

Augmenting Fact and Date of Death in Electronic Health Records using Internet Media Sources: A Validation Study from Two Large Healthcare Systems

Michele LeNoue-Newton^a; Mohammed al-Garadi^a; Kerry Ngan^b; Haritha Pillai^b; Ruth M. Reeves^a; Daniel Park^a; Dax M. Westerman^b; José J. Hernández-Muñoz^c; Xi Wang^c; Aida Kuzucan^c; Shirley V. Wang^b; Kueiyu Joshua Lin^b; Candace Fuller^d; Melissa McPheeters^e; Michael E. Matheny^{a,f}; Rishi J. Desai^b

^aDepartment of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; ^bDivision of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA; ^cOffice of Surveillance and Epidemiology, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD; ^dDepartment of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA; ^eRTI International, Research Triangle Park, NC, USA; ^fGeriatrics Research Education and Clinical Care Service, Tennessee Valley Healthcare System VA, Nashville, TN, USA

Correspondence:

Rishi J Desai, PhD

Associate Professor of Medicine,

Division of Pharmacoepidemiology and Pharmacoeconomics,

Department of Medicine, Brigham and Women's Hospital, Harvard Medical School,

1620 Tremont Street, Suite 3030-R, Boston, MA 02120, USA

Phone: 617-278-0932 | Fax: 617-232-8602 | Email: rdesai@bwh.harvard.edu

Counts:

Abstract words: 250

Main body of text words (excluding abstract, tables/figures, and references): 3,000

Figures: 2

Tables: 4

Abstract (250/250)

Objective: To evaluate the validity of death ascertainment from publicly available internet media (IM) sources by benchmarking against state and Federal vital statistics data for patients in two large healthcare systems from the US.

Methods: We extracted names and dates of birth and death from publicly available data—including obituaries and memorial websites—using previously developed natural language processing models. These data were probabilistically matched to electronic health records (EHRs) from Mass General Brigham (MGB) and Vanderbilt University Medical Center (VUMC) on first name, last name, and date of birth. Using reference standards from state vital statistics databases from MA, CT, and VT for MGB and the National Death Index (NDI) for VUMC patients, we reported positive predicted values (PPV) considering cases where dates of death from IM sources were within 7 days of the reference standard to be true positives. We also reported sensitivity of deaths ascertained from IM sources.

Results: When probabilistically matching 8.1 million deaths extracted from public data to 78,848 deaths observed in the reference standards across two sites, 30,607 (38.8%) matched exactly. A PPV of 98.2% for MGB and 98.9% for VUMC was observed for exact matches, while <6% for non-exact matches. Considering only the exact matches, IM sources led to an improvement in sensitivity of death capture by 24% in MGB and 18% in VUMC, compared to using EHRs alone for death ascertainment.

Conclusions: Using public information to augment mortality data increased capture of death meaningfully over reliance on EHR records alone.

Keywords: vital statistics, mortality, online media, data linkage

1.0 Background

In safety and effectiveness evaluations of medications, mortality is often an important outcome of interest to researchers, clinicians, and patients. However, information regarding deaths that occur outside of a clinical setting are frequently missing in real-world data sources such as electronic health records (EHRs) or insurance claims¹⁻³. Under-ascertainment of mortality can lead to bias in studies aimed at evaluating effectiveness or harms of given treatments or conditions³⁻⁸. Thus, substantial effort has gone into identifying alternate sources for death data to augment information available in claims or EHRs, by linking these data to external sources of information^{3, 9-11}.

In the United States, deaths are tracked at the state and national levels through vital statistics registries^{12, 13}. Reporting into electronic death registries occurs initially at the state-level usually within 24 to 48 hours with data then available for public use 1-2 years from the date of collection. Currently, data are not routinely linked between state death registries and EHRs or claims data. Several all-cause death databases exist, including the Social Security Death Index (SSDI), CenSoc^{14, 15}, the National Death Index (NDI), Veteran Administration data, and Center for Medicare Services (CMS) data¹⁶. Each of these databases have specific biases, delays in release, or are not readily available to researchers. CenSoc and SSDI data skew towards older populations and only contain individuals with social security numbers (SSN), which can lead to bias in racial-ethnic distributions^{14, 15, 17}. Missingness in SSDI has been exacerbated by regulatory changes made to the national SSDI in 2011 that prevent significant portions of state vital statistics data from being included^{18, 19}. NDI remains a highly accurate source of information but has significant time delays and is prohibitively expensive for large scale research studies and has some bias in linkage towards individuals with SSNs^{20, 21}.

