Foundational Models and Their Biomedical Applications

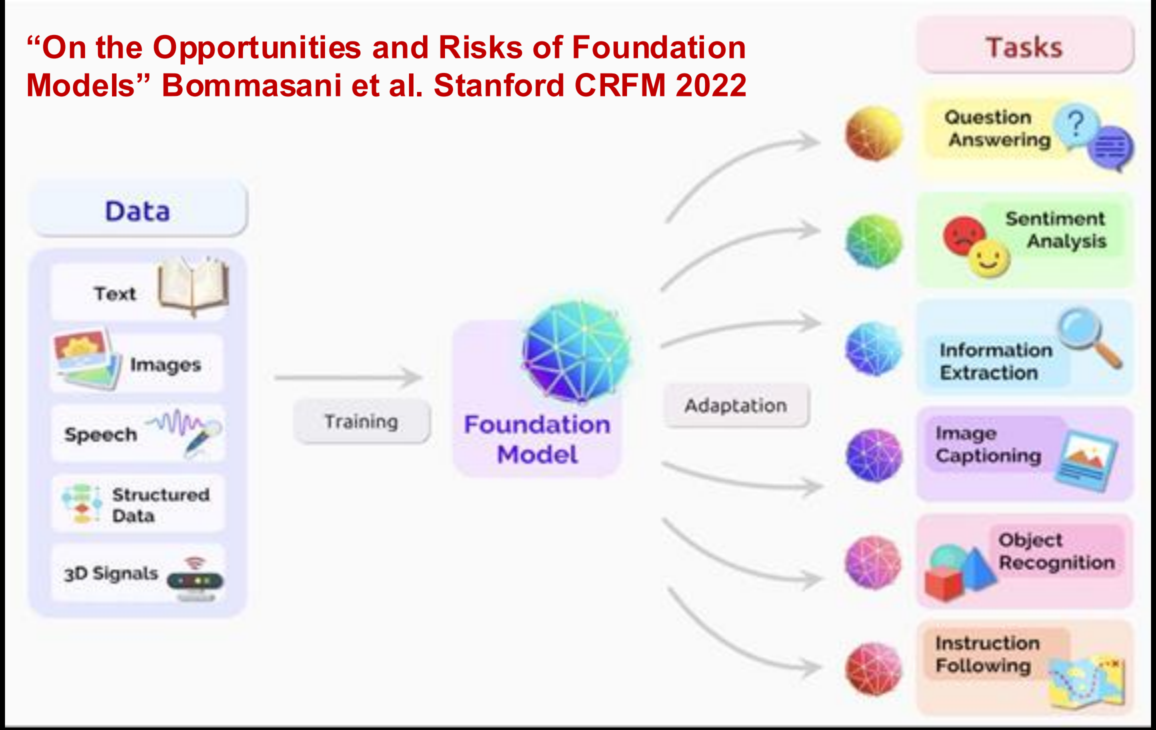
Dehan Kong (University of Toronto)
Yong Chen (University of Pennsylvania)
Linglong Kong (University of Alberta)
Katarzyna Reluga (Humboldt University of Berlin)
Hongtu Zhu (The University of North Carolina at Chapel Hill)

Statistics Up AI Alliance

https://statsupai.org

Foundation Models

- Foundation models are a replacement for task-specific models
- Large-scale pretraining on large unlabeled datasets
- Finetuning for diverse downstream tasks
- Self-supervised learning
- Transfer learning
- GPT-4, DALL-E 2, BERT, etc.



Foundation Models for Biomedical Sciences

1. Methods & Modeling

What unique methodological innovations are needed to build biomedical foundation models, beyond scaling architectures like transformers? How we incorporate causal reasoning, multimodal fusion, or domain-specific inductive biases from biology?

2. Data & Infrastructure

Biomedical foundation models require massive, high-quality data — but biomedical data is fragmented, noisy, and sensitive. How can we best address data scarcity, harmonization, and privacy while building scalable training pipelines?

3. Applications & Translation

Where could biomedical foundation models deliver the most immediate impact — in drug discovery, medical imaging, digital pathology, clinical trial design, or patient risk stratification? Which of these areas could realistically lead to both transformative products?

Toward Causal Generalist Medical AI (CGM-AI):

A Personal Perspective

Hongtu Zhu

University of North Carolina at Chapel Hill

Thanks to Drs. Xin Wang, Xin Wang, Qiao Liu, and Huaxiu Yao for co-teaching CGM-AI and to Drs. Jian Huang and Fei Wang for their slides.

https://www.med.unc.edu/big-s2

CONTENTS

Part I

Foundations of CGM-AI

"Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familiar tools and familiar problems, we will fail to grasp the new opportunities."

- Leo Breiman -

Biomedical Data Resources

Biobanks

- Large, deeply phenotyped cohorts
- Examples: UK Biobank

NIH-Funded Observational Cohorts

- Non-trial studies with rich multi-modal data
- Examples: All of Us, TOPMed, ARIC, ADNI

Clinical Trials & Registries

- Interventional protocols and real-world outcomes
- Examples: ClinicalTrials.gov, SEER cancer registry

Biomedical Data Resources

Healthcare Data

- Electronic Health Records (EHR) and claims
- Structured (ICD/CPT, labs) + unstructured (clinical notes)

Literature

- Peer-reviewed articles, preprints, case reports
- Sources: PubMed, bioRxiv, medRxiv

Ontologies

- Standard vocabularies for data harmonization
- Examples: UMLS, SNOMED CT, ICD-10, MeSH

Biomedical Data Types

Genetics & Omics

- DNA/RNA sequencing (WGS, WES, RNA-seq)
- Epigenomics (methylation, histone marks)
- Proteomics & metabolomics (mass-spec profiles)

Clinical & Administrative Records

- Electronic Health Records (ICD/CPT codes, labs, vitals)
- Claims & billing data
- Pharmacy orders & dispensing logs

Drug Information

- Prescription and utilization records
- Pharmacogenomic annotations (gene–drug interactions)
- Drug databases and adverse-event reports (e.g., FAERS)

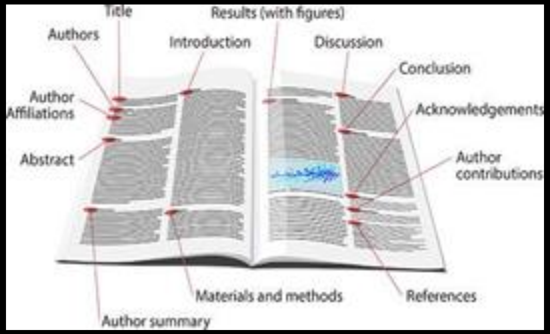
Biomedical Data Types

Medical Imaging

- Radiology (X-ray, CT, MRI, PET)
- Digital pathology and microscopy
- Functional modalities (fMRI, DTI)

Wearables & Remote Monitoring

- Physiologic waveforms (ECG, EEG)
- Continuous sensors (glucose monitors, activity/sleep trackers)
- Home-based vitals (BP, SpO_2)



Textual Data

- Unstructured clinical notes (discharge summaries, progress notes)
- Scientific literature & preprints (PubMed, bioRxiv)
- Patient-reported outcomes & survey responses

Generalist Medical AI (GMAI)

Definition: GMAI are **foundation models** trained via **self-supervision** on large, diverse biomedical datasets. They can flexibly solve **new**, **unseen** medical tasks with minimal or no task-specific labels by interpreting and reasoning across multiple data modalities.

Key Components of GMAI

In-Context Tasking: Natural-language prompts define new tasks on the fly

Multimodal Backbone: Single transformer handling images, EHR, labs, omics, text, and graphs

Knowledge Retrieval: On-demand access to KGs and literature for grounding and reasoning

Self-Supervised Pretraining: Masked and contrastive objectives on large biomedical corpora for zero-/few-shot transfer

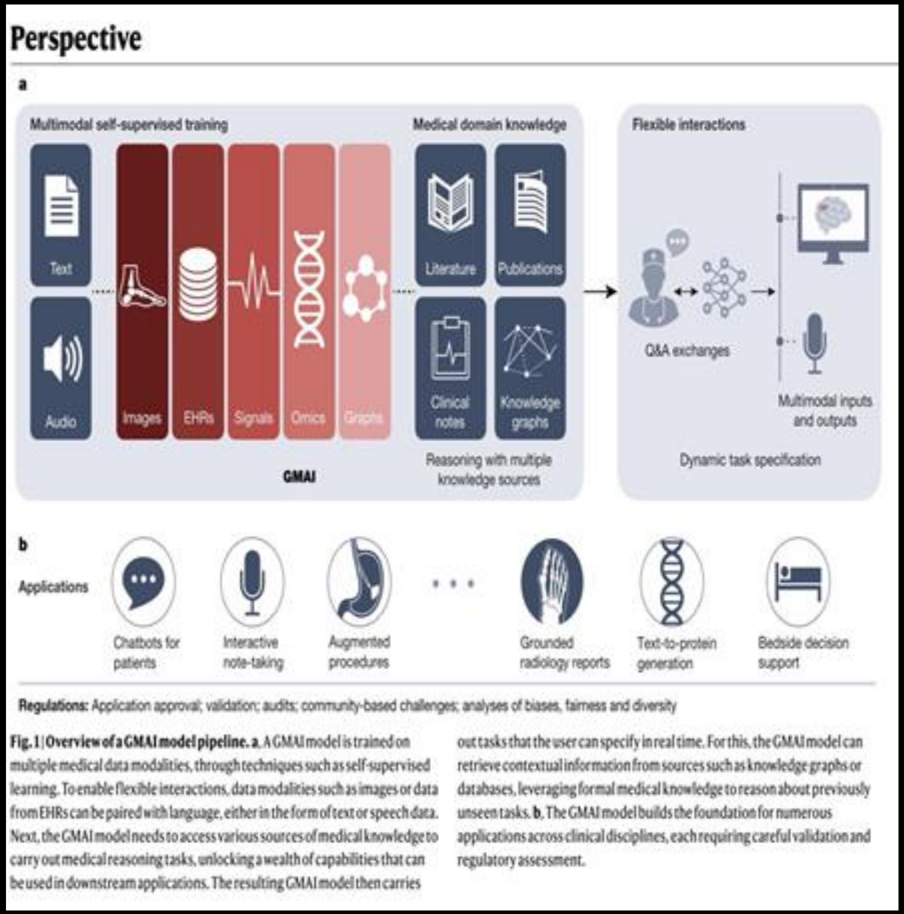


Fig. 1 | Overview of a GMAI model pipeline. a, A GMAI model is trained on multiple medical data modalities, through techniques such as self-supervised learning. To enable flexible interactions, data modalities such as images or data from EHRs can be paired with language, either in the form of text or speech data. Next, the GMAI model needs to access various sources of medical knowledge to carry out medical reasoning tasks, unlocking a wealth of capabilities that can be used in downstream applications. The resulting GMAI model then carries out tasks that the user can specify in real time. For this, the GMAI model can retrieve contextual information from sources such as knowledge graphs or databases, leveraging formal medical knowledge to reason about previously unseen tasks. **b**, The GMAI model builds the foundation for numerous applications across clinical disciplines, each requiring careful validation and regulatory assessment.

Moor, M., … ., Rajpurkar, P. (2023) Foundation models for generalist medical artificial intelligence. *Nature*.

Causal GMAI

Definition:

- Unified paradigm for integrating heterogeneous biomedical data (EHR, imaging, omics, text)
- Fuses causal inference (interventions, counterfactuals) with deep foundation models
- Generalizes across tasks: prevention, diagnosis, prognosis, and treatment planning

Vision:

- Transition from siloed, task-specific tools to a single, adaptable AI backbone
- Provide robust, interpretable decision support under uncertainty
- Accelerate translation from bench (research) to bedside (clinical)

GMAI v.s. Causal GMAI

Focus:

GMAI: Generalist pattern recognition

CGM-AI: Causal reasoning & valid interventions

Architecture:

GMAI: Foundation model only

CGM-AI: + SCM/DAG layers and causal constraints

Inference:

GMAI: Zero-/few-shot tasks

CGM-AI: + “What-if” and counterfactual queries, policy learning

Robustness:

- GMAI: Vulnerable to confounding
- CGM-AI: Confounding control via causal invariants

Explainability:

- GMAI: Attention-based insights
- CGM-AI: Explicit causal paths and do-calculus rationale

Part II

Foundation Models for Major Data Types

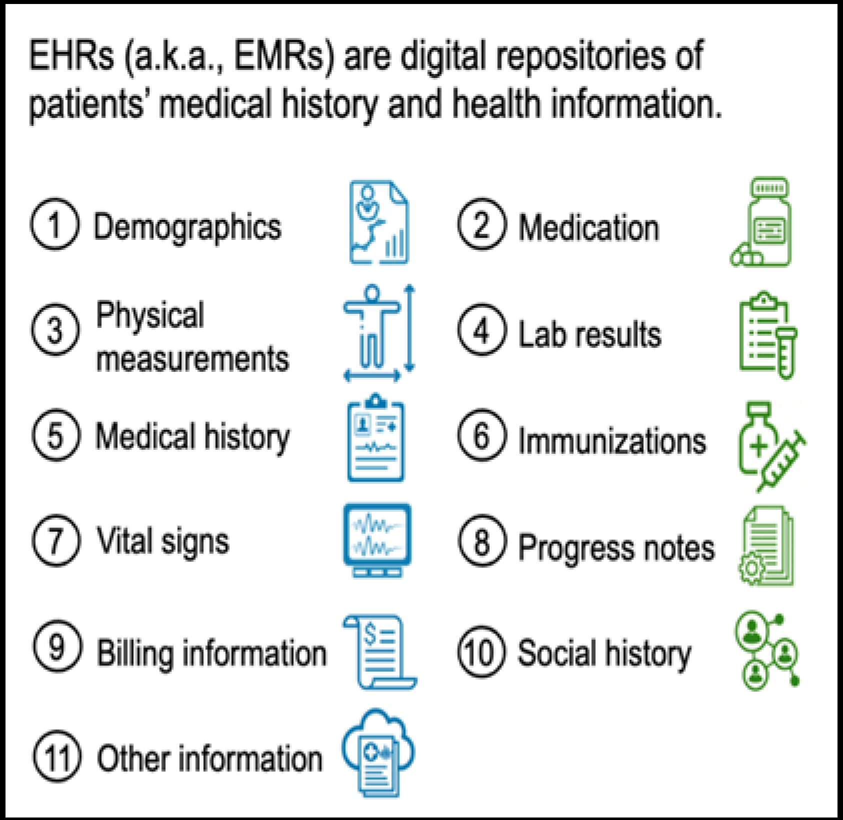
The best thing about being a statistician is that you get to play in everyone's backyard.
- John Tukey -

"If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."
- Leo Breiman -

Electronic Health Records

What and Why? Longitudinal Patient History: Comprehensive record across encounters

- **Multimodal Data Source: Structured codes (ICD, CPT), labs, vitals; unstructured notes**
- **Causal Insights: Temporal order of interventions and outcomes for SCMs**
- **Foundation for CGM-AI: Core modality for pretraining and downstream tasks**
- **Facilitates Care Coordination: Interoperable through standards (FHIR, HL7) across providers**



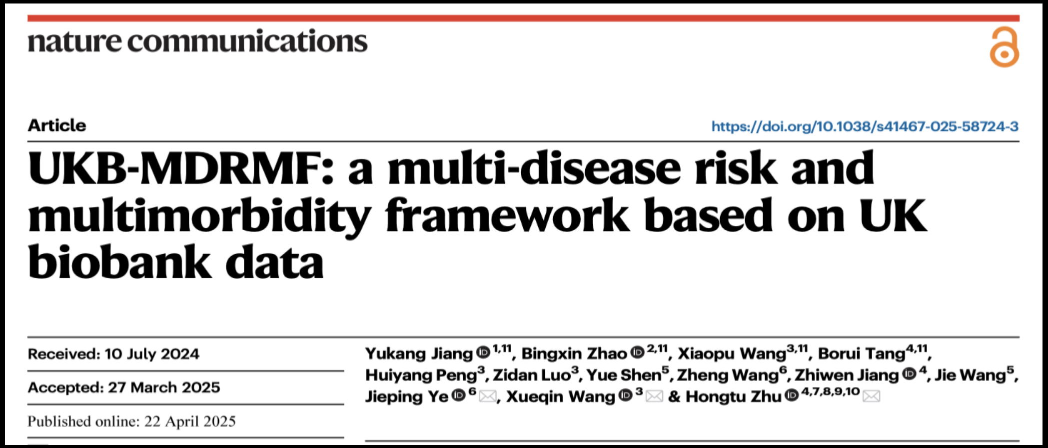
Electronic Health Records

Self-supervised Pretraining in EHR

- Masked Code Prediction: Randomly hide diagnosis/procedure codes, train model to reconstruct
- Next-Visit Forecasting: Predict future encounters, lab trends, or medication changes
- Temporal Contrastive Learning: Contrast segments of patient trajectories to learn robust embeddings
- Integration with Other Modalities: Joint objectives combining EHR and KG or imaging for cross-modal alignment

- Data Quality: Address coding errors, missing visits, and variable granularity
- Privacy Security: Federated learning and differential privacy for multi-center EHR
- Standardization: Adhere to FHIR and OMOP CDM for interoperability
- Cross-Attention Mechanisms: Fuse EHR embeddings with imaging, omics, and KG node representations
- Handling Missingness: Reconstruction losses and imputation for sparse code sequences
- Temporal Alignment: Synchronize EHR events with imaging timestamps and biomarker sampling
- Graph-Enhanced EHR: Augment code sequences with KG-derived entity embeddings for richer context

UKB-MDRMF


nature communications

Article

https://doi.org/10.1038/s41467-025-58724-3

UKB-MDRMF: a multi-disease risk and multimorbidity framework based on UK biobank data

Received: 10 July 2024

Accepted: 27 March 2025

Published online: 22 April 2025

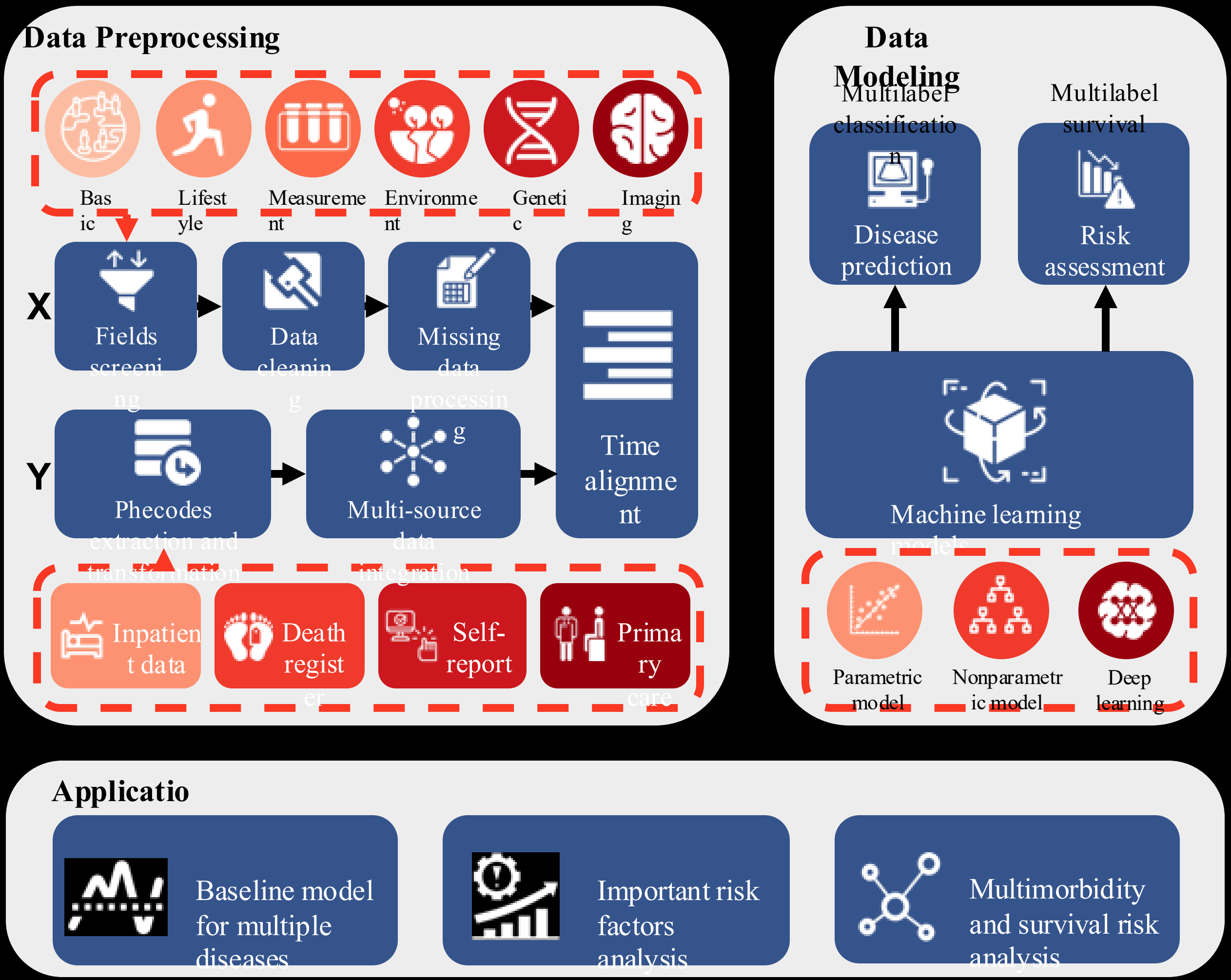
Yukang Jiang[1,11], Bingxin Zhao[2,11], Xiaopu Wang[3,11], Borui Tang[4,11], Huiyang Peng[3], Zidan Luo[3], Yue Shen[5], Zheng Wang[6], Zhiwen Jiang[4], Jie Wang[5], Jieping Ye[6], Xueqin Wang[3] & Hongtu Zhu[4,7,8,9,10]

UKB-MDRMF(a multi-disease risk and multimorbidity framework): Predicts and assesses risk for 1,560 diseases in a unified model.

