Proposal: NLP-driven interactive healthcare data visualization

Team 50: Stefan Lehman, Lei Han, Kelly Lewis, Ruby Truong, Ying Wang, Luke Williams

Abstract: The goal of this project is to transform raw clinical data, specifically de-identified free-text clinical notes, into visual reports for healthcare professionals and stakeholders with Natural Language Processing (NLP). The team aims to accelerate the identification of key medical phrases, such as conditions, treatments, procedures, and lab results to facilitate quick lookups and analysis for each patient. The novelty in our approach is the integration of advanced NLP techniques to extract and highlight important details such as medication names, conditions, and procedures from clinical notes. As the end-users of these interactive visualizations, healthcare workers can quickly filter and review key information, simplifying time-consuming interactions with electronic health records (EHRs).

Introduction - Clinical Need: Primary care physicians routinely spend more than half of their workday manually reviewing and entering EHRs. Of the average 11.4 hours worked per day, family medicine physicians spend 5.9 of those hours interacting with a computer [15]. Physicians spend an average of 16 minutes and 14 seconds in EHRs per patient encounter [10], competing with patients' time and contributing to physician burnout and job dissatisfaction [15]. Additionally, the shorter the visit, the greater the likelihood of patient harm with inappropriate prescriptions such as overprescription of antibiotics or harmful drug interactions like the co-prescription of opioids and benzodiazepines [12]. With our NLP-driven interactive healthcare data visualization tool, healthcare professionals can easily access a patient's information, allowing them to increase their time with patients and prevent adverse outcomes.

Methods – Processing and Analyzing Unstructured Data: Considering the diverse types and formats of EHRs, such as discharge summaries, procedural notes, lab reports, etc., as well as varying data structures among practices, and hospitals, and clinics, Robertson [8] and Morrison [9] point out the complexity of managing and processing unstructured healthcare data. Thus, the team will focus on analyzing discharge summaries for the EHR of interest due to the variety of treatments, medications, lab results, and procedures captured in this type of record. Ford [1], Abudiyab [2], and Chishtie [3] highlight the importance of preprocessing complex datasets to enable robust data analytics and visualizations downstream for healthcare end-users. The team will conduct thorough Exploratory Data Analysis (EDA) to understand variable relationships and anomalies. The team will then clean the data and parse each record into respective sections, such as family and social history, procedures, and medications.

Methods – Using NLP on EHRs: Our team is looking to advance the application of NLP in healthcare, focusing on a comprehensive approach for identifying and managing chronic diseases and their comorbidities in EHRs. Kop [7], Liu [17], Adarsh [18], and Suryanarayanan [19] highlight NLP's role in improving efficiency, accelerating data extraction, and enhancing clinical decisions. Inspired by Benavent's [13] and Grout's [16] research using NLP to identify and understand comorbidities for patients with rheumatoid arthritis and diabetes, the team will build upon this work by identifying more chronic diseases and comorbidities, facilitating a more holistic and effective patient care model. Shah-Mohammadi and Finkelstein (2024) [14] demonstrate the effectiveness of using NLP to accurately detect Chronic Obstructive Pulmonary Disease within the first 24 hours of hospital admissions for patients, achieving an 80% F1 score based on symptoms from real-time inferencing. Our team seeks to build upon this research by integrating more chronic conditions, such as heart disease, asthma, and arthritis, for real-time detection, expanding the scope of proactive management in chronic diseases. In particular, the team

will use industry standard NLP packages in Python (SpaCy, NLTK, and Scikit-Learn) for entity recognition and extraction of key medical terms, allowing for annotations and visualizations.

Visualizations for Healthcare: Integration of analytical tools into healthcare systems faces many challenges with one of the most significant being the lack of effective use of visualization when adopting these tools into the healthcare system. As highlighted by Velupillai [4], the adaptation and effective use of NLP for health outcomes research underscores the necessity for advancements in visualization techniques. Tian [5] and Shaikh [6] additionally underscore the importance of developing intuitive designs with data for interacting visualizations. The team will build a Flask app (JavaScript, Html, CSS) or Python Streamlit application to create a UI. We will build visualizations highlighting medical terms to allow clinicians to analyze patients with chronic diseases. These visualizations aim to assist healthcare professionals and stakeholders (hospitals, clinics, and billing departments) to access and interpret patient data effectively.

