npj | digital medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

The shaky foundations of large language models and foundation models for electronic health records

Michael Wornow[1]✉, Yizhe Xu[2], Rahul Thapa[2], Birju Patel[2], Ethan Steinberg[1], Scott Fleming[2], Michael A. Pfeffer[2,3], Jason Fries[2] and Nigam H. Shah[2,3,4,5]

The success of foundation models such as ChatGPT and AlphaFold has spurred significant interest in building similar models for electronic medical records (EMRs) to improve patient care and hospital operations. However, recent hype has obscured critical gaps in our understanding of these models' capabilities. In this narrative review, we examine 84 foundation models trained on non-imaging EMR data (i.e., clinical text and/or structured data) and create a taxonomy delineating their architectures, training data, and potential use cases. We find that most models are trained on small, narrowly-scoped clinical datasets (e.g., MIMIC-III) or broad, public biomedical corpora (e.g., PubMed) and are evaluated on tasks that do not provide meaningful insights on their usefulness to health systems. Considering these findings, we propose an improved evaluation framework for measuring the benefits of clinical foundation models that is more closely grounded to metrics that matter in healthcare.

npj Digital Medicine (2023)6:135 ; https://doi.org/10.1038/s41746-023-00879-8

INTRODUCTION

Foundation models (FMs) are machine learning models capable of performing many different tasks after being trained on large, typically unlabeled datasets[1]. FMs represent a paradigm shift in how machine learning (ML) models are developed—rather than developing a bespoke model for each specific use case (as was done traditionally), a single FM can instead be reused across a broad range of downstream tasks with minimal adaptation or retraining needed per task. FMs have received significant attention given their impressive range of capabilities across multiple domains, from text generation[2] and video editing[3] to protein folding[4] and robotics[5].

One of the most popular FMs has been OpenAI's ChatGPT, which surpassed 100 million users within two months of release[6]. ChatGPT is a large language model (LLM), a type of FM which ingests text and outputs text in response. Though ChatGPT was trained to simply predict the next word in a sentence—it is basically an advanced autocomplete— incredible capabilities "emerged" from this training setup which allow the model to perform a wide variety of complex tasks involving language[7]. Physicians were quick to apply the model to pass medical licensing exams[8–11], simplify radiology reports[12], and write research articles[13]. In addition to text, FMs built on structured EMR data have shown the ability to predict the risk of 30-day readmission[14], select future treatments[15], and diagnose rare diseases[16].

The breakneck progress of AI over the past year has made it difficult for healthcare technology professionals and decision-makers to accurately assess the strengths and limitations of these innovations for clinical applications. Beyond short demos being shared on social media, there is little systematic examination of what the best use cases for production-grade clinical FMs are, or how healthcare organizations should weigh their benefits against their substantial risks[1,17–19]. Clinical FMs lack the shared evaluation frameworks and datasets[20] that have underpinned progress in other fields, such as natural language processing (NLP) and computer vision[21]. This makes it difficult to quantify and compare these models' capabilities.

If we believe that FMs can help both providers and patients[22], then rigorous evaluations must be conducted to test these beliefs. In this review, we uncover notable limitations in how clinical FMs are evaluated and a large disconnect between their evaluation regimes and assumed clinical value. While adopting FMs into healthcare has immense potential[23], until we know how to evaluate whether these models are useful, fair, and reliable, it is difficult to justify their use in clinical practice. Inspired by recent efforts to holistically evaluate LLMs trained on non-clinical text for a range of capabilities beyond accuracy[24], we believe that a similar approach is necessary to tie the evaluation of FMs writ large with use cases that matter in healthcare.

To clarify these challenges, we reviewed over 80 different clinical FMs built from electronic medical record (EMR) data. We included all models trained on structured (e.g., billing codes, demographics, lab values, and medications) and unstructured (e.g., progress notes, radiology reports, other clinical text) EMR data, but explicitly excluded images, genetics, and wearables to manage the scope of this review. We refer to the combination of structured and unstructured EMR data (excluding images) as simply "EMR data" or "clinical data"[25]. We refer to FMs built on these forms of clinical data as "clinical foundation models" or "clinical FMs." Our primary contributions are:

1. *To our knowledge, we present the largest review of clinical FMs for structured and unstructured EMR data*. We organize these models into a simple taxonomy to clearly delineate their architectures, training data, capabilities, and public accessibility.
2. *We summarize the currently used evaluation frameworks for clinical FMs and identify their limitations*. We explain why current evaluation tasks provide little evidence for the purported benefits of FMs to a health system.

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA, USA. [3]Technology and Digital Services, Stanford Health Care, Palo Alto, CA, USA. [4]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [5]Clinical Excellence Research Center, Stanford University School of Medicine, Stanford, CA, USA. ✉email: mwornow@stanford.edu

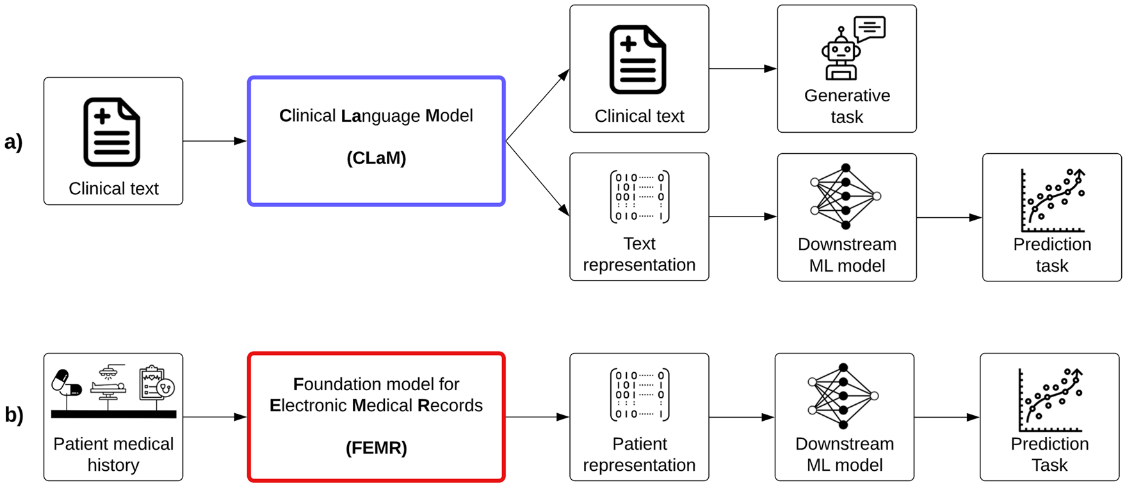


Fig. 1 The two types of clinical FMs. Overview of the inputs and outputs of the two main types of clinical FMs. **a** The inputs and outputs of Clinical Language Models (CLaMs). CLaMs ingest clinical text and output either clinical text or a machine-understandable representation of the input text, which can then be used for downstream prediction tasks. **b** The inputs and outputs of Foundation models for Electronic Medical Records (FEMRs). FEMRs ingest a patient's medical history—which is simply a sequence of medical events with some temporal ordering—and output a machine-understandable representation of the patient, which can then be used for downstream prediction tasks.

3. *We propose an improved framework for evaluating clinical FMs.* We advocate for metrics, tasks, and datasets that better capture the presumed value of clinical FMs.

We begin with a brief overview of clinical FMs and define their inputs, outputs, and capabilities in "What are clinical FMs?". In "Benefits of clinical FMs", we summarize the primary value propositions of FMs for health systems. In "State of published clinical FMs", we provide an overview of the training data behind clinical FMs, examine current evaluation regimens and identify their limitations, and propose a framework for improving these evaluations. Finally, we discuss the promise of clinical FMs for solving a diverse range of healthcare problems in Discussion.

WHAT ARE CLINICAL FMS?

A foundation model (FM) is a type of machine learning model that has been pre-trained on large amounts of unlabeled data and can be adapted to a broad range of downstream tasks[1]. FMs leverage a training procedure referred to as "pre-training," in which a "self-supervised" (i.e., no labels are required) learning objective is used to scale learning to immense amounts (i.e., terabytes) of unlabeled data. FMs also typically have significantly more parameters than traditional ML models—sometimes in the hundreds of billions of parameters—which requires significant computational resources to train (i.e., months of time on a supercomputer with hundreds of GPUs)[26]. The significantly larger size of FMs, coupled with their task-agnostic self-supervised learning objective, has sparked a paradigm shift in how ML models are developed, and has resulted in the "emergence" of unprecedented capabilities at sufficient model scale[27].

