NLP Detection of Mortality Information from Publicly Available Data Using Deep Learning Modeling

Mohammed A Al-Garadi[1], Michael E. Matheny[1,2], Rishi J Desai[3], Mirza S. Khan[1], Shirley V. Wang[3], Judith C. Maro[4], Candace C. Fuller[4], Kueiyu Joshua Lin[3,5], José J. Hernández-Muñoz[6], Aida Kuzucan[6], Xi Wang[6], Jill M. Whitaker[1], Jessica A. Deere[1], Michael F. McLemore[1], Dax Westerman[1], Josh Osmanski[1], Ruth Reeves[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA, [2]Geriatrics Research Education and Clinical Care Service & VINCI, Tennessee Valley Healthcare System VA, Nashville, TN, USA, [3]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA, [4]Harvard Pilgrim Health Care Institute, Harvard Medical School Department of Population Medicine, Boston, MA, USA, [5]Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA, [6]Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD

OBJECTIVE

To design and implement a NLP pipeline that can identify and extract the date of death, date of birth, and name of the deceased from multiple publicly available sources.

BACKGROUND

Mortality plays a crucial role in medical product outcome assessments. However, obtaining accurate and timely data on the date of death can be challenging. Additional information that enables linking these data to patient records, such as decedent and family member names, locations, and dates of birth, needs to be determined. There are potentially useful publicly available data sources with narrative text that could enhance the validity of death ascertainment, and advancements in NLP and machine learning techniques may show promise in extracting relevant death information.

METHODS

- Nationally, publicly available death information was collected from obituary records, Everloved, Tribute Archive, and GoFundMe from 2015 to 2021.
- Annotation guidelines were developed to identify information pertaining to the death events and identifiers in each source.
- Elements to be annotated were the decedent's name, date of birth, date of death, and any irrelevant dates.
- Three annotators were trained iteratively until they attained an acceptable agreement rate(>80%).
- 4,200 sampled documents, stratified by source, were annotated.
- Annotated documents were divided into training, testing, and validation datasets at proportions of 70%, 20%, and 10%, respectively.
- A variety of deep learning transformer-based language model approaches (BERT, RoBERTa, ALBERT, and BERTweet) were tested and compared as the candidate NLP pipeline.
- Meta-parameters and model tuning were completed through standard processes.
- The performance of the developed model was evaluated using sensitivity, positive predictive value (PPV), and F1-score (the harmonic mean of sensitivity and PPV).
- The RoBERTa model yielded the best performance.
- An overview of the study design workflow is shown in Figure 1 below.

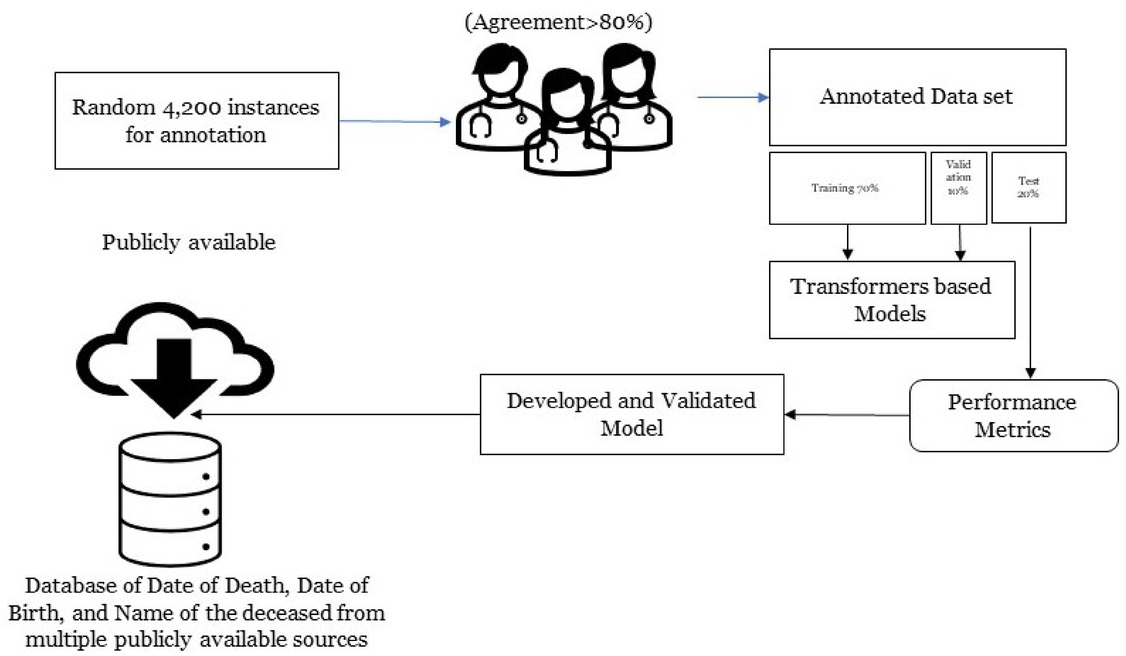

Figure 1. Study Design Workflow

RESULTS

The model achieved the best performance on the test set, as shown in Table 1, with an F1-score (strict), sensitivity, and PPV. We then applied this best-performing model to a large dataset collected from publicly available sources, and Figure 2 presents the number of state counts for death events.