Internet based resources have been used for public health surveillance, including pharmacovigilance studies of adverse events, but generally on a population rather than an individual level²²⁻²⁷. There are several publicly available data sources that could potentially be leveraged to augment death ascertainment. While cause of death information is largely unavailable outside of narrative notes, death certificates, and autopsy reports, the date of death has the potential to be detected through several mechanisms. Obituaries and memorial websites can offer near real-time updates on vital status^{26, 28-30}, especially if posted electronically where they can be easily accessed and processed using natural language processing (NLP) tools. Other sources of information on social media may also provide indications that an individual has died, including GoFundMe and similar sites on which family members or friends may share information.

Several prior studies have used on-line obituary data to extract fact and date of death;^{10, 11, 29-31} however, all of these studies were focused on cohorts of patients with specific diagnosis, for instance cancer or scleroderma. These prior studies also did not explore sources such as memorial websites and other social media in addition to obituaries. To that end, we aimed to provide health system level validation of death information retrieved using previously developed NLP models from various online sources. We linked NLP derived death information to EHRs from two large tertiary care medical centers using probabilistic linkage methodology and benchmarked against >78,000 deaths identified from state and Federal vital statistics data as the gold standard information.

2.0 Methods

2.1 Data Sources

2.1.1 Internet Media (IM) Data

Decedent names, birth dates, and dates of death were extracted from publicly available IM posts from Obituary.com, GoFundMe, and Everloved/TributeArchives. These data were compiled into a database to facilitate linkage with patient populations for analysis by both MGB and VUMC (Figure 1A). Data was extracted either from structured metadata or by using an XGBoost transformer model trained at VUMC based on manual annotations of 3150 documents (1050 from each source) by trained nurse annotators. Details on the training and models for extraction of mortality data from IM are described elsewhere³². The majority of the 8M+ IM records were derived from Obituary.com (91%) with some records from Everloved/TributeArchives (9%), and a small amount from GoFundMe (<1%) (Figure 1B). All records in the final IM cohort were required to have complete entries for first name, last name, date of birth, and date of death, with no missing data.

2.1.2 MGB Reference Cohort

The MGB reference cohort consisted of 65,139 deceased individuals from an MGB EHR-Medicare claims linked data asset. All of these deaths were recorded between Jan 1, 2015, and Dec 31, 2020, in one of 3 state vital statistic database (MA, CT, VT) and included patients who had ≥ 1 visit recorded in MGB EHRs within 180 days prior to death.

2.1.3 VUMC Reference Cohort

As it was not feasible to submit all VUMC patients within the eligibility cohort of individuals who had either an inpatient or outpatient encounter at VUMC between Oct 1, 2015 and Jan 1, 2020 (over 1M patients) to search for deaths records in NDI due to cost consideration, we submitted a subsample to NDI to obtain a “gold-standard” reference for deaths. Patients were initially stratified by availability of a recorded death date and year of last known contact in accordance with NDI submission protocols. We limited submission to those with last known contact dates or recorded death dates between Jan 1, 2019, and Dec 31, 2021, and to patients who had ≥ 1 visit in the year prior to the last date of contact recorded at VUMC. Over 300,000 patients met these criteria, 29,233 patients with recorded deaths and 278,576 patients with unknown vital status, from these we selected a subsample of 35,015 individuals. Patients with recorded deaths were oversampled to ensure adequate sample sizes for related studies using the cohort related to cause of death.

Of the 35,015 patients submitted, 12,395 patients had a death recorded within VUMC records, while the remaining 22,620 patients had an unknown vital status. Patients with unknown vital status were selected roughly equally from across the 3 submission years with approximately 7500 selected per year. Patients were selected randomly from each stratification grouping. NDI provided potential matches from the national registry along with match characteristics. We selected the highest probability match for each patient for assessment. We grouped potential matches based on the data elements matched (SSN, first name, last name, sex, and date of birth) and evaluated each group on concordance with recorded dates of death to establish deterministic match criteria that would result in the highest precision and recall. The following criteria were used to determine the final match cohort. Matches with a full SSN match and mismatches on a combination of two of the following elements: first name, last name, day of birth, month of birth, or year of birth. If DOB was the only mismatch, all 3 elements could be mismatched given that year of birth was within 10 years. If SSN was not a match, then all the following elements

needed to match exactly: first name, last name and DOB. Determination of the final matched cohort aligned criteria between institutions and established similar criteria for matching between NDI and EHR as used for IM to EHR.

