Duke & Chen Institute Joint Boot Camp for AI & AI Accelerated Medical Research 2025 Course Note

Patricia Xiao

May 20, 2025

Abstract

This is the course note I took after attending Duke & Chen Institute Joint Boot Camp for AI & AI Accelerated Medical Research in 2025 May. These notes reflect my personal understanding and may emphasize topics based on my individual interests. They do not capture the entirety of what the professors discussed and should not be considered an official record. The event's information can be found at: https://ai-bootcamp.cs.duke.edu/. Some of the course materials are shared throgh Google Drive at: https://drive.google.com/drive/folders/16WRmyM-fgZKEmIChbdypctr0huYGOE_s. Many thanks to all my classmates and all the professors who gave a talk. It won't be such fruitful without your efforts. Special thanks to Prof. Jian Pei for hosting this event. Special thanks to Prof. Hongtu Zhu who inspired us a lot and participate in discussion very actively. Special thanks to Amy Peters who helped us from all perspectives. Special thanks to Jianfeng Zhu and Haoyu Zhou. They helped me catch up on the class, so even when I spaced out in the late afternoon for a few times, I didn't miss the key points.

1 May 12

1.1 Challenges & Opportunities in Translating AI to Healthcare

Prof. May Dongmei Wang from Georgia Tech gave the talk in the morning. Under the objective of providing pHealth, which stands for:

$$\text{pHealth} := \begin{cases} \text{predictive} \\ \text{personalized} \\ \text{participatory} \\ \text{precision} \end{cases} \text{Health}\,,$$

we need continued advancements in **Biomedical Engineering and Biotechnology**. In recent years, **Biomedical Big Data and AI** have emerged as accelerators, enablers, and central hubs driving innovation in the field. This becomes especially critical as healthcare challenges linked to increased longevity arise, most notably, the rising prevalence of chronic diseases in the aging population. There is an urgent need to manage healthcare costs more effectively, and AI has the potential to play a pivotal role in achieving this goal. Modern medical practice has evolved from **experience-based** approaches to **evidence-based** methodologies, and is now increasingly transitioning toward **intelligence-based** methods powered by AI [1, 2, 3, 4]. Meanwhile, AI techniques are evolving fast, as is listed in Table 1.

Table 1. History of Modern AI

Step	Year	Algorithm	Capability
1	< 1960	Naive Algorithms	I repeat
2	1960 – 2010	Machine Learning	I imitate
3	2010 – 2018	Deep Learning	I learn
4	2018 – 2022	Deep Reinforcement Learning	I learn to learn
5	> 2022	Distributed Agents, Swarm deep reinforcement learning	I contribute, I exchange

Prof. May Dongmei Wang also provided us with an interesting analogy:

- **The Brain**: Foundation models, Multimodal Generative AI, Deep Reinforcement Learning, Digital Twin, etc.
- **The Peripheral Nervous System**: AI-Embedded Devices, Edge Computing, etc.
- **The Spinal Cord**: Federated Learning, Swarm Learning, etc.

And predicted the future of AI in Health:

- AI in Health 1.0 (2010 – 2025)
  - Machine Learning & Deep Learning
  - Computational Intelligence (machine dominant)
- AI in Health 2.0 (2025 – 2030)
  - Generative AI & Digital Twins
  - Convolutional Intelligence (machine and human synergy)
- AI in Health 3.0 (2030 – ?)
  - Cognitive Architecture
  - Cognitive Intelligence (human dominant)

With expectation in paradigm shift:

- Sickcare → Healthcare
- Reactive → Proactive
- Reverse → Preventive

And to make $1 + 1 \gg 1$.

And there are already many examples of AI applications that changed medical practice, and their influence is across the entire human lifespan:

- Embryo selection for IVF
- Genome interpretation sick newborns
- Voice medical coach via a smart speaker
- K+
- Mental health
- Paramedic dx of heart attack, stroke
- Assist reading of scans, slides, lesions, etc.
- Prevent blindness
- Classify cancer, identify mutations
- Promote patient safety
- Predict death in-hospital

However, critical challenges remain, particularly in ensuring that AI systems in biomedical contexts are safe, trustworthy, actionable, and responsible. In particular, Prof. May Dongmei Wang listed a few:

- Data Integrity (Provenance, Standardization, Bias, Reuse) ⇒ "Garbage in, Garbage out"
- Data Integration ⇒ better use of scientific insights
- Population-Based Learning + Case-Based Reasoning (Causal Inference) ⇒ Personalized Care
  - Causal inference is very hard, and still a long distance away from being solved;
  - Most of the related works right now are working on counterfactual simulation or so.
- Adaptive Learning ⇒ Real-Time Care Decision Making ⇒ Adaptive AI Algorithm
- Metric and Validation ⇒ Health Informatics Solutions Interpretability ⇒ Societal Impact
- Generative Model ⇒ Foundation Model ⇒ New Paradigm, New Evaluation

Besides, people's attitude towards the use of AI is not all optimistic [5]. See Figure 1 and 2. Another point that clinicians complain about is that, there are too many AI tools and they don't know where to begin.

Yet there is still need of AI assistance. Back in the year 2013, where the officially-reported cause of death is: 1-st heart disease (611k), 2-nd cancer (585k), 3-rd COPD (149k), they argue that they have an estimation of Medical Error casused death at about 251k, which is the real No. 3 cause of death that year. However, CDC is not even counting this cause. There is a lot AI can do to help reduce unnecessary death.

1. Data Integrity

Data integrity problem is closely associated with the 4-th point (social impact). One example is what's mentioned in a tech review, titled *Hundreds of AI tools have been built to catch covid. None of them*

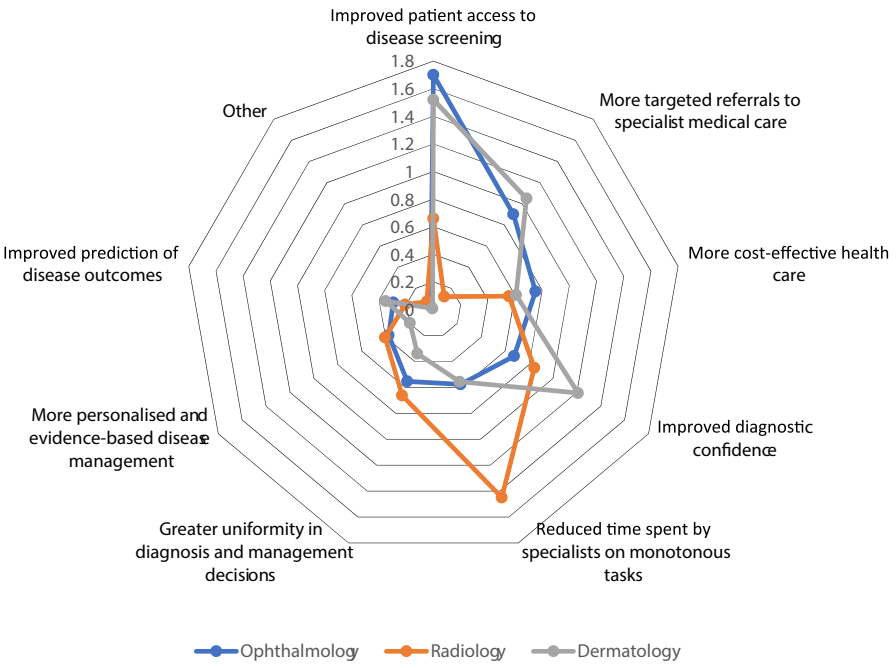

Figure 1. Radar plot showing the highest scoring responses for the greatest perceived advantages of the use of AI. Responses were selected from a list of set choices. Plot axes represent the average ranks for all respondents, with higher scores indicating a higher ranking/stronger preference. [5]

helped. [6]. Why is that happening? There are many reasons. First, most of the AI models we use right now are blackbox, the more powerful they are, the less explainable they are. Second, the data and results are not standardized well, causing a lot of overhead to use them. One way to solve them is to make data standardized, such as Fast Health Interoperability Resources (FHIR) [7].

In practice, data are collected from multiple resources, such as:

- Wearable Sensors
- ICU
- Medical Imaging
- EHRs
- Public Health Surveillance

while all of them share the same set of problems, such as:

- Missing Data
- Noise / Artifact
- Tabulation Errors
- Lack of Standardization

There are significant opportunities in data harmonization and quality improvement.

As for the necessity of data quality control, she also provided us with an example of how clinical diagnosis of WSIs can be affected by color batch effect. After doing proper normalization, results can be

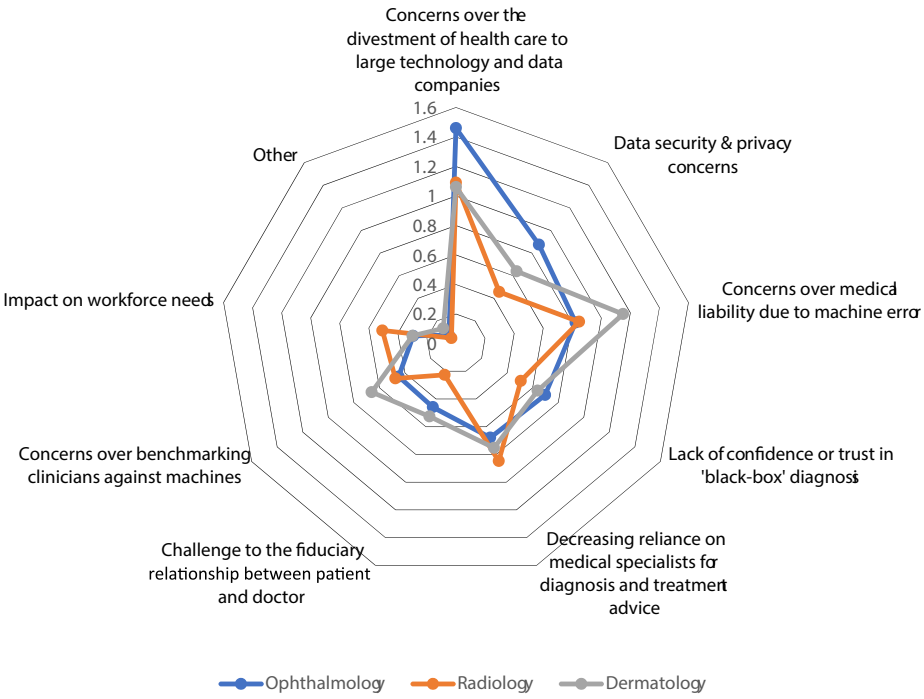

Figure 2. Radar plot showing the highest scoring responses for the perceived concerns or drawbacks of the use of AI. Responses were selected from a list of set choices. Plot axes represent the average ranks for all respondents, with higher scores indicating a higher ranking/stronger preference. [5]

much better.

2. Data Integration

Biomedical AI often faces multi-modality challenges [8, 9]. To enable holistic medical care (i.e., addressing the full complexity of a patient's condition rather than isolated symptoms) it is essential to integrate diverse data types, as shown in Figure 3. These include multi-omics data (e.g., genomics, transcriptomics, epigenomics, proteomics, and metabolomics), medical imaging (e.g., MRI, CT, PET, ultrasound), clinical records (both structured EHRs and unstructured physician notes), and so on. For instance, combining MRI and FDG-PET offers both structural and metabolic insights, aiding in Alzheimer's disease diagnosis. The key points of multi-modality learning are:

- to increase the value of data
  - every data modality has its strengths
  - some information may only be included in a specific modality
- to improve the integrity of data
  - make the model more robust
  - mitigate the impact of errors and inconsistencies

There are many challenges of multimodal machine learning, such as Representation (in a way that exploits the complementarity and redundancy of multiple modalities), Translation / Mapping (translate/map data from one modality to another), Alignment, Fusion, Co-Learning [10]. There are two principles to follow for multi-view learning (kind of used alternatively with multimodal learning in this talk):

- **Consensus Principle**: maximize the agreement on multiple distinct views
- **Complementary Principle**: each view may contain some knowledge that other views do not have

The multiple modalities are typically integrated at three different levels:

