Automated Extraction of Mortality Information from Publicly Available Sources Using Language Models

Mohammed Al-Garadi[1]; Michele LeNoue-Newton[1]; Michael E. Matheny[1,2]; Melissa McPheeters[3]; Jill M. Whitaker[1]; Jessica A. Deere[1]; Michael F. McLemore[1]; Dax Westerman[1]; Mirza S. Khan[1]; José J. Hernández-Muñoz[4]; Xi Wang[4]; Aida Kuzucan [4]; Rishi J. Desai[5]; Ruth Reeves[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [2]Geriatrics Research Education and Clinical Care Service, Tennessee Valley Healthcare System VA, Nashville, TN, USA; [3] RTI International, Research Triangle Park, NC, USA; [4] Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; [5] Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

Abstract

Background: Mortality is a critical variable in healthcare research, but inconsistencies in the availability of death date and cause of death (CoD) information limit the ability to monitor medical product safety and effectiveness.

Objective: To develop scalable approaches using natural language processing (NLP) and large language models (LLM) for the extraction of mortality information from publicly available online data sources, including social media platforms, crowdfunding websites, and online obituaries.

Methods. Data were collected from public posts on X (formerly Twitter), GoFundMe campaigns, memorial websites (EverLoved.com and TributeArchive.com), and online obituaries from 2015 to 2022. We developed a natural language processing (NLP) pipeline using transformer-based models to extract key mortality information such as decedent names, dates of birth, and dates of death. We then employed a few-shot learning (FSL) approach with large language models (LLMs) to identify primary and secondary causes of death. Model performance was assessed using precision, recall, F1-score, and accuracy metrics, with human-annotated labels serving as the reference standard for the transformer-based model and a human adjudicator blinded to labeling source for the FSL model reference standard.

Results: The best-performing model obtained a micro-averaged F1-score of 0.88 (95% CI, 0.86-0.90) in extracting mortality information. The FSL-LLM approach demonstrated high accuracy in identifying primary CoD across various online sources. For GoFundMe, the FSL-LLM achieved 95.9% accuracy for primary cause identification, compared to 97.9% for human annotators. In obituaries, FSL-LLM accuracy was 96.5% for primary causes, while human accuracy was 99.0%. For memorial websites, FSL-LLM achieved 98.0% accuracy for primary causes, with human accuracy at 99.5%.

Conclusions: These findings highlight the potential of leveraging advanced NLP techniques and publicly available data to enhance the timeliness, comprehensiveness, and granularity of mortality surveillance.

Funding statement: This project was supported by Task Order 75F40123F19010 under Master Agreement 75F40119D10037 from the US Food and Drug Administration (FDA). FDA coauthors reviewed the study protocol, statistical analysis plan, and the manuscript for scientific accuracy and clarity of presentation. Representatives of the FDA reviewed a draft of the manuscript for the presence of confidential information and accuracy regarding the statement of any FDA policy. The views expressed are those of the authors and not necessarily those of the US FDA.

Introduction

Mortality is a critical variable in healthcare research, and all-cause mortality is one of the most studied endpoints[1-4]. Accurate identification of the fact, timing, and cause of death (CoD) is essential for various types of medical research, including clinical trials, observational studies, and post-marketing surveillance programs such as the US FDA Sentinel System [5-8].

A recent report identified limitations in the availability of date and CoD information as a major cause for study insufficiency when considering the use of the Sentinel Active Risk Identification and Analysis (ARIA) system to address regulatory questions[9]. Failing to identify deaths may result in substantial underestimation of mortality outcomes related to medical products, so efforts to identify additional data sources to supplement current systems have far-reaching consequences. Vital statistics data, collected in the United States on death certificates and submitted at the state level, are the 'gold standard' for mortality information. Depending on state law (and sometimes the manner of death), death certificates may be completed by coroners, medical examiners, or physicians in the healthcare system. Once it is entered into the state reporting system, information from the death certificate is sent to the Centers for Disease Control and Prevention, which codes the underlying CoD and incorporates the information into national data. However, there is typically a lag of at least nine months after the end of a calendar year before vital statistics data become available, with the National Death Index (NDI) currently providing data for the calendar year two years prior.

There are other data sources for death information, including claims databases, and medical records, but each of these sources has limitations[10, 11]. Claims databases may underrepresent uninsured populations, while medical records often lack standardization between healthcare

providers, complicating data aggregation and comparison[12]. In most claims databases, death-related information, including occurrence and CoD, is often incomplete or not directly recorded. Similarly, healthcare system-based data sources, such as electronic health records (EHRs), frequently lack comprehensive mortality data, particularly when patients are not under the care of the healthcare system at the time of death. This incomplete ascertainment of death information poses significant challenges for researchers and clinicians relying on these data sources for epidemiological studies, outcomes research, and healthcare quality assessments.

The rise in the use of social media has introduced potential sources of mortality-related information, including online obituaries and the sharing of death information in social networks through Twitter and other channels. There is growing precedent for the use of social media in public health and other health-related research, and user posts have been used to track illnesses[13-17], measure behavioral risk factors[18-22], localize diseases geographically[21, 23, 24], and analyze symptoms and medication usage[25-30]. Nonetheless, a key challenge inherent in social media data for mortality information is the capacity to extract the date and CoD at scale and with replicable methods. These social media sources offer potential advantages in timeliness, context, and coverage compared to traditional mortality data sources.

In this study, we sought to develop a set of NLP tools to extract both the fact and CoD from publicly available records and to assess the relative information density of illness and death information within these records. This type of data, when combined with other sources, could improve ascertainment in downstream studies that require the use of the facts and causes of mortality among EHR and claims data analyses. A successful NLP pipeline, described here, was developed to identify, and extract information from publicly available sources.

The innovative approach leverages publicly available data to provide timely insights into population health trends, potentially enabling faster responses to emerging health threats. By linking social media and obituary data with patient records, the system could offer a more comprehensive view of health outcomes and risk factors, as well as system evaluation.