2.2 Data Linkage and Validation

2.2.1 Linkage of IM Records to EHR Records

Both MGB and VUMC utilized SPLINK³³ to determine probabilistic matches between IM records and their reference cohort. First name, last name, date of birth, and state of residence were used to match patient records between sources. The algorithm was trained independently at each site with the full IM dataset and each institution's reference cohort using identical parameters. The training determined weights for each variable and allowed for calculation of matching weights and probabilities for all matches within the dataset. Initial threshold for potential matching was set as a probability of 0.80 and the highest probability match for each reference patient was selected as the best match for further analysis.

2.2.2 Validating Linkage of IM Records to EHR Records

To determine the probability cut-off that yielded the matches with highest accuracy, we manually reviewed the potential matches identified using the probabilistic matching algorithm for validity using stratified random sampling of three hundred matches grouped by probability scores separately at each institution. Accuracy for the probabilistic matching algorithm was defined as the proportion of records predicted by the algorithm as matches that were confirmed by manual reviewers. For the initial round of validation, we limited data to data elements available to the match algorithm and assessed whether the match was potentially the correct patient. The following criteria were used in the manual review for determination of a potential match: 1) Exact first name with DOB within 1-5 days, 2) Exact match on uncommon first name and last name with different DOB (e.g. common names like John Smith would not match if DOB was not a match), 3) manual reviewers were free to use their discretion to account for nicknames (e.g. Sue instead of Susan), typos, and misspellings.

Positive predicted values (PPVs) were calculated based on alignment of date of death derived from the IM sources and the date of death from the reference standard for each match. For these calculations, we considered cases where dates of death from IM sources were within 7 days of the reference standard to be true positive cases. We also reported results based on exact match of date of death with the reference standard.

2.3 Ascertainment of Augmentation

Augmentation of date of death data from IM sources was determined at MGB by reporting the sensitivity of capture of deaths from IM records and EHRs as a proportion of total deaths identified in the reference cohorts. To calculate comparable metrics at VUMC, we adjusted for the oversampling of known deaths in the NDI-matched reference cohort. To estimate the actual impacts of IM sources on augmentation of death data, we first determined the percentage of patients with unknown death status submitted to NDI where a matched death record was identified to estimate the number of patients within the VUMC EHR that have unrecorded deaths. We then calculated the percentage of our reference cohort that had deaths identified within IM sources but not recorded in the EHR to estimate the number of deaths that would be identified if all patients were matched to public online data.

3.0 Results

3.1 NDI Match Results for patients sampled from the VUMC Research Derivative database

The final VUMC cohort contained 13,709 deaths identified in the NDI “gold-standard” reference. Of these, 11,934 had a death recorded in the EHR and 1,775 deaths were identified within the unknown cohort of 22,620 submitted to NDI (Table 1). This translates to 7.8% of persons with unknown death status matching with a deceased record.

3.2 Summary of linkage of IIM Records to EHR Records

The initial probability cutoff set within the probabilistic matching algorithm was 0.80 which resulted in the vast majority of the reference cohort having a potential match identified. The highest probability match for each patient within the reference cohort was assessed. We subset these based on match probability and assessed each of these grouping for match validity to determine final match criteria (Table 2). Exact matches on first name, last name, and date of birth were identified for 37.2% of the MGB cohort and 46.5% of the VUMC cohort.

3.3 Validation results for linkage of IM Records to EHR Records

Potential matches stratified by probability were manually reviewed to determine if they were likely to be a true match with an emphasis on reviewing matches ranked as non-exact matches. Both institutions found that matches with high probability ($>.99998$) matched on names and DOB were deemed a “true” match, but that matches with lower probability only met the criteria in our review for a “true” match in less than 50% of the reviewed pairs (Tables 3). This is even with a broad definition of match which only required first name match and DOB within 5 days or an uncommon full name match. The results indicated that to reduce the level of false positive matches, the match criteria needed to require a full name and DOB match between IM and the EHR.

We further reported PPVs by probability groups based on alignment of dates of death (Table 4). Predicted matches from the probabilistic matching procedure that matched exactly on the first name, last name, and date of birth showed PPV of 94-99%. Predicted matches that did not exactly match on these variables showed PPV of $<6\%$, suggesting that these matches were likely inaccurate matches.

3.4 Augmentation and Evaluation of Death Data

3.4.1 Augmentation of Death Data at MGB

Among the 65,389 deaths in the MGB reference cohort, the sensitivity of capture was 33% for EHR alone ($n=21,595$) and 37% for IM records ($n=24,228$) when using exact match. Overall, IM records using exact matching identified an additional 15,661 deaths compared to the EHR alone, increasing the sensitivity from 33% to 57% (Figure 2).

