Outlook

Re: Digital Twins Next Steps

From Mikey Ferguson <MFerguson@laureateinstitute.org>
Date Thu 6/5/2025 1:04 PM
To Martin Paulus <mpaulus@laureateinstitute.org>; Rayus Kuplicki <rkuplicki@laureateinstitute.org>

Fantastic, thank you for this plan! Seems a great way to give me more structure. I will give it my best and keep you both updated.

Get Outlook for iOS

From: Martin Paulus <mpaulus@laureateinstitute.org>
Sent: Thursday, June 5, 2025 12:55:50 PM
To: Mikey Ferguson <MFerguson@laureateinstitute.org>; Rayus Kuplicki <rkuplicki@laureateinstitute.org>
Subject: Re: Digital Twins Next Steps

Hi Mikey
Thanks for your email, I think your problem statement captures the essence, I took the liberty of refining it somewhat based on what I had seen, take a look at the following:

Refined, Testable Problem Statement
Goal: **Quantify how well different embedding + similarity pipelines capture clinically meaningful proximity between patient trajectories and determine how much narrative context is required for an LLM (e.g., Med-Gemma) to judge that proximity in agreement with expert clinicians.**

1. **Hypotheses**
   - *H1 (Geometry Match)*: Distances in embedding space will correlate monotonically ($\rho > 0.6$) with (a) expert similarity ratings and (b) similarity of downstream outcomes (e.g., next-visit diagnoses, 30-day readmission).
   - *H2 (Off-the-Shelf vs. Tuned)*: Clinical-domain transformers (ClinicalBERT, Bio+DischargeBERT, Med-Gemma) outperform generic English transformers and TF-IDF baselines on H1.
   - *H3 (Representation)*: Trajectory summaries that preserve temporal order (visit-level "sentence" tokens) outperform unordered bag-of-codes or isolated last-visit summaries.
   - *H4 (Context Length)*: Med-Gemma's agreement with experts plateaus once ≥ 6 most-recent visits (or ~2 k tokens) are provided.
   - *H5 (Rating Scale)*: A 0-10 Likert scale yields higher inter-rater reliability (ICC > 0.75) than coarser 1-5 scales.
2. **Data & Cohort Construction**

- **Dataset**: ~20 k adult patients with at least eight outpatient/inpatient encounters in SFHS 2015-2024, de-identified and split (train 60 %, dev 20 %, test 20 %).
- **Extracted Features per visit** (all ISO-date-time stamped): diagnoses (ICD-10-CM), procedures, meds started/stopped, lab-highlights, and free-text note embeddings (discharge or progress).
- **Ground-truth outcomes**: next-visit primary ICD-10, 30-day readmission flag, all-cause one-year mortality.

3. **Patient-Trajectory Representations (to be compared)**

ID	Method	Tokenization	Notes
R1	Bag-of-codes	Unordered ICD10+RxNorm	Baseline TF-IDF
R2	Visit-Sentence	[DATE] DX1	RX1
R3	Temporal Embedding	Same as R2 but add position IDs	Uses transformer positional encodings
R4	Summarized Narrative	Auto-summary (GPT-4o) of last k visits	≤ 4096 tokens
R5	Hybrid	Visit-Sentence for structured + summary of notes	Concatenate

4. **Embedding Generators**
   - **E1** TF-IDF (bag-of-ngrams).
   - **E2** Universal Sentence Encoder-Multilingual-QA (generic).
   - **E3** BioSentVec (PubMed + MIMIC).
   - **E4** ClinicalBERT-SentenceTransformer (HuggingFace).
   - **E5** Med-Gemma instruction-tuned embeddings.
   - **E6** E4 → domain-adaptive pre-training on SFHS notes (continue-PT for 1 epoch).
5. **Similarity Metrics**
   - Cosine, Euclidean, Manhattan, and **Learned Metric**: Siamese network trained with triplet loss on expert-labeled similar/dissimilar pairs.
6. **Evaluation Framework**
   - **Nearest-Neighbor Retrieval**
     - For each *index* patient in the test set, retrieve top N = 10 neighbors per (Representation × Embedding × Metric) triple.
   - **LLM Similarity Scoring**
     - **Prompt template** (system): “You are a clinical reasoning assistant.”
     - **User**: `\n\nCandidate Patient A: \n…\nRate from 0 (no resemblance) to 10 (almost identical) how similar each candidate’s clinical course is to the index.”
     - Record Med-Gemma’s scalar for each candidate.
   - **Expert Panel**
     - Three board-certified psychiatrists + one internist. Provide identical summaries and record 0-10 scores. Compute average (gold).
   - **Metrics**
     - *Correlation*: Spearman ρ between embedding distance and expert score.
     - *Agreement*: Med-Gemma vs. expert ICC (2-way random).
     - *Outcome Consistency*: proportion of retrieved neighbors sharing index’s next-visit ICD-10 chapter; AUROC for 30-day readmission similarity.
     - *Statistical Tests*: Friedman test across pipelines; post-hoc Wilcoxon with Holm correction.