Multimorbidity Modeling: Captures shared and unique risk-factor networks across diseases.

Performance: Outperforms single-disease models in predictive accuracy for all disease categories.

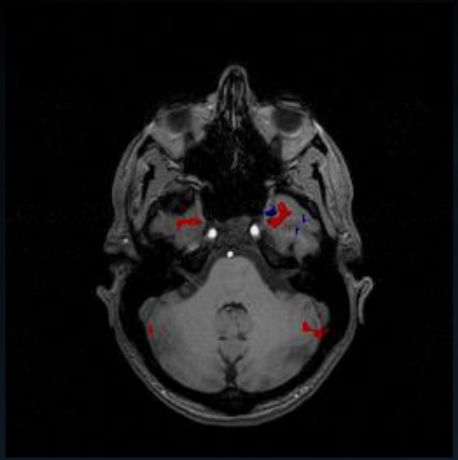
Insight: Provides a holistic perspective on health, revealing co-occurrence mechanisms and broadening disease understanding.



Medical Imaging

Medical imaging is the technique and process used to create images of the human body for clinical purposes or medical science. (https://en.wikipedia.org/)

- These imaging methods are essential for delineating the **structure and functionality of organs and tissues.** Each modality employs a distinct targeting agent, generates data in varying dimensions, extracts unique features, and serves specific purposes within clinical and research contexts.



- X-ray radiography
- Computerized tomography (CT)
- Magnetic resonance imaging (MRI)
- Ultrasound
- Positron emission tomography (PET)
- Electroencephalography (EEG)
- Magnetoencephalography (MEG)
- Functional near-infrared spectroscopy (fNIRS)
- Mammography
- Light microscopy images
- Fluoroscopy
- Echocardiography

Image Processing Analysis Methods

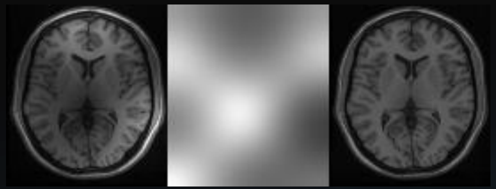
How to enhance and extract signals of interest in imaging data?

Deconvolution

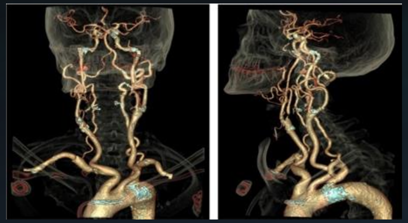
$$\frac{S(\mathbf{q})}{S_0} = \int P(\mathbf{r}, \Delta) e^{i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r}; \quad \mathbf{q} = \gamma \delta g \mathbf{u}$$

- **Image Reconstruction**

- **Image Enhancement**



- **Image Segmentation**



- **Image Registration**

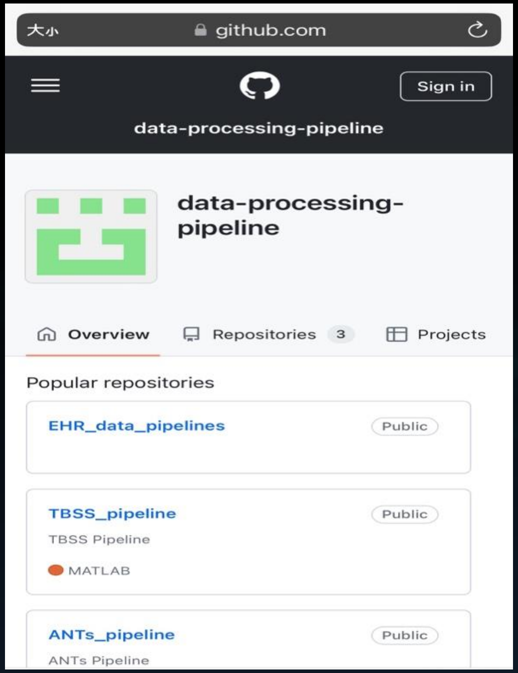
Structural Learning

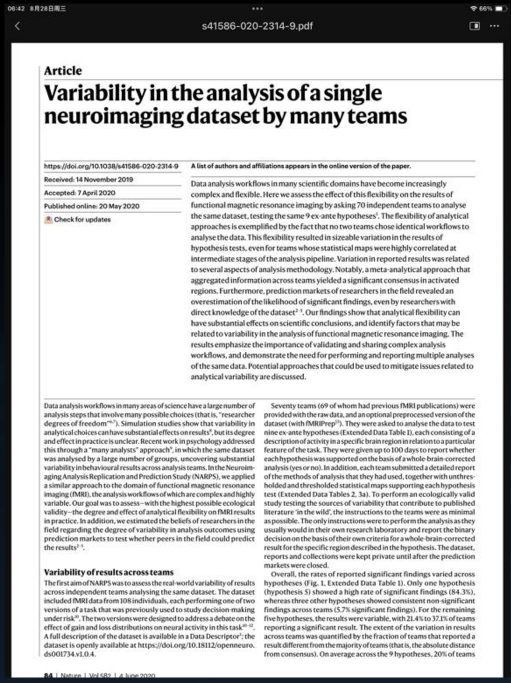
Engineering | Statistics | Machine Learning | Mathematics

Pipelines

More good data processing pipeline papers

https://github.com/data-processing-pipeline





Article

Variability in the analysis of a single neuroimaging dataset by many teams

https://doi.org/10.1038/s41586-020-2314-9
Received: 14 November 2019
Accepted: 7 April 2020
Published online: 20 May 2020
Check for updates

A list of authors and affiliations appears in the online version of the paper.

Data analysis workflows in many scientific domains have become increasingly complex and flexible. Here we assess the effect of this flexibility on the results of functional magnetic resonance imaging by asking 70 independent teams to analyse the same dataset, testing the same 9 ex-ante hypotheses[1]. The flexibility of analytical approaches is exemplified by the fact that no two teams chose identical workflows to analyse the data. This flexibility resulted in sizeable variation in the results of hypothesis tests, even for teams whose statistical maps were highly correlated at intermediate stages of the analysis pipeline. Variation in reported results was related to several aspects of analysis methodology. Notably, a meta-analytical approach that aggregated information across teams yielded a significant consensus in activated regions. Furthermore, prediction markets of researchers in the field revealed an overestimation of the likelihood of significant findings, even by researchers with direct knowledge of the dataset[2-5]. Our findings show that analytical flexibility can have substantial effects on scientific conclusions, and identify factors that may be related to variability in the analysis of functional magnetic resonance imaging. The results emphasize the importance of validating and sharing complex analysis workflows, and demonstrate the need for performing and reporting multiple analyses of the same data. Potential approaches that could be used to mitigate issues related to analytical variability are discussed.

Data analysis workflows in many areas of science have a large number of analysis steps that involve many possible choices (that is, "researcher degrees of freedom"[6,7]). Simulation studies show that variability in analytical choices can have substantial effects on results[8], but its degree and effect in practice is unclear. Recent work in psychology addressed this through a "many analysts" approach[9], in which the same dataset was analysed by a large number of groups, uncovering substantial variability in behavioural results across analysis teams. In the Neuroimaging Analysis Replication and Prediction Study (NARPS), we applied a similar approach to the domain of functional magnetic resonance imaging (fMRI), the analysis workflows of which are complex and highly variable. Our goal was to assess—with the highest possible ecological validity—the degree and effect of analytical flexibility on fMRI results in practice. In addition, we estimated the beliefs of researchers in the field regarding the degree of variability in analysis outcomes using prediction markets to test whether peers in the field could predict the results[2-5].

Variability of results across teams

The first aim of NARPS was to assess the real-world variability of results across independent teams analysing the same dataset. The dataset included fMRI data from 108 individuals, each performing one of two versions of a task that was previously used to study decision-making under risk[10]. The two versions were designed to address a debate on the effect of gain and loss distributions on neural activity in this task[10-12]. A full description of the dataset is available in a Data Descriptor[1]; the dataset is openly available at https://doi.org/10.18112/openneuro.ds001734.v1.0.4.

Seventy teams (69 of whom had previous fMRI publications) were provided with the raw data, and an optional preprocessed version of the dataset (with fMRIPrep[13]). They were asked to analyse the data to test nine ex-ante hypotheses (Extended Data Table 1), each consisting of a description of activity in a specific brain region in relation to a particular feature of the task. They were given up to 100 days to report whether each hypothesis was supported on the basis of a whole-brain-corrected analysis (yes or no). In addition, each team submitted a detailed report of the methods of analysis that they had used, together with unthresholded and thresholded statistical maps supporting each hypothesis test (Extended Data Tables 2, 3a). To perform an ecologically valid study testing the sources of variability that contribute to published literature 'in the wild', the instructions to the teams were as minimal as possible. The only instructions were to perform the analysis as they usually would in their own research laboratory and report the binary decision on the basis of their own criteria for a whole-brain-corrected result for the specific region described in the hypothesis. The dataset, reports and collections were kept private until after the prediction markets were closed.

Overall, the rates of reported significant findings varied across hypotheses (Fig. 1, Extended Data Table 1). Only one hypothesis (hypothesis 5) showed a high rate of significant findings (84.3%), whereas three other hypotheses showed consistent non-significant findings across teams (5.7% significant findings). For the remaining five hypotheses, the results were variable, with 21.4% to 37.1% of teams reporting a significant result. The extent of the variation in results across teams was quantified by the fraction of teams that reported a result different from the majority of teams (that is, the absolute distance from consensus). On average across the 9 hypotheses, 20% of teams

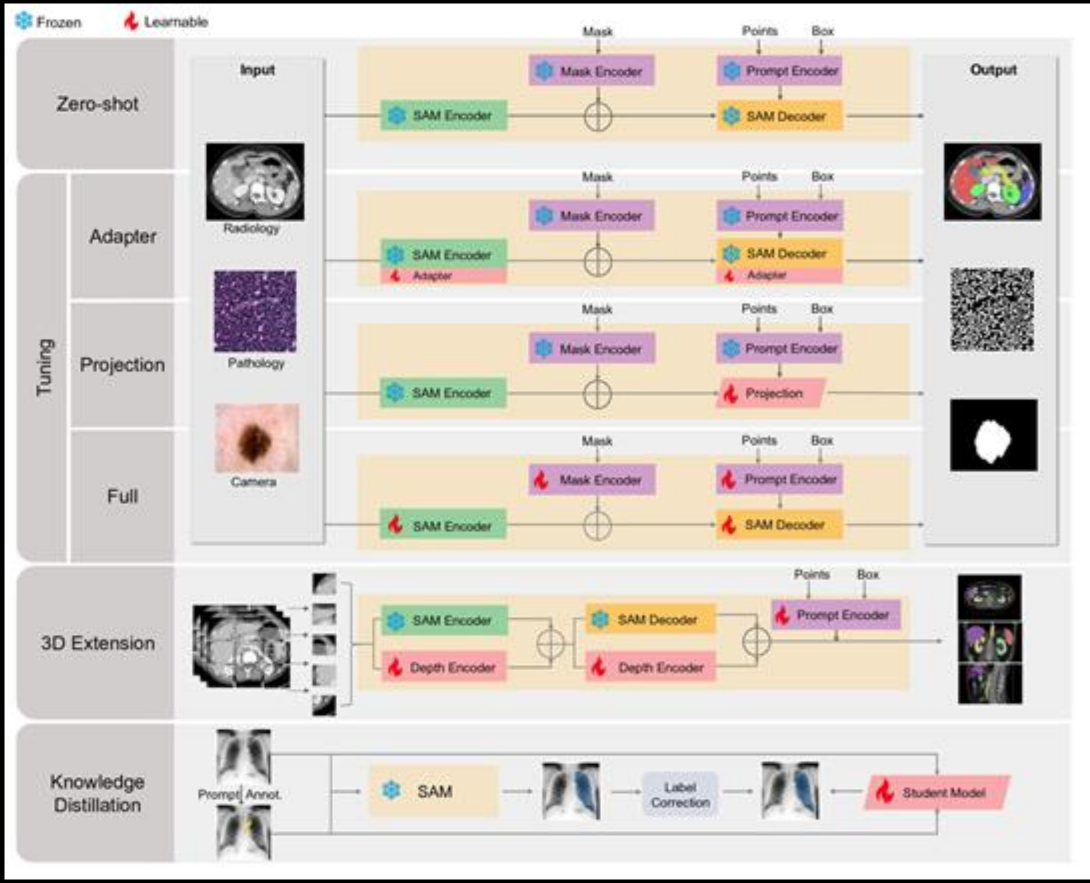
- Good datasets for comprehensive evaluation. There is no publicly available, high-quality imaging datasets with detailed annotation information that cover a large spectrum of segmentation tasks in health care.
- How to quantify the uncertainty and generalizability of brain atlas as well as segmentation and registration tools?
- How to develop foundational models for various segmentation and registration tasks?

Foundation Models for Segmentation

- **Multiple method have been proposed for the adaption of SAM to the medical domain.**

- **Zero-shot segmentation capabilities evaluation: Medical imaging presents unique challenges, distinguished by factors like varied imaging protocols and a wider range of patient demographics. These complexities are not as predominant in standard domain images, making SAM's adaptability in this context particularly intriguing.**

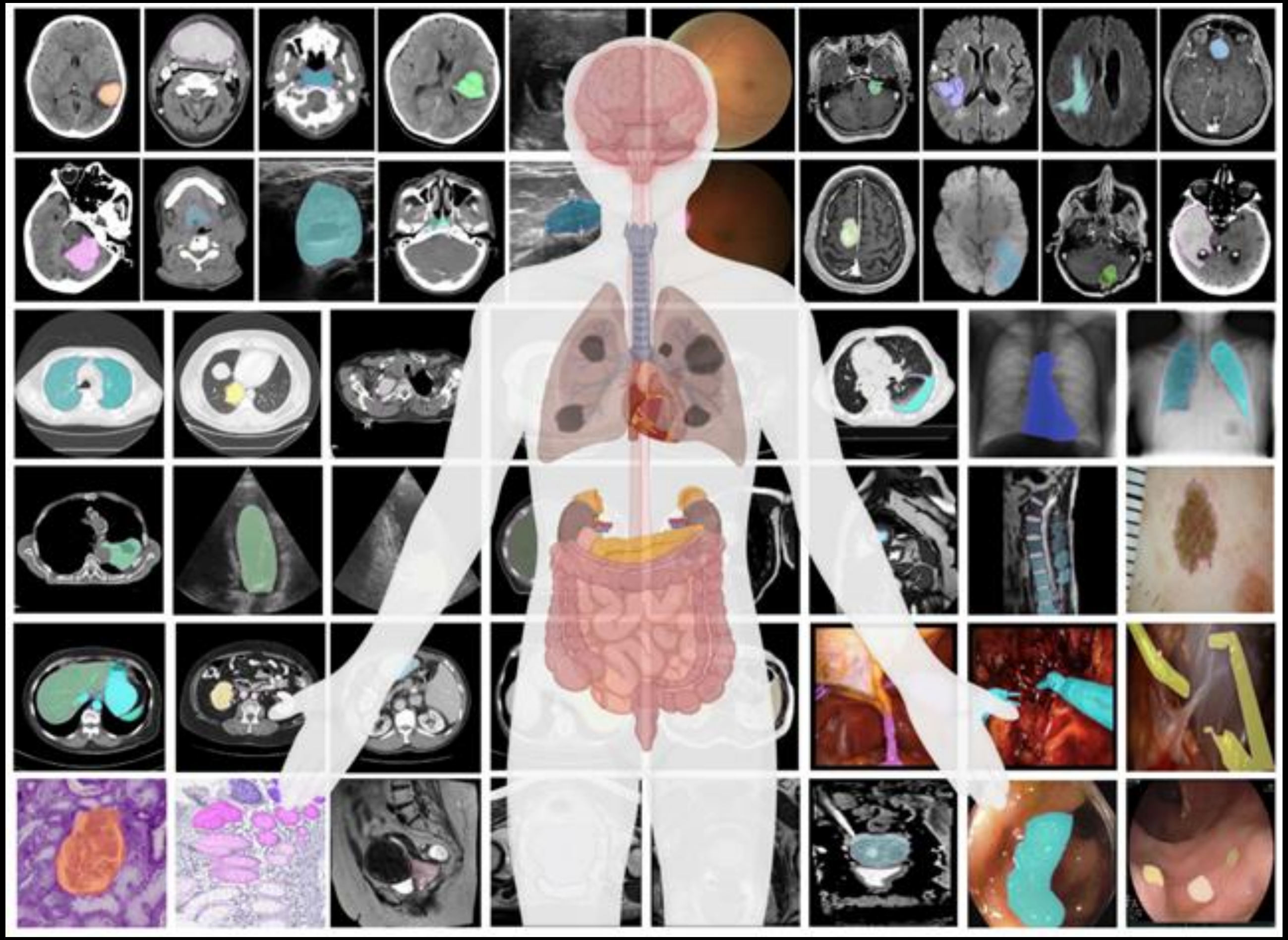
- **Domain-specific tuning: To address the varying results across different contrast appearances and organ morphologies, researchers have explored several domain-specific tuning strategies:**



Lee, H. H., Gu, Y., Zhao, T., Xu, Y., Yang, J., Usuyama, N., Wong, C., Wei, M., Landman, B. A., Huo, Y., Santamaria-Pang, A., & Poon, H. (2024). Foundation Models for Biomedical Image Segmentation: A Survey (No. arXiv:2401.07654). arXiv. http://arxiv.org/abs/2401.07654

MIFM for Segmentation

MedSAM: Segment Anything in Medical Images



1) Developed on a large-scale medical image dataset with 1,570,263 image-mask pairs, covering 10 imaging modalities and over 30 cancer types.

