Comparative Merits of Available Mortality Data Sources for Clinical Research

Evan T Sholle, MS[1,3], Marcos A Davila, MS[1], Kristin Kostka, MPH[2], Sajjad Abedian, MS[1], Marika Cusick, MS[1,4], Spencer Krichevsky, MS[3,6], Jyotishman Pathak, PhD[3], Thomas R Campion, Jr., PhD[1,3,5]

[1]Information Technologies & Services Department, Weill Cornell Medicine, New York, NY; [2]IQVIA, Durham, NC; [3]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY;[4]Stanford University, Stanford, CA; [5]Department of Pediatrics, Weill Cornell Medicine, New York NY; [6]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY

Abstract

Obtaining reliable data on patient mortality is a critical challenge facing observational researchers seeking to conduct studies using real-world data. As these analyses are conducted more broadly using newly-available sources of real-world evidence, missing data can serve as a rate-limiting factor. We conducted a comparison of mortality data sources from different stakeholder perspectives – academic medical center (AMC) informatics service providers, AMC research coordinators, industry analytics professionals, and academics - to understand the strengths and limitations of differing mortality data sources: locally generated data from sites conducting research, data provided by governmental sources, and commercially available data sets. Researchers seeking to conduct observational studies using extant data should consider these factors in sourcing outcomes data for their populations of interest.

Introduction

Recent changes in US legislation have spurred a significant degree of interest in the availability, accuracy, and scalability of what is often referred to as "real-world evidence" (RWE), a body of scientific knowledge generated by analysis of "real-world data," or "data that is derived from medical practice among heterogeneous sets of patients in real life practice settings, such as insurance claims data and clinical data from electronic health records," as opposed to data generated through randomized controlled trials (RCTs) (1). However, the use of these data sources comes with its own challenges, including data quality and availability, requiring researchers to carefully plan their strategy for data acquisition in order to ensure a methodologically valid and rigorous study. In some cases, securing high-quality real-world data is relatively straightforward. Researchers seeking to quantify the effect of certain somatic variants can secure funding to conduct next-generation sequencing on more cancer patients (3), develop natural language processing pipelines to extract genomic data from pathology reports already extant in the electronic health record (4), or partner with labs to generate a direct feed of structured genomic data.

However, the most important outcome for any clinical research study, without which one cannot construct a Kaplan-Meier curve or assess a five-year survival rate, is death. Despite the critical value of accurate, reliable information on who has expired when and from what cause, substantial obstacles remain for any researcher seeking to obtain this information and integrate it into a research study (8). Mortality data is difficult to obtain both on the population scale and at the individual patient level for a number of different factors, including regulatory concerns. Accordingly, the demand for this data has often yielded bizarre scenarios, such as research assistants conducting Google searches for the obituaries of patients enrolled in trials but lost to follow up, attempting to match pictures from funeral home websites to blurry snapshots stored in the electronic health record (9). To compensate for these difficulties, researchers have developed a number of techniques for specific studies. Eisenstein et al, for example, developed a hybrid "death identification and verification approach" for a specific trial, the ToRsemide compArisoN with furoSemide FOR Management of Heart Failure (TRANSFORM-HF), involving site-generated mortality data, data from the Social Security Adminstration, and the dedicated efforts of call center personnel – a resource unfortunately not available to many investigators seeking to conduct retrospective observational research (10). To mitigate the difficulty of obtaining this critical component of many research studies, we sought to provide an overview of the different forms of available information on patient-level mortality, including healthcare provider-generated mortality data, governmentally-generated mortality data, and commercially available sources of data on mortality, including data derived from insurance claims. We additionally sought to characterize each source of mortality data according to its strengths and

weaknesses, hypothesizing that a "fit for purpose" approach may lead to researchers preferring one data source to another for a specific use case.

