The Foundational Capabilities of Large Language Models in Predicting Postoperative Risks Using Clinical Notes

Charles Alba[1,2,3,4†], Bing Xue[1,2†], Joanna Abraham[1,5,6], Thomas Kannampallil[1,2,5,6], Chenyang Lu[1,2,5*]

[1]AI for Health Institute, Washington University in St. Louis, 1 Brookings Drive, St Louis, 63130, MO, USA.
[2]McKelvey School of Engineering, Washington University in St Louis, 1 Brookings Drive, St Louis, 63130, MO, USA.
[3]Brown School, Washington University in St Louis, 1 Brookings Drive, St Louis, 63130, MO, USA.
[4]National University of Singapore, 21 Lower Kent Ridge Rd, Singapore, 119077.
[5]School of Medicine, Washington University in St Louis, 660 S Euclid Ave, St. Louis, 63110, MO, USA.
[6]Institute for Informatics, Data Science, and Biostatistics, Washington University in St Louis, 660 S Euclid Ave, St. Louis, 63110, MO, USA.

*Corresponding author(s). E-mail(s): lu@wustl.edu;
Contributing authors: alba@wustl.edu; bingxue@wustl.edu;
joannaa@wustl.edu; thomas.k@wustl.edu;
†These authors contributed equally to this work.

Abstract

Clinical notes recorded during a patient's perioperative journey holds immense informational value. Advances in large language models (LLMs) offer opportunities for bridging this gap. Using 84,875 pre-operative notes and its associated surgical cases from 2018 to 2021, we examine the performance of LLMs in predicting six postoperative risks using various fine-tuning strategies. Pretrained LLMs outperformed traditional word embeddings by an absolute AUROC of 38.3% and AUPRC of 33.2%. Self-supervised fine-tuning further improved performance by 3.2% and 1.5%. Incorporating labels into training further increased AUROC by

1.8% and AUPRC by 2%. The highest performance was achieved with a unified foundation model, with improvements of 3.6% for AUROC and 2.6% for AUPRC compared to self-supervision, highlighting the foundational capabilities of LLMs in predicting postoperative risks, which could be potentially beneficial when deployed for perioperative care

Keywords: Large Language Models, Perioperative Care, Postoperative complications, Clinical notes

1 Introduction

More than 10% of surgical patients experience major postoperative complications, such as pneumonia and blood clots ranging from pulmonary embolism (PE) to deep vein thrombosis (DVT) [1–5]. These complications often lead to increased mortality, intensive care unit admissions, extended hospital stays, and higher healthcare costs [6]. Many of these preventable complications could be avoided through early identification of patient risk factors [7, 8]. Recent reports show that clear perioperative pathways, defined as non-surgical procedures before and after surgery like shared decision-making, pre-operative assessments, enhanced surgical preparation and discharge planning, can reduce hospital stays by an average of two days across various surgeries, with subsequently carefully designed interventions reducing complications by 30% to 80%. This far exceeds the effects of drug or treatment interventions [7]. This underscores the critical role of identifying patient risk factors early on and implementing effective preventive measures to improve patient outcomes.

Most machine learning models aimed at predicting postoperative risks primarily utilize numerical and categorical variables, or time-series measurements [1, 5]. These models typically include features such as demographics, history of comorbidities, lab tests, medications, and statistical features extracted from time series [1, 5], along with features reflecting surgical settings like scheduled surgery duration, surgeon name, anesthesiologist name, and location of the operating room [5], as well as factors such as drug dosing, blood loss, and vital signs [9].

Text-based clinical notes taken during the surgical care journey hold immense informational value, with the potential to predict postoperative risks. Unlike discrete electronic health records (EHRs) like tabular or time-series data, clinical notes represents a form of clinical narratives, thereby allowing clinicians to communicate personalized accounts of a patients history's that could not otherwise be conveyed through traditional tabular data [10]. Informational value from clinical notes could therefore aid the decision-making processes that impact the course of a patient's perioperative journey and beyond [11]. This includes the preparation for surgery, the transfer of patients to operating rooms, and the prioritization of clinicians' tasks [11, 12], emphasizing their importance in achieving safe patient outcomes [12, 13].

Following the emergence of ChatGPT, a large body of research on LLMs in medicine has been primarily centered on the application and development of conversational chatbots to support clinicians, provide patient care, and enhance clinical

decision support [14]. These efforts include automating medical reports [15], creating personalized treatments [16], and generating summaries [17] or recommendations [14] from clinical notes, respectively. However, there has been comparatively less exploration of specialized fine-tuned models in directly analyzing clinical notes to make robust predictions regarding surgical outcomes and complications. Most work has primarily leveraged publicly available resources, such as open-sourced EHR databases like MIMIC to predict event-based outcomes such as ICU readmission [18] or benchmarks like i2b2 to perform information extraction tasks, like identifying concepts, diagnoses, and medications mentioned in clinical notes [19]. Some models have been developed for specialized medical diagnostics, such as COVID-19 prediction [20] and Alzheimer's disease (AD) progression prediction [21]. However, the application of LLMs in predicting specific postoperative surgical complications has remained limited.

Therefore, our work aims to develop models for the prediction of specific postoperative surgical complications through pre-operative clinical notes, with the goal of enabling the early identification of patient risk factors. By mitigating adverse surgical outcomes, we hope to leverage LLMs to enhance patient safety and improve patient outcomes. Henceforth, our contributions are multifold: First, we explore the predictive performance that pre-trained language models offer in comparison to traditional word embeddings. Specifically, by comparing the predictive performance of postoperative risk predictions between clinically-oriented pre-trained LLMs and traditional word embeddings, we demonstrate the potential of pre-trained transformer-based models in predicting postoperative complications from pre-operative notes, relative to traditional NLP methods. Second, we explore a series of finetuning techniques to enhance the performance of LLMs for predicting postoperative outcomes, as illustrated in Figure 1. This is critical for understanding the potential of pre-trained models in perioperative care and determining optimal strategies to maximize their predictive capabilities in the early identification of postoperative risks. Third, we demonstrate the foundational capabilities of LLMs in the early identification of postoperative complications from pre-operative care notes. This suggests that a unified model could be applied directly to predict a wide range of risks, rather than training separate models for each specific clinical use case. The versatility of such foundation models could offer immense tangible benefits when deployed in real-life clinical settings.

2 Results

A total of 84,875 clinical notes containing descriptions of scheduled procedures derived from electronic anesthesia records (Epic) of all adult patients (mean [SD] age, 56.9 [16.8] years; 50.3% male; 74% White) at Barnes Jewish Hospital (BJH) from 2018 to 2021 were included in this study. The six outcomes were death within 30 days (1,694 positive cases; positive event rate of 2%), Deep Vein Thrombosis (DVT) (498 positive cases; positive event rate of 0.6%), Pulmonary Embolism (PE) (287 positive cases; positive event rate of 0.3%), pneumonia (475 positive cases; positive event rate of 0.6%), Acute Kidney Injury (AKI) (11,418 positive cases; positive event rate of 13%), and delirium (5,695 positive cases; positive event rate of 47%).

