge graphs encode biomedical entities and relationships in structured form, enabling reasoning and hypothesis generation (60-62). They are particularly useful for incorporating prior knowledge, representing biological relationships, and providing structured contexts in which causal hypotheses may be explored (63, 64). In these systems, biomedical entities such as genes, proteins, diseases, biomarkers, and patients are represented as nodes within a graph, while relationships—causal, associative, or temporal—are encoded as edges. This structured representation enables the organization of scientific knowledge in a way that supports querying, hypothesis generation, and reproducible interpretation.

These approaches typically rely on components such as ontologies that define hierarchical relationships between entities, knowledge graphs that encode structured relationships, logical reasoning engines that support inference, and query systems that enable exploration of the knowledge base. Their primary strengths lie in their interpretability, transparency, and ability to incorporate prior biological knowledge, particularly from the scientific literature. They are well suited for supporting structured reasoning and mechanistic hypothesis generation.

However, symbolic systems face challenges in scaling to high-dimensional raw data and often require substantial effort for curation, although this process is increasingly being augmented by AI. Their limited ability to capture complex statistical patterns restricts their standalone utility for predictive modeling. As a result, their greatest value lies in complementing data-driven approaches by providing structure, context, and interpretability.

Hybrid Neuro-Symbolic Architectures

Neuro-symbolic systems integrate statistical learning with structured reasoning to improve interpretability and generalization. They are particularly useful in biomedical settings where prior knowledge is essential. Hybrid neuro-symbolic architectures aim to bridge the gap between statistical learning and knowledge-driven reasoning. These systems are motivated by the recognition that many biomedical questions cannot be addressed solely through correlations extracted from data, nor can biomedical knowledge be maintained solely through manual curation. Instead, they seek to integrate the strengths of both paradigms (65, 66).

In practice, such systems may operate by allowing neural models to generate embeddings that are mapped into knowledge graphs, or by using structured knowledge to constrain or regularize neural learning. Iterative feedback loops are often employed, in which models propose hypotheses that are then evaluated, refined, or rejected using symbolic reasoning systems.

The advantages of this approach include improved interpretability, better generalization in settings with limited or noisy data, and the ability to incorporate biological priors into the modeling process. At the same time, these systems introduce significant complexity, require high-quality structured knowledge, and lack standardized implementation frameworks. Despite these challenges, they are particularly well suited for longitudinal disease modeling, where the goal is to link temporal statistical patterns with underlying biological mechanisms.

Agentic AI Systems for Research Workflows

Recent advances in agentic AI have renewed interest in treating research workflows, rather than individual models alone, as units of innovation and optimization. These systems consist of autonomous or semi-autonomous agents capable of performing a wide range of tasks and orchestrating complex workflows, including reading scientific literature, extracting structured knowledge, writing and executing code, running analyses, evaluating results and generating hypotheses (67-69).

Such systems are typically built around large language model–based agents that handle reasoning and planning, combined with tool-use capabilities that allow interaction with external resources such as APIs, databases, and computational environments. They also incorporate memory systems, including knowledge graphs and vector stores, and workflow orchestration mechanisms that enable multi-step pipelines. A common pattern is the implementation of structured workflows that move through stages such as source identification, data ingestion, analysis, and reporting.

The primary strength of agentic systems lies in their ability to dramatically accelerate research cycles, improve reproducibility through structured and auditable workflows, and enable large-scale integration of knowledge and data. They also facilitate new forms of human–AI collaboration. However, they introduce risks related to error propagation across automated pipelines, dependence on underlying foundation models, and challenges in ensuring trust, transparency, and validation(70, 71).

In disease modeling, agentic systems may support activities such as multimodal data assembly, hypothesis generation, workflow orchestration, and model evaluation. Their significance lies not only in automation, but in their potential to reshape how scientific work is conducted.

Comparative Perspective: Why No Single Architecture May Be Sufficient

Each architectural paradigm offers distinct strengths and limitations. Multimodal foundation models excel at large-scale pattern discovery but are often opaque and data-intensive. Knowledge graph approaches provide interpretability and support reasoning but are difficult to scale. Neuro-symbolic systems offer a balance between

Architecture Type	Strength	Weakness	Best Role
Multimodal Foundation Models	Pattern discovery at scale	Opaque, data-hungry	Representation learning
Knowledge Graphs	Interpretability, reasoning	Limited scalability	Mechanistic context
Neuro-Symbolic Systems	Balance of both	Complexity	Robust modeling
Agentic AI Systems	Workflow automation	Trust/validation	Research acceleration

Table 2. Different AI architectures offer distinct strengths and limitations

these approaches but introduce integration complexity. Agentic systems enable workflow automation and research acceleration but raise concerns about validation and trust.

The central implication is that no single architecture currently appears capable of meeting all the scientific and technical requirements of multimodal longitudinal disease modeling. Many emerging systems combine these approaches in a coordinated and modular manner. These differences can be summarized as follows (Table 2):

This comparison underscores the need for integrated, rather than singular, architectural approaches.

Toward an Integrated Scientific Operating System

One emerging direction is the development of integrated scientific operating systems in which different AI components serve distinct but interoperable roles. In such a system, foundation models can learn distributed representations from heterogeneous data, knowledge graphs can encode domain structure and biological relationships, and agentic systems can orchestrate workflows for data integration, modeling, and validation. Human experts remain essential for framing scientific questions, interpreting results, and adjudicating uncertainty.

This systems-level perspective is consistent with emerging views of AI as infrastructure rather than isolated tools (7, 14, 71). The strategic objective is therefore not architectural purity, but modular interoperability and coordination.

Key Takeaways for the Section

Multimodal longitudinal disease modeling appears to be a multi-paradigm problem that is unlikely to be addressed by any single class of models. Pure deep learning approaches may be insufficient for many disease-modeling tasks without complementary sources of structured knowledge, validation, or domain expertise. Agentic AI introduces a new dimension by enabling the automation of scientific workflows. Current trends suggest increasing interest in modular and interoperable AI ecosystems in which different components complement one another. An important area of current exploration is how improved integration among models, data resources, workflows, and scientific processes may complement advances in model scale and capability.

From Models to Workflows: AI as Scientific Infrastructure

AI has the potential to reshape not only individual modeling tasks, but the entire structure of the scientific research pipeline. Its significance lies in enabling integrated infrastructure that connects literature, data, computation, reasoning, validation, and reporting into continuous workflows.

AI-Enabled Knowledge Curation

Agentic AI systems can ingest scientific literature and extract structured knowledge, transforming publications into machine-readable resources that can be incorporated into computational workflows (72-75). This represents a significant shift in the role of scientific literature. Traditionally, publications have served as narrative outputs intended for human interpretation. With AI-enabled curation, the literature itself becomes a machine-readable substrate that can be directly incorporated into computational workflows.

This transformation allows scientific knowledge to be continuously updated, systematically organized, and explicitly linked to supporting evidence. As a result, literature curation evolves from a manual, episodic activity into an ongoing, computationally supported process.

Structured Scientific Reasoning Pipelines

AI-enabled systems can be conceptualized as structured reasoning pipelines, progressing from data ingestion to analysis and reporting (76). These pipelines typically begin with the identification of relevant sources, followed by the ingestion and integration of data. Subsequent stages involve extracting relationships and patterns, conducting computational analyses, and generating reports or visualizations (15, 71).

The distinguishing feature of this approach is not merely task automation, but the explicit structuring of the scientific process itself. Each stage contributes to an explicit chain of reasoning that can be inspected, challenged, validated, and reproduced. This structure is essential for maintaining scientific rigor as workflows become increasingly automated, enabling reproducibility and auditability of scientific reasoning (77).

AI as a "Computational Laboratory Notebook"

When combined with structured data representations and workflow systems, AI-enabled workflow systems can function as computational laboratory notebooks. In this role, they record not only results, but the full sequence of steps that produced them, including data sources, transformations, model configurations, and intermediate outputs (78-80).

This capability has important implications for reproducibility and collaboration. By preserving provenance and enabling transparent inspection of analytical processes, such systems support more reliable validation and facilitate shared scientific understanding across teams. In biomedical research, where conclusions often depend on complex chains of evidence, this level of traceability is essential for establishing trust and credibility (80-82).

These capabilities do not replace scientific expertise; rather, they provide mechanisms for organizing, documenting, and evaluating increasingly complex scientific workflows.

Data as Scientific Infrastructure: Integration, Engineering, and Provenance

A recurring assumption in the field is that theA recurring assumption in the field is that the primary barrier to progress in multimodal disease modeling is algorithmic. Yet many of the obstacles limiting scientific progress arise well before model development begins. Data are fragmented across institutions, collected under different protocols, annotated according to inconsistent conventions, and observed unevenly across time. Metadata are incomplete, provenance is often poorly documented, and substantial effort is required to reconcile heterogeneous sources into scientifically coherent datasets. These bottlenecks are not primarily computational; they are scientific, organizational, and infrastructural.

Addressing them demands a fundamental reorientation, one in which data integration, engineering, harmonization, and provenance are recognized not as preparatory logistics, but as first-class scientific activities deserving the same rigor, expertise, and institutional investment as model development itself (14).

The Fragmentation Problem

Biomedical datasets relevant to longitudinal disease modeling are heterogeneous not only in format, but in meaning, evidentiary status, and temporal structure. Imaging data differ in acquisition protocols, reconstruction algorithms, and field strengths (4, 9). Molecular and genomic assays vary across platforms and preprocessing pipelines. Clinical records reflect inconsistent coding practices, missing observations, and evolving diagnostic criteria. Digital biomarker streams operate at timescales and resolutions that rarely align with clinical assessment windows. The result is not merely a data management challenge; it is a scientific one. Combining these streams without careful reconciliation introduces biases, obscures signals, and produces models that may generalize poorly across cohorts or institutions.

