Knowledge-Guided Data Analytics for Determining Causal Chains of Death
Ying Sha*, Hang Wu*, May D. Wang

Abstract—Mortality data is one of the largest public health data that indicates major healthcare challenges for policy makers and researchers. Despite the importance of mortality data, the quality of them remains problematic because of the difficulty in determining a causal chain of death with limited information in a short time for physicians and other healthcare practitioners, especially for those who lack relevant experiences. Only a few research studies exist for developing data analytics to assist physicians in determining underlying causes of death, and no studies exist for identifying causal chains of death with data-driven approaches. Deriving insights from natural language processing, our study develops an advanced recurrent neural network for suggesting possible causal chains of death, based on multiple diagnostic codes associated with patients’ last hospital admission from a private dataset of Michigan. We achieve a Bilingual Evaluation Understudy (BLEU) score of 9.23.

I. INTRODUCTION

Accurate death reporting is essential for public health agencies, such as Center for Disease and Control (CDC), to understand disease epidemics and to identify public-health challenges in the U.S and other countries. Currently in the U.S., cases of death are primarily reported in death certificates, a state-wide standardized form including basic demographics, Underlying Causes of Death (UCOD), and a causal sequence of medical conditions leading to death [1]. Oftentimes, certifying physicians fill out the certificates based on medical records, clinical notes, and their familiarity with deceased patients. However, inadequate training of death reporting and unfamiliarity with patients in unexpected circumstances have resulted in subjective and inaccurate reporting [2].

One promising solution to improve the quality of mortality data is to facilitate death reporting with data-driven methods. Previously, only descriptive analysis of mortality data and preliminary work for identifying UCOD exist [3]–[5]. The recent availability of mortality data with corresponding medical history enables us to develop data-driven methods for identifying not only UCOD but also a complete causal chain of death. In addition, because existing mortality data contain human errors, data-driven models trained solely on mortality data may still replicate those errors. Therefore, to further reduce errors, we also need to incorporate medical domain knowledge to guide the learning process of models.

The study develops an advanced data-driven model that generates a causal sequence of death via state-of-the-art Natural Language Processing (NLP) techniques, and further improve the accuracy by incorporating medical knowledge.

II. METHOD AND RESULTS

For generating sequences of medical conditions leading to death, we have a dataset of n death cases as \(\{(X_i, Y_i)\}_{i=1}^{n}\). For the \(i\) th sample, \(X_i\) represents all relevant medical conditions of an individual, and \(Y_i\) is the target sequence of length \(L_i\) for conditions leading to death: \(Y_i = (Y_{i[0]}, ..., Y_{i[L_i]})\).

Sequence Generation Model for \(X \rightarrow Y\): We represent the input \(X_i\) by aggregating the low-dimensional embedding for each input condition \(X_{i[k]}\) via methods such as mean pooling. We then learn probability \(p(Y_{i[j+1]}|Y_{i[j]}, ..., Y_{i[1]}, X)\) using the Transformer model.

Guidance from Medical Ontology: We also impose additional ontology constraints for generating causal sequences. For conditions that could not be caused by a current condition, we set their probability, \(p(Y_{i[j+1]}|Y_{i[j]}, ..., Y_{i[1]}, X)\), to be zero. We then apply beam search to generate the most probable sequence.

Our method achieved a BLEU score of 9.23 and 9.22 with and without ontology respectively. We reasoned that as most causal sequences in the data conform to ontology, additional ontology constraints did not improve results significantly.

REFERENCES


The work was supported by the U.S. Department of Health and Human Services Centers for Disease Control and Prevention under Award HSIDS200015F652550B, the NIH National Center for Advancing Translational Sciences under Award UL1TR000454, and the National Science Foundation Award NSF1651360.

*The authors contributed to the work equally.

Y. Sha is with School of Biology, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: ysha8@gatech.edu).

H. Wu is with Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: hangwu@gatech.edu).

M. D. Wang is with Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332 USA (phone: 404-385-2954; e-mail: maywang@gatech.edu).