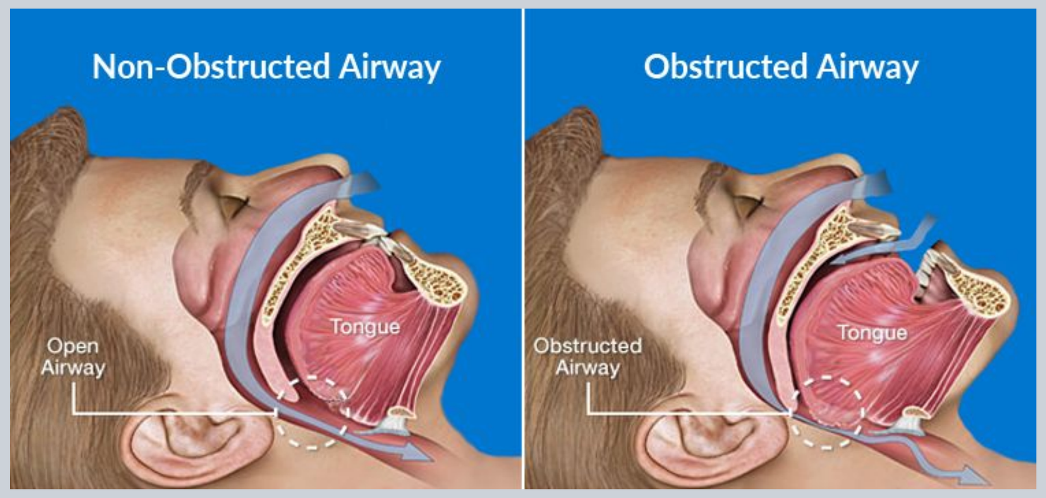
Utilizing Large Language Models for Disease Phenotyping in Obstructive Sleep Apnea

Ifrah Khurram[1,2], Rafael Zamora-Resendiz[2], Destinee Morrow[2], Silvia Crivelli[2]
[1]San Juan Bautista School of Medicine, [2]Lawrence Berkeley National Laboratory

ABSTRACT

Obstructive sleep apnea (OSA) impacts millions, linking to severe complications yet understanding its influence on comorbidities lags. Complications can be avoided by using expensive continuous positive airway pressure (CPAP) machines, but physicians cannot identify those at risk. Large language models (LLMs) have recently made impressive advancements in sequence modeling, and clinical applications are quickly emerging. However, the medical relevance of pre-trained LLM latent spaces remains uncertain. This study gauges 12 pre-trained clinical LLMs, clustering OSA-related phenotypes and comorbidities (atrial fibrillation, coronary artery disease, heart failure, hypertension, stroke, type 2 diabetes). Using 40 A100 GPUs on NERSC's Perlmutter, document-level embeddings for 331,793 MIMIC-IV discharge reports were computed for each LLM. K-Means models were ranked by clustering entropy of phenotype classes, guiding model selection. The top models successfully subset patients with similar histories and outcomes. This work will support ongoing OSA research by identifying phenotypes and assist physicians by informing CPAP allocation.

BACKGROUND



How does OSA interact with the disease progression of cardiovascular comorbidities?

- Obstruction during sleep → coronary blood flow does not increase proportionately with myocardial work
  - ↑ in coronary artery vascular resistance
  - **⇒ OSA has a mechanism to increase cardiovascular comorbidities in a dose-dependent fashion** [1]

Comorbidities explored:

- Atrial Fibrillation
- Coronary Artery Disease
- Heart Failure
- Hypertension
- Stroke
- Type 2 Diabetes Mellitus

LLMs used:

- BioBART
- BioBERT
- BioGPT
- BioMegatron
- Bio_ClinicalBERT
- Gatortron
- RadBERT

SIGNIFICANCE

Discover the **poorly understood disease progression pathways of OSA and its comorbidities** to enhance precision medicine for improved patient treatments

Gain efficiency in CPAP allocation and triaging Veterans Affairs (VA) patients who would benefit most from CPAP for the **optimization of patient care**

The world's elderly population grows yearly. Since older age is a known risk factor for OSA, this research also has **the potential to impact the broader world** [2]

RESEARCH QUESTION

Do documents within LLM clusters share common medical characteristics?