Cross-study harmonization compounds this challenge. Inconsistent ontologies, missing or non-standardized metadata, and incompatible cohort structures make direct integration unreliable. Temporal misalignment, particularly when datasets were not designed with longitudinal modeling in mind, further limits the biological interpretability of integrated analyses (83). These problems are not solved by acquiring more data; they require deliberate investment in harmonization infrastructure, shared standards, and iterative reconciliation workflows (84).

Integration as a Continuous, AI-Assisted Process

Conventional approaches tend to treat data harmonization as a one-time preprocessing step performed prior to modeling. This framing is insufficient for multimodal, longitudinal science, where both data and knowledge evolve over time (76). A more appropriate paradigm is to treat integration as a continuous and iterative process rather than a one-time step (14).

AI systems may play an important role in enabling this shift. They can support automated schema mapping across heterogeneous datasets, assist in aligning or generating ontologies, and facilitate cross-modal and cross-cohort entity resolution (85). More importantly, they enable the continuous reconciliation of datasets, metadata, and labels as new information becomes available. In this sense, the value of AI lies not only in automation, but in transforming integration into a dynamic and adaptive component of the scientific workflow.

Dataset Engineering as a Scientific Discipline

If data quality represents a major rate-limiting factor in disease modeling—as a growing body of evidence suggests—the scientific community must invest accordingly. Dataset engineering should therefore be recognized as a first-class scientific discipline rather than a technical preprocessing task.

This discipline encompasses a range of activities, including the annotation and enrichment of labels, the adjudication of ambiguous or conflicting data points, the detection and correction of errors, missing data handling, hard-negative identification, benchmark construction, the explicit representation of uncertainty, and the iterative refinement of cohort definitions (86). Reframing dataset engineering in this way shifts the focus of effort from purely model-centric optimization toward improving the fidelity, consistency, and scientific usability of the underlying data (14).

Annotation and label enrichment

Many longitudinal biomedical datasets suffer from sparse, inconsistent, or retrospectively applied labels. The challenge is not to collect more data, but to increase the information content of data that already exist. AI-enabled approaches, including weak supervision and label propagation, LLM-assisted extraction of structured variables from clinical narratives, and computer vision annotation of imaging features, can substantially expand the usability of existing datasets (87, 88). However, these approaches must themselves be validated: automatically generated labels carry their own error distributions, which must be characterized and managed rather than treated as ground truth (89).

Error detection and data auditing

Biomedical datasets routinely contain mislabeled samples, measurement artifacts, cohort inconsistencies, and cross-modal conflicts; for example, imaging findings inconsistent with clinical diagnosis. AI models can be deployed diagnostically as well as predictively: trained models can flag samples that strongly disagree with learned patterns, surface boundary cases for expert review, and identify systematic inconsistencies across data sources. This use of AI to audit and improve data quality, rather than simply to extract predictions from it, represents an underutilized and high-leverage application that should be treated as a standard component of modeling pipelines (90).

Missing data and temporal imputation

Longitudinal datasets are inherently incomplete. Participants drop out, assessments are missed, and modalities are not uniformly collected across cohorts or time points. Approaches including Gaussian process models, transformer-based temporal imputation, and cross-modal inference can partially address these gaps, but they carry a critical risk. When missingness is itself biologically or clinically informative (e.g., when cognitive decline drives attrition), imputation models may introduce systematic distortion. Missing data strategies must therefore be designed with explicit awareness of the mechanisms driving incompleteness, and imputed values should propagate uncertainty rather than assert false precision (91, 92).

Hard-negative mining and dataset refinement

Among the most scientifically valuable, and most underutilized, strategies in disease modeling is the deliberate identification and incorporation of hard negatives: individuals who resemble cases by available biomarker profiles but do not progress, or whose phenotypic boundaries are genuinely ambiguous. These samples are not noise; they sit at the real boundary of disease biology, and their inclusion can substantially sharpen model discrimination and early-signal detection. AI systems can identify samples near decision boundaries, detect potential mislabeling of apparent negatives, and prioritize cases for expert adjudication. Over multiple cycles, this creates a closed-loop refinement system in which model outputs improve dataset quality, and improved datasets in turn improve model performance, a virtuous loop that is potentially more impactful than many individual modeling advances (93).

AI-Enabled Feedback Loops for Dataset Improvement

The examples above point to a broader shift in how AI may contribute to longitudinal disease modeling. AI systems should not be viewed only as tools for generating predictions from finished datasets. They can also support the iterative improvement of the datasets themselves by identifying inconsistencies, surfacing uncertain cases, detecting missingness patterns, and

prioritizing samples for expert review.

In this framing, dataset construction and model development are not strictly sequential stages. Instead, they form a feedback loop. Models trained on existing data can identify cases that appear mislabeled, under-characterized, internally inconsistent, or unusually informative. These cases can then be reviewed by domain experts, reconciled against additional evidence, and incorporated into revised dataset versions. Updated datasets can then support improved models, which in turn can guide further refinement.

This feedback structure is especially important in longitudinal disease research, where labels may be uncertain, outcomes may evolve over time, and the boundary between risk, early disease, and overt clinical disease may be ambiguous. Rather than treating such ambiguity as noise to be removed, AI-enabled workflows can help make uncertainty visible and actionable. The goal is not to automate scientific judgment, but to focus expert attention on the cases, variables, and assumptions most likely to affect model validity.

Viewed in this way, AI-assisted curation becomes part of the scientific operating system itself: a mechanism for linking prediction, data quality, expert adjudication, provenance, and iterative learning. The value of such systems lies not simply in producing better models, but in creating datasets whose structure, assumptions, uncertainties, and revisions are more transparent over time.

Informative Samples and Iterative Refinement

In disease modeling, not all data points contribute equally to scientific insight. Particular value lies in samples that sit near the boundaries of disease definition, including individuals who resemble cases but do not progress, ambiguous or borderline phenotypes, and individuals who exhibit resilience despite high risk. These "hard negatives" and edge cases can sharpen model discrimination, improve early detection, and reveal biologically meaningful distinctions that are otherwise obscured.

AI systems can help identify such informative samples by detecting cases near decision boundaries or highlighting inconsistencies in labeling. This enables a closed-loop process of dataset refinement in which models are used to identify uncertain or misclassified cases, which are then re-evaluated by experts, incorporated back into the dataset, and used to retrain improved models. Over time, this iterative cycle leads to progressively higher-quality datasets, and more robust and generalizable models (93).

Benchmarking and Evaluation Ecosystems

Meaningful progress in the field requires benchmark datasets that reflect the true complexity of multimodal longitudinal disease modeling. Many existing benchmarks are limited in that they focus on a single modality, rely on cross-sectional data, or lack sufficient scale and diversity.

Next-generation benchmarks should integrate multiple data types, including imaging, omics, clinical records, and digital biomarkers, while capturing longitudinal trajectories over time (77). They should also incorporate standardized yet flexible outcome definitions and represent diverse populations to ensure generalizability.

A central challenge in this context is that ground truth is often uncertain and evolves over time, particularly in chronic and neurodegenerative diseases. Evaluation frameworks must therefore move beyond static labels and incorporate probabilistic representations, trajectory-based outcomes, and uncertainty-aware metrics. Measures such as time-to-event prediction accuracy, early detection performance, and calibration of uncertainty are likely to be more informative than traditional metrics such as accuracy or AUC alone (77, 94). Open benchmarking ecosystems, including shared evaluation platforms and reproducible pipelines, will be essential to support cumulative and transparent progress. The TRIPOD+AI reporting framework represents one emerging approach for transparent and reproducible reporting of prediction model performance (95).

Provenance, Metadata, and Reproducibility

Traceability of data and transformations is critical for scientific credibility. Robust data infrastructure requires that every data point, derived feature, and transformation be fully traceable. This includes not only the original source of the data, but also the sequence of preprocessing steps, transformations, and updates that it has undergone.

Machine-readable metadata standards and shared ontologies are critical for enabling interoperability across datasets and institutions

(78-80). In addition, reproducible data pipelines—supported by dataset versioning, containerized workflows, and transparent documentation—are essential for ensuring that results can be validated, reproduced, and extended (76, 81). These capabilities are not merely technical conveniences; they are foundational requirements for scientific credibility and trust. As AI systems increasingly contribute to data annotation, label generation, and dataset refinement, the provenance chain must extend to these AI-assisted transformations as well. A scientific operating system for longitudinal disease modeling must treat data lineage as a first-order design requirement, not an afterthought.

Key Takeaways

Across many disease-modeling applications, data quality, integration, engineering, and provenance appear to represent major constraints on progress, often rivaling or exceeding model complexity as limiting factors. Integration must therefore be treated as a continuous process rather than a one-time preprocessing step. Dataset engineering should be elevated to a core scientific discipline, supported by both human expertise and AI systems. Importantly, AI should be used not only to generate predictions, but also to improve data quality, expose uncertainty, and guide iterative refinement. The most effective systems will be those that tightly couple data curation, validation, and modeling in an ongoing feedback loop.

Continued advances in model architecture will remain important, but long-term progress is likely to depend equally on investments in scientific infrastructure, integration, engineering, benchmarking, and provenance.

Trust, Validation, and Reproducibility

AI systems can produce plausible but incorrect outputs, necessitating robust validation frameworks. Reliability is necessary in biomedical research and medical care, but is not sufficient (96). Biomedical AI therefore requires validation frameworks that extend beyond performance alone (97, 98).

