Foresight v2 - A Large Language Model Medical Forecaster

Anonymous ACL submission

Abstract

Foresight v2 (FS2) is a Large Language Model based on LLaMa v2 7B and fine-tuned on hospital data for modelling patient timelines. It is capable of understanding a patient's clinical notes and forecasting SNOMED codes for a wide range of biomedical use cases including disorder prediction, medication recommendation, risk prediction, procedure recommendation and many more. FS2 is trained on the free text portion of the MIMIC-III dataset, firstly through the extraction of biomedical concepts and then the creation of contextualised patient timelines, upon which the model is then fine-tuned. The results show significant improvement over the previous state-of-the-art for the next new biomedical concept prediction (P/R - 0.71/0.64 vs 0.52/0.32) and a similar improvement specifically for the next new disorder forecast (P/R - 0.66/0.59 vs 0.46/0.25). Finally, on the task of disorder forecast, we compare this model, to GPT-4-turbo, and show that FS2 performs significantly better on such tasks (P@5 - 0.84 vs 0.62). This highlights the need to incorporate real health data into LLMs and shows that even much smaller models when fine-tuned on high-quality specialised data outperform much larger ones.

1 Introduction and Related Work

Language plays a central role in healthcare and medical practice, with unstructured text representing the most prevalent data in Electronic Health Records (EHRs) (Jackson et al., 2018)). Yet today, AI models have largely failed to utilize this resource and mostly ignore the free text portion of the EHR. Recently, Large Language Models (LLMs, et al. (2023b,a); Touvron et al. (2023a,b); Bai et al. (2022)) have shown the potential to understand human language, but even the most advanced LLMs (as well as specialised medical LLMs Singhal et al. (2023b)) rarely use the free text portion of the EHR. What is more, most LLMs are not trained/tested/validated on real-world hospital data, but on medical quizzes and exams.

Today's large language models have seen a remarkable evolution. Initial models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2023), GPT-1 (Yenduri et al., 2023) and GPT-2 (Radford et al., 2019) set the stage. The BERT family notably changed natural language processing (NLP), largely replacing RNN-based models in tasks such as Named Entity Recognition (NER) and text classification. Meanwhile, the GPT series, focused solely on text generation, sought to predict the next word in a sequence. Despite initial limitations, these models showed potential. This set the groundwork for the recent revolution in NLP caused by highly capable general LLMs such as ChatGPT (Ouyang et al., 2022) and LLaMA 1&2 (Touvron et al., 2023a,b). These models enabled use cases that before were either extremely difficult or completely impossible. Tasks such as document summarization, text classification, programming, and question answering were now reduced to simple prompting. Today, state-of-the-art for a wide range of NLP tasks is being set almost with every release of a new LLM.

In the medical domain, the current LLM research can be split into three groups: 1)Using LLMs on medical tasks without any fine-tuning (via prompt engineering); 2) Fine-tuning existing LLMs for the medical domain; and 3) Training LLMs from the ground up on medical data.

Recent work is mostly focused on approaches from group 1, in other words evaluating existing models for different medical tasks. Khan et al. (2024) test GPT-4 (et al., 2023b) for Anesthesiology Board-style Examination Questions, in total they collected 884 questions and prompted GPT-4 for answers. They show promising results but note that GPT-4 is still lacking in this area and more research is needed for both validation and training. Murphy Lonergan et al. (2023) show similar re-

sults for surgery, they collected 23,035 questions from MedMCQA and prompted GPT-4 for answers. They note that GPT-4 shows promising results, but still requires more training and testing. On a similar note, Savage et al. (2024) explore how to construct prompts so that the reasoning style of GPT-4 matches that of clinicians, the dataset used is a modified MedQA USMLE, the conclusion being the same as for the other examples.

There is significantly less work from group 2, i.e. LLMs fine-tuned for the medical domain. Med-PaLM 1&2 (Singhal et al., 2023a,b) are closed-source and closed-access models from Google that build on the PaLM (Chowdhery et al., 2022; Anil et al., 2023) architecture. The models are trained on QA-style datasets and show state-of-the-art results on USMLE-style questions from MultiMedQA. MediTRON 70B (Chen et al., 2023) builds on top of LLaMA-2 70B and is finetuned on medical papers and clinical guidelines, the model is primarily tested on QA style datasets where it is shown to outperform general models like GPT-3.5, and comes close to closed source medical models like Med-PaLM.