	Precision (PPV)	Recall (sensitivity)	F1-score
Decedent Name	0.86	0.84	0.85
Date of Death	0.87	0.91	0.89
Date of Birth	0.95	0.93	0.94

Table 1. Performance metrics on test set

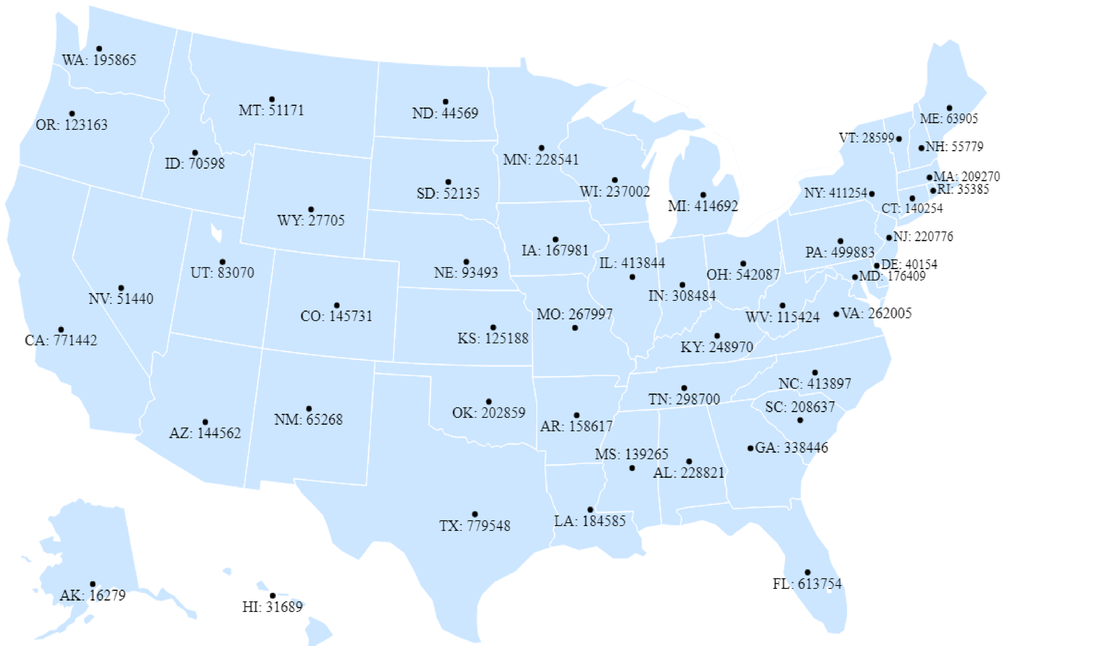

Figure 2: Per State Counts of Death Events Including All Fields (First Name, Last Name, Date of Birth, Date of Death, and State) Extracted from Public Data 2015-2021

CONCLUSION

The NLP model has demonstrated high accuracy in extracting entities related to death events from publicly available data in the independent testing set. The developed algorithm can be applied to large-scale datasets from multiple publicly available sources, thereby improving the timeliness and validity of death ascertainment in clinical research. However, it should be noted that data obtained from public sources may still contain gaps and inaccuracies. The reliance on these incomplete data sources can potentially introduce limitations in accurately identifying and extracting the date of death, date of birth, and name of the deceased. Overall, the NLP model's ability to accurately extract death-related entities from extensive and varied datasets has the potential to revolutionize death ascertainment in clinical research, enabling a more efficient approach that enhances the efficiency and accuracy of death ascertainment.

ACKNOWLEDGEMENTS/DISCLOSURES

This project was supported by Task Order 75F40119F19002 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA).CCF and JM are employees of the Harvard Pilgrim Health Care Institute. MAA, MEM, MSK, KJL, JMW, JAD, MFM, DX, JO, and RR are employees of Vanderbilt University Medical Center. RD and SVW are employees of Brigham and Women's Hospital, and JJH, AK, and XW are employees of the FDA. The views expressed in this presentation represent those of the presenters and do not necessarily represent the official views of the U.S. FDA.