Clinical FMs are foundation models built specifically for electronic medical record data. There are two broad categories of clinical FMs: Clinical language models (CLaMs) and Foundation models for EMRs (FEMRs).

Clinical language models (CLaMs)

The first category of FMs are clinical language models, or CLaMs, which are a subtype of large language models (LLMs). As shown in Fig. 1a, the unique attribute that separates CLaMs from general LLMs is their specialization on clinical/biomedical text—CLaMs are primarily trained on, ingest, and output clinical/biomedical text. For example, a CLaM could extract drug names from a doctor's note[28], automatically reply to patient questions[29], summarize medical dialogues[30], or predict mechanical ventilation needs based on clinical notes[31].

While general-purpose LLMs (e.g., ChatGPT, Bloom, GPT-4, etc.) trained on text scraped from the Internet can also be useful for clinical tasks, they tend to underperform CLaMs on domain-specific tasks[32,33], and thus we exclude them from this discussion. However, the conclusions from this review should also readily apply to these general-purpose models, as they suffer from the same limitations that we describe for CLaMs.

Foundation models for electronic medical records (FEMRs)

The second class of clinical FMs are foundation models for electronic medical records (FEMRs). These models are trained on the entire timeline of events in a patient's medical history. Given a patient's EMR as input, a FEMR will output not clinical text but rather a machine-understandable "representation" for that patient, as shown in Fig. 1b. This representation —also referred to as a "patient embedding"—is typically a fixed-length, high-dimensional vector which condenses large amounts of patient information[34]. A patient's representation can then be used as input to any number of downstream models for different tasks. These downstream models (built on the "foundation" of FEMR representations) tend to be more accurate and robust than traditional machine learning (ML) models on clinically relevant tasks, such as predicting 30-day readmission or long length-of-stay[35].

The input to a FEMR can include many aspects of a patient's medical history, such as structured codes, lab values, claims, and clinical text. In practice, however, FEMRs are typically limited to the single modality of structured codes, as discussed in "State of published clinical FMs".

Though CLaMs and FEMRs have remained fairly separate over the past several years, we note that the distinction between these two lines of work is becoming increasingly blurred as the next generation of foundation models for EMRs becomes more expressive and multimodal in nature.

Benefits of clinical FMs

Given the excitement around FMs in healthcare[23,36–41], we summarize their primary value propositions over traditional ML methods. These advantages could all be highly valuable to a health system. Thus, it is essential that our evaluation tasks, datasets, and metrics provide accurate assessments of these purported benefits.

1. **Clinical FMs have better predictive performance.** By using larger training datasets and more model parameters, FMs can achieve better predictive performance (e.g., higher sensitivity and specificity on classification tasks) than traditional ML models[34].
2. **Clinical FMs require less labeled data ("improved sample efficiency").** FMs enable superior model performance using fewer labeled data via "transfer learning"[42]. The core idea behind transfer learning is to first "pre-train" a model on large amounts of non-task-specific (and often unlabeled) data to teach the model general patterns. Then, the model is "fine-tuned" (i.e., continued to be trained) on a smaller dataset specific to the desired task. For example, a sentiment classification model pre-trained on the raw text of Wikipedia before being fine-tuned on a labeled dataset of 100 Tweets will outperform models solely trained on the smaller task-specific dataset of Tweets[42]. Additionally, some FMs can be directly applied to novel tasks without any additional fine-tuning via "zero-shot" or "few-shot" learning. In zero-shot learning, a model learns an entirely new task without being given any specific examples for that task—in other words, the model is given zero examples from which to learn and must instead rely on its general reasoning capabilities to complete the desired task. Similarly, in few-shot learning, the model is only provided with a few examples (typically less than 64) from which to learn. Zero/few-shot learning are particularly powerful capabilities, as they enable FMs to rapidly adapt to new tasks without the need for large, task-specific labeled datasets. Thus, by learning representations that are useful for many downstream tasks via self-supervised pre-training, FMs can greatly reduce the cost of developing ML models for a particular task.
3. **Clinical FMs enable simpler and cheaper model deployment.** After an FM is trained, it can help to decrease the time, talent, and resources required to build subsequent ML models by serving as the figurative "foundation" upon which these subsequent applications are built[1]. Numerous companies have already commercialized this "ML-as-a-Service" approach, in which a centralized FM is made available to end-users via a simple API[43]. A similar approach could work in healthcare, wherein a clinical FM allows informaticians to integrate AI-related capabilities into applications while avoiding the expensive data ingestion, preprocessing, model training, and deployment steps in a typical ML pipeline[44].
4. **Clinical FMs exhibit "emergent" capabilities that enable new clinical applications.** The large number of parameters in FMs has resulted in a phenomenon known as "emergence," in which previously intractable problems become tractable at sufficient model scale[7]. For example, CLaMs can now write coherent insurance appeals in ways thought impossible only a couple of years ago[45], while FEMRs can generate compact patient representations that enable time-to-event modeling of hundreds of outcomes simultaneously[46].
5. **Clinical FMs can more effectively handle multimodal data.** FMs can be designed to accept a wide range of data modalities (e.g., structured codes, lab values, clinical text, images, speech patterns, etc.) as inputs and incorporate them into a single unified representation[47]. Substantial prior work has shown that these models' large parameter counts and dataset sizes enable them to effectively model disparate modalities in the same shared latent space, thereby deriving richer representations for each modality than possible with unimodal models[48–51]. This is especially useful in medicine, given the many types of data produced by patients[52]. For example, a model might simultaneously consider an MRI scan, vital signs, and progress notes when predicting a patient's optimal treatment[53].
6. **Clinical FMs offer novel interfaces for human-AI interaction.** Via a technique called "prompting", a human can input natural language into an LLM and have the model respond in natural language[2]. This enables a two-way conversation between humans and machine, and allows for the decomposition of problems into smaller steps via techniques such as "chain-of-thought" prompting[54]. Prompting generalizes beyond natural language. For example, a FEMR could be prompted with a desired clinical end state (e.g., normal A1C level) to identify which medications should be prescribed to achieve it[55].

STATE OF PUBLISHED CLINICAL FMS

We identified 84 distinct clinical FMs published before March 1, 2023. Specifically, we identified 50 CLaMs and 34 FEMRs by following citations from several representative samples of recent work, as well as manual article curation. Given the rapid pace at which this field advances, we do not claim to include every possible model or cover every recent advancement in the clinical FM space, but rather aim at capturing the general narrative direction of the field. We believe the papers that we selected should adequately capture the general themes that would be identified in other types of reviews, as they are representative of the most recent work in the field, and therefore do not make any claims about the systematicity of our search process. We focus exclusively on models that utilize structured and unstructured EMR data (excluding images) to scope this review.