And lastly, there are only a few examples from the third group, i.e. LLMs trained from the ground up on medical data. Yang et al. (2022) train a large language model with 8.9B parameters on a dataset with >90 billion words (including >82B words of de-identified clinical text) and evaluate it on clinical NLP tasks including clinical concept extraction, medical relation extraction, semantic text similarity, natural language inference, and medical question answering. Similarly, Peng et al. (2023) train an LLM with up to 20B parameters on 82B words of clinical text and 195B words of general English text. The tests they performed were largely the same as shown in the work from Yang et al. (2022). Most other examples in this group are not what we would today consider LLMs, those include models like BioBERT (Lee et al., 2019) and ClinicalBERT (Huang et al., 2020).

Given the examples above, it is important to note that the vast majority of training/validation was performed on medical quizzes and exam questions, and not on real-world health data, highlighting the disparity between real-world use cases and LLM research in the medical domain. With some notable exceptions, like the work from Yang et al. (2022) which was trained on hospital data, but still tested mainly on public benchmarks for medical question answering, named entity recognition, and similar.

This paper builds on the recent work from Kraljevic et al. (2023), which presents Foresight v1 (FS1), a generative transformer for modelling patient timelines using derived structured concepts from unstructured text. The FS1 pipeline works as follows: 1) Collect all free text data from a hospital EHR; 2) Extract biomedical concepts (e.g., diseases, medications, procedures and symptoms) from the collected dataset; 3) Order the extracted concepts in time and group by the patient, i.e. create patient timelines; and 4) Train a generative transformer to predict the next concept in the timeline. The first weakness of FS1 is that the model does not know anything about the context in which a concept was mentioned (concepts are extracted from free text without their surrounding semantic context). The second problem is that FS1 was a pure empiricist with limited *a priori* biomedical or healthcare knowledge, in other words, the model was trained from the ground up on patient timelines consisting of only biomedical concepts.

To solve the aforementioned problems, we present Foresight v2 (FS2), a biomedical Large Language Model capable of extracting and modelling vast amounts of knowledge from EHRs. Foresight v2 is based on the LLaMAv2 7B (Touvron et al., 2023b) model and was fine-tuned on hospital data from the MIMIC-III (Johnson et al., 2016) dataset for the task of the next biomedical concept prediction in a patient timeline. The patient timelines in FS2 are contextualised, meaning a portion of the text where the concept was found is kept. FS2 is a general model capable of handling a wide range of use cases that are normally found in the free text portion of EHRs, including forecasting, medication suggestions, diagnosis and procedure suggestion - all tasks are based on the clinical notes, reflecting real word environments and not hand-made QA benchmarks.

2 Methods

Foresight v2 is a transformer-based model, built on top of a pretrained large LLM for temporal modelling of patient timelines. Formally, the task at hand can be defined as given a corpus of patients $U = \{u_1, u_2, u_3, ..\}$ where each patient is defined as a sequence of tokens $u_i = \{w_1, c_2, w_3, ...\}$ and each token is either a biomedical concept (c_n) or a text token (w_n - context where the concept was found), our objective is a modified language mod-

elling objective for supervised fine-tuning:

$$L(U) = \sum_i \sum_{j \in C^i} logP(c_j^i | k_{j-1}^i, k_{j-2}^i, ...k_0^i)$$

Where k_n^i is either a biomedical concept c_n or a free text token w_n belonging to the timeline from the patient i, and C^i is the list of all concept tokens from the timeline of patient i. In simpler terms, the model is not trained to predict text tokens, but only biomedical concept tokens given the past (concepts and text).

2.1 Data Preparation

The dataset used in this work is MIMIC-III (Johnson et al., 2016), we used all available free text from clinical notes totalling 2,083,179 documents from 46,520 patients.