This requirement becomes even more urgent as AI systems move from isolated analytical tasks toward end-to-end orchestration of literature review, data integration, model building, and reporting. The more capable the workflow, the greater the need for explicit validation and auditable control points.

Risks of AI-Generated Analysis

AI systems can generate analyses that are fluent, sophisticated, and apparently rigorous while still containing subtle conceptual, statistical, or implementation errors. In biomedical contexts, such errors are particularly concerning because they may not be immediately apparent, yet can propagate through downstream analyses leading to compounding failures, and create unwarranted confidence in incorrect conclusions (71).

As AI systems expand from isolated modeling tasks to orchestrating entire research workflows—including literature review, data integration, model development, and reporting—the scope and impact of these risks increase. Errors introduced at early stages of the pipeline may be amplified as they are incorporated into subsequent steps, especially when intermediate outputs are treated as reliable inputs (70). This creates the potential for compounding failure modes that are difficult to detect through conventional validation approaches (15, 99).

A further challenge is that AI-generated outputs are often optimized for coherence and apparent completeness rather than empirical correctness. As a result, systems may produce results that are internally consistent but not externally valid, particularly when operating at the limits of available data or outside their training distribution (100). In high-stakes biomedical settings, this combination of fluency and fallibility represents a significant risk, underscoring the need for validation frameworks that are both rigorous and deeply integrated into the workflow.

Verification Mechanisms

Validation must be embedded throughout the workflow, including uncertainty estimation and benchmarking, rather than appended after the fact. As AI systems take on larger portions of the scientific process, including data integration, model development, and analysis, the risk of error propagation increases (76). This makes it essential to embed verification mechanisms at multiple stages of the pipeline, ensuring that outputs are continuously evaluated rather than retrospectively checked (77).

One emerging approach is the use of multi-agent validation pipelines, in which different components of the system independently generate, review, and critique results. In such architectures, "worker" agents perform analyses while "checker" or "auditor" agents evaluate the validity of those outputs, helping to identify

inconsistencies, methodological errors, or unsupported conclusions (67). This introduces a form of internal redundancy that can improve robustness, particularly in complex, multi-step workflows (68).

In addition, uncertainty scoring provides a systematic way to assess the confidence of model outputs. By quantifying uncertainty at the level of predictions, models can signal when outputs are unreliable or fall outside the domain of the training data. This is particularly important in biomedical contexts, where overconfidence in incorrect predictions can have significant downstream consequences (94).

Benchmark datasets also play a critical role in verification by providing standardized reference points against which models can be evaluated (77). When combined with reproducible evaluation pipelines, benchmarks enable consistent comparison across models and ensure that performance claims are grounded in shared, transparent criteria.

Finally, human-in-the-loop oversight remains essential. While AI systems can automate many aspects of validation, expert judgment is required to interpret ambiguous cases, adjudicate conflicting evidence, and ensure that results are clinically and biologically meaningful (101). Effective verification frameworks therefore combine automated validation mechanisms with targeted human review, creating a layered approach to trust and reliability.

Provenance and Transparency

Trust in AI-enabled biomedical systems depends not only on the accuracy of their outputs, but also on the transparency of the processes that generate them. Provenance—the ability to trace results back through their underlying data sources, transformations, analytical steps, and model configurations—is therefore a foundational requirement for scientific credibility, reproducibility, and practical usability.

In complex AI-enabled workflows, provenance must extend across the entire analytical lifecycle, including data ingestion, preprocessing, model development, inference, validation, and reporting (78). Each stage introduces decisions and transformations that can influence downstream results. Without explicit traceability, it becomes difficult to diagnose errors, reproduce findings, assess reliability, or understand why a particular conclusion was reached (80). Provenance is therefore not simply a record-keeping exercise; it is an essential mechanism for understanding, evaluating, and validating the behavior of the system as a whole (79).

Achieving this level of traceability requires more than documentation alone. Structured metadata frameworks, versioned datasets, reproducible workflows, and machine-readable representations of data and model relationships provide the infrastructure necessary for end-to-end transparency (81). Knowledge graphs and related knowledge-management systems can further support these goals by linking data sources, analytical processes, assumptions, models, and outputs within a coherent and queryable framework.

Transparency also enables auditability and accountability. Systems should provide clear documentation of inputs, assumptions, model configurations, performance characteristics, and known limitations, allowing researchers, clinicians, regulators, and other stakeholders to evaluate the validity and applicability of the resulting outputs. This requirement becomes increasingly important as AI systems assume larger roles in biomedical research, clinical decision support, and drug development workflows.

Taken together, provenance and transparency transform AI systems from opaque computational tools into auditable scientific instruments. They support reproducibility, enable continuous validation, facilitate collaborative review, and provide mechanisms for integrating human judgment into increasingly automated workflows. As such, they are foundational components of any scientific operating system intended for high-stakes biomedical applications.

Key Takeaways

The central lesson is that validation cannot be treated as a discrete step applied after model development. Rather, it must be embedded throughout the scientific workflow, encompassing data quality, uncertainty quantification, provenance, transparency, reproducibility, and human oversight. In this sense, validation is not solely a property of individual models, but of the broader scientific system in which those models are developed, evaluated, deployed, and continuously updated.

As AI systems move from research environments into clinical and drug development settings, many of these scientific requirements increasingly intersect with formal regulatory

expectations. The following section examines how emerging regulatory frameworks seek to operationalize these principles in practice.

Regulatory Considerations and Clinical Translation

As disease models move from exploratory research into clinical and drug development contexts, validation requirements must become formalized within regulatory frameworks. In this context, AI systems are no longer evaluated solely on predictive performance, but on their ability to support reliable, auditable, and clinically meaningful use (48). Regulatory frameworks therefore represent not an external constraint, but a defining component of any scientific operating system intended for real-world biomedical impact (16, 102).

The validation requirements described in this section are not merely scientific best practices. For AI-driven disease models intended to support drug development, inform clinical trial design, or contribute to regulatory decision-making, they increasingly reflect formal regulatory expectations. The regulatory landscape governing AI in biomedicine has evolved rapidly, with agencies emphasizing risk-based evaluation, lifecycle management, and explicit specification of intended use (103). Research institutions and development teams that treat regulatory readiness as a downstream concern — something to address when a model is «ready» — face a structural disadvantage. Regulatory alignment must be designed into modeling workflows from the outset, not retrofitted after scientific conclusions are drawn.

Two frameworks are of particular relevance to multimodal longitudinal disease modeling: the frameworks governing AI/ML as components of drug development tools and clinical decision support, and the frameworks governing AI-based software as a medical device. Though distinct in their regulatory pathways, both share an emphasis on transparency, risk-proportionate validation, lifecycle management, and the explicit characterization of a model's intended context of use.

FDA Frameworks for AI in Drug Development

In January 2025, the FDA published draft guidance titled **Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products**— its first formal guidance specifically addressing AI use in drug development submissions (104). The guidance establishes a risk-based credibility assessment framework: the evidentiary standard required to establish confidence in an AI model's outputs is calibrated to the significance of the regulatory decision that those outputs inform. For disease progression models that contribute to trial design, endpoint selection, or patient stratification, this framework has direct implications. It is not sufficient to demonstrate that a model performs well on a held-out test set; sponsors must characterize the model's context of use, document its development and training data, demonstrate its performance under conditions representative of the intended application, and — critically — identify and disclose limitations that may affect the reliability of model-derived conclusions (16, 103).

For AI-enabled software functions that meet the definition of a medical device — including clinical decision support tools derived from disease models — the FDA's January 2025 draft guidance on **Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations** applies a Total Product Life Cycle (TPLC) approach (105). This framework requires documentation of model architecture and training data lineage, performance validation tied to specific clinical claims, bias assessment and mitigation strategies, and a plan for real-world performance monitoring following deployment. Institutions developing or deploying AI-derived disease models in clinical or trial contexts should treat these requirements not as external regulatory constraints, but as a design specification for trustworthy scientific infrastructure.

The Drug Development Tool Qualification Pathway: For research teams developing multimodal disease models with potential drug-development utility, the DDT pathway provides an important example of how AI-derived measurements may progress from exploratory tools toward formally qualified regulatory instruments. The evidentiary requirements are substantial, and qualification timelines have historically extended over multiple years.

Predetermined Change Control Plans: One of the most consequential regulatory developments for adaptive AI systems is the original issuance in December 2024 and update in August 2025 of the FDA's guidance on Predetermined Change

Control Plans (PCCPs) for AI-enabled device software functions (106). A PCCP allows developers to pre-specify, at the time of marketing submission, how an AI model may be updated post-deployment — including retraining on new data or expanding to new subpopulations — without requiring a full resubmission for each covered change. The plan must specify in advance the modification categories, the data and testing requirements that govern each change, performance acceptance criteria, and a monitoring and deployment protocol.

The PCCP framework is especially relevant to longitudinal disease models, which are expected to improve as new data accrue and whose performance characteristics may shift as patient populations, measurement technologies, or clinical standards evolve. Designing a disease modeling system with PCCP-compatible update governance — specifying in advance how the model will be retrained, what performance benchmarks must be maintained, and how users will be notified of version changes — is both a regulatory best practice and a scientific one. It encourages the level of specification that reproducible modeling requires.

EMA and International Regulatory Convergence

The European Medicines Agency published its final Reflection Paper on the Use of Artificial Intelligence in the Medicinal Product Lifecycle in September 2024, establishing expectations for AI transparency, traceability, and performance monitoring across the drug development continuum (107). The paper applies a risk-based approach calibrated to the regulatory impact of each AI application: tools that directly inform benefit-risk assessments, endpoint determination, or safety signals are subject to the highest evidentiary expectations, while tools used in exploratory or internal workflow contexts face lighter obligations.

