NEJM AI

DOI: 10.1056/AI-S2300233

SPONSORED CONTENT

Multimodal Generative AI for Precision Health

Hoifung Poon, Ph.D.[1]

Published: December 11, 2023

Executive Summary

The dream of precision health is to develop a continuous learning health system where new health information is instantly incorporated to optimize care delivery and accelerate biomedical discovery. Multimodal generative AI has the potential to drastically accelerate progress toward precision health by supercharging the structuring of health data and scaling population-level insight generation. In this article, we will highlight several prominent growth areas in this exciting frontier:

Application: Prior health AI applications often centered around diagnostics, but there are many low-risk yet high-value scenarios across the entire health system that are ripe for impact

Evaluation: Real-world use cases are often under-represented in existing health AI benchmarks; scaling realistic benchmark creation and evaluation is of increasing urgency

Modeling: Unlike standard contrastive learning, multimodal generative AI can benefit from gravitating in text as the "interlingua" of all modalities, given the vast amount of human knowledge captured in state-of-the-art large language models.

Multimodal Generative AI

Health data is inherently multimodal and longitudinal, comprising physical measurements and natural-language narratives. Only a fraction of such data is available in structured forms, as manual curation is costly and hard to scale. Multimodal generative AI can alleviate the curation bottleneck by leveraging self-supervision from cross-modal correspondence and help extract predictive signals not available in standalone modalities.

First-generation foundation models (e.g., ELMo, BERT, PubMedBERT) learn feature representation by pretraining on unlabeled data, but they still require task-specific supervised fine-tuning for individual applications. By contrast, modern generative AI takes self-supervision to the web scale, thus acquiring emergent capabilities not seen in smaller scale of pretraining. In a truly generalist fashion, large language models (LLMs) such as GPT-4 have demonstrated unprecedented prowess in conversation, reasoning, and instruction-following for new tasks without specialized training.

The author affiliations are listed at the end of the article.

In health applications, GPT-4 has already demonstrated its potentially transformative power from drafting in-basket responses to automating clinical documentation, and from scaling clinical trial matching to analyzing population-level real-world data. However, to fulfill the potential of generative AI in precision health, there are still monumental challenges, especially as we broaden the scope to multimodal data such as medical imaging and multi-omics.

Multimodal generative AI has the potential to drastically accelerate progress toward precision health by supercharging the structuring of health data and scaling population-level insight generation.

Scaling Benchmark Creation and Evaluation

Perhaps counterintuitively, the most fundamental challenge facing health generative AI lies less in model building, but in creating realistic benchmarks that reflect real-world use cases, as well as scaling evaluation for complex generation tasks. Without reliable goalposts, it's hard to measure progress, especially for modern foundation models with open-ended capabilities, such as GPT-4V. It is exciting to see a plethora of medical AI explorations in diagnostics. But the potential impact of health generative AI is much broader, encompassing many low-risk yet high-value scenarios across the entire health system, from clerical and documentation tasks in health delivery, to clinical trials and post-market surveillance in drug development, to early biomedical discovery and value-based care.

Existing biomedical AI benchmarks have produced tremendous value for the research community, but they barely scratch the surface given the breadth and depth of precision health applications. Additionally, evaluating generative AI systems is difficult, as semantically correct outputs may manifest in superficially varying forms. For example, the standard metrics of BLEU and ROUGE for evaluating radiology image report generation essentially assess keyword overlaps between the gold report and system-generated report, which over-index on stylistic variations and fail to reflect real-world usefulness. Factuality tests such as CheXbert and RadGraph aim to partially address the gap by probing the semantic correctness in the generated content, but they only cover limited slivers of health applications (a small number of radiology findings in emergency medicine). As automatic metrics, their own accuracy also has large room for improvement.

To sustain and accelerate progress in health generative AI, it is thus imperative to invest in comprehensive benchmark creation and evaluation that represent diverse populations and real-world scenarios. Manual effort is not scalable, especially for accurate evaluation. Interestingly, generative AI itself may be harnessed to address this challenge. Powerful LLMs such as GPT-4 have already demonstrated remarkable capabilities in self-verifying its own output. It may seem paradoxical at first sight, but verification is often easier than generation, akin to P vs NP in computer science. Developing self-improving LLM-based agent systems to supercharge domain experts in benchmark creation and evaluation represents a key direction forward for health generative AI.