We first perform entity recognition and linking on the collected free text. Extracted entities include disorders, symptoms, findings and medications (equivalent to Kraljevic et al. (2023)). Following extraction, these entities are chronologically organized into a timeline, reflecting their occurrence based on the document's creation date (The first part of Figure 1). An essential aspect of our methodology is the retention of contextual information for each extracted entity. For example, if an entity such as "hypertension" is identified, not only is the term itself preserved, but also the sentence in which it was found. This is crucial for two reasons: firstly, it allows us to capture qualifying information that could modify the understanding of the entity, such as severity (e.g., "severe hypertension"), and secondly, it enables the inclusion of negated or hypothetical concepts into the patient timeline (e.g., "no hypertension"). In instances where the boundaries of a sentence are ambiguous, we extract up to 50 tokens from each side of the entity, ensuring a comprehensive capture of context. In addition to contextual sentences, we also record the specific document ID for each concept, along with the absolute token IDs of words within the context. By tokenizing the entire document and assigning unique IDs to each token, we establish a precise reference system. This level of detail is instrumental in reconstructing the patient timeline (see the last step of Figure 1), as it allows for the accurate merging of contexts where concepts are closely related or appear within the same textual vicinity.

Once the concepts and their context are extracted, we further refine this data, employing a technique known as 'bucketing'. This process involves the aggregation of concepts within predefined time spans (we use 1 day) to eliminate repetitive mentions and reduce data noise. During bucketing we also identify potential errors; for instance, a concept mentioned only once within the whole patient EHR will be flagged as a probable NER+L mistake and removed.

After bucketing and cleaning, we add additional information to the patient timeline including age, ethnicity, sex and temporal separators. If the temporal difference between two concepts in a patient timeline is bigger than the size of the bucket (1 day in our case), we add a special token in between those concepts that tells the model how much time has passed (e.g. <1 day later>, *<7 days later>*, *<1 year later>*).

Lastly, from the concepts and their context, we reconstruct a single clinical note containing all the patient information. In this clinical note, all biomedical concepts are represented with SNOMED (Stearns et al., 2001) codes (as shown in green, bottom of Figure 1), while the context of those codes is free text. The size of this final prepared dataset is 39,591 patients in the train set and 2101 in the test set (the train/test split is 95/5).

2.2 Modelling of Patient Timelines

Foresight v2 is built on top of the LLaMA-2 7B model and fine-tuned for modeling of patient timelines. LLaMA-2 7B is a partially open-source model from Meta showing near state-of-the-art performance on a wide range of benchmarks. As it is a general large language model, it is not trained or fine-tuned for biomedical use cases and it does not have an understanding of SNOMED codes (the patient timelines consist of free text and SNOMED codes). To enable the LLaMA-2 model to handle SNOMED codes efficiently and effectively, we expand its tokenizer with the SNOMED concepts of interest (i.e. those SNOMED concepts that appear in our dataset). Usually, when adding new tokens to the tokenizer we would set the embeddings to be the average of all other tokens in the tokenizer. As the SNOMED codes are special we have slightly changed this approach. Every token we add is a SNOMED code with a unique name, so we first tokenise that name and then average the embeddings of the tokens in the name and set this as the embedding of the new token (i.e. SNOMED code). With this, every SNOMED code is represented as one token in our model.

