Patrick Meyers, M.D.

M.P.H. Candidate, Epidemiology Track
patrick.m.meyers.2@vanderbilt.edu

Practicum Site: V.A. Tennessee Valley Healthcare System

Practicum Site Supervisor: Steve Deppen, Ph.D.

Deciphering Smoke Signals: Advanced Text Analysis and Visualization of Smoking-Related Health Factors

Introduction: Assessing clinical factors is vital for population risk modeling, particularly for tobacco use and lung cancer. Within the Veteran Affairs (VA) Informatics and Computing Infrastructure (VINCI), Health Factors (HFs) include semi-structured patient data and clinical observations. Free text in HF Content Text offers insights but poses challenges due to sparseness, inconsistencies, and abbreviations. This study aims to improve the accuracy of tobacco-related data extraction from semi-structured HFs within the VA database using novel Natural Language Processing (NLP) algorithms.

Methods: We accessed HF data from the VA's common data model, VINCI OMOP, database. From an initial cohort of over 16 million veterans, we identified those diagnosed with lung cancer and screened via computed tomography using ICD9/10 codes. The HF comment dataset format was standardized for token identification and processing. We searched for variations on tobacco-related concepts (e.g., "cigarette," "pack per day," "current smoker") and analyzed token frequency within Content Text to construct similarity matrices. This included string searches, identifying variants and misspellings, context-specific usage analysis, and numeric searches (e.g., "3 packs per day until 2022"). Using a probabilistic approach, we applied edit distance and a novel "error distance" metric to correct misspellings, allowing standard NLP models without extensive preprocessing. Based on these findings, we customized HF categories for enhanced visualization and analysis through heatmapping.

Results: Our correction algorithm reduced data noise by clustering HF categories around common lexicon concepts. For example, our substring search for "cigarette(s)" identified 228,228 correct spellings and 84,132 potential misspellings/abbreviations. Applying our probabilistic algorithm resulted in 246,371 post-correction assignments, a 7.9% increase.

Conclusions: We developed a foundational pipeline for handling inconsistencies in complex datasets, enabling consistent data extraction and standardization by focusing on a cohort with dense and accurate data. We plan to expand this to the general VA population and make results available to the VA research community.