1) Evaluation on 86 internal validation tasks and 60 external validation tasks, demonstrating better accuracy and robustness than modality-wise specialist models.

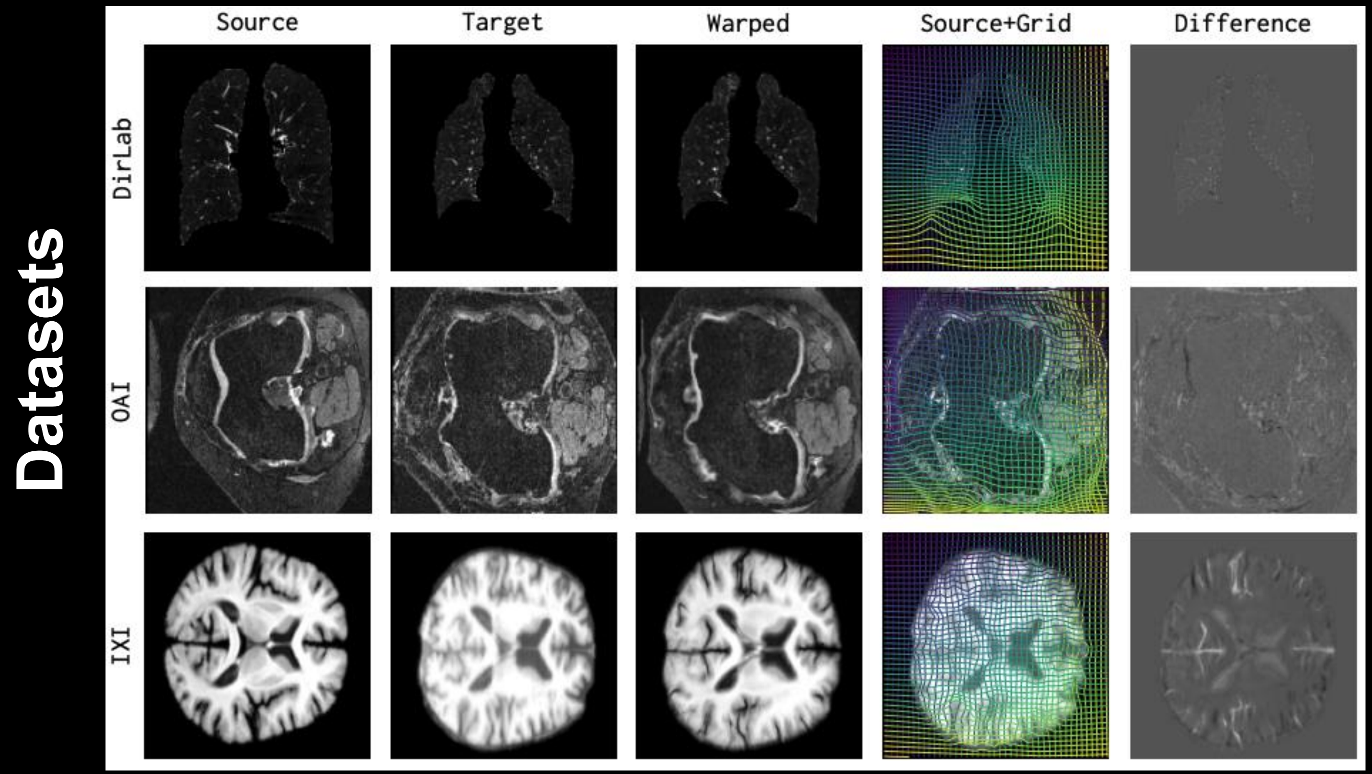
1) Delivering accurate and efficient segmentation across a wide spectrum of tasks.

Ma, …, Wang. Segment Anything in Medical Images. Nature Communications, 2024.

MIFM for Registration

uniGradICON：A Foundation Model for Medical Image Registration

Example uniGradICON Registrations



1) Great performance across multiple datasets which is Not feasible for current learning-based registration methods

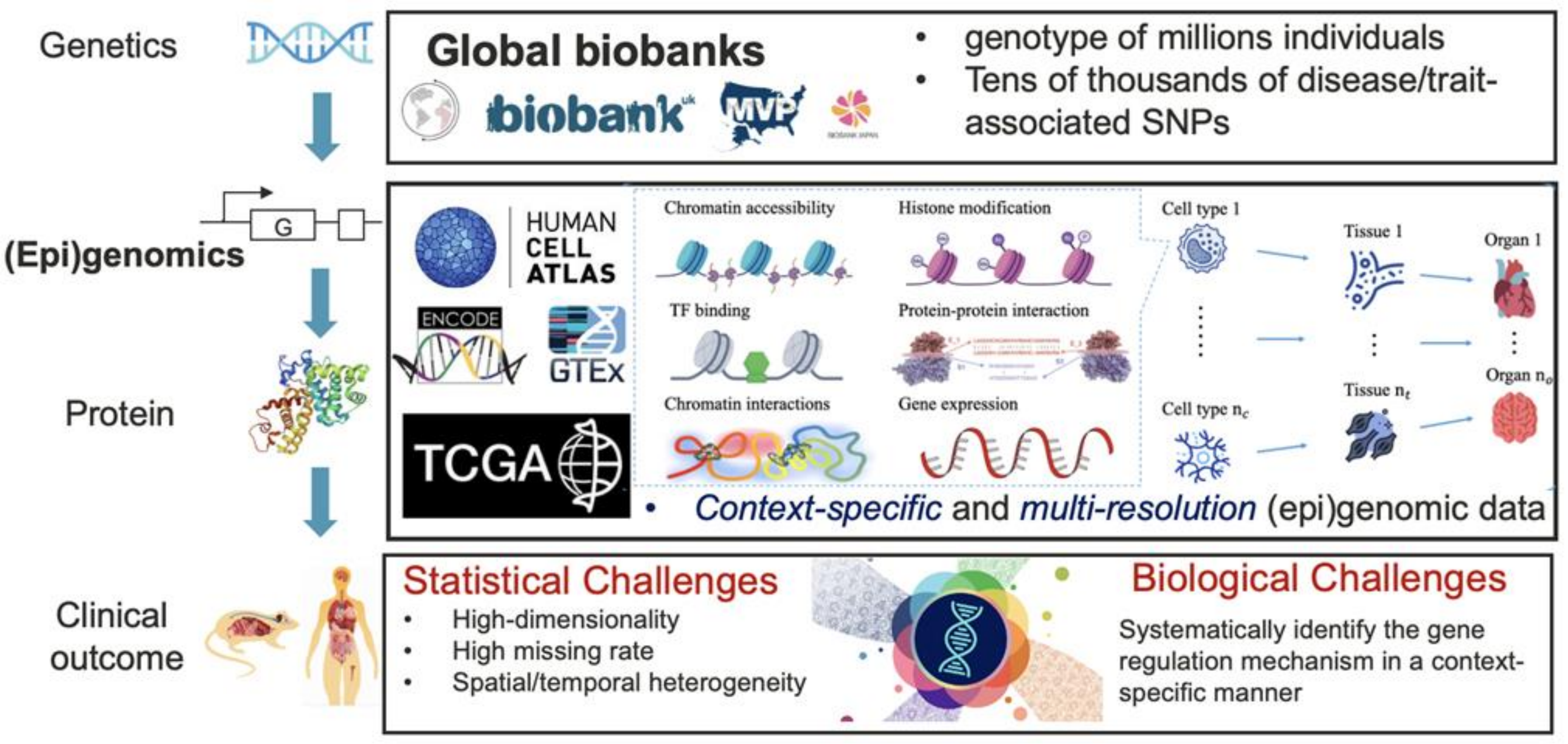
2) Zero-shot capabilities for new registration tasks suitable for different anatomical regions, and modalities

3) A strong initialization for finetuning on out-of-distribution registration tasks

Lin Tian, …, Marc Niethammer. uniGradICON: A Foundation Model for Medical Image Registration. MICCAI, 2024.

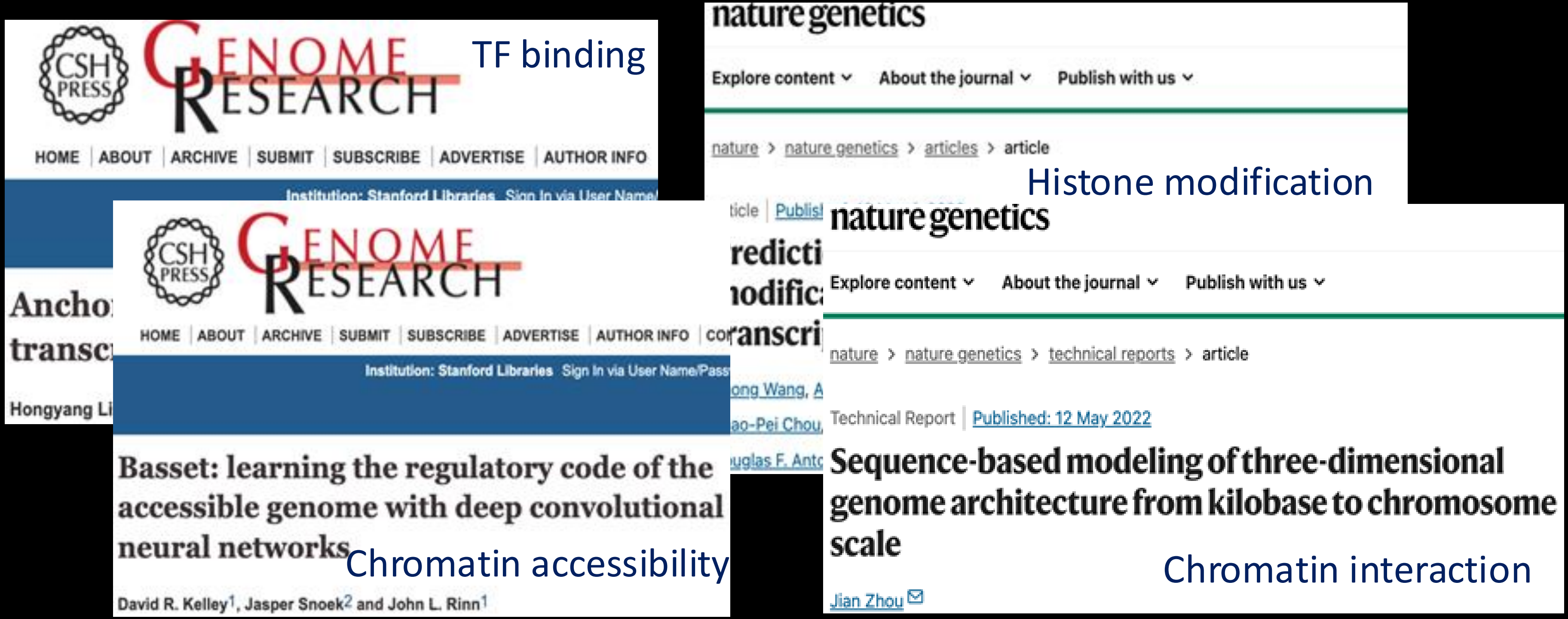
Foundation Models In Omics

- **Accumulated multi-omics data**
- **Challenges in data analysis**



Task-Specific Models For Omics Data

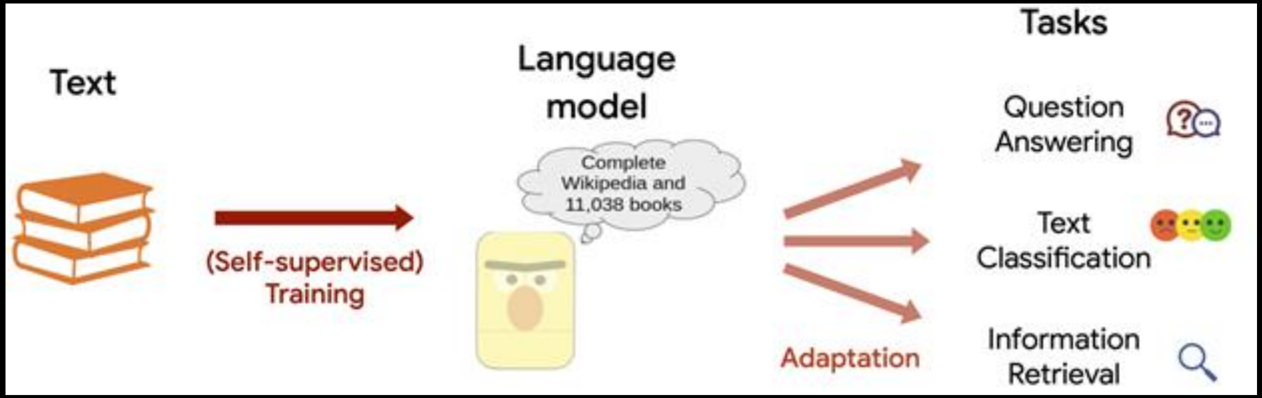
- Computational models are widely developed for analyzing a specific OMIC signal



These models are rather scattered in literature!

Can We Build A Unified FM for Omics?

- Foundation models have been rapidly developed in natural language processing (NLP)



- There is an intrinsic similarity between biological sequence and natural language

“word” ⟷ Nucleotide/amino acid

“sentence” ⟷ DNA/RNA/protein sequence

Recent FMs for Genomic Sequence

Model	# Parameters	Architecture	Training Data	Reference
Big Bird	127M	Transformer	Human ref genome	NeurIPS 2020
DNABERT	86-89M	Transformer	Human ref genome	Bioinformatics 2021
DNABERT2	117M	Transformer	135 species genome	ICLR 2024
Enformer	23M	CNN+Transformer	Human and mouse genome	Nat Methods 2021
Nucleotide Transformer	480M/2537M	Transformer	850 species genome	Nat Methods 2024
HyenaDNA	From 1k to 1M	Hyena operator(long conv)	Human ref genome	arXiv 2023
EpiGePT	71.3M	CNN+Transformer	Human ref genome + TF	Genome Biology 2024
Evo	7B	Striped Hyena Operator	Bacteria + archaea + virus + plasmid	Science 2024
PlantCaduceus	20M to 225 M	State space model	16 Angiosperm genomes	bioRxiv 2024
Evo2	1B/7B/40B	Striped Hyena2 Operator	128k genomes of Eukarya, Prokarya, Archaea	BioRxiv 2025

Recent FMs for Single Cells

Model	# Parameters	Architecture	Training Data	Reference
scGPT	51M	Transformer	33 M normal human cells (transcriptomics)	Nat. Methods 2024
scBERT	5M	Performer	1M+ cells (transcriptomics)	Nat. Mach. Intell. 2022
Geneformer	40M	Transformer	~30M cells (transcriptomics)	Nature 2023
scFoundation	100M	Transformer	~50M cells (transcriptomics)	Nat. Methods 2024
scGPT-spatial	51M	Transformer	~30M spatial profiles (cells/spots)	bioRxiv 2025
EpiFoundation	NA	Transformer	100k paired Multiome profiles	bioRxiv 2025
scLong	1B	Transformer	48M cells (transcriptomics)	bioRxiv 2024
UCE	650M	Transformer	36M cells (transcriptomics)	bioRxiv 2023
AIDO.Cell	From 3M to 650M	Transformer	50M cells (transcriptomics)	bioRxiv 2024

Grand Challenges

EHR

📝 Heterogeneity of structured/unstructured data
⏳ Irregular, sparse temporal trajectories
🔒 Data quality & privacy constraints
🎯 Prediction → causal inference

Medical Imaging

🖼 Multimodal variability (MRI, CT, PET)
📉 Scale vs. limited annotations
🏥 Site/scanner batch effects
⏱ Spatiotemporal progression modeling
🔎 Interpretability for clinical trust

Omics

🧬 Ultra-high dimensionality, small samples
📊 Noise & batch effects across platforms
🔗 Integration across omics layers
⚙ Causal grounding of biological drivers
🏥 Translation to imaging & outcomes

Part III

Vertical and Horizontal Medical Data Integration

The best thing about being a statistician is that you get to play in everyone's backyard.

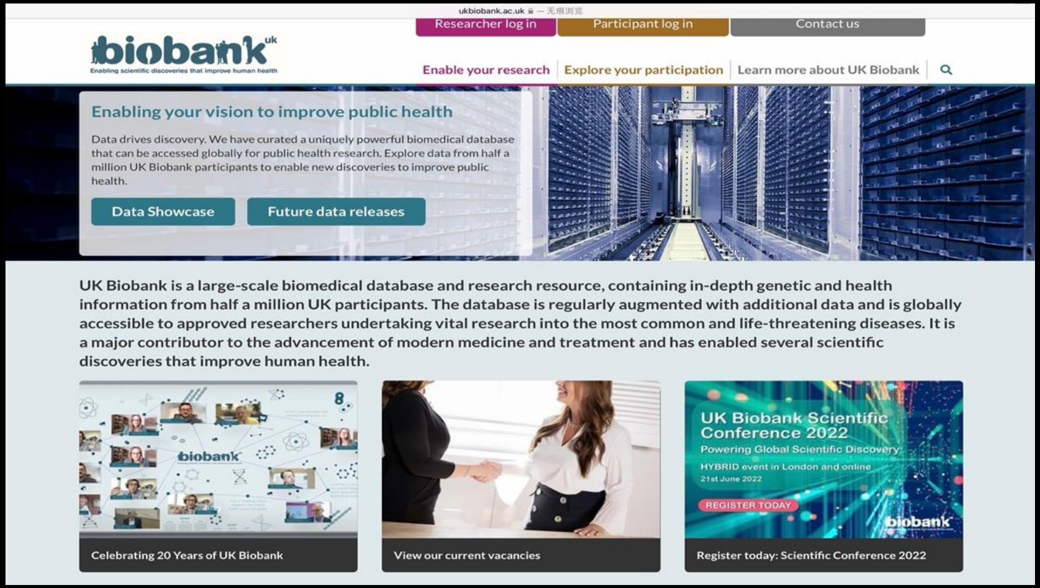
- John Tukey -

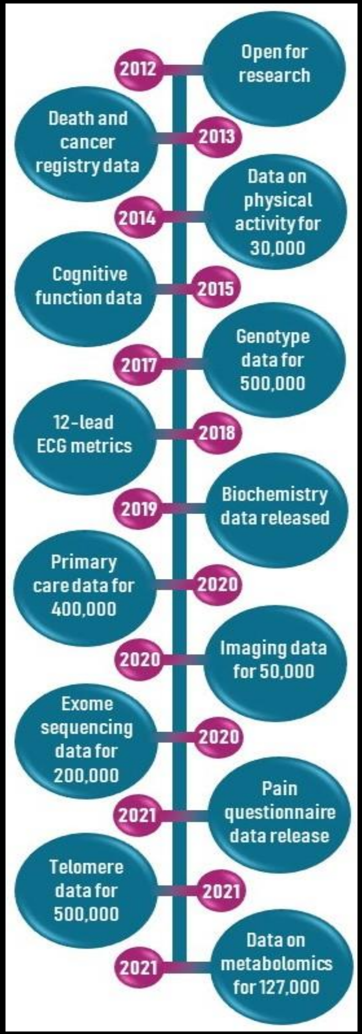
"If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."