Methods

To appropriately characterize the strengths and weaknesses of the three principal categories of death data available for clinical research, we conducted a comparison of these categories from multiple stakeholder perspectives. Academic medical center informatics service providers (ES, SA, MC, MD) conducted an analysis of local site-generated mortality data, given their experience extracting data from clinical and operational systems to support the research enterprise. An academic medical center research coordinator (SK) with experiencing using the National Death Index (NDI) to support clinical and translational research, along with an informatics service provider with experience extracting data from governmental sources (SA) conducted an analysis of available death data from governmental sources. An industry analyst (KK) employed by a contracted research organization (CRO) provided a perspective on commercially available sources of mortality data to support research. Academic informaticians (JP, TC), provided overall feedback on the relative strengths and weaknesses of each data source with an eye towards their larger context within the ecosystem of clinical research.

Results

Local site-generated mortality data

Researchers situated within academic medical centers or other healthcare provider organizations may turn first to sources of mortality data gathered at their specific institution. Individual healthcare organizations have access to a multitude of sources where mortality data is recorded, including structured elements within electronic health record systems, billing systems, and unstructured physician notes. Local-site generated mortality data is often precise, as providers will record patient deaths that occur at the point of care. However, patient deaths that take place outside of the organization may go undocumented, as there is no financial or legislative mandate to capture these deaths in a structured fashion. In order to address this gap, researchers may also seek to augment existing site-specific data sources with techniques such as natural language processing (NLP) on unstructured physician notes and patient death inference.

Structured Elements in Electronic Health Record (EHR) Systems

Within EHR systems, a patient is considered to be deceased if they have a death date and time populated within their record. While many institutions have a single shared EHR for their inpatient and outpatient setting, some institutions continue to use disparate systems. Because patient deaths are more likely to occur at the point of care in the inpatient setting, documentation of death data is more frequent within inpatient EHR systems. If there is no transfer of data between the inpatient and outpatient setting, the death dates may not be populated in patient records that exist in the outpatient EHR system. A recent study on the documentation of mortality data within a large academic medical institution concluded that less than 1% (33,295) of the 4 million patients that exist in a clinical data warehouse with both inpatient and outpatient data were documented as deceased. (11). This statistic is seemingly unlikely, as 43.6% and 43.1% of patients aged over 100 and 120, respectively, are documented to still be alive. Despite the gaps in data, these deaths and the dates associated with them are generally accurate, especially for in-hospital mortality. Other structured data elements contained within electronic health records can also be used to infer mortality status – for example, a patient without a recorded death date may, nonetheless, have a documented order for an autopsy, a procedure for which the subject is generally deceased.

Augmenting Structured EHR Mortality Data

To supplement existing structured data contained in electronic health record systems, researchers can rely on two principal techniques: natural language processing and imputation algorithms.

In the event of death at the point of care, physicians often document the date and time of death as well as the reason for death within a note. In some cases, this is easily detectable by filtering on document names, such as "Death Note." If using this technique, it possible to complete the death date as the authoring date and time, making the assumption that the provider authored the note shortly after the patient's death. However, these patients are likely to already have

their mortality status documented in a structured fashion in the EHR, rendering this technique of limited utility. More frequently, patients may be discharged from the hospital to hospice care or home, where they then die. This may be represented in a free text note representing a telephone encounter with a caregiver or relative, where the provider indicates that the patient has died. While rendering this data structured may seem to be tractable with the application of regular expressions and keywords, there are additional layers of complication, such as conditional mentions of death (e.g. treatment or medicine complication) or unrelated mentions of death (e.g. death of a family member or close friend). While NLP algorithms have proven success in extracting health information from unstructured clinical text (12), they remain an imperfect method of assessing overall patient mortality.

In addition to relying on natural language techniques, researchers can also conduct imputation of mortality based on existing information about the patient at the local site. These imputation algorithms can often range in complexity in terms of number of predictors and technique. At the simplest, one can infer a patient's death based on the patient's age and length of time since the patient's last encounter at the local site. However, there are many issues with this approach. Since life expectancy varies based on a number of demographic factors, it is impossible to determine a universal cutoff point beyond which one can assume a patient to be deceased. Furthermore, patients may be lost to follow-up due to factors other than death, such as relocation to a long-term care facility. Imputed mortality also suffers from the inability to determine a precise date of death, which limits its utility in the construction of survival curves and other analytic endpoints.