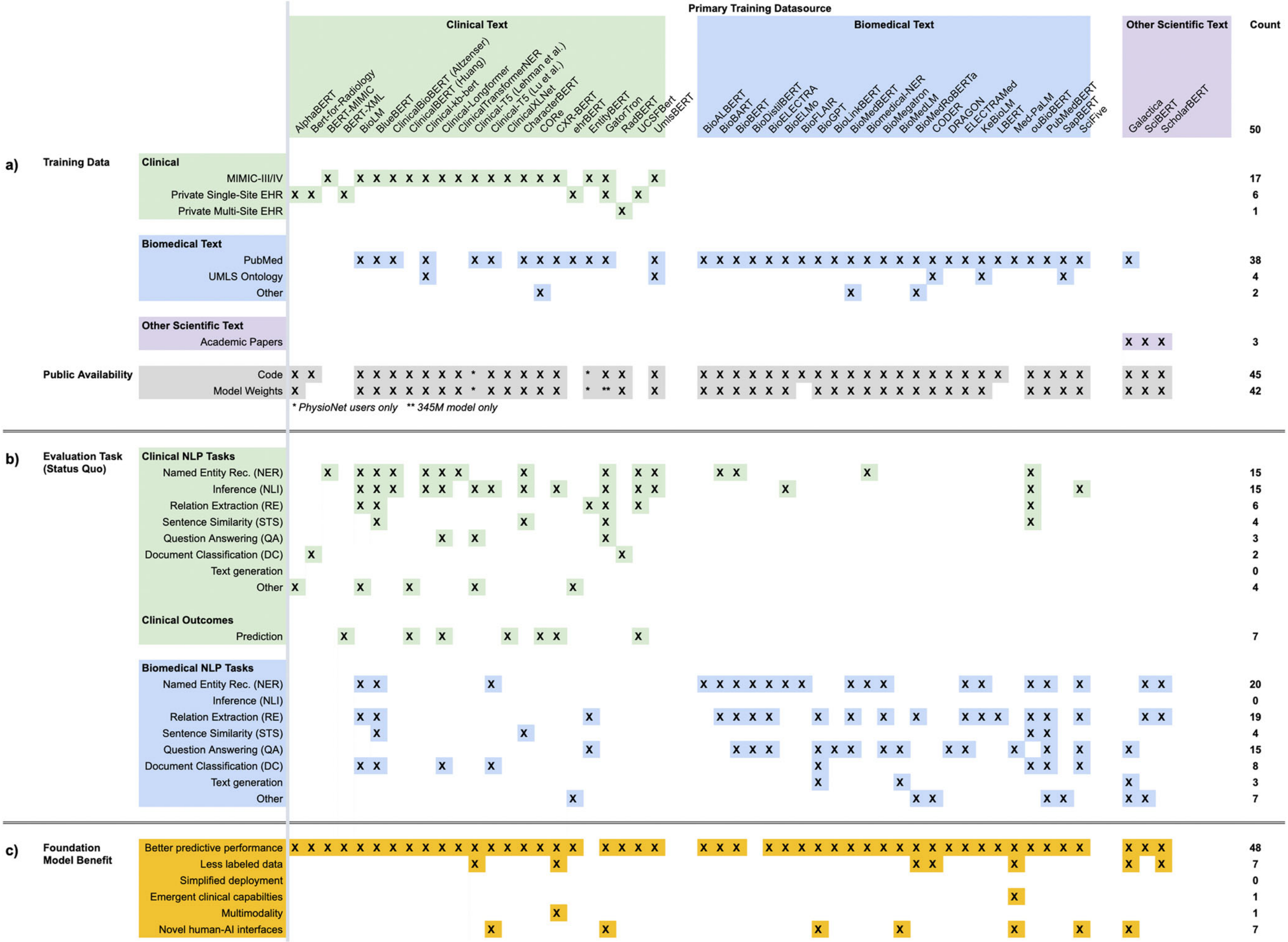
In the following section, we review the training data and public availability of both CLaMs and FEMRs.

CLaMs

Training data. CLaMs (Fig. 2a) are primarily trained on either clinical text (i.e., documents written during the course of care delivery) or biomedical text (i.e., publications on biomedical topics). Almost all CLaMs trained on the clinical text used a single database: MIMIC-III, which contains approximately 2 million notes written between 2001–2012 in the ICU of the Beth Israel Deaconess Medical Center[56]. CLaMs trained on biomedical text are virtually always trained on PubMed abstracts and/or full-text articles. While most CLaMs trained on clinical text are also trained on biomedical text, the converse is not true.

Model availability. Almost all CLaMs have been made publicly accessible via online model repositories like HuggingFace[57]. Unfortunately, the exceptions are the very CLaMs that seem to have the best performance[58] —ehrBERT[59], UCSF-Bert[58], and GatorTron[60]— as they were trained on private EMR datasets.

Takeaways. The high number of CLaMs published over the past several years may lead us to mistake motion for progress. Nearly all CLaMs have been trained on just two datasets -- MIMIC-III and PubMed, which respectively contain about 2 million clinical notes and 16 million abstracts with 5 million full-text publications. Combined, these two datasets contain about 18.5 billion words, which means models trained on them have substantial gaps in completeness (i.e., any scientific knowledge not contained within these corpora) and timeliness (i.e., any new diseases, treatments, or practices discovered after 2012 in the case of MIMIC-III). Empirically, we see that models trained on large-scale EHR data



Section	Group	Row	Count
a) Training Data	Clinical	MIMIC-III/IV	17
		Private Single-Site EHR	6
		Private Multi-Site EHR	1
	Biomedical Text	PubMed	38
		UMLS Ontology	4
		Other	2
	Other Scientific Text	Academic Papers	3
Public Availability		Code	45
		Model Weights	42
b) Evaluation Task (Status Quo)	Clinical NLP Tasks	Named Entity Rec. (NER)	15
		Inference (NLI)	15
		Relation Extraction (RE)	6
		Sentence Similarity (STS)	4
		Question Answering (QA)	3
		Document Classification (DC)	2
		Text generation	0
		Other	4
	Clinical Outcomes	Prediction	7
	Biomedical NLP Tasks	Named Entity Rec. (NER)	20
		Inference (NLI)	0
		Relation Extraction (RE)	19
		Sentence Similarity (STS)	4
		Question Answering (QA)	15
		Document Classification (DC)	8
		Text generation	3
		Other	7
c) Foundation Model Benefit		Better predictive performance	48
		Less labeled data	7
		Simplified deployment	0
		Emergent clinical capabilities	1
		Multimodality	1
		Novel human-AI interfaces	7

Fig. 2 Overview of CLaMs. A summary of CLaMs and how they were trained, evaluated, and published. Each column is a specific CLaM, grouped by the primary type of data they were trained on. Columnwise, the CLaMs primarily trained on clinical text are green ($n = 23$), those trained primarily on biomedical text are blue ($n = 24$), and models trained on general academic text are purple ($n = 3$). The last column is the count of entries in each row. An **X** indicates that the model has that characteristic. An * indicates that a model partially has that characteristic. **a** Training data and public availability of each model. The top rows mark whether a CLaM was trained on a specific dataset, while the bottom-most row records whether a model's code and weights have been published. Almost all CLaMs have had their model weights published, typically via shared repositories like the HuggingFace Model Hub. **b** Evaluation tasks on which each model was evaluated in its original paper. Green rows are tasks whose data were sourced from clinical text and blue rows are evaluation tasks sourced from biomedical text. The tasks are presented by the way they are commonly organized in the literature. CLaMs primarily trained on clinical text are evaluated on tasks drawn from clinical datasets, while CLaMs primarily trained on biomedical text are almost exclusively evaluated on tasks that contain general biomedical text (i.e., not clinical text). **c** Clinical FM benefits on which each model was evaluated in its original paper. The underlying tasks presented in this section are identical to those in (**b**), but here the tasks are reorganized into six buckets that reflect the six primary FM benefits described in Benefits of clinical FMs. While almost all CLaMs have demonstrated the ability to improve predictive accuracy over traditional ML approaches, there is scant evidence for the other five value propositions of clinical FMs.