- Leo Breiman -

The UK Biobank Study

UK Biobank has collected and continues to collect extensive environmental, lifestyle, and genetic data on half a million participants.



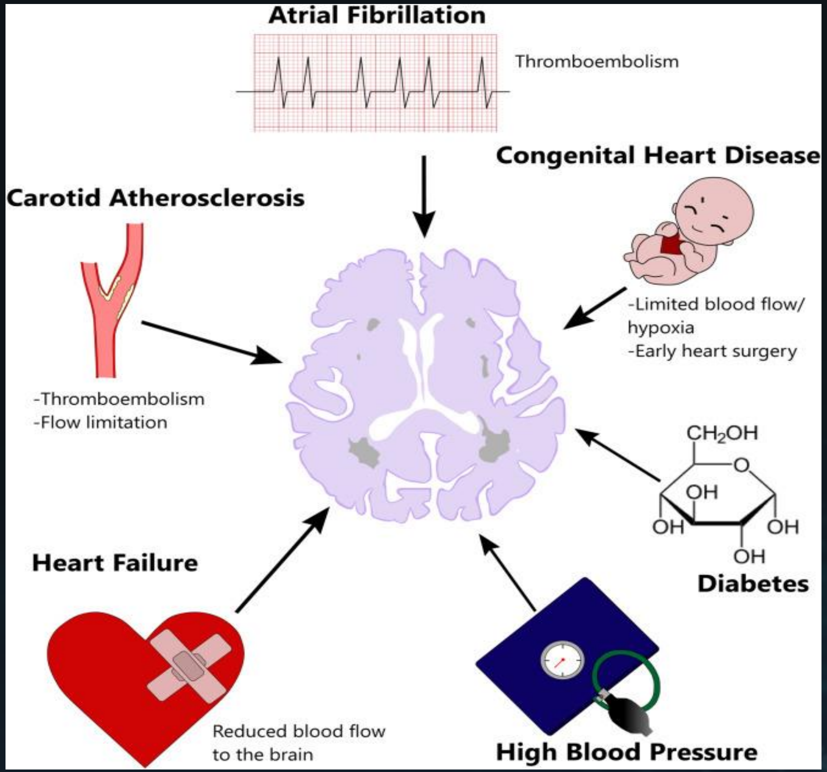


2006-now

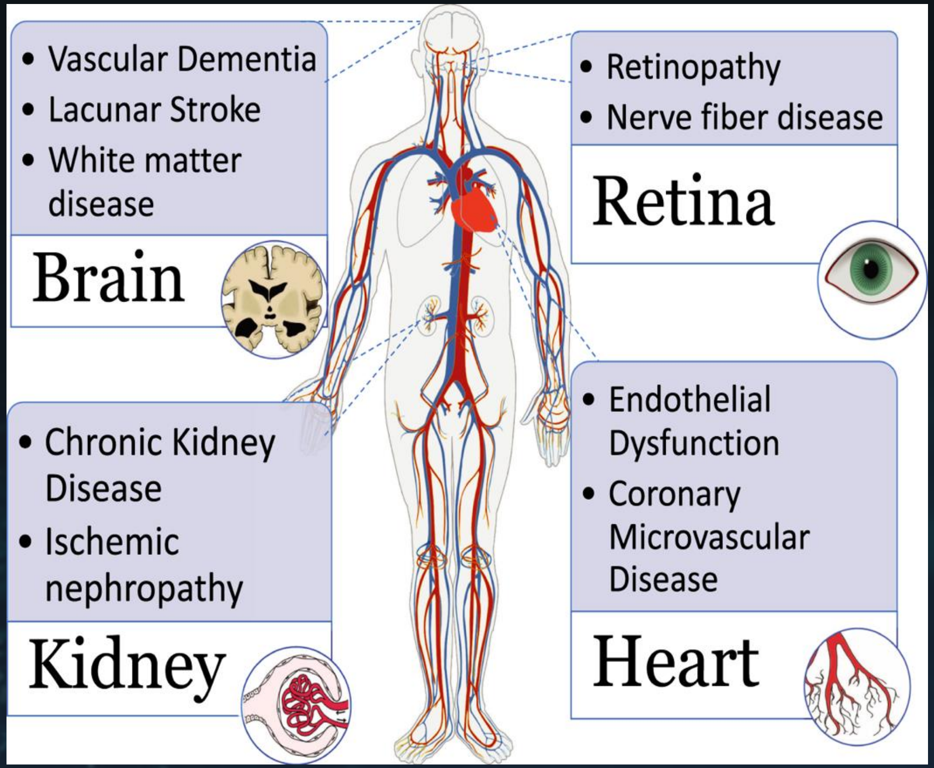
- **Imaging:** Brain, heart and full body MR imaging, plus full body DEXA scan of the bones and joints and an ultrasound of the carotid arteries. The goal is to image 100,000 participants, and to invite participants back for a repeat scan some years later.
- **Genetics:** Genotyping, whole exome sequencing & whole genome sequencing for all participants.
- **Health linkages:** Linkage to a wide range of electronic health-related records, including death, cancer, hospital admissions and primary care records.
- **Biomarkers:** Data on more than 30 key biochemistry markers from all participants, taken from samples collected at recruitment and the first repeat assessment.
- **Activity monitor:** Physical activity data over a 7-day period collected via a wrist-worn activity monitor for 100,000 participants plus a seasonal follow-up on a subset.
- **Online questionnaires:** Data on a range of exposures and health outcomes that are difficult to assess via routine health records, including diet, food preferences, work history, pain, cognitive function, digestive health and mental health.
- **Repeat baseline assessments:** A full baseline assessment is undertaken during the imaging assessment of 100,000 participants.
- **Samples:** Blood & urine was collected from all participants, and saliva for 100,000.

Multiorgan Dysfunction Syndromes

Imaging: help understand the complex interplay between brain and other human organs and their underlying genetic overlaps



Possible causal factors of brain structure changes, resulting in brain disorders like stroke, dementia and cognitive impairment



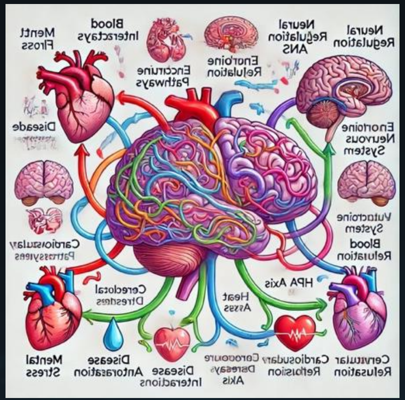
Many diseases (e.g., microvascular disease, high blood pressure) are multisystem disorders

The Brain-Heart Axis

The brain-heart axis refers to the bidirectional communication between the brain and the heart, playing a crucial role in regulating physiological functions and maintaining overall health.

Neural Regulation:

- **Autonomic Nervous System (ANS):** regulate heart rate, blood pressure, and cardiac output.
- **Vagus Nerve:** reduce heart rate and promoting relaxation.



Endocrine Pathways:

- **Hypothalamic-Pituitary-Adrenal (HPA) Axis:** Influences heart function through the release of hormones, affecting blood pressure and cardiovascular health.
- **Catecholamines:** Adrenaline and noradrenaline released during stress increase heart rate and cardiac output.

Blood Flow and Oxygen Supply:

- **Cerebral Perfusion:** The heart ensures a continuous supply of oxygenated blood to the brain, essential for cognitive functions and neural health.
- **Cerebral Autoregulation:** Mechanisms that maintain stable blood flow to the brain despite changes in systemic blood pressure.

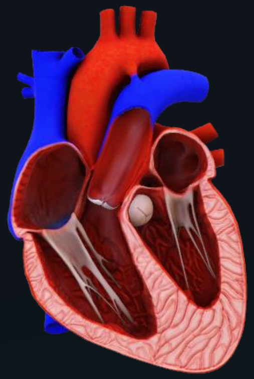
The Brain-Heart Axis

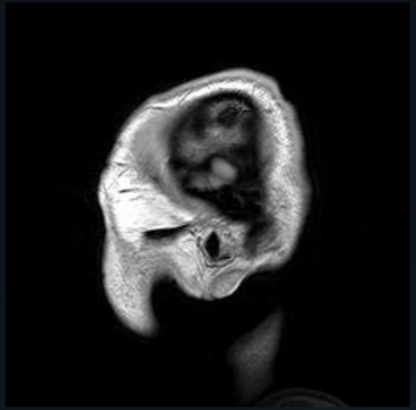
Disease Interactions:

- **Cardiovascular Diseases:** Conditions like atrial fibrillation and heart failure are linked to brain diseases such as stroke, dementia, and cognitive impairment due to reduced cerebral perfusion.
- **Mental Disorders:** Mental illnesses, including schizophrenia, bipolar disorder, epilepsy, and depression, increase the risk of CVD.

Acute Mental Stress:

- **Impact on Cardiovascular Health:** Acute stress can cause vascular inflammation and increase the risk of atherosclerosis due to stress-induced leukocyte migration.



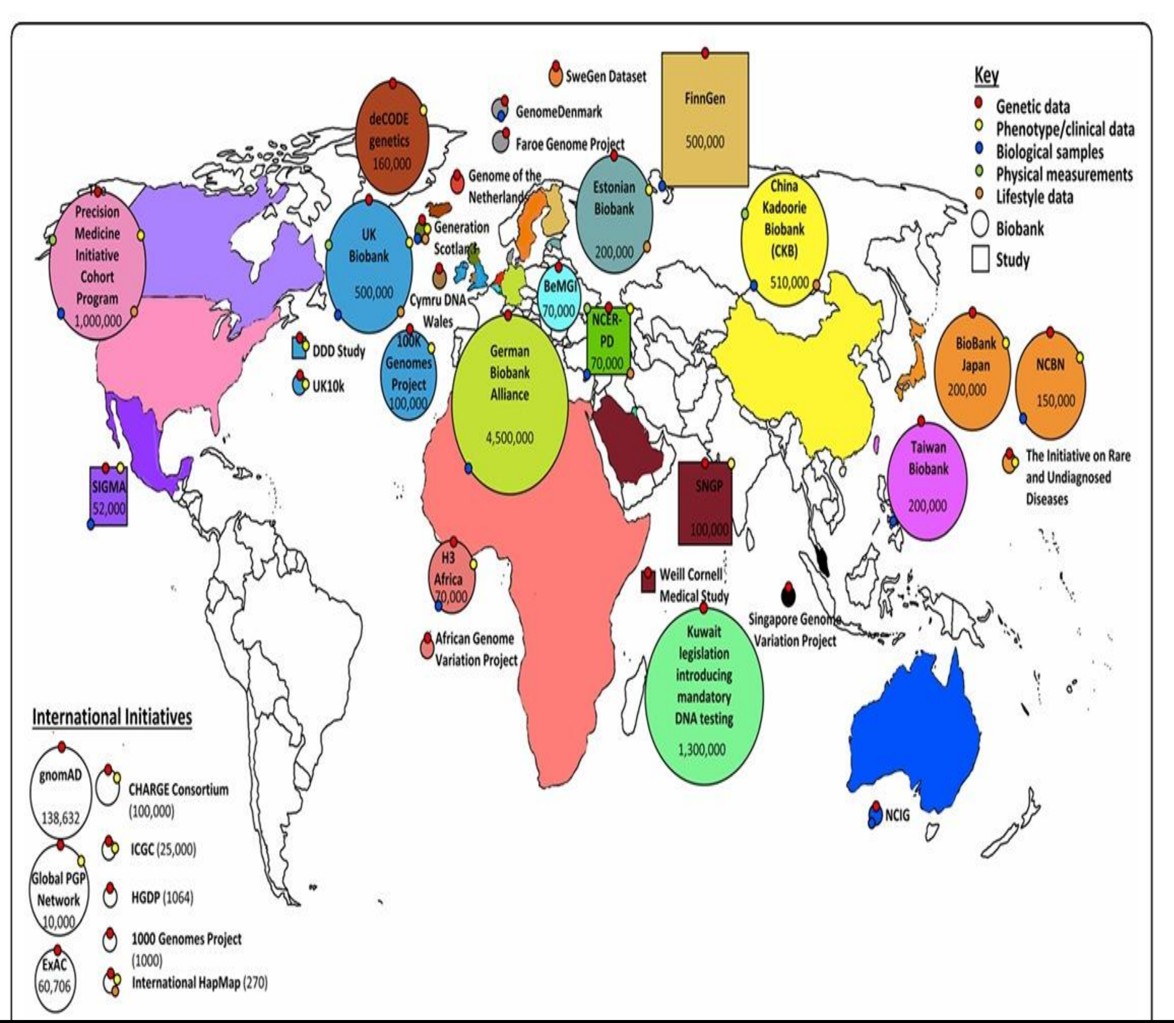


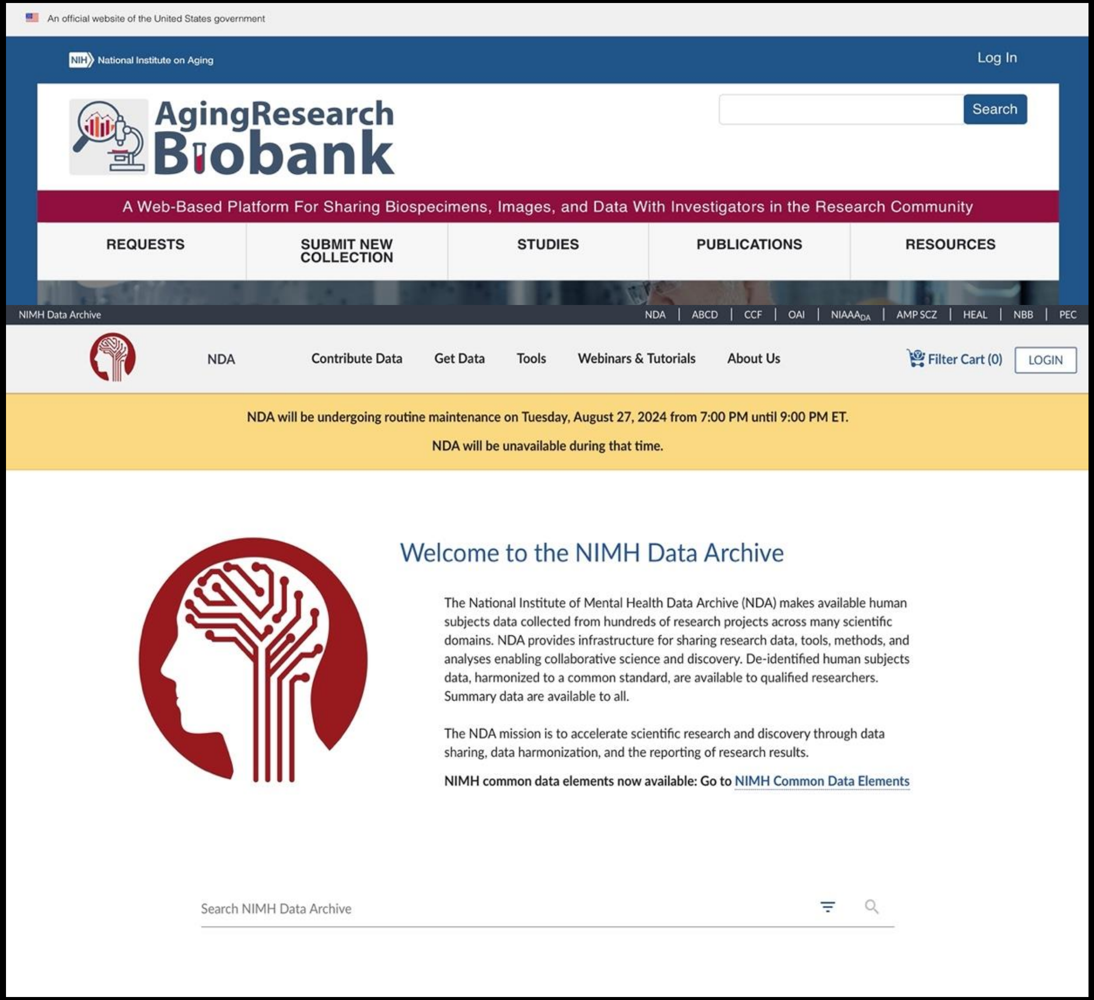
Research Significance:

•**Integrated Treatment Approaches:** Lead to better treatments for neurocardiological disorders.

•**Comprehensive Studies:** A need for larger studies to provide a complete picture of the structural and functional links between heart and brain health.

World Biobanks and NIH Data





Hannah Carress, Daniel John Lawson and Eran Elhaik. Population genetic considerations for using biobanks as international resources in the pandemic era and beyond. BMC Genomics. 2021.