Importantly, the FDA and EMA jointly published ten guiding principles for good AI practice in drug development in January 2026 — reflecting the outcome of sustained bilateral collaboration initiated at their April 2024 meeting (108). These principles articulate a shared framework emphasizing human-centric design, risk-proportionate validation, data governance, and transparent lifecycle management. Their joint publication signals a meaningful convergence in regulatory expectations across jurisdictions and may underpin future formal guidance from both agencies. For global drug development programs, designing AI-driven disease models against these shared principles — rather than against either jurisdiction's guidance in isolation — is both strategically and scientifically sound.

Additionally, the EU AI Act, which entered into force in August 2024 with phased application through 2026–2027, classifies AI systems used for clinical decision support and patient risk profiling as high-risk under its framework. High-risk AI systems are subject to mandatory conformity assessments, continuous performance monitoring, explainability requirements, and human oversight controls. For disease modeling systems deployed in European clinical or trial contexts, compliance with the EU AI Act's obligations will layer onto — and in some cases exceed — existing pharmaceutical regulatory requirements.

Model Calibration and Uncertainty as Regulatory Expectations

A consistent thread across FDA, EMA, and international frameworks is the treatment of model calibration and uncertainty quantification not as optional technical refinements, but as core evidentiary requirements. A model that produces well-calibrated probability estimates — in which stated confidence levels correspond to observed event rates — is far more defensible in a regulatory context than one optimized for classification accuracy alone (109). Poorly calibrated models may produce overconfident predictions that misrepresent risk, particularly when applied to populations or conditions that differ from the training data. Addressing this issue requires explicit evaluation of model performance across subgroups and under conditions of distribution shift (16, 58, 110).

For disease progression models, where predictions may inform consequential decisions about trial eligibility, intervention timing, or drug approval, the distinction between a model's apparent performance and its actual reliability under real-world conditions is a regulatory concern, not merely a scientific one.

Regulatory submissions involving AI-derived disease models should therefore include: explicit characterization of uncertainty in model predictions; calibration analyses across clinically relevant subgroups; temporal validation demonstrating that model performance holds across time periods not represented in training data; and assessments of performance under distribution shift — including across sites, cohorts, and patient populations that differ from those used for development. These requirements align directly with the scientific validation priorities described in the previous sections, underscoring that regulatory readiness and scientific rigor are not competing objectives. In many cases, the practices required to satisfy regulatory evidentiary standards also

Issuing Institution	Title and Description	Date
FDA	*Draft Guidance — AI in Drug Development* Considerations for the Use of AI to Support Regulatory Decision-Making for Drug and Biological Products. Risk-based credibility assessment framework.	January 2025
FDA	*Draft Guidance — AI-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations*. TPLC approach; bias, data lineage, monitoring requirements.	January 2025
FDA	*Final Guidance — PCCP* Marketing Submission Recommendations for a Predetermined Change Control Plan for AI-Enabled Device Software Functions; pre-specified post-market update governance.	December 2024
FDA	*DDT Qualification Program* 21st Century Cures Act, Section 507. Biomarker, COA, and ISTAND pathways. First AI-based DDT qualification (MASH histopathology, Dec. 2025).	December 2025
EMA	*Reflection Paper on AI* Use of Artificial Intelligence in the Medicinal Product Lifecycle. Risk-based approach; GxP alignment; transparency and traceability requirements.	September 2024
FDA-EMA	*Ten guiding principles for good AI practice in drug development* Shared framework underpinning future guidance from both agencies	January 2026

Table 3. Key regulatory instruments

strengthen scientific trustworthiness, reproducibility, and transparency.

Key Takeaways

The regulatory landscape for AI in drug development and clinical decision support has evolved rapidly since 2024. FDA, EMA, and international frameworks increasingly converge around a shared set of expectations: risk-proportionate validation, transparent lifecycle management, explicit context-of-use specification, uncertainty characterization, and continuous performance monitoring. These are not compliance hurdles to be addressed at the end of a modeling project; they are design requirements that influence how disease-modeling systems are conceived, developed, evaluated, and maintained.

More broadly, the convergence between scientific and regulatory expectations highlights a central theme of this field guide: trustworthy disease modeling depends not only on model performance, but on the infrastructure, workflows, governance mechanisms, and validation processes that surround the model. In this sense, regulatory alignment is not separate from the scientific operating system—it is one of the mechanisms through which scientific credibility is translated into real-world biomedical use.

Institutional Transformation and Scientific Control

Institutions that function primarily as consumers of external technologies may achieve short-term gains in efficiency, but can become increasingly dependent on external actors for standards, validation frameworks, workflow design, and technical infrastructure (14, 86). The scientific and technical vision described in this whitepaper — a modular, interoperable, continuously learning operating system for longitudinal disease modeling — cannot be built by any single institution. No organization, however well-resourced or scientifically distinguished, possesses the breadth of longitudinal data, the diversity of patient populations, the range of domain expertise, the computational infrastructure, and the governance legitimacy required to construct this ecosystem unilaterally. The history of biomedical science offers no successful precedent for solving problems of this scale and complexity in isolation.

Recognizing this reality is the necessary starting point for institutions deciding how to position themselves in relation to an emerging AI-enabled research infrastructure. That positioning decision is strategic, not cosmetic. Institutions that remain passive consumers of tools built elsewhere may gain short-term operational efficiency, but they surrender control over the standards, validation practices, data governance frameworks, and scientific priorities that will ultimately define the field.

Institutions that attempt to build everything internally risk duplicating effort, underinvesting in interoperability, and producing systems that cannot connect to the broader ecosystem upon which their value depends. The productive path lies between these extremes — and it requires coordinated, precompetitive engagement across institutional boundaries.

Why Precompetitive Collaboration Is Increasingly Important

Longitudinal disease modeling presents a fundamental collective action problem. The data required to build reliable, generalizable, and temporally validated predictive models cannot be generated by any single institution at the required scale, diversity, and follow-up duration. A single-site cohort, however carefully designed, reflects the demographics, clinical practices, measurement protocols, and disease ascertainment biases of that site, and models trained on such data generalize poorly (16). The history of biomarker discovery and disease staging in neurodegenerative and other chronic conditions is littered with findings that replicated within institutions and failed across them.

The same logic applies to benchmark construction, validation infrastructure, and ontology development. Shared evaluation frameworks require community agreement on outcome definitions, data standards, and performance criteria that no single institution can establish with the legitimacy needed for broad adoption. Precompetitive collaboration — in which institutions share infrastructure, standards, and scientific outputs without necessarily sharing competitive clinical programs or proprietary drug development assets — represents one of the few mechanisms capable of addressing these collective-action challenges at scale. Federated learning and distributed data networks enable institutions to collaborate on model development without sharing sensitive patient data, preserving privacy while enabling large-scale learning (84). Similarly, shared benchmarking ecosystems and reproducibility frameworks allow the field to converge on common standards for evaluation and validation (77).

Regulatory considerations reinforce this imperative. Emerging FDA frameworks for AI/ML-based tools used in drug development and clinical decision support increasingly emphasize generalizability, transparency, and post-deployment monitoring — properties that are most credibly demonstrated through multi-site, community-validated evidence. Institutions that help build the precompetitive infrastructure through which such evidence is generated will be better positioned to engage productively with evolving regulatory expectations.

What Precompetitive Infrastructure Looks Like in Practice

Precompetitive collaboration in this context is not abstract. It takes specific, investable forms. Federated data networks allow institutions to contribute to shared modeling efforts without transferring patient data — privacy-preserving federated learning and secure multi-party computation provide increasingly mature approaches to this challenge (84). Shared benchmark datasets and open evaluation platforms, modeled on initiatives such as ADNI, UK Biobank, and emerging challenge ecosystems, create the common ground on which algorithmic progress can be measured reliably (3).

Community-governed ontologies and metadata standards — developed through processes that include academic, clinical, industry, and regulatory stakeholders — provide the semantic interoperability that makes heterogeneous data mutually legible across institutions (80, 82).

Open or interoperable AI model registries, pre-trained on harmonized biomedical data and released under licensing terms that permit scientific use, can dramatically lower the barrier to high-quality modeling for institutions that lack the compute or data volume to train foundation models from scratch (53, 55). Shared validation frameworks — including agreed criteria for temporal validation, calibration assessment, and clinical utility evaluation — create accountability standards that benefit the entire field and support regulatory engagement (95).

Benchmarking Against Existing Biomedical AI and Precision Medicine Ecosystems

Existing initiatives already demonstrate many of the capabilities discussed throughout this field guide. Population-scale resources such as UK Biobank and the NIH All of Us Research Program illustrate the value of longitudinal, multimodal data infrastructure for disease modeling and precision medicine (3, 111). Research ecosystems such as EBRAINS and AI-Mind demonstrate approaches to interoperable scientific

infrastructure, federated collaboration, and large-scale integration of data, computational resources, and scientific workflows (112, 113).

At the same time, academic and industrial programs continue to advance foundation models, multimodal learning, knowledge representation, and AI-enabled scientific workflows (53, 55). Collectively, these efforts highlight the growing maturity of individual components required for longitudinal disease modeling.

The opportunity addressed by Scientific Operating Systems is not the creation of any single dataset, model, or platform. Rather, it is the coordination of these capabilities into an integrated, auditable, interoperable, and continuously learning scientific lifecycle. In this sense, Scientific Operating Systems should be viewed as complementary to existing biomedical AI and precision medicine initiatives, providing the architectural framework through which diverse resources, workflows, validation systems, and governance processes can function together as a coherent scientific ecosystem.