1. Raw data level
2. Feature level
3. Decision level

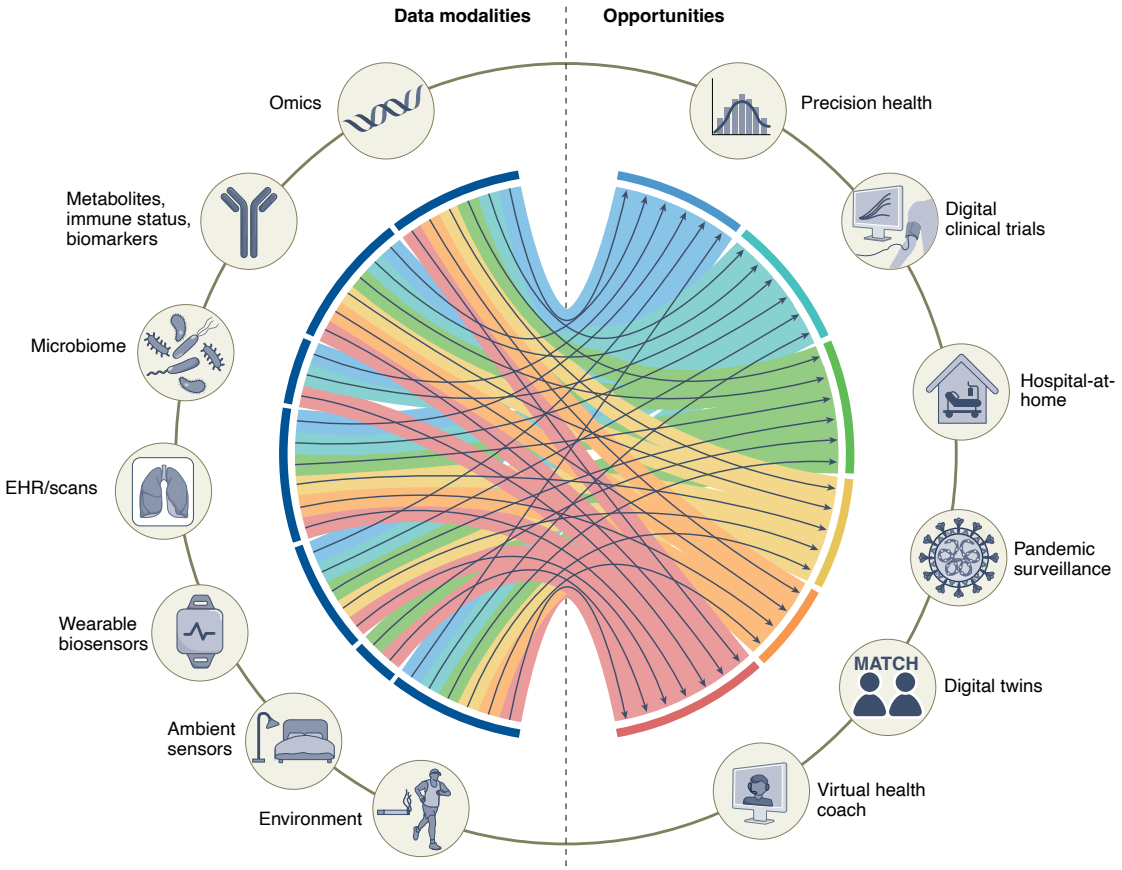

Figure 3. Data modalities and opportunities for multimodal biomedical AI. [9]

However, in many cases, instead of $1 + 1 > 2$, the situation can be even $1 + 1 < 1$, meaning that two modals working well on their own perform terribly worse when put together. So one should be very careful on what they do to integrate these data.

3. Adaptive Learning

The biggest challenge of real-time AI for public health and wellness lays in the responding time. Even 1 second will be way too slow in practice, and people want it even faster.

In theory, digital Twin [11] should be able to help solve the time delay, but it is not yet practical.

4. Metric and Validation

A common pitfall among computer science researchers is the "hammer looking for a nail" mindset (i.e., treating every problem as a technical task) and seeing ourselves, or being seen by others, merely as tool-builders. This narrow view often overlooks the bigger picture. To earn relevance and respect, we must engage deeply with real-world needs, translate them into AI solutions, and participate actively in problem formulation as equal intellectual contributors.

Validation requires interpretable (i.e., white-box, e.g., decision tree) and explainable (i.e., black-box, explain does not come with predictions, usually post-hoc, e.g., NN) AI. However, it is usually the case that the more accurate a model is, the less explainable the model will be, and therefore making them harder to be trusted. The lack of trust on clinicians prevents a wide adoption of AI models in real-world clinical trials, cuz we need to implement the following steps to make the AI models really helpful:

- Perform: the AI model needs to perform well;
- Trust: the clinicians needs to trust the AI results;
- Response: the clinicians needs to act based on the model results.

AI implementation science [12] has became an official academic discipline in recent years, aiming at putting AI to use in practice in the medical systems. They basically make sure to get through the loop of:

1. Technical Performance (Perform)
2. Usability Workflow (Trust)
3. Health Impact (Response, and then get back to improve Perform again)

Note that reprtition is important as conditions change.

All in all, we want the models to be actionable, that is able to:

1. provide the right information,
2. to the right people,
3. in the right format,
4. through the right channel,
5. at the right time of the workflow.

Note that translating biomedical problem to AI is far beyond dealing with the loop of "Problem Definition", "Information Extraction", "Knowledge Modeling", "Decision Making", it also involves people and policy. Policy affects what is allowed and what is not. As for people, Prof. May Dongmei Wang gave an interesting example, saying that her student once came across a patient with SSN 999-999-999 and this person is extremely sick. They investigated and found out that the nurse input 999-999-999 for every patient without a SSN to get the system proceed to the next step. Human actions can cause unexpected problems.

5. Foundation Model

With the foundation models' powerful generative ability, it is expected that by the year 2030, synthetic data will completely overshadow real data in AI models.[1] Synthetic data are not completely trustworthy (they have hallucination where they misunderstand everything or pretend to know things that they don't), but foundation models are useful tools without doubt, and are already outperforming many human doctors in identifying rare diseases.[2]

There are general-domain foundation models such as GPT, and also new foundation models for medical domain, but those medical foundation models usually have the following limitations:

[1] See Nvidia Blog for more: https://blogs.nvidia.com/blog/what-is-synthetic-data/
[2] https://www.today.com/health/mom-chatgpt-diagnosis-pain-rcna101843

- high demand for medical data and computational resources
- use data away from practice, such as medical QA from medical school entry exam
- less capable with white-box language models

In brief, according to some related works [13], the potential challenges are:

- Lack of domain knowledge
- High computational cost
- Privacy concern
- Hallucination
- Blackbox in logical reasoning

so the proposed solutions can be:

1. Retrieval augmented ChatGPT
2. Finetuning Llama + Chain-of-thoughts (CoT)

it seems that they prefer retrieval-augmented ChatGPT over CoT for it is less costly.

Impact driven LLMs can also help with communications between clinicians and patients, such as the idea of retrival-augmented generation (RAG) [14], where the LLMs refer authoritative knowledge source outside of training data source to generate responses.

Prof. May Dongmei Wang also introduced her work on agent AI, where LLMs server as autonomous agents. EHRAgent [15] are built to automatically generate code to load from EHR database according to human's inputs.

One should always be very careful on how foundation models can be misused. Misused AI can harm everyone. For example, there can potentially be a lot of biases (in data, in algorithm, in measurement, in objective, etc.) and other ethical concerns such as privacy issues. Remember that we need ethical and responsible AI (e.g., make sure it is fair, reliable, interpretable, clinical valid). Any problem exists in smaller models can be worsen when data or model scales up.

Conclusion

I have a few personal takeaways from this talk:

1. People won't trust your AI unless you provide sufficient evidence.
2. Pursue Need-Driven and Impact-Driven research pipelines.
3. Need to learn more about the application field, engage more with domain experts, and actively participate in the process of formulating meaningful research questions.

1.2 Medical Information Retrieval in the Era of Large Language Models (LLMs)

Prof. Wei Wang from University of California, Los Angeles (UCLA) introduced the recent advancements in medical large models.

She introduced foundation models, which are:

- large-scale machine learning models pre-trained on vast amount of data;
- can be adapted to perform a wide range of tasks.

For example, GPT-4, CLIP, SAM.

Among these types, LLMs are the main focus of today's talk. LLMs are built out of transformers:

$$\begin{cases} \text{Decoder-only: e.g. GPT} \\ \text{Encoder-only: e.g. BERT} \\ \text{Encoder-Decoder: e.g. T5, Flan-T5, Whisper} \end{cases}$$

As for foundation models, they can be roughly categorized into these classes:

1. Encoder LLMs with sequential data via masked language modeling;
2. (Encoder-) Decoder LLMs with sequential data via next token prediction (possibly with instruction tuning);
3. Mapping text and relevant sequential / graphs / images closer in the latent space via contrastive learning.

And she introduced CliBench [16], which aims at providing a benchmark of evaluating the large models' performances on biomedical tasks. She also mentioned additional findings comparing the strengths and limitations of LLMs, which are not included in CliBench and may not be formally published yet, but are potentially supported by evidence from other existing studies [17, 18], thus I list the most interesting findings (from my perspective) below:

- Domain-specific models are not outperforming general-domain models (e.g., Llama, GPT) on many tasks;
- In ablation study of the clinical data elements, they found that the importance (measured by amount of performance drop after removal) is: **medical record > patient profile > lab test results**, but the radiology report is kind of tricky because removing such fields sometimes do not result in performance drops (but neither did it significantly increased).
- Llama family has more hallucination than GPT family.
- There contain some certain bias, e.g., female and black are hard to predict, Medicare patients are somewhat easier to predict (perhaps because of that they are less diverse?)

Next, she introduced MERA [19], a model designed to predict the next diagnosis in a patient's visit by modeling the sequence of International Classification of Diseases (ICD) codes from past diagnoses. This transforms the task into a sequence-to-sequence prediction problem. MERA incorporates the following design components:

- Concept memorization: full names of diseases will be too long and exceed the LM's input length limit, but not providing the meaning of these concepts will harm the performance. In the end, they do fine-tuning for concept memorization, this is an effective knowledge injection approach.
- Defining loss on ranking: instead of forcing the model to predict an entire sequence of the output in order, it asks it to provide a ranking of the possible diseases and take the first few.
- Hierarchical contrastive learning: it applies an hierarchical contrastive learning approach on the ICD-9 ontology, to predict the disease groups from coarse-grain to fine-grain.
- Dynamic confidence threshold: they use a special token EOV to mark the end of visit, and train the model to generate this token at the proper time, so that no threshold needs to be set, the appearance of EOV itself indicates a low confidence.

The ablation study reveals that hierarchical contrastive learning contributes most significantly to performance, with the decoder design also playing a crucial role.

Derek is also exploring the potential of decoder-free generative models [20]. Where this decoder refers to the transformer decoder architecture.

STAR [21] is another work of Derek's, focusing on LLM for knowledge extraction and synthesis, taking the ontology prompt and unstructured documents as inputs to LLMs, do self-refinement by self-reflection (i.e., identify quality-issues automatically via self-reflection-question prompts), thus enhance low-resource information extraction performance (i.e., help learn rare cases better).

A few personal takeaways from this talk:

- There involves a lot of details in foundation model design.
- The idea of hierarchical contrastive learning is interesting.
- I really want to see the full report talking about the comparisons especially why Llama has more hallucination than GPT. Asked Derek and he said that's not part of CliBench findings. I'd have to wait and see if there comes other works.

2 May 13: Causal Generalist Medical Artificial Intelligence (CGM-AI)

Prof. Hongtu Zhu gave a talk that covers Section 2.1 to 2.5, Prof. Huaxiu Yao introduced Section 2.6, Prof. Qiao Liu introduced Section 2.7, and Prof. Xin Wang introduced Section 2.8.

2.1 What is Causal Generalist Medical AI?

This part begins with an introduction of the background, starting from data sets, Prof. Hongtu Zhu introduced a wide range of biomedical data sources, such as:

- Biobanks: large, deeply phenotypes cohorts (as for the meaning of deeply phenotype: [22]). e.g., UK Biobank [23].
- NIH-Funded Observational Cohorts: non-trivial studies with rich multi-modal data. e.g., ADNI [24], All of US [25], TopMed [26].
- Healthcare Data: e.g. Electronic Health Record
- Literature: peer-review articles. e.g., PubMed, bioRxiv, medRxiv.
- Ontologies: standard vocabularies for data harmonization. e.g., UMLS [27], SNOMED-CT [28], ICD-10 [29], MeSH [30].
- Clinical Trials
- etc.

And he listed the major biomedical data types, such as:

- Genetic & Omics: e.g., DNA/RNA, Epigenomics.
- Clinical & Administrative Records: e.g., EHR, claim, biling.
- Drug Information: e.g., prescription details, FAERS[3]
- Medical Imaging: e.g., CT, MRI, PET, X-ray, WSI, fMRI, Diffusion Tensor Imaging (DTI).
- Wearables & Remote Monitoring: e.g., Electrocardiogram (ECG), Electroencephalogram (EEG), Blood Pressure (BP), and Oxygen Saturation (SpO_2).
- Textual Data: e.g., PubMed, bioRxiv, medRxiv.

These datasets are introduced to emphasize the possibility and necessity of training large models on big data.

Foundation models are already introduced before in Section 1, where we learned that foundation models are:

- large-scale machine learning models pre-trained on vast amount of data;
- can be adapted to a wide range of downstream tasks.

Here, we learn that **Generalist Medical AI** models [31] are:

- foundation models pre-trained on large, **diverse** datasets;[4]
- can flexibly solve **new, unseen** medical tasks with **minimal or no** task-specific labels;
- via interpreting/reasoning across multiple data modalities.

And finally, Prof. Hongtu Zhu introduced Causal Generalist Medical AI (CGM-AI), which contains the following features:

- Data Integration: unified paradigm for integrating heterogeneous biomedical data;

[3] See: https://open.fda.gov/data/faers/

[4] It is mentioned in the GM-AI model that "general-purpose data sources can potentially be used to pretrain GMAI models".