Methods

Our study developed and evaluated natural language processing (NLP) techniques to extract mortality information from publicly available online sources. The research methodology included data collection, NLP model development, and performance assessment. Given its focus on public health surveillance using open-source information, this research qualifies for exemption from FDA and Vanderbilt University Medical Center (VUMC) Institutional Review Board (IRB) oversight.

Data Sources and Study Cohort

Data were collected from X (formerly Twitter), GoFundMe, and online obituaries (Obituaries.com), memorial websites (EverLoved.com, and TributeArchive.com) between the years 2015 and 2022 in the United States, which are publicly available and aggregated for research purposes in accordance with fair use. The online obituary sources provide more robust meta-data for determining inclusion criteria than records obtained from Twitter or GoFundMe. Our collection methods, therefore, differed by source.

Our search of X (Twitter) included around 50 derived keywords such as ‘death,’ ‘expired,’ and ‘deceased’ and used Twitter’s official research Application programming interface (API), yielding approximately 40 million tweets. Using similar keywords (provided in the appendix), we identified and retrieved posts from GoFundMe and from memorial websites (EverLoved.com and TributeArchive.com) containing mortality-related information. For Obituaries.com, we acquired reports from 2015 to 2022, which contained millions of records (see Table 4 for final counts across all sources). For the obituary data sources, we collected structured metadata (e.g., first name, last name, date of death, date of birth, location) and extracted accompanying textual information. NLP techniques were subsequently used on this textual content to supplement or complete missing or incomplete metadata fields. This approach allowed us to maximize the information extracted from each obituary, enhancing the overall quality and completeness of our dataset.

Reference Standard

To create a gold-standard reference for training and testing of the models, we developed annotation using the deceased's name, names of related individuals, dates relevant to the

deceased (death date, birth date, and other dates), and the CoD. First, annotators were instructed to accurately classify names with post-nominals, avoid names in Twitter handles, and use specific relationship attributes for related persons (e.g., spouse, sibling, child). Second, annotated dates included exact, partial, or relative expressions, with clear distinctions for death and birth dates. Third, the causes of death were annotated with attributes indicating assertion (positive, negative, uncertain) and patient vs. not-patient (reference to the deceased or to someone else). Finally, if no relevant data was found in a document, annotators classified the document as "No Data."

A corpus of 4,200 notes, 1,050 from each of the data sources, was randomly sampled. We split the 4,200 annotated posts from all data sources into training (70%), testing (20%), and validation (10%) datasets. The training data contained 81,082 tokens (words), and the test data contained 27,834 tokens (words).

Annotation

The data were annotated by three trained nurse annotators who closely followed a detailed annotation guideline, categorizing each post into first and last names, dates of birth, dates of death, and CoD. The training was initiated using records from Twitter, GoFundMe, memorial website (EverLoved/TributeArchives), and Obituaries.com, with all three annotators independently labeling the same documents in rounds of 15 documents from each source (n=45).

After each training round of annotation was completed by all three members, agreement rates were computed with F-measure (harmonic balance of precision and recall) between pairs of annotation sets. The overall inter-annotator agreement (IAA) was evaluated using Cohen's kappa[31], and annotators were required to achieve an overall IAA threshold of 0.80 on the training set before proceeding with the full annotation process. When the targeted threshold was not met, the annotation team performed a consensus annotation over each document in a given annotation round, discussing their differences, and updating or clarifying the annotation guidelines. At the point that the threshold was met, annotators were instructed to proceed to conduct independent annotation, with each member being assigned 1000 documents, 250 per source. An additional assessment of annotation reliability was assessed by randomly assigning 25

documents per source (n=100) to all three annotation team members and conducting an independent IAA for that subset.

Information density assessment

Annotations completed by nurse annotators were used to assess the information density of online sources, such as social media platforms like Twitter, to determine if they contained sufficient details for reliable patient linkage and augmentation of date of death in healthcare systems. Sources with inadequate information were excluded from further analysis.

Assessment of CoD availability was completed using the 600 document annotations utilized in the few-shot learning validation with verification by the nurse adjudicator of causes of death mentioned within the post.

NLP Development and Implementation

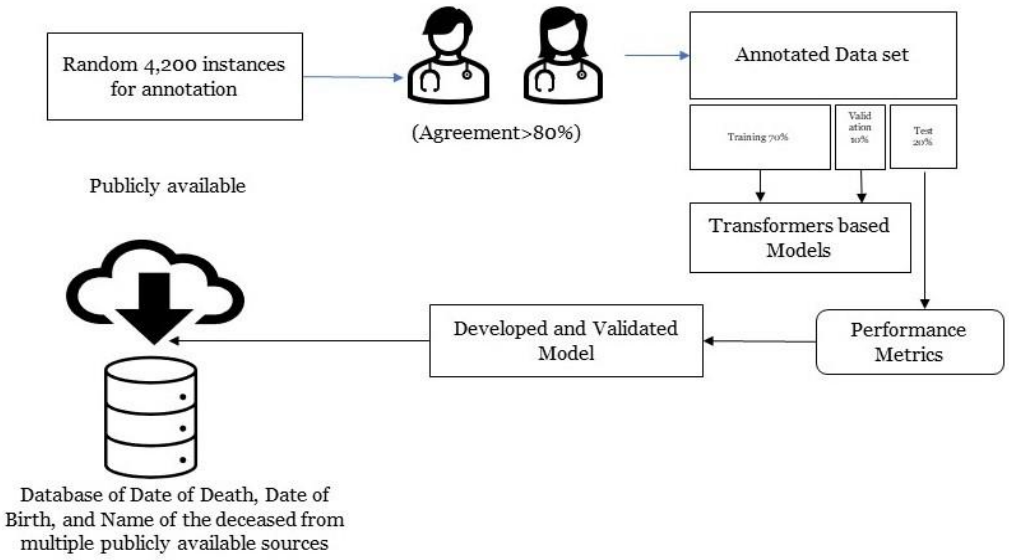


Figure 1: Workflow of the NLP pipeline development and evaluation process.