More complex imputation algorithms use probabilistic statistical models and a large number of predictor variables. In a recent study, researchers evaluated a patient-level model to predict whether the end of observations was due to death using US claims data. Using nearly 90,000 predictor variables, the model achieved reasonable performance, with a sensitivity of 62% and a positive predictive value of 74.8% (13). However, this approach relies on the availability of claims data, which limits its utility for researchers based at academic medical centers, who may not have access to such data sets. Accordingly, others have focused on an approach that predicts mortality at the patient cohort level rather than patient-level. Researchers at the University of Texas Anderson Cancer Center used a Bayesian approach with predictors from encounter and billing data in order to predict patient counts for a simulated patient population and were able to achieve error rates as low as 2.1% (14). While this technique may not be entirely effective for determining the death of a specific patient, it may be of use in increasing the accuracy of cohort discovery methods, improving clinical study feasibility analysis and planning.

Governmental data

National Vital Statistics System (NVSS)

The National Center for Health Statistics (NCHS), a division of the Centers for Disease Control and Prevention (CDC), formed in 1960 under the auspices of the US federal government, is the nation's primary body for the collection and dissemination of statistical health information. NCHS runs four major programs: the National Vital Statistics System (NVSS), the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHNES), and the National Health Care Surveys (NHCS), as well as other miscellaneous studies. NVSS is an inter-governmental data sharing system collecting data on births, deaths, marriages, divorces, and fetal deaths. The data is obtained from several programs under various jurisdictions nation-wide. While the publicly available data is limited and de-identified, NVSS offers a restricted-use "micro-data" set containing geographical variables about the decedent (15).

Although this dataset is fairly comprehensive, using this dataset to track mortality outcomes for each individual may not be a viable option. The exact date of death is only limited to researchers in federal agencies and their on-site contractors, with only a single exception: non-federal researchers may access the information through NCHS Research Data Centers. In addition to the relative difficulty of securing access to this data, its latency may serve as a hindering factor for studies requiring more up-to-date mortality information. The detailed mortality data is compiled 11-12 months after the end of a year; for example, the 2018 data was released in December of 2019 (16).

Social Security Death Master File (SSDMF)

Numident, or Numerical Identification System, is the Social Security Administration's (SSA) database containing all information recorded in each application for a United States Social Security number since their first issuance in 1936. The Social Security Death Master File (SSDMF) is a subset of the Numident database available through the

Department of Commerce's National Technical Information Service (NTIS). SSDMF provides several data points, such as Social Security Number (SSN), name, date of birth and death, as well as geographical identifiers for over 85 million individuals. NTIS offers weekly, monthly, and quarterly updates to the database. To reduce inaccuracies and keep local copies in line with the ever-growing file, users are strongly suggested to subscribe to at least one of the update schedules (17).

There two principal methods for accessing the SSDMF. In the fee per query option, the user is provided with the current version of the file accompanied by software to search and export queries. However, both users and queries incur financial cost: an annual subscription for one user with a 1 million query limit is $12,000, making large scale queries excessively expensive, especially for researchers without extensive funding allocated for data support. This option also does not afford the user access to the complete Death Master File. The second option allows the user to make a one-time purchase of the current dataset. However, users are required to subscribe to the weekly, monthly, or quarterly refresh package. In addition, the user must develop an application to read and query the dataset. Users with large scale queries may benefit from this option although it may require more effort to read, query, and maintain the dataset (18). The SSDMF is also prone to data quality issues corollary to existing accuracy problems relating to the capture of SSNs in the electronic health record – since the only linkage point between individual patients and the SSDMF is the SSN, patients with spurious or absent SSNs may not have death dates correctly associated with their records in a research database. Other research has also indicated that the SSDMF is prone to larger-scale data quality issues, suggesting that it "markedly underestimate[s] mortality rates, with variable undercapture among states and over time." (19)

National Death Index (NDI)

The National Death Index (NDI) is a centralized database housing death record information aggregated from state vital statistic agencies. The NDI was established by the National Center for Health Statistics (NCHS) to provision mortality data to epidemiologists, medical investigators, and other allied health professionals. Death records are available from 1979-2018 and the database is updated annually.