3.4.2 Augmentation of Death Data at VUMC

Within our reference cohort of 13,709 patient, 6,379 (46.5%) had an exact match within IM sources. Of 11,934 with recorded death dates, 5,629 (47.2%) had IM records and of the 1,775 with unknown death status in the EHR, 750 (42.3%) had a record in IM sources (Supplementary Figures 1A, B). Given that our submission was not a random representative sample, we needed to extrapolate the actual rates of death

in our cohort population based on the NDI results and from that calculate the rates of recorded deaths and the rate of augmentation from IM sources. As mentioned, the sample VUMC submitted to NDI to obtain our reference standard was skewed for known deaths with 35% of the total submission having recorded dates of death within the EHR. Based on a rate of 7.8% deaths within our submitted unknown sample and the numbers of patients with unknown death status within the full 2019-2021 VUMC cohort, we can estimate an additional 21,729 death over that time for a total of 50,952 deaths (Supplementary Figure 1C). Using that as the basis for calculation, we estimate that 57.3% of the total 2019-2021 VUMC deaths are captured within the EHR. In our reference data, IM death data existed for 42.3% of deceased patients which would extrapolate to an additional 9191 deaths, an increase of 18.0% (Figure 3B), comparable to the 24% increase seen at MGB.

4.0 Discussion

In this validation study conducted across two large healthcare systems from the US, we observed substantial missing data in the recording of deaths within EHRs. We further observed that leveraging NLP models to extract death information from publicly available IM platforms can accurately identify an additional 18-24% of deaths not captured in EHRs.

Mortality is a key outcome in observational studies, making capture of accurate death information of high importance. Many observational research studies rely upon data collected during the course of standard clinical care and available within a subject's EHR record, but these sources have missing mortality information^{3, 9}. Missing mortality information consistently leads to overestimate of survival with increasing levels of missingness relating to an increased bias in survival estimates^{3, 5, 34}. Even low levels of missing data potentially have significant impacts when assessing long-term survival and prevalence³⁵. Low sensitivity of mortality data must be addressed in order for the data to be considered fit for purpose^{4, 5, 8, 36}.

The standard sources for death data outside of the EHR include state vital statistics departments and NDI. NDI is often considered a "gold-standard" source for death information, but can face about 2 years lag before death data are made available¹⁸. For state vital statistics, depending on the study question, investigators may need to petition for access to multiple state registries and some gaps in coverage may still exist if patients move between states. For these reasons, death information extracted using NLP from publicly available internet sources such as obituaries, memorial archives, such as Everloved/TributeArchives, and other social media sharing sites, like GoFundMe, may be a useful supplement to EHR recorded death information. To support others in utilizing these models for similar purposes, we have made our NLP models publicly available for future use³⁷.

Prior studies have attempted to use IM data to extract fact of death, but these have primarily focused on obituary data in small cohorts^{10, 11, 29}. Our study expands on the literature by providing by far the largest validation of death information extracted from online sources including obituaries, memorial websites, and other social media, against gold standard information from state vital files or NDI. Importantly, our findings are likely more generalizable than prior studies as we conducted a population-based validation using deaths from two large healthcare systems in the US without restricting to any specific subpopulation with specific conditions.

There are some caveats and limitations to the use of IM sources for death ascertainment that merit discussion. Our results indicate that a high level of stringency needs to be used in determination of

matches to avoid false positives and erroneous assignment of deaths, which is in line with prior work^{29, 38}. We further acknowledge that even after successfully extracting the information from social media and linking them with EHRs, the sensitivity of mortality capture remains suboptimal. Finally, there is little literature on possible bias when using online memorial websites and obituaries for research purposes, but the possibility of non-representativeness in the sample with deaths recorded in internet media sources cannot be ruled out given the cost-related barriers associated with publishing obituaries and reliable internet access.

Overall, IM sources have the potential of increasing coverage of mortality outcomes data by 18-24% over reliance on EHR records alone with the ability to obtain this information in near real-time. For timely real-world surveillance, IM sources offer a potential solution to capture mortality information without the need for release of any PHI data outside of institutional firewalls, as public IM data can be scraped and brought into a secure environment for matching with EHRs.

5.0 Funding Sources

This work was supported by the US Food and Drug Administration Sentinel Innovation Center Contract (master agreement No. 75F40119D10037 to Drs Matheny, Smith, Al-Garadi, Davis, and Desai).