- Fuses Causal: fuses causal inference;
- Generalist Pretraining: generalizes across wide range of tasks, usually by transfer-learning via finetuning;
- Multimodal Integration: can apply cross-modal attention, the key challenge is to handle missing modalities via auxiliary reconstruction or imputation.

Compared to GM-AI, which focuses on general pattern recognition, CGM-AI cares more about causal reasoning & valid intervention. GM-AI is just a foundation model that works on zero- to few-shot learning, CGM-AI is a foundation model combined with Supply Chain Management (SCM) / Directed Acyclic Graph (DAG) layers and causal inferences and handles counterfactual tasks (e.g., "what if"). As a result, GM-AI is confounding vulnerable, but CGM-AI is relatively robust.

2.2 Biomedical Knowledge Graphs

There are various different ways of constructing Knowledge Graphs (KGs), such as:

- Database-interacted KGs: SPOKE [32], PrimeKG [33]
- Data Mining-based KGs: BIOS [34], BioKG [35]
- LLM-generated KGs

Causal discovery can be mined from KGs, but needs to be careful enough, such as manually verify, and perhaps test in lab. To help people verify, there are tools such as BioKDE[5], Big-KP[6], BIG40[7] and Enigma[8].

Prof. Hongtu Zhu's group has some recent works on building Alzheimer's disease knowledge graph (ADKG) [36] and also mentioned that they are working on mental health knowledge graph as well.

2.3 EHR Foundation Models

Electronic Health Record (EHR) records longitudinal patient history, usually:

- containing multimodal data sources;
- containing causal insights;
- serves as foundation of CGM-AI, since it is the core modality for pretraining and downstream tasks;
- facilitates care coordination: interoperable through standards (e.g., FHIR, HL7) across providers.

Self-supervised pretraining on EHR data can very similar to a masked LLM, for example, we can model it as code-prediction problem (where code means the diagnosis code), and do masked code prediction, next-visit forecasting, temporal contrastive learning, integrate it with other modalities (such as imaging), and so on. Their works along this path include CATI [37] and UKB-MDRFM [38].

2.4 Meta-Adaptive Multimodal Integration

The term "meta-adaptive" means that we use meta-learning to "learn how to learn". In this way, a model can quickly adapt to new data modalities, tasks, or study cohorts with minimal amount of extra training. The "multimodal" refers to EHR, imaging, omics, text, KGs, and so on.

So there Prof. Hongtu Zhu proposed some key components:

- Association Learner: techniques for high-dimensional X-Y relationships (e.g., Deep CCA, Regression, contrastive learning, GNN)
- Imputation Engine: fusion-aware mechanism for missing data (e.g., autoencoders, cross-modal attention, MICE)

[5] See: https://biokde.insilicom.com/
[6] See: https://bigkp.org/
[7] See: https://open.win.ox.ac.uk/ukbiobank/big40/
[8] See: https://enigma.ini.usc.edu/

- Domain Adaptation: methods to align feature distributions across cohorts and devices (e.g., adversarial alignment, MMD, batch-norm tuning)
- Modality Encoders: specialized network for each modal (e.g., Transformer for EHR, CNN for images)
- Fusion Modules: cross-attention or joint encoder layers to combine modality embeddings
- Meta-Learner: MAML or metric-based framework to optimize fusion initialization and similarity metrics
- Adaptation Layers: Meta-learned parameters (e.g., FiLM, adapters) for fast task/domain fine-tuning
- Transfer Head: Task-specific output layers, finetuned with few-shot labels for clinical outcomes

A very big challenge is missing data imputation. Normally there are three types of handling this:

1. infer missing value from existing values;
2. use average (I think he means "mean") (I also believe that for categorical data we shall do majority)
3. use a binary flag to mark the potentially missing values to handle them differently

A quick comment I have on this part is that:

1. Sometimes the medical experts will complain about how it is not reasonable to assume "mean" for every missing value;
2. Missing value issue is not special to medical field, in fact, previously when I studied social network data, we also need to deal with missing data. This is pretty interesting, I think my way was a simple implementation of "inferring missing value from existing values" [39]. In short, I made the missing features trainable parameters.

The missing data challenge is especially severe in biomedical models, because it suffers from missing not at random (MNAR) problem, thus the systematic biases are introduced and cinfidence intervals can be unstable.

Longitudinal data imputation brings even more challenges, such as:

- irregular asynchronous sampling
- cross-subject heterogeneousity
- complex temporal data

Prof. Hongtu Zhu has introduced that their group also work on heterogeneous GNN that aims at solving these problems [40, 41].

2.5 Causal Decision Making

Causal decision making will greatly benefit clinical trial phases, enabling precision medicine, and so on.

Ideally there's a pipeline that contains Causal Structure Learning (CSL), Causal Effect Learning (CEL), and Causal Policy Learning (CPL). CSL affects CEL and CPL, CEL affacts CPL, while CPL sometimes also affects CSL and CEL.

Here Prof. Hongtu Zhu introduced another work of his group's, using spatio-temporal causal graph to work on causal deepset [42].

2.6 CGM Large Language Models

The talk started from introducing some general-domain Large Vision Language Models (LVLMs) and explain how they lack the essential clinical reasoning capabilities. Prof. Huaxiu Yao then introduced a few examples of LVLMs in medical field, such as LLaVA-Med [43]. Most of these models use ViT backbone.

Prof. Huaxiu Yao emphasized the misalignment problems in those LVLMs, proposing to have either better questions to prompt, or better model, such as using external knowledge to aid question-answering.

However, if the external knowledge is incorrect, there is no way to fix it, and the task will fail. Therefore, they propose to fine-tune for LVLM alignment. The preference fine-tuning phase is either supervised or reinforcement learning. This line of their work is concluded in MMedPo [44].

They also explored an iterative Retrieval-Augmented Generation (RAG) framework for medical question answering, building on existing works such as i-MedRAG [45]. While incorporating external knowledge proves helpful, it remains insufficient on its own. Inspired by multimodal retrieval methods like CLIP, MARVEL, and RULE, they propose a domain-aware multimodal retrieval mechanism. This line of research is encapsulated in their work MMed-RAG [46].

Prof. Huaxiu Yao also introduced the concept of a Medical AI Agent, which builds upon the idea of a multi-agent RAG system. Building on MedAgents [47], which enables role-based collaboration among expert models, MDAgents [48] further advances this framework by introducing adaptivity, allowing the connections between agents to dynamically form or dissolve based on context.

He also mentioned that the future should move from static QA to more interactive simulation, and introduced simulated clinical environments, such as Agent Hospital [49], and AgentClinic [50]. These simulated environment can serve as benchmarks of evaluating AI's capabilities. Other benchmarks include: GMAI-MMBench [51], OmniMedVQA [52], and CARES (evaluate trustworthiness) [53].

2.7 Omics Foundation Models

Prof. Qiao Liu introduced the brief history of AI models, also mentioned that GPT-4 is likely an MoE, according to people's estimations. Then he answered why we need foundation models (FMs) in omics.

In brief, the study of omics faces challenges from both statistic and biological perspectives. Statistically, it is:

- high-dimensional,
- high missing-rate (and high noise),
- spatial and temporal heterogeneous.

And it also faces biomedical challenges that systematically identify the genetic regulation mechanism is a context-specific manner.

Therefore, facing so many task-specific omics models, it comes the question, can we build a unified model? The answer is yes and here's why:

- The rapid development of foundation models in NLP provide technical supports;
- There is intrinsic similarity between biological sequences and natural language sequences (as long as they are tokenized).

Then, Prof. Qiao Liu introduced three lines of work:

- Genomic LLMs:
  - modeling genomic sequence (DNA data): based on DNABERT [54], they improved the design, such as improving the tokenization methods and positional embedding methods, and propose DNABERT-2 [55].
  - modeling single cell data (DNA sequences within a cell): e.g., scBERT [56], ScGPT [57], and so on.
- RNA LLMs: not well-developed, because of the flexible 3D structure of RNAs, and their unknown functions, but there are still related works such as RNA-FM [58].
- Protein LLMs: can be used for structure determining, protein-protein interaction, mutation effect prediction, de novo protein design, e.g., ESM, ProtBERT, ESMFold, AlphaFold.

2.8 Medical Image Foundation Models

Prof. Xin Wang introduced:

- medical image segmentation;
- medical image registration;
- AI for heart-disease analysis.

For segmentation, he emphasized the key role of U-Net, introduced famous variations such as Trans-UNet, Swin-UNet, U-Mamba, followed by their work UU-Mamba [59]. They also introduced the powerful tool Segment Anything (SAM) [60], and its variants in medical field, such as MedSAM [61].

For registration tasks, the goal is to align at least two types of data within a shared coordinate system. Existing methods such as VoxelMorph [62] address this problem. One proposed approach frames it as an image-to-image translation (I2IT) task, evaluating potential solutions including deep learning-based, deformable-based, and reinforcement learning-based techniques. This led to the development of Stochastic Planner-Actor-Critic (SPAC) [63]. Additionally, the speaker introduced uniGradICON [64], a foundation model for medical image registration.

Regarding AI for image-based heart disease analysis, Prof. Xin Wang presented the ACDC dataset and highlighted that coronary artery disease (CAD) diagnosis via traditional methods is costly. To address this, he discussed CT-FFR simulation and introduced their work on DEEPVESSEL FFR,[9] which has received regulatory recognition and has been awarded in multiple venues.

2.9 Conclusions

A few personal takeaways:

- I think there is still a long distance towards really solving causal inference problem, but I am convinced that it is worth trying;
- For EHR datasets, I shall consider involving KG, instead of focusing only on the NLP algorithms.
- Different modalities are good for different aspects of diagnosis. Perhaps instead of seeking for combining more modalities, I shall first make sure they contain the necessary information to include.

3 May 14

3.1 Harnessing Real World Evidence for Better Clinical Trials with AI

Prof. Fei Wang from Weill Cornell Medicine (https://wcm-wanglab.github.io) gave a talk in the morning, focusing on clinical trials-related topics.

Taking drug discovery (see Figure 4) as an example, there are a lot of CS research works done on the green (left) side, but the amount of works on the red (right) side is not compatible. This phenomenon indicates these problems' difficulty, yet it also infers a lot of potential opportunities.

As is shown in Figure 5, there are many steps to go through, each of them takes time, and they are all essential in practice.

Target-Trial Emulation (TTE) Framework

When it comes to clinical trials, we usually think of randomized controlled tryouts (RCTs), where a treatment group and a control group with approximately the same amount of subjects under similar conditions are used to verify the effect of an intervention.

However, although RCT is the gold standard for evidence generation in medical decision-making, there are many challenges, such as:

- Unethical: sometimes the way people collect evidence is not ethical;
- Costly: it takes 12 million dollars for conducting an RCT on average;
- Untimely: hard to react to new diseases due to the lack of patients to recruit, e.g., to select long COVID patients.

Real World Data (RWD) is defined as [65]:

[9] See: https://www.accessdata.fda.gov/cdrh_docs/pdf21/K213657.pdf

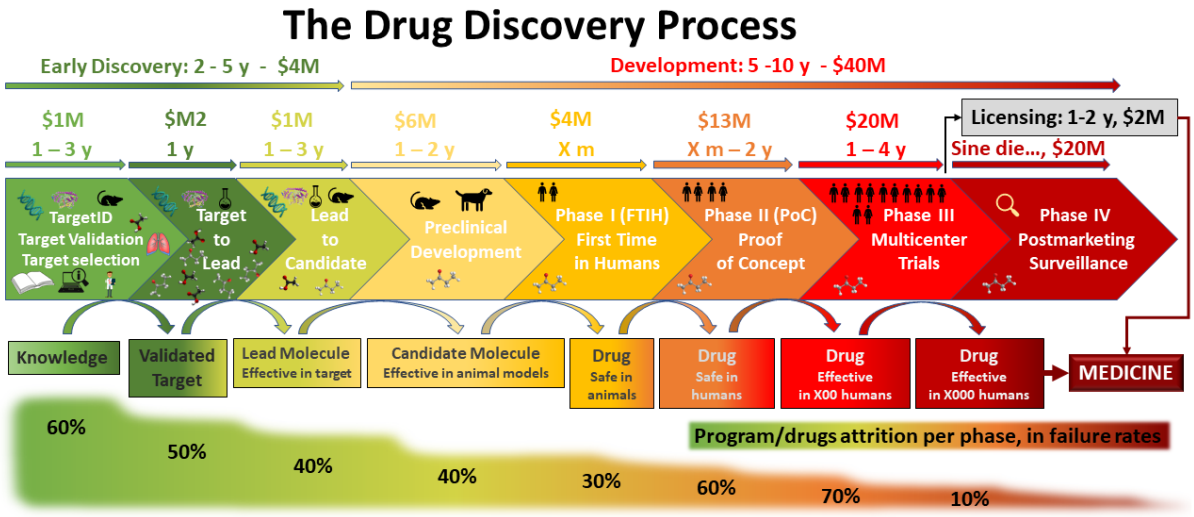

Figure 4. https://doctortarget.com/home/machine-learning-applied-drug-discovery/.