An Institutional Review Board (IRB) approved protocol supporting the need for NDI death data is required to initiate data transmission with a requesting institution. In addition, a comprehensive application form is completed and submitted, then reviewed by a 12-person panel. All necessary precautions are taken to ensure that protected health information (PHI) is transmitted between entities in a manner that complies with Health Insurance Portability and Accountability Act (HIPAA) policies and regulations.

Data transmission requests are structured according to NDI-mandated formats, which allocate a pre-determined number of character positions per identifying feature. Identifying features include last and first name, middle initial, social security number, date of birth, father's surname, suspected age at death, sex, race, marital status, state of residence, state of birth, and other optional data points. The example below illustrates these requirements:

```
12345678901234567890|123456789012345|6|789012345|67|89|0123|456789012345678901|2|34|5|6|7|89|01|2345678901|234567|890|
Doe                 |John           |A|111111111|10|31|1900|Doe                 |0|80|M|2|2|NY|NJ|          |      |   |
```

This string of identifiers is then parsed and used to match potential records to death data housed in the NDI database. Each requested record must contain information that satisfies ≥1 of the following criteria:

- Exact month and ±1 year of birth; first and last name
- Exact month and ±1 year of birth; initials of the first and middle name; last name
- Exact month and day of birth; first and last name
- Exact month and day of birth; initials of the first and middle name; last name
- Exact month and year of birth; first name; father's surname
- If female: exact month and year of birth; first name and last name; father's surname

Each NDI potential match record is attributed a probabilistic score where each weight is a base-2 logarithm of the inverse of the joint probability of occurrence. The score for each potential match is the sum of the weights for each identifying characteristic:

$$\boldsymbol{Score} = W_{SSN} + W_{firstname*sex*birthyear} + W_{middlenameinitial*sex} + W_{lastname} + W_{race} + W_{sex} + W_{martialstatus*sex*age} + W_{DOB} + W_{birthmonth} + W_{birthyear} + W_{birthstate} + W_{residencestate}$$

For example, males constitute 48.3% of the population aged 18 and over, so the corresponding weight is $\log_2(1/0.483)=1.05$.

To transmit data, the requesting institution generates a self-encrypted CD-ROM containing files with identifying characteristics for the entire population of interest. These CD-ROMs are sent to NCHS, by overnight delivery, reviewed, and overwritten with both accepted and rejected record matches. They are then returned via mail to the requesting institution.

While the NDI offers significant sophistication and access to a large degree of mortality data that might otherwise be unavailable to researchers, the costs associated with accessing and utilizing NDI data are not insignificant, and include a $350.00 initial service charge and a $0.15 per user record per year searched fee. For example, 500 records searched from 2008-2018 would cost $350 + ($0.15 * 500 * 10) = $1,100.00. Recent guidance from the National Institutes of Health (20) has indicated that NIH-supported investigators are eligible to receive reimbursement from NCHS for linkage of local research databases to the NDI, as long as the local research database is associated with the research aims funded by NIH. This may serve to render the NDI a more accessible resource for investigators without extensive funding who still need accurate mortality data beyond what is typically available within the confines of a local site. However, requesters are still limited to four requests per year, and any request in excess of $100,000 may necessitate approval by the NIH before submission of the request.

Commercially available death data

In addition to site-specific mortality data and mortality data aggregated by federal agencies, researchers can also avail themselves of a variety of syndicated, de-identified patient-level real-world data assets available for commercial licensing. A list of popular sources of US data sources routinely licensed by the pharmaceutical industry in real-world evidence generation studies are provided in Table 1. While many of these data sources draw from federally-aggregated data sets and local electronic health record data, they address some of the limitations of the previously mentioned data sources by aggregating and integrating data from multiple disparate sources, increasing the likelihood of obtaining high-quality data.