Table 1: Stratification of patients meeting inclusion criteria, submission cohort stratification, and final matched cohort stratification,

Description	N	% of patient population	% of submission	% of matches
Total patients meeting inclusion criteria	307,799	100%		
Recorded death in EHR	29233	9.5%		
Unknown status in EHR	278576	90.5%		
Total Submitted to NDI	35015	11.4%	100%	
Recorded death in EHR	12395	4.0%	35.4%	
Unknown Status in EHR	22620	7.3%	64.6%	
Final Matches	13709	4.5%	39.2%	100%
Recorded death in EHR	11,934	3.9%	34.1%	87.1%
Deceased with no recorded death in EHR	1775	0.6%	5.1%	12.9%

EHR = Electronic Health Record

Table 2: Matching between EHR and Social Media Records using a probabilistic linkage algorithm,

Description	MGB Data			VUMC Data		
	N	% of matched	% of total MGB death cohort	N	% of matched	% of total VUMC death cohort
Total death in MGB reference data	65,139	-	-	13,709	-	-
Total deaths in publicly available data	8,132,598	-	-	8,132,598	-	-
Total patients matched with $p \geq 0.80$	56,485	100%	86.7%	13343	100%	97.3%
<i>Match Probabilities</i>						
<0.975 match probability	20,281	35.9%	31.1%	12	.09%	.09%
≥ 0.975 , not exact match	11,976	21.2%	18.4%	6952	52.1%	50.7%
Exact match	24,228	42.9%	37.2%	6379	47.8%	46.5%

MGB = Mass General Brigham, VUMC= Vanderbilt University Medical Center

Table 3: Validation of probabilistic linkage of Internet Media Records to EHR Records by manual review

Match probabilities	Total records	Records manually reviewed	Confirmed matches by manual review	% Accuracy	Observations from manual annotation
MGB data					
<0.975	20,281	200	89	45%	In most matches, two out of four fields matched (first & last name; first name & DOB; last name & DOB)
>=0.975, not exact	11,976	32	15	47%	Most matched exactly on first name and last name + some matched closely on DOB (within 30 days) and state
Exact	24,228	68	68	100%	Matched exactly on first name, last name, DOB, and state
Total (>=0.8)	56,485	300	172	57%	-
VUMC data					
<0.975	12	12	12	100%	Matched on first name and date of birth, mismatched on last name and state. DOD did not match reference.
>=0.975, not exact	6952	225	13	5.7%	Most matched first name and last name but mismatch on DOB with some matches on state. The accepted matches matched on last name and DOB with variant first names (nickname/typo).
Exact	6379	63	63	100%	Matched on first name, last name, DOB, and state
Total (>=0.8)	13343	300	88	29.3%	

DOB = Date of birth, DOD = Date of death

Table 4: Positive predicted values for deaths identified with internet media records based on alignment of dates of death of reference standard

Match probabilities	Total records	True positives based on dates of death within 7 days of the reference standard	Positive predicted Value (death date within 7 days)	True positives based on same date of death as reference standard	Positive predicted Value (same death date)
MGB data					
<0.975	20,281	121	0.5%	23	0.1%
>=0.975, not exact	11,976	312	2.6%	243	2.0%
Exact	24,228	23,802	98.2%	22,979	94.7%
VUMC data					
<0.975	12	0	0.0%	0	0.0%
>=0.975, not exact	6,952	382	5.6%	347	5.1%
Exact	6,379	6,301	98.9%	6,135	96.3%