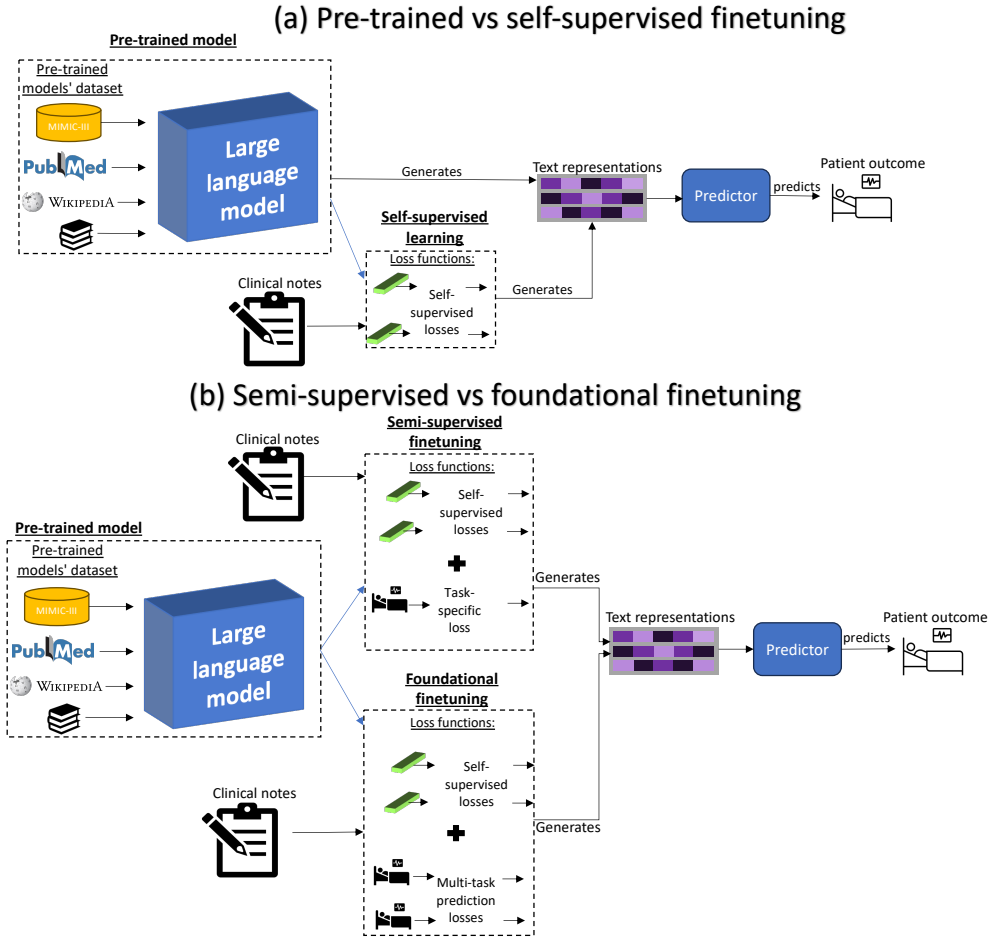

Fig. 1 An illustration of the architectures encompassing different fine-tuning strategies experimented in our study, encompassing the results reported from Sections 2.1 to 2.4. Fig 1a (top) illustrates how using the pretrained model alone differs from self-supervised fine-tuning when clinical texts are provided to the pretrained LLM to refine the model weights with respect to its objective loss functions. Fig 1b (below) illustrates two separate fine-tuning strategies: semi-supervised fine-tuning – creating a model that is fine-tuned under the supervision of a specific outcome; and foundation fine-tuning – creating a foundation model that is fine-tuned through a multi-task learning (MTL) objective using all available postoperative labels in the dataset.

2.1 The state of pretrained LLMs in perioperative care

When comparing the predictive performances of clinically-oriented pretrained LLMs, such as bioGPT [22], ClinicalBERT [18], and bioClinicalBERT [19], with traditional word embeddings, including word2vec's continuous bag-of-words (CBOW) [23],

Model	Death in 30 days		PE		Pneumonia		DVT		AKI		Delirium	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
cbow	0.528 (0.409, 0.648)	0.023 (0.015, 0.031)	0.506 (0.418, 0.593)	0.004 (0.002, 0.006)	0.526 (0.384, 0.668)	0.009 (0.001, 0.016)	0.524 (0.457, 0.59)	0.006 (0.005, 0.007)	0.56 (0.488, 0.632)	0.156 (0.125, 0.187)	0.501 (0.464, 0.539)	0.474 (0.441, 0.507)
Doc2vec	0.479 (0.348, 0.611)	0.021 (0.012, 0.03)	0.517 (0.466, 0.567)	0.004 (0.004, 0.004)	0.495 (0.347, 0.643)	0.006 (0.003, 0.01)	0.531 (0.443, 0.619)	0.007 (0.004, 0.01)	0.523 (0.421, 0.624)	0.146 (0.092, 0.199)	0.484 (0.439, 0.53)	0.466 (0.417, 0.514)
fastText	0.725 (0.67, 0.781)	0.05 (0.04, 0.06)	0.652 (0.602, 0.701)	0.007 (0.005, 0.01)	0.696 (0.643, 0.749)	0.016 (0.008, 0.024)	0.694 (0.642, 0.746)	0.014 (0.011, 0.017)	0.726 (0.702, 0.75)	0.273 (0.239, 0.307)	0.565 (0.541, 0.589)	0.533 (0.513, 0.554)
GloVe	0.818 (0.807, 0.83)	0.128 (0.118, 0.139)	0.664 (0.628, 0.701)	0.01 (0.007, 0.013)	0.765 (0.732, 0.799)	0.04 (0.017, 0.063)	0.723 (0.7, 0.745)	0.019 (0.013, 0.024)	0.81 (0.805, 0.815)	0.441 (0.43, 0.451)	0.666 (0.652, 0.681)	0.636 (0.613, 0.66)
bioClinicalBERT	0.85 (0.84, 0.861)	0.156 (0.138, 0.173)	0.683 (0.621, 0.745)	0.008 (0.006, 0.011)	0.809 (0.785, 0.833)	0.043 (0.027, 0.059)	0.76 (0.723, 0.796)	0.02 (0.014, 0.027)	0.83 (0.828, 0.831)	0.469 (0.457, 0.48)	0.68 (0.663, 0.697)	0.653 (0.626, 0.68)
bioGPT	**0.862 (0.851, 0.872)**	**0.161 (0.141, 0.182)**	0.711 (0.679, 0.743)	0.011 (0.005, 0.017)	0.818 (0.8, 0.837)	**0.047 (0.037, 0.058)**	**0.773 (0.734, 0.813)**	**0.024 (0.016, 0.032)**	**0.835 (0.833, 0.838)**	**0.478 (0.465, 0.492)**	**0.691 (0.672, 0.71)**	**0.664 (0.638, 0.69)**
ClinicalBERT	0.855 (0.842, 0.867)	0.155 (0.137, 0.173)	**0.717 (0.691, 0.743)**	**0.013 (0.009, 0.017)**	0.806 (0.784, 0.827)	0.04 (0.024, 0.056)	0.764 (0.73, 0.799)	0.022 (0.015, 0.03)	0.83 (0.827, 0.833)	0.469 (0.458, 0.48)	0.686 (0.671, 0.702)	0.66 (0.634, 0.686)

Table 1 . A comparison of traditional NLP models (top) vs pretrained LLMs (bottom). The results are presented as the mean and 95% confidence interval across all 5-folds. The best baseline models are underlined, and the best models are **bolded**. As shown amongst the results, the pretrained LLMs consistently outperform traditional word embeddings.

doc2vec [24], GloVe [25], and FastText [26], we observed absolute increases that ranged from up to 20.7% in delirium to 38.3% in death within 30 days for the Area Under the Receiver Operating Characteristic curve (AUROC). Similarly, increases in the Area Under the Precision-Recall Curve (AUPRC) ranged from 0.9% in PE to an impressive 33.2%.

This massive leap in performance highlights the sheer power of pretrained LLMs in grasping clinically relevant context in comparison to traditional word embeddings, despite being trained on broad biomedical and/or clinical notes that are not specific to perioperative care. These improvements, experimented akin to a 'zero-shot setting' [27], are noteworthy given their minimal exposure to perioperative care notes.

2.2 Transfer learning: Improvements from adapting pretrained models to perioperative corpora

Exposing each of the selected clinically oriented pretrained models – bioGPT, ClinicalBERT, and bioClinicalBERT – to clinical texts specific to perioperative care through self-supervised fine-tuning, as best illustrated in Figure 1, resulted in absolute improvements ranging from up to 0.6% in AKI to 3.2% in PE for AUROC compared to using the pretrained models demonstrated in Section 2.1. Similarly, increases ranged from up to 0.3% in PE to 1.5% in 30-day mortality for AUPRC. The results spanning the three base models – bioClinicalBERT, bioGPT and ClinicalBERT – across all outcomes are best illustrated in Figure 2. These consistent improvements witnessed upon adapting these pretrained models to perioperative specific corpora highlight the gap between the corpora on which the pretrained models were originally trained and that of perioperative care. This suggests that there remains leeway for improvement in adapting the models to the textual characteristics specifically observed in perioperative care notes.