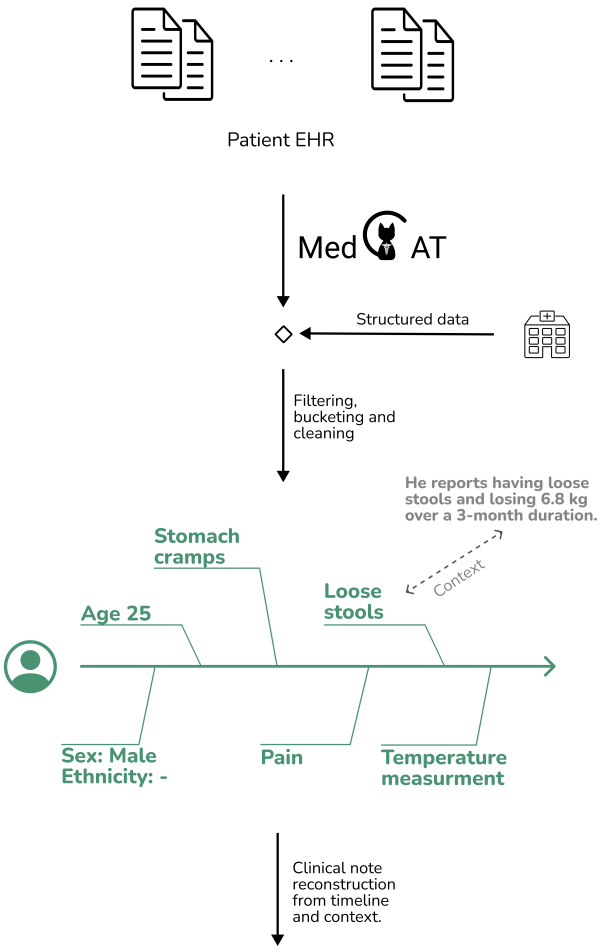

Figure 1: Data preparation workflow: 1) We collect all free text documents from the patient EHR; 2) Extract mentions of SNOMED-CT concepts and combine the concepts with static data like sex, ethnicity and age; 3) Clean, filter and bucketed concepts and turn them into a patient timeline; and lastly 4) From the concepts in the timeline, based on the context where each one was found, reconstruct a singular clinical note for each patient.

To finetune the model we used 4xA100 80GB GPUs (the training took around 1day), the hyperparameters were as follows: max_seq_len = 4096, learning_rate = 1e-5, gradient_accumulation_steps = 2, per_device_batch_size = 1, weight_decay = 0, warmup_ratio = 0.1, and the adamw_torch optimizer. To stabilise the model during training we set the adam_beta1 to 0.9 and adam_beta2 to 0.95. To speed up training and enable efficient training with long sequences we use Flash Attention 2 with PyTorch FSDP, without quantization. Importantly, we have disabled loading the model in bfloat16 with the Huggingface library, as enabling this significantly reduces the performance of the model.

All examples in the training set are packed, meaning a special token <s> is added at the beginning of each example, and they are then concatenated and split into sequences of length max_seq_len (4096 in our case). This was not done for the test set to preserve the timelines as they are. The labels are provided in a supervised fashion; only the concepts (SNOMED codes) themselves are trained on, while the labels for everything else (the free text part) are set to -100. As an example, at the bottom of Figure 1 in the reconstructed note, we would only train on the green parts of the text (i.e. SNOMED codes) while the labels for everything else would be set to -100 (i.e. no training would be performed on that part).

2.3 Metrics

The metrics used for the next concept prediction in the patient timeline are equivalent to those used in Kraljevic et al. (2023). In summary, the performance of the models is measured using custom metrics that are an extension of the standard precision (TP / TP + FP) and recall (TP / TP + FN) aiming to replicate what the model will be used for whilst also considering the limitations of the training data. There are four important parameters: 1) $T - days$ (30, 365, inf), if at timepoint T we are predicting the concept X it is considered correct if it appears anywhere in the window of length $T - days$ in the patient timeline; 2) Concept temporality, we make a distinction between concepts that never before appeared in a patient timeline (new concepts) and those that are recurring; 3) We add the notation $@N$ which denotes how many candidates we are taking from the model, if any one of the N candidates is correct then the example is considered a TP; and 4) When calculating the metrics, we filter the model output based on the type of the biomedical concept of the label, e.g. if the type of the label is *Disorder* then we filter the model output to only include disorders.

2.4 Second Stage Fine-tuning for Risk Forecasting

In addition to the contextualised patient timelines we also create timelines for disease or symptom forecasting. We do this by taking a patient timeline, splitting it in the middle (or at most after 50

concepts) and taking the first part of the timeline as is, while for the second part, we extract unique diseases that appear in the first month. So our task is, given a patient timeline (first part of it) predict the disorders that will affect the patient in the first month of the subsequent patient timeline. We take only patients that, after the timeline split, have at least one month of data in the future, and have at least 5 different disorder concepts appearing in that month (this is >90% of patients). This reduces the dataset to 13,651 in the train set and 727 in the test set.

When fine-tuning FS2 on this data, all training parameters are kept the same as for the initial training, we also run for only 1 epoch - anything above this led to overfitting to the training set. We prepare the same timelines for GPT-4-turbo, the primary difference is that all SNOMED codes are replaced with proper names (e.g. 73211009 $-$ $>$ Diabetes mellitus), find the full prompt used with GPT-4-turbo in Appendix A.

3 Results

The primary task that we test the Foresight v2 model on is the prediction of the next concepts in a patient timeline. In this task, the model showed a significant improvement over the Foresight v1 model, including a jump of 19% for the prediction of the next new concept (concept type 'All'), and 20% for the prediction of the next new disorder (Table 1). These results were achieved using the objective function shown in Section 2; if we modify this function to a standard language modelling objective (i.e. input_ids == labels) and train the model, the performance is slightly worse (overall 2-3% worse than that shown in Table 1). Also, if we remove the context from the patient timelines (i.e. leave only the biomedical concepts) the performance drops drastically (on average around 40% compared to the results in Table 1). In addition, in Table 2, we show the top 10 and bottom 10 concepts with respect to precision for prediction of new Disorders.