RESULTS

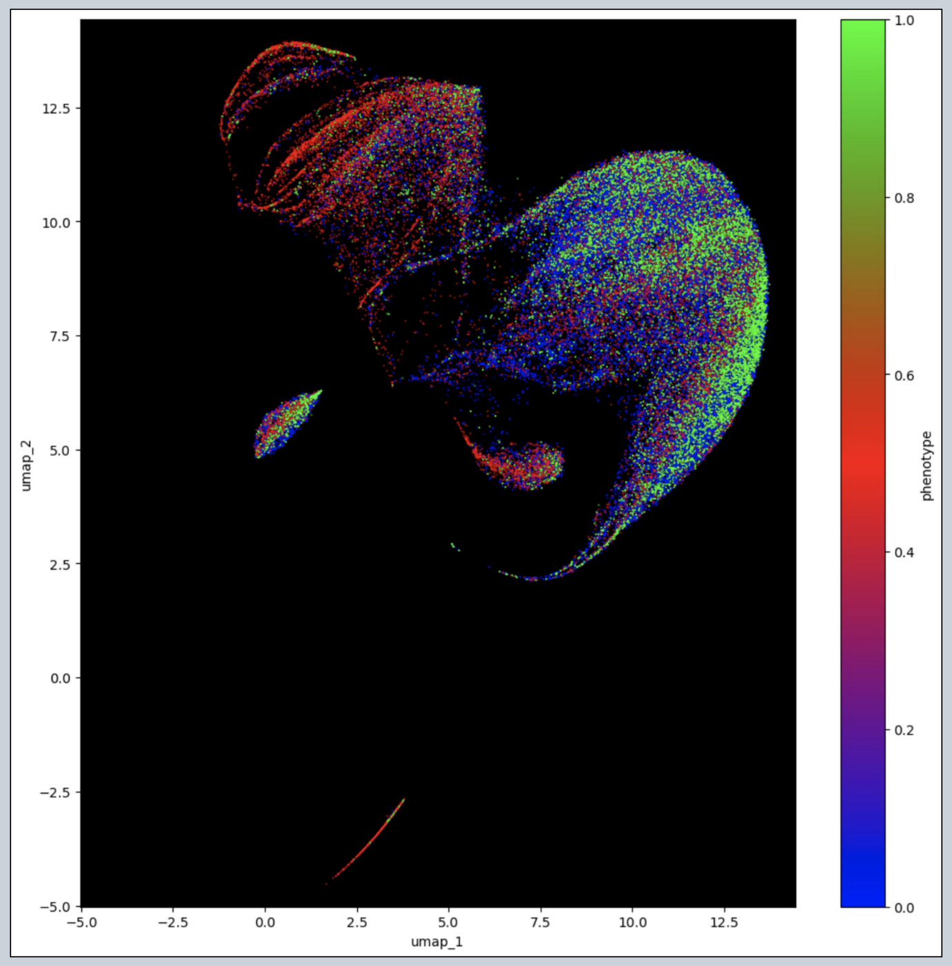


Figure 1: Gatortron_base, layer = 1, K = 1024 on HF + OSA

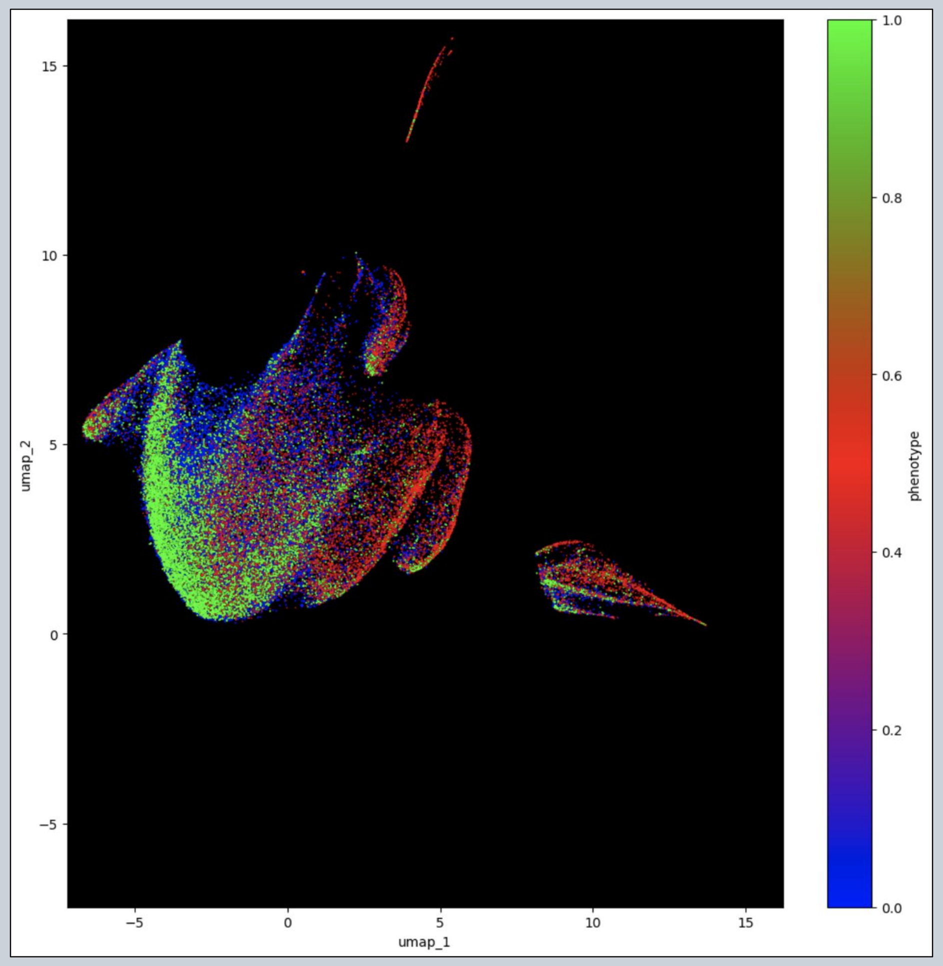


Figure 2: BioGPT_large, layer = 0, K = 1024 on HF + OSA

Each point is a patient document. The comorbidity of focus is *heart failure*. There is clustering of the documents by color, separating the comorbidities: HF + OSA. Our UMAP analysis shows potential for clustering comorbidities as the color clusters are prominent.

Table 1: Ranking the models by Shannon entropy within each independent OSA + comorbidity combination. The red box highlights the entropy of the *heart failure* comorbidity, the focus of this poster. Teal shows the best model for that comorbidity: Gatortron_base.

rank	model	layer	nb_clusters	entropy_osa.afib	entropy_osa.cad	entropy_osa.hf	entropy_osa.htn	entropy_osa.strokeb	entropy_osa.stroken	entropy_osa.t2dm
1	Gatortron_base	1	1024	0.727537	0.713387	0.700127	0.478009	0.667173	0.577617	0.742396
2	Gatortron_base	0	1024	0.733768	0.720652	0.708736	0.479609	0.668317	0.576819	0.743659
3	BioGPT_large	0	1024	0.744778	0.726301	0.717526	0.48123	0.677021	0.589504	0.745349
4	BioGPT_base	0	1024	0.747872	0.724421	0.716554	0.478673	0.679932	0.593591	0.747183
5	BioGPT_base	1	1024	0.746139	0.726391	0.723233	0.478709	0.682348	0.595837	0.747154