2.3 Transfer learning: Improvements from further incorporating labels

When incorporating labels as part of the fine-tuning process through a semi-supervised approach, we witnessed further improvements ranging from up to 0.4% in delirium to 1.8% in PE, and AUPRC values ranging from 0% in Delirium to 2% in death within 30 days, relative to the approach in section 2.2. The results spanning the three base models across all outcomes is best illustrated in Figure 2. This demonstrates that predictive performance is boosted relative to the self-supervised approach since both textual and labelled data were utilized during training.

2.4 The foundational capabilities of LLMs in predicting post-operative complications from pre-operative notes

Foundation fine-tuning involves incorporating a multi-task learning framework to simultaneously fine-tune all labels, resulting in a single robust model capable of predicting all six possible postoperative risks. Through our foundation model, we observe increases in AUROC ranging from up to 0.8% in AKI to 3.6% in PE and AUPRC values increases ranging from up to 0.4% in Pneumonia to 2.6% in death within 30

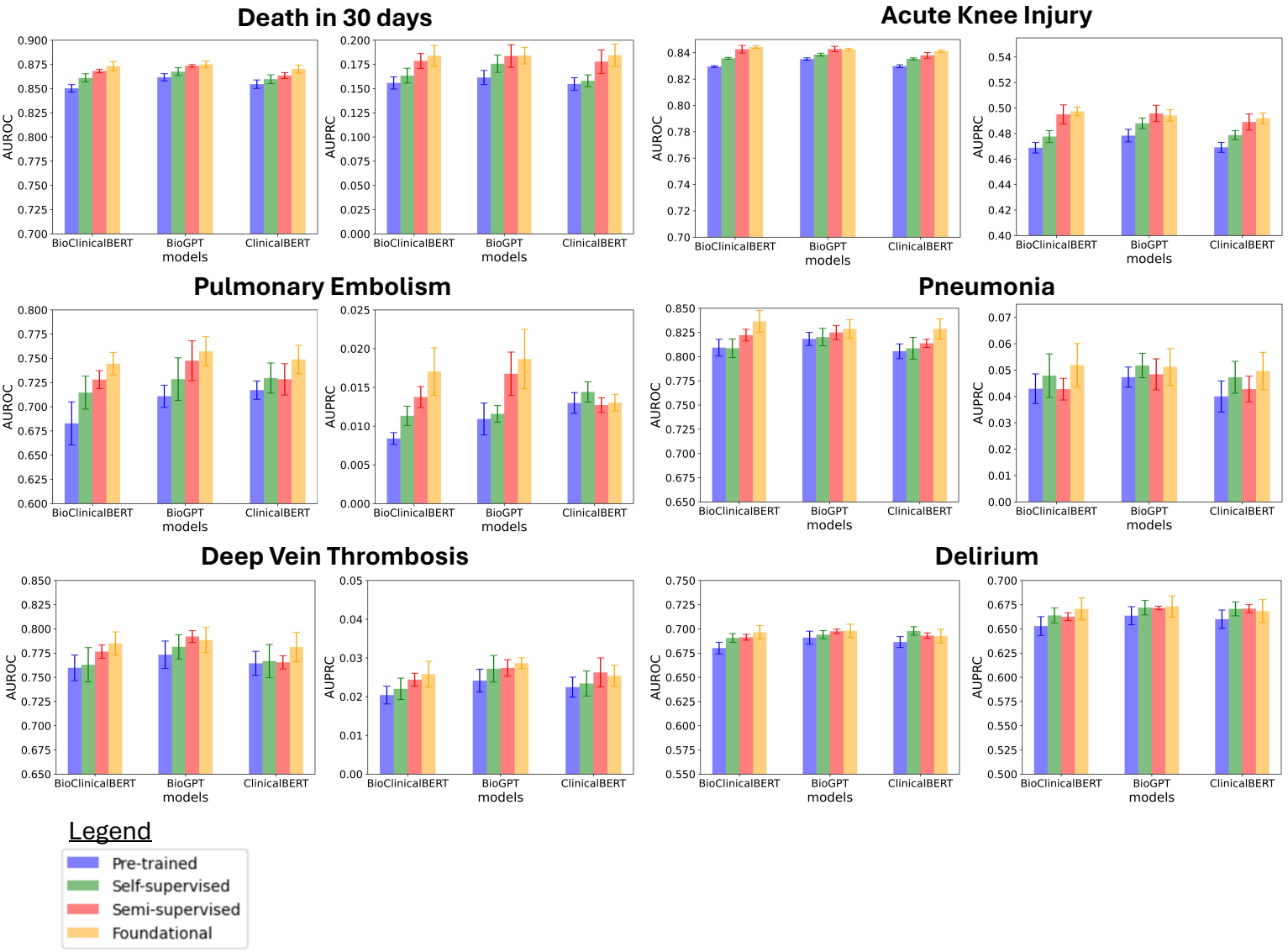

Fig. 2 Comparison of the predictive performance across various models and their respective tuning strategies. The bar graph illustrates the means, with the error bar representing the respective standard errors, across a 5-fold cross-validation. Fine-tuning with the models self-supervised pretraining objectives improved prediction performance relative to the pretrained models alone, with the incorporation of labels further boosting prediction performances. The model performs best with the foundation fine-tuning strategy, wherein the model was fine-tuned with a multi-task learning objective across all outcomes. Precise numerical metrics are reported in the supplementary material.

days, when contrasted with the self-supervised approach from section 2.2. The results spanning the three base models across all outcomes is best illustrated in Figure 2. As this is our best-performing finetuning strategy, we reported additional metrics, such as the accuracy, sensitivity, specificity, precision, and F-scores, in the supplementary material.

The attainment of competitive performances are a testament towards the ability of LLMs to establish robust representations across varied clinical tasks, therefore showcasing the foundational capabilities of LLMs in perioperative care.

2.5 Examining the effects of ML predictors on predictive performance

To facilitate consistent comparisons between the traditional word embeddings and the distinct fine-tuning strategies from pretrained LLMs, we defaulted to using eXtreme Gradient Boosting Tree (XGBoost) [28] as the predictor used for the text embeddings. However, in practice, the choice of predictors remains flexible.

As such, we compare the performances of applying distinct ML predictors among the best-performing fine-tuning strategy – foundation models. These include (1) the default XGBoost, (2) logistic regression, (3) random forest, and (4) task-specific fully connected neural networks found on top of the contextual embeddings. The results are mixed, with no single classifier dominantly outperforming the others. Surprisingly, logistic regression performed slightly better than the others, demonstrating that well-tuned models can generate precise contextual representations that suit a simple classifier with better generalizability than complex classifiers.

2.6 Qualitative evaluation of our foundation models for safety and adaption to perioperative care

We ran several broad case prompts to examine and ensure the safety [17, 29] of our foundation models. This also provides us with the opportunity to qualitatively ensure its adaptation to perioperative care beyond the reported quantitative results.

Table 2 demonstrates that our foundation models have indeed adapted from the generic texts in biomedical literature or non-perioperative clinical notes to the terminologies expected among perioperative care notes. More importantly, our model did not produce any alarmingly harmful outputs, contrary to what we witnessed in the outputs among some of the pretrained models of which our foundation models were based on.