Strategic Positioning for Research Institutions

Within this precompetitive landscape, individual institutions face genuine strategic choices about where to invest and what kind of contributor to become. Research institutions may occupy different positions within the emerging ecosystem, and each involves real trade-offs.

An institution that positions itself primarily as a **consumer** of existing AI tools and data ecosystems minimizes upfront investment and captures near-term efficiency gains — but cedes influence over standards, workflows, and scientific direction to others. An institution that invests in becoming a **developer** of domain-specific AI systems, validated pipelines, or specialized data infrastructure acquires deep capability and potential scientific leadership — but risks building systems that remain isolated if interoperability is not designed in from the outset. A **collaborative builder** orientation — one in which an institution actively contributes to shared infrastructure, participates in governance of community standards, and selectively develops domain-specific capabilities that feed back into the common ecosystem — may provide a balance of scientific influence, ecosystem participation, and long-term adaptability.

The choice of orientation should be driven by an honest assessment of existing institutional assets: the depth and longitudinal completeness of available cohort data, the maturity of computational infrastructure, the presence of expertise spanning clinical, biological, and computational domains, and the institution's existing relationships with regulatory bodies, industry partners, and peer research organizations. Institutions with distinctive longitudinal datasets occupy a particularly strategic position: that data, contributed to shared federated networks under appropriate governance frameworks, may represent a more consequential scientific contribution than any model built on it in isolation (114).

Governance and Scientific Sovereignty

Precompetitive collaboration raises genuine governance challenges. Shared data infrastructure requires agreements on access, use limitations, attribution, and benefit sharing. Community-developed models raise questions about intellectual property, maintenance responsibility, and version governance. Federated networks require trust frameworks that span institutional, national, and regulatory boundaries. These are not insurmountable problems — the open science community, clinical trial consortia, and international biobank networks have developed workable models — but they require deliberate investment in governance design, not just scientific infrastructure.

A related concern is scientific sovereignty: the risk that dependence on a small number of large foundation model providers, proprietary cloud infrastructure, or commercially controlled data platforms creates structural vulnerabilities for publicly funded biomedical research (114). Open scientific AI infrastructure — including community-governed model repositories, open benchmarking platforms, and publicly accessible compute resources — is therefore not merely a matter of cost efficiency. It may become an important prerequisite for the kind of transparent, reproducible, and independently auditable science that the field's credibility depends on (80, 115).

More broadly, institutions seeking to contribute to the next generation of predictive medicine may find increasing value in investing in shared data standards, validation ecosystems, governance frameworks, and interoperable infrastructure alongside their own scientific programs. The scientific operating system envisioned in this whitepaper is, by definition, a collective construction. Its realization requires institutions to act not only as competitors for discovery, but

as co-architects of the systems through which discovery becomes possible.

Limitations, Risks, and Counterarguments

The argument for a scientific operating system for AI-driven disease modeling is deliberately ambitious. It proposes that reliable prediction of disease trajectories will require not only better models, but also shared data infrastructure, reproducible workflows, validation systems, governance mechanisms, and regulatory alignment. However, this vision also raises important limitations and counterarguments. A credible implementation strategy must address these concerns directly rather than treating infrastructure, automation, and standardization as unqualified goods.

Risk of over-centralization

A significant concern is that the operating system metaphor may imply excessive centralization. Biomedical science has historically advanced through methodological pluralism, local expertise, disease-specific specialization, and investigator-driven experimentation. If poorly designed, a single centralized infrastructure could constrain innovation, impose unsuitable standards, or foster dependency on a limited number of institutional or commercial platforms.

This risk supports adopting a modular, federated architecture rather than a monolithic system. The objective should not be to enforce a single model, database, or workflow across all disease areas. Instead, the operating system should establish interoperable interfaces, shared documentation standards, validation criteria, provenance requirements, and governance principles. Disease-specific communities should maintain the flexibility to develop specialized models, endpoints, ontologies, and workflows, provided their outputs remain auditable, reproducible, and comparable across different contexts.

Therefore, the most appropriate analogy is not a single closed platform, but rather an interoperable scientific ecosystem. This approach combines shared protocols and infrastructure, as needed, with local specialization when scientifically justified.

Risk of premature standardization

A second counterargument asserts that premature standardization may entrench immature methods before superior approaches are identified. This concern is particularly pertinent in rapidly evolving domains such as multimodal foundation models, agentic scientific workflows, neuro-symbolic reasoning, causal representation learning, and adaptive clinical AI systems. Early standards may privilege legacy methods, marginalize emerging techniques, or incentivize superficial compliance at the expense of genuine scientific validity.

This concern warrants careful consideration. Accordingly, the proposed operating system should differentiate between stable and evolving standards. Stable standards should govern fundamental scientific requirements, including data provenance, cohort definition, versioning, leakage assessment, validation reporting, uncertainty documentation, model-change tracking, and human accountability. In contrast, evolving standards should address model architectures, optimization methods, foundation model design, agentic workflow patterns, and adaptive learning approaches.

In summary, the system should standardize the evidence trail rather than prematurely standardize scientific methodologies. While innovation must remain unrestricted, the processes for making, testing, documenting, and updating claims should adhere to rigorous standards.

Cost, complexity, and operational burden

Developing a scientific operating system for disease modeling necessitates substantial investment. The costs associated with data harmonization, ontology mapping, privacy-preserving infrastructure, external validation networks, monitoring systems, audit logs, regulatory documentation, and multidisciplinary governance are considerable. For many academic groups, health systems, and small biotechnology companies, these requirements may appear burdensome or unattainable.

This concern is not trivial. Infrastructure-intensive scientific approaches may create barriers to entry if participation is limited to large institutions. Rather than lowering scientific standards, the field should prioritize developing shared infrastructure, reusable templates, open benchmarks, federated validation networks, and community-maintained tools. Public-private partnerships and precompetitive consortia are particularly important, as no single laboratory or company can efficiently construct all necessary components independently.

Accordingly, the operating system should not be depicted as a solution that each institution must construct independently. A more feasible objective is the creation of a layered ecosystem in which common components, metadata standards, validation protocols, benchmark datasets, documentation templates, workflow tools, and model-monitoring frameworks are reusable across diverse disease areas and institutional contexts.

Data sharing, privacy, and intellectual property constraints

The proposed approach relies on access to longitudinal, multimodal, and frequently sensitive biomedical data. However, these data are subject to constraints imposed by privacy laws, institutional policies, consent limitations, patient expectations, commercial interests, and cross-border data-transfer restrictions. Furthermore, organizations that invest significantly in data generation may be unwilling to share assets that provide scientific or commercial advantages.

These constraints are structural rather than incidental. Consequently, a credible operating system should support multiple collaboration models, such as centralized data sharing where feasible, federated learning when data mobility is restricted, secure research environments, privacy-preserving analytics, synthetic or partially de-identified benchmark datasets, and controlled-access evaluation servers. Additionally, it should establish clear guidelines for attribution, intellectual property, publication rights, benefit sharing, and downstream use of the model.

Precompetitive collaboration is most feasible when the shared layer is clearly defined. Institutions may collaborate on standards, validation protocols, benchmark tasks, metadata schemas, reference ontologies, and regulatory evidence frameworks, while maintaining competition in proprietary models, therapeutic hypotheses, clinical programs, and implementation strategies.

Risk of encoding bias and inequity

A scientific operating system alone does not ensure that artificial intelligence is fair, generalizable, or clinically beneficial. When underlying data reflect unequal access to care, underrepresentation of specific populations, biased diagnostic practices, or inconsistent follow-up, models trained on such data are likely to reproduce or amplify existing inequities. Furthermore, multimodal disease modeling may disproportionately benefit populations with access to advanced imaging, molecular testing, specialty care, digital devices, or comprehensive longitudinal health records.

This consideration is particularly relevant for chronic disease modeling, where biological processes, social context, healthcare access, environmental exposures, and treatment availability influence disease trajectories. A model that accurately predicts risk within one health system may not generalize to another, as observed trajectories often reflect differences in care pathways rather than underlying biological mechanisms.

Consequently, the operating-system approach should regard bias assessment as a fundamental scientific requirement rather than an optional ethical consideration. Dataset documentation, subgroup performance analysis, subgroup calibration, missingness analysis, site-stratified validation, fairness evaluation, and patient-centered governance must be integrated throughout the model lifecycle. In cases where populations are underrepresented, targeted data generation may be more appropriate than relying solely on technical corrections.

Regulatory alignment may slow exploratory science

A potential objection is that regulatory alignment may introduce excessive formality into early-stage research. Exploratory modeling, hypothesis generation, and biological discovery require flexibility, rapid iteration, and tolerance for uncertainty. Mandating regulatory-grade documentation for all AI research from the outset could inhibit scientific creativity.

A more effective approach is to differentiate between levels of evidentiary maturity. Not all models require the same degree of regulatory documentation. For example, a model used for exploratory clustering of disease subtypes should not be subject to the same standards as one employed to select patients for clinical trials or to guide patient-specific care. The model's context of use should determine the appropriate standard, the potential consequences of error, the extent of human oversight, and the maturity of supporting evidence.

Nevertheless, even exploratory models should maintain basic provenance, data lineage, cohort definitions, and version control. Early documentation should not be burdensome, but must be sufficient to enable reconstruction of

the research process, including the data used, assumptions made, and known limitations. Regulatory alignment should be conceptualized as a graduated lifecycle process rather than a uniform threshold applied to all scientific activities.