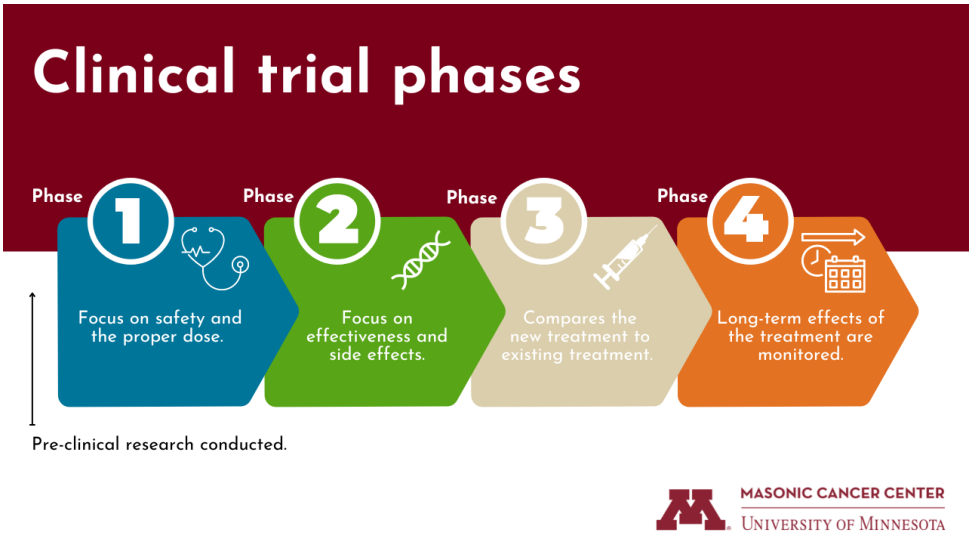

Figure 5. https://cancer.umn.edu/news/what-are-cancer-clinical-trials.

- data relating to patient health status and/or the delivery of health care;
- routinely collected from a variety of sources.

They can include family history, claims and bills, social media content, and so on. Real World Evidence (RWE) [66] is emerging together with RWD in recent years. They bring both opportunities and challenges, here we list a few:

- Opportunities: timely (quick response to new diseases) and long-term data, more ethical, bigger patient data, better generalizability, increase throughout (instead of case by case), rare outcomes available, InForm trials and more and more new applications, etc.

- Challenges:
  - Observational data's quality is questinable;
  - No randomized sampling, and the data contains all kinds of biases;
  - Problems like data missing, censoring, and the data is often longitudinal, complex, multi-modal, etc.

The future of evidence-based medicine is promising, while the currently-implemented techniques are almost just tip of an iceberg, leaving a lot more work to be done in the future [67].

Many existing works list out outline of a Target-Trial Protocol, listing [68]:

- The Protocol Component: e.g., treatment assignment
- Description: e.g., How will eligible persons be assigned to the interventions?
- Specification: e.g., Eligible persons will be randomly assigned to one strategy and will be aware of which strategy they were assigned to
- Emulation using Observational Cohorts: e.g., Eligible persons will be assigned to the strategies with which their data were compatible at the time of eligibility

This sort of justification will be done for each protocol component (e.g., eligibility criteria, treatment assignment, outcomes, follow-up, causal estimand ...). The goal is to use non-randomized observation data to serve as evidence, and it is very important to justify before you use any dataset. For instance, you can't use ICU data (such as MIMIC) to train chronic diseases' prediction model.

Confounding is a very big issue in RWD that we want to get rid of. For example, some subgroups (e.g., gender, race) are under-represented. In those cases, people usually do Inverse Probability of Treatment Weighting (IPTW). These strategies aim at balancing each subgroup under $T = 0$, and also do the same for $T = 1$ (where T here indicates treatment). Say that we have a group A and group B (e.g., male and female), then the easiest way of assigning such weight is:

$$w(T=1,A) = \frac{1}{\mathbb{P}(T=1|A)}, \qquad w(T=0,A) = \frac{1}{\mathbb{P}(T=0|A)},$$
$$w(T=1,B) = \frac{1}{\mathbb{P}(T=1|B)}, \qquad w(T=0,B) = \frac{1}{\mathbb{P}(T=0|B)}.$$

Of course there exists many other ways of assigning the weight, there are a lot of papers in recent years working on optimizing this weight, each with a lot of theoretical supports (we may find some in ICML proceedings). Despite all these efforts being made, in real world, there are still many problems with IPTW.

While research efforts initially focused on understanding and addressing COVID-19, there has been a growing shift in recent years toward studying long COVID, reflecting the evolving challenges in post-pandemic healthcare [69]. Prof. Fei Wang's team have worked on using data-driven (data: EHR from RECOVER [10]) approach to analyze long COVID [70]. In this work they provided formal definition of long-COVID (long-term consequences of COVID-19, conditions with significantly higher risk in COVID+ than in COVID- patients in the past-acute period), and explained why diagnosis code isn't reliable – clinicians are very careful in diagnosis, therefore, mild symptoms are often ignored. Without diagnosis code to rely on, nor well-established gold standard to study COVID-19, they have to mine from EHR data.

One of the findings I personally find interesting about their work is the effectiveness of Paxlovid in preventing/mitigating long-COVID. Their findings suggest that Paxlovid is not very helpful to low-risk patients any how (although it is theoretically not recommended to low-risk patients in the first place, some low-risk patients took the medicine and left record in the system). And also, it is also interesting to hear about (in the previous introduction part) vaccinated population has fewer rate of long COVID. Although I personally doubt if there is some biases in here: say, perhaps those who take vaccine shot are more serious in preventing infections and are more healthy in general?

[10] See: https://recovercovid.org/data

Their next work is also under the target-trial framework, but for a different purpose: drug repurposing hypothesis generation. Drug reporposing aims at revealing the other effects of drugs that were not the original intention of them. Alzheimer's disease (AD) is a big threat to the health of the aging population in recent years, and their team propose to mine the EHR data of Mild Cognitive Impairment (MCI) patients (from a large cohort, that potentially have taken various kinds of drugs) to find repurposed drugs [71].

They indeed have found some drugs that can potentially help mitigating AD. The interesting part of this work (to me) is that, Prof. Fei Wang compared the machine learning propensity score (ML-PS) methods used for effective inference of treatment effects by adjusting for confounding issues within the observational data, and at least for this particular task, he found that simple strategies, namely logistic regression (LR), outperforms the complex ones (e.g., GBM, MLP, LSTM). His insight was that, because the goal is to "simulate" the randomization which usually contains a significant overlap between the treated and controlled groups' data distributions, but the way-too-powerful models are separating the two groups away with way-too-clear decision boundaries, and therefore resulting in worse performances.

Federated Learning

Theoretically, federated learning is a good way of solving a lot of problems in today's medical AI. For example, can we have larger dataset? Can hospitals collaborate with each other efficiently and effectively? In practice, it is hard to obtain medical data, and there always come with a lot of privacy constraints. Federated learning enables agents to optimize locally and to transmute some intermediate parameters to a central server for joint learning. In this way, the sensitive data never left the hospitals.

In practice, it is usually hard to convince the hospitals to participate in federated learning, for many reasons:

- Privacy: Hospitals prefer to keep their data their own. Besides, there are always privacy concerns, cuz no one can say for sure that the intermediate layers won't obtain any sensitive information. For instance, the use of LLMs will not be very suitable in federated learning, cuz the encoder tells too many secrets in their embeddings, and it is totally possible to reconstruct some of theses.
- Benefit: In fact, small (and maybe poorer) hospitals typically benefit more from the federated learning design among hospitals, cuz it'll be really impossible to train a powerful model on their own data. But for the large and rich hospitals, they can afford training on their own and the performance is usually good. Introducing data of unknown quality from other sources means injecting noise to them in most cases, and the model performance can even drop.

Prof. Fei Wang's team proposed a federated learning model that introduces a collaborate network in the middle, to allow for hospitals working on their own, and don't have to share [72]. In this way, the performance shall not drop, at least in theory.

Federated learning is much harder than meta analysis (who share final results instead of intermediate parameters) to convince hospitals to participate in. There are some other very influential works related to federated learning, such as swarm Learning [73], who basically replace the central server as a blockchain. This method might make some participants feel more comfortable to join.

Diseases are Heterogeneous

Diseases can develop, and patients with similar diseases can be completely different with each other, and react differently to drugs.

With similar TTE framework, they analyzed the progression of SEPSIS, aiming at building a good predictor [74]. This time MIMIC data becomes super helpful (cuz this is not chronic disease).

There are existing works studying the progression heterogeneity of Parkinson's Disease (PD) [75]. Prof. Fei Wang's team propose to focus on the sutypes of PD [76]. This work involves a lot of data modalities, including EHR, imagings, multi-omics, and so on.

Individual Treatment Effect (ITE)

Patients will have different effect even when provided with the same treatment. There's a trend of shifting from TTE to Individual Treatment Effect (ITE) studies. And Prof. Fei Wang shared two works of theirs'.

The first one focus on Muscle Invasive Bladder Cancer (MIBC) data. For some patients, neoadju-

vant chemotherapy have very good outcomes and they probably won't need a surgery. For some other patients, it is not very effective. This work try predicting the outcomes of neoadjuvant chemotherapy more accurately for each individual, so that in case it is unlikely to work, consider surgery as soon as possible [77].

The second topic is lung cancer [78]. There are already some existing works on immune checkpoint inhibitor treatment for non-small cell lung cancer (NSCLC) [79]. Prof. Fei Wang's team uses a data-driven approach for analyzing NSCLC [80]. Note that they filtered the raw data (namely selected 4,666 patients out of 17,265 candidates) to create a much cleaner dataset to use.

Multi-Agent System for Clinical Trial

Prof. Fei Wang also introduced a multi-agent system (MAS) his team developed, which aims at facilitating and accelerating clinical trial design (CTD) via automated RWE extraction and refinement from EHRs [81]. This system contains:

- multiple autonomous AI agents with specialized roles;
- structured conversations and analysis that allows agents to collaborate through;
- powered by LLMs;
- autonomous, iterative refinement of trial protocols.

They work on MMIC-IV dataset, compared GPT-4o, Phi-4, DeepSeek-R1, Gemma 3, and the results show that GPT-4o performs the best on their tasks. Their evaluation task includes a structured report generation, and the results look reasonable to me.

Conclusions

A few personal takeaways:

- It seems that studying long COVID has now becoming a trending topic?
- Clinical trials always try to answer causal questions, although causal inference is hard.
- Don't think that you can work on this on your own. Always have domain experts involved.
- I can also consider multi-agent system for report generation.

3.2 Foundation Agents

Prof. Bang Liu from University of Montreal (UdeM) shared his slides with us: https://drive.google.com/file/d/1zmcuH6WD2IeBXKdqIjp69_vMAZs5VsiE/view?usp=sharing from his personal website https://www-labs.iro.umontreal.ca/~liubang/index.html. We follow the definition of agent in literature [82]:

> An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.

Based on the definition, an agent is usually defined as Figure 6.[11]

Human beings seek to build AI models that are even smarter than our selves. Although AI models face a lot of challenges, such as that it is hard to keep safety+trustworthy, while keeping capability+efficiency, these models are way too

Basically what Prof. Bang Liu introduced in today's talk are all included in his survey paper [83]. This is a very very long survey! The first time I ever seen a server exceeding 200 pages, that's almost a textbook.

A framework of powerful foundation agent is proposed, as is illustrated in Figure 7.

Besides, Prof. Bang Liu also introduced agent for medical and health, from two perspectives:

- LLM-based Medical Agents: e.g., MAGDA [84], SurgBox [85], MedAgents [47], Medchain [86], MDAgents [48]
- AI Hospital: e.g., Agent Hospital [49]

[11] I found the figure from https://www.ques10.com/p/47996/intelligent-agent-agents-and-environments-1/

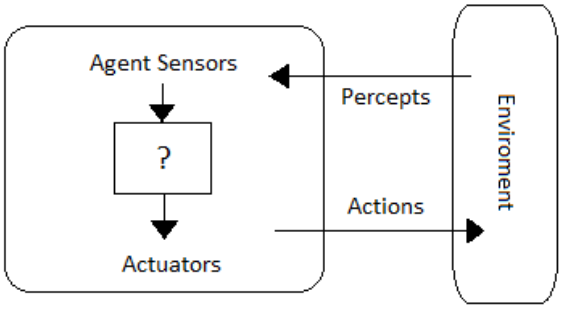

Figure 6. An illustration of basic architecture of an agent.