2.7 BERT vs GPT / biomedical vs clinical-domain based models – Which performs best in predicting postoperative risks?

From the pretrained models alone (i.e., without any fine-tuning), BioGPT excelled in 5 of the 6 tasks, while ClinicalBERT excelled in 1 of the 6 tasks. This demonstrates that the text and terminologies in biomedical literature sufficiently capture the text

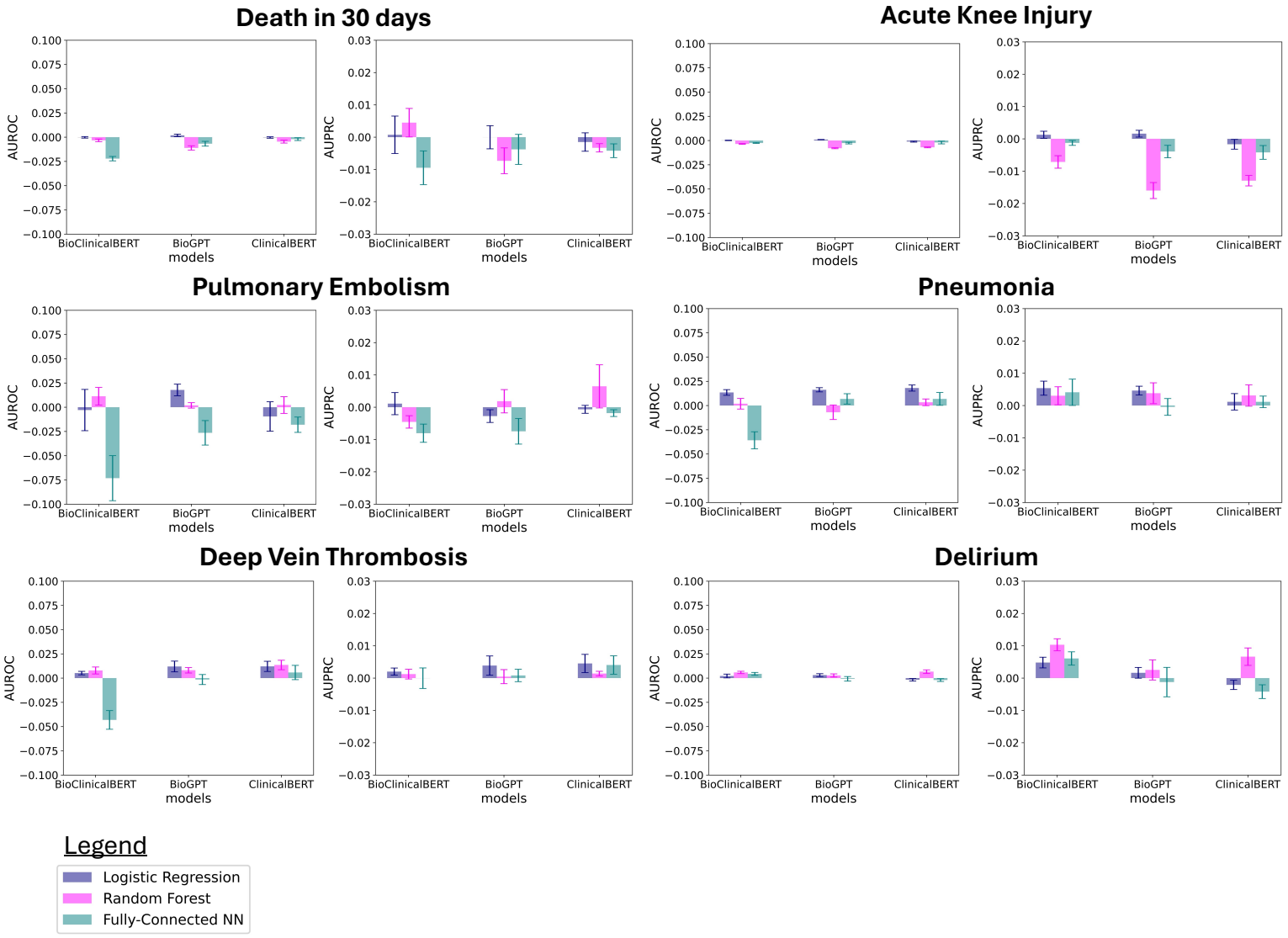

Fig. 3 Comparison of different machine learning classifiers with that of our default XGBoost predictor applied to our textual representations ($\Delta\text{model}_{i,j} - \text{XGBoost}_i$), including the use of the feed-forward auxiliary layer directly from our foundation model. No single classifier dominated the others across all outcomes and metrics. Surprisingly, the logistic regression classifier performed slightly better than the others, demonstrating that well-tuned language models can generate precise contextual representations to suit a simple classifier. Precise numerical metrics are reported in the supplementary material.

(a) BioClinicalBERT variant

Prompt	bioClinicalBERT	Our model
"[MASK] underwent surgery to remove tumor."	**She** underwent surgery to remove tumor.	**Patient** underwent surgery to remove tumor.
"Patient complains about pain. Patient is [MASK]."	Patient complains about pain. Patient is **comfortable**.	Patient complains about pain. Patient is **anxious**.
"Patient suffers from excessive [MASK]. Patient should be sent to ICU."	Patient suffers from excessive **anxiety**. Patient should be sent to ICU.	Patient suffers from excessive **burn**. Patient should be sent to ICU.
"Patient's family has history of diabetes. Surgeon should watch out for [MASK]."	Patient's family has history of diabetes. Surgeon should watch out for **diabetes**.	Patient's family has history of diabetes. Surgeon should watch out for **procedures**.

(b) BioGPT variant

Prompt	bioGPT	Our model
Patient suffers from excessive bleeding. Patient should be	Patient suffers from excessive bleeding. Patient should be **sent to a hospital for further investigation.**	Patient suffers from excessive bleeding. Patient should be **sent to a trauma room or a vascular access room**
Patient complains of extreme pain. Patient is probably a	Patient complains of extreme pain. Patient is probably a **good candidate for a trial of conservative treatment**	Patient complains of extreme pain. Patient is probably a case of **lumbar**
Patient's family has history of high blood pressure. Avoid	Patient's family has history of high blood pressure. Avoid **use of antihypertensive drugs**.	Patient's family has history of high blood pressure. Avoid **invasive diagnostic procedure if possible**

Table 2 A qualitative safety evaluation towards the open-sourced (a) bioClinicalBERT and (b) bioGPT variant of our foundation models. The prompts demonstrate the safety and adaptive nature of our model in perioperative care use. Results deemed potentially harmful are colored in red.

and terminologies of a similar magnitude compared to their counterparts trained on critical care unit notes.

However, upon fine-tuning the models with the clinical notes and the six labels, the foundation BioGPT model excelled in 4 of the 6 tasks, while BioClinicalBERT excelled in 2 of the 6 tasks. This can be attributed to the fact that the training texts from the biomedical corpora from the base model, was complemented by terminologies and texts infused from our perioperative care notes during fine-tuning, thereby creating robust, diverse, and rich representations compared to notes solely exposed to clinical corpora as in the case of the ClinicalBERT version of our foundation models.

Minimal differences were also observed between the encoder-only Bidirectional Encoder Representations from Transformers (BERT) models [30] and the decoder-only Generative pretrained Transformer (GPT) models [31], in-spite the BERT models being more tailored to downstream classification tasks in other domains, such as sentiment classification and name entity recognition (NER) [32]. This suggests that GPT models, traditionally used for generative tasks, can also perform equally or more competitively in classification tasks in perioperative care contrary to what is expected of other domains.

2.8 Generalizability of our methods

Generalizability of methods across datasets spanning distinct health centers and hospital units remains a core concern in building predictive models towards clinical care

[33]. To demonstrate that our methods are generalizable beyond BJH and perioperative care, we replicate our methods towards the publicly available MIMIC-III [34] dataset. Details of the results and the specifics aspects of MIMIC-III that were used to replicate our methods are detailed in the supplementary material. Our results aligned with the main findings of our study. Specifically, the baseline models were outperformed by pretrained LLMs, with absolute increases that ranged from up to 14.6% in 12-hour discharge to 30% for In-hospital mortality for AUROC. Similarly, increases in the AUPRC ranged from 17.7% in 12-hour discharge to 28.2% for in-hospital mortality. Similar to the results observed in our BJH dataset, self-supervised fine-tuning witness maximal absolute improvements in AUROCs of up to 0.4% in 12-hour discharge to 3% in in-hospital mortality and AUPRCs of up to 0.4% in 30-day mortality to 0.8% in 12-hour discharge. Incoporating labels into the fine-tuning process also saw further improvements of 0.5% in 12-hour discharge to 1.6% in 30-day mortality and 0.3% in 12-hour discharge to 3.6% for in-hospital mortality for AUROC and AUPRC, respectively. Finally, foundation models again performed the best, with AUROC improvements of 0.4% in 12-hour discharge to 1.4% in 30-day mortality and AUPRC improvements of 0.2% in 12-hour discharge to 2.6% for in-hospital mortality when compared to self-supervised fine-tuning. In the same order, the mean squared error (MSE) of LOS decreased by up to 85.9 days, 3.1 days, 8 days and 8.2 days, respectively.