Table 1. Commonly used commercially available real-world data sources

Data source	**Type of data**
Optum© de-identified EHR Dataset (PANTHER)	Integrated claims, EHR data
Optum© de-Identified Clinformatics® Data Mart	Claims, lab results, confinement data, provider data
IBM® MarketScan® Commercial Claims and Encounters (CCAE)	Commercial insurance claims
IBM® MarketScan® Multi-State Medicaid Database (MDCD)	Medicaid claims
IBM® MarketScan® Medicare Supplemental Database (MDCR)	Medicare claims
IQVIA® Longitudinal Patient-Level Prescriptions (LRx) and Medical Claims (Dx)	Pre-adjudicated claims
IQVIA® Hospital Charge Data Master (CDM)	Hospital charge data from short-term, acute care, and non-federal hospitals
IQVIA® Real World Data Adjudicated Claims US (formerly known as PharMetrics Plus)	Adjudicated claims
IQVIA® Oncology EMR	EMRs data from community-based, oncology practices
IQVIA® Ambulatory Electronical Medical Records (AEMR)	EMR data from over 70,000 physicians (50% primary care)
Quest Diagnostics	Lab data from over 2,200+ centers that perform 3,500 tests

Cerner® HealthFacts Data Warehouse	De-identified EHR data

Despite the ease of procuring de-identified US patient data, patient death may be sparsely documented in patient records or medical claims sold for secondary use. Discharge status and certain diagnosis codes can be used to deduce mortality status of a patient (21). Optum® is the only known vendor to offer a linked real-world data asset with date-of-death (DOD) with its Clinformatics data mart offering. However, prior research suggests that relying on other a linkage with public SSA death master file (DMR) may still under-identify deaths (11).

As many data owners seek to maintain HITRUST certification, inclusive of HIPAA regulations, there are additional mandates the redaction of personally identifiable information including death information that could be potentially re-identifiable. Thus, as commercial vendors de-identify patient data for syndication, they often struggle with the ability to provide complete information on how patients leave their data. In some engagements, commercial vendors may offer limited, bespoke querying of a database to answer a research question behind their own firewalls. In these circumstances, only aggregate results adhering to a minimum cell size can be shared. As the emphasis for preserving patient privacy increases, resellers of de-identified patient-level data will continue to face challenges in their ability to appropriate capture death information. Additionally, the costs associated with obtaining this data may render it prohibitive for individual use cases, as investigators not already affiliated with an institution that has secured access to these data sets may find it impracticable to obtain access for a specific study.

Conclusion

The undeniable potential of real-world data to provide timely, accurate, and comprehensive evidence on the safety and efficacy of various treatment modalities is limited only by the data quality issues intrinsic to its nature as a secondary outcome of existing clinical and organizational workflows. In order to provide support for real-world data applications, researchers need accurate, comprehensive, and high-quality information not only about patient exposures, but also about patient outcomes, including death.

Here we have summarized the strengths and weaknesses of three primary categories of death data available to researchers seeking to make use of real-world data for observational research. Site-specific mortality data is likely to be of high quality – especially in the case of in-hospital mortality, where researchers may even be able to accurately assess or obtain the proximate cause of death. However, what it gains in depth it sacrifices in breadth – patients who die outside the care of the healthcare organization are not likely to have their death reflected in these data sources. Governmental data sources offer the potential to address these issues by relying on centralized data capture methodologies, including those maintained by federal mandate, but suffer from their own data quality issues and may be unaffordable for investigators without significant funding support. Finally, commercially available data sources aggregate data from local electronic health records, claims data sets, and others to provide a maximally comprehensive portrait of patient mortality – however, the cost associated with securing access to them may be out of reach for many investigators outside the confines of the organizations that generated them.