Biomedical Data Resources

Literature

- ❖ Peer-reviewed articles, preprints, case reports
- ❖ Sources: PubMed, bioRxiv, medRxiv

Ontologies

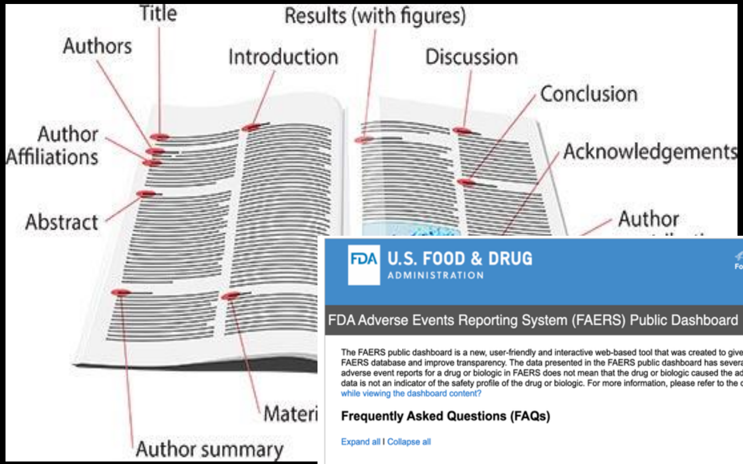
- ❖ Standard vocabularies for data harmonization
- ❖ Examples: UMLS, SNOMED CT, ICD-10, MeSH

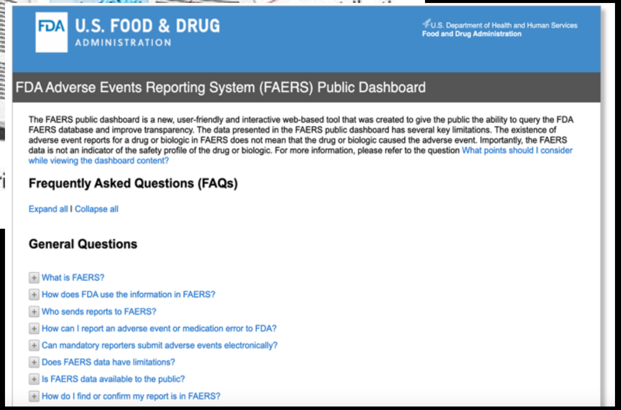
Drug Information

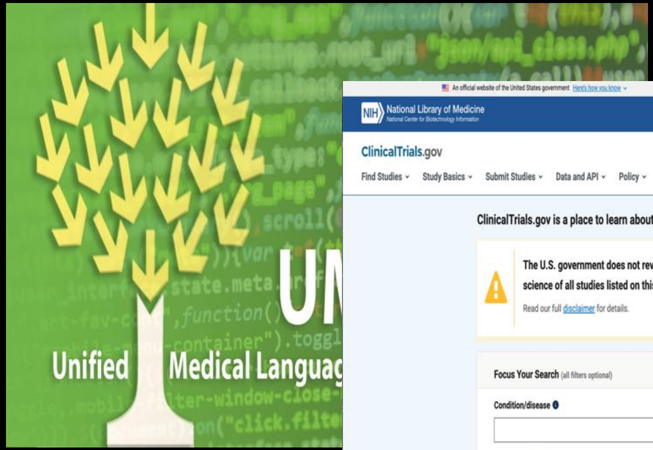
- ❖ Prescription and utilization records
- ❖ Pharmacogenomic annotations (gene–drug interactions)
- ❖ Drug databases and adverse-event reports (e.g., FAERS)

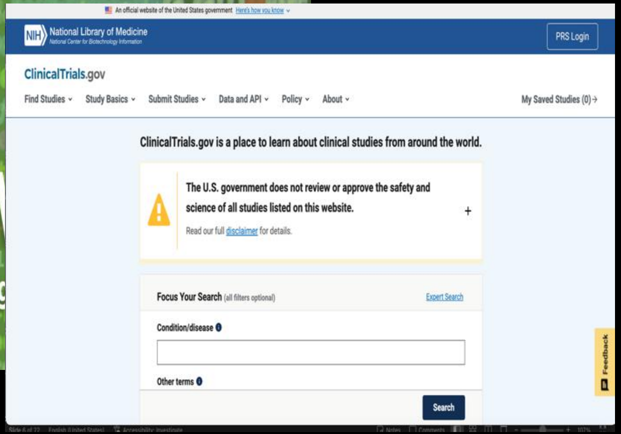
Clinical Trials & Registries

- ▪ Interventional protocols and real-world outcomes
- ▪ Examples: ClinicalTrials.gov, SEER cancer registry



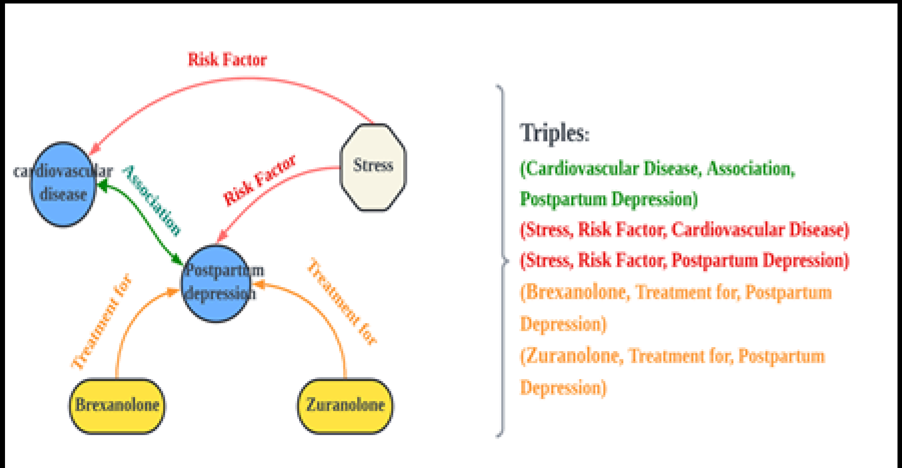






Biomedical Knowledge Graph (BKG)

- **What?** A graph of nodes (entities) and edges (relations) capturing biomedical facts
- Entities: diseases, genes, proteins, drugs, phenotypes
- Relations: ”causes”, ”interacts with”, ”treats”, ”associated with”
- Attributes: provenance, confidence scores, timestamps
- Enables multi-hop reasoning and semantic queries



- **Why?** Summary the old knowledge and discover new insights
- Complex Network Data: Biomedical inherently contains many complex network data, such as gene networks and brain networks.
- Scattered Knowledge: Knowledge is scattered across various data sources.
- Advanced Algorithm Integration: Graph structures can be combined with advanced algorithms to facilitate downstream applications, such as drug repurposing, drug discovery, and disease prediction.

Knowledge Graph Construction

Computers in Biology and Medicine 192 (2025) 110285

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ELSEVIER

Alzheimer's disease knowledge graph enhances knowledge discovery and disease prediction

Yue Yang [a], Kaixian Yu [b], Shan Gao [c], Sheng Yu [d], Di Xiong [e], Chuanyang Qin [c], Huiyuan Chen [c], Jiarui Tang [a], Niansheng Tang [c], Hongtu Zhu [a,*]

nature communications

Article

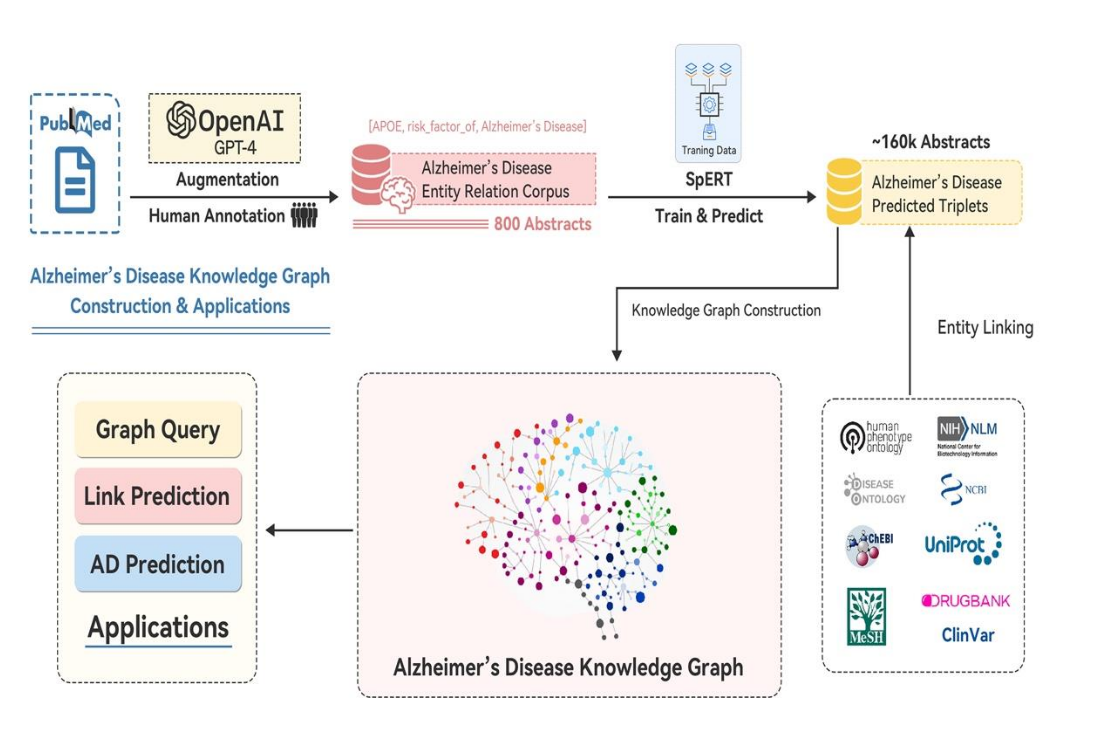
https://doi.org/10.1038/s41467-025-62781-z

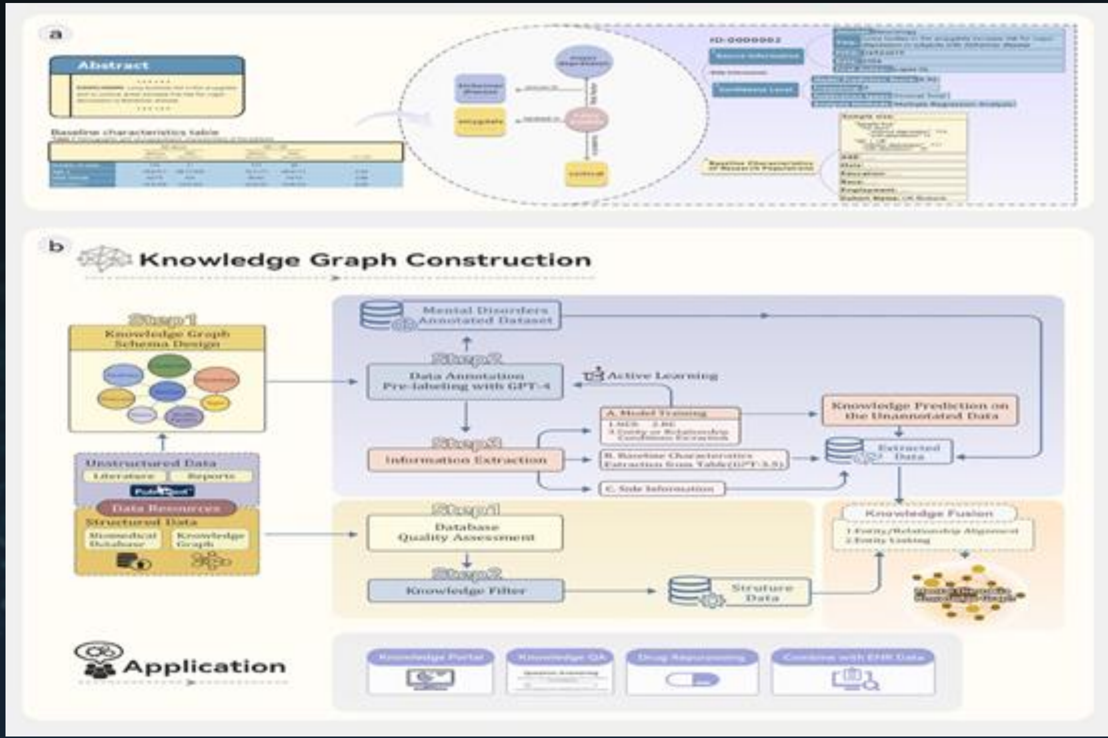
Large language model powered knowledge graph construction for mental health exploration

Received: 26 August 2024

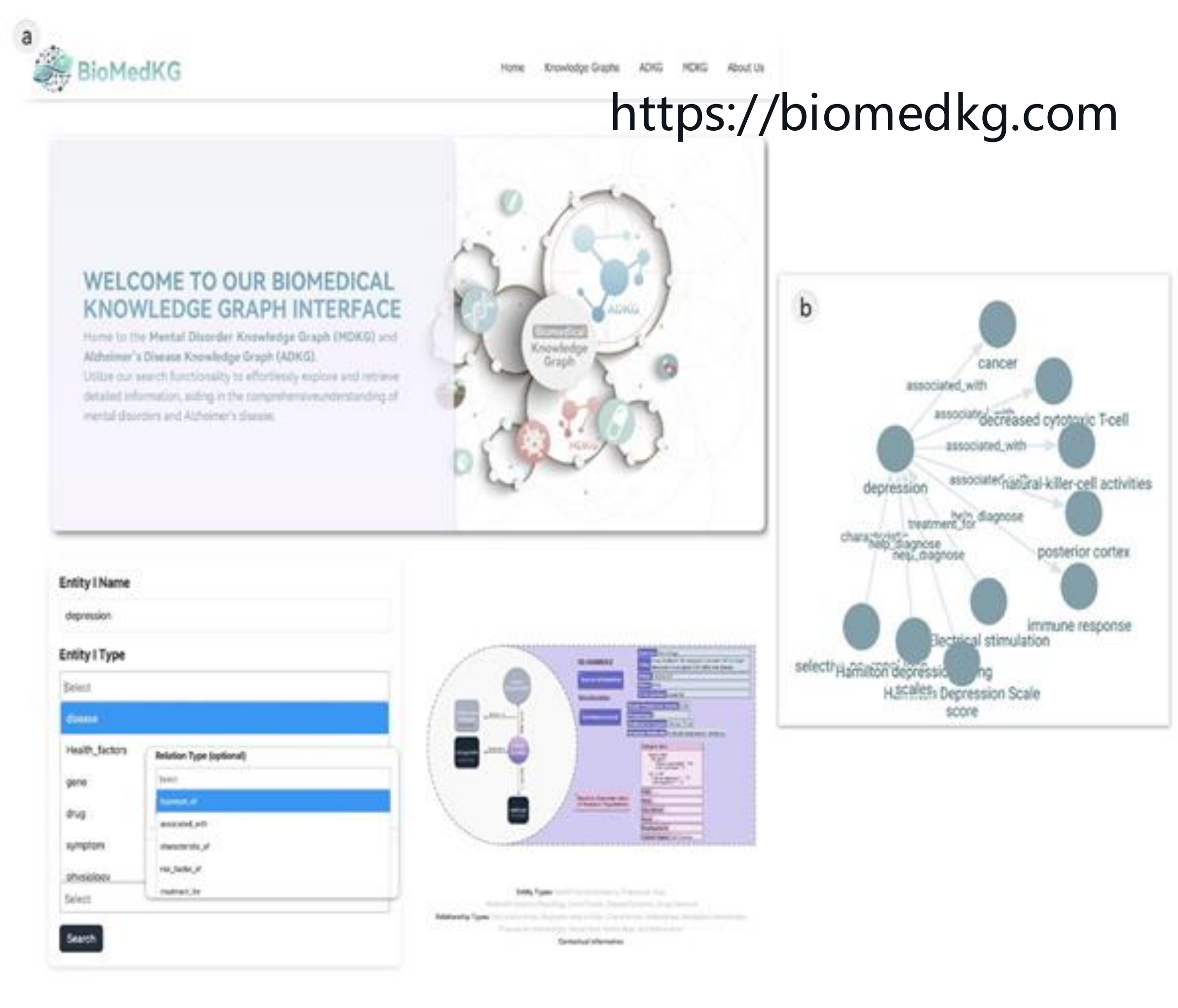
Accepted: 29 July 2025

Shan Gao[1], Kaixian Yu[2], Yue Yang[3], Sheng Yu[4], Chenglong Shi[5], Xueqin Wang[6], Niansheng Tang[1] & Hongtu Zhu[3,7,8,9,10]

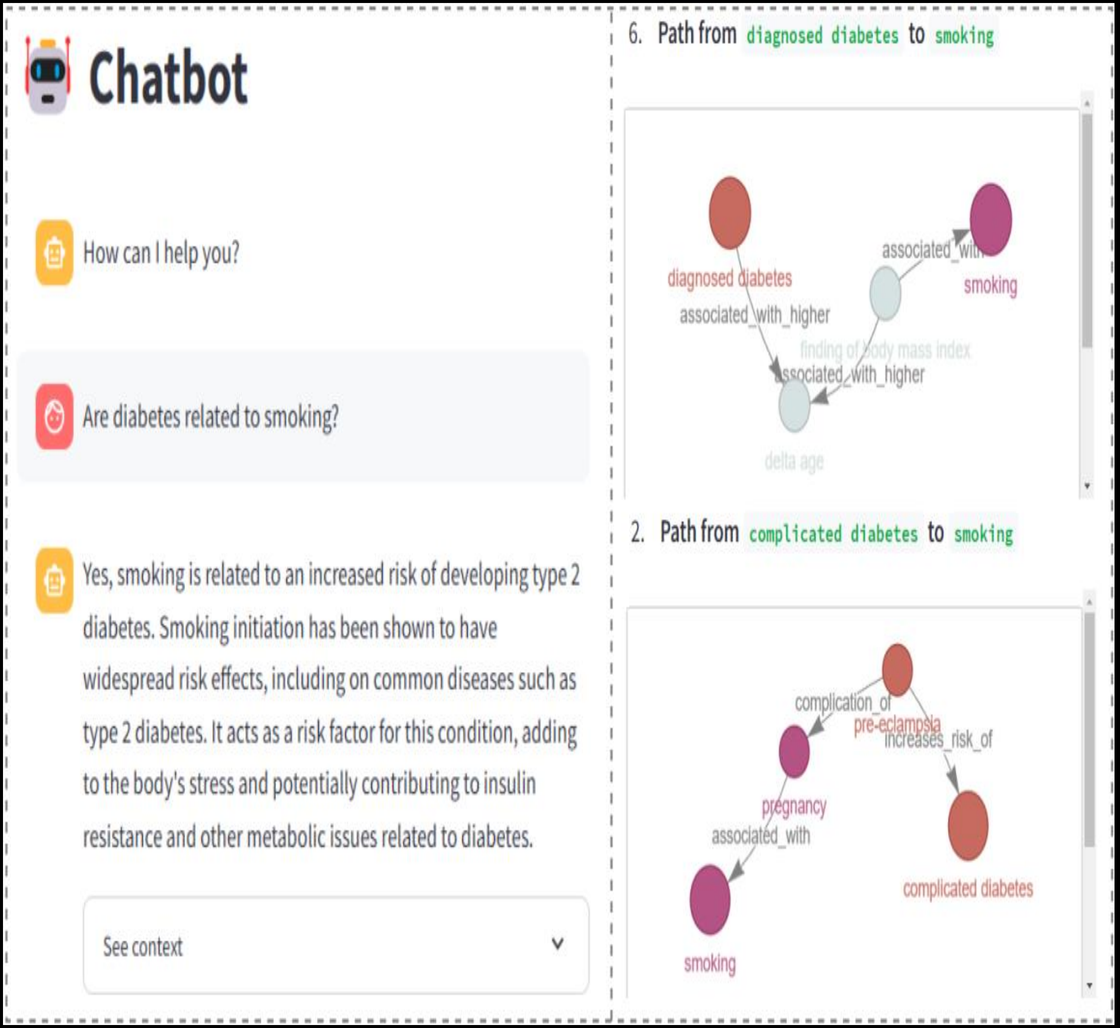




Our Knowledge Graph



https://biomedkg.com



Grand Challenges

Vertical + High Dimensionality

- Omics + imaging + EHR feature explosion
- Multi-scale alignment
- Dimensionality reduction pipelines

Horizontal + High Dimensionality

- Sparse & heterogeneous site-level features
- Non-overlapping variable sets
- Scalable harmonization

Vertical + Other Challenges

- Cross-scale causal inference
- Biological interpretability
- Data harmonization