1. **Experiments**

Exp	Focus	Variables	Fixed Parameters
A	Representation	R1-R5	E4 + cosine
B	Embeddings	E1-E6	Best R from Exp A + cosine
C	Metric	cosine, Euclidean, Manhattan, Learned	Best R,E
D	Context Ablation	visits = 2,4,6,8,full	Best R,E,metric
E	Rating Scale	0-10 vs 1-5	50 random seeds, compute ICC

2. **Success Criteria**
    1. Any pipeline with $\rho \geq 0.7$ **and** ICC (Med-Gemma vs experts) ≥ 0.8 is deemed clinically aligned.
    2. Learned metric beating cosine by ≥ 5 % on both correlation and outcome consistency is considered a significant gain.
2. **Refinement Approaches (if H2 fails)**
    1. **Contrastive Fine-Tuning**: Use expert-rated pairs to fine-tune embedding model with InfoNCE loss.
    2. **Outcome-Supervised**: Jointly optimize embeddings to predict next-visit ICD codes (multi-label) and 30-day readmission.
    3. **Prompt Engineering**: Provide chain-of-thought exemplars for similarity judging.
    4. **Augmented Tokens**: Add demographics and high-level comorbidity flags as special tokens.
3. **Deliverables & Timeline (≈12 weeks)**

Week	Milestone
1	Data extract + cohort freeze
2-3	Representation scripts (R1-R5)
4-5	Generate embeddings E1-E5; continue-PT for E6
6	Implement similarity retrieval APIs
7-8	Collect expert ratings (200 index patients × 10 neighbors)
9	Run Experiments A-C
10	Run Experiments D-E
11	Statistical analysis; draft plots
12	Technical report, code repo, reproducibility checklist

3. **Risk Mitigation & Checks**

- **Data leakage**: verify patients in test set absent from fine-tuning (E6).
- **Prompt bias**: randomize neighbor order; blind experts to embedding type.
- **Compute limits**: cap token length to 4 k; chunk summaries if exceeded.
- **Ethics**: de-ID review, HIPAA-compliant storage, IRB exemption or approval recorded.

The Art of War: In the midst of chaos, there is also opportunity.
Martin

Scientific Director and President, Laureate Institute for Brain Research
6655 S Yale Ave Tulsa, OK 74136-3326 P 918 502 5120 F 918 502 5135
email: mpaulus@laureateinstitute.org X:@mpwpaulus

web: http://www.laureateinstitute.org

Professor of Neuroscience, Oxley College of Health and Natural Sciences
University of Tulsa, email: mpp4692@utulsa.edu

Adjunct Professor, Department of Psychiatry, University of California, San Diego
email: mpaulus@health.ucsd.edu

From: Mikey Ferguson <MFerguson@laureateinstitute.org>
Date: Thursday, June 5, 2025 at 12:18 PM
To: Martin Paulus <mpaulus@laureateinstitute.org>, Rayus Kuplicki <rkuplicki@laureateinstitute.org>
Subject: Digital Twins Next Steps

Hi Dr.'s Paulus and Kuplicki,
Apologies for not reaching out to you both sooner about this, but it has taken me some time to organize my thoughts. I was feeling rather lost, but Dale and I talked this morning, and he's been super helpful in helping me to organize my next steps. Some exciting news for us students – Google Gemini Pro is free for us for the next year! After a year must re-verify our student statuses again… presumably it will stay free as long as we're students? I'd sure love that. As such, I called on my friend Gemini to help summarize the next steps that Dale helped me to decide on:
Here's a quick summary of our plan:

1. **The Core Problem:** We recognize that the quality of our LLM's predictions for future patient visits heavily relies on how effectively we identify "closest" historical patient trajectories. This hinges on our chosen method for turning visit sequences into numerical vectors (embeddings) and the metric we use to measure their similarity.
2. **Proposed Testing Framework:** We will develop a systematic framework to compare different vectorization methods and similarity metrics.
    - **Vectorization Methods to Test:** We will experiment with various models, including our current BioBERT-based Sentence Transformer, other medically focused Sentence Transformers from Hugging Face, and potentially simpler baselines like TF-IDF.
    - **Similarity Metrics to Test:** We will compare cosine similarity (our current method) with other metrics like Euclidean distance.
    - **LLM-based Evaluation:** The novelty of our approach lies in using a powerful, domain-specific LLM (specifically, Med-Gemma, which Dale found on HuggingFace, and we plan to integrate) to objectively evaluate the "relevance" of the identified nearest neighbors.
        - For a given patient's history, we will provide the LLM with narratives of that patient's journey and the narratives of the top 'N' similar patients found by each vectorization/similarity combination.
        - The LLM will then be prompted to provide a 1-10 rating on "How close are these nearby neighbors really?" This score will serve as our primary metric for determining the effectiveness of each embedding and similarity approach.
3. **Rationale for this approach:** This iterative evaluation, leveraging the nuanced

understanding of a medical LLM, will allow us to objectively determine which combination of vectorization and similarity truly captures the most clinically relevant relationships between patient trajectories. Improving this foundational step is critical for enhancing the overall accuracy and clinical utility of our Digital Twin predictions.

Back to Mikey talking:

I just wanted to let you know about the plan – if you have any suggestions, I am ears, and I can certainly make any additional meetings either of you feel we should have – especially if talking with Wes or Chun about these next steps could be helpful.

Thanks!

Mikey