References

1. Network for Excellence in Health Innovation. Real World Evidence: A New Era for Health Care Innovation [Internet]. NEHI. 2015 [cited 2020 Mar 24]. Available from: https://www.nehi.net/publications/66-real-world-evidence-a-new-era-for-health-care-innovation/view
2. Moore TJ, Zhang H, Anderson G, Alexander GC. Estimated Costs of Pivotal Trials for Novel Therapeutic Agents Approved by the US Food and Drug Administration, 2015-2016. JAMA Intern Med. 2018 Nov 1;178(11):1451–1457.
3. Song J, Hussaini M. Adopting solutions for annotation and reporting of next generation sequencing in clinical practice. Practical Laboratory Medicine. 2020 Mar;19:e00154.
4. García-Remesal M, Maojo V. Integration of Relational and Textual Biomedical Sources. … of information in …. 2010;
5. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science. 2016 Dec 23;354(6319).

6. Sholle E, Krichevsky S, Scandura J, Sosner C, Campion TR. Lessons learned in the development of a computable phenotype for response in myeloproliferative neoplasms. IEEE Int Conf Healthc Inform. 2018 Jun;2018:328–331.
7. Karystianis G, Nevado AJ, Kim C-H, Dehghan A, Keane JA, Nenadic G. Automatic mining of symptom severity from psychiatric evaluation notes. Int J Methods Psychiatr Res. 2018;27(1).
8. Curtis MD, Griffith SD, Tucker M, Taylor MD, Capra WB, Carrigan G, et al. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. Health Serv Res. 2018 May 14;53(6):4460–4476.
9. Huang K. Difficulties of clinical and translational research. 2020.
10. Eisenstein EL, Prather K, Greene SJ, Harding T, Harrington A, Gabriel D, et al. Death: the simple clinical trial endpoint. Stud Health Technol Inform. 2019;257:86–91.
11. Jones B, Vawdrey DK. Measuring mortality information in clinical data warehouses. AMIA Jt Summits Transl Sci Proc. 2015 Mar 25;2015:450–455.
12. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009 Oct;42(5):760–772.
13. Reps JM, Rijnbeek PR, Ryan PB. Identifying the DEAD: Development and Validation of a Patient-Level Model to Predict Death Status in Population-Level Claims Data. Drug Saf. 2019;42(11):1377–1386.
14. Myers RB, Herskovic JR. Probabilistic techniques for obtaining accurate patient counts in Clinical Data Warehouses. J Biomed Inform. 2011 Dec;44 Suppl 1:S69–77.
15. Centers for Disease Control and Prevention. About the National Vital Statistics System [Internet]. National Center for Health Statistics. 2016 [cited 2020 Mar 25]. Available from: https://www.cdc.gov/nchs/nvss/about_nvss.htm
16. Centers for Disease Control and Prevention. Restricted-Use Vital Statistics Data [Internet]. National Center for Health Statistics. 2019 [cited 2020 Mar 25]. Available from: https://www.cdc.gov/nchs/nvss/nvss-restricted-data.htm
17. United States Department of Commerce. Limited Access Death Master File [Internet]. National Technical Information Service. 2020 [cited 2020 Mar 25]. Available from: https://classic.ntis.gov/products/ssa-dmf/#
18. United States Department of Commerce. Limited Access Death Master File Available Through Value-Added Online Products [Internet]. National Technical Information Service. 2020 [cited 2020 Mar 25]. Available from: https://classic.ntis.gov/products/ssa-online/
19. Navar AM, Peterson ED, Steen DL, Wojdyla DM, Sanchez RJ, Khan I, et al. Evaluation of mortality data from the social security administration death master file for clinical research. JAMA Cardiol. 2019 Apr 1;4(4):375–379.
20. Office of Behavioral and Social Sciences Research. Notice of Information: National Death Index Linkage Access for NIH-Supported Investigators [Internet]. Grants.gov. 2020 [cited 2020 Mar 25]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-057.html
21. Friends of Cancer Research. Establishing a Framework to Evaluate Real-World Endpoints [Internet]. Friends of Cancer Research. 2018 [cited 2020 Mar 25]. Available from: https://www.focr.org/sites/default/files/RWE_FINAL%207.6.18.pdf?eType=EmailBlastContent&eId=45c28471-3adc-4f10-baa5-2b1181858d97