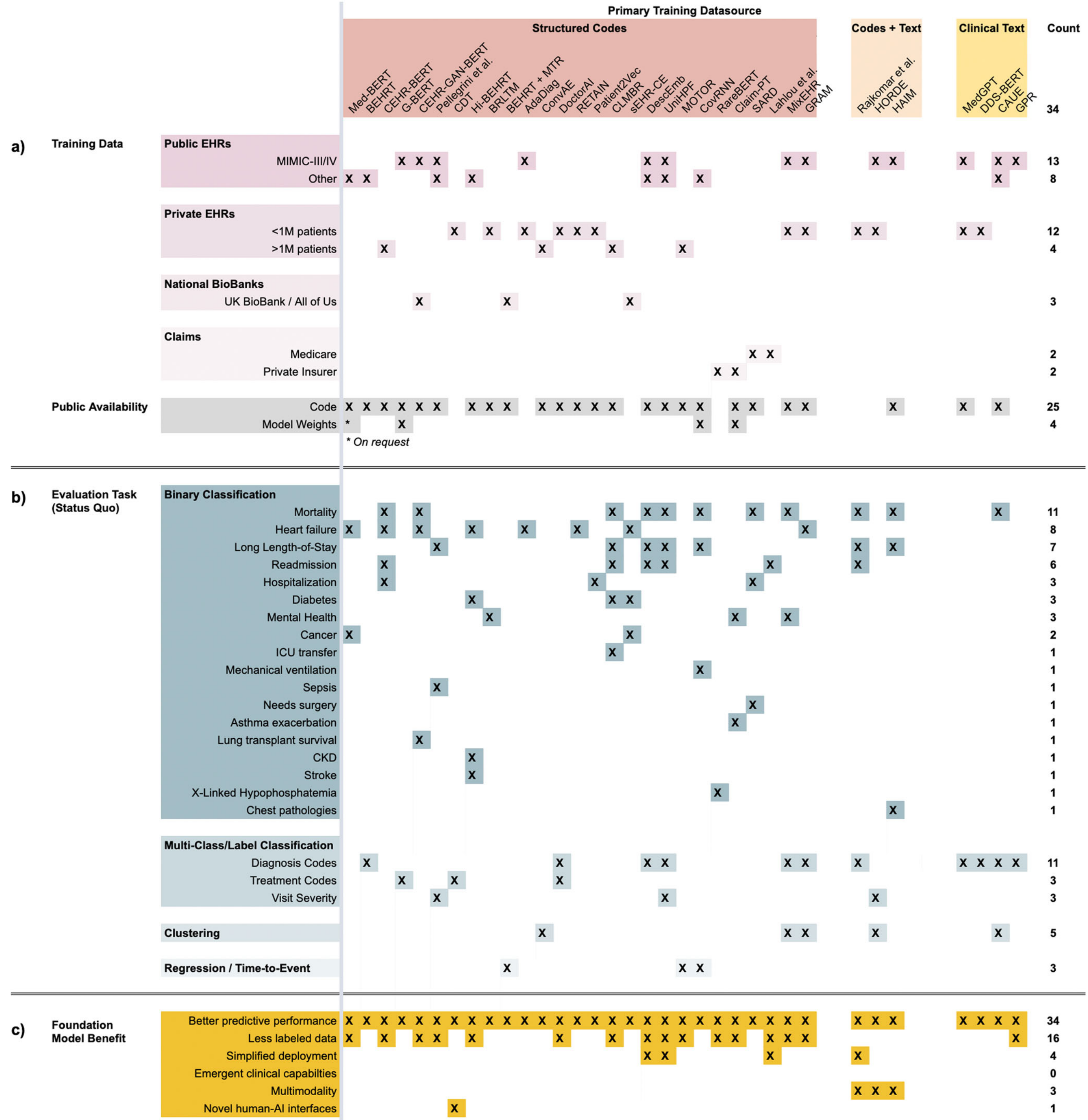
outperform CLaMs trained on shared public datasets across-the-board on out-of-domain data distributions[32,58].

FEMRs

Training data. Most FEMRs (Fig. 3a) are trained on either small, publicly available EMR datasets or a single private health system's EMR database. Again, the most popular public dataset is MIMIC-III, which contains less than 40,000 patients[56]. Other public datasets vary greatly in size, from eICU's 139,000 patients[61] to the CPRD's longitudinal records on 7% of all patients in the UK[62]. Several FEMRs have been trained on insurance claims, which are typically larger in size and more diverse than EMR data but contain less granular information[63]. Examples of claims datasets include Truven Health MarketScan (170 million patients)[64] and Partners For Kids (1.8 million pediatric patients)[65]. In terms of data modalities, most FEMRs are unimodal as they only consider structured codes (e.g., LOINC, SNOMED, etc.).

Model accessibility. FEMRs lack a common mechanism like HuggingFace for distributing models to the research community, as can be seen in the sparsity of the bottom-most row in Fig. 3a compared to the density of the bottom-most row in Fig. 2a. Few FEMRs have had their model weights published, meaning researchers must re-train these models from scratch on local EMR data to verify their performance.

Takeaways. The overreliance on structured codes limits the generalizability of FEMRs across health systems that use different EMR systems and coding practices. Some models, such as DescEmb, address this problem by first converting coded data into their textual descriptions, thus detaching the model from the



			Primary Training Datasource																																		Count
			Structured Codes																											Codes + Text			Clinical Text				
			Med-BERT	BEHRT	CEHR-BERT	G-BERT	CEHR-GAN-BERT	Pellegrini et al.	CDT	Hi-BEHRT	BRLTM	BEHRT + MTR	AdaDiag	ConvAE	DoctorAI	RETAIN	Patient2Vec	CLMBR	sEHR-CE	DescEmb	UniHPF	MOTOR	CovRNN	RareBERT	Claim-PT	SARD	Lahlou et al.	MixEHR	GRAM	Rajkomar et al.	HORDE	HAIM	MedGPT	DDS-BERT	CAUE	GPR	34
a) Training Data	Public EHRs	MIMIC-III/IV				X	X	X					X							X	X							X	X		X	X	X		X	X	13
		Other	X	X				X		X										X	X		X												X		8
	Private EHRs	<1M patients							X		X		X		X	X	X											X	X	X	X		X	X			12
		>1M patients			X									X				X				X															4
	National BioBanks	UK BioBank / All of Us					X					X							X																		3
	Claims	Medicare																								X	X										2
		Private Insurer																						X	X												2
Public Availability		Code	X	X	X	X	X	X		X	X	X		X	X	X	X	X		X	X	X	X		X	X		X	X			X	X		X		25
		Model Weights	*			X																	X		X												4
b) Evaluation Task (Status Quo)	Binary Classification	Mortality			X		X											X		X	X		X			X		X		X		X			X		11
		Heart failure	X		X		X			X			X			X			X										X								8
		Long Length-of-Stay						X										X		X	X		X							X		X					7
		Readmission			X													X		X	X						X			X							6
		Hospitalization			X												X									X											3
		Diabetes								X								X	X																		3
		Mental Health									X														X			X									3
		Cancer	X																X																		2
		ICU transfer																X																			1
		Mechanical ventilation																					X														1
		Sepsis						X																													1
		Needs surgery																								X											1
		Asthma exacerbation																							X												1
		Lung transplant survival					X																														1
		CKD								X																											1
		Stroke								X																											1
		X-Linked Hypophosphatemia																						X													1
		Chest pathologies																														X					1
	Multi-Class/Label Classification	Diagnosis Codes		X											X					X	X							X	X	X			X	X	X	X	11
		Treatment Codes				X			X						X																						3
		Visit Severity						X													X										X						3
	Clustering													X														X	X		X				X		5
	Regression / Time-to-Event											X										X	X														3
c) Foundation Model Benefit		Better predictive performance	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	34
		Less labeled data	X		X		X	X		X					X			X		X	X	X		X	X		X	X	X							X	16
		Simplified deployment																		X	X						X			X							4
		Emergent clinical capabilites																																			0
		Multimodality																												X	X	X					3
		Novel human-AI interfaces							X																												1

* On request

Fig. 3 Overview of FEMRs. A summary of FEMRs and how they were trained, evaluated, and published. Each column is a specific FEMR, grouped by the primary type of data they were trained on. Columnwise, the FEMRs primarily trained on structured EMR codes (e.g., billing, medications, etc.) are red ($n = 27$), those trained on both structured codes and clinical text are orange ($n = 3$), and models trained only on clinical text are yellow ($n = 4$). The last column is the count of entries in each row. An **X** indicates that the model has that characteristic. An * indicates that a model partially has that characteristic. **a** Training data and public availability of each model. The top rows mark whether a FEMR was trained on a specific dataset, while the bottom-most row records whether a model's code and weights have been published. Very few FEMRs have had their model weights published, as they are limited by data privacy concerns and a lack of interoperability between EMR schemas. **b** Evaluation tasks on which each model was evaluated in its original paper. From top to bottom, the evaluation tasks are binary classification, multi-class/label classification, clustering of patients/diseases, and regression tasks like time-to-event. The tasks are presented by the way they are commonly organized in the literature. FEMRs are evaluated on a very broad and sparse set of evaluation tasks—even the same nominal task will often have different definitions across papers. **c** Clinical FM benefits on which each model was evaluated in its original paper. The underlying tasks presented in this section are identical to those in (**b**), but here the tasks are reorganized into six buckets that reflect the six primary FM benefits described in "Benefits of clinical FMs". While almost all FEMRs have demonstrated the ability to improve predictive accuracy over traditional ML approaches, and a significant number have demonstrated improved sample efficiency, there is scant evidence for the other four value propositions of clinical FMs.

specific codes on which it was trained[66]. An additional limitation of relying on coded data is that it contains inconsistencies and errors[67], and often provides an incomplete picture of patient state[68]. Some FEMRs have tackled this problem by combining unstructured EHR data (i.e., text) with structured EMR data to boost performance on specific phenotyping and prediction tasks[69,70]. However, the key unsolved challenge of how to publicly share pre-trained FEMRs continues to hinder the field's progress and precludes the primary value proposition of FMs—namely, being able to build off a pre-trained model.