To showcase the capabilities of the model, we manually go through the MIMIC-III notes and find examples of different tasks that the model had to solve during the prediction of the next concept in a sequence. The results are shown in Table 3, for the input (column Patient) we only show a brief summary of the patient's condition as we are not able to show real patient data. The Prompt column shows the prompt used for FS2, it is what really was found in the clinical note for this patient. For GPT-4-turbo the prompt was slightly adjusted to be more natural and concrete, for example, we pre-pended every prompt shown in the table with an explanation that this is a medical quiz, that the model should try to answer in a way a doctor would and that it should be as precise as possible. The Ground Truth comes from the patient's EHR and it represents what really happened to the patient.

3.1 Risk Forecasting

Both GPT-4-turbo and Foresight v2 were given the task of predicting the top 5 disorders that the patient is at risk for in the next month. The dataset consisted of 100 random patients from the test set prepared for the risk prediction task (727 patients in total). As seen in Table 4, out of the 5 predictions on a dataset of 100 patients, for GPT-4-turbo in 62% of patients, at least one prediction was correct, compared to 84% in the case of Foresight v2.

To make sure the reconstructed timelines were not problematic for GPT-4-turbo, we have also taken the first 10 patients (from the 100 above) and fed the full patient history (complete clinical notes) until the timepoint T and prompted the model to predict risk in the next month (this was possible with GPT-4-turbo because the current maximum sequence length is 128K tokens). The results for these 10 patients were 10% worse in the case of full timelines compared to the reconstructed timelines.

4 Conclusion and Discussion

Foresight v2 is an LLM fine-tuned on hospital data for modelling and understanding patient timelines. It is capable of understanding clinical notes and predicting SNOMED codes for a wide range of biomedical use cases including disorder prediction, medication recommendation, symptom forecast, procedure recommendation and many more. Foresight v2 marks a significant advancement in the modelling of patient timelines over the previous state of the art (Foresight v1), enhancing the precision and effectiveness of LLMs for healthcare.

There are four primary reasons why SNOMED codes were added to the tokenizer (and the model): 1) It allows us to standardise patient timelines, remove noise and repetitions (Searle et al., 2021) and provides a way to train and benchmark LLMs on hospital data. 2) It allows us to easily rank the predictions of the model based on probability. At

			FS2 - P/R		FS1 - P/R			
Type	**T - days**	**@**	**New**	**Recurring**	**New**	**Recurring**	**Sup. R.**	**Sup. N.**
All	30	1	0.71/0.64	0.95/0.95	0.52/0.32	0.83/0.67	245265	114922
All	30	5	0.91/0.85	1.00/1.00	0.84/0.59	0.98/0.92	245265	114922
All	30	10	0.95/0.90	1.00/1.00	0.91/0.70	1.00/0.97	245265	114922
All	365	1	0.71/0.64	0.95/0.96	0.54/0.33	0.85/0.70	245265	114922
All	inf	1	0.71/0.64	0.95/0.96	0.55/0.33	0.86/0.70	245265	114922
Disorders	30	1	0.66/0.59	0.94/0.94	0.46/0.25	0.79/0.60	109019	51675
Disorders	30	5	0.88/0.81	1.00/1.00	0.79/0.51	0.98/0.89	109019	51675
Disorders	30	10	0.94/0.87	1.00/1.00	0.88/0.62	0.99/0.96	109019	51675
Disorders	365	1	0.67/0.59	0.95/0.95	0.49/0.26	0.83/0.64	109019	51675
Disorders	inf	1	0.67/0.59	0.95/0.95	0.50/0.26	0.84/0.65	109019	51675
Findings	30	1	0.74/0.67	0.95/0.96	0.52/0.29	0.83/0.66	71007	33772
Findings	30	5	0.94/0.88	1.00/1.00	0.85/0.58	0.99/0.93	71007	33772
Findings	30	10	0.97/0.93	1.00/1.00	0.92/0.70	1.00/0.98	71007	33772
Findings	365	1	0.75/0.67	0.95/0.96	0.54/0.29	0.85/0.67	71007	33772
Findings	inf	1	0.75/0.67	0.95/0.96	0.55/0.29	0.85/0.68	71007	33772
Substances	30	1	0.63/0.53	0.95/0.94	0.52/0.32	0.84/0.70	39578	19172
Substances	30	5	0.88/0.79	1.00/1.00	0.85/0.61	0.99/0.94	39578	19172
Substances	30	10	0.94/0.87	1.00/1.00	0.92/0.73	1.00/0.99	39578	19172
Substances	365	1	0.63/0.53	0.95/0.95	0.53/0.32	0.84/0.71	39578	19172
Substances	inf	1	0.63/0.53	0.95/0.95	0.53/0.32	0.85/0.71	39578	19172
Procedures	30	1	0.92/0.90	0.98/0.99	0.79/0.67	0.94/0.92	7831	3379
Procedures	30	5	0.99/0.99	1.00/1.00	0.97/0.94	1.00/1.00	7831	3379
Procedures	30	10	1.00/1.00	1.00/1.00	0.99/0.99	1.00/1.00	7831	3379
Procedures	365	1	0.93/0.90	0.98/0.99	0.81/0.67	0.95/0.93	7831	3379
Procedures	inf	1	0.93/0.90	0.98/0.99	0.81/0.67	0.95/0.94	7831	3379