REFERENCES

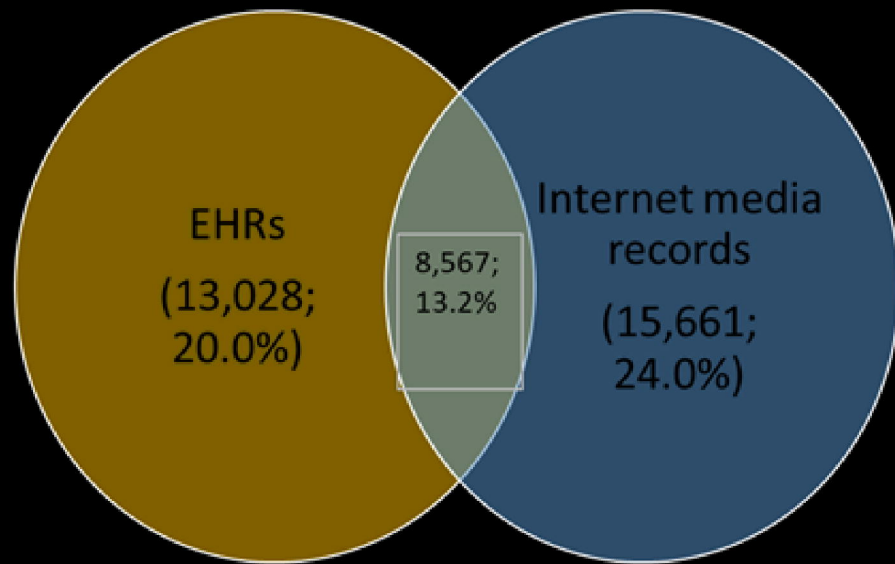
1. Hannan EL, Samadashvili Z, Cozzens K, Chikwe J, Adams DH, Sundt TM, 3rd, Girardi L, Smith CR, Lahey SJ, Gold JP, Wechsler A, Jordan D, Ashraf MH, Kurlansky PA. Out-of-Hospital 30-day Deaths After Cardiac Surgery Are Often Underreported. *Ann Thorac Surg*. 2020;110(1):183-8. Epub 20191109. doi: 10.1016/j.athoracsur.2019.09.061. PubMed PMID: 31715155.
2. Svartbo B, Olov Bygren L, Gunnarsson T, Steen L, Ribe M. Survival during and after hospitalization: a medical record linkage. *Int J Health Care Qual Assur Inc Leadersh Health Serv*. 1999;12(1):13-7. doi: 10.1108/09526869910249631. PubMed PMID: 10351019.
3. Curtis MD, Griffith SD, Tucker M, Taylor MD, Capra WB, Carrigan G, Holzman B, Torres AZ, You P, Arnieri B, Abernethy AP. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. *Health Serv Res*. 2018;53(6):4460-76. Epub 20180514. doi: 10.1111/1475-6773.12872. PubMed PMID: 29756355; PMCID: PMC6232402.
4. Desai RJ, Levin R, Lin KJ, Paterno E. Bias Implications of Outcome Misclassification in Observational Studies Evaluating Association Between Treatments and All-Cause or Cardiovascular Mortality Using Administrative Claims. *J Am Heart Assoc*. 2020;9(17):e016906. Epub 20200826. doi: 10.1161/jaha.120.016906. PubMed PMID: 32844711; PMCID: PMC7660765.
5. Carrigan G, Whipple S, Taylor MD, Torres AZ, Gossai A, Arnieri B, Tucker M, Hofmeister PP, Lambert P, Griffith SD, Capra WB. An evaluation of the impact of missing deaths on overall survival analyses of advanced non-small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiol Drug Saf*. 2019;28(5):572-81. Epub 20190314. doi: 10.1002/pds.4758. PubMed PMID: 30873729; PMCID: PMC6594237.
6. Butler CW, Snyder M, Wood DE, Curtis JR, Albert RK, Benditt JO. Underestimation of mortality following lung volume reduction surgery resulting from incomplete follow-up. *Chest*. 2001;119(4):1056-60. doi: 10.1378/chest.119.4.1056. PubMed PMID: 11296169.
7. Hsu WC, Crowley A, Parzynski CS. The impact of different censoring methods for analyzing survival using real-world data with linked mortality information: a simulation study. *BMC Med Res Methodol*. 2024;24(1):203. Epub 20240913. doi: 10.1186/s12874-024-02313-3. PubMed PMID: 39272007; PMCID: PMC11395225.
8. Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, Mulla S, Lamontagne F, Bassler D, Vera C, Alshurafa M, Katsios CM, Zhou Q, Cukierman-Yaffe T, Gangji A, Mills EJ, Walter SD, Cook DJ, Schünemann HJ, Altman DG, Guyatt GH. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *Bmj*. 2012;344:e2809. Epub 20120518. doi: 10.1136/bmj.e2809. PubMed PMID: 22611167.
9. Boyd JH, Randall SM, Ferrante AM, Bauer JK, McInnery K, Brown AP, Spilsbury K, Gillies M, Semmens JB. Accuracy and completeness of patient pathways--the benefits of national data linkage in Australia. *BMC Health Serv Res*. 2015;15:312. Epub 20150808. doi: 10.1186/s12913-015-0981-2. PubMed PMID: 26253452; PMCID: PMC4529694.
10. Lund JL, Cacciatore J, Tylock R, Su IH, Sharma S, Hinton SP, Smith S, Nowels MA, Chen X, Duberstein PR, Hanson LC, Mohile SG. Development and Evaluation of a Multisource Approach to Extend Mortality Follow-Up for Older Adults With Advanced Cancer Enrolled in Randomized Trials. *JCO Clin Cancer Inform*. 2024;8:e2300183. doi: 10.1200/cci.23.00183. PubMed PMID: 38564692; PMCID: PMC11000769.
11. Shao P, Tepsick JG, Walker B, Ray HE. Improving Real-World Mortality Data Quality in Oncology Research: Augmenting Electronic Medical Records With Obituary, Social Security Death Index, and Commercial Claims Data. *JCO Clin Cancer Inform*. 2023;7:e2300014. doi: 10.1200/cci.23.00014. PubMed PMID: 37695983; PMCID: PMC10569778.