Next, we considered the common evaluation frameworks for clinical FMs. The common thread between most of these evaluations is that they are relatively straightforward to conduct in an automated fashoin. While these tasks provide diagnostic insights on model behavior, they provide limited insight into the claims of FMs being a "categorically different" technology[71,72], and offer little evidence for the clinical utility achieved by these models. Taking inspiration from the broader ML community's push towards Holistic Evaluation of Language Models[24], we do a critical evaluation of the evaluations currently used to evaluate clinical FMs.

CLaMs

Evaluation of standard tasks and datasets. We collected every evaluation task that a CLaM was evaluated on in its original publication in Fig. 2b, and grouped these tasks as they are commonly reported in the literature. Most CLaMs are being evaluated on traditional NLP-style tasks such as named entity recognition, relation extraction, and document classification on either MIMIC-III (clinical text) or PubMed (biomedical text)[73,74]. Given that clinical text has its own unique structure, grammar, abbreviations, terminology, formatting, and other idiosyncrasies not found in other domains[75], it is alarming that roughly half of all CLaMs surveyed were not validated on clinical text, and thus may be overestimating their expected performance in a healthcare setting.

When NLP tasks are sourced from clinical text, they can be useful measures of a model's linguistic capabilities. However, these NLP tasks are greatly limited by their overreliance on the same handful of data sources[74], small dataset sizes (typically thousands of examples)[74,76], highly repetitive content[77], and low coverage of use cases[20]. As a result, strong performance on a clinical NLP task does not provide compelling evidence to a hospital looking to deploy a CLaM— claiming that *"Model A achieves high precision on named entity recognition on 2,000 discharge notes from MIMIC-III*" is very different than "*Model A should be deployed across all of Health System X to identify patients at risk of suicide*".

Evaluation on FM benefits. To illustrate the disconnect between current evaluation tasks and the loftier promises of clinical FMs, we reorganized the rows of evaluation tasks from Fig. 2b—originally presented as they are typically grouped in the literature—along the six primary FM value propositions from "Benefits of Clinical FMs". The result is Fig. 2c, which identifies which CLaMs were evaluated against any of the six core benefits of clinical FMs. Most CLaMs have only shown evidence for one FM value proposition: improved predictive accuracy on certain tasks. However, there is little evidence supporting the other purported benefits of FMs, such as simplified model deployment or reducing the need for labeled data. For example, while zero- and few-shot prompting techniques have been rigorously studied for general-purpose LLMs as an important method for achieving improved performance, few CLaMs have been evaluated across different prompting strategies and fine-tuning techniques. In other words, there is a gap in our understanding of what CLaMs *can do* versus what CLaMs can do that is valuable to a health system and which traditional ML models cannot do.

FEMRs

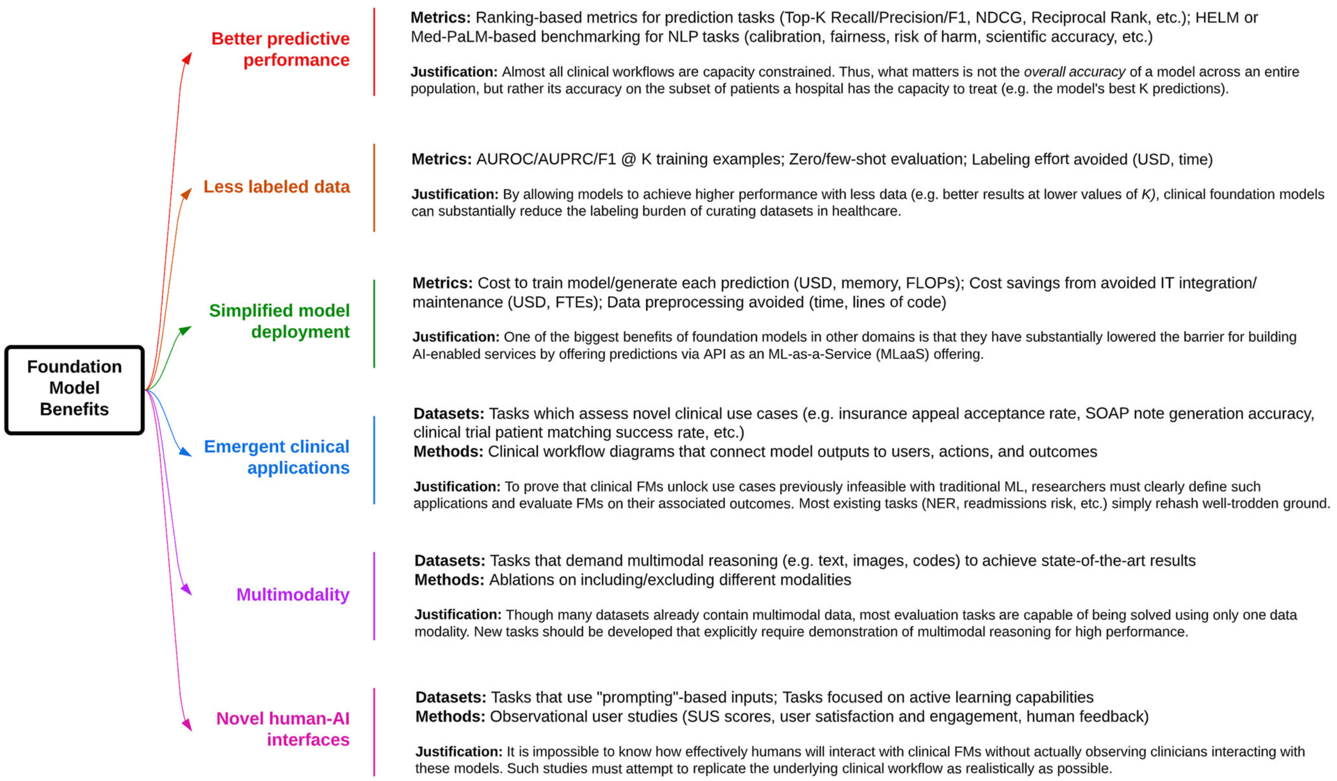
Evaluation on standard tasks and datasets. We collected the original tasks on which each FEMR was evaluated in Fig. 3b and bucketed them as they are typically presented in the literature. Evaluation of FEMRs is in an even poorer state than that of CLaMs. While CLaMs benefit from the NLP community's adoption of standardized task formats, FEMRs lack a similar set of "canonical" evaluations. Instead, FEMRs are evaluated on an extremely sparse set of tasks with little-to-no overlap across publications. This makes it highly non-trivial to compare the performance of different FEMRs.

These tasks are typically grouped by how each task is formulated, e.g., binary classification v. multi-label classification v. regression. The most popular prediction tasks are binary classification tasks such as mortality, heart failure, and long length-of-stay, but even the same nominal task can have widely divergent definitions across papers[78].

Evaluation on FM benefits. We reorganized the rows of evaluation tasks from Fig. 3b along the six primary value propositions of clinical FMs listed in "Benefits of Clinical FMs". The result is Fig. 3c, which shows that almost all evaluations of FEMRs have been focused on demonstrating their superior predictive accuracy over traditional ML models. Notably, the ability to use less labeled data (i.e., sample efficiency) has been fairly well-documented with FEMRs. However, the other four potential benefits of FMs have gone largely unstudied. And while evaluations of predictive accuracy are straightforward to perform, it is not the sole property of FMs that would justify their adoption by a health system.

Finally, to better quantify the ability of clinical FMs to achieve the six key benefits of FMs outlined in "Benefits of clinical FMs", we propose several improved evaluation metrics and tasks in Fig. 4. Our suggestions are by no means comprehensive, but rather meant to spark a further discussion on how to align model evaluation with the demonstration of clinical value.