We developed in parallel two NLP tools for information extraction from the previously described social media sources. First, we adapted four deep learning transformer-based methods including BERT (Bidirectional Encoder Representations from Transformers)[32], RoBERTa (Robustly Optimized BERT Pretraining Approach)[33], ALBERT (A Lite BERT)[34], BERTweet[35] to extract the decedent's name, date of birth, and date of death, and to exclude any irrelevant dates. The technical pipeline overview for the transformer-based model is illustrated in Figure 1.

To identify CoD, we used a few-shot learning approach to leverage an open-source Large Language Model (LLM) (Figure 2). The decision not to use transformer models for this portion of the information extraction task was predicated on the fact that predicting a concept like CoD relies on a robust understanding of both the extracted cause and the context. We used an iterative prompting approach with examples and guidelines to define both primary and secondary CoD using high-quality annotation labels where at least two annotators agreed on the assigned label for prompting.

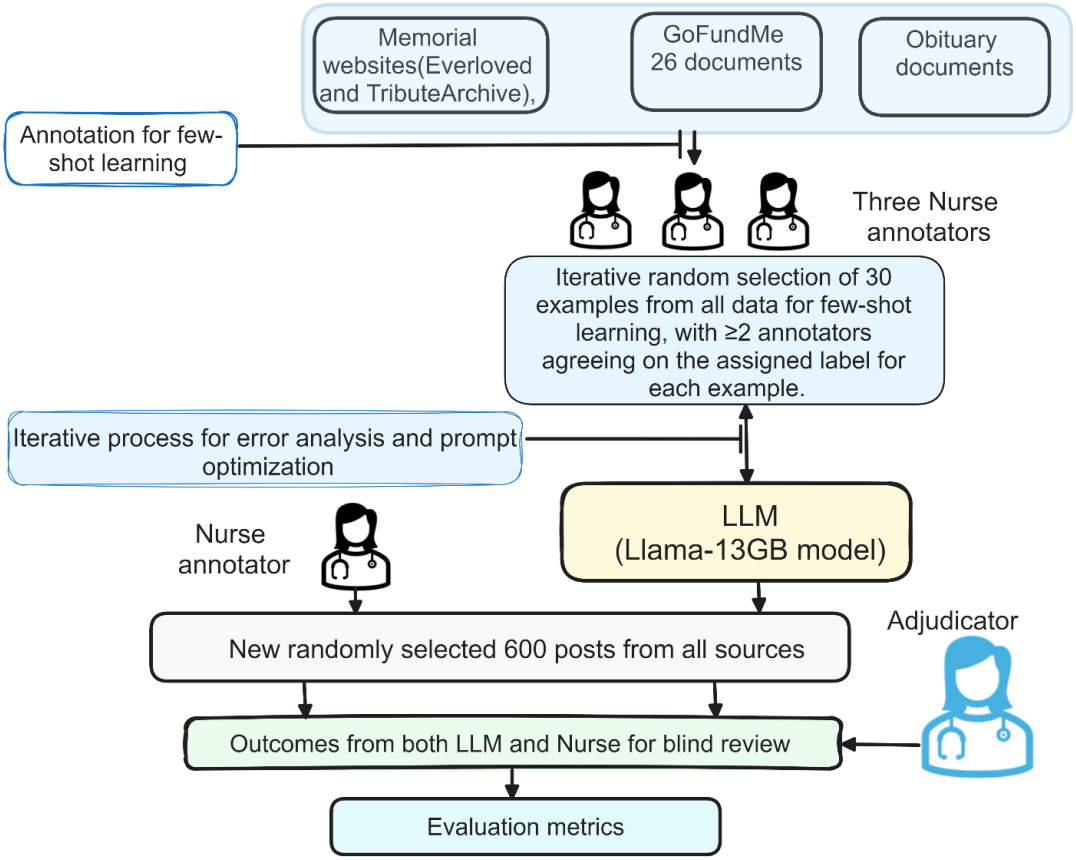


Figure 2: Workflow for Few-Shot Learning and Evaluation

For example, in the post, "Jane Smith died from a severe infection following surgery. She also had diabetes and hypertension, which contributed to her deteriorating health," the main cause would be noted as "Severe Infection Following Surgery," and the secondary causes as "Diabetes" and "Hypertension." The initial prompt engineering stage ensures the LLM properly formulates the type of information to extract/predict. We utilized the LLaMA model, a 13 GB language model developed by Facebook AI Research, for processing the data[36]. LLaMA, which stands for "Large Language Model Meta AI," is a foundational language model that exhibits remarkable

performance across various NLP tasks[36]. A smaller version, such as the 13GB variant, of the Llama model can be run locally on a machine with sufficient computational resources, making it more accessible and efficient for certain applications. We started with 30 randomly selected examples from the manually annotated data (training split) for prompting and 30 for assessing the model's performance, where at least two annotators agreed on the annotated instance. The prompts and assessment examples went through several iterations of LLM refinement, totaling four iterations, until the identified CoD was correct in most cases across the various assessment sets. The accuracy during the prompting process was evaluated qualitatively to understand where the model performed correctly and where it made errors.

During the testing phase, we evaluated our final prompting design on a new set of 600 examples. The evaluation process involved three steps:

- A nurse annotator identified the CoD in these examples following the provided guidelines.
- Simultaneously, our refined language model (LLM) automatically extracted the CoD from the same 600 examples.
- A second trained nurse, acting as an adjudicator, independently reviewed both sets of results. This review ensured that the annotations adhered to the guidelines and that the primary CoD was accurately identified in each case.

Following the evaluation, we analyzed the results by determining the accuracy of primary CoD and additional identified causes from both the human annotator and the LLM per the adjudication. We then determined true positives, true negatives, false positives, and false negatives to compute relevant statistical metrics, allowing us to assess the accuracy and effectiveness of both human and automated CoD identification methods.