Limits of agentic AI and automated scientific workflows

Agentic AI systems may accelerate literature review, data preparation, model comparison, documentation, and hypothesis generation. However, they also introduce new risks. Agents can hallucinate citations, misuse tools, propagate earlier errors, overfit workflow choices to benchmark metrics, or generate plausible but incorrect scientific explanations. In complex biomedical settings, an automated workflow can fail silently if it produces outputs that appear coherent but are based on flawed assumptions.

For this reason, agentic systems should be framed as workflow accelerators rather than autonomous scientific authorities. Their outputs must be auditable, their tool use must be logged, and their claims must be traceable to data, code, literature, or human review. Human experts should remain responsible for scientific interpretation, clinical relevance, and regulatory judgment.

A mature scientific operating system would evaluate agents as components of the research workflow. This means benchmarking not only final model performance, but also task completion accuracy, citation fidelity, data-curation quality, error-detection ability, reproducibility, and robustness under distribution shift. Without such evaluation, agentic automation may accelerate biomedical research while reducing its reliability.

Predictive performance is not equivalent to biological understanding

A central risk in AI-driven disease modeling is conflating prediction with explanation. A model may predict disease progression accurately by exploiting correlations that are not biologically causal. Conversely, a biologically meaningful mechanism may not produce strong individual-level predictive performance in available data. This distinction matters for clinical translation, therapeutic development, and scientific inference.

The operating-system framework should therefore separate several tasks that are often blended: risk prediction, disease subtyping, causal inference, mechanism discovery, treatment-effect estimation, clinical decision support, and regulatory qualification. Each task requires different data, assumptions, validation methods, and evidentiary standards.

For example, a prognostic model used to enrich a clinical trial may require strong calibration and external validation but may not need to fully explain disease biology. A model used to nominate therapeutic targets requires stronger mechanistic support and causal evidence. A model used in patient-specific clinical decision support requires usability testing, safety analysis, monitoring, and human oversight. Treating these use cases as interchangeable would weaken both the scientific and regulatory arguments.

Sustainability and maintenance

A final limitation is that scientific infrastructure must be sustained over time. Research platforms often fail not because the underlying concept is flawed, but because long-term funding, governance, documentation, software maintenance, data updates, and community engagement are neglected. Disease-modeling systems are particularly vulnerable to this challenge because clinical practice, diagnostic criteria, biomarker technologies, treatment options, coding systems, and population health patterns continuously evolve.

A scientific operating system must therefore incorporate lifecycle management from the outset. Models should be monitored for drift, datasets versioned, ontologies updated, benchmarks revised, and retired models retained in a traceable and auditable form. Maintenance should be recognized as a scientific function rather than merely an operational expense.

This challenge extends beyond technology to institutional incentives. Academic promotion systems, grant mechanisms, and many public-private partnerships often reward novelty, publication, and new model development more readily than infrastructure stewardship. Yet the long-term value of disease-modeling systems depends on the continued investment in data curation, benchmark maintenance, validation, documentation, governance, and software sustainability. If these activities are not supported, even technically successful platforms may gradually lose scientific relevance, reproducibility, and trustworthiness.

Risk or counterargument	Why it matters	Design response
Over-centralization	May suppress local expertise and disease-specific innovation	Modular, federated architecture with interoperable interfaces
Premature standardization	May freeze immature methods	Standardize provenance, validation, and documentation rather than model architecture
High cost and complexity	May exclude smaller institutions	Shared tooling, templates, benchmarks, and precompetitive infrastructure
Privacy and IP constraints	May limit data access and collaboration	Federated analysis, secure environments, controlled-access benchmarks
Bias and inequity	May reproduce unequal care and underrepresentation	Subgroup calibration, representativeness audits, targeted data generation
Regulatory burden	May slow exploratory research	Graduated evidence requirements based on the context of use
Agentic AI failure modes	May introduce hallucinations or silent workflow errors	Human review, audit logs, tool-use tracking, agent benchmarking
Prediction-explanation confusion	May overstate biological meaning	Separate prediction, causality, mechanism, and decision-support use cases
Sustainability	Platforms may decay without maintenance	Lifecycle management, versioning, drift monitoring, governance funding

Table 4. Key risks and design responses

Summary

The limitations described above do not undermine the case for a scientific operating system for disease modeling; they help define the conditions under which such a system may be scientifically useful, operationally sustainable, and socially trustworthy. The central challenge is not simply building more capable models, but creating infrastructure that supports reproducibility, validation, transparency, interoperability, and continuous improvement without constraining scientific discovery.

Accordingly, any successful scientific operating system will need to balance competing objectives: standardization without rigidity, collaboration without loss of autonomy, automation without loss of accountability, and innovation without sacrificing reproducibility. These tensions are not obstacles to be eliminated, but design constraints that must be managed deliberately. The long-term value of such systems will depend less on their technical sophistication than on their ability to support reliable scientific learning across institutions, populations, and time.

Future Research Directions

The preceding sections suggest that the next phase of progress in predictive medicine will depend as much on advances in scientific infrastructure as on advances in model performance. While substantial attention has been devoted to developing increasingly capable AI systems, many of the most important open questions now concern how such systems can be integrated into reliable, reproducible, and clinically meaningful scientific workflows. Several areas appear particularly important for future investigation.

Multiscale Disease Modeling

A central scientific challenge remains the integration of observations across molecular, cellular, physiological, behavioral, and clinical scales. Future systems will need to represent interactions across these levels while preserving both biological meaning and temporal context (116, 117). Progress may depend not only on larger datasets and more powerful models, but also on improved representations that connect mechanisms operating across scales.

Earlier and More Reliable Prediction

One longstanding objective in neurodegenerative disease research is the identification of predictive signals years or decades before overt disease onset (118, 119). Achieving this goal will require distinguishing early pathogenic processes from normal aging, transient variation, and mechanisms of resilience. Although recent work demonstrates the feasibility of early risk prediction using longitudinal and multimodal data, important challenges remain regarding data quality, temporal sparsity, generalizability, and validation across populations (120).

Validation Science and Trustworthy AI

As AI systems become increasingly integrated into biomedical research workflows, validation itself may emerge as an important area of scientific investigation. Questions concerning uncertainty quantification, calibration, temporal robustness, reproducibility, provenance, and performance under distribution shift remain incompletely understood. Future work may benefit from treating validation frameworks as scientific infrastructure rather than as secondary evaluation procedures applied after model development.

AI-Augmented Scientific Workflow

Agentic and workflow-oriented AI systems may increasingly support literature review, dataset construction, benchmarking, hypothesis generation, model development, and evidence synthesis (121-123). Their long-term significance may lie less in automating individual tasks than in enabling more structured, iterative, and transparent forms of scientific inquiry. Realizing this potential will require workflows that preserve provenance, support reproducibility, and enable rigorous evaluation of both outputs and underlying assumptions.

Scientific Operating Systems as a Research Discipline

A broader opportunity lies in studying scientific operating systems themselves. Much of contemporary AI research focuses on model architectures and predictive performance, while comparatively little attention has been devoted to understanding how data infrastructure, validation systems, workflow architectures, governance mechanisms, and regulatory frameworks interact to influence scientific outcomes. Future work may therefore benefit from evaluating and comparing entire scientific systems rather than individual models alone.

International Infrastructure and Collaboration

The scale and complexity of longitudinal disease modeling make international collaboration increasingly important. The development of shared datasets, federated learning networks, benchmarking ecosystems, metadata standards, and governance frameworks will likely require sustained cooperation across institutions and countries (84, 124). The long-term success of predictive medicine may depend as much on the quality of these collaborative infrastructures as on advances in algorithms themselves.

Taken together, these directions suggest that future progress will depend on more than the development of increasingly capable models. Equally important will be the creation of the scientific, technical, and institutional infrastructure required to make those models reliable, reproducible, interpretable, and useful in practice.

Conclusion: Toward a Scientific Operating System for Predictive Medicine

The convergence of multimodal biomedical data, advanced AI architectures, and large-scale computational infrastructure has created a rare and consequential opportunity. The scientific foundations for predictive, longitudinal disease modeling now exist. What remains is not a single technical breakthrough, but a coordinated effort to build the systems through which these capabilities can be translated into reliable, cumulative, and clinically meaningful knowledge.

This white paper argues that the central task facing the field is the construction of a scientific operating system for disease modeling. This is a scientific, engineering, and governance challenge requiring deliberate design, sustained investment, and coordinated action across institutions, sectors, and regulatory bodies.

A credible path forward requires progress across six interdependent components.

First, data must be treated as scientific infrastructure. Dataset engineering, harmonization, provenance tracking, metadata management, and AI-assisted curation should be recognized as core scientific activities. The quality and longitudinal integrity of data ultimately constrain what models can learn and what conclusions can be trusted.

Second, validation must be embedded throughout the system. Validation cannot remain a retrospective exercise applied to finished models. Calibration, uncertainty quantification, temporal validation, benchmarking, and explicit characterization of model limitations should be integrated throughout the lifecycle of data, models, and workflows.

Third, AI architectures must be modular and interoperable. No single modeling paradigm is sufficient. Progress will likely depend on systems that combine foundation models,

knowledge graphs, neuro-symbolic approaches, and agentic workflows within coherent and auditable frameworks.

Fourth, workflow infrastructure must become a first-class scientific asset. Research pipelines should be structured, traceable, reproducible, and auditable, preserving the full chain of scientific reasoning from data acquisition through analysis and reporting.