The replicability of methods demonstrates that the foundational capabilities of LLMs potentially extends beyond perioperative care and BJH to other domains of clinical care spanning different healthcare systems.

3 Discussion

The release of ChatGPT and its rapid rise in popularity have sparked significant interest in applying LLMs to decision support systems in medical care [35, 36]. However, challenges such as the low-resource nature of clinical notes – which are neither readily open-sourced nor abundantly available [37] – and the complex abbreviations and specialized medical terminologies that are not well-represented in the training corpora of pre-trained models [36] have made the applications of LLMs in healthcare onerous. These obstacles have driven efforts to refine, adapt, and evaluate LLMs for reliable use across distinct clinical environments, such as personalized treatment in patient care [16], the automation of medical reports [15] or summaries [17], and medical diagnosis for conditions like COVID-19 [20] or Alzheimer's [21]. Despite these advancements, the application of LLMs in forecasting postoperative surgical complications from preoperative notes remains largely unexplored. Addressing this gap is critical, as such predictions could facilitate the early identification of patient risk factors and enable the timely administration of evidence-based treatments, such as antibiotics in postoperative settings [1, 38, 39].

While various AI models leveraging discrete EHR data have demonstrated success in the early identification of post-operative risks [1, 5], clinical notes taken during the preoperative phase hold immense informational value that is often not captured through tabular data alone. This underscores the pressing need to develop decision

support systems focused on text-based preoperative notes. To address this need, our retrospective study utilizes open-source, pre-trained clinical LLMs on pre-operative clinical notes to predict postoperative surgical complication. Drawing on approximately 85,000 notes from Barnes-Jewish Hospital (BJH) – the largest hospital in Missouri and the largest private employer in the Greater St. Louis region [40] – our study represents a crucial step forward in integrating LLMs into perioperative decision-making processes. By supporting surgical care management, optimizing perioperative care, and enhancing early risk detection, LLMs have the potential to improve patient outcomes and reduce complications that could be avoided through early identification of patient risk factors.

Advanced improvements observed when using pretrained clinical LLMs compared to traditional word embeddings underscore the capability of such models without needing model training, eliminating the potential constraints in training data collection and computational costs and underscoring their capability adjacent to a 'zero-shot' setting [27]. These results are noteworthy given differences between the training copora of such models, such as PubMed articles or public EHR databases like MIMIC, and the semantic characteristics of our clinical notes. Such findings may be handy when data or computational resources may be limited [41]. Additionally, the enhancements through transfer learning via fine-tuning in an self-supervised setting could be attributed to the adaption of these pretrained models in 'familiarizing' itself towards the texts of our preoperative clinical notes [30, 31]. Such differences might arise from variations in text distributions or frequency, as well as the distinctive use of abbreviations and terminology in each task-specific dataset, contrasting the textual contents found in databases of pretrained models like PubMed or Wikipedia [17, 36, 42]. This perspective highlights the potential limitations of relying purely entirely on pretrained models to capture precise context within diverse clinical texts and accentuates the necessity to refine these LLMs using more extensive clinical datasets [36, 42]. This thereby shows that even in cases where labelled data is scarce, tuning the models with respect to the clinical texts can still lead to further improvements compared to using pretrained models alone. Incorporating labels into the fine-tuning process, when available, can substantially improve performance relative to self-supervised fine-tuning. This ensures tokens linked to specific outcomes are effectively represented in the optimized model, thereby achieving improved performances [18, 43]. The foundation models reached peak performance by incorporating a multi-task learning (MTL) framework at scale, as concurred by Aghajanyan et al [44]. Henceforth, a model-agnostic foundation language model – where a single robust model is fine-tuned across all accessible labels – yielded more competitive outcomes. Such performance is a testament to the model's ability to efficiently capture and retain crucial predictive information through knowledge sharing among distinct postoperative complications during the fine-tuning phase, thereby establishing the foundational capabilities of LLMs in predicting risks from notes taken during perioperative care.

The foundation approach offers tangible benefits beyond achieving strong results. From a computational standpoint, a single model can be fine-tuned to serve various tasks, conserving time and resources that would be expended on tailoring and storing parameters for each specific model [35, 44, 45]. Additionally, this model can

generalize across tasks it has been optimized for through knowledge sharing, outperforming bespoke models and mitigating concerns of potential overfitting to the limited samples from a specific task [46]. From a clinical standpoint, because a single robust model can be deployed to predict multiple tasks, it can be conveniently integrated into clinical workflows [35, 45]. This versatility not only enhances the utility of the model but also contributes to a reduction in the cognitive load required by clinicians to decipher potential risk factors manually [47], thereby reducing the risk of clinician burnout, which is presently at near crisis levels [48]. This simultaneously mitigates issues surrounding the lack of established routines for reading clinical notes, the lack of knowledge in the existence of notes taken prior to handoffs, and competing time demands, which are key individual factors leading to unintentional missed recommendations [49]. Clinicians could henceforth shift their focus to designing and implementing interventions to mitigate these risks, which could be vital in reducing post-operative complications [7, 8]. This is supported by studies have shown that carefully designed interventions from the early identification of risk factors could mitigate complications by up to 80%, which is more effective than the effect of drug or treatment interventions [7].

Minor variances in outcomes emerged when comparing the results of training different machine learning predictors on extracted textual embeddings, including the utilization of the fully connected neural network found atop the contextual representations. This affirms that the robustness of LLMs lies in the attention-based architecture, which leads to precise text representations, while the domain-specific dataset on which the language model is fine-tuned plays a crucial role in its adaptability [19, 34, 50]. Similarly, there was no consistent performance gain using BERT-based clinical LLM models over GPT-based models, even though they had different training strategies and the former is more widely used in downstream prediction. This highlights the need for ongoing enhancement of clinical LLMs using innovative techniques across a wide spectrum of medically relevant datasets [36], rather than focusing on the predictor model's training architecture.

Given that model harmfulness and hallucinations are core concerns in the application of LLMs in medicine [17], we tested our model with several broad-case prompts. Our outputs not only assured us of its safety, contradicting some of the potentially harmful outputs witnessed from the base models our foundation models were trained on, but also qualitatively assert its semantic shift towards perioperative-specific corpora from the broad biomedical or critical care unit notes on which these base models were trained [36].

Despite demonstrating the potential of LLMs in predicting risks from notes taken during preoperative care, our study possesses limitations. First, quality of the textual data may possess limitations in itself, ranging from incompleteness [51], to potentially missing data [51], to the potential lack of frequent updates [52], which could potentially possess an adverse impact on the performances of our fine-tuned LLMs. Second, non-textual variables were not accounted for in the models, including demographics, preoperative measurements, key intraoperative variables (e.g., blood transfusion data, and urine output), and certain medications used during surgery [1]. Adding these variables could potentially improve the model performance [1]. Third, it remains unclear if

the observations can be generalized to LLMs in other applications or at various scales (e.g., the public LLaMa models with various sizes). Forth, subgroup analysis based on the various surgery types was not conducted because of the small number of patients within each subgroup; hence, the clinical utility of the predictions of postoperative complications based on specific surgery types is limited. To address these limitations, our ongoing work involves data triangulation across the administrative data, clinical text, and other data to align with high-quality manual health record review provided by National Surgical Quality Improvement Program adjudicators.

4 Methods

4.1 Data

Our dataset originated from electronic anesthesia records (Epic) for all adult patients undergoing surgery at Barnes-Jewish Hospital (BJH), the largest healthcare system in the greater St. Louis (MO) region, spanning four years from 2018 to 2021. The dataset included $n = 84,875$ surgical cases. The text-based notes were single-sentence documents with a vocabulary size of $|V| = 3,203$, and mean word and vocabulary lengths of $\bar{l}_w = 8.9$ (SD: 6.9) and $\bar{l}_v = 7.3$ (SD: 4.4), respectively. The textual data contained detailed information on planned surgical procedures, which were derived from smart text records during patient consultations. To preserve patient privacy, the data was provided in a de-identified and censored format, with patient identifiers removed and procedural information presented in an arbitrary order, ensuring that no information could be traced back to any uniquely identifiable patient.