Statistical metrics for model evaluation

For the transformer-based model evaluation, we calculated sensitivity, positive predictive value, and the F1-Score to evaluate model performance, and we computed micro averages for each to compute the average metric for a global measure of performance. We used bootstrap to calculate the confidence interval by resampling the test set, calculating the required metrics for each

resample, and using percentiles of these metrics to form the confidence interval. We also assessed the information density of online posts from each data source to determine their adequacy for reliable patient linkage and mortality information augmentation in healthcare systems.

For the LLM CoD information extraction module, we calculated the F-score, accuracy, precision, and recall for the primary CoD. However, for all potential CoD, due to the variation in the number of causes and the challenge of measuring performance using traditional NLP metrics, we asked the adjudicator to qualitatively assess the number of cases where the LLM correctly identified all the contributing CoD mentioned in the posts and to determine if the LLM and human annotators correctly identified the primary CoD. As such, we addressed the ambiguity of using a static output for each social media post. The adjudicator focused on whether the predicted CoD was accurate, regardless of whether it was explicitly mentioned in the post or inferred from the overall understanding of the post.

Phrases classified as "No CoD" indicated no specific medical CoD. These included "brief/sudden/extended/chronic illness," "unexpected" or "sudden death/passing," "natural causes," "no mention" of cause, "none," and "unknown/unspecified reasons/cause." Posts containing only such phrases were categorized as "No CoD". Correct identification of these cases by the language model counted as true negatives in the CoD identification process. This approach ensured that vague or non-medical descriptions were not misclassified as specific causes of death.

Application of NLP and Final Data Collection

The final phase of our study involved compiling the extracted data into a comprehensive dataset ready for analysis. We applied a series of cleaning filters and NLP techniques to ensure that only documents with reliable mortality-related information were included. This thorough process resulted in a dataset, as detailed in Table 4 of the Results section. This dataset, enriched with mortality information from various sources, is poised to serve as a valuable resource for public health surveillance and future research efforts.

Results

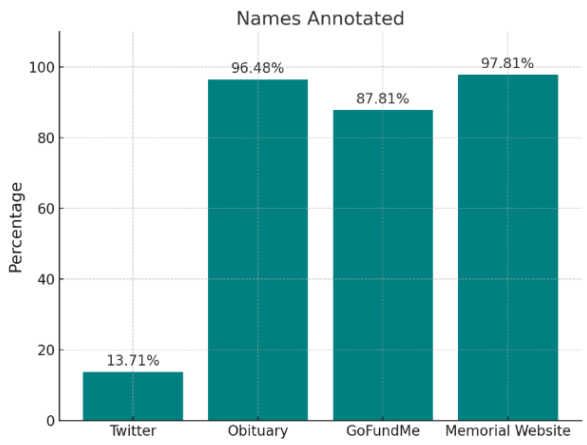
Annotation inter-annotator agreement

Overall IAA with respect to GoFundMe achieved a 92.5% agreement rate in the final iteration while the IAA within Twitter data maintained an 85.7% agreement rate after 3 rounds of assessment. IAA achieved within data sourced from the obituary websites demonstrated strong overall agreement, with a 91.5% agreement rate after the third round of assessment.

Information Density of Patient Identification in Social Media

Analysis of information density in online posts revealed varying levels of utility for patient linkage and mortality information augmentation in healthcare systems. Among the examined sources, three demonstrated high information density for annotated names, ranging from 87.81% to 97.81% (Figure 3). These sources provided sufficient detail for reliable patient identification and mortality data enhancement, whereas X (formerly Twitter) had low information density for patient identification and was excluded from subsequent analysis.

Figure 3: Information Density of Annotated Names Across Different Online Public Sources



Extracting Mortality Information Results

When evaluated on the manually annotated test data, the RoBERTa model demonstrated the highest overall performance for extracting the inquired information, with a micro-averaged F1-score of 0.88 (95% CI, 0.86-0.90) (Table 1). This model outperformed others in all three tasks, achieving an F1-score of 0.85 (95% CI, 0.84-0.86) for Decedent Name, 0.89 (95% CI, 0.88-0.90) for Date of Death, and 0.94 (95% CI, 0.92-0.94) for Date of Birth

The ALBERT model attained an F1-score of 0.87 (95% CI, 0.86-0.89) for Date of Death, and F1-scores of 0.83 (95% CI, 0.82-0.86) for Decedent Name, and 0.91 (95% CI, 0.90-0.93) for Date of

Birth. BERTweet achieved an F1-score of 0.90 (95% CI, 0.89-0.91) for Date of Birth, with scores of 0.82 (95% CI, 0.81-0.83) and 0.85 (95% CI, 0.84-0.86) for Decedent Name and

Date of Death, respectively. BERT's performance was marginally lower, with F1-scores of 0.81 (95% CI, 0.80-0.83), 0.84 (95% CI, 0.82-0.86), and 0.89 (95% CI, 0.88-0.90) for Decedent Name, Date of Death, and Date of Birth, respectively.

Table 1:Performance comparison of finetuned transformer models (on named entity recognition tasks (Decedent Name, Date of Death, and Date of Birth)

	RoBERTa			BERT			ALBERT			BERTweet		
	Precision (PPV)	*Recall (Sensitivity)*	*F1-score (95% CI)*	*Precision (PPV)*	*Recall (Sensitivity)*	*F1-score (95% CI)*	*Precision (PPV)*	*Recall (Sensitivity)*	*F1-score (95% CI)*	*Precision (PPV)*	*Recall (Sensitivity)*	*F1-score (95% CI)*
Decedent Name	0.86	0.84	0.85 (0.84-0.86)	0.81	0.80	0.81 (0.80-0.83)	0.84	0.82	0.83 (0.82-0.86)	0.83	0.81	0.82 (0.81-0.83)
Date of Death	0.87	0.91	0.89 (0.88-0.90)	0.82	0.87	0.84 (0.82-0.86)	0.86	0.88	0.87 (0.86-0.89)	0.86	0.84	0.85 (0.84-0.86)
Date of Birth	0.95	0.93	0.94 (0.92-0.94)	0.90	0.89	0.89 (0.88-0.90)	0.92	0.91	0.91 (0.90-0.93)	0.91	0.89	0.90 (0.89-0.91)
Micro Avg	0.88	0.88	0.88 (0.86-0.90)	0.83	0.85	0.84 (0.82-0.86)	0.85	0.87	0.86 (0.84-0.86)	0.84	0.85	0.84 (0.82-0.86)