Horizontal + Other Challenges

- Cohort heterogeneity
- Site/scanner effects
- Privacy & governance

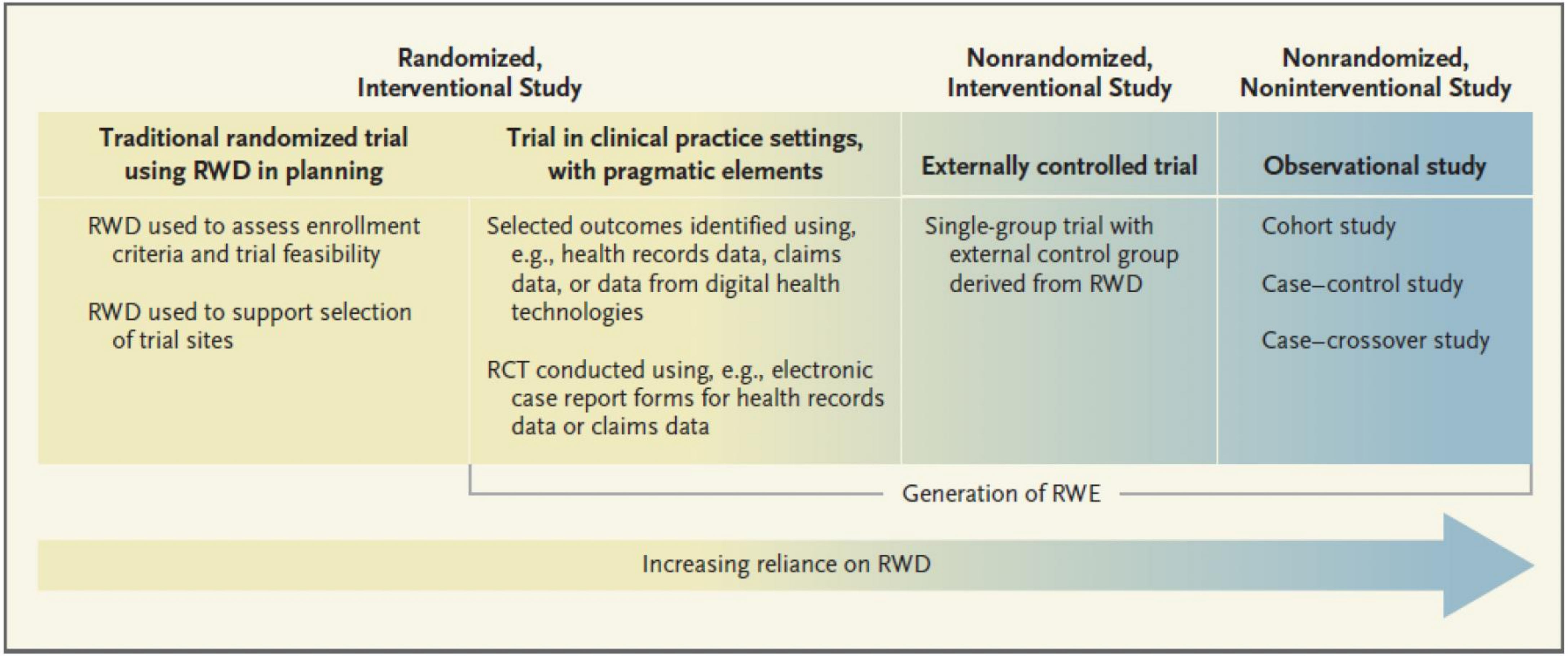
Part IV

Causal Decision Making and Drug Discovery

“Causation is not merely a useful concept, it is fundamental to our understanding of the world. Without causal inference, we are merely describing patterns, not explaining them.”

-Judea Pearl-

Real World Evidence (RWE)

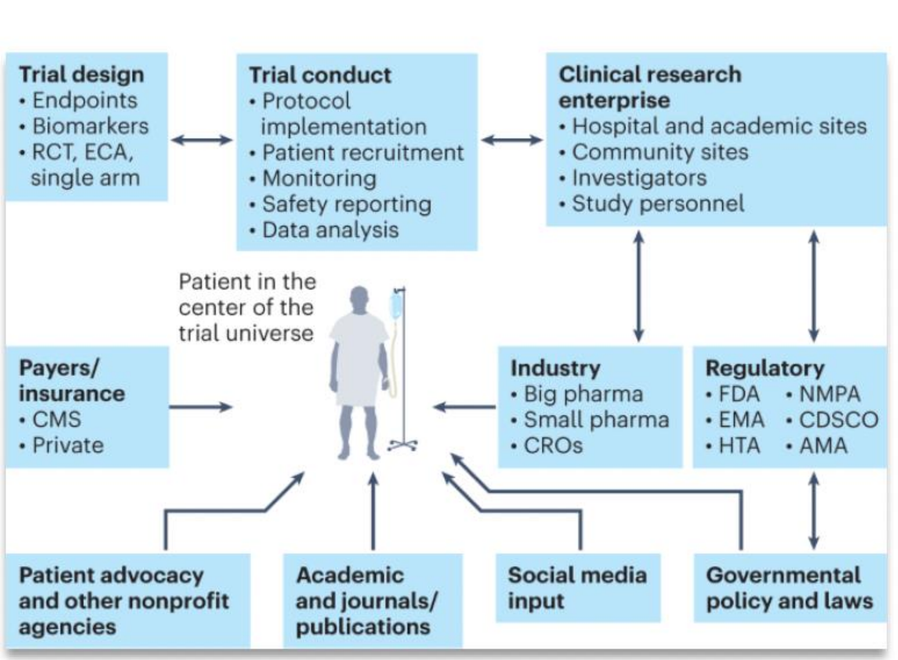


Reliance on RWD in Representative Types of Study Design.

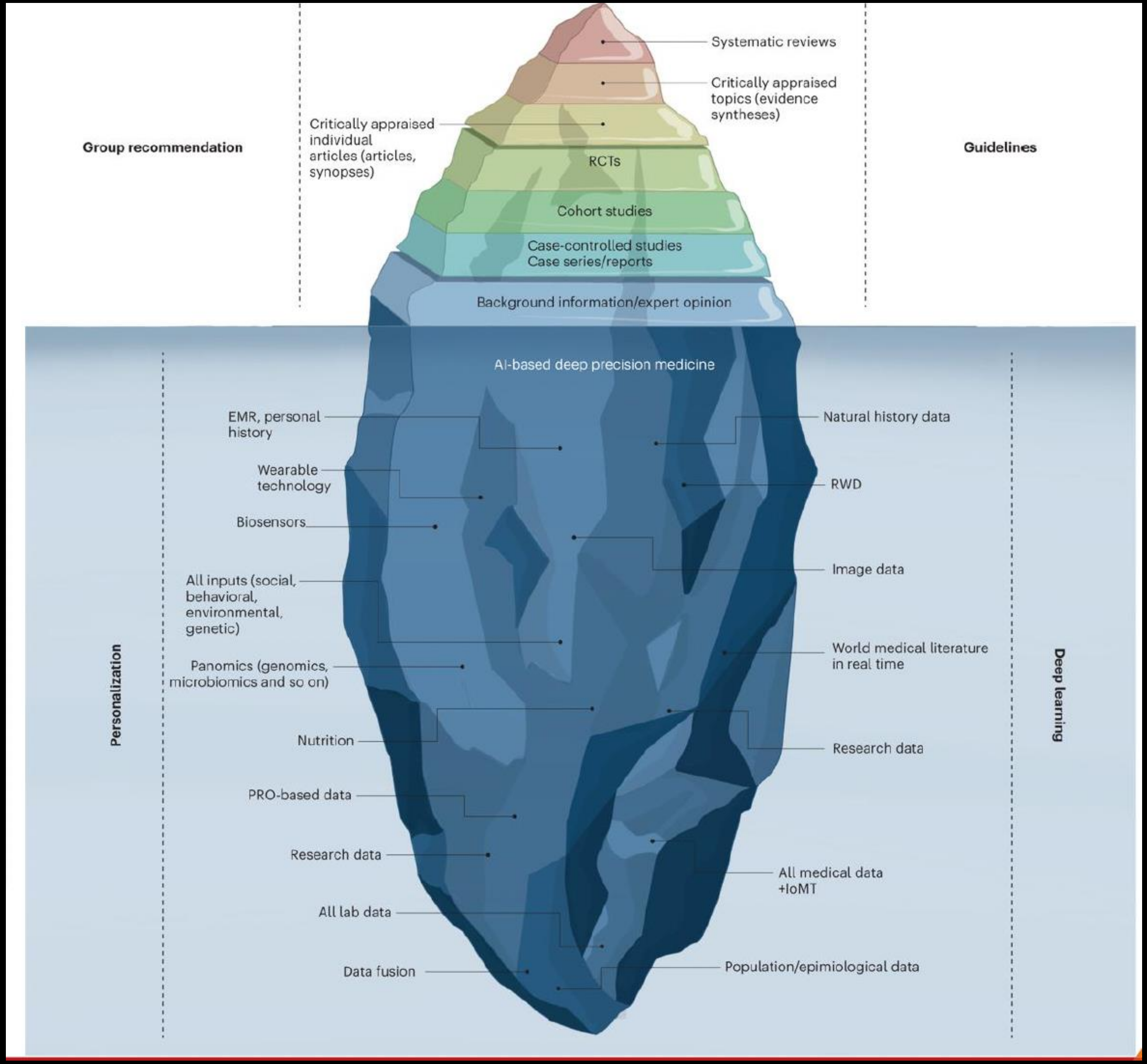
RCT denotes randomized, controlled trial; RWD real-world data; and RWE real-world evidence.

Concato, John, and Jacqueline Corrigan-Curay. "Real-world evidence—where are we now?." *New England Journal of Medicine* 386, no. 18 (2022): 1680-1682.

Evidence-based Medicine

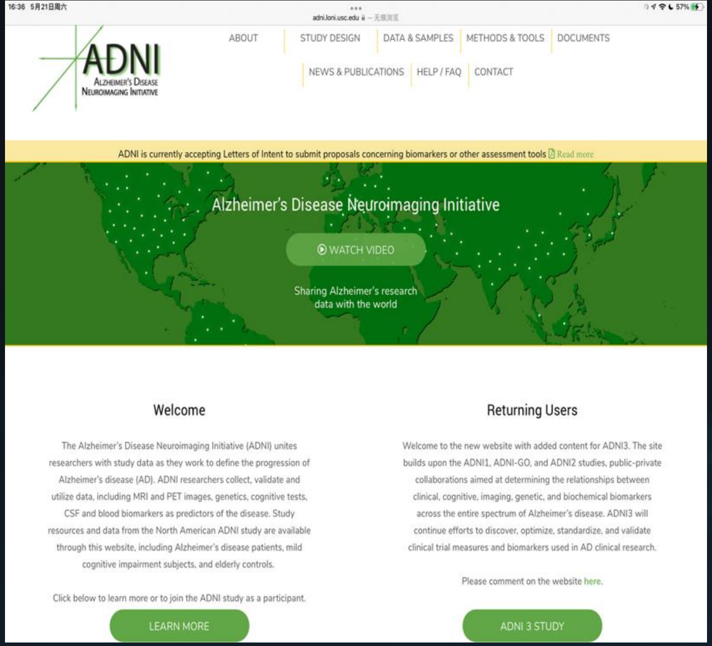


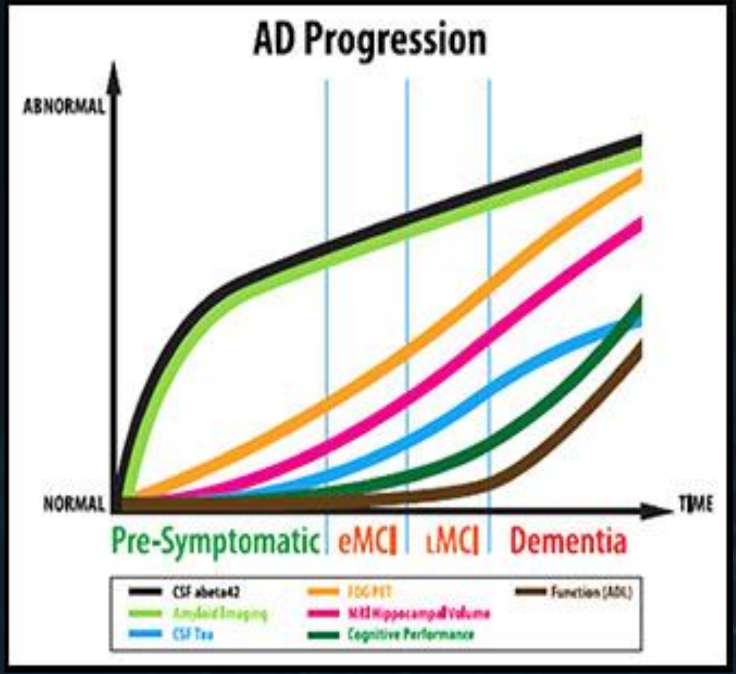
Subbiah, Vivek. "The next generation of evidence-based medicine." *Nature medicine* 29, no. 1 (2023): 49-58.

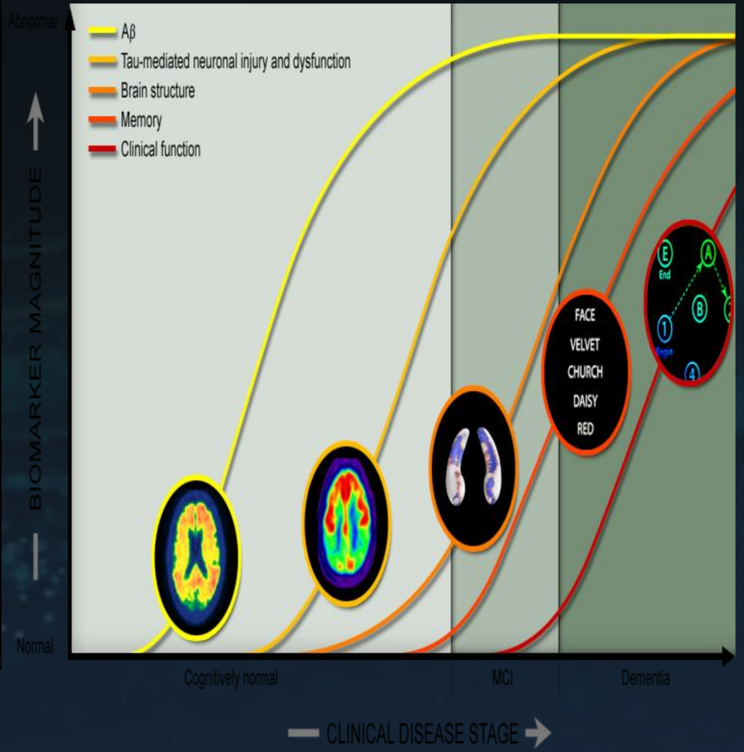


Alzheimer's Disease Neuroimaging Initiative

The overall goal of ADNI is to validate potentially useful biomarkers for AD clinical treatment trials. ADNI is a longitudinal, disease-targeted, multi-center clinical cohort and actively supports the investigation and development of treatments that may slow or stop the progression of AD https://adni.loni.usc.edu/study-design. Researchers across 63 sites in the US and Canada have been tracking the progression of AD through clinical, imaging, genetic and biospecimen biomarkers, starting from normal aging, early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) to dementia or AD.





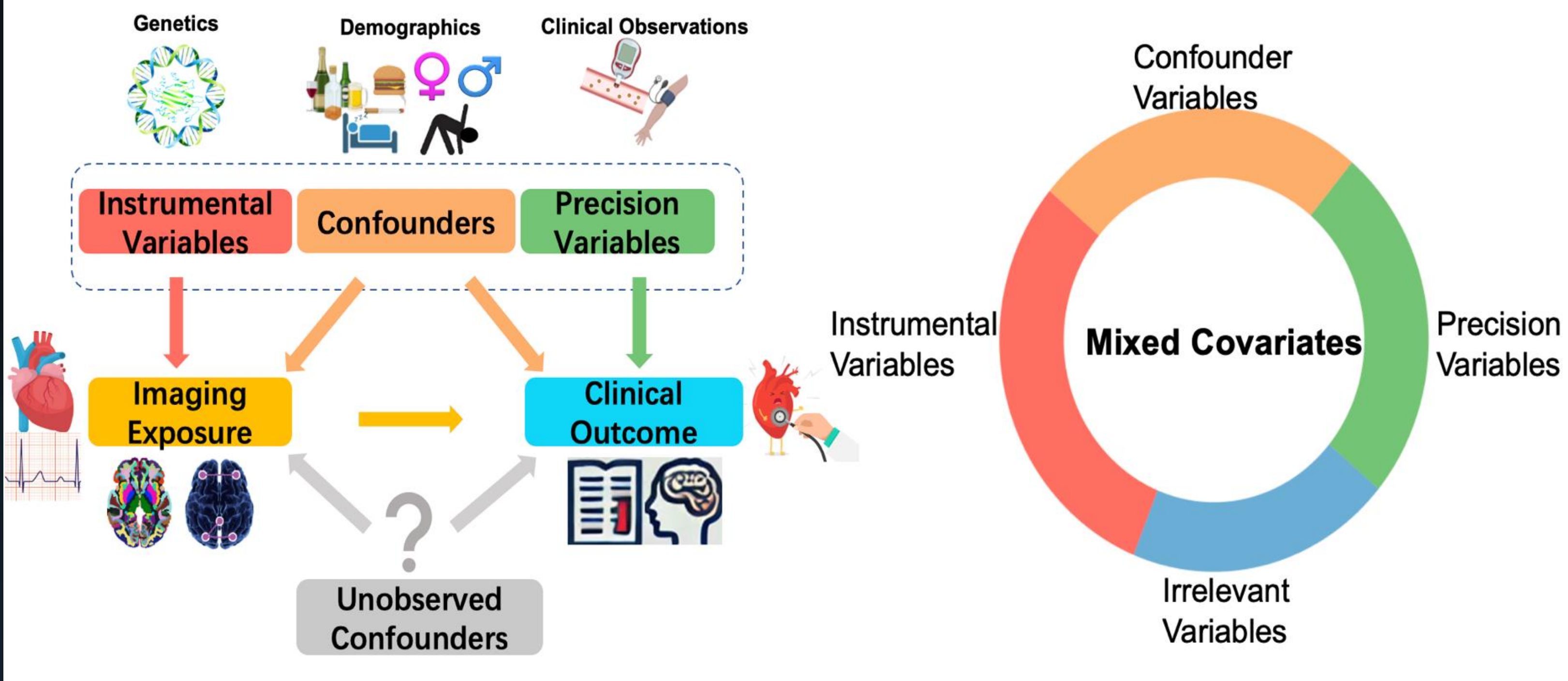


2004-now

Causal Imaging Genetic Models

Outcome generating model. $Y_i = \sum_{l=1}^{s} x_{il}\,\beta_l + < \boldsymbol{Z}_i, \boldsymbol{B} > + \epsilon_i$