1. **Better predictive performance:** The most thoroughly studied property of clinical FMs has been their improved predictive performance on classification and regression tasks based on AUROC, AUPRC, F1 Score, and Accuracy. These metrics assume an infinite capacity to act on a model's predictions. In reality, clinical workflows are capacity constrained—a nursing team may only be able to act on a handful of model predictions per day[79,80]. Thus, a health system should only care about a model's accuracy on patients for which it has the capacity to intervene. We, therefore, recommend that researchers adopt ranking-based metrics (e.g., Top-K precision/recall/F1, reciprocal ranking, etc.), which are commonly used for recommendation systems[81]. Additionally, we propose examining not just a model's ability to classify patients correctly but also its calibration across subgroups, fairness, and alignment with clinical best practices[24,29]. The human evaluation may also be necessary in some cases, such as evaluating a CLaM's ability to accurately generate answers to clinical questions[82]. Traditional NLP metrics such as ROUGE, METEOR, and BLEU —which simply count n-gram overlap between generated and reference text—are known to poorly correlate with human evaluations of natural language generations[83–85]. We also lack automated metrics for evaluating the more qualitative aspects of a model's "alignment" with human values (e.g., helpfulness or harmlessness)[86], even as the importance of human feedback during training has been repeatedly demonstrated via techniques like Reinforcement


Foundation Model Benefits
Better predictive performance
Less labeled data
Simplified model deployment
Emergent clinical applications
Multimodality
Novel human-AI interfaces

Metrics: Ranking-based metrics for prediction tasks (Top-K Recall/Precision/F1, NDCG, Reciprocal Rank, etc.); HELM or Med-PaLM-based benchmarking for NLP tasks (calibration, fairness, risk of harm, scientific accuracy, etc.)

Justification: Almost all clinical workflows are capacity constrained. Thus, what matters is not the *overall accuracy* of a model across an entire population, but rather its accuracy on the subset of patients a hospital has the capacity to treat (e.g. the model's best K predictions).

Metrics: AUROC/AUPRC/F1 @ K training examples; Zero/few-shot evaluation; Labeling effort avoided (USD, time)

Justification: By allowing models to achieve higher performance with less data (e.g. better results at lower values of *K)*, clinical foundation models can substantially reduce the labeling burden of curating datasets in healthcare.

Metrics: Cost to train model/generate each prediction (USD, memory, FLOPs); Cost savings from avoided IT integration/ maintenance (USD, FTEs); Data preprocessing avoided (time, lines of code)

Justification: One of the biggest benefits of foundation models in other domains is that they have substantially lowered the barrier for building AI-enabled services by offering predictions via API as an ML-as-a-Service (MLaaS) offering.

Datasets: Tasks which assess novel clinical use cases (e.g. insurance appeal acceptance rate, SOAP note generation accuracy, clinical trial patient matching success rate, etc.)
Methods: Clinical workflow diagrams that connect model outputs to users, actions, and outcomes

Justification: To prove that clinical FMs unlock use cases previously infeasible with traditional ML, researchers must clearly define such applications and evaluate FMs on their associated outcomes. Most existing tasks (NER, readmissions risk, etc.) simply rehash well-trodden ground.

Datasets: Tasks that demand multimodal reasoning (e.g. text, images, codes) to achieve state-of-the-art results
Methods: Ablations on including/excluding different modalities

Justification: Though many datasets already contain multimodal data, most evaluation tasks are capable of being solved using only one data modality. New tasks should be developed that explicitly require demonstration of multimodal reasoning for high performance.

Datasets: Tasks that use "prompting"-based inputs; Tasks focused on active learning capabilities
Methods: Observational user studies (SUS scores, user satisfaction and engagement, human feedback)

Justification: It is impossible to know how effectively humans will interact with clinical FMs without actually observing clinicians interacting with these models. Such studies must attempt to replicate the underlying clinical workflow as realistically as possible.

Fig. 4 Better evaluations of clinical FMs. Proposals for how to demonstrate the value of CLaMs and FEMRs for achieving the six primary value propositions of FMs to health systems over traditional ML models.

Learning from Human Feedback[87]. This is especially worrying in medical settings, where patient safety is paramount. CLaMs that might impact clinical decisions should be evaluated much more rigorously than automated metrics can provide, across axes such as agreement with scientific consensus, minimization of the extent and risk of harm, possibility of bias, and the clinical utility of the advice[29,82].

2. **Less labeled data:** The simplest way for researchers to demonstrate how clinical FMs exhibit improved sample efficiency is to replace evaluation metric "*X*" with the more nuanced metric "*X using K training examples*". For example, replacing "*AUROC*" with "*AUROC using 1000 labeled radiology reports for fine-tuning.*" Ideally, a clinical FM would enable similar model performance at low values of *K* as at high values of *K*. Another way to demonstrate improved sample efficiency is to measure *zero-shot* and *few-shot* model performance, in which a model is given either zero or <100 examples, respectively, for the task on which it is evaluated. Researchers should also consider measuring the performance difference between fine-tuning versus prompting, where the former has been known to achieve higher accuracy, but the latter represents a much simpler and more flexible deployment option (as the model weights remain frozen)[88]. One could also measure the *total dataset annotation time saved* by using a clinical FM, measured in terms of dollars or hours.
3. **Simplified model deployment:** To quantify the value of FMs in lowering the barrier for building task-specific models[1,89], one possible metric is the cost of hardware/compute/memory needed to train a model or generate a prediction. More broadly, we can measure the overall cost savings of using a clinical FM in terms of full-time equivalents (FTEs) or resource hours saved when downstream models (e.g., risk of inpatient mortality) are built on top of a clinical FM versus training a task-specific model from scratch. We recognize, however, that this evaluation may be the most challenging to conduct, as it requires buy-in from the business, clinical, and IT units of a health system. Health systems with dedicated ML Operations ("MLOps") teams may be better positioned to realize these benefits[90].
4. **Emergent clinical applications:** Clinical FMs can perform entirely novel tasks thought to be beyond the reach of machines even just a year ago, e.g., summarizing MRI reports in patient-accessible terms, writing discharge instructions, or generating differential diagnoses[45,91]. "Emergence" is a term of art used by ML researchers to describe the phenomenon by which FMs trained on large datasets are able to perform tasks that were impossible for smaller ML models to accomplish[7]. While this greatly broadens the range of clinical problems addressable via machine learning, it is still unproven whether these capabilities provide tangible utility to health systems in production settings[92]. Thus, we must explicitly define the scenarios in which the emergent capabilities of clinical FMs achieve their purported benefits. For example, LLMs such as GPT-4 can produce new USMLE exam questions, which are indistinguishable from human-authored questions. However, whether the use of these questions results in better-prepared medical students, or a lower burden for creating exam questions, remains to be quantified[93].
5. **Multimodality:** Currently, the majority of evaluation tasks span one data modality[78], even though models that simultaneously use multiple data modalities show substantial gains[94]. There is a strong unmet need for evaluation scenarios which explicitly require multimodal representations. Many datasets already include multimodal data (e.g., MIMIC-III, eICU, private EMRs, etc.), but evaluation tasks are not constructed in ways that require the demonstration of

multimodal reasoning across both structured data and unstructured text. A great example of datasets that accomplish this are the Holistic AI Framework (HAIM), which builds on top of MIMIC-III to enable truly multimodal evaluation scenarios[95].

6. **Novel human-AI interfaces:** Human evaluation and usability studies are needed to quantify the utility of interacting with FMs via prompts[1]. Metrics include user satisfaction, engagement, system usability scale scores, qualitative interview feedback, and the time/effort required to achieve stated goals[96–98]. Measuring the skill level necessary to operate a model can also shed light on its ability to empower providers to perform a multitude of roles. For FEMRs, an accepted paradigm for "prompting" does not yet exist, so developing a framework for prompting a patient's medical history would represent a significant step forward. One exception is the Clinical Decision Transformer, which used a desired clinical end state (e.g., normal A1C levels) as a prompt to generate medication recommendations[55].

DISCUSSION

Our review of 50 CLaMs and 34 FEMRs shows that most clinical FMs are being evaluated on tasks that provide little information on the potential advantages of FMs over traditional ML models. While there is ample evidence that clinical FMs enable more accurate model predictions, Figs. 2, 3 show that minimal work has been conducted to validate whether the other, potentially more valuable benefits of FMs will be realized in healthcare. These benefits include reducing the burden of labeling data, offering novel human-AI interfaces, and enabling new clinical applications beyond the reach of traditional ML models, among others outlined in "Benefits of Clinical FMs". To help bridge this divide, we advocate for the development of new evaluation tasks, metrics, and datasets more directly tied to clinical utility, as summarized in Fig. 4.