The accuracy of the primary CoD identification and all CoD identification for both FSL-LLM and human identification are as follows: For GoFundMe, FSL-LLM achieved an accuracy of 95.9% for primary cause and 56.4% for all causes, while human accuracy was 97.9% for primary cause and 93.3% for all causes. For Obituary, FSL-LLM accuracy was 96.5% for primary and 96.0% for all causes, with human accuracy at 99.0% for primary cause and 98.5% for all causes. For Memorial websites, FSL-LLM accuracy was 98.0% for primary causes and 93.5% for all causes, whereas human accuracy was 99.5% for primary causes and 99.0% for all causes (Table 2).

Table 2: Accuracy of CoD Identification (FSL-LLM vs Human)

Source	*FSL-LLM: Primary CoD Identification Accuracy (%)*	*Human: Primary CoD Identification Accuracy (%)*	*LLM: All CoD Identification Accuracy (%)*	*Human: All CoD Identification Accuracy (%)*
GoFundMe	95.9	97.9	56.4	93.3
Obituary	96.5	99.0	96.0	98.5
Memorial websites	98.0	99.5	93.5	99.0

The precision, recall, and F-score for the LLM's vs human detection of the primary CoD were computed for each source. The metrics are presented below (Table 3).

*Table 3: Precision, Recall, And F-Score for FSL- LLM Vs human (**Primary CoD**)*

Sources	FSL-LLM Precision	*FSL-LLM Recall*	FSL-LLM F1-Score	*Human Precision*	*Human Recall*	*Human F1-Score*
GoFundMe	0.97	0.95	0.96	1.00	0.98	0.99
Obituary	0.61	1.00	0.76	1.00	0.82	0.90
Memorial Websites	0.94	0.98	0.96	1.00	0.98	0.99

Assessment of CoD Availability and Classification Error Analysis across Social Media Sources

As shown in Figure 4, social media sources varied significantly in the availability of CoD information, with Obituaries having a very low density of CoD (6%), Everloved posts primarily having a single potential CoD, GoFundMe being the richest source of potential CoD information with 43% containing a single CoD and 50% containing multiple potential CoD mentions, though not all potential causes of death/conditions pertained to the deceased subject of the social media posting.

Figure 4: Availability of CoD within Social Media Posts

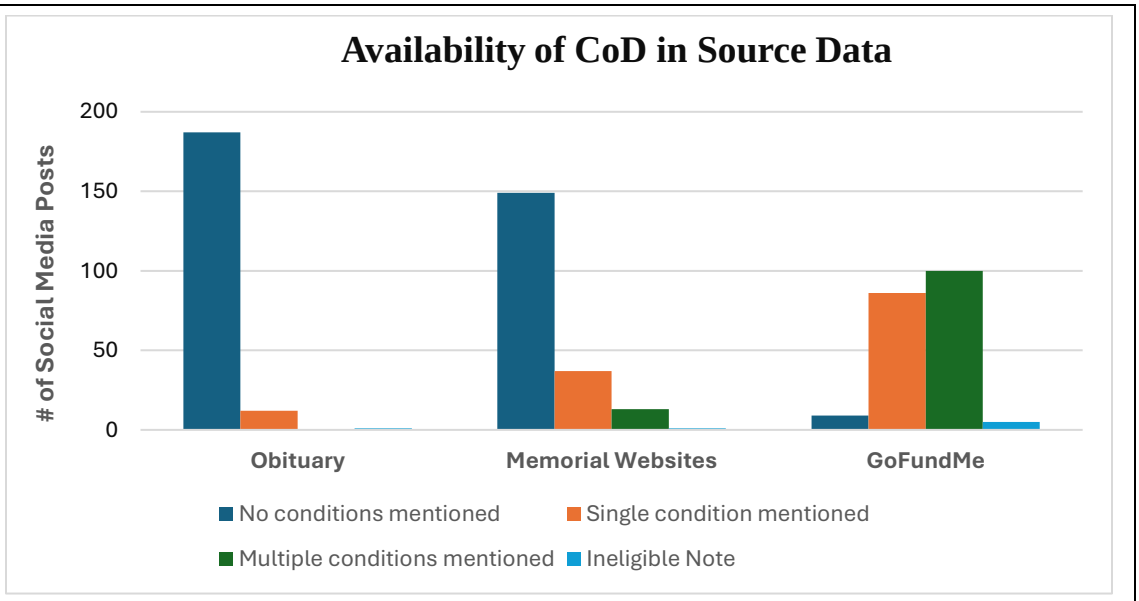
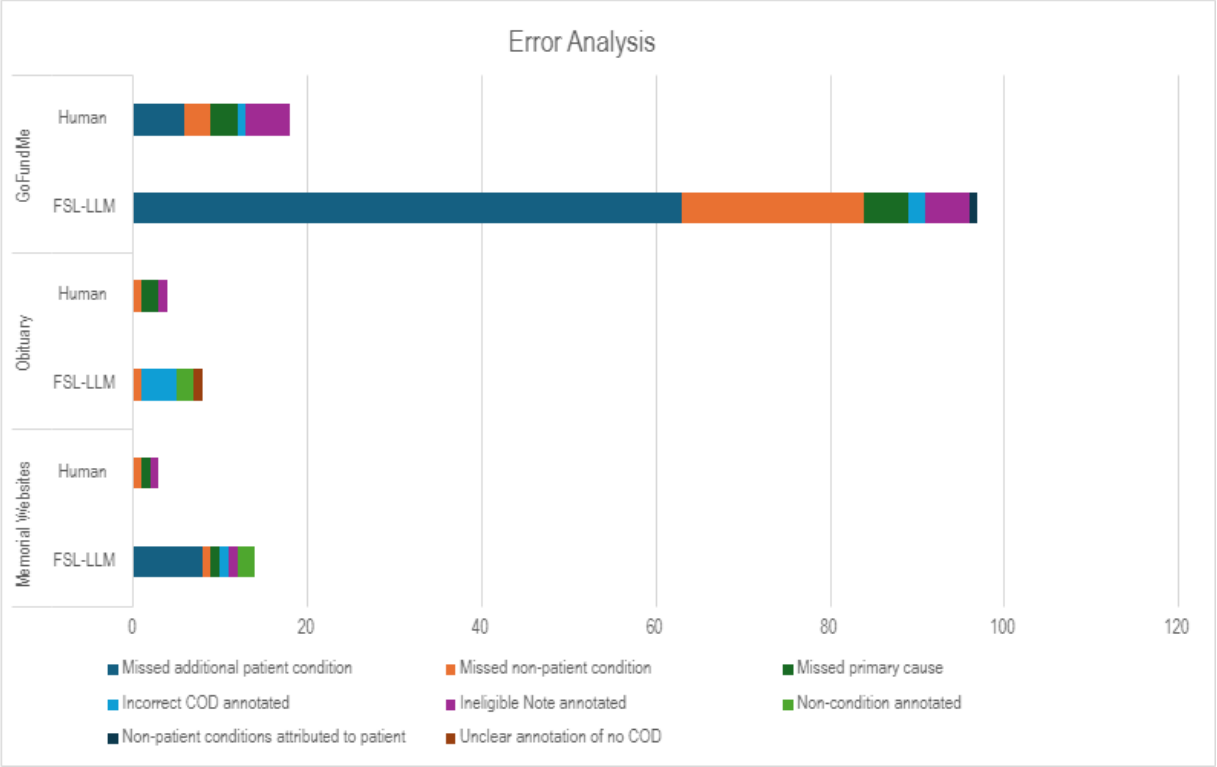