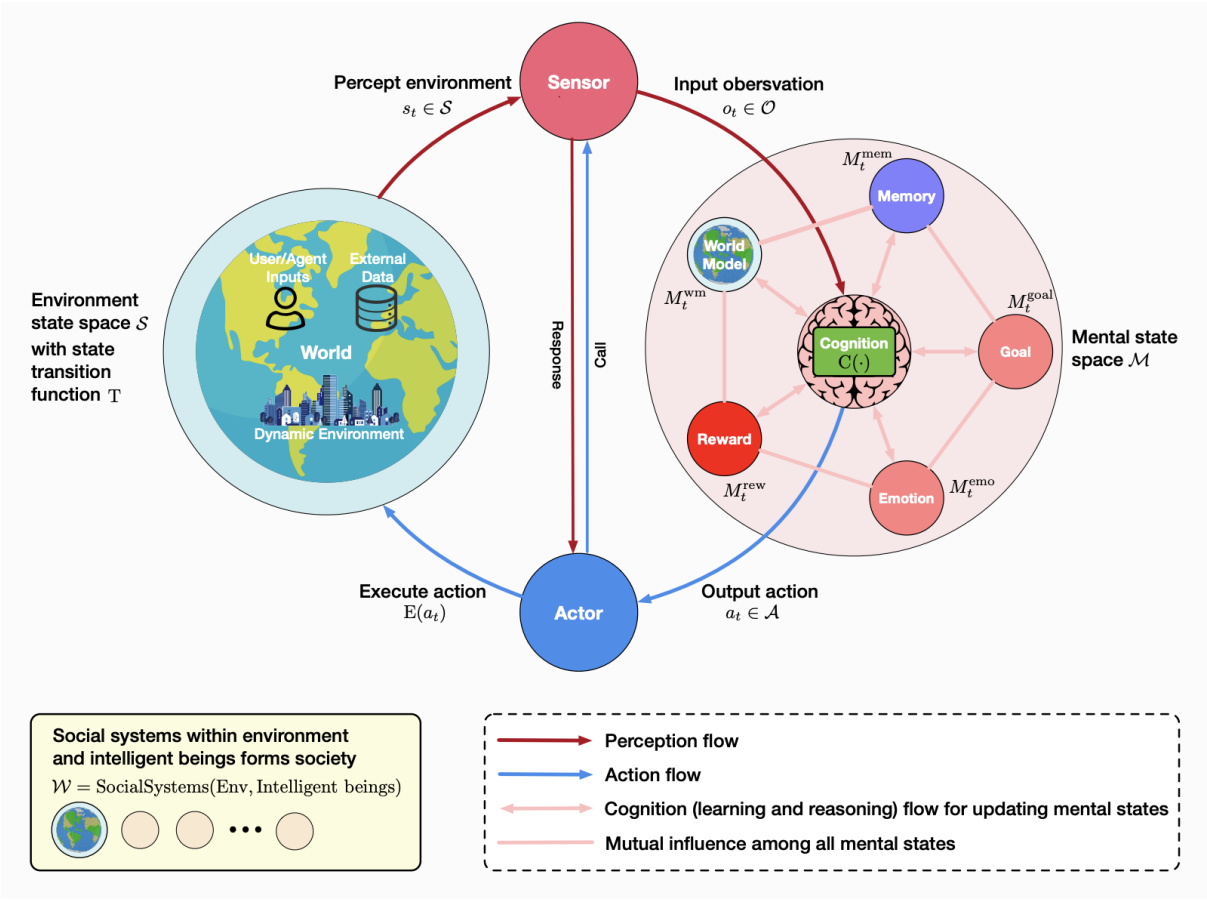

Figure 7. An overview of their general framework for describing an intelligent agent loop and agent society [83].

These are largely overlapped with Section 2.6, which was introduced on day 2. He mentioned a survey on LLM-based agents in medicine [87], and a survey on LLM-based Multi-Agent AI Hospital [88], and conclude that the AI agents are helpful in:

- Simulating Specific Scenarios
- Solving Complex Tasks
- Evaluating Agents & Synthesizing Data for Model Training

Conclusions

A few personal takeaways:

- It is very important to have a big picture of your works. I shall also start thinking about my own big pictures. The way Prof. Bang Liu organize things up is really inspiring.
- This is a pretty good design of AGI. No wonder why Prof. Jian Pei appreciated Prof. Bang Liu that much.
- The survey is a good one. I can not imagine how much efforts have been made into this. Worth reading carefully. Can be transformed into a course.

4 May 15

4.1 Scalable and Responsible Natural Language Processing to Transform Healthcare

Prof. Monica Agrawal from Duke University introduced her work on providing better support on analyzing patient data (centered around the use of EHR datasets).

She showed us that language is embedded across medicine, deeply involved at every step of the pipeline, such as:

- clinical notes,
- medical information,
- patient comments/instruction,
- clinical trials,

and that these notes are important to researchers, clinicians, and patients. For example, the Emergency Department (ED) Document typically includes: Triage Assessment, Chief Complaint, Nurse Notes, Doctor Notes, Discharge Summary. Among these fields, only Cheif Complaint (i.e., one-phrase summary describing reason of visit) is structured data, others are all free-text.

As for the medical free-text data, understanding even the most basic building blocks requires multiple hops of logic. For example:

> On carbo ia for TNBC. Will dc.

In this example, after reasoning back and forth under the other context information, an experienced human expert can conclude that carbo turns out to be Carboplatin, ia stands for Intra-Arterial, TNBC means Triple-Negative Breast Cancer, and dc means discontinue. It is not very easy to understand, cuz when treated separately:

- carbo can be carboplatin or carbodome, etc.
- ia can be Intra-Arterial or Intra-Articular, etc.
- dc can also have many other meanings.

These notes are messy, and can be horribly hard to interprete. Prof. Monica Agrawal provided us with a few extreme examples that Pete Szolovits[12] collected:

Primer on clinical foundation models

Prof. Monica Agrawal covers three sub-topics:

- basics of language models,
- clinical fine-tuning and data behind LLMs,
- foundation models of other modalities.

[12] https://groups.csail.mit.edu/medg/people/psz/home/Pete_MEDG_site/Home.html

Table 2. How Bad Can it be to Interpret

Note	Meaning
3/11/98 IPN	(date of) Intern Progress Note
SOB & DOE ↓	the patient's shortness of breath and dyspnea on exertion are decreased
VSS, AF	the patient's vital signs are stable and the patient is afebrile
CXR ⊕ LLL ASD no Δ	a recent chest xray shows a left lower lobe air space density that is unchanged from the previous radiograph
WBC 11K	a recent new white blood cell count is 11,000 cells per cubic milliliter
S/B Cx ⊕ GPC c/w PC, no GNR	the patient'd sputum and blood cultures are positive for gram positive cocci consistent with pneumococcus, no gram negative rods have grown
D/C Cef → PCN IV	so the plan is to discontinue the cefazolin and then begin penicillin treatment intravenously

When using structured (tabular) data to form a vector, usually even simple models such as a linear regression works well.

When modeling text information, it is natural to think of treating each token differently and embed them as one-hot vectors into N-dimensional space, where N is the vocabulary size so it is typically very high-dimensional.

Reducing their dimensionality, people seek to learn the words' embeddings (n-dimensional, where $n \ll N$). Then we need to use the text context to train work(/token) embeddings. Using co-occurrences as training signals, we learn dense-vector text embeddings from unlabeled text corpus, in an unsupervised, or self-supervised manner (e.g., Word2Vec, GloVE). The intuition is that similar words shall have similar context, such as:

> h.o of htn, taking metropolol to lower bp.
>
> ...
>
> Has high bp. One metropolol to treat hypertension.

Typically, there are three major types of training data:

- raw text (e.g., clinical notes),
- example task inputs and outputs (e.g., medical Q&A),
- human feedback on responses.

And they often stack on top of each other, and we use them together.

The biggest problem of having fixed vectors is that their meanings are not contextualized, never adapted according to different context. For example:

> "a history of **depression** and celexa administration"
>
> "**depression** of the right hemidiaphragm"
>
> "NSR with 1mm ST **depression**"

Then there comes the era of contextualized word embeddings, which we also refer to as language models (LM). Such as ElMo, who base on RNN architecture, and many other CNN approaches as well. Since the invention of Transformer [89], there comes a lot more larger language models based on the transformer architecture, which we call large language models.

Training of LMs (e.g., Med-PaLM [90]) typically involves three phases [91]:

1. Unsupervised Pre-Training: usually under the paradigm of either masked language model (MLM, blank-filling) or autoregressive (AR, next-token-prediction) training.
2. Instruction Tuning: usually involve multiple tasks, for example, train for medical Q&A. ("learn to do things correct")

3. Reinforcement Learning with Human Feedback: this step is proven to be very important in practice, typically letting human to rank the generated answer, telling which is better and which is worse, and then train the model with reward to encourage it to satisfy human-defined criteria on the samples with feedback. ("learn to do things well")

There is a growing trend of applying pretrained LLM on clinical text [92]. However, there also comes the question: since LLMs trained on general domain datasets have already achieved good performance in the clinical domain, and are hard to beat, are medical LLMs really needed [93, 94]?

The current conclusion, in general, is that, by rephrasing the prompts to a more suitable form, the medical LLMs performance is much better than we though they were. The key point lies in that the training corpus of a biomedical model is usually well-constructed and thus make the model less robust.

The other trouble is that, a dataset does not always test what you think [95]. This can also cause problem in the performance.

General-purpose models are performing quite well in clinical domain, but how? Where do they get the training from? Maybe from dataset leakage (online test sets), maybe from publicly available literature (e.g., PubMed), maybe something else.

All these problems motivate us for better understanding of our data:

- For general models: how do we understand the content of what is/isn't in the pre-training data?
- For clinical models: how can we augment their pre-training data?

Prof. introduced WIMBD [96] as a useful tool to analyze the training datasets for majority open-sourced LLMs. They did a work that follow these steps:

1. Test Clinical Knowledge on a LLM,
2. Discover the performance's correlation with the LLM's pretraining data,
3. Analyze source data.

In this way, we understand the training corpus better.

There are other modalities of clinical foundation models, such as vision-language models, like PLIP [97] who follows CLIP [98], and LLaVa-Med [43] who follows LLaVa [99]. As an interesting example, to solve the problem of lacking paired text to each image, LLaVa-Med [43] uses GPT-4o to prompt answers and have enough training corpus in Phase #2: instruction fine-tuning.

And there are also foundation models for EHR data [100]. They are regarded as a different type of FM, FEMRs. Namely, there are two broad categories of clinical FMs: Clinical language models (CLaMs) and Foundation models for EMRs (FEMRs). Most FEMRs are unimodal as they only consider structured codes (e.g., LOINC, SNOMED, etc.).

Accelerating clinical research

Prof. Monica Agrawal covers sub-topics:

- introduction to information extraction,
- boosting LLM performance,
- human-in-the-loop paradigm.

There are huge potential of EHR data, for example, it can fuel studies on:

- causal inference + reinforcement learning: what treatment will lead to the best outcome?
- disease progression modeling: what is the patient's expected disease trajectory?

Answering them typically requires knowing additional variables such as: comorbidities, side effects, treatment efficacy, drug status.

Extracting information from free-text can help organize knowledge better. For instance, a piece of free-text clinical note might be:

> "pt progressed after 5 mos of CarboTaxo for EC. Will dc and discuss pembro"

It can be translated into full expression such as:

> "Patient progressed after 5 months of carboplatin/pacilitaxel for endometrial cancer. Will discontinue and discuss pembrolizumab."

Then it can be further organized into a tabular form of medical record, as is shown in Table 3. Many of

Table 3. Result of deciphering clinical text.

Field	Before	After
Medication	Carboplatin + Paclitaxel	Pembrolizumab
Reason	Endometrial Cancer	Endometrial Cancer
Status	discontinued	starting (implicit)
Reason for Stop	progression	
Duration	past 5 months	

her past works focus on extracting intended information from clinical notes, as is shown in Table 4. The general pipeline is that, they start from clinical note, do partial chart review, train ML (NLP) model, form structured data, and use the structured data for medical research. There are several hurdles:

Table 4. Statys quo for information extraction.

Paper	Variable	# Training Data
[101]	Start/Stop Data for Oral Medications	6,000+
[102]	Binary Metastasis	17,000+
[103]	Binary Reason for Stopping Treatment	8,000+ and 1,500+

- Avoid Label Leakage: their solution is re-annotating the existing public dataset to create benchmark for fair judgment on few-shot performances [104].
- Evidence-Backed Output: seek to provide not only medications, but also reason, dosage, frequency, duration, as available.
- Deployability: the biggest concern will be: (1) compliance + unwieldy size of models, and (2) sensitivity to wording + model mis-calibration.

We can boost LLM performance by having better prompts. There are two promising settings [105]:

1. Single Prompt: LLM as weak supervision, (1) get LLM outputs on public data (2) identify confident outputs (typically the most homogeneous regions) (3) train smaller model on confident outputs (4) run smaller model on same or new datasets.
2. Multiple Prompts: Combining multiple prompts by co-training framework [106].

Many studies first require constructing a timeline of events, with many dates only found in unstructured text notes from EHRs. In these cases, human-in-the-loop framework can be a good solution. There, domain experts involve in the process of labeling and verifying the initially unlabeled timeline of notes.

Her conclusion indicates:

- A lot of research's bottleneck is that the huge amount of information lives only in clinical notes are unrevealed;
- LLMs can get us a huge amount of the way there, but naive use of LLMs can be suboptimal;
- Both ML and human-in-the-loop approaches can improve performance.

Streamlining point-of-care

Prof. Monica Agrawal covers sub-topics:

- smarter EHRs,
- Case Study: summarization,
- automation bias.