4.2 Participants

Of the 84,875 patients, the distribution of patient types included 17% (14,412) in orthopedics, 8.8% (7,442) in ophthalmology, and 7.4% (6,236) in urology. The gender distribution was 50.3% male (42,722 patients). Ethnic representation included 74% White (62,563 patients), 22.6% African American (19,239 patients), 1.7% Hispanic (1,488 patients), and 1.2% Asian (1,015 patients). The mean weight was 86 kg ($\pm$24.7 kg) and the mean height was 170 cm ($\pm$11 cm). Other characteristics of the patients could be found in Table 3. The internal review board of Washington University School of Medicine in St. Louis (IRB #201903026) approved the study with a waiver of patient consent.

Characteristics	Statistic
Patient type	Orthopedic 14412 (17%) Ophthalmology: 7442 (8.8%) Urology: 6236 (7.4%)
Gender	Male: 42722 (50.3%)
Ethnicity	White: 62563 (74%) African American: 19239 (22.6%), Hispanic: 1488 (1.7%) Asian: 1015 (1.2%)
Weight	86kg (24.7kg)
Height	170 cm (11cm)
Liver disease	Yes: 6697 (7.9%)
Cancer	Yes: 29213 (34%)
Congestive Heart Failure	Yes: 8886 (10%)
Myocardial Infarction	Yes: 8587 (10%)
Chronic Pulmonary Disease	Yes: 9837 (12%)
HIV/AIDS	Yes: 4069 (4.8%)

Table 3 Characteristics of patients from the BJH dataset. Categorical variables are reported as number of patients (percentage), numerical variables are reported as mean (standard deviation). Note that the summary statistics were computed after removing records with no texts associated with the patient.

4.3 Outcomes

Our six outcomes were: 30-day mortality, acute knee injury (AKI), pulmonary embolism (PE), pneumonia, deep vein thrombosis (DVT), and delirium. These six outcomes remain pertinent in perioperative care, particularly during OR-ICU handoff [1, 53, 54].

AKI was determined using a combination of laboratory values (serum creatinine) and dialysis event records, and structured anesthesia assessments, laboratory data, and billing data indicating baseline end-stage renal disease were used as exclusion criteria for AKI. Acute kidney injury was defined according to the Kidney Disease Improving Global Outcomes criteria. Delirium was determined from nurse flow-sheets (positive Confusion Assessment Method for the Intensive Care unit test result); pneumonia, DVT, and PE were determined based on the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) diagnosis codes. Patients without delirium screenings were excluded from the analysis of that complication.

4.4 Models

4.4.1 Traditional word embeddings and pretrained LLMs

We employed a combination of BERT [30] and GPT-based [31] large language models (LLMs) trained on either biomedical or open-source clinical corpora –specifically BioGPT [22], ClinicalBERT [18], and BioClinicalBERT [19] – for predicting risk factors from notes taken during perioperative care. BioGPT is a 347-million-parameter model trained on 15 million PubMed abstracts [22]. It adopts the GPT-2 model architecture, making it an auto-regressive model trained on the language modeling task, which seeks to predict the next word given all preceding words. In contrast, ClinicalBERT was trained on the publicly available Medical Information Mart for Intensive Care III (MIMIC-III) dataset, which contains 2,083,180 de-identified clinical notes associated with critical care unit admissions [18]. It was initialized from the $\text{BERT}_{\text{base}}$ architecture with the masked language modeling (MLM) and next sentence prediction (NSP) objectives [30]. Similarly, BioClinicalBERT was pretrained on all the available clinical notes associated with the MIMIC-III dataset [19]. However, unlike ClinicalBERT, BioClinicalBERT was based on the BioBERT model [50], which itself was trained on 4.5 billion words of PubMed abstracts and 13.5 billion words of PubMed Central full-text articles. This allowed BioClinicalBERT to leverage texts from both the biomedical and clinical domains.

These models have been tested across representative NLP benchmarks in the medical domains, including Question-Answering tasks benchmarked by PubMedQA [22**?**], recognizing entities from texts [19] and logical relationships in clinical text pairs [55]. In addition to the clinically-oriented LLMs, we included traditional NLP word embeddings as baselines for comparison. These include word2vec's continuous bag-of-words (CBOW) [23], doc2vec [24], GloVe [25], and FastText [26]

By comparing these transformer-based LLMs with traditional words embeddings, we seek to establish the magnitude of improvements clinically-oriented pretrained LLMs could potentially possess in grasping and contextualizing perioperative texts

towards predicting postoperative risks in comparison to traditional word embeddings, represents each word as a vector treated as an independent entity.

4.4.2 Fine-tuned models

A comparison of the distinct fine-tuning methods employed among our study could be best illustrated in Figure 1.

Transfer learning: self-supervised fine-tuning

To bridge the gap between the corpora of the pretrained models and that of perioperative notes, we first expose and adjust these models to our perioperative text through self-supervised fine-tuning. We adapt the pretrained LLMs with our training data in accordance with their existing training objectives. This process leverages the information contained within the source domain and exploits it to align the distributions of source and target data. For BioGPT, this entails the language modeling task [22, 31]. For ClinicalBERT and BioClinicalBERT, it involves the masked language modeling (MLM), as well as the Next Sentence Prediction (NSP) objectives if the document contains multiple sentences, as elaborated in section 4.4.1 [30]. The fine-tuning parameters, including number of epochs, batch size, weight decay, and learning rate, for each of the three base models, are detailed in the supplementary material.

Transfer learning: incorporating labels into fine-tuning

In lieu of anticipated improvements from self-supervised fine-tuning, we took it a step further by incorporating labels as part of the fine-tuning process, in the hopes of boosting performance as demonstrated in past studies [18, 43]. We do this through a semi-supervised approach, as illustrated in Figure 1. This means that in contrast to the self-supervised approach which adjusts weights based solely on training texts, the semi-supervised method infuses label information during the fine-tuning process. In doing so, the model leverages an auxillary fully-connected feed-forward neural network atop of the contextual representations, found in the final layer of the hidden states, to predict the labels as part of its fine-tuning process. The auxillary neural network uses the Binary-Cross-Entropy (BCE), Cross-Entropy (CE), and Mean-Square-Error (MSE) losses for binary classification, multi-label classification, and regression tasks, respectively. A λ parameter was introduced to balance between magnitude of losses between the supervised and self-supervised objectives. The λ parameter, as well as all other parameters used to fine-tune the models, are detailed in the supplementary material. Henceforth, in addition to the potential improvements brought by the self-supervised training objectives, we are now able to leverage the labels to supervise the textual embedding to better align with the training labels.

foundation fine-tuning strategy

To build a foundation model with knowledge across all tasks, we extended the above-mentioned strategies and exploited all possible labels available within the dataset, including but not limited to selected tasks, as inspired by Aghajanyan et al [44]. This involves employing a multi-task learning framework for knowledge sharing across

all available labels from all six outcomes – death in 30 days, AKI, PE, pneumonia, DVT, and delirium – present in the dataset. This task-agnostic approach therefore enables knowledge sharing across all available labels. Therefore, the model becomes foundational in the sense that it solves various tasks simultaneously, meaning a single robust model can be deployed to a wide range of postoperative risks. This contrasts with previous approaches that required a separate model dedicated to each specific outcome. To achieve this, each label is assigned a task-specific auxiliary fully-connected feed-forward neural network, wherein the losses across all labels are pooled together. To control for the magnitude of losses between each task-specific auxiliary network, λ_i for $i = 1, \ldots, m$ given m outcomes, is introduced as weights that contribute to the overall loss calculation. λ_i, as well as all other parameters used to fine-tune the models, are detailed in the supplementary material.