Figure 4: Availability of CoD in social media: Social media posts were annotated for potential causes of death using the annotation guidelines outlined in Appendix A. Social media sources were assessed for information density related to cause of death information by categorizing a post as having either no medical conditions mentioned, a single medical condition mentioned, or having multiple unique medical conditions within the post.

The distribution and comparison of errors made by LLM and human annotators across the test dataset is illustrated in Figure 5. Each post may have multiple errors or error types. The analysis focuses on discrepancies in both primary and additional CoD annotations, providing a detailed breakdown of error types and frequencies. The errors include missed additional patient conditions, missed non-patient conditions, missed primary causes, incorrect CoD annotated, ineligible notes annotated, non-patient conditions attributed to the patient, and unclear annotation of no CoD.

Figure 5. Types of errors between LLM vs. human annotation on the test dataset (errors per post)



The disparate information density across the data sources (see Figure 4) influenced the types of errors found within the annotations though the human annotator consistently had higher rates of agreement with the adjudicator than the computer annotations. Obituaries had low density of CoD information and very low error rates. The most common error made by the FSL algorithm was in annotation of a CoD that was not mentioned in the post (3.5%), whereas the human annotator missed a mentioned CoD in 1% of posts. For Memorial websites, both human and

FSL-LLM annotations exhibited a small number of errors. FSL-LLM annotations missed mentions of medical conditions in 4.5% of posts and attributed primary causes incorrectly in only 2% of posts, whereas human annotators had an error rate of less than 1% for any category. For GoFundMe, which regularly mentions multiple patient and non-patient conditions, the FSL-LLM model has similar error rates to human annotation, except for "Missed non-patient condition" (10.5%) and " Missed additional patient conditions" (31.5%) categories, indicating a performance gap compared to human annotations (1.5% and 3%, respectively) in identification of all potential medical conditions within the post though very low error rate in identification of the primary CoD.

Final Collected Records

After applying the cleaning filters and NLP techniques, we successfully identified and extracted mortality-related information from a substantial number of documents across various sources. Table 4 below provides a summary of the total documents retained from each source.

Table 4: Number of Documents with mortality-related information identified from Each Source

Source	**Total Documents**
GoFundme	23615
Memorial website (Tribute archive; Ever loved)	733754
Obituaries.com	7375229
Total	8,132,598

Discussion

We employed a novel approach to extract mortality data from online sources using transformer-based NLP models and few-shot learning with LLMs. Our analysis demonstrated the effectiveness of finetuned transformer-based NLP models in extracting mortality data from diverse online sources, showcasing their potential to enhance traditional data collection methods. We also developed a few-shot learning approach with LLMs to effectively identify primary CoD from online unstructured text data, achieving high agreement with human annotators. By

leveraging publicly available online data, our approach has the potential to supplement conventional mortality databases, facilitating a more timely, comprehensive, and granular understanding of population-level mortality trends and risk factors.

Our study is consistent with other published papers that uses social media data generally and obituary data specifically to improve the ability of health and healthcare research to accurately measure outcomes at the population level. For example, some studies have successfully used data from the Twitter platform to predict opioid overdose [37] and heart-disease mortality [38], outperforming traditional demographic and health risk factors in predicting mortality. Additional studies have used GoFundMe data to identify disease categories in 89,645 medical crowdfunding campaigns [39] and to identify factors associated with cancer fundraising success [40]. An additional set of studies has used a range of techniques in online obituaries specifically, including for automated surveillance of cancer mortality trends [41], extraction of kinship data for genetic research [42], and reporting of drug overdose [43].

Our study extends the existing literature by using transformer-based NLP models, which enhanced the extraction of key components of mortality data across public sources. Models such as RoBERTa, ALBERT, BERTweet, and BERT showed strong performance in handling unstructured data to extract decedent names (first and last), dates of birth, and dates of death, with RoBERTa achieving the highest micro-averaged F1-score of 0.88 (95% CI, 0.86-0.90).