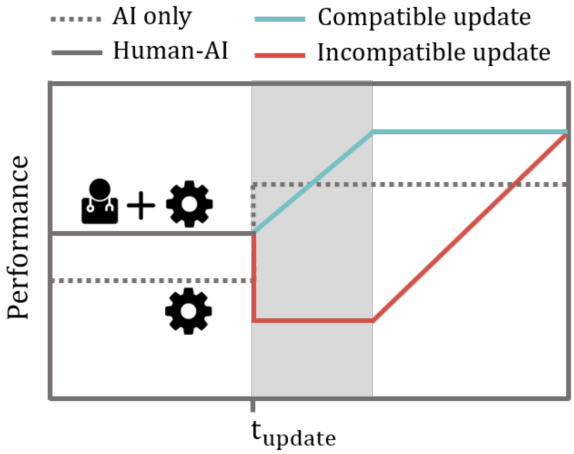

Figure 8. Schematized view of human-AI teams in the presence of AI updates. Human-AI teams perform better than either alone, but when the AI is updated its behavior may violate human expectations [111].

Clinical documentation has always been a challenge [107, 108]. The EHR data is messy and has some usability problems in them. Doctors hate their computer for it is time-consuming to either enter data or retrieve information.

Streamling data entry is a solution they proposed, in short, the contextual autocompletion can help save time [109].

In such a human-AI team, AI can also help with more efficient information retrieval. Their team's work is MedKnowts [110], which tracks the Audit logging and learns from the domain experts' (clinicians') read and write actions when they analyze cases. In their work, they learn condition-procedure relations.

Also note that in a human-AI team, collaboration don't always lead to better outcomes, and things might not be smooth, but there are ways to make it work [111] (see Figure 8).

As for text summarization, its necessity lies in the redundancy of clinical note and the time-saving need of clinicians. To automatically measure the quality of generated (summarization) text, there are two type of ways to evaluate [112]:

- intrinsic: measures quality directly, using similarity of text, or some fact extractor to compare the information included, valid when ground truth is available, e.g., **??**.
- extrinsic: measures utility, by Q&A or other methods, e.g., HARVEST [113]
- both: e.g., Patient Portal Drafts [114], Prof. MA's team's work [115].

The wider issue regarding autonomous bias is that, as we move towards incorporating AI into new clinical workflows, it is important to minimize unintended consequences.

For example, when clinicians are exhausted, which is usually the case in real life, they'll be more casual on the note they take and will let go of nonfatal errors, and do more copy & paste, leaving redundant, even conflicting content in the same note. While introducing AI can help reduce workload, AI does not really know how to correct mistakes, and can be easily biased.

Improving accessibility of health information

Prof. Monica Agrawal covers sub-topics:

- readability,
- online health information.

Improving accessibility to personal EHR is a promising direction to go. Meanwhile, without sufficient medical training backgroup, the clinical note need to be translated into more readable form to be

understood by patients. Although there hasn't been a perfect standard yet (right now: mostly using the simplicity of vocabulary usage), existing works (also Prof. MA team: [116]) have already started working on solving this problem. However, an interesting fact is that, although getting well-explained, some patients are not more satisfied with the latest version of notes. They feel that these notes introduce more concerns and make them more worried.

And that is the problem: online health information. Some websites are helpful for understanding medical terms,[13], while some others are not accurate, even using AI-generated content as ground truth. These information are actually dangerous, with selective answers and unclear (even fake) references, made-up supporting points, and so on. There includes so much misinformation and disinformation, resulting in an urgent call to more responsible AI models [117].

Conclusion

Personal takeaways:

- Prof. Monica Agrawal introduced the difficulties of using EHR text data properly, to which I strongly agree with, due to some previous experience dealing with EHR data.
- Directly using EHR data have made some of our previous projects stuck in the middle. Perhaps we shall move towards our goal step-by-step, cleaning EHR text data itself can be turned into several challenging projects.
- Our projects are actually using human-in-the-loop framework a lot. Unfortunately, domain experts' available time is way too limited, and that is our new bottleneck.
- It is so important to understand our data before feeding it into any model.
- EHR free-text note contains a lot of information that can not be found elsewhere.

4.2 Advanced Machine-Learning-Enabled Imaging-Omics Analysis

Prof. Heng Huang from University of Maryland gave a talk concluding his work in the past two decades, mostly focusing on introducing computer science basis to the students with strong medical backgrounds.

I tried filling his outline in with his top-cited works. As for the computational and mathematical background, I believe textbooks explain better than my note.

There is no direct 1-on-1 relation between his topics and his publications, he tried to summarize them all up. To dig in the details of his works, I would suggest visit his publications and read the papers directly: https://scholar.google.com/citations?user=4OqLaDwAAAAJ&hl=en.

Biomarker Identification

This line of work is basically doing feature engineering on biomarkers (i.e., the minimal SNPs that distinguishe cases from controlled group or any other phenotypes). Prof. Heng Huang introduced many tricks in implementation, such as using LASSO regularization term to encourage sparsity.

His works include: [118], etc.

Multi-Modal Genotypes and Phenotypes Integration

This line of work deals with heterogeneous data, with multi-modality or so.

These works include: [119, 120, 121], etc.

Longitudinal Biomedical Data Analysis

This line of work adds sequential modeling. Many medical datasets include timesteps, like EHR or video sequences. These scenarios are suitable to apply these models.

These works include: [122], etc.

Distributed and Federated Learning

This line of work focus on distributed computing. Prof. Heng Huang shared with us how hard it is to let hospitals share their data, and therefore, to train at a large scale, federated learning is one of the most

[13] e.g., https://www.pharmgkb.org/clinicalAnnotations

feasible options.

These works include: [123, 124], etc.

Conclusion

Personal takeaways:

- AI application in biomedical studies are not falling behind the frontier of AI, instead, it almost always uses the trending models and popular strategies.
- I got a quick review on the implementation details of some models, but I do think that many interesting parts are skipped (due to the time limit).
- Prof. Heng Huang is very experienced. For example, he said if your dimensionality is not at least at thousands level, the tabular data perform well with XGBoost, but not as good with transformers. This is indeed the case.

5 May 16

5.1 Foundation Models and Knowledge Graphs for Consolidating our Knowledge Regarding the Human Genome

Prof. Jie Liu from UMich gives a talk in the morning, basically about human genome modeling. His slides are shared at: https://drive.google.com/file/d/1Hj4rayx3NY-x-Sxy66M19O6kQ1Wa5k34/view.

Background of Genomic Modalities

Prof. Jie Liu started from introducing human gene expression.

As we all knows, DNA is the blueprint of life. The human gene program that was conducted at about 1970s – early 2000 triggered the develop of biomedical studies. And there are now many resources around like GTEx project.[14]

Something that we might not have been told is that among all existed human gene (approximately 3 Billion, he said), only about 5% are functional. Of these functional genes, only about 1%–2% coding region, and the remaining are non-coding region, that might be related to encoder, promoter, etc.

Genes are blueprint, but are not directly the answer. One genome can relate to hundreds of cell types, and identical genes can result in different gene expression.

The central dogma of biology is:

- Genes (DNA) transcribe to RNA;
- RNA translate to proteins;
- Proteins carry out biochemical functions (in the cell);
- Different type of cell expresses genes differently, thus have different proteins, and different functions.

As a matter of fact, gene-expression is complex and multi-step. Each step is carefully regulated. The stepwise gene expression might change over time. Whenever dysregulated gene expression happens, there usually comes disease, such as cancer. Waddington's landscape [125] is a metaphor of how a cell's developmental trajectory is guided by genes and environment. A cell is like a ball rolling down through specific pathways (chreods) to reach a state. Cells become different step-by-step ("differenciation"), and may be "reprogrammed" by manipulating gene expression. One interesting application of such theory is organoid [126]. FDA is now discouraging testings on human and animal organs, instead, encouraging the use of organoid.

By the way, if you accept that cancer is related to gene expression, then it is not hard for you to see how heterogeneous cancer can be (i.e., different patients have different symptoms).

In fact, some special proteins, called transcription factors (TFs), control DNA expression. TFs work together to activate or repress transcription. TFs are proteins encoded by genes, therefore, we can also say that genes form regulatory networks, where genes activate and repress each other's expression.

[14] https://www.gtexportal.org/home/

DNA is tightly coiled in the nucleus (with help from histone, a protein that provides structural support for a chromosome.), while different portions of it vary a lot in degree of compaction – tightly compressed parts cannot be expressed, cuz proteins can not access it to work on it. Therefore, the 3D genome, or 3D chromatin organization, matters a lot to gene expression.

High-throughput sequencing usually result in text file containing DNA sequences (i.e., "reads"), with each nucleotide assigned a quality/confidence score (meaning that not 100% sure they are correct). High-throughput sequencing techniques can rapidly and cheaply sequence billions of molecules, and determining the sequence of nucleotides in a DNA molecule is a fundamental experiment. Some studies on DNA molecule mostly care about the sequence of nucleotides. Interpreting these sequences is hard, even with help of tools like bowtie2 or bwa, it is still challenging to handle:

- large genome size,
- existence of read errors,
- alignment of reads and genome.

There are many other kinds of sequences, like RNA sequencing (RNA-seq), Bisulfite sequencing (BS-seq) measuring DNA methylation, and Assay of Transposase-Accessible Chromatin Sequencing (ATAC-seq) measuring DNA accessibility, and Chromatin immunoprecipitation sequencing (ChIP-seq) measures transcription factor binding.

Hi-C (Chromosome Conformation Capture followed by High-throughput Sequencing) is a technique using high-throughput sequencing to analyze the 3D genome structure. It's smallest units are bins, where the 3 Billion gene is divided into bins at about 10 K length.

Knowledge like the two types of loops, CTCF-cohesion and E-P loops, are introduced, saying that CTCF means larger loops that can be far away, while E-P loops are smaller, closer loops.

Computational Models to Impute 3D Chromatin Organization

They first introduced another technique called Micro-C. Compared to Hi-C who uses restriction enzymes to digest the DNA and thus results in larger DNA fragments, Micro-C employs MNase, which digests DNA in regions not stably bound by proteins, leading to smaller fragments (mono, di, or tri-nucleosome sized). They use CNN architecture to build their CAESAR [127] model, train it on Micro-C data, and achieve good performances on tasks such as loops and stripes detection. The model can also be used to impute high-resolution 3D chromatin contact maps.

Then, Prof. Jie Liu introduced Region Capture Micro-C (RCMC), which is much more accurate than Hi-C but it is not scalable. They propose to learn from paired Micro-C data and RCMC data, and to generalize (impute) RCMC to the entire sequence. Their solution is a model named Clepatra [128].

Foundation Model for Jointly Predicting Multiple Genomic Modalities

With all the different modalities of human genome data, they propose to use one model to predict all the modalities together. Compared with autoregressive models like GPT, they decided to adapt BERT-based architecture, for its bi-directional nature fits better to the gene expression setting. Their solution is called EPCOT [129].

In the next stage, they found multi-task setting suitable for handling even more genomic modalities. Instead of translating one modality to another, multi-task framework can translate to more. Among all LLMs, they found T5 useful [130], and they build a EPCOT v2 model on top of it. This paper is said to be out on BioRxiv (to be out soon?), but I haven't found it online.

Knowledge Graph for Human Genome

Knowledge graph captures relational data that traditional tabular datasets can not handle well. Besides, it makes it much easier to connect different resources of data together.

Their group's work on building and maintaining human genome knowledge graphs, namely GenomicKB [131] and GLKB [132], are still ongoing. These knowledge graphs, though not initially designed for assisting LLMs, turn out to be very helpful for LLMs in medical domain Q&As (GLKB already supports LLM access). He also mentioned another team's work: PanGraph.[15]

They are planning to include images and other data later on.

[15] https://github.com/neherlab/pangraph

Prof. Jie Liu mentioned that this is not an easy work to do. A lot of labor forces are needed just for sanity check. Meanwhile, the automated pipeline of streaming data won't always work.

Besides, as for conflict resolving, in case there's any contradictory inputs, the system will listen to the majority, and regard it as ground truth.

Conclusion

Personal takeaways:

- This talk is a very good example of how to organize a series of work under a consistent storyline.
- Will try to log all the points in case I work on geneme data in the future.

5.2 Machine Learning for Large-Scale, Multi-Modal, Biomedical Data

Prof. Sriram Sankararaman from UCLA gives a talk in the afternoon, centering around his team's recent explorations on applying AI models to solve biomedical problems.

His talk discussed their work on 3 modalities and then causal inference.

Genetics

This line of work build (approximated) mathematical models for gene expression. Their variance components models basically treat heritability and environmental variance component as random variable sampled from normal distribution, and the standard deviations of the distributions are to be estimated.

In this way, they define the problem as:

$$
\begin{aligned}
y &= \sum_{k=1}^{K} \mathbf{X}_k \beta_k + \varepsilon \\
\beta_k &\sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_k^2}{M_k} \mathbf{I}_{\mathbf{M_k}}\right), k \in \{1, 2, \ldots, K\} \\
\varepsilon &\sim \mathcal{N}\left(\mathbf{0}, \sigma_e^2 \mathbf{I}_N\right)
\end{aligned}
$$

Here, y is the phenotype, $\mathbf{X}$ is the genotype, β denotes the effective SNPs, and the goal is to estimate the variance components $\sigma_1, \ldots \sigma_K$ and σ_e.