Results - Measuring Success and Risk Factors: While data security risks can be minimized with de-identification, such as the Safe Harbor method that removes 18 protected health identifiers (PHI) [11], this tradeoff may reduce the performance and reliability of the NLP algorithm by eliminating information. Our team will adhere to this standard, using a de-identified dataset, generating dummy data, and implementing other privacy protective measures while still arriving at a demonstrable final product. Another risk is that we may lack the subject matter expertise to define the most optimal and useful features for this NLP visualization tool. Therefore, we are treating this tool as a proof-of-concept expanding upon existing implementations, ensuring we integrate human-in-the-loop design to receive input, feedback, and verification from clinicians. To measure success, we will determine if time savings are statistically significant when comparing manual use of EHRs to our NLP-derived visualization tool. We will recruit a clinician and/or analyze historical data and benchmark against existing manual processes.

Plan of Activities and Cost: Local development will cost $0. AWS's total costs for 8 weeks of development include $72.00 for SageMaker, $12.24 for EC2, and $0.076 for S3, totaling $84.31. GCP's costs would be $314.64 for Vertex AI, $79.20 for AppEngine, $0.066 for Cloud Storage, summing up to $393.9. **All team members have contributed a similar amount of effort to this proposal.**

Week	Activity	Members	Key Milestones	Description
March 3 - 6	Conduct EDA (Part 1)	All members	Deliverable 1: Initial Insights	Initial exploratory data analysis to understand dataset characteristics and data quality issues.
March 7 - 11	Conduct EDA (Part 2)	All members	Deliverable 2: EDA Report	Continue EDA focusing on identifying patterns, anomalies, and relationships between variables.
March 12 - 17	Data Cleaning	Kelly, Han	Deliverable 3: Cleaned Dataset	Clean the dataset based on insights from EDA, handling missing values, outliers, and errors.
March 19 - 24	EHR Structuring	Han, Ying	Midterm 1: Process EHR Data	Break down the electronic health records into structured sections for each record, preparing for efficient querying and analysis.
March 25 - 31	Handling Big Data with SQLite or Pickle	Ruby, Ying	Midterm 1: Data Split for ML	Implement SQLite or use Pickle files for managing big data. Split the dataset into training, testing, and validation datasets.
April 1 - 6	SpaCy Development for Structured EHR Data	Kelly, Stefan	Deliverable 4: NLP Model Development	Develop and train custom SpaCy models for NLP tasks specific to the structured EHR data.
April 7 - 11	Integration of Data Querying with NLP	Luke, Ruby	Final: NLP Integration with Data	Ensure efficient querying of structured data for use with NLP models, optimizing for performance.
April 12 - 19	App Development with Visualization	Stefan, Luke	Final: Visualization Web App	Develop a web application for visualizing EHR data and NLP analysis results, focusing on UI
April 20 - 21	Conduct Statistical Analysis and Impact	All members	Final: Final Analysis Report	Apply statistical methods and machine learning models for in-depth analysis of EHR data. Interpret results to derive insights.

References:

[1] Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016 Sep;23(5):1007-15. doi: 10.1093/jamia/ocv180. Epub 2016 Feb 5. PMID: 26911811; PMCID: PMC4997034. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4997034/

[2] Abudiyab NA, Alanazi AT. Visualization Techniques in Healthcare Applications: A Narrative Review. Cureus. 2022 Nov 11;14(11):e31355. doi: 10.7759/cureus.31355. PMID: 36514654; PMCID: PMC9741729. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9741729/

[3] Chishtie J, Bielska I, Barrera A, Marchand J, Imran M, Tirmizi S, Turcotte L, Munce S, Shepherd J, Senthinathan A, Cepoiu-Martin M, Irvine M, Babineau J, Abudiab S, Bjelica M, Collins C, Craven B, Guilcher S, Jeji T, Naraei P, Jaglal S. Interactive Visualization Applications in Population Health and Health Services Research: Systematic Scoping Review. J Med Internet Res 2022;24(2):e27534. https://www.jmir.org/2022/2/e27534

[4] Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, Osborn D, Hayes J, Stewart R, Downs J, Chapman W, Dutta R. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform. 2018 Dec;88:11-19. doi: 10.1016/j.jbi.2018.10.005. Epub 2018 Oct 24. PMID: 30368002; PMCID: PMC6986921. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6986921/

[5] Tian D, Jiang S, Zhang L, Lu X, Xu Y. The role of large language models in medical image processing: a narrative review. Quant Imaging Med Surg. 2024 Jan 3;14(1):1108-1121. doi: 10.21037/qims-23-892. Epub 2023 Nov 23. PMID: 38223123; PMCID: PMC10784029. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10784029/