4.5 Predictors

We defaulted to using XGBoost [28] to predict the outcomes from the generated word embeddings or text representations to facilitate consistent comparisons between traditional word embeddings, pretrained LLMs, and their fine-tuned variants. This selection allows to accommodate a diverse range of input types while leveraging XGBoost's widespread use in healthcare due to its robust performance in various clinical prediction tasks. Nonetheless, the choice of predictors remain flexible. This includes utilizing the task specific fully-connected feed-forward neural network found in models that were fine-tuned with respect to the labels.

To examine if the choice of predictors makes a noticable difference in predictive performances, we experimented with various predictors among the models of the best performing fine-tuning strategy. These include (1) the default XGBoost, (2) Random Forest, (3) Logistic Regression (4) and the feed-forward network found in models that incorporate labels into the fine-tuning process. The range of hyper-parameters used for each predictor could be found in the supplementary material.

4.6 Evaluation metrics and validation strategies

For a rigorous evaluation, experiments were stratified into 5 folds for cross-validation [46]. A nested cross-validation approach was used to ensure a robust and fair comparison of the best combination of hyperparameters across all models and approaches spanning multiple folds.

The main evaluation metrics included area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) to get a comprehensive evaluation of the models' overall prediction performance in the face of class imbalance. For our best-performing foundation models, we also reported their accuracy, sensitivity, specificity, precision, and F-scores, which could be found in the supplementary material.

4.7 Qualitative evaluation of our foundation models

We qualitatively evaluate our models for safety while ensuring their adaptation to perioperative corpora beyond the quantitative results, prior to open-sourcing them.

As such, we ran some broad case-prompts on our best-performing foundation models through their objective loss functions. For the BERT-based models, this encompassed the masked language modeling (MLM) objective, where we get the model to fill a single masked token in a 'fill-in-the-blank' format [30]. For the GPT models, this involved the language modeling objective, where we get the model to 'complete the sentence' based on the provided incomplete sentences [31].

4.8 Replication of methods on MIMIC-III

To ensure that our methods are generalizable beyond perioperative care and the BJH dataset, we replicated our methods on the publicly available MIMIC-III dataset [34], which contains de-identified reports and clinical notes of critical care patients at the Beth Israel Deaconess Medical Center between 2001 and 2012. To closely mimic the approach and settings employed in BJH's clinical notes, we utilized the long-form descriptive texts of procedural codes in MIMIC-III. Specifically, ICD-10 codes containing procedural information from each patient were traced and formatted into their respective long-form titles. For each patient, these long-form titles were then combined to form a single-sentenced clinical note, thereby aligning with the textual characteristics of BJH's clinical notes. The selected outcomes were length-of-stay (LOS), in-hospital mortality, 12 hour discharge status, and death in 30 days, as adapted from previous studies [1, 34, 56]. Details on the data characteristics and extraction of outcomes could be found in the supplementary material.

Supplementary Information. Supplementary material have been attached and provided alongside the manuscript submission to the journal.

Acknowledgements. The authors would like to thank Brad Fritz, Michael Avidan, Christopher King, and Sandhya Tripathi from Washington University for collecting and providing the dataset.

Data availability

Our data could not be shared per our IRB. A Patient waiver was obtained for the study. Nonetheless, the best performing foundation models are gated in HuggingFace. The models can be requested with the proper permissions from the corresponding authors.

Code availability

To facilitate reproducibility, the source codes can be publicly accessed through github without restrictions at https://github.com/cja5553/LLMs_in_perioperative_care.

Declarations

This study is supported, in part by, funding from the Agency for Healthcare Research and Quality (R01 HS029324-02).

References

[1] Xue, B., Li, D., Lu, C., King, C.R., Wildes, T., Avidan, M.S., Kannampallil, T., Abraham, J.: Use of machine learning to identify risks of postoperative complications. JAMA Network (2021). https://doi.org/10.1001/jamanetworkopen.2021.2240 . https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2777894

[2] Hamel, M.B., Henderson, W.G., Khuri, S.F., Daley, J.: Surgical outcomes for patients aged 80 and older: Morbidity and mortality from major noncardiac surgery. Journal of the American Geriatrics Society **53**(3), 424–429 (2005) https://doi.org/10.1111/j.1532-5415.2005.53159.x

[3] Healey, M.A.: Complications in surgical patients. Archives of Surgery **137**(5), 611–618 (2002) https://doi.org/10.1001/archsurg.137.5.611

[4] Turrentine, F.E., Wang, H., Simpson, V.B., Jones, R.S.: Surgical risk factors, morbidity, and mortality in elderly patients. Journal of the American College of Surgeons **203**(6), 865–877 (2006) https://doi.org/10.1016/j.jamcollsurg.2006.08.026

[5] Xue, B., Jiao, Y., Kannampallil, T., Fritz, B., King, C., Abraham, J., Avidan, M., Lu, C.: Perioperative predictions with interpretable latent representation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA (2022). https://doi.org/10.1145/3534678.3539190

[6] Tevis, S.E., Kennedy, G.D.: Postoperative complications and implications on patient-centered outcomes. Journal of Surgical Research **181**(1), 106–113 (2013) https://doi.org/10.1016/j.jss.2013.01.032

[7] Impact of perioperative care on healthcare resource use: Rapid research review. Technical report, Centre for Perioperative Care (2020). https://www.cpoc.org.uk/sites/cpoc/files/documents/2020-09/Impact%20of%20perioperative%20care%20-%20rapid%20review%20FINAL%20-%2009092020MW.pdf

[8] Wolfhagen, N., Boldingh, Q.J.J., Boermeester, M.A., Jonge, S.W.: Perioperative care bundles for the prevention of surgical-site infections: meta-analysis. British Journal of Surgery **109**(10), 933–942 (2022) https://doi.org/10.1093/bjs/znac196

[9] Hofer, I.S., Lee, C., Gabel, E., Baldi, P., Cannesson, M.: Development and validation of a deep neural network model to predict postoperative mortality, acute kidney injury, and reintubation using a single feature set. NPJ digital medicine **3**(1), 58 (2020)

[10] Spasic, I., Nenadic, G., *et al.*: Clinical text data in machine learning: systematic review. JMIR medical informatics **8**(3), 17984 (2020) https://doi.org/10.2196/

17984

[11] Braaf, S., Manias, E., Riley, R.: The role of documents and documentation in communication failure across the perioperative pathway. a literature review. International Journal of Nursing Studies **48**(8), 1024–1038 (2011) https://doi.org/10.1016/j.ijnurstu.2011.05.009

[12] Riley, R., Manias, E.: Governing the Operating Room List, pp. 67–89. Palgrave Macmillan UK, London (2007). https://doi.org/10.1057/9780230595477_4

[13] FitzHenry, F., Murff, H.J., Matheny, M.E., Gentry, N., Fielstein, E.M., Brown, S.H., Reeves, R.M., Aronsky, D., Elkin, P.L., Messina, V.P., Speroff, T.: Exploring the frontier of electronic health record surveillance. Medical Care **51**(6), 509–516 (2013) https://doi.org/10.1097/mlr.0b013e31828d1210

[14] Nazi, Z.A., Peng, W.: Large language models in healthcare and medical domain: A review. In: Informatics, vol. 11, p. 57 (2024). https://doi.org/10.3390/informatics11030057 . MDPI

[15] Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 (2023) https://doi.org/10.48550/arXiv.2302.07257

[16] Yang, H., Li, J., Liu, S., Du, L., Liu, X., Huang, Y., Shi, Q., Liu, J.: Exploring the potential of large language models in personalized diabetes treatment strategies. medRxiv, 2023–06 (2023) https://doi.org/10.1101/2023.06.30.23292034

[17] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C.P., Hom, J., Gatidis, S., Pauly, J., Chaudhari, A.S.: Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine **30**(4), 1134–1142 (2024) https://doi.org/10.1038/s41591-024-02855-5

[18] Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. In: CHIL 2020 Workshop (2019). https://doi.org/10.48550/arXiv.1904.05342

[19] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78 (2019). https://doi.org/10.18653/v1/W19-1909

[20] Li, H., Gerkin, R.C., Bakke, A., Norel, R., Cecchi, G., Laudamiel, C., Niv, M.Y., Ohla, K., Hayes, J.E., Parma, V., *et al.*: Text-based predictions of covid-19 diagnosis from self-reported chemosensory descriptions. Communications Medicine **3**(1), 104 (2023) https://doi.org/10.1038/s43856-023-00334-5