Table 1: Results for the next concept prediction task. Sup N and Sup R is the support for recurring and new concepts, FS2 = Foresight v2 model, FS1 = Foresight v1 model, P = Precision, R = Recall. $T - days$ is the size of the temporal window in days.

every point where we want to predict the next concept in a timeline, we can easily see what is the most probable, or what are the top N predictions. 3) It makes sure the model predictions are part of a standardised widely accepted medical ontology, as opposed to having a model generate free text and then needing another step to map backwards into standardised forms for compatibility with existing healthcare informatics systems. This compatibility with biomedical ontologies is important as general-purpose LLMs are prone to hallucinate realistic-looking standardised output (e.g. academic citations, Zhou et al. (2024)) which is addressable with greater exposure to standardised ontologies (Wang et al., 2024) in the way shown in this manuscript. 4) The model predictions are inherently privacy-preserving as the model was not directly trained on text, it can only output healthcare concepts within intentionally constrained healthcare vocabulary (SNOMED), it cannot predict any personally identifiable information, like names, addresses or other HIPAA-defined protected health information. Without this guarantee, no model trained on hospital data should be made publicly available.

During benchmarking between GPT4 and FS2, some leniency was provided to GPT4 by allowing for some poecilonymic predictions (e.g. 'gastrointestinal haemorrhage' being predicted when the ground truth was 'gastric haemorrhage'). At the same time, Foresight v2 was only scored on exact predictions. This means the real-world performance of FS2 may be underestimated as poecilonymic predictions may be sufficient for real-world utility (especially for human-in-the-loop implementations). It is important to note the difficulty of the task, which in turn can explain why models like GPT-4-turbo are performing signif-

Disorder	P	TP	FP
Stress ulcer	1.00	175	0
Postcholecystectomy s.	1.00	32	0
Left atrial dilatation	1.00	35	0
Muscle atrophy	1.00	22	0
Rubella	1.00	19	0
Conjunctival edema	1.00	16	0
Mediastinal shift	1.00	11	0
Diastolic hypertension	1.00	12	0
Mitral valve regurgitation	0.98	687	12
Systolic hypertension	0.98	338	6
Hypercholesterolemia	0.31	46	105
Left bundle branch block	0.30	12	28
Kidney stone	0.30	16	38
Gastroesophageal reflux d.	0.30	18	43
Gastrointestinal hemorrhage	0.30	18	43
Hyperlipidemia	0.29	62	155
Right bundle branch block	0.28	23	60
Hypothyroidism	0.27	41	109
Asthma	0.23	18	61
Benign prostatic hyperplasia	0.19	29	123

Table 2: Top and Bottom 10 concepts with respect to precision for prediction of new disorders.

icantly worse. Real-world EHR data is messy, noisy, extremely complex and filled with duplicated text. Within this noisy data, predicting the next event can prove to be a very difficult task with the added factors of patient complexity, multimorbidity, polypharmacy and acute clinical instability of patients. Complications can develop as a result of their severe underlying disease or as an iatrogenic event secondary to procedures and medications. Of note, the median age of patients in MIMIC-III was 66 years old, with a mortality of 23.2% and a median hospital stay of 2.1 days (Q1-Q3: 1.2–4.1) (Dai et al., 2020). Predicting the next concept in such a highly unstable cohort of patients over such a short time span is exceptionally difficult.