For primary CoD identification, our FSL-LLM approach demonstrated high accuracy across all sources (GoFundMe: 95.9%, obituaries: 96.5%, memorial websites: 98.0%), approximating human annotator performance (97.9%, 99.0%, and 99.5% respectively). Detailed performance metrics revealed robust results for GoFundMe (precision=0.97, recall=0.95, F1=0.96) and memorial websites (precision=0.94, recall=0.98, F1=0.96). Obituaries achieved high accuracy, though the precision-recall pattern (precision=0.61, recall=1.0, F1=0.76) suggests potential for optimization in processing such data format. These findings demonstrate the model's effectiveness while highlighting opportunities for source-specific improvements

FSL-LLM demonstrated equivalent performance to human annotations for CoD identification across all sources; there remains room for further enhancement to identify potential contributing causes of death. The error analysis indicates that FSL-LLM exhibited higher error in categories

such as "Missed non-patient condition" and "Missed additional patient condition," whereas it exhibited very low rates of error in identifying primary CoD or appropriately classifying a note as having no specific CoD noted. This was primarily noted in GoFundMe data, as it was the only data source with significant posts containing more than one medical condition. Targeted improvements in the model's ability to identify non-patient conditions and additional potential contributing causes are necessary to reduce these errors. The observed variation in error rates underscores the need for data-specific tuning to optimize model accuracy across different sources. To further enhance the FSL-LLM's performance, focused finetuning on the identified error types and the integration of more diverse training datasets are recommended.

An additional finding was the low information density observed in the data from X platform (Twitter at data collection time) relative to the other data sources allowing linkage to specific persons. Absence of reliable person identification in the data hinders reliable patient linkage, an essential element in the augmentation of mortality information and subsequent integration into healthcare system. We therefore excluded Twitter data from the analysis, after the annotation phase.

Automated extraction of key mortality information from online sources has the potential to significantly improves traditional mortality databases, which often experience delays and incomplete data. This approach enables the timely collection of crucial details surrounding mortality such as decedent names, dates of birth, and dates of death, which could enable linkage to other healthcare data sources such as EHRs to facilitate clinical research. For instance, in studies monitoring medical product safety and effectiveness using insurance claims and EHR data such as in the FDA Sentinel system, mortality information from publicly available sources using approaches described here could allow investigators to study inferential questions regarding the impact of medical products on overall and cause-specific mortality.

Limitations

Despite promising results, this study has several limitations. First, social media data may not fully represent all population segments due to usage and sharing biases. Second, while the NLP pipeline achieved high accuracy, the inherent ambiguity and scarcity of specific CoD mentions in the source data with the resultant underdetermination of some portions of the targeted

information as represented by the human reference standard reviewers, the NLP system may still misclassify some data points. Finally, CoD identification from text remains challenging, often requiring an understanding of context and relationships between mentioned conditions. While the few-shot learning with the LLM algorithm performed well in identifying primary CoD, further work is needed to improve its ability to extract multiple contributing causes from individual posts.

Future Directions

At the population level, future research could focus on comparing CoD derived from online public data with those reported by official agencies. This comparison could help validate the accuracy and timeliness of online-sourced mortality information. If validated, such data could potentially provide near real-time insights into emerging mortality trends, particularly for rapidly spreading causes such as infectious diseases or environmental exposures.

The integration of online-sourced mortality data into existing surveillance systems would require careful validation against official records to ensure accuracy and reliability. This process would likely involve collaboration between researchers and public health agencies. Such collaborations could help develop protocols for effectively incorporating online data into public health surveillance and decision-making processes, potentially enhancing the speed and breadth of public health responses.

Conclusion

We have demonstrated a promising application of advanced NLP techniques, including transformer-based models and few-shot learning with LLMs, to extract critical mortality information and identify causes of death from diverse online public data sources. The successful development of an NLP pipeline and the strong performance of the few-shot learning algorithm highlight the potential of these approaches to address limitations in traditional mortality databases and improve the timeliness, comprehensiveness, and granularity of mortality monitoring. However, the study acknowledges several limitations, such as potential biases in online data representation and challenges in extracting multiple contributing causes of death. Future research should focus on validating the usefulness of these methods in real-world settings,

studying the correlation between online-derived causes of death and official records, and improving the integration of online data into public health surveillance systems. Addressing these challenges and opportunities will strengthen the application of advanced NLP techniques to online public data for enhancing mortality surveillance.

References

[1] N. S. Weiss, "All-cause mortality as an outcome in epidemiologic studies: proceed with caution," (in eng), *European journal of epidemiology,* vol. 29, no. 3, pp. 147-9, Mar 2014, doi: 10.1007/s10654-014-9899-y.

[2] K. M. Flegal, B. K. Kit, H. Orpana, and B. I. Graubard, "Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories: A Systematic Review and Meta-analysis," *JAMA,* vol. 309, no. 1, pp. 71-82, 2013, doi: 10.1001/jama.2012.113905.

[3] A. Berrington de Gonzalez *et al.*, "Body-Mass Index and Mortality among 1.46 Million White Adults," *New England Journal of Medicine,* vol. 363, no. 23, pp. 2211-2219, 2010, doi: 10.1056/NEJMoa1000367.

[4] B. Starfield, "Is US Health Really the Best in the World?," *JAMA,* vol. 284, no. 4, pp. 483-485, 2000, doi: 10.1001/jama.284.4.483.

[5] M. N. Mieno *et al.*, "Accuracy of death certificates and assessment of factors for misclassification of underlying cause of death," *Journal of epidemiology,* vol. 26, no. 4, pp. 191-198, 2016.

[6] M. S. Lauer, E. H. Blackstone, J. B. Young, and E. J. Topol, "Cause of death in clinical research: time for a reassessment?," *Journal of the American College of Cardiology,* vol. 34, no. 3, pp. 618-620, 1999.

[7] J. D. Hart *et al.*, "Improving medical certification of cause of death: effective strategies and approaches based on experiences from the Data for Health Initiative," *BMC medicine,* vol. 18, pp. 1-11, 2020.