[6] Shaikh TA, Dar TR, Sofi S. A data-centric artificial intelligent and extended reality technology in smart healthcare systems. Soc Netw Anal Min. 2022;12(1):122. doi: 10.1007/s13278-022-00888-7. Epub 2022 Sep 1. PMID: 36065420; PMCID: PMC9434088. https://link.springer.com/article/10.1007/s13278-022-00888-7

[7] Kop, R., Liao, K. P., Blei, D. M., Teh, Y. W., Bodenreider, O., Bentsen, B. G., Savova, G. K., Zeng, T., Aronson, A. R., Salleb-Aouissi, A., & Lehman, L. (2016, March 31). Utilizing uncoded consultation notes from Electronic Medical Records for Predictive Modeling of Colorectal Cancer. https://www.sciencedirect.com/science/article/abs/pii/S093336571530066X

[8] Robertson, A. R. R., Fernando, B., Morrison, Z., Kalra, D., & Sheikh, A. (2014). Structuring and coding in health care records: a qualitative analysis using diabetes as a case study. Journal of Innovation in Health Informatics, 22(2), 275-283. https://doi.org/10.14236/jhi.v22i2.90et

[9] Zoe Morrison, Bernard Fernando, Dipak Kalra, Kathrin Cresswell, Aziz Sheikh, National evaluation of the benefits and risks of greater structuring and coding of the electronic health record: exploratory qualitative investigation, Journal of the American Medical Informatics Association, Volume 21, Issue 3, May 2014, Pages 492–500, https://doi.org/10.1136/amiajnl-2013-001666

[10] J. Marc Overhage & David McCallie Jr. (2020, January 14). Physician Time Spent Using the Electronic Health Record During Outpatient Encounters. Annals of Internal Medicine. https://www.acpjournals.org/doi/10.7326/M18-3684

[11] Noor Abu-el-rub, Jay Urbain, George Kowalski, Kristen Osinski, Robert Spaniol, Mei Liu, Bradley Taylor, & Lemuel R. Waitman. (2022, May 23). Natural Language Processing for Enterprise-scale De-identification of Protected Health Information in Clinical Notes. AMIA Summits on Translational Science Proceedings 2022: 92–101. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9285160/

[12] Hannah T. Neprash, John F. Mulcahy, Dori A. Cross, et al. (2023, March 10). Association of Primary Care Visit Length with Potentially Inappropriate Prescribing. JAMA Health Forum. https://jamanetwork.com/journals/jama-health-forum/fullarticle/2802144

[13] Benavent, D., & Plasencia-Rodríguez, C. (2024). Redefining comorbidity understanding in rheumatoid arthritis through novel approaches using real-world data. Journal of Clinical Rheumatology and Immunology. https://www.explorationpub.com/Journals/emd/Article/100732

[14] Shah-Mohammadi, F., & Finkelstein, J. (2024). NLP-Assisted Differential Diagnosis of Chronic Obstructive Pulmonary Disease Exacerbation. Journal of Respiratory Medicine. https://pubmed.ncbi.nlm.nih.gov/38269877/

[15] Brian G. Arndt, John W. Beasley, Michelle D. Watkinson, Jonathan L. Temte, Wen-Jan Tuan, Christine A. Sinsky & Valerie J. Gilchrist. (2017). Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. The Annals of Family Medicine September 2017, 15 (5) 419-426. https://www.annfammed.org/content/15/5/419.full

[16] Grout, R., Gupta, R., Bryant, R., Elmahgoub, M. A., Li, Y., Irfanullah, K., Patel, R. F., Fawkes, J., & Inness, C. (2024). Predicting disease onset from electronic health records for population health management: a scalable and explainable Deep Learning approach. Frontiers in Artificial Intelligence. https://dx.doi.org/10.3389/frai.2023.1287541

[17] Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, Ravikumar KE, Wu ST, Kullo IJ, Chute CG (2013). An information extraction framework for cohort identification using electronic health records. AMIA. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3845757/

[18] Adarsh Kumar and ANJALI GOND. NATURAL LANGUAGE PROCESSING: HEALTHCARE ACHIEVING BENEFITS VIA NLP . ScienceOpen Preprints. 2023. https://doi.org/10.1234/jhi.2021.12345

[19] Suryanarayanan P, Epstein EA, Malvankar A, Lewis BL, DeGenaro L, Liang JJ, Tsou CH, Pathak D. Timely and Efficient AI Insights on EHR: System Design. AMIA Annu Symp Proc. 2021 Jan 25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075522/