An earlier and simpler version of it is:

$$
\mathbf{Y} = \mathbf{X}\beta + \varepsilon\,,
$$

where $\mathbf{Y} \in \mathbb{R}^N$ is the phenotype, $\mathbf{X} \in \mathbb{R}^{N \times M}$ is the genotype, $\beta \in \mathbb{R}^M$ is the effectiveness of SNPs, $\varepsilon \in \mathbb{R}^N$ represents the environmental variance component.

Their work on Randomized Heritability-Environment Regression (RHE) is already published in 2020 [133]. There is said to be a paper from Boyang Fu (first author) and Pazokitoroudi (second author) to be published soon in Nature Genetics 2025. But I haven't find a preprint yet. The later work, SUM-RHE, is said to require only summary data, which should be more easily permitted in practice.

When discussing the scaling of biomedical dataset and mentioning the different biobanks around the world (and how people can potentially use all these datasets), Prof. Hongtu Zhu commented that it is hard to obtain data from those biobanks due to a lot of regulations. Normally, only internal access is provided. There is still a distance to go towards diverse data collaboration in biomed. Right now, biobanks usually play on their own.

Tabular Phenotype Data

Data missing in biobanks are quite common, and the missing isn't at random.

- The easier ones are less likely to be missing;
- The missing is structured, e.g., if you do not fill in A you do not need to answer B, then the A&B values are often missed together.

They propose to impute the missing phenotype. A classmate asked why don't we just ignore the data column if it can be inferred from the other parts. Prof. Sriram Sankararaman said that it is

not really inferring, and showed us the results, indicating the imputed data can not really simulate the ground truth well (he used depression's estimation in mental health as an example). They proposed AutoComplete [134], which automatically impute the tabular data, while integrating some priors on what columns can be potentially "structured" and "correlated" (tend to be missing together), and do a copy-masking trick (which is hard to be impute in DL models directly) by masking off these columns together. AutoComplete improve the downstream tasks performances, and actually outperformed DL models of comparable sizes.

It is said that there is another work from Gorla et al. to be out soom in ICLM 2025. Haven't find a preprint yet. This new paper is said to further improve imputation by using transformers.

Another aspect of the finding shows that the models' performance (ranking) on biobank data is fundamentally different from that on machine learning benchmark datasets. One should be very careful on this difference.

Also, due to the different locations, a model that works perfect on one biobank can easily fail on other biobanks.

Imaging

This part focus on introducing the use of foundation models for medical volumes.

Simple feature extractor always fail to handle medical data. Powerful ones like DINOv2 performs well on 2D images. There hasn't been an established foundation model for medical volumes (3D data). Current methods do not scale well in volumns.

There is said to be a paper from Ulzee An (first author) and Jeong (second author) to be published soon in ICML 2025. But I haven't find a preprint yet. It is said that this model is called RAPTOR. This is a post-hoc approach in terms of foundation model training, meaning that they do not train anything (train-free). Instead, they propose a cleverer way of selecting the 2D slices of the 3D volumn (flexible choice of compression ratio), and build up a 3D representation both efficiently and effectively.

Causal Inference

They do causal inference from observational data [135].

Correlation isn't causation, because of the existence of confounders. Their idea is to condition on all the confounders, so as to break the dependency.

One key feature they introduce, is to simulate a patient's MR-Twin (Digital Twin) by predicting the potential genes (DNA-seq) of their siblings. Then they simulate the cases of these siblings, and use them as a cohort together with the original patient. In this way, they can break the dependency.

Conclusion

Personal takeaway:

- The mathematical formula of genetic study is pretty interesting.
- The design of MR Twin is cool, I would like to see the paper published and take a look.
- I am personally interested in checking the details of RAPTOR (I need both paper and code, so I shall wait for the published version).

References

[1] National Research Council, Division on Earth, Life Studies, Board on Life Sciences, and Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. 2011.

[2] Francis S Collins and Harold Varmus. A New Initiative on Precision Medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[3] Euan A Ashley. Towards Precision Medicine. *Nature Reviews Genetics*, 17(9):507–522, 2016.

[4] Michael R Kosorok and Eric B Laber. Precision Medicine. *Annual review of statistics and its application*, 6(1):263–286, 2019.

[5] Jane Scheetz, Philip Rothschild, Myra McGuinness, Xavier Hadoux, H Peter Soyer, Monika Janda, James JJ Condon, Luke Oakden-Rayner, Lyle J Palmer, Stuart Keel, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports*, 11(1):5193, 2021.

[6] Will Douglas Heaven. Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*, 6, 2021.

[7] Muhammad Ayaz, Muhammad F Pasha, Mohammed Y Alzahrani, Rahmat Budiarto, and Deris Stiawan. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR medical informatics*, 9(7):e21929, 2021.

[8] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[9] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal Biomedical AI. *Nature medicine*, 28(9):1773–1784, 2022.

[10] Tadas Baltruvsaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[11] Yishan Zhong, Benoit Marteau, Andrew Hornback, Yuanda Zhu, Wenqi Shi, Felipe Giuste, Joseph J Krzak, Adam Graf, Ross Chafetz, and May D Wang. IDTVR: A Novel Cloud Framework for an Interactive Digital Twin in Virtual Reality. In *2022 IEEE 2nd International Conference on Intelligent Reality (ICIR)*, pages 21–26. IEEE, 2022.

[12] Amir Kamel Rahimi, Oliver Pienaar, Moji Ghadimi, Oliver J Canfell, Jason D Pole, Sally Shrapnel, Anton H van der Vegt, and Clair Sullivan. Implementing AI in Hospitals to Achieve a Learning Health System: Systematic Review of Current Enablers and Barriers. *Journal of medical Internet research*, 26:e49655, 2024.

[13] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Large Language Models for Healthcare Data Augmentation: An Example on Patient-Trial Matching. In *American Medical Informatics Association (AMIA) Annual Symposium*, 2023.

[14] Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry Iwinski, Michael Wattenbarger, and May Dongmei Wang. Retrieval-Augmented Large Language Models for Adolescent Idiopathic Scoliosis Patients in Shared Decision-Making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10, 2023.

[15] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D Wang. EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records. *arXiv preprint arXiv:2401.07128*, 2024.

[16] Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. CliBench: A Multifaceted and Multigranular Evaluation of Large Language Models for Clinical Decision Making. *CoRR*, 2024.

[17] Yuhong Sun, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao. Benchmarking Hallucination in Large Language Models based on Unanswerable Math Word Problem. *arXiv preprint arXiv:2403.03558*, 2024.

[18] David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. Benchmarking Llama2, Mistral, Gemma and GPT for Factuality, Toxicity, Bias and Propensity for Hallucinations. *arXiv preprint arXiv:2404.09785*, 2024.

[19] Mingyu Derek Ma, Xiaoxuan Wang, Yijia Xiao, Anthony Cuturrufo, Vijay S Nori, Eran Halperin, and Wei Wang. Memorize and Rank: Elevating Large Language Models for Clinical Diagnosis Prediction. *arXiv preprint arXiv:2501.17326*, 2025.

[20] Mingyu Derek Ma, Yanna Ding, Zijie Huang, Jianxi Gao, Yizhou Sun, and Wei Wang. Inferring from Logits: Exploring Best Practices for Decoding-Free Generative Candidate Selection. *arXiv preprint arXiv:2501.17338*, 2025.

[21] Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P Jeffrey Brantingham, Nanyun Peng, and Wei Wang. STAR: Boosting Low-Resource Information Extraction by Structure-to-Text Data Generation with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18751–18759, 2024.

[22] Cathryn M Delude. Deep phenotyping: the details of disease. *Nature*, 527(7576):S14–S15, 2015.

[23] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[24] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[25] All of Us Research Program Investigators. The "All of Us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[26] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, 2021.

[27] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[28] Kevin Donnelly et al. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279, 2006.

[29] JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. ICD-10: History and Context. *American Journal of Neuroradiology*, 37(4):596–599, 2016.

[30] Carolyn E Lipscomb. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[31] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[32] John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, et al. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080, 2023.

[33] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[34] Sheng Yu, Zheng Yuan, Jun Xia, Shengxuan Luo, Huaiyuan Ying, Sihang Zeng, Jingyi Ren, Hongyi Yuan, Zhengyun Zhao, Yucong Lin, et al. BIOS: An Algorithmically Generated Biomedical Knowledge Graph. *arXiv preprint arXiv:2203.09975*, 2022.

[35] Brian Walsh, Sameh K Mohamed, and Vít Nová*v*cek. BioKG: A Knowledge Graph for Relational Learning On Biological Data. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3173–3180, 2020.

[36] Yue Yang, Kaixian Yu, Shan Gao, Sheng Yu, Di Xiong, Chuanyang Qin, Huiyuan Chen, Jiarui Tang, Niansheng Tang, and Hongtu Zhu. Alzheimer's disease knowledge graph enhances knowledge discovery and disease prediction. *Computers in Biology and Medicine*, 192:110285, 2025.

[37] Yue Shen, Jie Wang, Zhe Wang, Zhihao Shi, Hanzhu Chen, Zheng Wang, Yukang Jiang, Xiaopu Wang, Chuandong Cheng, Xueqin Wang, et al. CATI: A medical context-enhanced framework for diagnosis code assignment in the UK Biobank study. *Artificial Intelligence in Medicine*, page 103136, 2025.

[38] Yukang Jiang, Bingxin Zhao, Xiaopu Wang, Borui Tang, Huiyang Peng, Zidan Luo, Yue Shen, Zheng Wang, Zhiwen Jiang, Jie Wang, et al. UKB-MDRMF: a multi-disease risk and multimorbidity framework based on UK biobank data. *Nature Communications*, 16(1):3767, 2025.

[39] Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2258–2268, 2020.

[40] Zhaoyang Zhang, Hongtu Zhu, Ziqi Chen, Yingjie Zhang, and Hai Shu. Enhancing Missing Data Imputation through Combined Bipartite Graph and Complete Directed Graph. *arXiv preprint arXiv:2411.04907*, 2024.

[41] Zhaoyang Zhang, Ziqi Chen, Qiao Liu, Jinhan Xie, and Hongtu Zhu. Sampling-guided Heterogeneous Graph Neural Network with Temporal Smoothing for Scalable Longitudinal Data Imputation. *arXiv preprint arXiv:2411.04899*, 2024.

[42] Runpeng Dai, Jianing Wang, Fan Zhou, Shikai Luo, Zhiwei Qin, Chengchun Shi, and Hongtu Zhu. Causal deepsets for off-policy evaluation under spatial or spatio-temporal interferences. *arXiv preprint arXiv:2407.17910*, 2024.

[43] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[44] Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. MMedPO: Aligning Medical Vision-Language Models with Clinical-Aware Multimodal Preference Optimization. *arXiv preprint arXiv:2412.06141*, 2024.

[45] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific, 2024.

[46] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMED-RAG: VERSATILE MULTIMODAL RAG SYSTEM FOR MEDICAL VISION LANGUAGE MODELS. *arXiv preprint arXiv:2410.13085*, 2024.

[47] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

[48] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An Adaptive Collaboration of LLMs in Medical Decision Making. 2024.

[49] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957*, 2024.

[50] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

[51] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[52] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMed-VQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[53] Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. CARES: A Comprehensive Benchmark of Trustworthiness in Medical Vision Language Models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

[54] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[55] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv preprint arXiv:2306.15006*, 2023.

[56] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

[57] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024.

[58] Jiayang Chen, Zhihang Hu, Siqi Sun, Qingxiong Tan, Yixuan Wang, Qinze Yu, Licheng Zong, Liang Hong, Jin Xiao, Tao Shen, et al. Interpretable RNA foundation model from unannotated data for highly accurate RNA structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

[59] Ting Yu Tsai, Li Lin, Shu Hu, Ming-Ching Chang, Hongtu Zhu, and Xin Wang. UU-Mamba: uncertainty-aware u-mamba for cardiac image segmentation. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 267–273. IEEE, 2024.

[60] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[61] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment Anything in Medical Images. *Nature Communications*, 15(1):654, 2024.

[62] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[63] Ziwei Luo, Jing Hu, Xin Wang, Shu Hu, Bin Kong, Youbing Yin, Qi Song, Xi Wu, and Siwei Lyu. Stochastic Planner-Actor-Critic for Unsupervised Deformable Image Registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1917–1925, 2022.

[64] Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, Sylvain Bouix, Richard Rushmore, and Marc Niethammer. uniGradICON: A Foundation Model for Medical Image Registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 749–760. Springer, 2024.

[65] Fang Liu and Demosthenes Panagiotakos. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Medical Research Methodology*, 22(1):287, 2022.

[66] John Concato and Jacqueline Corrigan-Curay. Real-World Evidence—Where Are We Now? *New England Journal of Medicine*, 386(18):1680–1682, 2022.

[67] Vivek Subbiah. The next generation of evidence-based medicine. *Nature medicine*, 29(1):49–58, 2023.

[68] Miguel A Hernán. Methods of Public Health Research – Strengthening Causal Inference from Observational Data. *New England Journal of Medicine*, 385(15):1345–1348, 2021.