12. Ventura SJ. The U.S. National Vital Statistics System: Transitioning Into the 21st Century, 1990-2017. *Vital Health Stat 1*. 2018;(62):1-84. PubMed PMID: 30248018.
13. Statistics. NRCUCoN, editor. The U.S. Vital Statistics System: The Role of State and Local Health Departments 2009; Washington (DC): National Academies Press (US).
14. Breen CF, Osborne M, Goldstein JR. CenSoc: Public Linked Administrative Mortality Records for Individual-level Research. *Sci Data*. 2023;10(1):802. Epub 20231115. doi: 10.1038/s41597-023-02713-y. PubMed PMID: 37968265; PMCID: PMC10651897.
15. Finlay K, Genadek KR. Measuring All-Cause Mortality With the Census Numident File. *Am J Public Health*. 2021;111(S2):S141-s8. doi: 10.2105/ajph.2021.306217. PubMed PMID: 34314212; PMCID: PMC8495642.
16. Mues KE, Liede A, Liu J, Wetmore JB, Zaha R, Bradbury BD, Collins AJ, Gilbertson DT. Use of the Medicare database in epidemiologic and health services research: a valuable source of real-world evidence on the older and disabled populations in the US. *Clin Epidemiol*. 2017;9:267-77. Epub 20170509. doi: 10.2147/clep.S105613. PubMed PMID: 28533698; PMCID: PMC5433516.
17. Huntington JT, Butterfield M, Fisher J, Torrent D, Bloomston M. The Social Security Death Index (SSDI) most accurately reflects true survival for older oncology patients. *Am J Cancer Res*. 2013;3(5):518-22. Epub 20131101. PubMed PMID: 24224129; PMCID: PMC3816971.
18. da Graca B, Filardo G, Nicewander D. Consequences for healthcare quality and research of the exclusion of records from the Death Master File. *Circ Cardiovasc Qual Outcomes*. 2013;6(1):124-8. doi: 10.1161/circoutcomes.112.968826. PubMed PMID: 23322808.
19. Navar AM, Peterson ED, Steen DL, Wojdyla DM, Sanchez RJ, Khan I, Song X, Gold ME, Pencina MJ. Evaluation of Mortality Data From the Social Security Administration Death Master File for Clinical Research. *JAMA Cardiol*. 2019;4(4):375-9. doi: 10.1001/jamacardio.2019.0198. PubMed PMID: 30840023; PMCID: PMC6484792.
20. Miller EA, McCarty FA, Parker JD. Racial and Ethnic Differences in a Linkage with the National Death Index. *Ethn Dis*. 2017;27(2):77-84. Epub 20170420. doi: 10.18865/ed.27.2.77. PubMed PMID: 28439177; PMCID: PMC5398181.
21. Cowper DC, Kubal JD, Maynard C, Hynes DM. A primer and comparative review of major US mortality databases. *Ann Epidemiol*. 2002;12(7):462-8. doi: 10.1016/s1047-2797(01)00285-x. PubMed PMID: 12377423.
22. Barros JM, Duggan J, Rebholz-Schuhmann D. The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review. *J Med Internet Res*. 2020;22(3):e13680. Epub 20200313. doi: 10.2196/13680. PubMed PMID: 32167477; PMCID: PMC7101503.
23. Singh L, Gresenz CR, Wang Y, Hu S. Assessing Social Media Data as a Resource for Firearm Research: Analysis of Tweets Pertaining to Firearm Deaths. *J Med Internet Res*. 2022;24(8):e38319. Epub 20220825. doi: 10.2196/38319. PubMed PMID: 36006693; PMCID: PMC9459834.
24. Wessel D, Pogrebnyakov N. Using Social Media as a Source of Real-World Data for Pharmaceutical Drug Development and Regulatory Decision Making. *Drug Safety*. 2024;47(5):495-511. doi: 10.1007/s40264-024-01409-5.
25. Klein AZ, Cai H, Weissenbacher D, Levine LD, Gonzalez-Hernandez G. A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes. *J Biomed Inform*. 2020;112s:100076. Epub 20200808. doi: 10.1016/j.yjbix.2020.100076. PubMed PMID: 34417007.
26. Tourassi G, Yoon HJ, Xu S. A novel web informatics approach for automated surveillance of cancer mortality trends. *J Biomed Inform*. 2016;61:110-8. Epub 20160401. doi: 10.1016/j.jbi.2016.03.027. PubMed PMID: 27044930; PMCID: PMC5048742.