Fifth, precompetitive collaboration must expand. Shared infrastructure—including federated data networks, benchmarking ecosystems, ontologies, metadata standards, and validation frameworks—will be essential for building systems that are robust, generalizable, and scalable across institutions and populations.

Sixth, regulatory alignment must be incorporated from the outset. Emerging frameworks from the FDA, EMA, and other regulatory bodies increasingly emphasize context of use, lifecycle management, uncertainty characterization, transparency, and post-deployment monitoring. These considerations are becoming integral components of scientifically credible disease-modeling systems.

These components are not independent. Data quality constrains model performance. Validation defines trust. Architecture determines what can be learned. Workflows determine what can be reproduced. Collaboration determines what can scale. Regulatory frameworks influence what can be deployed in real-world settings. Weakness in any one dimension limits the effectiveness of the broader system.

The implications extend beyond technical performance. As AI systems increasingly inform drug development, clinical trial design, and patient care, requirements such as provenance, validation, transparency, lifecycle governance, and regulatory alignment become closely intertwined with scientific rigor itself. Systems that do not address these requirements may struggle not only in regulatory settings, but also in their ability to generate trustworthy and actionable knowledge.

The field now faces an opportunity to move beyond isolated modeling efforts toward more integrated scientific infrastructure. Continued advances in algorithms will remain important, but long-term progress may depend equally on the development of shared systems that support reproducibility, interoperability, validation, and continuous learning.

The central argument of this field guide is that advances in predictive medicine will depend not only on more capable models, but also on the infrastructure through which data, models, workflows, validation systems, governance mechanisms, and regulatory frameworks interact. Whether this vision ultimately proves successful remains an empirical question. What is already clear, however, is that reliable, reproducible, and clinically meaningful disease modeling will require greater attention to these system-level foundations. Building them will depend on sustained collaboration across disciplines, institutions, industry, and regulatory communities.

Acknowledgement
The authors thank Varun Ojha, PhD, Associate Professor of Artificial Intelligence, School of Computing, Newcastle University, Newcastle upon Tyne, England, United Kingdom, for his review and commentary during the preparation of this manuscript.

Funding
This scientific article is part of the CLARA project, which has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101136607.

Conflict of interest
The authors report the following disclosures: ASK and ZSK report relationships involving Brain Watch Coalition, PAD 20/20, Khachaturian and Associates, Inc., Alzheon, CTAD, Eli Lilly & Company, International Neurodegenerative Disorders Research Center, and Serdi Publishing; VS and JMB report relationships involving International Neurodegenerative Disorders Research Center; MWL reports NIH grant support paid to his institution; OCF reports manuscript support and grants from multiple public and foundation sources and consulting fees from CLAIRIgene; JH reports speaker honoraria from Eli Lilly and is co-founder of the Alzheimerchain Foundation/Terrapino mobile app; IRJHH reports Horizon Europe/H2020 grant support and royalties or licenses related to BrainSymph AS; SZL reports Israel Science Foundation grant support; and SS reports no conflicts of interest. NPP and JT have no relevant disclosures. Ara S. Khachaturian, the journal's Editor-in-Chief and a coauthor of this article, was recused from its peer review and editorial decision. The manuscript was handled by another editor.

Ethical standards
This article is a conceptual review and field guide. It did not involve human participants, human tissue, or animal subjects, and no primary or identifiable patient-level data were collected, generated, or analyzed. Accordingly, institutional review board approval and informed consent were not required. All work was conducted in accordance with the ethical principles of the Declaration of Helsinki, as applicable, and with ICMJE and COPE recommendations on authorship, disclosure, and publication integrity. All named authors meet ICMJE authorship criteria, have reviewed and approved the final manuscript, and consent to its publication. The authors declare no plagiarism, duplicate publication, or undisclosed prior dissemination of this work. Generative AI tools were used only for language editing and reference formatting; all scientific content, interpretation, and conclusions are the authors' own, and no AI tool is listed as an author.

Data Availability Statement
No new datasets were generated or analyzed in the preparation of this article. All sources supporting the arguments presented are cited in the reference list.

Open Access
This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide

a link to the Creative Commons license and indicate if changes were made.

References

1. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances. 2021;49:107739.
2. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020;14:117793221989905.
3. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12(3):e1001779.
4. Zitnik M, Nguyen F, Wang B, et al. Machine learning for integrating data in biology and medicine. Nature Reviews Methods Primers. 2022;2:39.
5. Dong A, Toledo JB, Honnorat N, et al. Heterogeneity of neuroanatomical patterns in Alzheimer's disease. Nature Communications. 2021;12:1161.
6. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1(1):18.
7. Wang E, Yu K, Cao J, Wang M, Katsel P, Song W-m, et al. Multiscale proteomic modeling reveals protein networks driving Alzheimer's disease pathogenesis. Cell. 2025;188(22):6186–204. e13.
8. Young AL, Marinescu RV, Oxtoby NP, Bocchetta M, Yong K, Firth NC, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nature Communications. 2018;9(1):4273.
9. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nature Medicine. 2022;28(9):1773–84.
10. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology. 2017;18:83.
11. Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Medicine. 2020;3(1):136.
12. Lim B, Alaa AM, van der Schaar M. Forecasting treatment responses over time using recurrent marginal structural networks. Advances in Neural Information Processing Systems; 2018.
13. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8.
14. Sambasivan N, Kapania S, Highfill H, et al. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Proceedings of CHI; 2021. doi:10.1145/3411764.3445518.
15. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden Technical Debt in Machine Learning Systems. Advances in Neural Information Processing Systems; Montreal, Canada: Neural Information Processing Systems Foundation; 2015. p. 2503-11.
16. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
17. Insel PS, Ossenkoppele R, Gessert D, et al. Time to Amyloid Positivity and Preclinical Changes in Brain Metabolism, Atrophy, and Cognition: Evidence for Emerging Amyloid Pathology in Alzheimer's Disease. Front Neurosci. 2017;11:281. doi:10.3389/fnins.2017.00281.
18. Khachaturian AS, Khachaturian, Z.S. The Calcium Signaling System: Aiming for a Comprehensive Explanation of Aging and the Dementia-Alzheimer's Syndrome. Vitality, Medicine & Engineering. 2023;6.
19. Khachaturian ZS. A unifying hypothesis on the etiology of Alzheimer's disease. In: Fisher A, Nagatsu, T.,Yoshida, M., editor. Alzheimer's and Parkinson's disease II: basic and therapeutic strategies. New York: Plenum Press; 1989.
20. Alzheimer's Association Calcium Hypothesis W. Calcium Hypothesis of Alzheimer's disease and brain aging: A framework for integrating new evidence into a comprehensive theory of pathogenesis. Alzheimers Dement. 2017;13(2):178–82 e17.
21. Khachaturian ZS, Kuller LH, Khachaturian AS. Strategic goals and roadmap for dementia prevention by stroke prevention. Alzheimers Dement. 2019;15(7):865–9.
22. Hachinski V, Einhäupl K, Ganten D, et al. Preventing dementia by preventing stroke: The Berlin Manifesto. Alzheimers Dement. 2019;15(7):961-984.
23. Jack CR Jr, Knopman DS, Jagust WJ, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 2010;9(1):119-128.
24. Sperling RA, Aisen PS, Beckett LA, et al. Toward defining the preclinical stages of Alzheimer's disease. Alzheimers Dement. 2011;7(3):280-292.
25. Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031.
26. Pronk NP, Woodard C, Zimmerman FJ, Arena R. An ecological framework for population health and well-being. Prog Cardiovasc Dis. 2025;90:13–21.
27. Braveman P, Egerter S, Williams DR. The social determinants of health: coming of age. Annual Review of Public Health. 2011;32:381–98.
28. Magnan S. Social Determinants of Health 101 for Health Care: Five Plus Five. NAM Perspectives. 2017;7(10).
29. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers. NPJ Digital Medicine. 2019;2:14.
30. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics. 2018;19(6):1236–46.
31. Alexander DC, Dyrby TB, Nilsson M, Zhang H. Imaging brain microstructure with diffusion MRI: practicality and applications. NMR Biomed. 2019;32(4):e3841.
32. Kitano H. Systems biology: a brief overview. Science. 2002;295:1662–4.
33. Jack CR Jr, Knopman DS, Jagust WJ, et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 2013;12(2):207-216.
34. Toledo JB, Cairns NJ, Da X, Chen K, Carter D, Fleisher A, et al. Clinical and multimodal biomarker correlates of ADNI neuropathological findings. Acta Neuropathol Commun. 2013;1:65.
35. Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics. 2017;56:229–38.
36. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035.
37. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record analysis. Journal of Biomedical Informatics. 2017;83:168–85.
38. Zimmerman FJ, Pronk NP. Socioeconomic milieu and culture: Forcing factors and the Most fundamental determinant of health. Prog Cardiovasc Dis. 2025;90:62–70.
39. Stringhini S, Carmeli C, Jokela M, Avendaño M, Muennig P, Guida F, et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. The Lancet. 2017;389(10075):1229–37.
40. Pronk NP, Wang S, Woodard C, Bhatt T, Arena R. Multilevel prediction of Alzheimer's disease dementia in the United States: An artificial intelligence analysis. Alzheimers Dement (Amst). 2026;18(1):e70274.
41. Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler NE, et al. Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence. SSM - Population Health. 2020;10:100526.
42. Galea S, Tracy M, Hoggatt KJ, Dimaggio C, Karpati A. Estimated deaths attributable to social factors in the United States. Am J Public Health. 2011;101(8):1456–65.
43. Yadav S, Mondal J, Saha M. Systems biology approaches for multi omics integration using artificial intelligence. Academia Biology. 2026;4(1).
44. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13):i457-i466. doi:10.1093/bioinformatics/bty294.
45. Lim B, Arık SÖ, Loeff N, Pfister T. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting. 2021;37(4):1748–64.
46. Trapnell C. Defining cell types and states with single-cell genomics. Nature Methods. 2015;12:381–6.
47. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
48. Theodorakis N, Feretzakis G, Tzelves L, Paxinou E, Hitas C, Vamvakou G, et al. Integrating Machine Learning with Multi-Omics Technologies in Geroscience: Towards Personalized Medicine. JPM. 2024;14(9):931.
49. Bateman RJ, Xiong C, Benzinger TLS, et al. Clinical and biomarker changes in dominantly inherited Alzheimer's disease. New England Journal of Medicine. 2012;367:795–804.
50. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data,