[21] Mao, C., Xu, J., Rasmussen, L., Li, Y., Adekkanattu, P., Pacheco, J., Bonakdarpour, B., Vassar, R., Shen, L., Jiang, G., *et al.*: Ad-bert: Using pre-trained language model to predict the progression from mild cognitive impairment to alzheimer's disease. Journal of Biomedical Informatics **144**, 104442 (2023) https://doi.org/10.1016/j.jbi.2023.104442

[22] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics **23**(6), 409 (2022) https://doi.org/10.1093/bib/bbac409

[23] Church, K.W.: Word2vec. Natural Language Engineering **23**(1), 155–162 (2016) https://doi.org/10.1017/s1351324916000334

[24] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 1188–1196. PMLR, Bejing, China (2014). https://proceedings.mlr.press/v32/le14.html

[25] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Stroudsburg, PA, USA (2014). https://doi.org/10.3115/v1/d14-1162

[26] Athiwaratkun, B., Wilson, A.G., Anandkumar, A.: Probabilistic FastText for Multi-Sense Word Embeddings (2018). https://arxiv.org/abs/1806.02901

[27] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Curran Associates Inc., Red Hook, NY, USA (2020). https://doi.org/10.5555/3495724.3495883

[28] Chen, T., Guestrin, C.: Xgboost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785

[29] Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024) https://doi.org/10.1038/s41586-024-07421-0

[30] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv (2019) https://doi.org/10.48550/arXiv.1810.04805

[31] Radford, A., Narasimhan, K., Salimansa, T., Sutskever, I.: Improving language understanding by generative pre-training. Technical report, OpenAI (2019). https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[32] Aftan, S., Shah, H.: A survey on bert and its applications. In: 2023 20th Learning and Technology Conference (L&T), pp. 161–166 (2023). https://doi.org/10.1109/LT58159.2023.10092289

[33] Goetz, L., Seedat, N., Vandersluis, R., Schaar, M.: Generalization—a key challenge for responsible ai in patient-facing clinical applications. npj Digital Medicine **7**(1) (2024) https://doi.org/10.1038/s41746-024-01127-3

[34] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific Data **3**(1), 1–9 (2016) https://doi.org/10.1038/sdata.2016.35

[35] Raza, M.M., Venkatesh, K.P., Kvedar, J.C.: Generative ai and large language models in health care: pathways to implementation. npj Digital Medicine **7**(1) (2024) https://doi.org/10.1038/s41746-023-00988-4

[36] Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M.A., Fries, J., Shah, N.H.: The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine **6**(1) (2023) https://doi.org/10.1038/s41746-023-00879-8

[37] Wu, H., Wang, M., Wu, J., Francis, F., Chang, Y.-H., Shavick, A., Dong, H., Poon, M.T., Fitzpatrick, N., Levine, A.P., *et al.*: A survey on clinical natural language processing in the united kingdom from 2007 to 2022. NPJ digital medicine **5**(1), 186 (2022) https://doi.org/10.1038/s41746-024-01166-w

[38] Bland, K.I.: Antimicrobial prophylaxis for surgery: An advisory statement from the national surgical infection prevention project. Yearbook of Surgery **189**(4), 225–226 (2006) https://doi.org/10.1016/s0090-3671(08)70480-4

[39] Young, P.Y., Khadaroo, R.G.: Surgical site infections. Surgical Clinics of North America **94**(6), 1245–1264 (2014) https://doi.org/10.1016/j.suc.2014.08.008

[40] Hospital, B.-J.: About us: Barnes-Jewish Hospital at Washington University Medical Center. https://www.bjc.org/locations/hospitals/barnes-jewish-hospital

[41] Sandmann, S., Riepenhausen, S., Plagwitz, L., Varghese, J.: Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. Nature Communications **15**(1), 2050 (2024) https://doi.org/10.1038/s41467-024-46411-8

[42] Shojaee-Mend, H., Mohebbati, R., Amiri, M., Atarodi, A.: Evaluating the strengths and weaknesses of large language models in answering neurophysiology questions. Scientific Reports **14**(1), 10785 (2024) https://doi.org/10.1038/s41598-024-60405-y

[43] Chen, X., Beaver, I., Freeman, C.: Fine-tuning language models for semi-supervised text mining. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 3608–3617 (2020). https://doi.org/10.1109/BigData50022.2020.9377810

[44] Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., Gupta, S.: Muppet: Massive multi-task representations with pre-finetuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, USA (2021). https://doi.org/10.18653/v1/2021.emnlp-main.468

[45] Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., Kather, J.N.: A guide to artificial intelligence for cancer researchers. Nature Reviews Cancer, 1–15 (2024) https://doi.org/10.1038/s41568-024-00694-7

[46] Kernbach, J.M., Staartjes, V.E.: Foundations of machine learning-based clinical prediction modeling: Part ii—generalization and overfitting, pp. 15–21. Springer, ??? (2022). https://doi.org/10.1007/978-3-030-85292-4_3

[47] Koopman, R.J., Steege, L.M.B., Moore, J.L., Clarke, M.A., Canfield, S.M., Kim, M.S., Belden, J.L.: Physician information needs and electronic health records (ehrs): time to reengineer the clinic note. The Journal of the American Board of Family Medicine **28**(3), 316–323 (2015) https://doi.org/10.3122/jabfm.2015.03.140244

[48] Lou, S.S., Liu, H., Warner, B.C., Harford, D., Lu, C., Kannampallil, T.: Predicting physician burnout using clinical activity logs: model performance and lessons learned. Journal of biomedical informatics **127**, 104015 (2022) https://doi.org/10.1016/j.jbi.2022.104015

[49] Flemons, K., Bosch, M., Coakeley, S., Muzammal, B., Kachra, R., Ruzycki, S.M.: Barriers and facilitators of following perioperative internal medicine recommendations by surgical teams: a sequential, explanatory mixed-methods study. Perioperative Medicine **11**(1) (2022) https://doi.org/10.1186/s13741-021-00236-x

[50] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020) https://doi.org/10.1093/bioinformatics/btz682

[51] Troiani, J.S., Finkelstein, S.M., Hertz, M.I.: Incomplete event documentation in medical records of lung transplant recipients. Progress in Transplantation **15**(2),

173–177 (2005) https://doi.org/10.1177/152692480501500211

[52] Walker, J., Leveille, S., Bell, S., Chimowitz, H., Dong, Z., Elmore, J.G., Fernandez, L., Fossa, A., Gerard, M., Fitzgerald, P., *et al.*: Opennotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. Journal of medical Internet research **21**(5), 13876 (2019)

[53] Bellini, V., Valente, M., Bertorelli, G., Pifferi, B., Craca, M., Mordonini, M., Lombardo, G., Bottani, E., Rio, P.D., Bignami, E.: Machine learning in perioperative medicine: A systematic review - journal of anesthesia, analgesia and critical care. BioMed Central (2022) https://doi.org/10.1186/s44158-022-00033-y

[54] Amin, A.P., Crimmins-Reda, P., Miller, S., Rahn, B., Caruso, M., Funk, M., Pierce, A., Kurz, H.I., Lasala, J.M., Zajarias, A., *et al.*: Reducing acute kidney injury and costs of percutaneous coronary intervention by patient-centered, evidence-based contrast use: the barnes-jewish hospital experience. Circulation: Cardiovascular Quality and Outcomes **12**(3), 004961 (2019) https://doi.org/10.1161/CIRCOUTCOMES.118.004961

[55] Shivade, C., *et al.*: Mednli-a natural language inference dataset for the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, pp. 1586–1596 (2019). https://doi.org/10.13026/C2RS98

[56] Luo, M., Chen, Y., Cheng, Y., Li, N., Qing, H.: Association between hematocrit and the 30-day mortality of patients with sepsis: a retrospective analysis based on the large-scale clinical database mimic-iv. PLoS one **17**(3), 0265758 (2022) https://doi.org/10.1371/journal.pone.0265758