While we focused this review on FMs developed specifically for clinical data, we recognize that there has been significant recent progress in adapting general-purpose LLMs to medical knowledge tasks[11]. As these general-purpose models continue to improve, the need and value of having clinical-specific models remain an open question[32]. However, it is worth emphasizing that the evaluation of these general-purpose LLMs suffers from the same exact limitations as evaluations of clinical LLMs, and the critiques described in this review still apply. While general-purpose LLMs continue to improve on specific clinical tasks, e.g., clinical knowledge and board certification benchmarks, it remains unclear how well they perform for broader applications in the hospital and what is achievable without training on some degree of in-domain data (e.g., EHRs). For example, the fact that GPT-4 passes the USMLE does not necessarily mean the model is useful for the types of questions clinicians care about in practice[82]. We believe more work needs to be done to assess the clinical reasoning capabilities of these general-purpose systems, and to develop a better theoretical understanding of how a model's skills in other domains strengthen or worsen its performance on clinical tasks. There are also considerations beyond overall accuracy, such as scalability and inference cost, that may have different trade-offs in smaller, more targeted clinical-specific FM deployments[99].

In addition to the potential benefits listed in "What are Clinical FMs?", FMs present numerous risks that must also be considered and investigated. Data privacy and security are significant concerns with FMs, as they may leak protected health information through model weights or prompt injection attacks[100,101]. FMs are also more difficult to interpret, edit, and control due to their immense size[102]. They require high up-front costs to create, and while these costs can be amortized over multiple downstream applications, their value may take longer to realize than a smaller model developed for a single high-value task[103]. Additionally, FMs may fall under Software-as-a-Medical-Device guidelines regulating their usage in the clinic[104]. And similar to traditional ML models, FMs are susceptible to biases induced by miscalibration or overfitting[105], as well as inducing "automation bias" in which clinicians defer to a model's outputs even when they are obviously incorrect[106]. Developing frameworks for determining a model's overall worth remains indispensable[79].

Despite these challenges, clinical FMs hold immense promise for solving a diverse range of healthcare problems. We invite the research community to develop better evaluations to help realize their potential for benefiting both patients and providers[22].

DATA AVAILABILITY

We do not have any data beyond what is depicted in the Figures of this paper.

Received: 21 March 2023; Accepted: 13 July 2023;
Published online: 29 July 2023

REFERENCES

1. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at arXiv: 2108.07258 (2021).
2. Brown, T. B. et al. Language models are few-shot learners. Preprint at arXiv:2005.14165 (2020).
3. Esser, P., Chiu, J., Atighehchian, P., Granskog, J. & Germanidis, A. Structure and content-guided video synthesis with diffusion models. Preprint at arXiv: 2302.03011 (2023).
4. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. Jiang, Y. et al. VIMA: general robot manipulation with multimodal prompts. Preprint at arXiv: 2210.03094 (2022).
6. Eysenbach, G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ.* **9**, e46885 (2023).
7. Wei, J. et al. Emergent abilities of large language models. Preprint at arXiv: 2206.07682 (2022).
8. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
9. Gilson, A. et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ. (*2023)
10. Liévin, V., Hother, C. E. & Winther, O. Can large language models reason about medical questions? Preprint at arXiv: :2207.08143 (2022).
11. Nori, H., King, N., Mc Kinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. Preprint at arXiv: 2303.13375 (2023).
12. Jeblick, K. et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Preprint at arXiv: 2212.14882 (2022).
13. Macdonald, C., Adeloye, D., Sheikh, A. & Rudan, I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. *J. Glob. Health* **13**, 01003 (2023).
14. Pang, C. et al. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. Machine Learning for Health. *PMLR* (2021)
15. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F. & Sun, J. Doctor AI: predicting clinical events via recurrent neural networks. Preprint at arXiv: 1511.05942 (2015).
16. Prakash, P. K. S., Chilukuri, S., Ranade, N. & Viswanathan, S. RareBERT: transformer architecture for rare disease patient identification using administrative claims. *AAAI* **35**, 453–460 (2021).
17. Cascella, M., Montomoli, J., Bellini, V. & Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J. Med. Syst.* **47**, 33 (2023).
18. Shen, Y. et al. ChatGPT and other large language models are double-edged swords. *Radiology* **307**, 230163 (2023).
19. Wójcik, M. A. Foundation models in healthcare: opportunities, biases and regulatory prospects in Europe. In *Electronic Government and the Information Systems Perspective*: *11th International Conference, EGOVIS 2022 Proceedings* 32–46 (Springer-Verlag, 2022).

20. Blagec, K., Kraiger, J., Frühwirt, W. & Samwald, M. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *J. Biomed. Inform.* **137**, 104274 (2023).
21. Donoho, D. 50 years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).
22. Topol, E. When M.D. is a machine doctor. https://erictopol.substack.com/p/when-md-is-a-machine-doctor (2023).
23. Robert, P. 5 Ways ChatGPT will change healthcare forever, for better. *Forbes Magazine* (13 February 2023).
24. Liang, P. et al. Holistic evaluation of language models. Preprint at arXiv [cs.CL] (2022).
25. Mohsen, F., Ali, H., El Hajj, N. & Shah, Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Sci. Rep.* **12**, 17981 (2022).
26. BigScience Workshop, et al. BLOOM: a 176B-Parameter open-access multilingual language model. Preprint at arXiv [cs.CL] (2022).
27. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at arXiv [cs.CL] (2023).
28. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. Preprint at arXiv [cs.CL] (2022).
29. Singhal, K. et al. Large language models encode clinical knowledge. Preprint at arXiv [cs.CL] (2022).
30. Chintagunta, B., Katariya, N., Amatriain, X. & Kannan, A. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proc. Second Workshop on Natural Language Processing for Medical Conversations* 66–76 (Association for Computational Linguistics, 2021).
31. Huang, K. et al. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. *Proceedings of the 3rd Clinical Natural Language Processing Workshop* (2020).
32. Lehman, E. et al. Do we still need clinical language models? Preprint at arXiv [cs.CL] (2023).
33. Moradi, M., Blagec, K., Haberl, F. & Samwald, M. GPT-3 models are poor few-shot learners in the biomedical domain. Preprint at arXiv [cs.CL] (2021).
34. Steinberg, E. et al. Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).
35. Guo, L. L. et al. EHR foundation models improve robustness in the presence of temporal distribution shift. *Sci. Rep.* **13**, 3767 (2022).
36. Fei, N. et al. Towards artificial general intelligence via a multimodal foundation model. *Nat. Commun.* **13**, 3094 (2022).
37. Si, Y. et al. Deep representation learning of patient data from Electronic Health Records (EHR): a systematic review. *J. Biomed. Inform.* **115**, 103671 (2021).
38. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
39. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1419–1428 (2018).
40. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Health. J.* **6**, 94–98 (2019).
41. Bohr, A. & Memarzadeh, K. The rise of artificial intelligence in healthcare applications. *Artif. Intell. Healthcare* **25** (2020).
42. Howard, J. & Sebastian, R. Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (2018).
43. Chen, L. et al. HAPI: a large-scale longitudinal dataset of commercial ML API predictions. Preprint at arXiv [cs.SE] (2022).
44. Huge 'foundation models' are turbo-charging AI progress. *The Economist* (15 June 2022).
45. Canes, D. The time-saving magic of Chat GPT for doctors. https://tillthecavalryarrive.substack.com/p/the-time-saving-magic-of-chat-gpt?utm_campaign=auto_share (2022).
46. Steinberg, E., Xu, Y., Fries, J. & Shah, N. Self-supervised time-to-event modeling with structured medical records. Preprint at arXiv [cs.LG] (2023).
47. Kline, A. et al. Multimodal machine learning in precision health: a scoping review. *NPJ Digit. Med.* **5**, 171 (2022).
48. Baevski, A. et al. Data2vec: A general framework for self-supervised learning in speech, vision and language. *International Conference on Machine Learning. PMLR* (2022).
49. Girdhar, R. et al. ImageBind: one embedding space to bind them all. Preprint at arXiv [cs.CV] (2023).
50. Boecking, B. et al. Making the most of text semantics to improve biomedical vision--language processing. Preprint at arXiv [cs.CV] (2022).
51. Radford, A. et al. Learning transferable visual models from natural language supervision. Preprint at arXiv [cs.CV] (2021).
52. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit. Med.* **3**, 136 (2020).
53. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
54. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* (2022).
55. Lee, S., Da Young, L., Im, S., Kim, N. H. & Park, S.-M. Clinical decision transformer: intended treatment recommendation through goal prompting. Preprint at arXiv [cs.AI] (2023).
56. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
57. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. *EMNLP 2020* (2020).
58. Sushil, M., Ludwig, D., Butte, A. J. & Rudrapatna, V. A. Developing a general-purpose clinical language inference model from a large corpus of clinical notes. Preprint at arXiv [cs.CL] (2022).
59. Li, F. et al. Fine-tuning bidirectional encoder representations from transformers (BERT)-based models on large-scale electronic health record notes: an empirical study. *JMIR Med. Inf.* **7**, e14830 (2019).
60. Yang, X. et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.27.22271257 (2022).
61. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
62. Li, Y. et al. Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE J. Biomed. Health Inform.* **27** (2022).
63. Zeltzer, D. et al. Prediction accuracy with electronic medical records versus administrative claims. *Med. Care* **57**, 551–559 (2019).
64. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, 86 (2021).
65. Zeng, X., Linwood, S. L. & Liu, C. Pretrained transformer framework on pediatric claims data for population specific tasks. *Sci. Rep.* **12**, 3651 (2022).
66. Hur, K. et al. Unifying heterogeneous electronic health records systems via text-based code embedding. *Conference on Health, Inference, and Learning, PMLR* (2022).
67. Tang, P. C., Ralston, M., Arrigotti, M. F., Qureshi, L. & Graham, J. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. *J. Am. Med. Inform. Assoc.* **14**, 10–15 (2007).
68. Wei, W.-Q. et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* **23**, e20–e27 (2016).
69. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 1–10 (2018).
70. Lee, D., Jiang, X. & Yu, H. Harmonized representation learning on dynamic EHR graphs. *J. Biomed. Inform.* **106**, 103426 (2020).
71. Ateev, H. R. B. A. ChatGPT-assisted diagnosis: is the future suddenly here? https://www.statnews.com/2023/02/13/chatgpt-assisted-diagnosis/ (2023).
72. Raths, D. How UCSF physician execs are thinking about ChatGPT. *Healthcare Innovation* (17 February 2023).
73. Fries, J. et al. Bigbio: a framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems* 35 (2022).
74. Gao, Y. et al. A scoping review of publicly available language tasks in clinical natural language processing. *J. Am. Med. Inform. Assoc.* **29**, 1797–1806 (2022).
75. Leaman, R., Khare, R. & Lu, Z. Challenges in clinical natural language processing for automated disorder normalization. *J. Biomed. Inform.* **57**, 28–37 (2015).
76. Spasic, I. & Nenadic, G. Clinical text data in machine learning: systematic review. *JMIR Med. Inf.* **8**, e17984 (2020).
77. Yue, X., Jimenez Gutierrez, B. & Sun, H. Clinical reading comprehension: a thorough analysis of the emrQA dataset. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 4474–4486 (Association for Computational Linguistics, 2020).
78. McDermott, M. et al. A comprehensive EHR timeseries pre-training benchmark. In *Proc. Conference on Health, Inference, and Learning* 257–278 (Association for Computing Machinery, 2021).
79. Shah, N. Making machine learning models clinically useful. *JAMA* **322**, 1351 (2019).
80. Wornow, M., Gyang Ross, E., Callahan, A. & Shah, N. H. APLUS: a Python library for usefulness simulations of machine learning models in healthcare. *J. Biomed. Inform.* **139**, 104319 (2023).