4.1 Limitations and Risk

There are limitations to ontological classification systems such as SNOMED or ICD-10 - these systems may not cover all details and nuances within the clinical text. For example, there will be diseases or concepts that don't fall within the defined boundaries of available terminology or do not yet exist as formal concepts in codified terminologies (highly prevalent in fields with rapid scientific progress, e.g. cancer genetics and precision medicine). This challenge is to the most extent resolved because FS2 is capable of understanding free text next to SNOMED concepts. Exploring this area in detail is left for future work.

As this model is trained on a relatively small dataset without any human preference alignment, the prompts are more similar to GPT-3 rather than recent LLMs (e.g. GPT-4). The prompts have to reflect the way the clinical notes are written, and the model cannot answer general questions or hold conversations. For example, in the notes, we will often have the phrase "The patient was discharged with: " the model will know that after this it has to predict discharge medications. Q&A-style prompting popularised by ChatGPT like "What are the discharge medications for this patient?" would not work without further human preference alignment.

It is also important to note that while the results obtained are very good, these models are still in the early stages of research and testing, and are not yet suitable to be Software as a Medical Device (SaMD). There is a temptation to imagine the predictions to be used for clinical care or decision support - this is still premature as Foresight v2 is derived from historical practice so would not always be expected to be consistent with contemporary best practice.

Lastly, significantly larger hospital datasets as well as general medical literature are needed to better cover all possible biomedical concepts found in SNOMED, as well as prevent biases or inaccuracies that can stem from using a single hospital as the training dataset. Future work should explore expanding the training data with medical guidelines, textbooks, and definitions and if possible include multiple hospitals.

4.2 Potential Utility

We note alerting systems as a use-case for which models like Foresight v2 could be well-suited. Table 2 shows that there is a wide range of conditions with a precision of 100% and such conditions are particularly well suited in the context of designing alert systems. This high precision ensures that when an alert is issued, it is almost invariably relevant. Importantly, this high-precision approach minimises the clinician 'alert fatigue', a scenario that might arise if high recall was favoured over high precision.

Another utility of FS2 is for risk prediction and prognosis, this can be used to guide primary or

Patient	Prompt	Foresight v2	GPT-4-turbo	Ground Truth
Middle-aged male patient with swelling and fracture of ankle.	Rule out:	DVT	DVT	DVT
Older male patient with obesity and sleep apnoea.	Recently increased somnolence and dyspnoea, likely a sign of	Hypercapnia	Acute Respiratory Distress Syndrome	Hypercapnia (later confirmed to really be Hypercapnia)
Older female patient with a complex mental health history.	Given the parapsychotic nature of the depression, started on	Risperidone	Aripiprazole or Lurasidone	Risperidone
A young female patient with a long medical history and current visit for gastrointestinal issues.	The patient was discharged with scripts for:	Omeprazole (One of the top 3 predictions)	Proton Pump Inhibitors (PPIs) or H2 Blockers (one of top 3 predictions)	Omeprazole
Infant with hypertension	<list of problems>* evaluate with	Echo	Echo	Echo

Table 3: Examples of tasks found in the MIMIC-III dataset, and the predictions by Foresight v2 and GPT-4-turbo. The *Patient* column represents a very brief summary of the patient's past for privacy reasons, during the tests models were fed the real patient timelines. The prompts are original pieces of text taken from the patient's timeline. The Ground Truth is taken from the clincal notes for the patient. *We redacted the full list of problems to avoid re-identification risk, in the prompts used with GPT-4-turbo and Foresight v2 the list was kept as found in the clinical note.