6	BioGPT_large	1	1024	0.744199	0.728984	0.720886	0.479312	0.682256	0.596699	0.745871
7	Gatortron_medium	0	1024	0.744878	0.725628	0.718788	0.481929	0.683699	0.595467	0.749013
8	Gatortron_medium	1	1024	0.745283	0.729098	0.723508	0.48276	0.687068	0.596232	0.750483
9	Gatortron_s	1	1024	0.753914	0.731327	0.724027	0.482158	0.68991	0.603036	0.751927
10	Gatortron_s	0	1024	0.755611	0.733709	0.728364	0.484576	0.696359	0.608374	0.753862

DATA & METHODS

MIMIC-IV Database

- 1 Boston medical center
- Over 10 years of data from 2008 to 2019
- 299,712 patients from the hospital as well as the ICU with ≥ 1 admission
- 331,793 unique discharge reports

Qualitative Analysis ⟷ ***Quantitative Analysis***

UMAP

- Reduces dimensions of latent space
- Preserves data structure and relations
- Embeds documents plottable on a (x,y) coordinate plane

Clustering Entropy

- Entropy means disorganization
- ↓ entropy favored
- ↓ Shannon entropy score = ↑ cluster organization, purity, and cohesion

CONCLUSIONS

Overall:

- Gatortron_base's second to last layer (layer=1) had the lowest entropy across 6 out of 7 comorbid 3-class problems when dividing the corpus envelope into 1024 clusters with K-Means.
- Gatortron and BioGPT outperformed other LLMs.