81. Tamm, Y.-M., Damdinov, R. & Vasilev, A. Quality metrics in recommender systems: Do we calculate metrics consistently? *Proceedings of the 15th ACM Conference on Recommender Systems* (2021).
82. Dash, D. et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. Preprint at arXiv [cs.AI] (2023).
83. Reiter, E. A structured review of the validity of BLEU. *Comput. Linguist.* **44**, 393–401 (2018).
84. Hu, X. et al. Correlating automated and human evaluation of code documentation generation quality. *ACM Trans. Softw. Eng. Methodol.* **31**, 1–28 (2022).
85. Liu, Y. et al. G-Eval: NLG evaluation using GPT-4 with better human alignment. Preprint at arXiv [cs.CL] (2023).
86. Thomas, R. & Uminsky, D. The problem with metrics is a fundamental problem for AI. Preprint at arXiv [cs.CY] (2020).
87. Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. Preprint at arXiv [cs.CL] (2022).
88. Gao, T., Fisch, A. & Chen, D. Making pre-trained language models better few-shot learners. Preprint at arXiv [cs.CL] (2020).
89. Kaufmann, J. Foundation models are the new public cloud. *ScaleVP* https://www.scalevp.com/blog/foundation-models-are-the-new-public-cloud (2022).
90. Kashyap, S., Morse, K. E., Patel, B. & Shah, N. H. A survey of extant organizational and computational setups for deploying predictive models in health systems. *J. Am. Med. Inform. Assoc.* **28**, 2445–2450 (2021).
91. Abdullah, I. S., Loganathan, A., Lee, R. W. ChatGPT & doctors: the Medical Dream Team. *URGENT Matters* (2023).
92. Lee, P., Goldberg, C. & Kohane, I. *The AI Revolution in Medicine: GPT-4 and Beyond*. (Pearson, 2023).
93. Fleming, S. L. et al. Assessing the potential of USMLE-like exam questions generated by GPT-4. Preprint at medRxiv https://doi.org/10.1101/2023.04.25.23288588 (2023).
94. Husmann, S., Yèche, H., Rätsch, G. & Kuznetsova, R. On the importance of clinical notes in multi-modal learning for EHR data. Preprint at arXiv [cs.LG] (2022).
95. Soenksen, L. R. et al. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digit. Med.* **5**, 149 (2022).
96. Peng, S., Kalliamvakou, E., Cihon, P. & Demirer, M. The impact of AI on developer productivity: evidence from GitHub copilot. Preprint at arXiv [cs.SE] (2023).
97. Noy, S. et al. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf (2023).
98. Perry, N., Srivastava, M., Kumar, D. & Boneh, D. Do users write more insecure code with AI assistants? Preprint at arXiv [cs.CR] (2022).
99. Zhang, X., Zhou, Z., Chen, D. & Wang, Y. E. AutoDistill: an end-to-end framework to explore and distill hardware-efficient language models. Preprint at arXiv [cs.LG] (2022).
100. El-Mhamdi, E.-M. et al. SoK: on the impossible security of very large foundation models. Preprint at arXiv [cs.LG] (2022).
101. Carlini, N. et al. Quantifying memorization across neural language models. Preprint at arXiv [cs.LG] (2022).
102. Mitchell, E., Lin, C., Bosselut, A., Manning, C. D. & Finn, C. Memory-based model editing at scale. Preprint at arXiv [cs.AI] (2022).
103. Sharir, O., Peleg, B. & Shoham, Y. The cost of training NLP models: a concise overview. Preprint at arXiv [cs.CL] (2020).
104. Yaeger, K. A., Martini, M., Yaniv, G., Oermann, E. K. & Costa, A. B. United States regulatory approval of medical devices and software applications enhanced by artificial intelligence. *Health Policy Technol.* **8**, 192–197 (2019).
105. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **27**, 2020–2023 (2020).
106. Wickens, C. D., Clegg, B. A., Vieane, A. Z. & Sebok, A. L. Complacency and automation bias in the use of imperfect automation. *Hum. Factors* **57**, 728–739 (2015).

ACKNOWLEDGEMENTS

M.W. is supported by an NSF Graduate Research Fellowship. M.W., Y.X., R.T., J.F., E.S., S.F., M.A.P., and N.H.S. also acknowledge support from Stanford Medicine for this research.

AUTHOR CONTRIBUTIONS

M.W., Y.X., J.F., and N.H.S. conceptualized and designed the study; M.W., Y.X., and R.T. extracted data; M.W., Y.X., B.P., R.T., J.F., and N.H.S. conducted the analysis and wrote the manuscript. M.W., Y.X., B.P., R.T., J.F., N.H.S., and M.A.P. revised the manuscript. E.S. and S.F. contributed to the analysis. All authors approved the final version of the manuscript and take accountability for all aspects of the work.

COMPETING INTERESTS

B.P. reports stock-based compensation from Google, LLC. Otherwise, the authors declare that there are no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Michael Wornow.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.