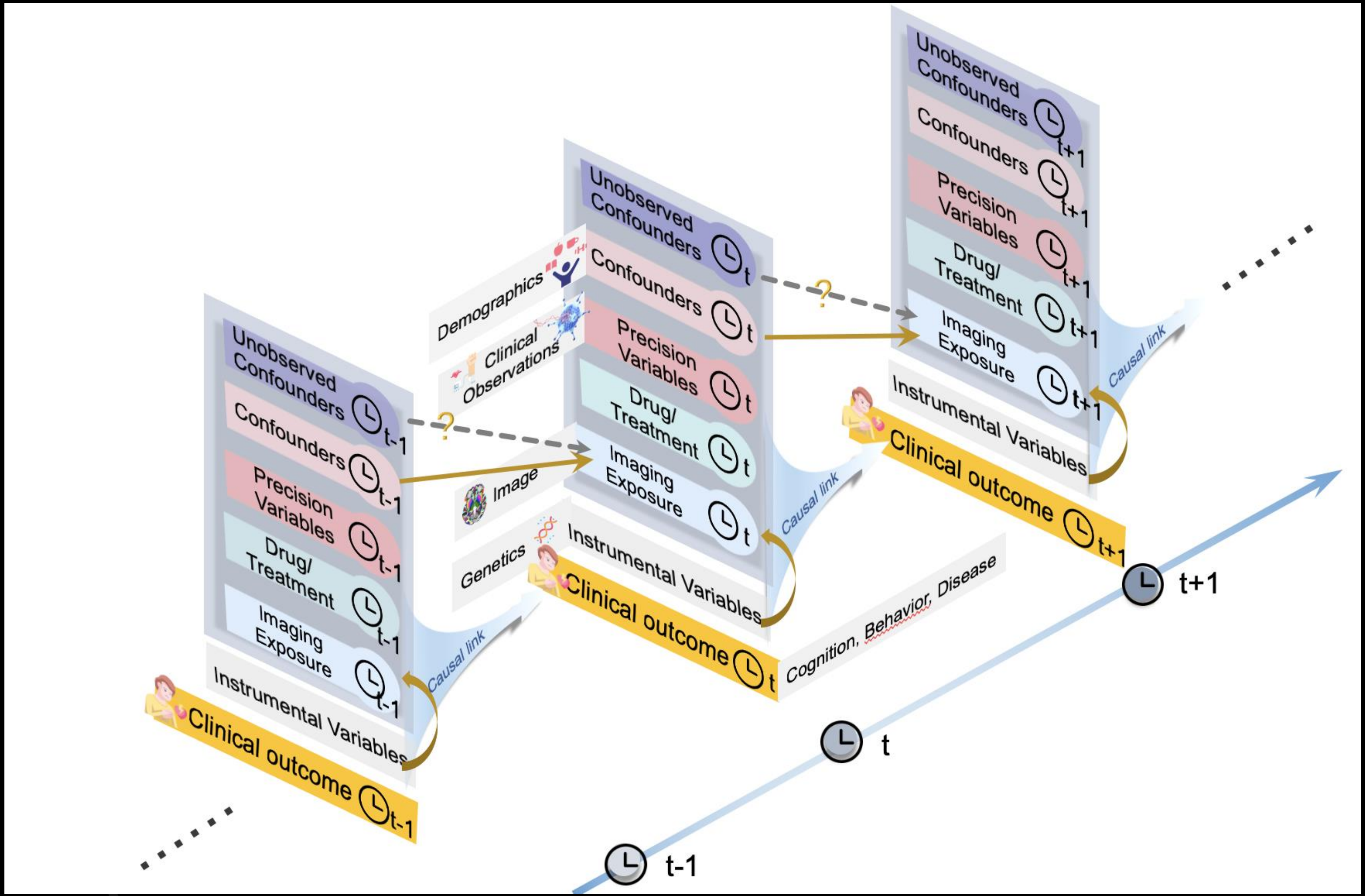
Exposure generating model. $\boldsymbol{Z}_i = \sum_{l=1}^{s} \mathrm{x}_{\mathrm{il}} * \boldsymbol{C}_l + \boldsymbol{E}_i$



(a) Directed acyclic graph

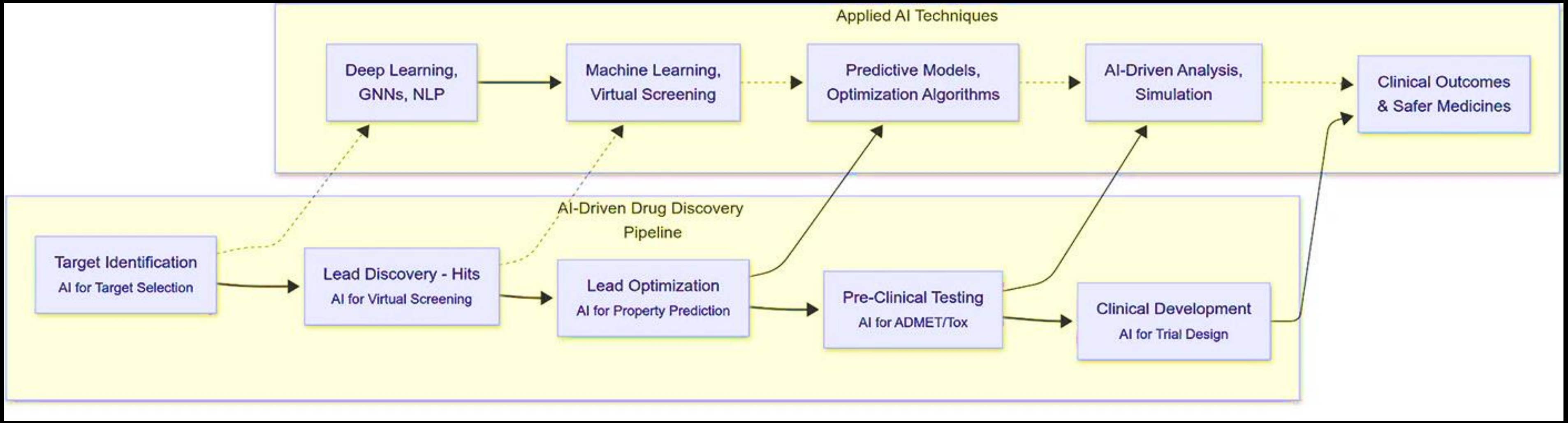
(b) Four groups of the scalar covariates

Dynamic Causal Imaging Genetic Models

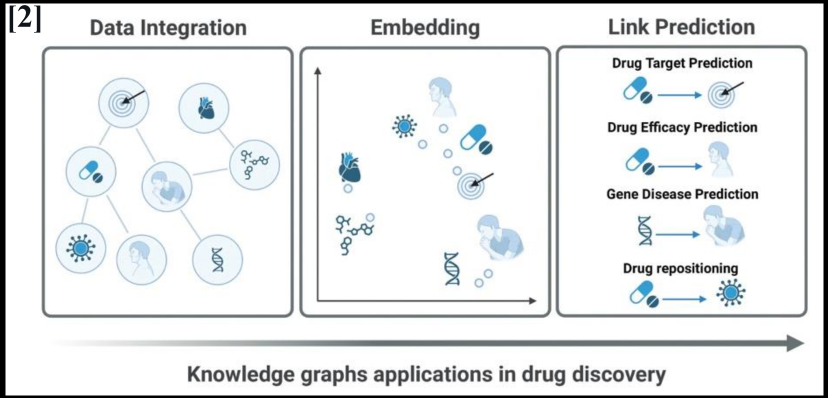


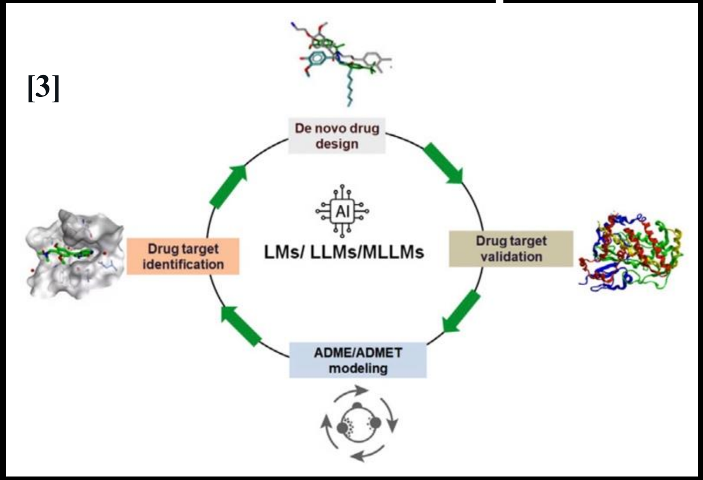
AI-driven Drug Discovery

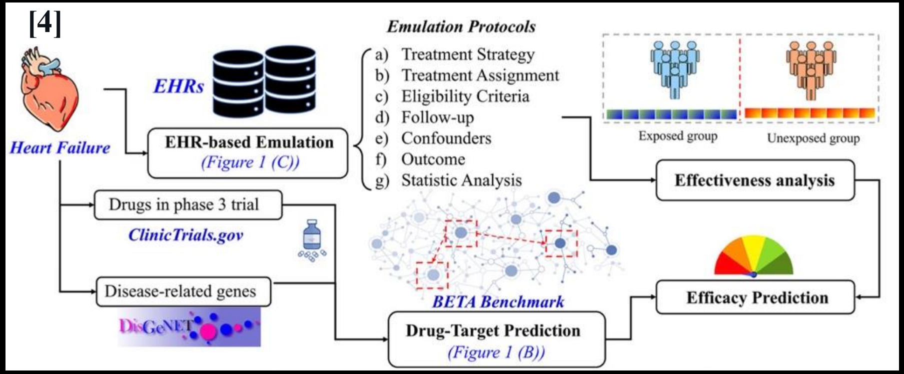
[1]



KG, LLM/Agents, EHR enhance different phases in drug discovery



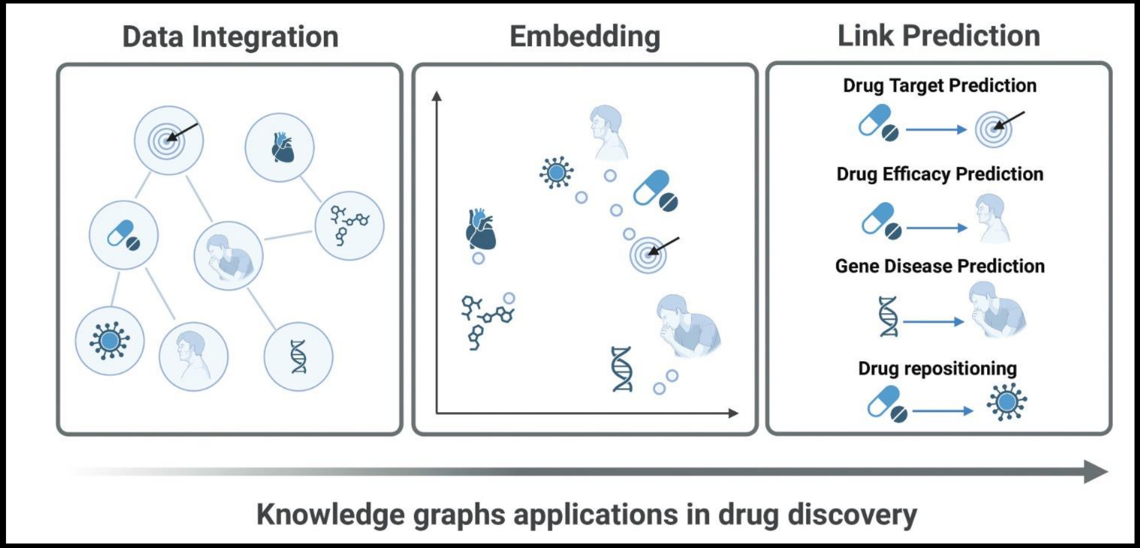


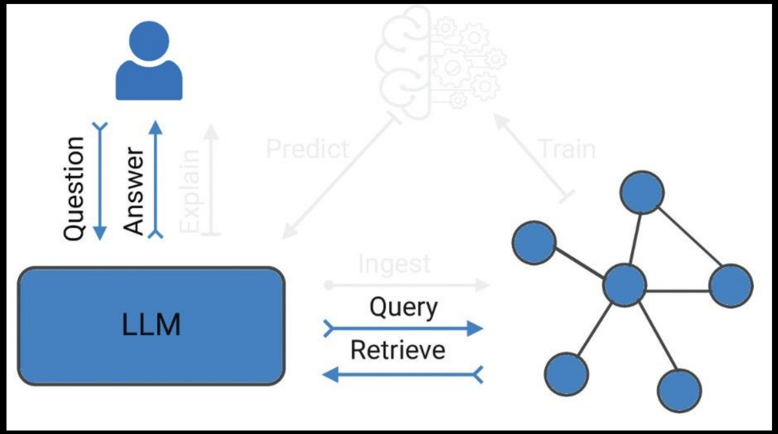


[1] Ferreira F J N, Carneiro A S. AI-Driven Drug Discovery: A Comprehensive Review[J]. ACS omega, 2025.
[2] An update on knowledge graphs and their current and potential applications in drug discovery
[3] Chakraborty C, Bhattacharya M, Pal S, et al. Ai-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development[J]. Journal of Advanced Research, 2025.
[4] Zong N, Chowdhury S, Zhou S, et al. Advancing efficacy prediction for electronic health records based emulated trials in repurposing heart failure therapies[J]. npj Digital Medicine, 2025, 8(1): 306.

Biomedical Knowledge in Drug Discovery

[1] [1]





Core tasks

Cross-source integration & standardization: entity alignment, ontology harmonization.

Relation modeling & link prediction: DTI / DDR / DRD completion; pathway hierarchies

Affect which R&D stages

Target discovery: pathway + causal evidence → prioritized, druggable targets.

Hit/lead: mechanism / network-constrained screening and repurposing.

Preclinical: Evidence packages; potential adverse / interaction (DDI) paths.

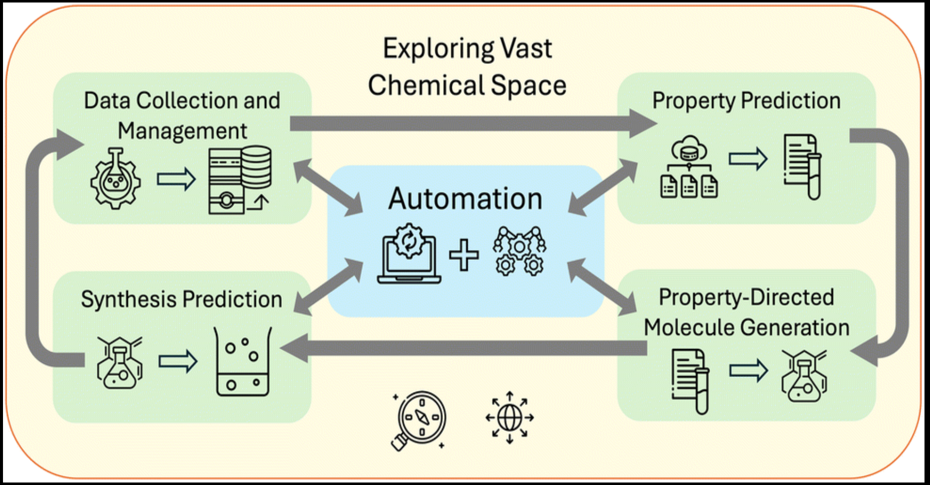
Interface with LLM/Agents

RAG & tool orchestration: LLMs use KG for retrieval/constraints;

[1] An update on knowledge graphs and their current and potential applications in drug discovery

LLM/Agent in Drug Discovery

[1]



For Agent:

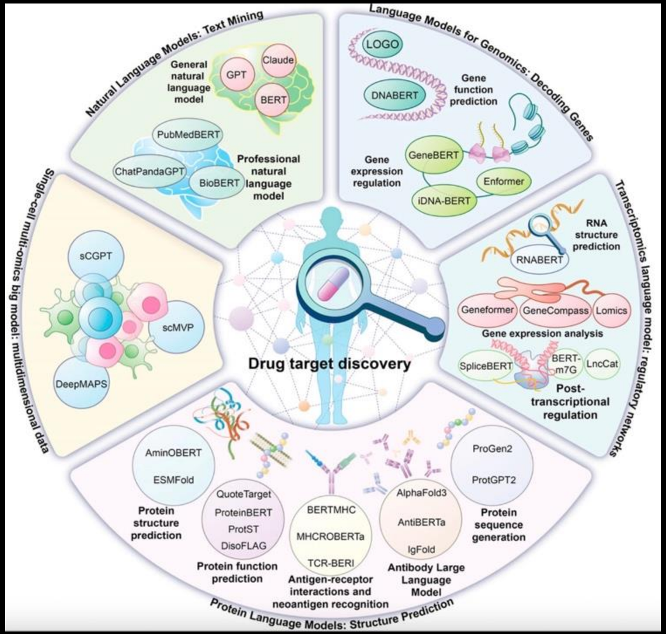
Core tasks

Tool orchestration: Coordinate molecule generation, docking, etc.
Autonomous execution: Drive DMTA and trial-emulation loops;

Core roles

Shrink search space & speed : Accelerate from design to validation.
Reproducibility & scalability: Standardized pipelines for robust reuse.

[2]



For LLM:

Core tasks

Knowledge anchoring: RAG linking KG/EHR to reduce hallucinations.
Protocol & documentation automation: Draft study protocols, variable dictionaries and auditable reports.

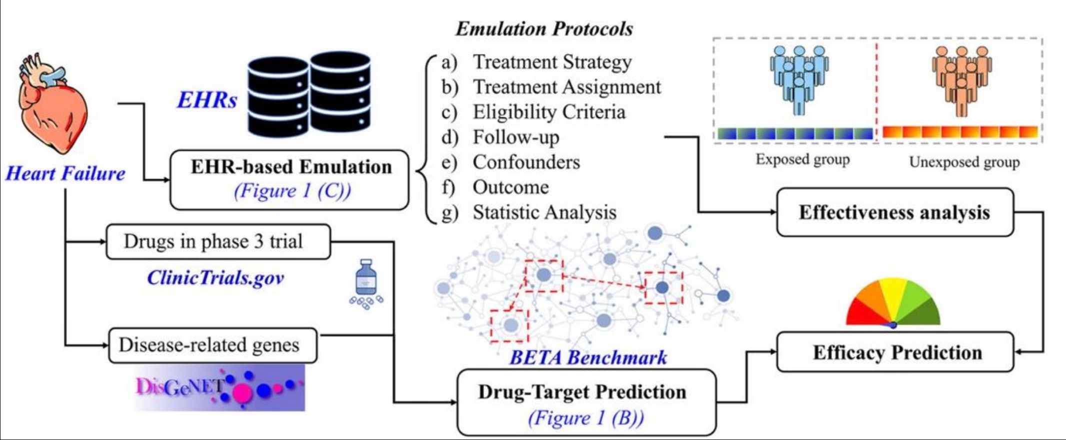
Core roles

Rapid hypothesis generation & explanation: Turn multi-source evidence into testable mechanisms and trial feasibility.
Efficiency & compliance: Cut prep time/costs and improve traceability.