Bridging the Health Competency Gap in Large Multimodal Models (LMMs)

Generalist LLMs such as GPT-4 have demonstrated amazing capabilities in various health applications, but they still have significant growth areas, such as hallucinations. In multimodal generative AI, the potential gap in health competency may be even larger, as there is scarce availability of multimodal, longitudinal health data in the public web for pretraining.

A standard approach to multimodal learning is contrastive learning. Contrastive learning has a long history in NLP and computer vision, initially focusing on unimodal scenarios (e.g., by creating semantic-equivalent variants of an image and training the encoder to map them closer to each other). It can be extended to the multimodal regime, as exemplified by CLIP, Florence, and many contemporary explorations, by pushing corresponding instances across modalities (e.g., image-text pairs) closer and non-corresponding ones farther apart, akin to machine translation.

To sustain and accelerate progress in health generative AI, it is thus imperative to invest in comprehensive benchmark creation and evaluation that represent diverse populations and real-world scenarios.

For precision health, a major bottleneck lies in the scarcity of biomedical data in common resources for multimodal pretraining. One way to address this deficiency is by extracting cross-modal data from publicly available resources, as exemplified by BiomedCLIP. By extracting biomedical figures and their corresponding captions and citations from PubMed-Central articles, the BiomedCLIP project has assembled one of the largest datasets for biomedical vision-language pretraining, with 15 million image-text pairs across diverse categories and 46 million fine-grained pairs after splitting composite figures.

Standard contrastive learning treats all modalities equally. In practice, text captures a large swath of digitized knowledge, whereas data for other modalities is much scarcer by comparison. Multimodal learning can become more efficient by anchoring on the text modality, as exemplified by LLaVA-Med. LLaVA-Med uses cross-modal data to train a projection layer for mapping non-text modalities to the text space, effectively treating text as the interlingua of all modalities. Multimodal input is tokenized and embedded in the natural-language instruction for output generation by a large language model. Subsequent biomedical large multimodal models (e.g., MedPaLM-M and Med-Flamingo) all follow the same general recipe to align image and text, but LLaVA-Med also stands out by using GPT-4 to generate multimodal instruction-following data. Similar approaches can also be applied to other modalities such as single-cell data and protein sequences.

Moonshot: Population-Level Health LMM

Ultimately, routinely collected real-world clinical data represents the biggest growth opportunity for multimodal generative AI. For example, a cancer patient may have hundreds of clinical notes with their corresponding imaging, multi-omics, and other medical tests, across the disease trajectory from diagnosis to treatment to outcome. Such naturally occurring multimodal, longitudinal data can be leveraged to bridge the health competency gap in LMMs, following the general recipe of LLaVA-Med.

For precision health, a major bottleneck lies in the scarcity of biomedical data in common resources for multimodal pretraining.

Microsoft Research's ongoing digital pathology exploration has demonstrated the promise in this direction. In this work, Microsoft and Providence researchers have collaborated to train real-world pathology foundation models using over one billion images from hundreds of thousands of slides along with reports at Providence, with initial promising results in pathomics and progression modeling. Meanwhile, the BiomedJourney project extended the LLaVA-Med approach to image generation, by using GPT-4 to generate instruction-following data from multimodal patient journeys. This enables BiomedJourney to help answer what-if questions for in silico medical imaging: given a prior image and natural-language description of disease progress, simulate what the new image might look like.

By encapsulating multimodal and longitudinal dependencies from population-level real-world data, LMMs can potentially unlock a slew of high-value applications, from cross-modal retrieval and generation, to multimodal conversational copilots (enabling stakeholders to "talk to" multimodal data), to the ultimate dream of precision health: learning from multimodal real-world observational data to predict longitudinal treatment outcome and adverse events, as in real-world evidence.

Author Affiliations

[1] General Manager, Health Futures, Microsoft Research