Model	At least 1	At least 2	At least 3
GPT-4-turbo	62%	23%	6%
Foresight v2	84%	56%	28%

Table 4: Risk prediction results for GPT-4-turbo and Foresight v2, both models were prompted to predict the top 5 disorders a patient is at risk for in the next month. The column 'At least N' shows in the dataset of 100 patients, the percentage of patients where at least N out of the 5 predictions are correct.

secondary disease prevention or determine management course. In medicine, there are countless validated risk and prognostic scores designed for disease-specific scenarios; e.g. QRISK (Hippisley-Cox et al., 2017) for stratification of cardiovascular disease, CHADSVASC (Lip et al., 2010) score for stroke risk, CURB65 (Lim, 2003) for pneumonia severity; these require large-scale calibration for generalisability and ongoing feature-engineering for more variables. Our approach with FS2 is more fine-grained and high-dimensional as it models temporally ordered sequences of comorbidities, and additional features (e.g. medications, social determinants of health, complications and outcomes) are included with limited *a priori* assumptions.

Lastly, various use cases in medical education, clinical co-pilots (for medications, procedures, disorders, etc.), synthetic data generation and reconstruction of patient timelines are all possible with models like FS2. In effect, models such as FS2 that are trained on whole hospitals and medical literature (*a priori* for now in FS2) will become a model of an entire healthcare system and the corresponding population.

The code for Foresight v2 will be open-sourced upon paper acceptance (because of the anonymity period), we will try to publish the models also, but this depends on MIMIC-III and the rules and regulations they will apply to such models.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Zheng Dai, Siru Liu, Jinfa Wu, Mengdie Li, Jialin Liu, and Ke Li. 2020. Analysis of adult disease characteristics and mortality on MIMIC-III. *PLoS One*, 15(4):e0232176.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team et al. 2023a. Gemini: A family of highly capable multimodal models.

OpenAI et al. 2023b. Gpt-4 technical report.

Julia Hippisley-Cox, Carol Coupland, and Peter Brindle. 2017. Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, 357.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. Clinicalbert: Modeling clinical notes and predicting hospital readmission.

Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, Damian Lewsley, Doug Northwood, Amos Folarin, Robert Stewart, and Richard Dobson. 2018. Cogstack - experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. *BMC Medical Informatics and Decision Making*, 18(1):47.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Adnan A. Khan, Rayaan Yunus, Mahad Sohail, Taha A. Rehman, Shirin Saeed, Yifan Bu, Cullen D. Jackson, Aidan Sharkey, Feroze Mahmood, and Robina Matyal. 2024. Artificial intelligence for anesthesiology board-style examination questions: Role of large language models. *Journal of Cardiothoracic and Vascular Anesthesia*.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfie Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J Dobson. 2023. Foresight – generative pretrained transformer (gpt) for modelling of patient timelines using ehrs.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

W S Lim. 2003. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*, 58(5):377–382.

Gregory Y H Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry J G M Crijns. 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Rebecca Murphy Lonergan, Jake Curry, Kallpana Dhas, and Benno I Simmons. 2023. Stratified evaluation of GPT's question answering in surgery reveals artificial intelligence (AI) knowledge gaps. *Cureus*, 15(11):e48788.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. 2023. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7(20).

Thomas Searle, Zina Ibrahim, James Teo, and Richard Dobson. 2021. Estimating redundancy in clinical text. *Journal of Biomedical Informatics*, 124:103938.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023b. Towards expert-level medical question answering with large language models.

Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. Snomed clinical terms: overview of the development process and project status. *Proceedings. AMIA Symposium*, pages 662–6.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay

Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. 2024. Drg-llama : tuning llama model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, 7(16).

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1).

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. 2023. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Principles, applications, and challenges.

A The prompt used for GPT-4-turbo

```
<system prompts, these are appended
as system messages to gpt-4-turbo>

You are now playing the role of a
medical doctor taking an exam,
your goal is to be as accurate
as possible and make sure you do
not make any mistakes. If you
are unsure about something, think
step by step and then answer.
You have to follow the instructions
precisely.

Your first question in this medical
 quiz will consist of a patient history,
your goal is to predict 5 specific disorders
the patient is at risk for in the next
month. Please take care that the disorders
you are predicting cannot be part of
the patient's past. They
have to be new disorders that will most
likely affect the patient in the next month.
You have to predict specific disorders,
for example: you should never say
"pulmonary problems"
as this is not a specific disorder,
but you can say "pneumonia" as that
is a specific disorder. Your output
should be in .json format and consists
of a list of disorder names and
explanations
(e.g. [('<disorder_1>', '<explanation>'),
...])
</system prompts>

{history}
Given the above patient history. What
specific new disorders is this
patient at risk for in the next month?
```