[8] M. Ter-Minassian, S. S. Basra, E. S. Watson, A. J. Derus, and M. A. Horberg, "Validation of US CDC National Death Index mortality data, focusing on differences in race and ethnicity," *BMJ Health & Care Informatics,* vol. 30, no. 1, 2023.

[9] L. Antolini *et al.*, "Spontaneous ARIA-like events in cerebral amyloid angiopathy–related inflammation: a multicenter prospective longitudinal cohort study," *Neurology,* vol. 97, no. 18, pp. e1809-e1822, 2021.

[10] G. F. Riley, "Administrative and claims records as sources of health care cost data," *Medical care,* vol. 47, no. 7_Supplement_1, pp. S51-S55, 2009.

[11] E. R. Haut, P. J. Pronovost, and E. B. Schneider, "Limitations of administrative databases," *Jama,* vol. 307, no. 24, pp. 2589-2590, 2012.

[12] G. B. Taksler *et al.*, "Opportunities, pitfalls, and alternatives in adapting electronic health records for health services research," *Medical Decision Making,* vol. 41, no. 2, pp. 133-142, 2021.

[13] A. E. Aiello, A. Renson, and P. Zivich, "Social media-and internet-based disease surveillance for public health," *Annual review of public health,* vol. 41, p. 101, 2020.

[14] S. Abdullah and T. Choudhury, "Sensing technologies for monitoring serious mental illnesses," *IEEE MultiMedia,* vol. 25, no. 1, pp. 61-75, 2018.

[15] L. E. Charles-Smith *et al.*, "Using social media for actionable disease surveillance and outbreak management: a systematic literature review," *PloS one,* vol. 10, no. 10, p. e0139701, 2015.

[16] V. Dey, P. Krasniak, M. Nguyen, C. Lee, and X. Ning, "A pipeline to understand emerging illness via social media data analysis: case study on breast implant illness," *JMIR Medical Informatics,* vol. 9, no. 11, p. e29768, 2021.

[17] B. Chapman, B. Raymond, and D. Powell, "Potential of social media as a tool to combat foodborne illness," *Perspectives in public health,* vol. 134, no. 4, pp. 225-230, 2014.

[18] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the international AAAI conference on web and social media*, 2013, vol. 7, no. 1, pp. 128-137.

[19] D. Centola, "Social media and the science of health behavior," *Circulation*, vol. 127, no. 21, pp. 2135-2144, 2013.

[20] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 3267-3276.

[21] M. A. Al-Garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," *Journal of biomedical informatics*, vol. 62, pp. 1-11, 2016.

[22] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner, "Social media and mental health: benefits, risks, and opportunities for research and practice," *Journal of technology in behavioral science*, vol. 5, pp. 245-257, 2020.

[23] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal*, vol. 78, pp. 319-338, 2013.

[24] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic," *PloS one*, vol. 8, no. 12, p. e83672, 2013.

[25] A. Sarker *et al.*, "Utilizing social media data for pharmacovigilance: a review," *Journal of biomedical informatics*, vol. 54, pp. 202-212, 2015.

[26] M. A. Al-Garadi *et al.*, "Text classification models for the automatic detection of nonmedical prescription medication use from social media," *BMC medical informatics and decision making*, vol. 21, pp. 1-13, 2021.

[27] R. Thackeray, B. L. Neiger, A. K. Smith, and S. B. Van Wagenen, "Adoption and use of social media among public health departments," *BMC public health,* vol. 12, pp. 1-6, 2012.

[28] R. N. Vereen, R. Kurtzman, and S. M. Noar, "Are social media interventions for health behavior change efficacious among populations with health disparities?: A meta-analytic review," *Health communication,* vol. 38, no. 1, pp. 133-140, 2023.

[29] A. Gough *et al.*, "Tweet for behavior change: using social media for the dissemination of public health messages," *JMIR public health and surveillance,* vol. 3, no. 1, p. e6313, 2017.

[30] I. C.-H. Fung, Z. T. H. Tse, and K.-W. Fu, "The use of social media in public health surveillance," *Western Pacific surveillance and response journal: WPSAR,* vol. 6, no. 2, p. 3, 2015.

[31] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica,* vol. 22, no. 3, pp. 276-282, 2012.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[33] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692,* 2019.

[34] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942,* 2019.

[35] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English Tweets," *arXiv preprint arXiv:2005.10200,* 2020.

[36] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971,* 2023.

[37] S. Giorgi *et al.*, "Predicting US county opioid poisoning mortality from multi-modal social media and psychological self-report data," *Scientific reports,* vol. 13, no. 1, p. 9027, 2023.

[38] J. C. Eichstaedt *et al.*, "Psychological language on Twitter predicts county-level heart disease mortality," *Psychological science,* vol. 26, no. 2, pp. 159-169, 2015.

[39] S. S. Doerstling, D. Akrobetu, M. M. Engelhard, F. Chen, and P. A. Ubel, "A disease identification algorithm for medical crowdfunding campaigns: validation study," *Journal of Medical Internet Research,* vol. 24, no. 6, p. e32867, 2022.

[40] X. Zhang, X. Tao, B. Ji, R. Wang, and S. Sörensen, "The success of cancer crowdfunding campaigns: project and text analysis," *Journal of Medical Internet Research,* vol. 25, p. e44197, 2023.

[41] G. Tourassi, H.-J. Yoon, and S. Xu, "A novel web informatics approach for automated surveillance of cancer mortality trends," *Journal of biomedical informatics,* vol. 61, pp. 110-118, 2016.

[42] K. He, L. Yao, J. Zhang, Y. Li, and C. Li, "Construction of genealogical knowledge graphs from obituaries: Multitask neural network extraction system," *Journal of medical Internet research,* vol. 23, no. 8, p. e25670, 2021.

[43] K. Warren, "nd "From Death Notice to the Cyber Obit: The History of the Overdose Obituary."," *Unpublished manuscript. https://projects. iq. harvard. edu/files/historyopioidepidemic/files/katherine_warren_paper. pdf.*