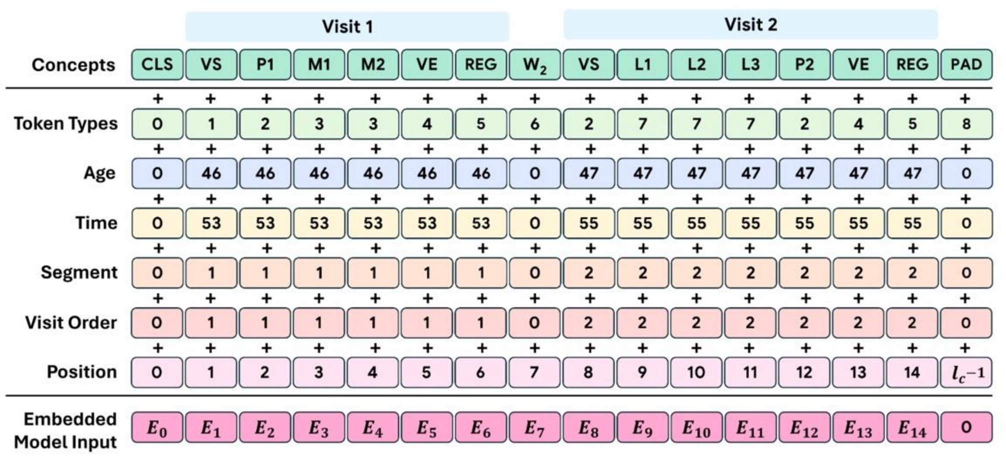
[1] Ramos M C, Collison C J, White A D. A review of large language models and autonomous agents in chemistry[J]. Chemical science, 2025.
[2] Liu X, Zhang J, Wang X, et al. Application of artificial intelligence large language models in drug target discovery[J]. Frontiers in Pharmacology, 2025, 1

Electronic Health Records (EHR) in Drug Discovery

[1]



[2]



Core tasks

Cohort construction & target trial emulation: eligibility, exposure, comparators.

Causal inference & bias control: propensity methods; sensitivity/negative controls.

Affect which R&D stages

Candidate prioritization & repurposing from real-world effectiveness signals.

Clinical design refinement: eligibility, endpoints and comparators.

Integrative view

LLM / Agents orchestrate the loop (Hypothesis → Modeling → RWD validation → KG update) to cut waste and clarify MoA.

Need to Combine:

-- Knowledge summarization & data integration (Knowledge Graph)

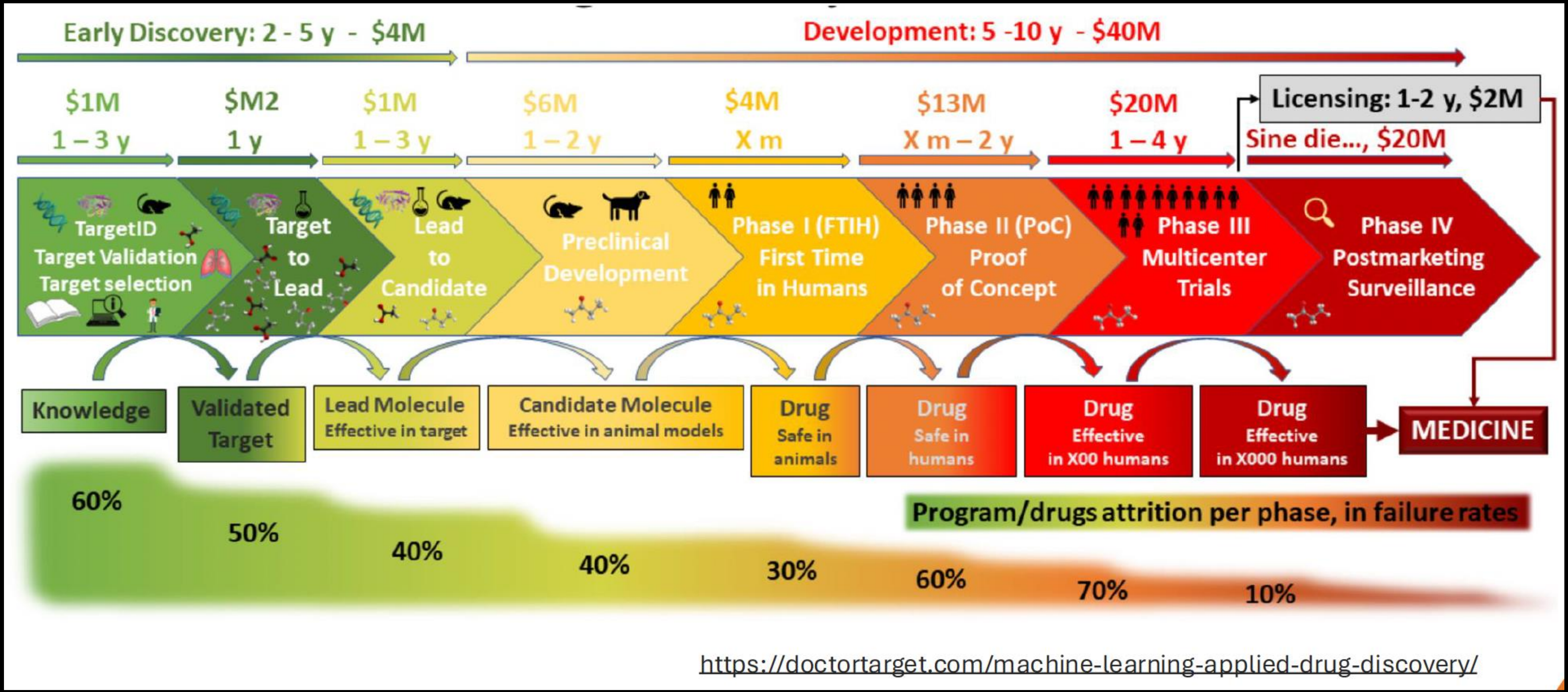
-- Real-world data feedback (EHR)

-- Reasoning & induction (LLM)

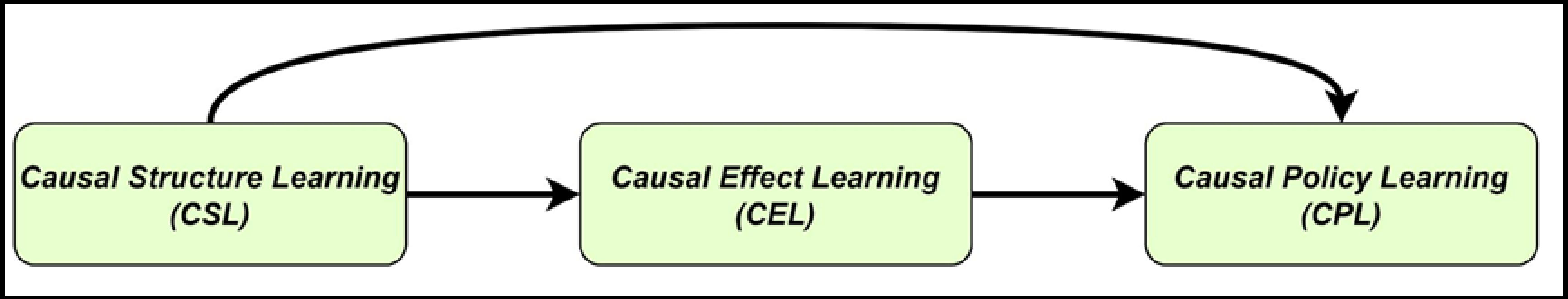
[1] Zong N, Chowdhury S, Zhou S, et al. Advancing efficacy prediction for electronic health records based emulated trials in repurposing heart failure therapies[J]. npj Digital Medicine, 2025, 8(1): 306.

[2] Fallahpour A, Alinoori M, Ye W, et al. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records[J]. arXiv preprint arXiv:2405.14567,

Electronic Health Records (EHR) in Drug Discovery



Causal Decision Making



CSL>>>CPL

- Identify Treatment Nodes:
- Guard Against Bias
- Eliminate Spurious Links
- Focus on Decision-Relevant Variables:
- Outcome: A reduced, validated causal subgraph that informs robust policy optimization for effective decision making.

CSL>>>CEL>>>CPL

- Estimate Intervention Effects
- Detect Unmeasured Confounding
- Adjust potential spillover
- Filter Prioritize Treatments
- Action Optimization via CEL

Grand Challenges

Integration of Heterogeneous Data

Combining genomics, imaging, EHR, and trial data into coherent causal frameworks.

Causal Discovery at Scale

Identifying robust causal structures from high-dimensional, noisy biomedical data.

Personalized Decision Making

Tailoring causal insights to patient-level diversity and comorbidities.

Translating AI to Drug Development

Bridging causal inference with practical drug design, validation, and regulation.

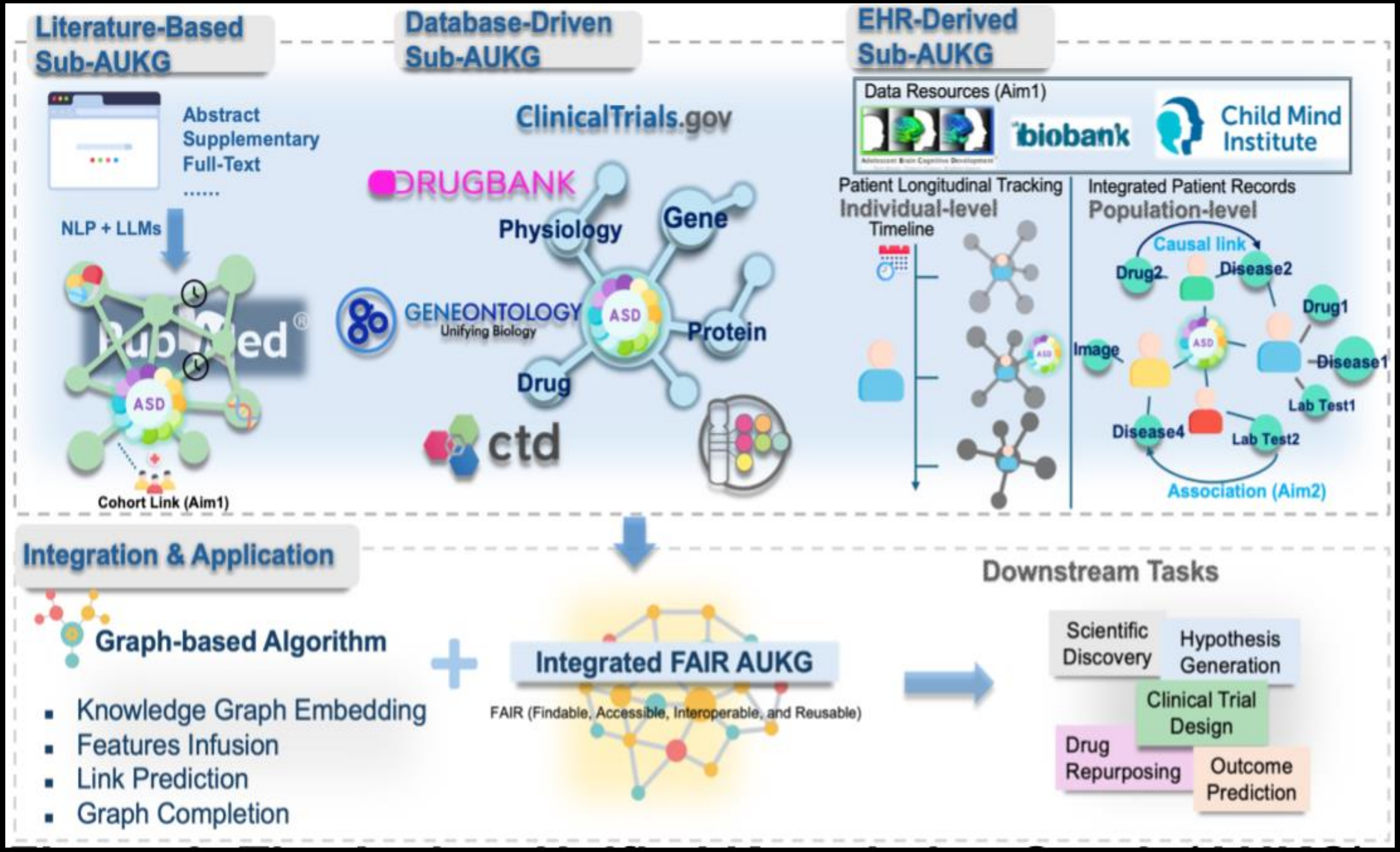
Part V

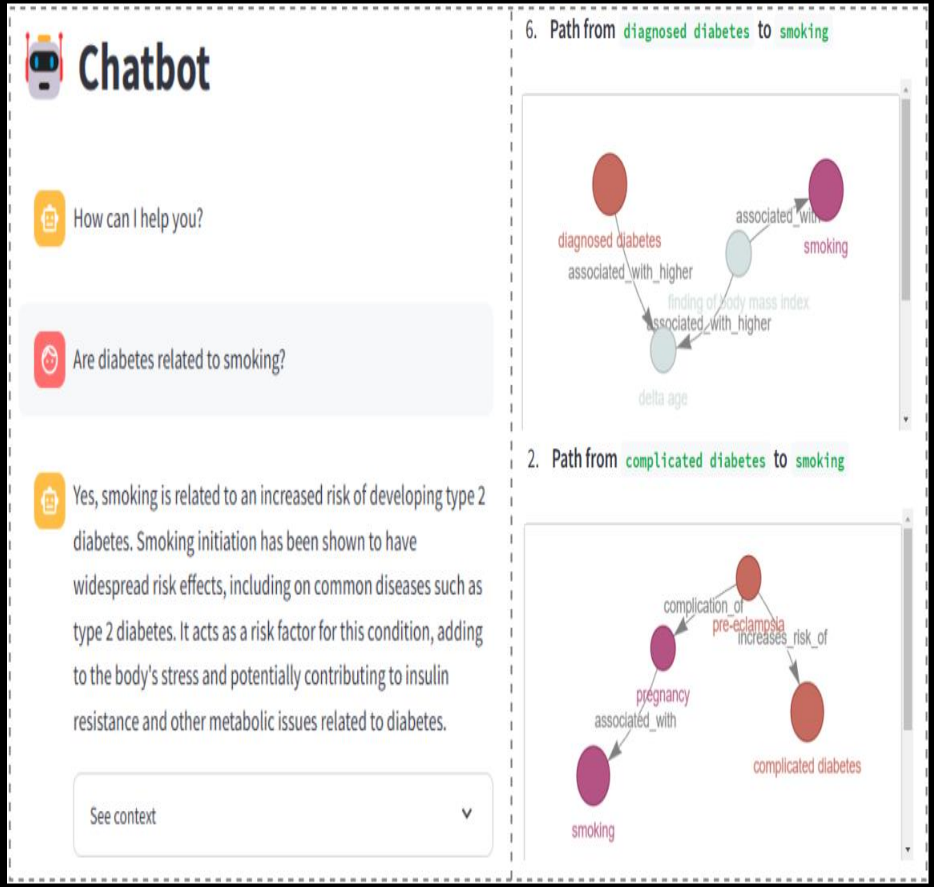
Large Language Models and AI Agent for Biomedical Data Analysis

"Language models capture words; agents must capture intentions. Without grounding, we are merely echoing text, not creating understanding."

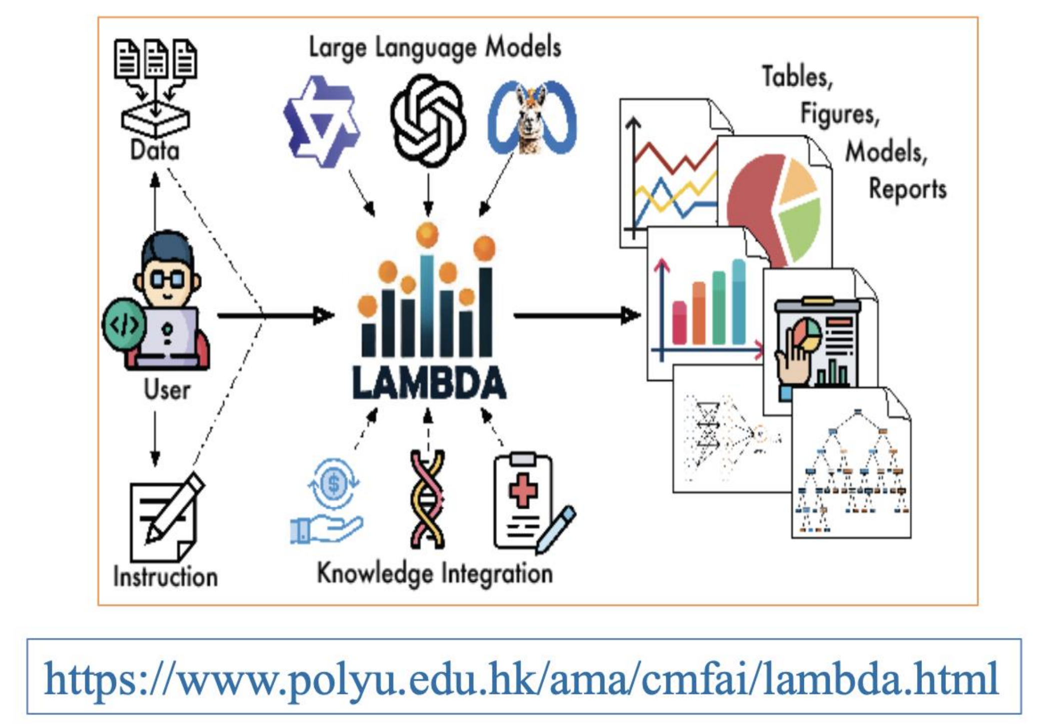
— **Yoshua Bengio**

KG Empowered LLM





LAMBDA: A Large Model Based Data Agent for Statistical Analysis



Core Capabilities

Human Language Interface

Expertise in Statistics and Data Science

Methods Recommendation

Automated Report Generation

Generative Visualization

Empowerment by Domain Large Models

....

Key Features

Reproducibility: Statistically consistent outputs given same prompts

Portability: Adapt to a variety of LLMs.

Scalability: Knowledge Integration of customized methods

Medical AI Agent

A **Medical AI Agent** generally refers to an AI system designed to act autonomously (or semi-autonomously) to support medical tasks by integrating data, reasoning, and decision-making

Core Characteristics

- ❖ **Data Processing and Feature Engineering** ⚙️
- ❖ **Data Integration** 🧩: Multimodal (EHR, imaging, genomics, wearables) and different resources.
- ❖ **Reasoning Ability** 🧠: Beyond pattern recognition, includes causal inference.
- ❖ **Autonomy & Adaptivity** 🤖: Co-pilot role for clinicians, context-aware.
- ❖ **Digital Twin** 📊 **:** Simulated Clinical Environments
- ❖ **Interoperability** 🔗: Seamless with hospital systems and workflows.

Grand Challenges

📊 Data Integration

- Heterogeneous sources
- High dimensionality
- Standardization

🧬 Causal Reasoning

- Beyond correlation
- Personalized medicine
- Uncertainty quantification

🔒 Trust & Safety

- Interpretability
- Robustness
- Regulatory approval

🤝 Human–AI Collaboration

- Augmenting clinicians
- Patient engagement
- Ethical concerns

Acknowledgement

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Brain Imaging Genetics Knowledge Portal (BIG-KP)

Genetics Discoveries in Human Brain by Big Data Integration

bigkp.org

Funding: U.S. NIH Grants R01AG082938, 1R01AG085581, 1R01MH136055, and R01AR082684.

Pictures: Copyrights belong to their own authors and/or holders.

Data: We thank Bingxin Zhao, Tengfei Li and other members of the UNC BIG-S2 lab (https://med.unc.edu/bigs2/) for processing the neuroimaging data.

UK Biobank resource application number: 22783.