27. Lavertu A, Hamamsy T, Altman RB. Quantifying the Severity of Adverse Drug Reactions Using Social Media: Network Analysis. *J Med Internet Res*. 2021;23(10):e27714. Epub 20211021. doi: 10.2196/27714. PubMed PMID: 34673524; PMCID: PMC8569532.
28. Buonanno P, Puca M. Using newspaper obituaries to "nowcast" daily mortality: Evidence from the Italian COVID-19 hot-spots. *Health Policy*. 2021;125(4):535-40. Epub 20210127. doi: 10.1016/j.healthpol.2021.01.006. PubMed PMID: 33579561; PMCID: PMC8529307.
29. Sylvestre E, Bouzille G, Breton M, Cuggia M, Campillo-Gimenez B. Retrieving the Vital Status of Patients with Cancer Using Online Obituaries. *Stud Health Technol Inform*. 2018;247:571-5. PubMed PMID: 29678025.
30. Soowamber ML, Granton JT, Bavaghar-Zaeimi F, Johnson SR. Online obituaries are a reliable and valid source of mortality data. *Journal of Clinical Epidemiology*. 2016;79:167-8. doi: 10.1016/j.jclinepi.2016.05.012.
31. Zhang Q, Gossai A, Monroe S, Nussbaum NC, Parrinello CM. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. *Health Serv Res*. 2021;56(6):1281-7. Epub 20210517. doi: 10.1111/1475-6773.13669. PubMed PMID: 33998685; PMCID: PMC8586476.
32. Al-Garadi M, LeNoue-Newton M, Matheny ME, McPheeters M, Whitaker JM, Deere JA, McLemore MF, Westerman D, Khan MS, Hernández-Muñoz JJ, Wang X, Kuzucan A, Desai RJ, Reeves R. Automated Extraction of Mortality Information from Publicly Available Sources Using Language Models. *medRxiv*. 2024.10.28.24316027. doi: 10.1101/2024.10.28.24316027.
33. Linacre R, Lindsay S, Manassis T, Slade Z, Hepworth T. Splink: Free software for probabilistic record linkage at scale. *International Journal of Population Data Science*. 2022;7. doi: 10.23889/ijpds.v7i3.1794.
34. Pinheiro PS, Morris CR, Liu L, Bungum TJ, Altekruze SF. The impact of follow-up type and missed deaths on population-based cancer survival studies for Hispanics and Asians. *J Natl Cancer Inst Monogr*. 2014;2014(49):210-7. doi: 10.1093/jncimonographs/lgu016. PubMed PMID: 25417234; PMCID: PMC4841164.
35. Dahm S, Barnes B, Kraywinkel K. Detection of missed deaths in cancer registry data to reduce bias in long-term survival estimation. *Front Oncol*. 2023;13:1088657. Epub 20230309. doi: 10.3389/fonc.2023.1088657. PubMed PMID: 36969013; PMCID: PMC10034313.
36. Krantz MJ, Kaul S. The ATLAS ACS 2–TIMI 51 Trial and the Burden of Missing Data: (Anti-Xa Therapy to Lower Cardiovascular Events in Addition to Standard Therapy in Subjects With Acute Coronary Syndrome ACS 2–Thrombolysis In Myocardial Infarction 51). *Journal of the American College of Cardiology*. 2013;62(9):777-81. doi: <https://doi.org/10.1016/j.jacc.2013.05.024>.
37. al-Garadi M, Park D, Westerman D, Reeves R, Matheny M. IC_DI5_Matheny. Sentinel Innovations Center. 2024. https://dev.sentinelssystem.org/projects/IC/repos/ic_di5_matheny.
38. Jewell A, Broadbent M, Hayes RD, Gilbert R, Stewart R, Downs J. Impact of matching error on linked mortality outcome in a data linkage of secondary mental health data with Hospital Episode Statistics (HES) and mortality records in South East London: a cross-sectional study. *BMJ Open*. 2020;10(7):e035884. Epub 20200707. doi: 10.1136/bmjopen-2019-035884. PubMed PMID: 32641360; PMCID: PMC7342822.



MGB Total = 65,139



VUMC Total = 50,952