[69] Ani Nalbandian, Kartik Sehgal, Aakriti Gupta, Mahesh V Madhavan, Claire McGroder, Jacob S Stevens, Joshua R Cook, Anna S Nordvig, Daniel Shalev, Tejasav S Sehrawat, et al. Post-acute COVID-19 syndrome. *Nature medicine*, 27(4):601–615, 2021.

[70] Chengxi Zang, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozyuk, Edward J Schenck, Dhruv Khullar, Anna S Nordvig, Elizabeth A Shenkman, Russell L Rothman, et al. Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative. *Nature Communications*, 14(1):1948, 2023.

[71] Chengxi Zang, Hao Zhang, Jie Xu, Hansi Zhang, Sajjad Fouladvand, Shreyas Havaldar, Feixiong Cheng, Kun Chen, Yong Chen, Benjamin S Glicksberg, et al. High-throughput target trial emulation for Alzheimer's disease drug repurposing with real-world data. *Nature communications*, 14(1):8180, 2023.

[72] Haoyang Li, Chengxi Zang, Zhenxing Xu, Weishen Pan, Suraj Rajendran, Yong Chen, and Fei Wang. Federated Target Trial Emulation using Distributed Observational Data for Treatment Effect Estimation. *medRxiv*, pages 2025–05, 2025.

[73] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021.

[74] Suraj Rajendran, Zhenxing Xu, Weishen Pan, Chengxi Zang, Ilias Siempos, Lisa Torres, Jie Xu, Jiang Bian, Edward J Schenck, and Fei Wang. Corticosteroids for infectious critical illness: A multicenter target trial emulation stratified by predicted organ dysfunction trajectory. *medRxiv*, 2024.

[75] Sophie L Farrow, Antony A Cooper, and Justin M O'Sullivan. Redefining the hypotheses driving Parkinson's diseases research. *NPJ Parkinson's disease*, 8(1):45, 2022.

[76] Chang Su, Yu Hou, Jielin Xu, Zhenxing Xu, Manqi Zhou, Alison Ke, Haoyang Li, Jie Xu, Matthew Brendel, Jacqueline RMA Maasch, et al. Identification of Parkinson's disease PACE subtypes and repurposing treatments through integrative analyses of multimodal data. *npj Digital Medicine*, 7(1):184, 2024.

[77] Zilong Bai, Mohamed Osman, Matthew Brendel, Catherine M Tangen, Thomas W Flaig, Ian M Thompson, Melissa Plets, M Scott Lucia, Dan Theodorescu, Daniel Gustafson, et al. Predicting response to neoadjuvant chemotherapy in muscle-invasive bladder cancer via interpretable multimodal deep learning. *npj Digital Medicine*, 8(1):174, 2025.

[78] Pierre Saintigny and Jan A Burger. Recent advances in non-small cell lung cancer biology and clinical management. *Discovery medicine*, 13(71):287–297, 2012.

[79] Zhangfeng Huang, Wenhao Su, Tong Lu, Yuanyong Wang, Yanting Dong, Yi Qin, Dahai Liu, Lili Sun, and Wenjie Jiao. First-Line Immune-Checkpoint Inhibitors in Non-Small Cell Lung Cancer: Current Landscape and Future Progress. *Frontiers in Pharmacology*, 11:578091, 2020.

[80] Weishen Pan, Deep Hathi, Zhenxing Xu, Qiannan Zhang, Ying Li, and Fei Wang. Identification of predictive subphenotypes for clinical outcomes using real world data and machine learning. *Nature Communications*, 16(1):1–14, 2025.

[81] Haoyang Li, Weishen Pan, Suraj Rajendran, Chengxi Zang, and Fei Wang. TrialGenie: Empowering Clinical Trial Design with Agentic Intelligence and Real World Data. *medRxiv*, pages 2025–04, 2025.

[82] Stuart J Russell and Peter Norvig. *Artificial Intelligence: a modern approach*. pearson, 2016.

[83] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems. *arXiv preprint arXiv:2504.01990*, 2025.

[84] David Bani-Harouni, Nassir Navab, and Matthias Keicher. MAGDA: Multi-Agent Guideline-Driven diagnostic Assistance. In *International workshop on foundation models for general medical AI*, pages 163–172. Springer, 2024.

[85] Jinlin Wu, Xusheng Liang, Xuexue Bai, and Zhen Chen. SurgBox: Agent-Driven Operating Room Sandbox with Surgery Copilot. In *2024 IEEE International Conference on Big Data (BigData)*, pages 2041–2048. IEEE, 2024.

[86] Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. Medchain: Bridging the Gap Between LLM Agents and Clinical Practice through Interactive Sequential Benchmarking. *arXiv preprint arXiv:2412.01605*, 2024.

[87] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*, 2025.

[88] Zonghai Yao and Hong Yu. A Survey on LLM-based Multi-Agent AI Hospital. 2025.

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in neural information processing systems*, 30, 2017.

[90] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards Generalist Biomedical AI. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[91] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375*, 2023.

[92] Daniel P Jeong, Saurabh Garg, Zachary C Lipton, and Michael Oberst. Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? *arXiv preprint arXiv:2411.04118*, 2024.

[93] Daniel P Jeong, Pranav Mani, Saurabh Garg, Zachary C Lipton, and Michael Oberst. The Limited Impact of Medical Adaptation of Large Language and Vision-Language Models. *arXiv preprint arXiv:2411.08870*, 2024.

[94] Yahan Li, Keith Harrigian, Ayah Zirikly, and Mark Dredze. Are Clinical T5 Models Better for Clinical Text? *arXiv preprint arXiv:2412.05845*, 2024.

[95] Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical Large Language Model Benchmarks Should Prioritize Construct Validity. *arXiv preprint arXiv:2503.10694*, 2025.

[96] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's In My Big Data? *arXiv preprint arXiv:2310.20707*, 2023.

[97] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature medicine*, 29(9):2307–2316, 2023.

[98] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[99] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[100] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135, 2023.

[101] Monica Agrawal, Griffin Adams, Nathan Nussbaum, and Benjamin Birnbaum. TIFTI: A Framework for Extracting Drug Intervals from Longitudinal Clinic Notes. *arXiv preprint arXiv:1811.12793*, 2018.

[102] Benjamin Birnbaum, Nathan Nussbaum, Katharina Seidl-Rathkopf, Monica Agrawal, Melissa Estevez, Evan Estola, Joshua Haimson, Lucy He, Peter Larson, and Paul Richardson. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. *arXiv preprint arXiv:2001.09765*, 2020.

[103] Matthew S Alkaitis, Monica N Agrawal, Gregory J Riely, Pedram Razavi, and David Sontag. Automated NLP extraction of clinical rationale for treatment discontinuation in breast cancer. *JCO Clinical Cancer Informatics*, 5:550–560, 2021.

[104] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large Language Models are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

[105] Hunter Lang, Monica N Agrawal, Yoon Kim, and David Sontag. Co-training Improves Prompt-based Learning for Large Language Models. In *International Conference on Machine Learning*, pages 11985–12003. PMLR, 2022.

[106] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. COLT: Proceedings of the Workshop on Computational Learning Theory, 1998.

[107] Barbara B Anderson and Naomi Sager. Grammatical Compression in Notes and Records: Analysis and Computation. *American Journal of Computational Linguistics*, pages 68–81, 1975.

[108] Steven J Davidson, Frank L Zwemer Jr, Larry A Nathanson, Kenneth N Sable, and Abu NGA Khan. Where's the beef? The promise and the reality of clinical documentation. *Academic Emergency Medicine*, 11(11):1127–1134, 2004.

[109] Divya Gopinath, Monica Agrawal, Luke Murray, Steven Horng, David Karger, and David Sontag. Fast, Structured Clinical Documentation via Contextual Autocomplete. In *Machine Learning for Healthcare Conference*, pages 842–870. PMLR, 2020.

[110] Luke Murray, Divya Gopinath, Monica Agrawal, Steven Horng, David Sontag, and David R Karger. MedKnowts: Unified Documentation and Information Retrieval for Electronic Health Records. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 1169–1183, 2021.

[111] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2429–2437, 2019.

[112] Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *arXiv preprint arXiv:2104.13498*, 2021.

[113] Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. HARVEST, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2015.

[114] Siru Liu, Allison B McCoy, Aileen P Wright, Babatunde Carew, Julian Z Genkins, Sean S Huang, Josh F Peterson, Bryan Steitz, and Adam Wright. Leveraging large language models for generating responses to patient messages–a subjective analysis. *Journal of the American Medical Informatics Association*, 31(6):1367–1379, 2024.

[115] Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models. *arXiv preprint arXiv:2402.15422*, 2024.

[116] Niklas Mannhardt, Elizabeth Bondi-Kelly, Barbara Lam, Hussein Mozannar, Chloe O'Connell, Mercy Asiedu, Alejandro Buendia, Tatiana Urman, Irbaz B Riaz, Catherine E Ricciardi, et al. Impact of Large Language Model Assistance on Patients Reading Clinical Notes: A Mixed-Methods Study. *arXiv preprint arXiv:2401.09637*, 2024.

[117] Lionel Wong, Ayman Ali, Raymond Xiong, Shannon Zeijang Shen, Yoon Kim, and Monica Agrawal. Retrieval-augmented systems can be dangerous medical communicators. *arXiv preprint arXiv:2502.14898*, 2025.

[118] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and Robust Feature Selection via Joint ℓ2,1-Norms Minimization. *Advances in neural information processing systems*, 23, 2010.

[119] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 977–986, 2014.

[120] Feiping Nie, Xiaoqian Wang, Michael Jordan, and Heng Huang. The Constrained Laplacian Rank Algorithm for Graph-Based Clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[121] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*, volume 13, pages 2598–2604, 2013.

[122] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019.

[123] Bin Gu, An Xu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Privacy-preserving asynchronous vertical federated learning algorithms for multiparty collaborative learning. *IEEE transactions on neural networks and learning systems*, 33(11):6103–6115, 2021.

[124] Xidong Wu, Feihu Huang, Zhengmian Hu, and Heng Huang. Faster Adaptive Federated Learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10379–10387, 2023.

[125] James E Ferrell. Bistability, bifurcations, and Waddington's epigenetic landscape. *Current biology*, 22(11):R458–R466, 2012.

[126] Giuliana Rossi, Andrea Manfrin, and Matthias P Lutolf. Progress and potential in organoid research. *Nature Reviews Genetics*, 19(11):671–687, 2018.

[127] Fan Feng, Yuan Yao, Xue Qing David Wang, Xiaotian Zhang, and Jie Liu. Connecting high-resolution 3D chromatin organization with epigenomics. *Nature communications*, 13(1):2054, 2022.

[128] Clarice KY Hong, Fan Feng, Varshini Ramanathan, Jie Liu, and Anders S Hansen. Genome structure mapping with high-resolution 3D genomics and deep learning. *bioRxiv*, pages 2025–05, 2025.

[129] Zhenhao Zhang, Fan Feng, Yiyang Qiu, and Jie Liu. A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. *Nucleic Acids Research*, 51(12):5931–5947, 2023.

[130] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[131] Fan Feng, Feitong Tang, Yijia Gao, Dongyu Zhu, Tianjun Li, Shuyuan Yang, Yuan Yao, Yuanhao Huang, and Jie Liu. GenomicKB: a knowledge graph for the human genome. *Nucleic Acids Research*, 51(D1):D950–D956, 2023.

[132] Yuanhao Huang, Zhaowei Han, Xin Luo, Xuteng Luo, Yijia Gao, Meiqi Zhao, Feitong Tang, Yiqun Wang, Jiyu Chen, Chengfan Li, et al. Building a literature knowledge base towards transparent biomedical AI. *bioRxiv*, pages 2024–09, 2024.

[133] Ali Pazokitoroudi, Yue Wu, Kathryn S Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. Efficient variance components analysis across millions of genomes. *Nature communications*, 11(1):4020, 2020.

[134] Ulzee An, Ali Pazokitoroudi, Marcus Alvarez, Lianyun Huang, Silviu Bacanu, Andrew J Schork, Kenneth Kendler, Päivi Pajukanta, Jonathan Flint, Noah Zaitlen, et al. Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nature Genetics*, 55(12):2269–2276, 2023.

[135] Nathan LaPierre, Boyang Fu, Steven Turnbull, Eleazar Eskin, and Sriram Sankararaman. Leveraging family data to design Mendelian randomization that is provably robust to population stratification. *Genome Research*, 33(7):1032–1041, 2023.

[136] Xiaochuan Wang, Yuqi Fang, Qianqian Wang, Pew-Thian Yap, Hongtu Zhu, and Mingxia Liu. Self-supervised graph contrastive learning with diffusion augmentation for functional MRI analysis and brain disorder detection. *Medical Image Analysis*, 101:103403, 2025.

[137] Chen Li, Hui Geng, Linhua Ji, Xiaojing Ma, Qichao Yin, and Hua Xiong. ESM-1: A novel tumor biomaker and its research advances. *Anti-Cancer Agents in Medicinal Chemistry-Anti-Cancer Agents)*, 19(14):1687–1694, 2019.

[138] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, 2020.