Machine Learning, and Clinical Medicine. New England Journal of Medicine. 2016;375(13):1216–9.
51. Khachaturian AS, Cassin B, Finney GR. Using Clinical Decision Intelligence Applications to Improve Pathways For Earlier Detection Of Underrecognized Cognitive Disorders. J Aging Res Lifestyle. 2023;12:14-17. doi:10.14283/jarlife.2023.3.
52. Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019;41:423–43.
53. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258. 2021.
54. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems; 2020.
55. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65.
56. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. International Conference on Learning Representations (ICLR); 2021.
57. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems; 2017.
58. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.
59. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1:206–15.
60. Gaudelet T, Day B, Jamasb AR, Soman J, Regep C, Liu G, et al. Utilizing graph machine learning within drug discovery and development. Briefings in Bioinformatics. 2021;22(6):bbab159.
61. Himmelstein DS, Lizee A, Hessler C, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017;6:e26726.
62. Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nature Biotechnology. 2022;40(5):692–702.
63. Nicholson DN, Greene CS. Constructing knowledge graphs for precision medicine. Brief Bioinform. 2020;21(4):1073-1087.
64. Schölkopf B. Causality for Machine Learning. arXiv:1911.10500. 2019.
65. Garcez AdA, Lamb LC, Gabbay DM. Neural-symbolic learning systems: foundations and applications. Nature Machine Intelligence. 2021;3:554–61.
66. Mao J, Gan C, Kohli P, et al. The neuro-symbolic concept learner. International Conference on Learning Representations (ICLR); 2019.
67. Park JS, O'Brien JC, Cai CJ, et al. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442. 2023.
68. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A Survey on Large Language Model based Autonomous Agents. Front Comput Sci. 2024;18(6):186345.
69. Xi Z, Chen W, Guo X, et al. The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv:2309.07864. 2023.
70. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete Problems in AI Safety. arXiv; 2016.
71. Sendak MP, Gao M, Nichols M, et al. A path for translation of machine learning products into healthcare delivery. BMJ. 2020;370:m3215.
72. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. Proceedings of EMNLP-IJCNLP; 2019.
73. Booeshaghi AS, Luebbert L, Pachter L. Science should be machine-readable. Scientific Communication and Education; 2026.
74. Burns G, Groth P. Complex Knowledge Curation using Agentic Ontological Notebook Memory. Proceedings of CAIS '26; 2026.
75. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
76. Kreuzberger D, Kühl N, Hirschl S. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. arXiv; 2022.
77. Pineau J, Vincent-Lamarre P, Sinha K, et al. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. Journal of Machine Learning Research. 2021;22:1–20.
78. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86–92.
79. Mons B, Neylon C, Velterop J, et al. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use. 2017;37(1):49–56.
80. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018.
81. Polyzotis N, Zinkevich M, Roy S, Breck E, Whang S. Data Validation for Machine Learning. Proceedings of MLSys; 2019.
82. Sansone S-A, Gonzalez-Beltran A, Rocca-Serra P, et al. FAIRsharing as a community approach to standards, repositories and policies. Nature Genetics. 2019;51:169–70.
83. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. NeuroImage. 2018;167:104–20.
84. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. npj Digit Med. 2020;3:119. doi:10.1038/s41746-020-00323-1.
85. Hogan A, Blomqvist E, Cochez M, D'Amato C, Melo GD, Gutierrez C, et al. Knowledge Graphs. ACM Comput Surv. 2022;54(4):1–37.
86. Zha D, Bhat ZP, Lai K-H, et al. Data-centric artificial intelligence: A survey. AI Magazine. 2021;42(3):64–77.
87. Fries JA, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nature Communications. 2021;12(1):2017.
88. Ratner A, Bach SH, Ehrenberg H, Fries J, Wu S, Ré C. Snorkel: rapid training data creation with weak supervision. Proc VLDB Endow. 2017;11(3):269–82.
89. Northcutt C, Jiang L, Chuang I. Confident Learning: Estimating Uncertainty in Dataset Labels. jair. 2021;70:1373–411.
90. Pang G, Shen C, Cao L, Hengel AVD. Deep Learning for Anomaly Detection: A Review. ACM Comput Surv. 2022;54(2):1–38.
91. Chen Z, Tan S, Chajewska U, Rudin C, Caruana R. Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help? : arXiv; 2023.
92. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 3rd ed. Hoboken, NJ: Wiley; 2019.
93. Swayamdipta S, Schwartz R, Lourie N, et al. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. Proceedings of EMNLP; 2020.
94. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports. 2017;7(1):17816.
95. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
96. Hooker SA, Rossom RC, Haapala J, et al. External Validation of Epic's Risk of Opioid Abuse and Overdose Model Among Primary Care Patients in Three Health Systems. J Gen Intern Med. 2026. doi:10.1007/s11606-026-10257-1.
97. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? Proceedings of FAccT; 2021.
98. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023;55(12):1–38.
99. Espinosa-Sánchez A, Carnero A. Artificial Intelligence in Tumor Evolution: Understanding Cancer Complexity Through Multi-Modal Data Integration in Precision Oncology. Cells. 2026;15(11):1031. doi:10.3390/cells15111031.
100. Maynez J, Narayan S, Bohnet B, McDonald R. On Faithfulness and Factuality in Abstractive Summarization. Proceedings of ACL; 2020.
101. Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for Human-AI Interaction. Proceedings of CHI; 2019.
102. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trials evaluating artificial intelligence interventions. Nature Medicine. 2020;26:1364–74.
103. Vokinger KN, Gasser U. Regulating AI in medicine in the United States and Europe. Nature Machine Intelligence. 2021;3(9):738–9.
104. US Food and Drug Administration. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products. Draft guidance; 2025.
105. US Food and Drug Administration. AI-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft guidance; 2025.
106. US Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. Final guidance; originally issued December 4, 2024; updated August 18, 2025.
107. European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle.

EMA/CHMP/CVMP/83833/2023; September 2024.
108. European Medicines Agency; US Food and Drug Administration. Guiding Principles of Good AI Practice in Drug Development. 2026.
109. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group "Evaluating Diagnostic Tests and Prediction Models" of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
110. Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. arXiv; 2019.
111. All of Us Research Program Investigators. The All of Us Research Program. N Engl J Med. 2019;381(7):668-676. doi:10.1056/NEJMsr1809937.
112. Amunts K, Axer M, Banerjee S, Bitsch L, Bjaalie JG, Brauner P, et al. The coming decade of digital brain research: A vision for neuroscience at the intersection of technology and computing. Imaging Neurosci (Camb). 2024;2.
113. Haraldsen IH, Hatlestad-Hall C, Marra C, Renvall H, Maestu F, Acosta-Hernandez J, et al. Intelligent digital tools for screening of brain connectivity and dementia risk estimation in people affected by mild cognitive impairment: the AI-Mind clinical study protocol. Front Neurorobot. 2023;17:1289406.
114. Birhane A, Kalluri P, Card D, et al. The Values Encoded in Machine Learning Research. Proceedings of FAccT; 2022.
115. Stall S, Yarmey L, Cutcher-Gershenfeld J, Hanson B, Lehnert K, Nosek B, et al. Make scientific data FAIR. Nature. 2019;570(7759):27–9.
116. Khachaturian ZS. Models and modeling systems in Alzheimer disease drug discovery. Alzheimer Dis Assoc Disord. 2002;16 Suppl 1:S9–S12.
117. Khachaturian ZS, Lombardo J. In silico modeling system: a national research resource for simulation of complex brain disorders. Alzheimers Dement. 2009;5(1):1–4.
118. Hampel H, Lista S, Khachaturian ZS. Development of biomarkers to chart all Alzheimer's disease stages: the royal road to cutting the therapeutic Gordian Knot. Alzheimers Dement. 2012;8(4):312–36.
119. Geerts H, Dacks PA, Devanarayan V, Haas M, Khachaturian ZS, Gordon MF, et al. Big data to smart data in Alzheimer's disease: The brain health modeling initiative to foster actionable knowledge. Alzheimers Dement. 2016;12(9):1014–21.
120. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine. 2021;4(1):86.
121. Bran AM, Cox S, Schilter O, et al. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6:525-535. doi:10.1038/s42256-024-00832-8.
122. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
123. Merchant A, Batzner S, Schoenholz SS, et al. Scaling deep learning for materials discovery. Nature. 2023;624:80–5.
124. Van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. 2014;14(1):1144..

© *The Authors 2026*

How to cite this article: Ara S. Khachaturian, Nicolas P Pronk, Sudhir Sivakumaran et al. 2026 Field Guide to Scientific Operating Systems for Longitudinal Disease Research: Architectures, Workflows, Validation Frameworks, and Emerging Directions in ML-Enabled Biomedical Research. VM&E 2026;9:1-27; http://dx.doi.org/10.14283/VME.2026.1