Specific Novel Findings:

- When ranking clusters by class purity, clinical notes were sampled (n=100) from clusters of high rates of OSA patients with heart failure.
  - These notes describe admissions of patients with a history of OSA, dyspnea as a chief complaint, and no prescription/adherence to CPAP treatment.
- As the number of K-means clusters ↑, model performance ↑
- Models trained on larger sets of data perform better at organizing corpora by clinically relevant measures.
- This work contributes to advancing precision medicine as it allows physicians to understand how to better treat patients based on the patient's specific characteristics.

HPC

Document-level embeddings of the MIMIC-IV corpus were computed using NERSC's Perlmutter. Each GPU node contains 4 NVidia A100s (40 GB of onboard mem) and all explored clinical LLMs were able to fit in on a single GPU. Each model was benchmarked to ascertain the inference throughput and runtime. Data parallelism was employed to distribute embedding jobs.

Table 2: Parameter count, input throughput, and inference runtime for clinical LLMs

model	parameters	max batch throughput	runtime (n=100)
BioBart_base	16,404,864	12	0.6037 s
BioBart_large	442,270,720	5	0.4925 s
BioBert_base	108,340,804	36	0.2109 s
BioBert_large	364,360,308	14	0.2988 s
BioGPT_base	346,763,264	6	0.3272 s
BioGPT_large	1,571,188,800	1	0.3272 s
BioMegatron_base	333,640,704	10	0.2831 s
Bio_ClinicalBERT	108,310,272	42	0.2626 s
Gatortron_base	355,267,584	33	0.2740 s
Gatortron_s	355,267,584	33	0.2707 s
Gatortron_medium	3,912,798,720	5	0.1275 s
RadBERT_2m	109,514,298	36	0.2127 s

REFERENCES

1. Wang, X., Ouyang, Y., Wang, Z., Zhao, G., Liu, L., & Bi, Y. (2013). Obstructive sleep apnea and risk of cardiovascular disease and all-cause mortality: a meta-analysis of prospective cohort studies. *International journal of cardiology*, *169*(3), 207–214. https://doi.org/10.1016/j.ijcard.2013.08.088
2. Leatherby, L. (2023, July 16). *How a vast demographic shift will reshape the world*. The New York Times. https://www.nytimes.com/interactive/2023/07/16/world/world-demographics.html?unlocked_article_code=gEr0R8XlhQ2zWZcWTFGTWvp3Wbjzy2-ly9BbypsB-PLkFMlqA822cigQTph1CtBy2XxJ1Wlei8UFE8b77wrw5r1y-yGAmHfi7sPG7geSCxBXaYAf6Xwc7lQoBvUn7mPhpDvhE5-2rTZ1RXWq4fkJCGqzmq87dxf3rpYweKorl-tQhlCkQWQ41LOa1qesg9OtR2iI4Myr72-AI8DHGYsdbGCG_DmR6wgB9BmSsV335AvU_gSz-z1HGtD1dk4BCknKGHpwIeyrezdItaEA3ccj2AoBQdtLJ_dXYAxE2aG_MmMCy6ri19VonUkt0TesGKhpIDI14vRS60L2MZrkc80M5eJf6S2ywdVWTMY&smid=url-share
3. OSA Image: *Obstructive sleep apnea treatment: Beverly Hills: Los Angeles: Sleep study clinic*. Sleep Study Clinic | Home Sleep Apnea Testing. (2019, August 8). https://losangelessleepstudyclinic.com/obstructive-sleep-apnea-treatment/
4. Hernandez, B., Stiff, O., Ming, D. K., Ho Quang, C., Nguyen Lam, V., Nguyen Minh, T., Nguyen Minh, N., Nguyen Quang, H., Phung Khanh, L., Dinh The, T., Huynh Trung, T., Wills, B., Simmons, C. P., Holmes, A. H., Yacoub, S., & Georgiou, P. (2023). *Learning meaningful latent space representations for patient risk stratification: Model development and validation for dengue and other acute febrile illness*. Frontiers in Digital Health, 5, 1057467. https://doi.org/10.3389/fdgth.2023.1057467

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) Program and the Applied Mathematics and Computational Research Division (AMCR) of the Lawrence Berkeley National Lab.